

Line Outage Localization using Phasor Measurement Data in Transient State

Manuel Garcia, Thomas Catanach, Scott Vander Wiel, Russell Bent, and Earl Lawrence

Abstract—This work introduces a statistical classifier that quickly locates line outages in a power system utilizing only time series phasor measurement data sampled during the system’s transient response to the outage. The presented classifier is a linear multinomial regression model that is trained by solving a maximum likelihood optimization problem using synthetic data. The synthetic data is produced through dynamic simulations which are initialized by random samples of a forecast load/generation distribution. Real time computation of the proposed classifier is minimal and therefore the classifier is capable of locating a line outage before steady state is reached, allowing for quick corrective action in response to an outage. In addition, the output of the classifier fits into a statistical framework that is easily accessible. Specific line outages are identified as being difficult to localize and future improvements to the classifier are proposed.

Index Terms—Estimation, transient response, power system faults, uncertainty

I. INTRODUCTION

A recent increase in power load and the rapid integration of renewable energy into the power grid necessitates an improvement in situational awareness. State estimation has been targeted as one of the causes of most of the major blackouts in North America [1], as the system operator makes crucial decisions based on the possibly incorrect estimate of the state. This manuscript addresses state estimation by providing a line outage localization procedure that utilizes time series data provided from phasor measurement units (PMUs) during the system’s transient response to the outage. The proposed classifier requires minimal real time computation and accurately quantifies uncertainty by providing a probability distribution over all potential line outages.

The state estimation process begins with the Network Topology Processor which estimates the static parameters of the system model including line statuses [2]. This static model of the system is then assumed to be accurate while it is utilized to compute the analog state estimate via methods such as Weighted Least Squares (WLS) [2]. For this reason, an incorrect estimate of the network topology will cascade through the state estimation process, making topology estimation a crucial part of generalized state estimation. The classical method of determining the network topology involves estimating the analog state for every possible network topology and then labeling the topology that corresponds to the smallest measurement residual as the true topology of the system. However, more computationally tractable approaches for topology estimation have been proposed in literature [3], [4].

Some recent work improves topology estimation by providing procedures that detect and localize line outages in the system using PMU data [5]–[8]. The cited work addresses

the issue of both change point detection and localization of line outages. Although the change point detection methods do utilize time series PMU data sampled during the transient response of the system, the localization methods do not. Instead the localization procedures utilize only measurement data sampled at a steady state and rely heavily on the DC approximations [9]. In fact, most line outage classification techniques utilize steady state assumptions and inherit two fundamental drawbacks:

- 1) *Information Provided by Transients*: Steady state based outage localization procedures do not utilize the information provided by PMU measurement data during the transient response of the system. This information may be used to improve the performance of a localization procedure.
- 2) *Slow Acting Classification*: Severe line outages may cause instability in the system and as a result a steady state may not be reached until equipment failures have occurred. Classification techniques that are based on steady state assumptions cannot begin their execution until a steady state is reached and thus cannot provide information used for corrective action in the presence of an unstable outage.

This manuscript provides a localization procedure that corrects these two fundamental drawbacks. However, the procedure presented assumes knowledge of more information about the system than do standard steady state based classification methods. In addition to using PMU data, system topology data, and transmission line and transformer parameter data, dynamic generator and motor data are required. These additional data are necessary to perform the simulations that provide us with synthetic observations to train the classifier. In addition, we assume that a forecast distribution of system load and system generation is available as well as a list of all possible line outages.

To effectively utilize time series data that is sampled during the transient response of the system, two estimation problems must be addressed. First, change point detection must accurately identify the time at which a disturbance occurs in the system. Second, a classification method must accurately identify the disturbance that occurred. This manuscript does not address change point detection, which has been studied in other works [5], [10]. Instead we assume that change point detection is capable of providing accurate results, and we develop a classification method that localizes line outages in the system.

Plenty of recent work has focused on developing state estimation procedures in statistical frameworks, providing

probability distributions that are easily accessible. By placing our localization procedure in a clear statistical framework, the presented classifier can be merged with existing estimation procedures that rely on steady state assumptions effectively improving steady state-based classifiers [8], [11].

The classification procedure presented in this paper is perhaps most similar to the work done by [11]. Both techniques quantify uncertainty in the system topology and utilize forecast load/generation distributions which are assumed to be available a priori. However, localization procedures provided by [11] utilize steady state assumptions.

It is important to note that our methods can easily be extended to the classification of any type of disturbance in the system. However, the presented classifier may only consider a finite number of disturbances and cannot account for every disturbance in the system. This work considers only single-line outages, allowing our results to be compared to existing line outage localization methods. Future work will consider multiple types of disturbances.

II. CLASSIFICATION METHODS: MULTINOMIAL REGRESSION

Consider a transmission level power system with m transmission lines and n buses, b of which are equipped with PMUs. Each PMU produces time series data that are measurements of the voltage angle for their corresponding bus with respect to a nominal rotating reference frame (ie. 60 Hz). The system is initially in steady state and is perturbed by some disturbance. We assume that change point detection has successfully determined that a disturbance has occurred in the system and has produced the exact time at which the disturbance has occurred, t_0 . The measurement data is collected from each PMU until the time at which the classifier computation begins, $t_0 + T$. Utilizing the data available at time $t_0 + T$, the classifier will quickly localize the line outage.

This section defines the available PMU measurement data and describes in detail a probabilistic classifier. The classifier utilizes the time series measurement data to predict the line outage that produced it. To learn how to do this, an analysis is performed on PMU measurement data traces for which the true line outage is known. Such data is not available from a true power system in large quantity. Thus our work utilizes a large number of synthetic data traces that are produced using forecast load/generation distribution and dynamic simulation. Specifically, the synthetic data traces are collected from dynamic simulations whose pre-fault state is determined from a random sample of the forecast load/generation distribution. A classifier is then trained using the synthetic data produced from the simulated events. In this work, the classifier of choice is a linear multinomial regression model and the predictors include the post fault frequency domain PMU data obtained through discrete Fourier transforms.

Vectors are denoted with bold font (ie. \mathbf{v}) and subscripts denote the element index of a vector (ie \mathbf{v}_i). The vector of ones and the identity matrix are denoted $\bar{\mathbf{1}}$ and \mathbf{I} respectively and are of appropriate dimension. The set of real and complex numbers are denoted \mathbb{R} and \mathbb{C} respectively. The exponential

TABLE I: Table of Observations and Predictors. $FFT(\mathbf{a}_1^{(k_0:k_{Tf})})$ represents the complex Fourier transform frequency domain data of the signal $\mathbf{a}_1^{(k_0:k_{Tf})}$.

	Slack Bus PMU	All Other PMUs
Observations $j \in [2, \dots, b]$	$\mathbf{a}_1^{(k_{-\infty}:k_{Tf})}$	$\mathbf{d}_j^{(k_{-\infty}:k_{Tf})}$
Predictors $j \in [2, \dots, b]$	$FFT(\mathbf{a}_1^{(k_0:k_{Tf})})$ $\mathbf{a}_1^{(k_0)} - \mathbf{a}_1^{(k_{-1})}$	$FFT(\mathbf{d}_j^{(k_0:k_{Tf})})$ $\mathbf{d}_j^{(k_0)} - \mathbf{d}_j^{(k_{-1})}$

and natural logarithm functions are written $exp\{\cdot\}$ and $ln(\cdot)$ respectively.

A. PMU data

The set of PMU indices is denoted by $\mathbb{P} = [1, \dots, b]$. We assume that the measurement sample times are identical across each PMU. Each measurement sample occurs at a time $t^{(k)} = k/f$ where f is the sampling frequency and k is the integer time step. The time at which the disturbance has occurred is denoted by $t_0 = k_0/f$. The time over which data is collected after the fault, T , is chosen based on the desired performance of the classifier and is chosen such that $T = k/f$ for some integer k . The available measurement time samples are written as $[k_{-\infty}, \dots, k_{Tf}]$.

Let $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^b$ denote the vector of voltage angles at each PMU bus for the k^{th} step in the time series with respect to a nominal rotating reference frame. $\boldsymbol{\theta}^{(k_a:k_b)}$ represents the sequence of vectors for the time steps k_a, k_{a+1}, \dots, k_b . The PMU measurement vector is modeled as

$$\mathbf{a}^{(k)} = \boldsymbol{\theta}^{(k)} + \mathbf{n}^{(k)} \quad (1)$$

where the random noise vector is denoted $\mathbf{n}^{(k)} \sim \mathcal{N}(0, \Omega)$. The bus equipped with PMU 1 is arbitrarily chosen as the slack bus. The relative voltage angles are defined as the difference between $\mathbf{a}_1^{(k)}$ and the angle measurements on each of the other PMU buses. This vector of angle differences is written as

$$\mathbf{d}^{(k)} = \mathbf{a}^{(k)} - \bar{\mathbf{1}}\mathbf{a}_1^{(k)} \quad (2)$$

A simulation-based example of the observation $\mathbf{a}_1^{(k_{-10}:k_{30})}$ is shown in Figure 1a. The black curve shows the true signal trajectory which exhibits a typical response to a transmission line outage as an instantaneous jump occurs at the time of the outage, k_0 , followed by some sinusoidal oscillations. The gray dots show synthetic measurement samples with co-variance matrix $\Omega = 4.84 \times 10^{-4} * 10^{-4}\mathbf{I}$ in units of degrees. Table I illustrates the observations available from PMUs.

B. Producing Synthetic PMU Measurement Data

Utilizing the knowledge of a forecast load distribution a set of simulations is performed to produce likely time trajectories of the PMU data. Define the random vector $\mathbf{Z} \in \mathbb{R}^{2n}$ to be the real and reactive loads at each bus in the system. Load forecast error is taken into account by assuming \mathbf{Z} follows a known forecast distribution. The required forecast distribution has no restrictions and may be difficult to identify. If sampled

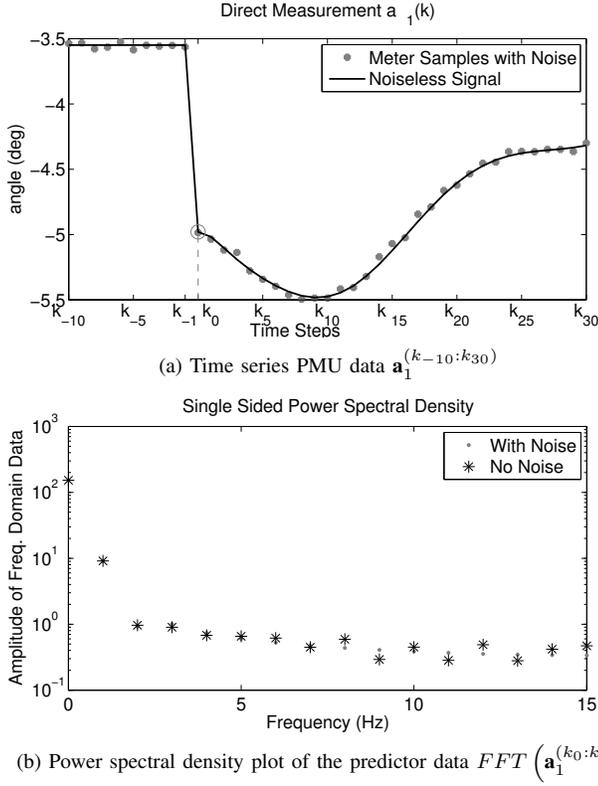


Fig. 1: Example of PMU data and frequency domain observations. a) Noiseless signal and measurements samples $\mathbf{a}_1^{(k_{-10}:k_{30})}$. b) Single sided power spectral density plot for the frequency domain predictors $FFT(\mathbf{a}_1^{(k_0:k_{30})})$ with and without noise.

properly, non-normal forecast distributions should not effect the classifier performance; however, in this work we use a Gaussian forecast distribution with some known mean ν and co-variance matrix Λ .

$$\mathbf{Z} \sim \mathcal{N}(\nu, \Lambda) \quad (3)$$

This forecast distribution of loads is used to simulate a set of likely PMU observations for a given line outage. The first step in executing a simulation is to initialize the static state of the system by solving the AC power flow equations for a given set of randomly sampled loads to determine the complex voltages at each bus. To provide the standard input to the power flow solver, the real power generation at each bus is scaled accordingly to match the random increase or decrease in the total system load. Using these complex voltages the dynamic state of the system (ie. generator and motor states) is initialized by way of a newton raphson algorithm [12]. A disturbance (ie line outage) can be simulated using any integration method which computes time traces of the system states including the voltage angles. Artificial noise is added to the voltage angle trajectories to obtain synthetic time series PMU data. This PMU data and the known line outages constitute the synthetic training data used to train the regression model.

TABLE II: Steps Required to Construct the Training Data

For $i \in \mathbb{L}$	
For $j \in [1, \dots, s]$	
2	Sample $\mathbf{z}^{(i,j)}$ from forecast distribution (3)
3	Using $\mathbf{z}^{(i,j)}$ injections, solve power flow and initialize system state
4	Simulate outage on line i
5	Collect PMU data and add artificial noise, Equation (1)
6	Construct vector of predictors, $\mathbf{x}^{(i,j)}$, as in Section II-C
7	Construct response variable $y^{(i,j)} = i$
end	
end	

C. Classifier Model

In machine learning, the multinomial regression classification method is used when the response variable is categorical. In this case the discrete outcomes to be predicted include only potential single-line outages indexed by the set of integers $\mathbb{L} = [1, \dots, L]$. As each outage corresponds to a transmission line, this set will have $L \leq m$ elements. The response variable Y realizes an integer $y \in \mathbb{L}$ that represents an index indicating the outed line.

The predictors are constructed from the angle data produced from the PMU at the slack bus and the angle difference data produced from each additional PMU on the time interval $[k_{-1}, \dots, k_{Tf}]$. The predictors directly include the instantaneous jump experienced by this time series observation data in response to the outage and indirectly include the post-outage time series data on the time interval $[k_0, \dots, k_{Tf}]$. Rather than utilizing the post-outage time series data directly, our predictors include the complex parts of the discrete Fourier transform frequency domain data of the signal. A typical single sided power spectral density plot of the voltage angle data is shown in Figure 1b. The predictors are concatenated into a vector denoted by $\mathbf{x} \in \mathbb{R}^{(Tf+2)b}$ and are stated explicitly in Table I. (Note: frequency domain data comprises of the same number of elements, $Tf + 1$, as the time domain data)

Under the assumption that change point detection has been triggered by a line outage represented in \mathbb{L} , the multinomial regression model approximates the probability distribution of the categorical random variable Y conditioned on the predictors \mathbf{x} . This distribution is a function of the predictors \mathbf{x} . Using short hand notation we have

$$Pr(Y = y|\mathbf{x}) = \mathbf{p}_y(\mathbf{x}) \quad \forall y \in \mathbb{L}$$

where $\mathbf{p}_y(\mathbf{x}) \geq 0$ is the y^{th} element of the vector valued function $\mathbf{p}(\mathbf{x})$ and $\bar{1}^T \mathbf{p}(\mathbf{x}) = 1$. The standard generalized linear model is used as the response function for the multinomial response variable.

$$\mathbf{p}_y(\mathbf{x}) = \frac{\exp\{\gamma_y + \beta_y^T \mathbf{x}\}}{\sum_{\ell=1}^L \exp\{\gamma_\ell + \beta_\ell^T \mathbf{x}\}} \quad \forall y \in \mathbb{L} \quad (4)$$

where $\gamma \in \mathbb{R}^L$ is a vector of parameters and $\beta_j \in \mathbb{R}^{(Tf+2)b}$ is the j^{th} column vector of a matrix of parameters $\beta \in \mathbb{R}^{(Tf+2)b \times L}$. These model parameters will be found via multinomial regression. The probability density function is

$$f(Y = y|\mathbf{x}) = \prod_{k=1}^L (\mathbf{p}_k(\mathbf{x}))^{[y=k]} \quad (5)$$

where $[y = k]$ evaluates to 1 if $y = k$ and 0 otherwise. The log-linearity of the probability density function greatly simplifies the training procedure.

D. Training the Regression Model

Table II provides the steps required to build the set of synthetic training data. Each of the L potential line outages will be simulated s times. The response variable, the vector of predictors, and the vector of power injections corresponding to a simulation are realized as $y^{(i,j)}$, $\mathbf{x}^{(i,j)}$, and $\mathbf{z}^{(i,j)}$ respectively, where $i \in [1, \dots, L]$ indexes the line being outed and $j \in [1, \dots, s]$ indexes the randomly sampled initial load. Since each of the sL observations are independent, their joint probability function can be written

$$\begin{aligned} g(Y^{(1,1)} = y^{(1,1)}, \dots, Y^{(L,s)} = y^{(L,s)} | \mathbf{x}^{(1,1)}, \dots, \mathbf{x}^{(L,s)}) &= \\ &= \prod_{j=1}^s \prod_{i=1}^L Pr(Y^{(i,j)} = y^{(i,j)} | \mathbf{x}^{(i,j)}) \\ &= \prod_{j=1}^s \prod_{i=1}^L \prod_{k=1}^L (\mathbf{p}_k(\mathbf{x}^{(i,j)}))^{[y^{(i,j)}=k]} \end{aligned}$$

The joint log-likelihood function, written as a function of the model parameters, can be derived to be

$$\begin{aligned} \ln(\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})) &= \sum_{j=1}^s \sum_{i=1}^L \left(\sum_{k=1}^L [y^{(i,j)} = k] (\boldsymbol{\gamma}_k + \boldsymbol{\beta}_k^T \mathbf{x}^{(i,j)}) \right. \\ &\quad \left. - \ln(\sum_{\ell=1}^L \exp(\boldsymbol{\gamma}_\ell + \boldsymbol{\beta}_\ell^T \mathbf{x}^{(i,j)})) \right) \end{aligned}$$

The maximum likelihood problem is written as

$$\min_{\boldsymbol{\gamma}, \boldsymbol{\beta}} - \ln(\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})) \quad (6)$$

In this paper problem (6) is solved using elastic net regularization paths from [13]. This algorithm is a homotopic algorithm that solves a sequence of grouped elastic net optimization problems that converge to the maximum likelihood problem (6). Warm start initial guesses are utilized to solve successive optimization problems via coordinate descent method.

Under the assumption that the number of iterations of the coordinate descent algorithm and the number of grouped elastic net optimization problems required to be solved does not scale with the size of the problem, the computation required to solve problem (6) scales linearly with the number of predictors, the number of events, and the number of classification outcomes, and $O((Tf + 2)bsL^2)$ operations are required.

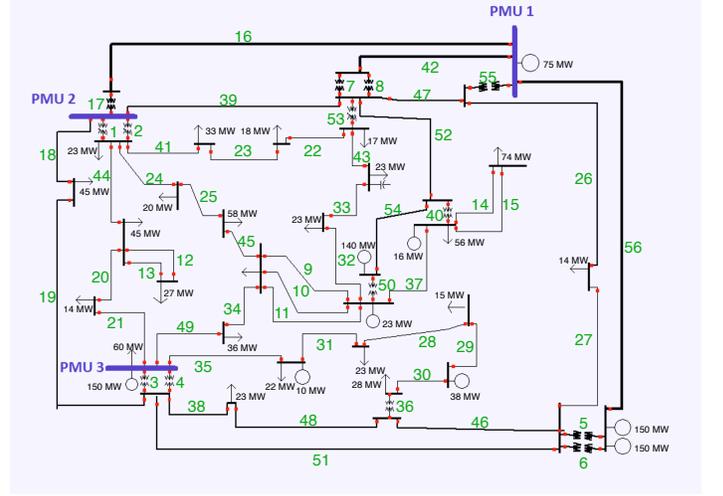


Fig. 2: 37 bus system

III. EMPIRICAL RESULTS

Our methodology is demonstrated on two power system models. Using the 37 bus test case from [5], empirical results are provided to illustrate the ability of our methodology in classifying outages quickly enough to provide stabilizing action. Specific outages in the 37 bus system are identified as being difficult to localize and both *topological* and *operating point* classification difficulties are discussed. Second, a 150 bus test case is constructed to analyze the scalability and computational tractability of our classification technique. Using this system, PMU placement is briefly investigated and *PMU configuration* classification difficulties are discussed.

A. 37 Bus Test Case

Our methodology is first demonstrated on the 37 bus test case analyzed in [5], utilizing PMUs on the same three buses. The one line diagram is shown in Figure 2. In this figure each line is numbered so we are able to correctly identify outed lines. The bus equipped with PMU 1 is designated the slack bus from which we attain observations $\mathbf{a}_1^{(k-\infty:kTf)}$. The measurement co-variance is set to the standard value of $\Omega = 4.84 \times 10^{-4} \mathbf{I}$ and has units of degrees [14]. The PMUs sample at a frequency of $f = 30$ Hz and the nominal voltage frequency is 60 Hz. The mean of the prior load distribution, $\boldsymbol{\nu}$, is set to the standard load provided by test case and the co-variance is defined to be $\Lambda = \text{diag}(.2\nu)$.

There are 9 generators and 25 loads in the system illustrated by circles and arrows respectively. The system includes two generator models, four excitation system models, and four turbine governor models. The most complex dynamic generator model is that of the generator placed on the bus that is equipped with PMU 1. This model is completely described by 16 dynamic state variables. Loads are modeled as being constant. Additionally, there exist multiple shunt capacitors and transformers in the system.

1) *Probability Evolution in Time*: We begin by analyzing the line outage of transmission line 46 in Figure 2, the same line outage studied in [5]. To study the behavior of

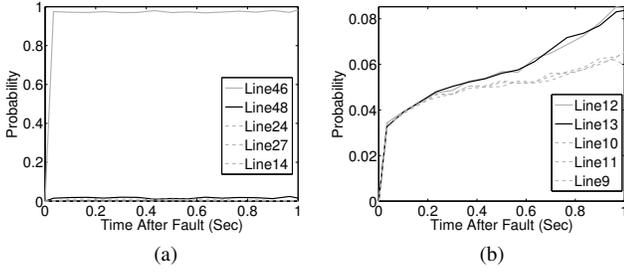


Fig. 3: Only those trajectories corresponding to the top five probabilities are shown. (a) Probability evolution in time for an outage on line 46. (b) Probability evolution in time for an outage on line 13.

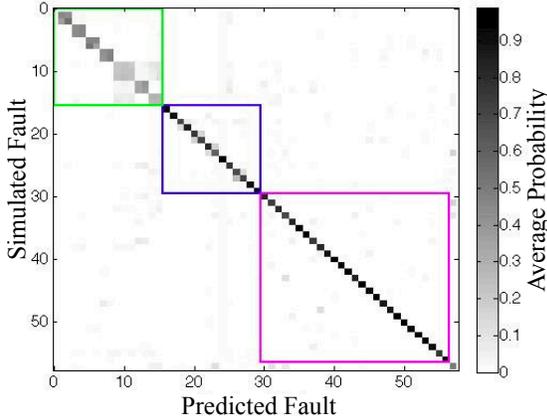


Fig. 4: Analyzing model performance: The y-axis represents the simulated fault and the x-axis represents the predicted fault. The green square outlines *multiple* lines. The blue square outlines *trail* lines. The magenta square represents all other outages and are ordered in terms of their line flow at the mean load ν . The final outage represents the null outage. Each row of the matrix represents the average probability vector over each of the 100 samples, $\mathbf{p}(\cdot)$. Each row sums to 1.

our classifier as a function of time horizon, T , 20 different classifiers are built using horizons ranging from .033 seconds to 1 second. Each of the classifiers were built using $s = 100$ events for $L = 57$ possible line outages. Included in the set \mathbb{L} is the set of $m = 56$ outages on lines and the *null outage* which simulates no disturbance to the system.

Figure 3a shows an example of how the probability vector evolves in time using an initial load sample that was not used in the training data. The plot contains trajectories representing elements of the vector of probabilities, $\mathbf{p}(\cdot)$. Only those trajectories corresponding to the top five probabilities at a one second horizon are shown. The sum of probabilities across all possible outages equals 1 at every time instant. The probabilities across all outages are $1/57$ at the instant the fault occurs; however, 0.033 second after the fault occurs it becomes apparent that the outage on line 46 is significantly more likely to have occurred than all others. With minimal real-time computation, the outage in this example is correctly classified in less than 0.04 seconds. It is important to note

that the quasi-steady state classifier presented in [5] cannot begin execution until a steady state is reached which does not happen in this example until after 4 seconds have passed.

Our results are compared with Tables I and II from [5]. Like those results, we correctly identified the outed line. Aside from this line, our next four most probable lines do not overlap with the next four highest ranked lines from [5]. It is difficult to determine if this difference in ranking is a result of the transient information used by our classifier or of the classification method itself. Additionally, our probabilistic results have a more straightforward interpretation than the NAD quantity presented in [5].

Figure 3b shows an example of how the probabilities evolve in response to an outage on line 13. Only the five outages with the largest probabilities at a one second horizon are shown. These five outages have very similar probabilities throughout the 1 second horizon, making it difficult to accurately choose the one specific line that has been outed. A deterministic localization procedure may choose line 12 as the estimated outage because it has the largest probability one second after the fault occurrence. However, corrective action should consider the possibility of an outage on lines 9-13.

2) *Classification Difficulties*: As illustrated in the section III-A1, some outages are easier to localize than others. Classifying an outage on line 46 is relatively easy as compared to an outage on line 13. In this section we begin to identify reasons that a line outage may be difficult to localize. For demonstration purposes, a relatively long time horizon of $T = 1$ second is used to identify line outages that are difficult to localize. Once again, the classifier was built using $s = 100$ events for each of the $L = 57$ possible line outages. Using validation data that was constructed from a separate set of random load samples, the probability vector, $\mathbf{p}(\cdot)$, was calculated 100 times for each outage $\ell \in \mathbb{L}$. The average probability vector for each of the 57 outages were concatenated vertically to attain a matrix that is depicted by the image plot in Figure 4. In this plot the rows correspond to the simulated outage and columns correspond to the predicted outage. The value of each element in the matrix represents a probability and each row sums to 1. If the diagonal elements were all 1, then our classifier would perform perfectly. We will analyze this figure with one concept in mind: it is difficult for the classifier to distinguish between two outages that result in similar steady state post-fault power flows throughout the system.

Two different *topological difficulties* are illustrated by the squares along the diagonal of the image plot in Figure 4. As seen in section III-A1, the classifier has difficulties distinguishing between lines that connect the same pair of buses. Let's refer to these groups of lines as *multiple* lines. The groups of *multiple* lines in the system include lines (1,2), (3,4), (5,6), (7,8), (9,10,11), (12,13), and (14,15). The green square highlights the groups of *multiple* lines in the system. When one of these lines is outed, the power flows throughout the network do not change significantly. Rather, the corresponding *multiple* lines assume the power flow lost by the outed line, steady state pre-fault and post-fault voltage angles are nearly identical and the dynamic response of the voltage angles exhibit oscillations

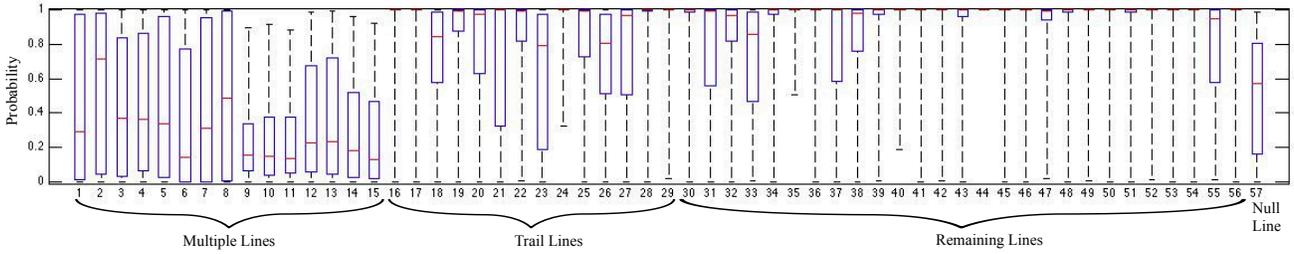


Fig. 5: Illustration of the distribution of probabilities produced from the classifier over each of the 100 validation events. For each line, ℓ , this box plot describes the distribution of the 100 probabilities produced for the outed line, $\mathbf{p}(\cdot)$. The x-axis displays the line number, ℓ , of the outed line. The same line groups are outlined as in Figure 4. Lines outages with a high mean and a low lower bound experience occasional *operating point* classification difficulties resulting in low outliers.

of small magnitude. A resulting low signal-to-noise ratio often causes single line outages located on *multiple* lines to be indistinguishable from each other and from the null outage which excites no dynamic response. Illustrating this point are the dense off-diagonal elements located within the green square of this plot as well as the elements that couple the *multiple* lines with the null outage.

Notice the seven dense diagonal blocks coupling the groups of *multiple* lines. These dense blocks show that *multiple* lines contained in the same group cannot be well distinguished from each other; however, the classifier is able to identify the group of *multiple* lines that contains the outed line. This suggests that it may be beneficial to consolidate each group of *multiple* lines into separate classification outcomes.

The blue square outlines groups of lines that also exhibit an ambiguously coupled behavior with respect to the classifier, although the dense blocks on the diagonal are less apparent. These lines have common topological characteristics as each pair represent trail edges of the system graph, the adjoining bus is not connected to a third transmission line, and the adjoining bus experiences little or no load/generation. These lines are referred to as *trail* lines. The groups of *trail* lines in the system whose adjoining buses generate or consume 20 MW of real power or less include (16,17), (18,19), (20,21), (22,23), (24,25), (26,27), and (28,29). When one of these lines is outed the power flows throughout the network may change significantly and dynamic oscillations of the observed voltage angles may experience large magnitude. However, outages located on grouped *trail* lines tend to excite nearly identical dynamic system responses. To understand the intuition, consider the extreme scenario where a group of two *trail* lines have an adjoining bus that does not inject or extract power and their line resistances are zero. In this case an outage on either line will drop the real power flowing on both lines to zero and result in identical real power flows throughout the system.

The box plot shown in figure 5 illustrates the distribution of probabilities produced for the correctly outed line, $\mathbf{p}(\cdot)$, over each of the 100 validation events. The mean for each line outage represents the diagonal elements of the matrix depicted in Figure 4.

The ambiguous coupling between line outages on *multiple* lines as well as *trail* lines has already been introduced. The classifier performs poorly with respect to these outages because the system's response to each line in an ambiguous

group is nearly identical. Outages located on *multiple* lines are particularly difficult to classify because they not only excite a similar response to their paired outage, but also excite a similar response to the null outage. On the other hand, outages on *trail* lines do not typically excite a similar response to the null outage, are thus easier to classify, and exhibit a high mean probability in figure 5. However, outages on *trail* lines are occasionally misclassified as their paired outages and thus such outages often exhibit a low 75th percentile. Line outages 37-56 do not present topological classification difficulties and exhibit both a high mean probability and a high 75th and 25th percentile bounds.

For each outage, outlier events exist resulting in misclassification. This is illustrated in figure 5 as most line outages experience lower probability bounds that are nearly zero. Many of these outliers occur when the randomly sampled initial state is such that the power flowing across the outed line is nearly zero. During such events, the system experiences a very small perturbation and the transient oscillations are low in magnitude. A resulting low signal-to-noise ratio causes these outliers to be indistinguishable from the null outage. These outlier events will be labeled as *operating point* classification difficulties. Lines 30-33 and line 23 are most vulnerable to *operating point* classification difficulties because they experience the lowest power flow at the mean power injections, ν .

B. 150 Bus Test Case

To demonstrate our methodology on a larger test case and to investigate PMU placement, we developed a 150 bus test case using the PSAT toolbox in MATLAB [12]. This large test case is represented by the graph shown in Figure 6. There are six regions in the system illustrated by blue and black dots each connected by eight tie lines. Regions 5 and 6 consist of only one bus containing a load of 2/8 p.u. and 1/8 p.u. respectively. Regions 1 through 4 are duplicate 37 bus systems containing the same graphical structure as in Figure 2. The PMU locations shown in Figure 2 are used for each of the regions 1 through 4 and no PMUs exist in regions 5 and 6. The tie lines that connect regions have low impedance and are connected to the bus equipped with PMU 1. Figure 6 assigns indices to each of the tie lines. The slack bus for the entire network is designated to be the bus in region 3 that is equipped with PMU 1. The slack bus picks up the additional load that is introduced in

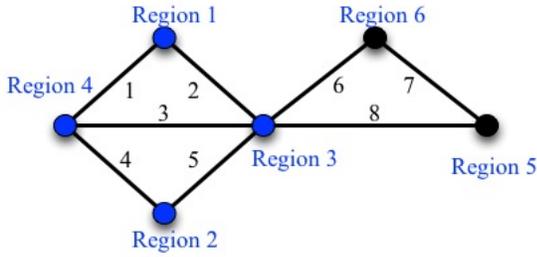


Fig. 6: 150 bus system. Regions are labeled in blue and tie lines are labeled in black.

regions 5 and 6. There are 12 total PMUs in the system, and there are 232 lines in the system. Each PMU samples at a frequency of $f = 30$ Hz and the measurement co-variance is set to $\Omega = 4.84 \times 10^{-4} \mathbf{10}^{-4} \mathbf{I}$.

1) *Computational Scalability*: An important feature of the proposed classifier is the minimal real time computation. Once an outage has been detected two major computations must be carried out. First, each of the PMU signals must be transformed into the frequency domain. The computation time of a discrete Fourier transform has been extensively studied [15], [16]. Signals considered in this paper include few measurement samples and Fourier transform computations are executed on the order of 10^{-5} seconds. Second, each of the probabilities must be calculated. The algebraic probability calculation, equation (4), requires little effort to evaluate and are executed in fractions of a second for very large systems. In fact, the vector of probabilities scales linearly with the number of available PMUs, b , and number of potential outages, L .

In this example we consider $L = 233$ potential line outages, one for each of the $m = 232$ lines and one for the null outage. Using a time horizon of 1 second results in 63 predictors for each of the 12 PMUs and 756 total predictors. Transforming each of the 12 signals into the frequency domain requires roughly 10^{-5} seconds of computation. The 233 algebraic probability calculations, require a total of roughly 0.14 seconds. This computational time was measured using a single processor on a standard 2.7 GHz laptop computer. Assuming detection occurs within one second of the outage occurrence, our procedure will provide probabilities 1.14 seconds after the fault has occurred.

Additional computation is required to solve the optimization problem (6) and train the classifier. This computation is executed a priori and is further discussed in section IV-B. In this example, data was collected from $s = 100$ simulations for each of the $L = 233$ outages. The model is fit to the resulting 23,300 events. The fitting process (solving problem (6)) required two hours to complete using a single 2.7 GHz processor. This computation time may be improved by utilizing additional computational resources including hardware and custom algorithms. Each of the 23,300 simulations require less than 1.5 seconds of computation time on a single 2.7 GHz processor using simulation code that is not computationally efficient. Utilizing four processors can lower the simulation computation to less than two and a half hours. Utilizing

computationally efficient simulation software can significantly decrease this computational burden.

2) *Classification Difficulties*: Once again, we consider a classifier that utilizes a time horizon of $T = 1$ second. The classifier was built using $s = 100$ events for each of the $L = 233$ possible line outages. Using validation data that was constructed from a separate set of random load samples, the probability vector, $\mathbf{p}(\cdot)$, was calculated 100 times for each outage $\ell \in \mathbb{L}$. Figure 7 illustrates the performance of the classifier. For each outed line, ℓ , this bar graph presents the average probability for the outed line, $\mathbf{p}(\cdot)$, over the 100 probabilities produced. This metric is used to analyze the classifier's performance for individual line outages. The performance characteristics of each outage in regions 2, 3, and 4 are nearly identical to those contained in region 1. For this reason, the analysis of outages in region 1 are directly applicable to outages in regions 2, 3, and 4.

This section intuitively explains classification difficulties experienced by the 41 outages that are most difficult to classify in the 150 bus system according to the provided metric, 27 of which are contained in regions 2, 3, and 4 and are thus omitted from figure 7. These 41 outages account for all single line outages that produce an average probability of less than .1. In attempt to explain only the most severe classification difficulties, this section does not address outages that produce average probabilities greater than .1. The value of .1 is chosen only for illustration purposes.

There are fourteen outages shown in figure 7 that fall below an average probability of .1, nine of which are contained in region 1. The *topological difficulties* identified in the last example explain why seven of these fourteen outages are difficult to localize. Specifically, outages on lines 9 through 15 in region 1 are *multiple* lines. The *operating point difficulties* identified in the last example explain why 4 of these fourteen outages are difficult to localize. Specifically, tie lines 4 and 5 and lines 23 and 31 in region 1 are four of the five lines that have the lowest pre-outage power flow at the mean injection values, ν . The remaining 3 difficult to localize outages are left unexplained. This example will demonstrate that the classifier cannot localize these 3 outages because of a deficiency in the PMU configuration.

PMU placement has been briefly studied in attempt to improve the classifier's performance. In this example, we are interested in placing one additional PMU in the system to boost the performance of the classifier with respect to the 14 outages that are difficult to localize. Each bus connected to one of these 14 lines were considered as candidate placement locations. For each of the 10 candidate PMU locations a different classifier was built. None of the 10 classifiers were able to raise the performance metric above the threshold for those 11 outages that have been labeled as being difficult to localize due to *operating point difficulties* and *topological difficulties*. However, 2 of the 10 classifiers were able to easily detect the 3 outages whose difficulties have been left unexplained, tie lines 6-8. These 3 outages are labeled as having *PMU configuration difficulties*. One of the improved classifiers utilized PMU data from the bus in region 6, and the corresponding results are shown in figure 7.

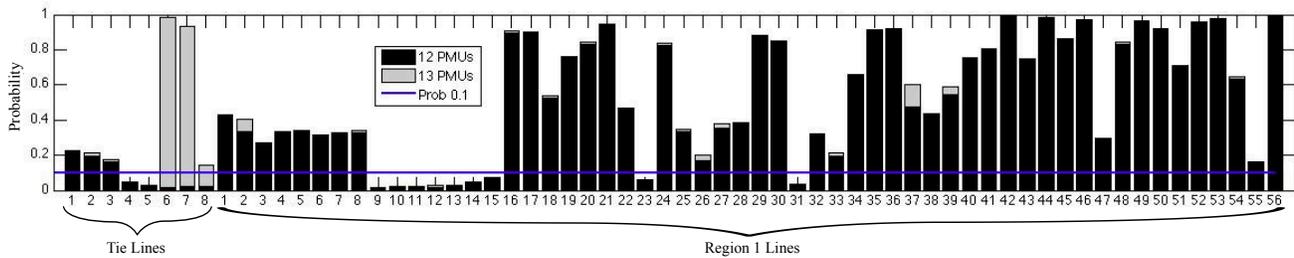


Fig. 7: Classification performance for the 150 bus system. Outages on each tie line and each line in region 1 are presented. Outages on lines in regions 2 through 4 are omitted because they experience nearly identical performance characteristics as those in region 1. For each outed line, ℓ , this bar graph presents the average probability for the outed line, $\mathbf{p}(\cdot)$, over the 100 probabilities produced. The x-axis displays the line number, ℓ , of the outed line. The gray bars show results when an additional PMU is placed in region 6. The analysis of section III-B2 addresses classification difficulties experienced by all outages whose average probability falls below the horizontal line drawn at the probability .1.

IV. DISCUSSION AND FUTURE WORK

This section begins by discussing the classification difficulties identified in Section III. Beyond these difficulties, additional complications arise when classifier is implemented into actual system operation. Discussed first is the classifier update process. As the power injections in the system and the power injection forecasts evolve in time the classifier must be updated in a manner that fits well into existing state estimation procedures. Second, the classifier must be adjusted to perform well in the presence of an imperfect dynamic simulation model. To address this problem, a method of feature selection is discussed.

A. Classifier Difficulties

Section III empirically identifies three difficulties that deter the performance of the classifier. Future work will adapt the classifier to accommodate these difficulties. *Operating point* classification difficulties are attributed to small system perturbations, causing the system to exhibit transient oscillations that are low in magnitude. Due to their unobtrusive nature, these outages often do not require immediate corrective action and may be better suited for a steady state-based classifier. Eliminating these unobtrusive events supports the purpose of fast acting classification which is to perform immediate corrective action.

Topological classification difficulties that were presented are attributed to pairs of lines that when outed produce a nearly identical dynamic system response. Since the system response is similar for each of these pairs, it is likely that the same corrective action would be employed for either line outage. With this in mind, each group of *multiple* and *trail* lines may be merged into one classification outcome. Indeed the empirical results of section III-A2 showed that the presented classifier was able to successfully locate the ambiguous group that contained the outed line.

PMU configuration classification difficulties may be identifiable using theory provided by existing cyber-security research. For a given PMU configuration, reference [17] identifies unobservable *breaker and jammer attacks* which can be interpreted as transmission line outages exhibiting *PMU configuration* classification difficulties. Assuming that every

power injection is metered in the 150 bus test case and 12 PMUs are located as in our example, the only unobservable breaker and jammer attacks that compromise exactly one breaker status must compromise the status on tie line 6, 7, or 8. These are the same lines that present *PMU configuration* difficulties. To protect against these attacks, methods from [17] propose the placement of one PMU in either region 5 or 6. This PMU placement agrees with the findings in this manuscript and further suggests that cyber-security-based research may be relevant in determining PMU placement to eliminate *PMU configuration* classification difficulties.

Additional existing quasi-steady state-based PMU placement heuristics may be relevant to improving our classifier. Reference [18] formulates an integer programming problem that identifies the PMU configuration that maximizes the change in observed quasi-steady state voltage angle differences in response to a list of potential outages. Such a PMU configuration will not only correct *PMU configuration* classification difficulties but may also improve the classifiers general performance.

B. Classifier Updates

The power system operator's day ahead market schedules hourly bulk generation based on the load and renewable generation forecasts roughly one day in advance. The forecasts are provided as constant values over the hour intervals of interest. We propose that distribution (3) be created using these forecasts along with the bulk generation scheduled through the day ahead market. The mean of the distribution (3), ν , is set to the forecast injections plus the scheduled bulk generation at each bus. The co-variance matrix, Λ , is directly related to the expected forecast accuracy. The classifier training computation begins immediately after the day ahead market has cleared. Along with the scheduled bulk generation, the classifier is updated on the hourly time intervals considered by the day ahead market.

It is critical for the construction of the classifier to require less than 24 hours of computation when using the proposed update method. Scaling to large systems (i.e. 100,000 bus systems) may be difficult. The number of simulations required to construct the training data grows linearly with the number

of line outages considered by the classifier; however, this computation is executable in parallel. Of greater concern is the computational burden of solving the optimization problem (6) to train the classifier. This computation cannot be executed in parallel and the order of operations required to solve this problem grows quadratically with the number line outages considered and linearly with the number of PMU meters. In reference to the 150 bus test case the computation time required to construct the training data using 4 processors and train the classifier was roughly four hours. As a result, four models must be built in parallel at all times throughout the day in order to update the classifier each hour. This would require four processors on a single computer. Future work will focus on this computational burden.

C. Feature Selection and Imperfect Simulation Models

Rather than solving the maximum likelihood problem (6) to build the regression model, a grouped elastic net problem can be solved to encourage a sparse parameter matrix β and effectively perform feature selection. This method of feature selection has been investigated. Using the analysis framework provided in this paper, we have shown that reducing the number of predictors via feature selection deters the performance of the classifier. However, removing predictors at specific frequencies provides a filtering effect that may remove high frequency oscillations introduced by higher order dynamic models. As a result, feature selection could possibly improve the performance of the classifier when the simulation model is not a perfect representation of the actual system. Future work will focus on feature selection in the presence of inaccurate dynamic simulation models using field measurements in a true power system.

V. CONCLUSION

In this paper we have presented a statistical classifier that localizes line outages using time series PMU data that is sampled during the transient response of the system. The classifier is a linear multinomial regression model that is trained by solving a maximum likelihood problem. The synthetic training data is constructed from time domain simulations that are initialized through random sampling of a forecast distribution on system power injections. The real time evaluation of the classifier requires minimal computation and produces probabilities over each potential line outage. The empirical results section analyzes two systems. The 57 bus system is used to compare our results to existing quasi-steady state classifiers and to illustrate the ability of the classifier to quickly identify

line outages before a steady state is reached. The 150 bus system is used to illustrate the computational tractability of implementing the classifier into a real system. Classification difficulties were identified based on the topology of the system, system operating point, and the PMU configuration. Future work is proposed to improve the classification performance with respect to these difficulties. Additional future work will investigate feature selection and PMU placement.

REFERENCES

- [1] (2004) Us-canada task force presents final report on blackout of august 2003. [Online]. Available: http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/Blackout_Press_Release_April_5_2004.pdf
- [2] A. Monticelli, "Electric power system state estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, 2000.
- [3] —, "Modeling circuit breakers in weighted least squares state estimation," *Power Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 1143–1149, 1993.
- [4] B. Clewer, M. Irving, and M. Sterling, "Topologically independent state estimation," in *Proceedings of the IFAC symposium on Power systems, modeling and control applications*, vol. 17, no. 1, 1988, pp. 17–1.
- [5] J. E. Tate and T. J. Overbye, "Line outage detection using phasor angle measurements," *Power Systems, IEEE Transactions on*, vol. 23, no. 4, pp. 1644–1652, 2008.
- [6] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.
- [7] Y. Zhao, R. Sevlian, R. Rajagopal, A. Goldsmith, and H. V. Poor, "Outage detection in power distribution networks with optimally-deployed power flow sensors," in *Power and Energy Society General Meeting (PES), 2013 IEEE*. IEEE, 2013, pp. 1–5.
- [8] Y. Zhao, J. Chen, A. Goldsmith, and H. Poor, "Identification of outages in power systems with uncertain states and optimal sensor locations," 2014.
- [9] A. Abur and A. G. Exposito, *Power system state estimation: theory and implementation*. CRC Press, 2004.
- [10] L. Xie, Y. Chen, and P. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," 2014.
- [11] E. Lawrence, S. V. Wiel, and R. Bent, "Model bank state estimation for power grids using importance sampling," *Technometrics*, vol. 55, no. 4, pp. 426–435, 2013.
- [12] F. Milano, "Power system analysis toolbox quick reference manual for psat version 2.1. 2," 2008.
- [13] J. Friedman, T. Hastie, and R. Tibshirani, "glmnet: Lasso and elastic-net regularized generalized linear models," *R package version*, vol. 1, 2009.
- [14] K. Martin, G. Benmouyal, M. Adamiak, M. Begovic, R. Burnett Jr, K. Carr, A. Cobb, J. Kusters, S. Horowitz, G. Jensen *et al.*, "Ieee standard for synchrophasors for power systems," *Power Delivery, IEEE Transactions on*, vol. 13, no. 1, pp. 73–77, 1998.
- [15] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.
- [16] P. Duhamel and M. Vetterli, "Fast fourier transforms: a tutorial review and a state of the art," *Signal processing*, vol. 19, no. 4, pp. 259–299, 1990.
- [17] M. Garcia, A. Giani, and R. Baldick, "Smart grid data integrity attacks: Observable islands," in *Power and Energy Society General Meeting*. IEEE, 2015.
- [18] Y. Zhao, A. Goldsmith, and H. V. Poor, "On pmu location selection for line outage detection in wide-area transmission networks," in *Power and Energy Society General Meeting, 2012 IEEE*. IEEE, 2012, pp. 1–8.