Bad Data Detection in the Context of Leverage Point Attacks in Modern Power Networks

Ankur Majumdar, *Student Member, IEEE,* Bikash C. Pal, *Fellow, IEEE*

***Abstract*—This paper demonstrates a concept to detect bad data in state estimation when the leverage measurements are tampered with gross error. The concept is based on separating leverage measurements from non-leverage measurements by a technique called diagnostic robust generalized potential (DRGP), which also takes care of the masking or swamping effect, if any. The methodology then detects the erroneous measurements from the generalized studentized residuals (GSR). The effectiveness of the method is validated with a small illustrative example, standard IEEE 14-bus and 123-bus unbalanced network models and compared with the existing methods. The method is demonstrated to be potentially very useful to detect attacks in smart power grid targeting leverage points in the system.**

***Index Terms*—distribution management system (DMS), remote terminal unit (RTU), state estimation (SE), leverage measurements, bad data detection (BDD), generalized studentized residuals (GSR), diagnostic-robust generalized potentials (DRGP)**

I. Introduction

POWER networks all over the world are witnessing significant scale of development. The major drivers are shift of technology of generation towards renewables (mainly solar and wind) and new forms of demand such as electric transportation, district heating, etc. The uncertain nature of generation and new type of demand need to be dealt with by more active energy management strategy [1]. There is an increasing adoption of smart instrumentation such as phasor measurement units (PMUs), intelligent metering, etc. in transmission networks and smart meters in distribution networks with information and communication technology (ICT) infrastructure. As a result, the integrity of data and information is exposed to risk and the power system is more prone to malicious attacks from adversaries. Tampered data will obviously affect the outcome of network control and computing functions such as state estimation, security analysis, volt var control (VVC), etc.

To enable the effective control of the power network, the states of the system need to be observed properly. The energy/distribution management systems (EMS/DMS) will play a crucial role in control and operation of the smart power systems. Central to every EMS/DMS are two functional blocks: the state estimator and the control scheduling block. The state estimation provides a real-time estimate of system states, based on the measurements obtained from meters and sensors in the remote terminal units (RTUs).

Over the years, the state estimation has been developed to deal with gross error in data because of inaccuracy of the measurements. Any tampering with data and maliciously operating switch will also result in gross errors in data. So in principle, the effect of malicious attack can be detected through bad data detection. Depending on the state estimation methodology the bad data detection can be part of state estimation process or a post estimation computation as shown in Fig.1. As long as these errors are part of overly measured systems (more measurements than the number of states to be estimated) and do not belong to the critical measurement and leverage points (measurements that significantly influence the state estimation solution), eliminating them to get a clear and accurate estimate is not difficult. However, if these bad data belong to the meters in the leverage set-they need to be handled carefully. The leverage measurements help in improving the state variables of the system by providing enough redundancy. The critical measurements are those, whose removal affects the system observability. Reference [2] and [3] have talked about protection of some or all of basic measurements or critical k-tuples in the system. However, the leverage measurements are not protected. The leverages can occur both in transmission and distribution networks [4]. This requires to develop a methodology to deal with the situation. The primary motivation of this research is driven by such possible scenarios when the hacking of the data in meters is related to the measurements of the leverage points.

A schematic of a typical energy/distribution management system is shown in Fig.1. The state estimation block in Fig.1 provides the estimate of power system states on the basis of measurements obtained from the supervisory control and data acquisition (SCADA) system. Usually, the measurement errors are assumed to be random and independently and identically distributed and obey normal distribution. As a result, weighted least-squares (WLS) estimator is used to find the best estimates of the states [5].

It has been discussed in [6] that an adversary can inject malicious data into the system without being detected by classical bad data detection techniques. In the case the adversary performs an unobservable attack, it is important to know how vulnerable the power system operation is to these attacks.

In recent years, there have been growing interests in the false data injection to power system and dealing with those attacks and the vulnerabilities [7]–[11]. The basic idea of false data injection attack is to add a non-zero attack vector into the measurements [7]. It has been reported in the literature [7]–

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) UK through the PV2025-Potential Costs and Benefits of Photovoltaics for UK-Infrastructure and Society (EP/K02227X/1): Data supporting this publication can be obtained on request from cap-publications@imperial.ac.uk.

A. Majumdar and B. C. Pal are with the Department of Electrical and Electronic Engineering, Imperial College London SW72AZ, U.K. (e-mail: ankur.majumdar@imperial.ac.uk;b.pal@imperial.ac.uk)

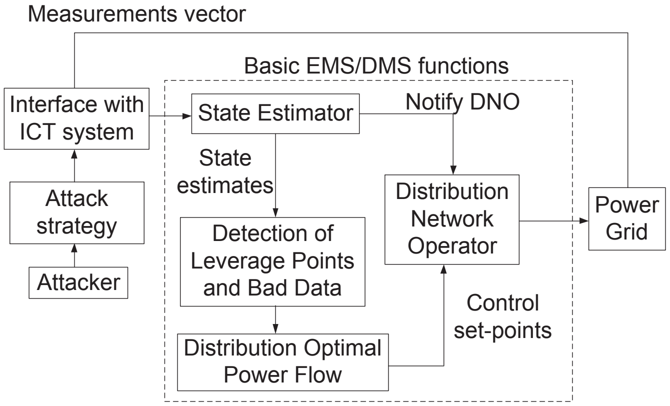

Fig. 1. A typical energy/distribution management system architecture

[11] that how an adversary can synthesize an attack vector just to bypass the normalized residual test in the dc state estimator. Reference [7] further explains different types of attack that can be synthesized. Reference [8] illustrates the strategies for malicious attacks to be incorporated and henceforth, uses the concept of generalized likelihood ratio test (GLRT) to detect the gross error. Reference [9] reports how to optimize the number of measurements to be tampered in order to compromise a given number of state variables. Bi and Zhang [10] have proposed a greedy algorithm for the sequential protection of state estimation against the malicious tampering. Hug and Giampapa [11] show how an attack can be hidden by tampering with a number of measurement data and they assess the vulnerability associated with this threat. Kim and Poor in [12] proposed placing secure PMUs on strategic buses to identify data injection attacks and Giani et al [13] used known-secure PMUs as a countermeasure to unobservable attacks. However, none of [7]–[13] have addressed the situation when a particular influential or leverage measurement is compromised. Hence, this provides the basis to have an identification procedure against such an unobservable attack.

Traditionally, the detection of bad data has been carried out by largest normalized residuals (LNR), which performs pretty well when there is a single bad measurement or multiple non-interacting bad measurements in the system [14]. However, it fails in case of influential or leverage measurements [15]. Reference [5] and [16] have proposed a χ^2 test for the identification of bad measurement. Chen and Abur have proposed the method of placement of PMUs to enable bad data detection in state estimation [17]. Reference [18] and [19] have devised the concepts of robust distances and influence functions in regression analysis of measurement equations to identify leverage points in a system. But they did not discuss about detecting bad data in leverage data points.

In other technology areas, the gross measurement errors have been treated as outliers in factor space of linear regression analysis [20], [21]. However, sometimes some influential measurements called leverage points may look like outliers in factor space as they lie outside the regression line [5], [22]. Therefore, it is hard to identify errors in leverage points. As a result, it is necessary to distinguish the leverage points from the outliers. It has been reported that there are number of ways one can identify the leverage points from the diagonal elements of the hat matrix, Mahalanobis distance (MD) of measurements, projection statistics (PS) [5], [16], etc. Reference [19] has devised the concept of influence function as a combination of influence of residuals and influence of position in factor space. Thus, looking at the influence function one can identify bad data even for influential measurements. The concept of finding outliers in multivariate data has been suggested in [20] and the concept of identifying outliers based on different residual diagnostics in linear regression model of a system has been reported in other applications [21]. However, they have not addressed the masking and swamping effect of leverage points when there are multiple leverage points.

In the field of applied statistics, there has been research on identifying the multiple high leverage points in multivariate analysis. It has been pointed out that due to the presence of more than one high leverage point, the leverage structure may change in such a way that the leverage diagnostics for single leverage point like twice-the-mean rule, thrice-the-mean rule, Cook's distance, Welsch and Kuh's distance, etc. will not be able to identify the real high leverage points. Nurunnabi, Hadi and Imon in [23] have used a modified Cook's distance and Habshah, Norazan and Imon [24] have proposed a robust diagnostic potential to address this issue. Reference [24] have applied the technique on Hawkins, Bradu and Kass data and Brownlee's stack loss data to illustrate the simultaneous identification of outliers or erroneous data and high leverage points. To the best of the knowledge of the authors, this methodology has never been applied in power system bad data detection context. This paper presents a robust bad data detection technique when the leverage measurements are compromised and shows that it can take care of masking/swamping phenomenon in sparse systems like power systems that existing methods cannot. This methodology is applied, for the first time, to robustly detect bad data in regards to state estimation of power system.

The power system state estimation measurement equations in this context are looked at as linearized regression equations of state variables at each operating point. This paper presents this concept and proposes a technique to diagnose bad data and leverage measurements simultaneously from the rest of the regression data. The paper is organised as follows- Section II describes a brief review of distribution system state estimation, bad data detection and the regression in regards to state estimation and the attack strategies if an adversary intends to attack. This section also explains the phenomena of masking and swamping. Section III explains in detail the concept of robust generalized potentials and the method to identify the outliers and leverage data points at the same time with generalized potentials and studentized residuals. Section IV presents case studies and analyses the results. The performances of this method have been tested on a small illustrative example and on standard IEEE-14 bus test system and IEEE-123 bus unbalanced distribution system. Section V summarizes the contributions and conclusions of the work.

II. Regression Analysis in relation to State Estimation

A. State Estimation

The unbalanced power distribution system state estimation (SE) is a process which estimates real-time states based on the measurements. The transmission system is, however, a special case of unbalanced system where, the system is balanced and hence, the number of state variables and equations are reduced. The problem is usually solved by Weighted Least Squares (WLS) estimation algorithm. This WLS algorithm is formulated as a minimization of the following objective function.

$$J = [z - h(x)]^T R^{-1} [z - h(x)] \tag{1}$$

Subject to:

$$c\,(x) = 0 \tag{2}$$

Where,

x State variables such as voltage magnitudes and angles.
m Number of measurements per phase.
R Measurement error covariance matrix, $z = \left[z_1^a\ z_1^b\ z_1^c\ \ldots\ z_i^a\ z_i^b\ z_i^c\ \ldots\ z_m^a\ z_m^b\ z_m^c\right]$.
z_i Measured value of i^{th} measurement.
$h(x)$ vector of measurement as a function of state x
$c(x)$ vector of zero injection measurements.

In three phase system $x = \left[\delta_i^k\ V_i^k\right]^\top$, where $V_i^k = \left[V_i^a\ V_i^b\ V_i^c\right]^\top$ is the vector of three-phase voltage magnitude at bus i, $\delta_i^k = \left[\delta_i^a\ \delta_i^b\ \delta_i^c\right]^\top$ denotes the phase angles of bus i except the reference bus and R=$Cov(\mathbf{e})$=$E(\mathbf{ee^T})$=$diag(\sigma_1^2, \ldots \sigma_i^2, \ldots, \sigma_m^2)$, where $\sigma_i = \left[\sigma_i^a\ \sigma_i^b\ \sigma_i^c\right]^\top$. Eq. (1) and (2) can be solved by Newton's method, which translates into solving the following equation at each iteration

$$[G(x^k)]\Delta x^{k+1} = H^T R^{-1} [z - h(x^k)] \tag{3}$$

where, n is the total number of state variables in the system. $\Delta x^{k+1} = x^{k+1} - x^k$ and $H(x^k) = [\frac{\partial h}{\partial x}]$ is the Jacobian matrix of dimension $3m \times (n-3)$ and $G(x^k) = H^T(x^k) R^{-1} H(x^k)$ is the Gain matrix in the k^{th} iteration.

Usually, in the distribution system there are not enough available measurements. But to have an estimate of the states, the number of measurements should be more than the number of state variables. Hence, to facilitate this the unmeasured loads of the distribution system are predicted from the load history of the load bus. Moreover, there can be some buses in the system where there are no injections or measurements available. These buses are known as zero-injection buses.

B. Leverage Points and Bad Data

The state estimation problem is linearized around an operating point and is expressed as the following regression model.

$$\Delta z = H \Delta x + \mathbf{e} \tag{4}$$

where z is considered the output of the regression model and x vector is the regressor, predictor or the factor in the regression model and $\mathbf{e}$ is the random error vector, which are random and assumed to be independently and identically distributed (*i.i.d.*), in the regression model. The matrix H is known as the coefficient or regressor matrix. The detection, assessment and understanding of influential points are the main areas of study in the regression model building. The factor variables or the explanatory variables in the regression model are solved by least squares estimation as in equations (1) and (2). From Eq. (3), the estimated measurement vector is derived as

$$\Delta \hat{z} = H(H^T R^{-1} H)^{-1} H^T R^{-1} \Delta z = K \Delta z \tag{5}$$

where, K is called the hat or high-leverage matrix. A large diagonal entry of the hat matrix implies that the particular measurement has more leverage or influence on the estimated states than others and they can be called as *leverage points*. If the influence is high enough the corresponding diagonal entry may be close to 1. In other words, according to equation (4), each observation (Δz_i,H_i) is a point in the factor space of regression, where H_i is a row of the H matrix. When there is an outlier in the X-space or H_i-space or the regressor variable space, it is said to have an undue influence on the state estimates and is called a leverage measurement.

The concept of bad data and outliers go hand in hand in the context of regression analysis. *Bad data* usually refers to an erroneous measurement due to various reasons. Due to the integration of PMUs, intelligent and smart metering with ICT infrastructure, bad data or gross errors can occur during the data transfer over the telecommunication medium. Telecommunication system failures or noise caused by unexpected interference also lead to large deviations in recorded measurements. Random errors usually exist in measurements due to the finite accuracy of meters when the meters have biases, drifts or wrong connections. So, these bad data or gross errors can be looked as outliers in the measurement space. However, a measurement, which may or may not contain errors, such as leverage points, may also appear as an outlier due to the structure of the corresponding regression equation. As a result, it is essential to differentiate the leverage points from bad data and identify the error, if any, in leverage points.

In the modern power system, more and more advanced communication and cyber technologies are getting incorporated [15]. Therefore, the possibility of an adversary to tamper with the measurements to drive the state estimator to wrong estimates is also high. Theoretically, the bad data detection (BDD) technique using normalized residual is a post-estimation process. Essentially, the largest normalized residual (LNR) method is used to detect, identify and eliminate bad measurement data. The largest normalized residual refers to the test where the largest normalized residual corresponds to the bad measurement data. Normalized residual based approaches for identification of bad data have been reported in [5] and [16]. In the case of one erroneous measurement data, the largest normalized residual works perfectly fine. Even it has been reported in the literature [5], [16] that LNR also works on both non-interacting and interacting non-conforming multiple bad measurement data. However, it fails to detect the bad data if there are multiple interacting and conforming bad data [14],where the errors are in agreement, and if they are part of the leverage set. Moreover, the residuals are given as

$$r = \Delta z - \Delta \hat{z} \tag{6}$$

Eq.(6) can be rewritten as:

$$r = (\mathbf{I} - K)\Delta z \tag{7}$$

Therefore, the measurement residuals with large diagonal entries of the hat matrix are small even if they are contaminated with gross error.

C. Attack Strategies

In power systems, the state estimator as mentioned in Fig. 1 takes three kinds of inputs-the meter measurement data (power injection and power flow), the network topology information data (on/off status of switches) and the parameter data (branch impedance and variances of measurement errors). Typically, these inputs are either sent from meters to control center or stored in the databases. It is assumed that the adversary can access and manipulate all the three kinds of inputs.

The leverage measurements occur when there are injection measurements on a bus, which has more number of branches connected to it compared to others, injection measurements on a bus incident to branches with very different impedance, and the line power flow measurements on relatively short lines. In large meshed distribution systems these leverages can occur due to the presence of line power flow measurements on short lines [4] and also due to the lower redundancy of measurements.

An adversary can take advantage of this situation and attack the high leverage points to influence the estimates of the state variables of the system and hence, can hide the attack from being detected. The leverage points affected by gross errors are called bad leverage points. Though bad leverage points are harmful to many estimators, good leverage points are particularly useful in improving the variance of the estimates.

1) Attacking power flow measurements: Power flow measurements are normally placed between buses to monitor the flow of the branches. Leverage power flow measurements are formed when the measurements are placed on relatively short or long lines. An attacker, if he/she intends to make the attack invisible, makes changes to the value of the corresponding diagonal element of the hat matrix by applying Theorem 2 and rule 1 and rule 2 as given in [25].

2) Attacking power injection measurements: Power injection measurements are placed at bus to monitor the active and reactive power injections from a load or generations at a particular bus. A node/bus is particularly vulnerable to leverage attack if that has more connecting branches connected to it or in other words there are more non-zero elements in that row of the H as in (3) matrix compared to other rows. If an adversary wishes to attack an injection leverage measurements he/she should increase the particular diagonal element of the hat matrix to make the attack undetectable by applying Theorem 2 and rule 1 and rule 2 as given in [25].

The sample high and low leverage points are shown by arrow marks in Fig.2. In the figure, the line flow measurement flow 5-4 is a high leverage measurement. To make a successful attack, the attacker makes changes to the value of the diagonal element of the hat matrix K_{ii} corresponding to the flow measurement 5-4 by applying Theorem 2 and rule 1 and rule 2. The theorems 1,2 and 3 are stated in Appendix A. Theorem 1 states how a successful attack can be made on a measurement by changing K_{ii}. While Theorem 2 shows how much K_{ii} has to be increased to make an attack on a single measurement z_i. To make an attack on multiple measurements, the attacker will perturb the measurements one at a time and apply Theorem 2 repeatedly. Finally, Theorem 3 suggests how to increase the value of K_{ii}.

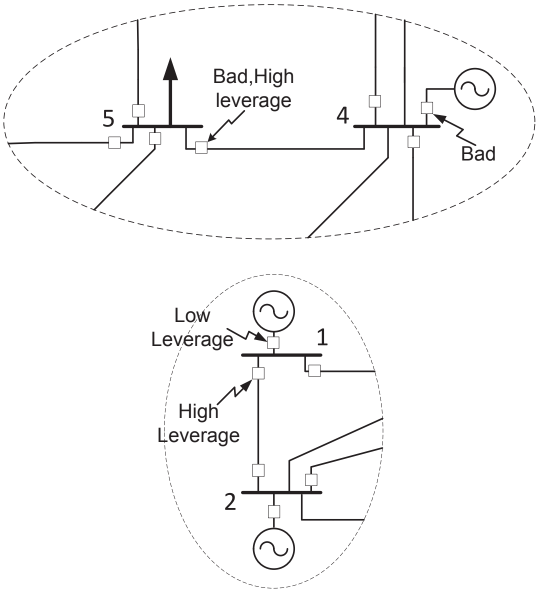

Fig. 2. The sample high and low leverage points in IEEE-14 bus system

D. Masking and Swamping

The leverage points in regression studies carry with them an inherent difficulty. When there are more than one influential point, some of them may remain undetected. This phenomenon is known as masking. On the other hand, some of the non-influential points may be wrongly detected as influential points, which is known as swamping. The masking/swamping can be explained by the following equation.

The residual for the i^{th} measurement with two high leverages at z_i and z_k is expressed as

$$r_i = (1 - K_{ii})\Delta z_i - K_{ik}\Delta z_k - \sum_{\substack{j=1 \\ j\neq i, j\neq k}}^{m} K_{ij}\Delta z_j \tag{8}$$

So, if the first two terms in equation (8) are opposite in sign a bad leverage may appear like a good leverage. This is known as masking. On the other hand, if the second and the third terms in the same equation are added up the good leverage may become a bad leverage. This is known as swamping.

The masking and swamping phenomena have been reported in the literature as in Hawkins, Bradu and Kass (HBK) data, Brownlee's stack loss data [23], [24], Hadi and Simonoff (HS) data, Belgian Telephone data, etc. In the HBK data there are 75 observations, 14 high leverage points and 10 outliers with points 11-14 are swamped cases. The Brownlee's data shows, however, that there are 21 observations with 4 outliers (cases 1,3,4,21) and 4 high leverage points (cases 1,2,3,21). There are two points which are masked and on the other hand point 17 is swamped. This swamping or masking phenomenon is, however, not present when there is only one influential measurement. This phenomenon is similar to the largest normalized residual (LNR) test to identify outliers. Hence, when there are multiple outliers or bad data or influential points the largest normalized residual test is deemed unsuitable. To the best of my knowledge, the masking/swamping phenomenon has not been investigated in the context of power system state estimation. This sets up the motivation to devise a method by which the high leverage points, low leverage points and outliers or bad data are completely separated and identified.

III. Detection of Leverage and Bad Data Points

Leverage values are normally denoted as measures of influential observations in the X-space. The X-space is in regard to the regressor variables. The hat matrix in (5) gives a measure of how a particular measurement is influential or not. The ones which have higher influence are called high leverages and ones which have lower influence are called low leverages. The twice-the-mean rule and thrice-the-mean rule on the diagonal elements of the hat matrix have been reported in the literature to identify the leverage points. Reference [23] has mentioned the Cook's distance and Welsch and Kuh's distance to detect and identify the single leverage point. The Mahalanobis distance based on the projection pursuit algorithm for minimum volume ellipsoid cannot be applied to sparse systems. Since the electric power system is a sparse system, the projection pursuit algorithm has to be modified in order to be applied to the sparse power system. However, due to masking or swamping effect it becomes difficult to identify the group of high leverage points.

A. Diagnostic robust generalized potentials (DRGP)

This technique, an adaptive approach to identify the group of leverage points, is a unified approach of diagnostic and robust approaches. The robust approach identifies the suspected high leverage points and the diagnostic approach confirms the above suspicion. The robust approach identifies the leverage points by the corresponding potentials of the data. The potential of a data is defined by Hadi [26] as the diagonal element of the hat matrix with the i^{th} data deleted. It is denoted by

$$pot_{ii} = h_i^T (H_{(i)}^T H_{(i)})^{-1} h_i \tag{9}$$

The points having a potential value more than the robust cut-off $Median(pot_{ii}) + c.MAD(pot_{ii})$ is said to be a high leverage point, where, MAD is the median absolute deviation from the median and c is a constant equal to 2 or 3. However, this method is not robust against swamping. Habshah et al [24] have proposed a robust method to identify high leverage points. The robust Mahalanobis distance (RMD_i) is defined as

$$RMD_i = \sqrt{[h_i - \bar{H}]^T [C(H)]^{-1} [h_i - \bar{H}]} \tag{10}$$

where, $\bar{H}$ is the mean of the l points for which determinant of the covariance matrix (MCD) is minimum or $\bar{H}$ is the centre of the minimum volume ellipsoid (MVE) covering these points, and $C(H)$ is the corresponding covariance matrix. The cut-off value for a normal distributed multivariate data is $\sqrt{\chi^2_{n,\alpha}}$, but, for general non-normal data the cut-off value as suggested in [24], [27] is given by

$$Median(RMD_i) + 3MAD(RMD_i) \tag{11}$$

The observations are grouped in two sets. Those which have robust Mahalanobis distance greater than the cut-off as in Eq.(11) are considered to be in set D and the rest in set R. The robust potentials for the observations in two sets are given as

$$pot^*_{ii} = \begin{cases} \dfrac{K_{ii}^{-(D)}}{1 - K_{ii}^{-(D)}} \ \forall\ i \in R \\ K_{ii}^{-(D)} \ \forall\ i \in D \end{cases} \tag{12}$$

$K_{ii}^{-(D)}$ denotes the i^{th} diagonal element of the hat matrix with data as in set D deleted. There exists no theoretical distribution for pot^*_{ii} and hence, there is no finite upper bound. However, [24], [27] suggested a suitable confidence bound type cut-off like

$$Median(pot^*_{ii}) + c.MAD(pot^*_{ii}) \tag{13}$$

The Mahalanobis distances of the multivariate data are first calculated. The Mahalanobis distance, however, is prone to the masking effect of multiple leverage data points [28]. Fig. 3 shows the step-by-step procedure for the identification of leverage points.

1) The robust Mahalanobis distances of the observations of the multi-variate data are carried out based on minimum volume ellipsoid (MVE) or minimum covariance determinant (MCD). Conceptually, MVE is the ellipsoid with minimum volume that contains l data points. MCD is, however, the minimum of the determinant of the covariance matrix which contains l points. l is typically equal to $[3m/2] + 1$ (where $3m$ is the number of data points). MVE has been considered here.
2) The multi-variate data are grouped into two separate subsets R and D. The observations which have a distance higher than the cut-off as in (11) are deleted from the main set and kept in a separate set called the deleted set D. The rest of the data are kept as it is in a set called R.
3) The generalized robust potentials for both the sets are computed.
4) If all the observations in the deleted set D have their generalized potentials higher than the cut-off, then the leverage points are identified. If not, data are put back to set R sequentially starting with the one which has the least generalized robust potential value.
5) The generalized potential values are recalculated with the new subsets.
6) This process continues till all the data in the set D have generalized potential values more than the cut-off.

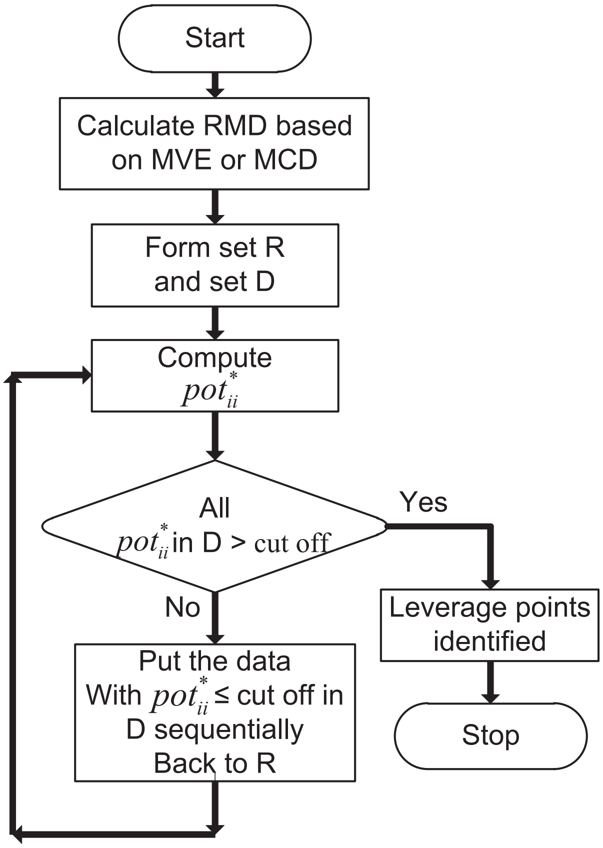
Start
Calculate RMD based on MVE or MCD
Form set R and set D
Compute pot^*_{ii}
All pot^*_{ii} in D > cut off
Yes
No
Leverage points identified
Stop
Put the data With pot^*_{ii} ≤ cut off in D sequentially Back to R

Fig. 3. A flowchart showing the identification of leverage points

By this process, the masking and swamping effects, if present, are completely taken care of and the high leverage and non-leverages are separated from each other.

B. Identification of gross error and high leverage points simultaneously

The measurements in a generic power system can be easily tampered with for nefarious purposes. The physical meters in the system can be compromised by introducing a large error by intelligent hackers. As discussed in Section II-B, the residuals as in (7) for leverage measurements are close to zero. The important class of M-estimators, including the LAV estimator, cannot handle bad leverage data points. Hence, it is very difficult to identify the gross error in case of leverage measurements. Other estimators like LMS, LTS, RLS, Iterative RLS, BOFOLS, etc. are computationally intensive. Table I shows that DRGP has better performance in terms of computational time and complexity.

TABLE I
COMPUTATIONAL COMPLEXITY OF ESTIMATORS

Algorithm	Computational time (secs)	Complexity
LMS	5.543	$O(L)$
RLS	10.234	$O(L^2)$
DRGP	2.436	$O(L)$

The residuals in the measurement data are functionally related to the leverage values of the data. This method is a combination of direct and indirect approach of multiple outlier detection. The low leverages and high leverages are separated first based on DRGP and then generalized studentized residuals (GSR) is calculated for the entire data set to identify the outliers. So, an outlier in set R will not be confused with an outlier in set D. They are defined as

$$r^*_{st,i} = \begin{cases} \dfrac{r_i^{-(D)}}{\hat{\sigma}_{R-i}\sqrt{1-K_{ii}^{-(D)}}} \ \forall\, i \in R \\ \dfrac{r_i^{-(D)}}{\hat{\sigma}_{R}\sqrt{1+K_{ii}^{-(D)}}} \ \forall\, i \in D \end{cases} \tag{14}$$

where, $\hat{\sigma}^2$ is the least square estimate of variance. $r_i^{-(D)}$ represents the residual of i^{th} measurement with D data set deleted. R is the data set without the high leverages.

The GSR is a form of a Student's t-statistic with $(3m-n-3-1)$ degrees of freedom and 97.5% detection confidence probability. One could, therefore, use a t-table to get the exact cut-off values. But since the degrees of freedom are usually quite large, the rule of thumb that absolute value of externally studentized residuals is greater than 3 is used [29]. The GSR is a type of an externally studentized residual. This is a way of determining the i^{th} residual except the i^{th} observation. If the i^{th} observation is a serious outlier it may influence the least square function and may influence to move it close to the i^{th} observation. So, if it is removed, the i^{th} residual on the new model will indicate that this observation is an extreme value. The mathematical background for the studentized residuals are given in Appendix B. All the observations for both the data sets are then plotted in a DRGP-GSR plot. High leverage points are the points which have higher DRGP values and bad data are those data which have higher GSR values. This leverage-residual plot shows that most of the data will be clustered around the origin and the masking/swamping effects do not come into picture. The DRGP-GSR plot clearly separates and identifies the bad measurement data and high leverages. Even if high leverage measurements are adulterated with gross errors the graphical plot clearly identifies the measurement errors. Based on this concept, the next section shows some case studies both for power transmission system and power distribution system and thus justifies the effectiveness of the procedure.

IV. RESULTS AND DISCUSSIONS

A. Case Studies

The problem formulation shown in Section II is a three-phase formulation suitable for generic distribution systems. However, the formulation for balanced transmission systems can be taken as a special case of the above formulation, where, the number of state variables and the number of equations as given in Section II-A are reduced due to the balanced nature of the system. The voltage magnitudes and angles for a particular bus will be the same for the three different phases. The proposed approach has been performed on test systems: a small illustrative example, the IEEE 14-bus system

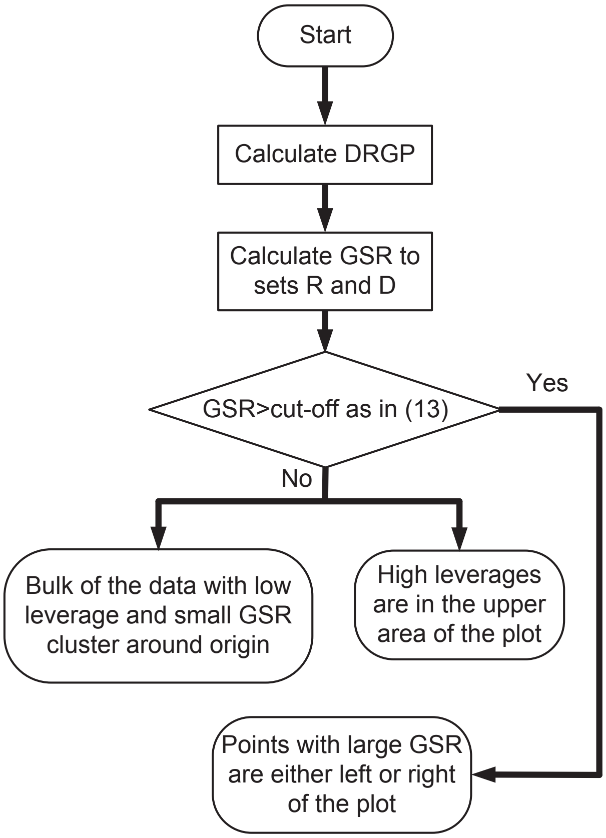
Start
Calculate DRGP
Calculate GSR to sets R and D
GSR>cut-off as in (13)
Yes
No
Bulk of the data with low leverage and small GSR cluster around origin
High leverages are in the upper area of the plot
Points with large GSR are either left or right of the plot

Fig. 4. A schematic diagram of the DRGP-GSR plot

and the IEEE 123-bus distribution system. The algorithm was implemented in MATLAB and run on a system with Intel Xeon processor @3.33 GHz and 12 GB RAM.

1) Illustrative example: Fig. 5 shows a basic four bus system with possible power injection and branch power flow measurements. All branches are assumed to have a reactance of j0.1 pu. The state variables of the system are considered as voltage magnitudes and voltage angles of buses. Since, there are four buses in the system altogether there are eight state variables. However, the voltage angle for bus #1 is taken as the reference. Table II presents the measurements for the given system. The system, currently, has no leverage points. However, if the line between 1-2 is shortened by decreasing the reactance to j0.01 pu, the measurements flow 1-2 and inj 1 become isolated leverage points. An attacker can introduce a leverage point attack on the system by tampering with the reactance of the line 1-2, should he/she wishes to attack inj 1 and/or flow 1-2. If the line 2-3 is shortened and the injection measurement is on bus 1 instead of 3, the measurements flow 3-2 and inj 3 will become leverage points.

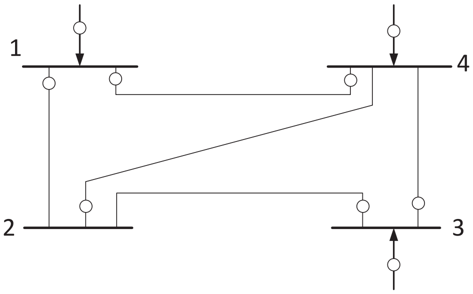

Fig. 5. A 4-bus system for illustrative example

These two measurements become bad leverage points in the factor space. The largest normalized residuals (LNR) method fails to identify these two bad leverage points. It turns out from Table II that the generalized studentized residuals clearly detect or identify the bad measurements in case of leverage points. The value of the studentized residual corresponding to the bad measurements with respect to other measurements is much higher compared to that of the normalized residual with respect to other measurements. Table III further shows the masking/swamping effect of leverage points, if any. It also compares the leverage diagnostics of diagonal elements of the hat matrix with the DRGP technique proposed in Section III-A. It depicts that while the leverage measure (diagonal element of the hat matrix) fails to identify the leverage points due to masking/swamping effect the DRGP technique can easily identify them. Table IV and Table V present the results for the active power injection and reactive power flow measurements when the line 2-3 is shortened.

TABLE II
REAL POWER MEASUREMENTS AND RESIDUALS FOR THE 4-BUS SYSTEM WHEN THE LINE 1-2 IS SHORTENED

Measurement type	Measurement with no bad data	Measurement with bad data	Normalized/ Internally studentized residuals	$\lvert GSR \rvert$ (2.228)
flow 1-2	1.50882	**1.00892**	0.550	**2.578**
flow 1-4	0.49119	0.49119	0.4793	0.4328
flow 2-4	0.33966	0.33966	0.2987	0.578
flow 3-2	-0.56915	-0.56915	1.2921	1.374
flow 3-4	-0.23084	-0.23084	0.2373	0.4328
flow 4-1	-0.49119	-0.49119	0.7821	0.8921
inj 1	2.00011	**1.50011**	0.3034	**2.781**
inj 3	-0.800	-0.800	0.5082	0.7811
inj 4	-0.600	-0.600	0.6821	0.852

TABLE III
LEVERAGE POINTS AND MASKING/SWAMPING EFFECT FOR REAL POWER MEASUREMENTS WHEN LINE 1-2 IS SHORTENED

Measurement type	Masking or Swamping effect	Leverage (0.726)	DRGP (0.823)	Bad Data
flow 1-2	Yes	0.3172	**0.8763**	Yes
flow 1-4	No	0.2988	0.3126	No
flow 2-4	No	0.6309	0.4312	No
flow 3-2	No	0.3180	0.2182	No
flow 3-4	No	0.3257	0.5278	No
flow 4-1	No	0.5587	0.6721	No
inj 1	Yes	0.3272	**0.8450**	Yes
inj 3	No	0.2238	0.2994	No
inj 4	No	0.4592	0.2994	No

2) IEEE 14-bus system: Fig.6 shows a typical IEEE 14-bus system. It is a typical meshed transmission network. The network parameters and load data are given in [30]. There are five generation buses in the system. The loads are modelled as a combination of constant impedance (Z), constant current (I) and constant power (P) loads, which is known as the ZIP model. The measured variables are power injection and branch

TABLE IV
REAL POWER MEASUREMENTS AND RESIDUALS FOR THE 4-BUS SYSTEM WHEN THE LINE 2-3 IS SHORTENED

Measurement type	Measurement with no bad data	Measurement with bad data	Normalized/ Internally studentized residuals	$\lvert GSR \rvert$ (2.228)
flow 1-2	1.50882	1.50882	0.5813	0.1243
flow 1-4	0.49119	0.49119	0.5343	0.7923
flow 2-4	0.33966	0.33966	0.2453	1.265
flow 3-2	-0.56915	**-0.17119**	0.497	**2.567**
flow 3-4	-0.23084	-0.23084	1.2643	1.8809
flow 4-1	-0.49119	-0.49119	0.8702	0.811
inj 1	2.00011	2.00011	0.7982	0.1284
inj 3	-0.800	**-0.400**	0.530	**2.879**
inj 4	-0.600	-0.600	0.7033	0.4252

TABLE V
LEVERAGE POINTS AND MASKING/SWAMPING EFFECT FOR REAL POWER MEASUREMENTS WHEN LINE 2-3 IS SHORTENED

Measurement type	Masking or Swamping effect	Leverage (0.726)	DRGP (0.823)	Bad Data
flow 1-2	No	0.3810	0.4491	No
flow 1-4	No	0.3279	0.4318	No
flow 2-4	No	0.3692	0.3268	No
flow 3-2	Yes	0.6523	**0.856**	Yes
flow 3-4	No	0.5781	0.7284	No
flow 4-1	No	0.5432	0.3067	No
inj 1	No	0.4789	0.3104	No
inj 3	Yes	0.4872	**0.894**	Yes
inj 4	No	0.5890	0.4321	No

TABLE VI
COMPARISON OF STUDENTIZED RESIDUALS WITH OTHER RESIDUALS FOR 14-BUS SYSTEM

Measurement	Semi-studentized residuals (2.3)	Internally studentized residuals (3.0)	Externally studentized residuals (3.0)	DFFITS (1.782)	Cook's distance (1.00)
flow 2-1	0.8864	0.8208	0.9445	0.4176	0.4279
flow 3-2	2.2952	1.4276	2.1169	0.6811	0.7892
flow 2-4	0.8099	0.9821	0.349	0.9031	0.1404
flow 1-5	0.4656	0.5793	1.5759	1.2042	0.4107
flow 5-2	2.9818	**3.7311**	2.0244	1.4321	0.1201
flow 5-4	**2.7264**	**3.1437**	**5.4399**	0.6478	0.4197
flow 5-6	1.7476	1.3681	0.3057	1.1573	0.1691
flow 4-7	0.8080	0.6435	1.9444	0.7921	0.3198
flow 8-7	0.6419	0.8092	1.2097	0.7695	0.4180
flow 9-7	1.0385	1.4952	0.4564	1.4502	0.3179
flow 9-10	0.1676	0.1280	0.4745	1.2998	0.0981
flow 6-11	0.7222	0.8211	1.1889	0.8931	.4193
flow 13-6	0.4754	0.1704	0.5142	1.672	0.1801
flow 10-11	0.7417	0.7411	1.0561	0.8701	0.3153
flow 13-14	1.5130	1.3711	1.3913	0.7982	0.3172
inj 1	1.6799	**3.4143**	**6.0186**	0.7921	0.8793
inj 4	0.3914	0.3719	**4.3473**	1.2983	0.9168
inj 8	0.6934	0.5489	0.9061	1.4042	0.9082
inj 10	0.5051	0.4301	0.3792	1.2763	0.4193
inj 12	0.0713	0.1032	0.0393	0.6822	.3812
inj 14	1.8547	1.432	2.0724	1.4731	0.8932
flow 1-2	2.0240	2.4191	2.0664	0.7291	0.8911
flow 5-1	1.6224	1.4522	1.5051	0.4321	0.7821
flow 4-3	1.8094	1.2480	2.3539	0.4126	0.1794
flow 7-8	1.6933	1.3421	2.1443	1.3279	0.7891
flow 9-4	0.3187	0.2819	1.1517	1.2792	0.6871
flow 10-9	0.1167	0.3179	0.057	0.9110	0.7981
flow 14-9	0.9419	0.3183	0.3926	0.2479	0.4729
flow 13-12	0.2060	0.1261	1.3136	1.593	0.7911
inj 2	2.6871	0.4271	1.9129	0.495	0.4792
inj 6	0.2100	0.2721	2.3154	1.110	0.6871
inj 7	1.4043	1.3211	0.234	0.4729	0.8862
inj 11	0.4903	0.4302	0.1057	0.4380	0.6621
inj 13	0.7720	0.8711	1.342	0.1793	0.6911

TABLE VII
THE GSR-DRGP APPROACH AND LNR APPROACH

Measurement	Normalized Residuals	Leverages identified by DRGP	GSR-DRGP approach	Bad Data
flow 5-4	2.7264	Yes	5.4399	Yes
inj 4	0.3914	Yes	4.3473	Yes
flow 1-2	2.0240	Yes	2.0664	No
inj 2	2.6871	No	1.9129	No
inj 1	1.6799	No	6.0186	Yes

power flows. The measurements are shown in Table VI. The measurements are generated by adding random Gaussian noise to the single-phase load flow results. The gross errors are generated by changing the value of the corresponding diagonal element of the hat matrix K_{ii}. The change in the K_{ii} value reflects a change in the corresponding measurement z_i. The details are given in Appendix A.

The cut-off values for all the potential values and the studentized residuals are shown in Table VIII. It shows that DRGP correctly identifies the leverage data points while the potential and the leverage values (i.e. diagonal entries of the hat matrix) fails to identify the leverage measurements correctly and instead swamps some non-leverage measurements as leverage and masks some leverage measurements as non-leverage for 14-bus system. Table VII justifies the fact, with some key measurements shown with text arrows in Fig. 8, 9, that DRGP technique with GSR properly identifies the bad data for leverage measurements, however, the normalized residuals fail to do so. Table VIII further shows the GSR of the measurements and thus, validates the effectiveness of the strategy. Table IX justifies the fact that DRGP is robust against swamping or masking effect. While the robust Mahalanobis distance masks some high leverage points as low leverages, the DRGP identifies all the high leverages correctly. The above strategy is robust against the size of the system and can be applied to larger standard systems such as IEEE-30 and IEEE-118 bus system. The next subsection provides the results for a standard large but meshed distribution 123-bus system.

3) Distribution system: The IEEE 123-bus test distribution system has also been considered for this study. The network parameters and load data are obtained from [31], [32]. The topologies of the test systems are shown in Fig. 7. The voltage level of the system is 4.16 kV. There are both three-phase and single-phase loads. Thus, the system is inherently unbalanced. The three-phase loads are either star or delta connected. The loads are either constant current or constant impedance or constant power. The loads in the system have been modelled as ZIP-model. The test system consists of both overhead lines and underground cables. The overhead lines and underground cables have been modelled with modified Carson's equations [33]. The distribution feeder is either three-phase or three-phase with a grounded neutral or single or

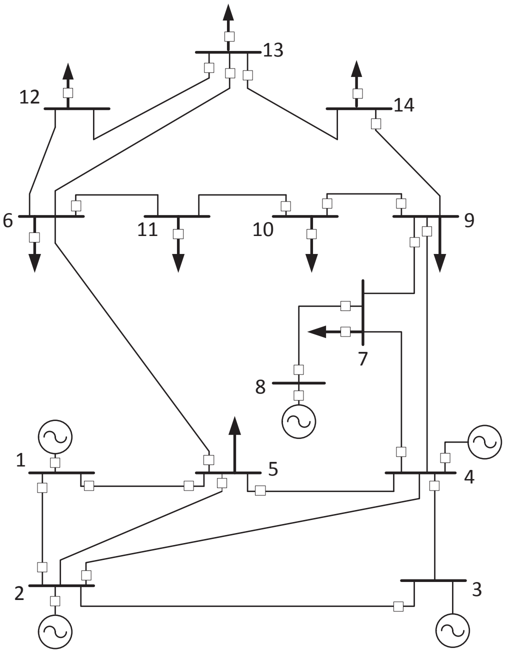

Fig. 6. IEEE-14 bus system

two-phase laterals. Therefore, the impedance of each overhead line or underground cable is represented as either a 3x3 or a 4x4 matrix compared to a single element in single phase representation. However, the 4x4 matrix for three-phase lines with grounded neutral is converted to 3x3 matrix by Kron's reduction [33]. A three-phase transformer is modelled as three individual single-phase transformers. The tap changers are considered to have fixed taps. The switches between buses 14 and 117, 61 and 118, 19 and 116, 98 and 119 are considered closed. The closed switches between buses 55 and 95 and 123 and 121 makes the system meshed in nature.

The IEEE-123 test system has been modified to incorporate some leverage data points in the measurement data set. Injection measurements are added on buses 14, 19 and 55 and the lines 9-14 and 19-22 are made short.

The measurements are generated by adding random Gaussian noise to the three-phase load flow results. The percentage error in real measurements is 3-5% and that in pseudo measurements is 20%. It should be noted that the number of pseudo measurements is larger than real measurements (about 60% of the total). With the increase in percentage error of measurements the estimation error will be high. So, for 30% and 40% error in pseudo measurements the estimation error will be much higher than for 20% error in pseudo measurements. The accuracy of the estimates of voltage magnitudes have been shown in Fig. 13-15. Table XI further shows the sum of squares of error for 20% and 30% error in pseudo measurements. The gross errors are generated by changing the value of K_{ii} as explained in Section II-C. Appendix A states the details.

The main advantage of this method is that it can separate and simultaneously identify the bad data points (outliers) and the leverages and, therefore, can be easily applied to the measurement set even if the high leverages are affected by gross error. These are reported here.

TABLE VIII
GENERALIZED POTENTIALS AND STUDENTIZED RESIDUALS FOR 14 BUS SYSTEM

Measurement No.	Measurement	Leverage (0.758)	DRGP (0.927)	GSR(3.0)
1	flow 2-1	0.5907	0.0297	-0.9445
2	flow 3-2	0.1943	0.05	2.1169
3	flow 2-4	0.6339	0.3652	-0.349
4	flow 1-5	**0.8152**	0.0677	-1.5759
5	flow 5-2	0.6124	0.5727	2.0244
6	**flow 5-4 (bad, high leverage)**	0.2519	**1.6442**	**5.4399**
7	flow 5-6	0.23	0.5057	-0.3057
8	flow 4-7	0.0729	0.1049	-1.9444
9	flow 8-7	0.461	0.4617	-1.2097
10	flow 9-7	0.5467	0.363	-0.4564
11	flow 9-10	0.4373	0.6811	-0.4745
12	flow 6-11	0.3406	0.4543	-1.1889
13	flow 13-6	0.5156	0.5602	0.5142
14	flow 10-11	0.5181	0.3177	1.0561
15	flow 13-14	0.5432	0.3027	-1.3913
16	**inj 1 (bad, low leverage)**	**0.7782**	0.0777	**6.0186**
17	**inj 4 (bad)**	**0.9065**	**0.9884**	**-4.3473**
18	inj 8	0.1671	0.0932	-0.9061
19	inj 10	0.5674	0.1214	-0.3792
20	inj 12	0.189	0.2737	0.0393
21	inj 14	0.0955	0.582	-2.0724
22	**flow 1-2 (good, high leverage)**	**0.8927**	**2.5896**	2.0664
23	flow 5-1	0.3601	0.0423	1.5051
24	flow 4-3	0.367	0.2795	-2.3539
25	flow 7-8	0.5296	0.3688	2.1443
26	flow 9-4	0.6162	0.2918	1.1517
27	flow 10-9	0.2802	0.4598	-0.057
28	flow 14-9	0.5296	0.4396	0.3926
29	flow 13-12	0.3329	0.2044	-1.3136
30	inj 2	**0.9115**	0.9173	-1.9129
31	inj 6	0.6663	0.0108	2.3154
32	inj 7	0.2052	0.6888	0.234
33	inj 11	0.4908	0.117	0.1057
34	inj 13	0.4658	0.0744	-1.342

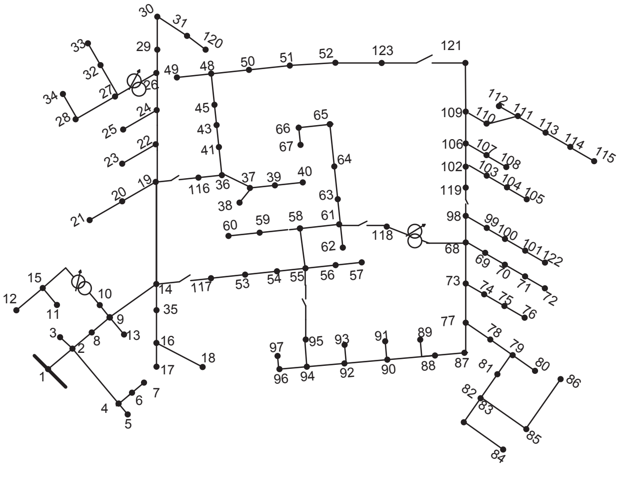

Fig. 7. IEEE 123-bus distribution system

The proposed methodology is carried out for other scenarios

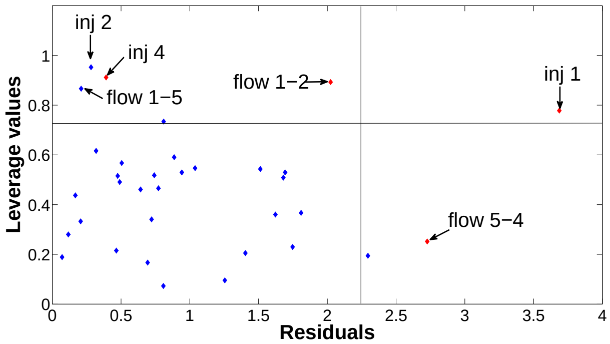

Fig. 8. Leverage vs Residual plot for 14 bus system

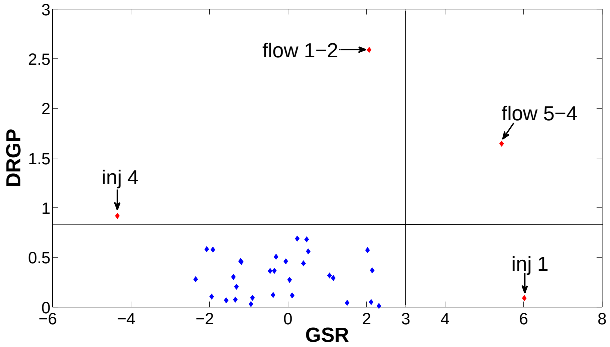

Fig. 9. DRGP vs GSR plot for 14 bus system

TABLE IX
MASKING OR SWAMPING EFFECT FOR 14-BUS SYSTEM

Measurement	Identified by RMD	Identified by DRGP	Actual lever-ages	Bad Data
flow 5-4	No	Yes	Yes	Yes
inj 4	Yes	Yes	Yes	Yes
flow 1-2	Yes	Yes	Yes	No
inj 2	No	No	No	No
inj 1	Yes	No	No	Yes

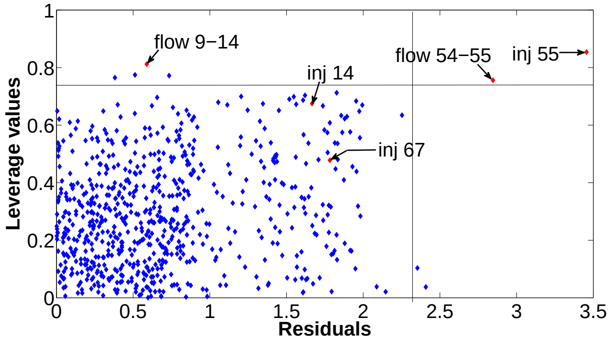

Fig. 10. Leverage vs Residual plot for 123 bus system

TABLE X
DRGP AND GSR FOR 123 BUS SYSTEM

Measurement	Leverage (0.736)	DRGP (0.853)	GSR(3.0)	Bad Data
inj 55	0.854	1.7924	4.586	Yes
flow 54-55	0.756	1.8595	3.673	Yes
inj 14	0.675	0.9595	-4.457	Yes
flow 9-14	0.812	1.2595	2.0670	No
inj 67	0.478	0.5595	5.5465	Yes
inj 36	0.798	0.657	1.967	No

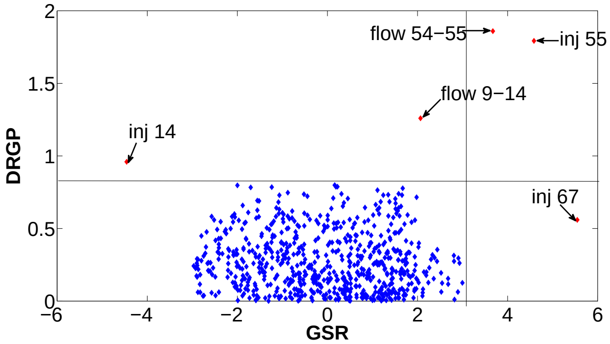

Fig. 11. DRGP vs GSR plot for 123 bus system

in a Monte Carlo simulation for both IEEE 14 bus system and IEEE 123 bus system. The boxplot for two scenarios for IEEE 14 bus system is presented in Fig.16. It shows that the outliers in both scenarios are correctly identified.

B. Discussions

The DRGP vs GSR graphs for 14 bus and 123 bus are shown in Fig.9 and Fig.11 respectively. The positions of high leverage points, low leverage points, outliers on high leverage points and outliers on low leverage points are shown clearly. The high leverages and the bad data points are shown in red in the figures. As the bulk of the data are low leverages with low residuals, most of the data points lie around the origin. The points with high leverages are located in the upper area of the plot and the data points with large residuals lie either in the left or right of the plot. This is explained in the schematic in Fig.4. The measurements marked in red are highlighted in bold in Table VIII. Table X shows the measurements marked in red for 123-bus system. The high leverage measurement (flow 5-4 in 14-bus system and inj 55 and flow 54-55 in 123-bus system) which contains gross error are located at the top right corner of the graph. The low leverage (inj 1 in 14-bus system and inj 67 in 123-bus system) with gross error is located at the extreme right end of the $x-axis$ of the graph. The high leverages (flow 1-2 in 14-bus system and flow 9-14 in 123-bus system) which are not contaminated with gross errors are located at the top of the graph. Fig.8 and Fig.10 show the plot of the leverage values (i.e. diagonal entries of the hat matrix) against the square of the normalized residuals. The same cases shown in red in Fig.9 and Fig.11 are shown in red here. However, here, the cases (inj 1, inj 2 and flow 1-5 in 14-bus system and inj 14 in 123-bus system) are swamped and the case (flow 4-3 in 14-bus system and inj 67 in 123-bus system) shows a large normalized residual. It is evident from the figure that it is difficult to differentiate the outliers from the high leverage points. Due to masking/swamping effect some measurements are misrepresented as high leverages and vice versa. The key measurement points are shown with red data points and text arrows in the Fig. 8-11.

The above method has been applied to small 4-bus example, balanced 14-bus system and unbalanced IEEE 123-bus systems. The generalized studentized residual has been used instead of the normalized residuals to identify the bad

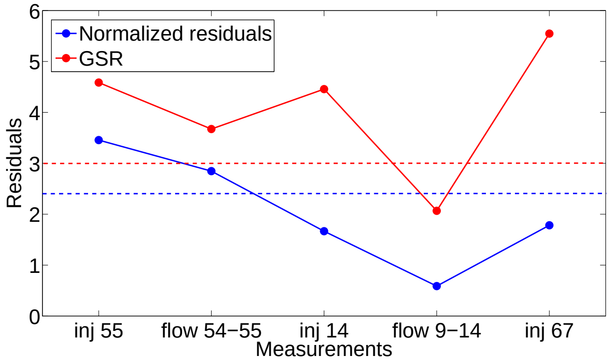

Fig. 12. Comparison of GSR and normalized residuals of key measurements for 123-bus system

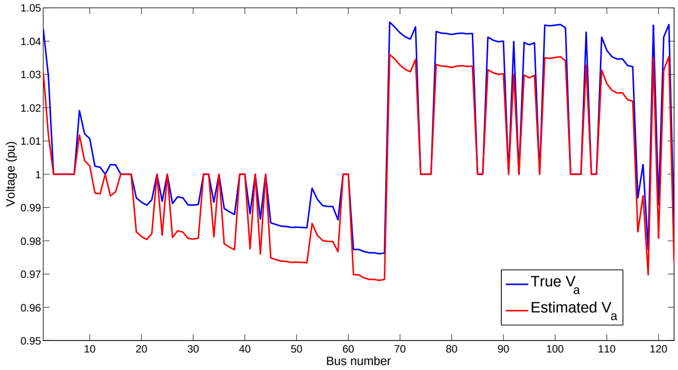

Fig. 13. True and estimated voltages of phase a for IEEE 123 bus system

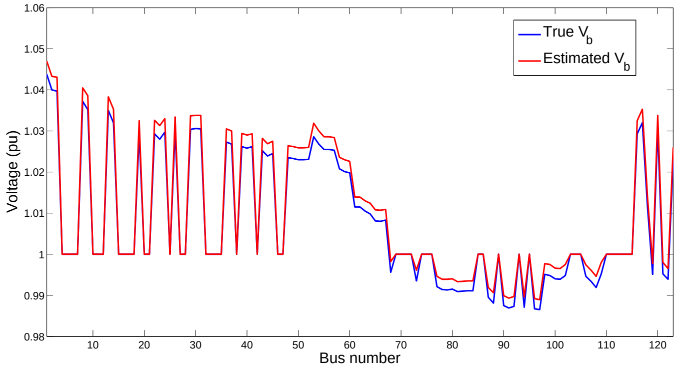

Fig. 14. True and estimated voltages of phase b for IEEE 123 bus system

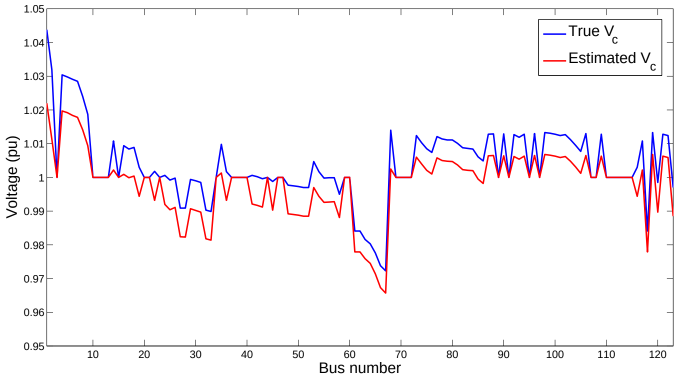

Fig. 15. True and estimated voltages of phase c for IEEE 123 bus system

TABLE XI
COMPARISON OF ACCURACY FOR ESTIMATED VOLTAGES

Results	20% error in pseudo measurements	30% error in pseudo measurements
SSE of bus voltages at phase a	0.00277	0.00315
SSE of bus voltages at phase b	0.00545	0.00587
SSE of bus voltages at phase c	0.00596	0.00645

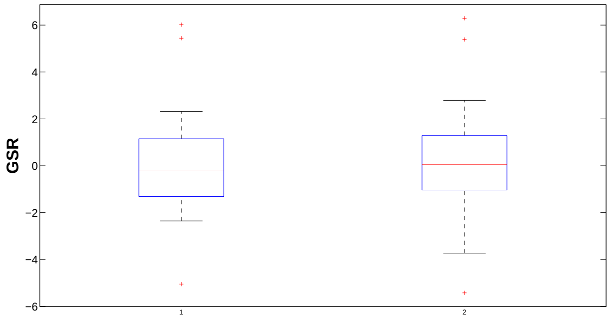

Fig. 16. Boxplot of scenarios in a Monte Carlo simulation

data. Even if the normalized/internally studentized residuals are low the GSR, for the false data, is significant. However, the above method has been compared with the normalized residual test to identify bad data. Fig. 9,11 and Fig. 8,10 justify the effectiveness of the above algorithm. Table VI compares the normalized/internally studentized residuals and other measures with externally studentized residuals for the 14-bus case, while Fig. 12 shows the comparison of normalized residuals and GSR for key measurements in case of 123-bus distribution system. From both cases it can be inferred that while the largest normalized residual test fails to separate the outliers from the high leverages and thus fails to identify the bad data when there are multiple influential data points, the simultaneous technique of DRGP and GSR clearly separates the high leverages, low leverages and measurement outliers from one other and also prevents the masking or swamping effect in the presence of multiple influential data points. Thus the method has the capability to deal with deliberate man-made attack.

V. CONCLUSION

It is always necessary to detect erroneous measurements in active power network operation. Due to growing deployment of ICT and automation technologies to operate modern power systems the measurements can be tampered for mala fide intentions. The attacker will always try to influence the states of the system by hiding the attack from the detection algorithm, which is possible if the high leverage measurements are especially targeted. The high leverages can occur in both transmission and distribution networks. The research reported here has used the concept of regression analysis to identify the outliers and influential measurements in the system. It has been found that identifying the bad data for leverage measurements is particularly difficult due to the low value of residuals even if they are infected with gross errors. In addition, if there are multiple leverage measurements some of the high leverage measurement points may be masked or some of the non-influential measurement points may be swamped. Hence, in order to take care of this masking and swamping effect, the concept of diagnostic-robust generalized potential has been proposed to separate the leverage measurements from rest of the measurements and then the studentized residuals are applied on the measurements to identify the bad data for multiple high leverage measurements. Moreover, even if there are large errors in high leverage measurements it will be possible to identify them. Comprehensive results and comparative studies on both transmission and distribution systems/balanced and unbalanced systems further show the advantages of this methodology against other existing residual techniques to identify bad data against leverage attack. The proposed method can assist the EMS/DMS in taking control and operation decisions in these scenarios.

Appendix A
Theorems on Attack Strategy

Let, $\bar{H}$ is defined as $\bar{H} = R^{-1/2}H$, and $\bar{H}_i$ is the i^{th} row of $\bar{H}$

Theorem 1: Let ϵ be the threshold and $\sigma_{i=1,\ldots,3m}$ be the variance of errors in the $J(\hat{x})$ test. Given any set of measurements z, it is guaranteed to pass the $J(\hat{x})$ test when $\sum_{i=1}^{3m}(1-K_{ii})\sum_{j=1}^{3m}(z_j^2/\sigma_j^2) \leq \epsilon$.

Theorem 2: Suppose the original set of measurements z can bypass the $J(\hat{x})$ test. When the measurement z_i in z is perturbed into $z_i^{attacked}$ by the attacker, there always exists a new value $K_{ii}^{attacked} \in (K_{ii}, 1]$, such that the new measurement set $z^{attacked}$ is guaranteed to bypass the $J(\hat{x})$ test.

Theorem 3: Let K_{ii} be the i^{th} diagonal element of hat matrix K, then,

$$(1-K_{ii})^2 \leq \frac{\left\| \begin{bmatrix} \bar{H}_p \\ \bar{H}_f \end{bmatrix} \right\|_2^2}{\left\| \bar{H}_i^T \right\|_2^2}$$

where $\bar{H}$ is partitioned as: $\bar{H} = \left[\bar{H}_p \bar{H}_i \bar{H}_f\right]^T$.

An attacker can increase the value of K_{ii} by just increasing the l_2-norm of $\bar{H}_i^T$. Since, $\bar{H}_i = 1/\sigma_i.H_i$, it gives rise to three rules

Rule 1: Increase the absolute values of elements in H_i.

Rule 2: Decrease the value of σ_i.

Rule 3: Increase the number of non-zero elements in H_i.

The proofs of the theorems are given in [25].

Therefore, there is a relationship between the measurement z_i and the corresponding K_{ii}. Let the attacked measurement be denoted by $z_i^{attacked}$ and the attacked corresponding diagonal element of the hat matrix be $K_{ii}^{attacked}$. Then, $\Delta K_{ii} = K_{ii}^{attacked} - K_{ii}$. Hence, the change in the value of K_{ii} reflects a change in the value of the corresponding z_i.

A smaller σ_i indicates a higher accuracy measurement. A higher accuracy measurement is more likely to become a leverage measurement and thus has a higher chance of getting attacked. From Theorem 2, it is clear that with a small change in the value of K_{ii} can make the attack successful against measurements with larger value of K_{ii}. Hence, the leverage measurements are more susceptible to successful attacks.

Appendix B
Studentized Residuals

The diagnostics for single case influential observations are ineffective in case of multiple influential observations due to masking/swamping effects. The phenomenon of masking and/or swamping has been explained in Section II-D. Let the set of deleted cases be D and the set of remaining cases be R. When a group of observations is deleted

$$K_{ii}^{-(D)} = h_i^T(H_R^T H_R)^{-1}h_i$$

$K_{ii}^{-(D)}$ is the i^{th} diagonal element of the $H(H_R^T H_R)^{-1}H^T$ matrix. Most of the outlier detection methods separate the clean observations from the potential outliers.

When an additional point i is added to the set R, according to [27], [34]

$$K_{ii}^{-(D)+i} = h_i^T(H_R^T H_R + h_i h_i^T)^{-1}h_i = \frac{K_{ii}^{-(D)}}{1+K_{ii}^{-(D)}}$$

The new state variables with the additional point i in the set R is given by

$$\begin{aligned}\Delta\hat{x}_{R+i} &= (H_R^T H_R + h_i h_i^T)^{-1}(H_R^T \Delta z_R + h_i \Delta z_i) \\ &= \Delta\hat{x}_R + \frac{(H_R^T H_R)^{-1}h_i}{1+K_{ii}^{-(D)}} r_{st,i}^*\end{aligned}$$

Let $r_i^{-(D)}$ be the i^{th} deletion residual.

$$r_{st,i}^{*,R+i} = \frac{r_i^{-(D)}}{\hat{\sigma_R}\sqrt{1+K_{ii}^{-(D)}}}$$

The variances of the observations in the basic subset and outside the basic subset are given [35] as:

$$\begin{aligned}1 - h_i^T(H_R^T H_R)^{-1}h_i,&\ i \in R \\ 1 + h_i^T(H_R^T H_R)^{-1}h_i,&\ i \notin R\end{aligned}$$

The studentized residuals for the two subsets are given as

$$\frac{r_i^{-(D)}}{\hat{\sigma_R}\sqrt{1 - h_i^T(H_R^T H_R)^{-1}h_i}},\ i \in R$$

$$\frac{r_i^{-(D)}}{\hat{\sigma_R}\sqrt{1 + h_i^T(H_R^T H_R)^{-1}h_i}},\ i \notin R$$

References

[1] C. Gomez-Quiles, A. Gomez-Exposito, and A. de la Villa Jaen, “State estimation for smart distribution substations,” *IEEE Transactions on Smart Grid*, vol. 3, pp. 986–995, June 2012.

[2] S. Bi and Y. J. Zhang, “Graphical methods for defense against false-data injection attacks on power system state estimation,” *IEEE Transactions on Smart Grid*, vol. 5, pp. 1216–1227, May 2014.

[3] K. C. Sou, H. Sandberg, and K. Johansson, “Computing critical k -tuples in power networks,” *IEEE Transactions on Power Systems*, vol. 27, pp. 1511–1520, Aug 2012.

[4] M. Gol and A. Abur, "Lav based robust state estimation for systems measured by pmus," *IEEE Transactions on Smart Grid*, vol. 5, pp. 1808–1814, July 2014.

[5] A. Abur and A. Exposito, *Power System State Estimation, Theory and Implementation*. CRC Press, 2004.

[6] F. Pasqualetti, F. Dorfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC), 2011*, pp. 2195–2201, Dec 2011.

[7] O. Kosut, L. Jia, R. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, pp. 645–658, Dec 2011.

[8] L. Jia, O. Kosut, R. Thomas, and L. Tong, "Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures," in *First IEEE International Conference on Smart Grid Communications (SmartGridComm), 2010*, pp. 220–225, Oct 2010.

[9] Q. Yang, J. Yang, W. Yu, N. Zhang, and W. Zhao, "On a hierarchical false data injection attack on power system state estimation," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pp. 1–5, Dec 2011.

[10] S. Bi and Y. Zhang, "Defending mechanisms against false-data injection attacks in the power system state estimation," in *GLOBECOM Workshops (GC Wkshps), 2011 IEEE*, pp. 1162–1167, Dec 2011.

[11] G. Hug and J. Giampapa, "Vulnerability assessment of ac state estimation with respect to false data injection cyber-attacks," *IEEE Transactions on Smart Grid*, vol. 3, pp. 1362–1370, Sept 2012.

[12] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Transactions on Smart Grid*, vol. 2, pp. 326–333, June 2011.

[13] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks: characterizations and countermeasures," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, pp. 232–237, Oct 2011.

[14] E. Asada, A. Garcia, and R. Romero, "Identifying multiple interacting bad data in power system state estimation," in *Power Engineering Society General Meeting, 2005. IEEE*, pp. 571–577 Vol. 1, June 2005.

[15] P. Koponen, L. D. Saco, N. Orchard, T. Vorisek, J. Parsons, C. Rochas, A. Z. Morch, V. Lopes, and M. Togeby, "Definition of smart metering and applications and identification of benefits," *Deliverable D3 of the European Smart Metering Alliance ESMA (available at www. esmahome. eu, members area)*, 2008.

[16] A. Monticelli, *State Estimation in Electric Power Systems, A Generalized Approach*. Luwer's power Electronics and power Systems Series, 2004.

[17] J. Chen and A. Abur, "Placement of pmus to enable bad data detection in state estimation," *IEEE Transactions on Power Systems*, vol. 21, pp. 1608–1615, Nov 2006.

[18] A. Narvaez and S. Grijalva, "Robust state estimator applied to the ecuadorian electric power system," in *Transmission and Distribution Conference and Exposition: Latin America, 2008 IEEE/PES*, pp. 1–6, Aug 2008.

[19] L. Mili, V. Phaniraj, and P. Rousseeuw, "Least median of squares estimation in power systems," *IEEE Transactions on Power Systems*, vol. 6, pp. 511–523, May 1991.

[20] W. J. Egan and S. L. Morgan, "Outlier detection in multivariate analytical chemical data," *Analytical chemistry*, vol. 70, no. 11, pp. 2372–2379, 1998.

[21] M. Meloun and J. Militky, "Detection of single influential points in ols regression model building," *Analytica Chimica Acta*, vol. 439, no. 2, pp. 169–191, 2001.

[22] L. Mili, M. Cheniae, N. S. Vichare, and P. Rousseeuw, "Robust state estimation based on projection statistics of power systems," *IEEE Transactions on Power Systems*, vol. 11, pp. 1118–1127, May 1996.

[23] A. Nurunnabi, A. S. Hadi, and A. Imon, "Procedures for the identification of multiple influential observations in linear regression," *Journal of Applied Statistics*, vol. 41, no. 6, pp. 1315–1331, 2014.

[24] M. Habshah, M. Norazan, and A. Rahmatullah Imon, "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression," *Journal of Applied Statistics*, vol. 36, no. 5, pp. 507–520, 2009.

[25] S. Tan, W.-Z. Song, M. Stewart, and L. Long, "Lpattack: Leverage point attacks against state estimation in smart grid," in *Global Communications Conference (GLOBECOM), 2014 IEEE*, pp. 643–648, Dec 2014.

[26] A. S. Hadi, "A new measure of overall potential influence in linear regression," *Computational Statistics & Data Analysis*, vol. 14, no. 1, pp. 1–27, 1992.

[27] A. Rahmatullah Imon, "Identifying multiple influential observations in linear regression," *Journal of Applied statistics*, vol. 32, no. 9, pp. 929–946, 2005.

[28] L. Mili, M. Cheniae, N. Vichare, and P. Rousseeuw, "Robust state estimation based on projection statistics [of power systems]," *IEEE Transactions on Power Systems*, vol. 11, pp. 1118–1127, May 1996.

[29] "Advanced regression diagnostic methods." https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/pt3_adv_regression_diagnostics.pdf.

[30] "Power systems test case archive." http://www.ee.washington.edu/research/pstca/.

[31] "Distribution test feeders." http://ewh.ieee.org/soc/pes/dsacom/testfeeders/index.html.

[32] W. Kersting, "Radial distribution test feeders," in *Power Engineering Society Winter Meeting, 2001. IEEE*, vol. 2, pp. 908–912 vol.2, 2001.

[33] W. H. Kersting, *Distribution system modeling and analysis*. CRC press, 2012.

[34] C. R. Rao, *Linear statistical inference and its applications*, vol. 22. John Wiley & Sons, 2009.

[35] A. S. Hadi and J. S. Simonoff, "Procedures for the identification of multiple outliers in linear models," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1264–1272, 1993.

Ankur Majumdar (S'12) received his PhD in Electrical Power Engineering in June 2016 from Imperial College London and his B.E.E.(with honours) degree from Jadavpur University, Calcutta, India, in Electrical Engineering in 2009 and the M.Tech. degree from the Indian Institute of Technology, Delhi, India in Electric Power Systems in 2011. Currently, he is a Research Associate in the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. His research interests include state estimation, smart grid security and power system analysis.

Bikash C. Pal (M'00-SM'02-F'13) received the B.E.E.(with honours) degree from Jadavpur University, Calcutta, India, the M.E. degree from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree from the Imperial College London, London, U.K, in 1990, 1992, and 1999, respectively, all in electrical engineering. Currently, he is a Professor in the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. His current research interests include state estimation, power system dynamics, and flexible ac transmission system controllers. He is the Editor-in-Chief of IEEE TRANSACTIONS ON SUSTAINABLE ENERGY and a Fellow of the IEEE for his contribution to power system stability and control.