Real-time Fault Diagnosis for Large-Scale Nonlinear Power Networks

Wei Pan, Ye Yuan†, Henrik Sandberg, Jorge Gonçalves and Guy-Bart Stan

***Abstract*— In this paper, automatic fault diagnosis in large scale power networks described by second-order nonlinear swing equations is studied. This work focuses on a class of faults that occur in the transmission lines. Transmission line protection is an important issue in power system engineering because a large portion of power system faults is occurring in transmission lines. This paper presents a novel technique to detect, isolate and identify the faults on transmissions using only a small number of observations. We formulate the problem of fault diagnosis of nonlinear power network into a compressive sensing framework and derive an optimisation-based formulation of the fault identification problem. An iterative reweighted ℓ_1-minimisation algorithm is finally derived to solve the detection problem efficiently. Under the proposed framework, a real-time fault monitoring scheme can be built using only measurements of phase angles of nonlinear power networks.**

I. INTRODUCTION

Power networks are large-scale spatially distributed systems. Being a critical infrastructure, they possess strict safety and reliability constraints [1]. The behaviour of a synchronous electrical motor located in a bus of the power networks can be described by the so-called *swing equation* [2]. In [3], [4], [5], a common technique to detect the faults in the system described above is to generate a set of residuals to indicate the presence of faults either in a centralised or distributed fashion. However, it is well-known that a large portion of power system faults occurring in transmission lines doesn't involve additive faults only. In addition, the introduction of more components into the network will increase the vulnerability of the transmission lines. Thus it is critically important in power networks to automatically diagnose the faults that occur.

In terms of fault diagnosis, one of the major goals is to detect, isolate and identify the faults as soon as possible. One objective here is then to present a method that allows to address the recommendations provided by the IFAC Technical Committee *SAFEPROCESS*, namely, we want to determine whether there is an occurrence of a fault and the time of its occurrence (*detection*), determine the kind, location and time of detection of a fault (*isolation*), determine the size and time-variant behaviour of a fault (*identification*).

W. Pan and G.-B. Stan are with Centre for Synthetic Biology and Innovation and the Department of Bioengineering, Imperial College London, United Kingdom.Y. Yuan and J. Gonçalves are with Control Group, Department of Engineering, University of Cambridge, United Kingdom. H. Sandberg is with Automatic Control Laboratory, School of Electrical Engineering, Royal Institute of Technology (KTH), Sweden.

The authors gratefully acknowledge the support of Microsoft Research through the PhD Scholarship Programs of Wei Pan. Ye Yuan and Jorge Gonçalves acknowledge the support from EPSRC (project EP/I03210X/1). Guy-Bart Stan gratefully acknowledges the support of the EPSRC Centre for Synthetic Biology and Innovation at Imperial College London through the Science and Innovation award (project EP/G036004/1).

†Corresponding author, email: yy311@cam.ac.uk.

The dynamics of the buses in the power networks can be described by the so-called swing equation where the active power flows are nonlinear functions of the phase angles. Due to the typical large scale of power networks and the introduction of nonlinearities, fault diagnosis over the transmission lines is a very challenging problem. To the best knowledge of the authors, such a problem is seldom addressed from a system and control perspective. In this paper, our approach draws inspiration from the fields of signal processing and machine learning by combining compressive sensing and variational Bayesian inference techniques together to offer an efficient method for fault diagnosis.

Contributions: We formulate the problem of fault diagnosis of nonlinear power networks with additive noise into a sparse signal recovery problem. We derive a sparse Bayesian formulation of the fault identification problem which is casted into a nonconvex optimisation problem. We relax the nonconvex optimisation problem into a convex problem and develop an iterative reweighted ℓ_1-minimisation algorithm to solve it efficiently.

II. MODEL FORMULATION

A. Power Networks Model

Power systems are examples of very complex systems in which generators and loads are dynamically interconnected. Thus they can be seen as networked systems, where each bus can be viewed as a node in the network. We assume that all the buses in the network are connected to synchronous machines (motors or generators). Based on these common assumptions, we provide the nonlinear model for the active power flow in a power grid branch connected between bus i and bus j. For $i = 1, \dots, N$, the behaviour of bus/node i can be represented by the swing equation [2]

$$m_i \ddot{\delta}_i(t) + d_i \dot{\delta}_i(t) - P_{mi}(t) = -\sum_{j=1}^{n} P_{ij}(t), \quad (1)$$

where δ_i is the phase angle of bus i, m_i and d_i are the inertia and damping coefficients respectively, P_{mi} is the mechanical input power, P_{ij} is the active power flow from bus i to j.

Considering that there are no power losses nor ground admittances and letting $V_i = |V_i|e^{j\delta_i}$ be the complex voltage of bus i, the active power flow between bus i and bus j, P_{ij}, is given by:

$$P_{ij}(t) = w_{ij}^{(1)} \cos(\delta_i(t) - \delta_j(t)) + w_{ij}^{(2)} \sin(\delta_i(t) - \delta_j(t)), \quad (2)$$

where $w_{ij}^{(1)} = |V_i||V_j|G_{ij}$ and G_{ij} is the branch conductance between bus i and bus j; $w_{ij}^{(2)} = |V_i||V_j|B_{ij}$ and B_{ij} is the branch susceptance between bus i and bus j.

If we let $\xi_i(t) = \delta_i(t)$ and $\zeta_i(t) = \dot{\delta}_i(t)$, each bus i can be assumed to have double integrator dynamics

$$\begin{aligned} \dot{\xi}_i(t) &= \zeta_i(t), & (3)\\ \dot{\zeta}_i(t) &= u_i(t) + v_i(t), & (4) \end{aligned}$$

where ξ_i, ζ_i are the scalar states, $v_i(t)$ is a scalar known external input, and u_i is the control given by the nonlinear control law

$$v_i(t) = \frac{P_{mi}(t)}{m_i} \quad (5)$$

$$u_i(t) = -\frac{d_i}{m_i}\zeta_i(t) - \frac{1}{m_i}\sum_{j=1}^{N}[w_{ij}^{(1)}\cos(\xi_i(t) - \xi_j(t)) + w_{ij}^{(2)}\sin(\xi_i(t) - \xi_j(t))]. \quad (6)$$

The variables ξ_i and ζ_i can be interpreted as phase and frequency in the context of power networks.

In [6] and other papers, the $\cos(\cdot)$ terms (no branch conductance between buses) is not considered and further assumption such as phase angles are close is made. In this case, system (1) can be linearised around some equilibrium point as

$$m_i\ddot{\delta}_i(t)+d_i\dot{\delta}_i(t)-P_{mi}(t) = -\sum_{j\in N_i} w_{ij}^{(2)}(\delta_i(t)-\delta_j(t)). \quad (7)$$

If each bus in the power network corresponds to a node then the underlying graph of this network can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{i\}_1^N$ is the set of nodes and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the edge set of the graph. The undirected edge $\{i, j\}$ is incident on vertices i and j if nodes i and j share a transmission line, and a positive weight is associated with this link. Moreover, $N_i = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$ is the neighborhood set of i. Each bus i is also assumed to have double integrator dynamics as described in (3) and (4). The difference is that $u_i(t)$ in (6) becomes as a linear equation

$$u_i(t) = -\frac{d_i}{m_i}\xi_i(t) - \frac{1}{m_i}\sum_{j\in N_i} w_{ij}^{(2)}(\xi_i(t) - \xi_j(t)). \quad (8)$$

B. Fault Diagnosis in Power Networks

For the linearised system (8), a bus $k \in \mathcal{V}$ is faulty if for some functions $f_{\xi k}(t)$ and $f_{\zeta k}(t)$ not identical to zero either $\dot{\xi}_i(t) = \zeta_i(t) + f_{\xi k}(t)$, or $\dot{\zeta}_i(t) = u_i(t) + v_i(t) + f_{\zeta k}(t)$. The functions $f_{\xi k}(t)$ and $f_{\zeta k}(t)$ correspond to fault signals. A fruitful model-based or observer-based fault diagnosis methods are available for power networks (see [6] and reference therein). However, specific aspects need careful consideration when dealing with fault diagnosis in power networks. Firstly, the simplified linear model can be often applied in practice when the phase angles are close. However, when the system is strained and faults are of large magnitude, angles can be far apart. As a result, the linear model cannot used to approximate the nonlinear model in (1) anymore. Secondly, power networks are highly distributed and interconnected, not more than one transmission line tend to be faulty simultaneously. Thirdly, to be more realistic, the noise should be incorporated into (1) as

$$m_i\ddot{\delta}_i(t) + d_i\dot{\delta}_i(t) - P_{mi}(t) = -\sum_{j=1}^{n} P_{ij}(t) + \eta_i(t), \quad (9)$$

Then we can rewrite the state space mode (3) and (4) as:

$$\begin{aligned} \dot{\xi}_i(t) &= \zeta_i(t), & (10)\\ \dot{\zeta}_i(t) &= u_i(t) + v_i(t) + \eta_i(t), & (11) \end{aligned}$$

where n_i are assumed to be Gaussian noise with zero mean and known variance σ^2.

III. PROBLEM FORMULATION

Given the model and explanation above, we primarily focus on the following setting in this paper.

Definition 1: If the dynamics of a power network can be described by (10) and (11), the transmission line between bus i and bus j is faulty when $w_{ij}^{(1)}$ changes to a new scalar $\tilde{w}_{ij}^{(1)}$ and/or $w_{ij}^{(2)}$ changes to a new scalar $\tilde{w}_{ij}^{(2)}$, where $w_{ij}^{(1)}$ and $w_{ij}^{(2)}$ are the weight for $\cos$ and $\sin$ terms defined in (6).

Problem 1: Based on the considerations above and Definition 1, the problem that we are interested in solving is the following: knowing the measurements and the distribution of the noise, how to estimate for the faults, namely, $\forall i, j$, $w_{ij}^{(1)} - \tilde{w}_{ij}^{(1)}$ and $w_{ij}^{(2)} - \tilde{w}_{ij}^{(2)}$ using a very small number of samples.

To address Problem 1, we have the following assumption.

Assumption 1: The power networks described by (10) and (11) is fully measurable, i.e., the phase angles of all the buses can be measured.

A. Model Transformation

Applying the standard forward Euler discretisation to (10) and (11) and assuming the discretisation step $t_{k+1} - t_k = \Delta t$ is constant for all k, we obtain the following discrete-time system approximation to the continuous-time system:

$$\begin{aligned} \frac{\xi_i(t_{k+1}) - \xi_i(t_k)}{\Delta t} &= \zeta_i(t_k), & (12)\\ \frac{\zeta_i(t_{k+1}) - \zeta_i(t_k)}{\Delta t} &= u_i(t) + v_i(t) + \eta_i(t_k). & (13) \end{aligned}$$

If we let

$$e_i(t_{k+1}) \triangleq -\left(\frac{(\zeta(t_{k+1}) - \zeta(t_k))}{\Delta t} + \frac{d_i\zeta(t_k)}{m_i} - \frac{P_{mi}(t_k)}{m_i}\right), \quad (14)$$

we have

$$e_i(t_{k+1}) = \frac{1}{m_i}\sum_{j\in N_i}[w_{ij}^{(1)}\cos(\xi_i(t_k) - \xi_j(t_k)) + w_{ij}^{(2)}\sin(\xi_i(t_k) - \xi_j(t_k))] + \eta_i(t_k), \quad (15)$$

where e_i, the power flow measurement, is treated as the output of the system.

We let $\mathbf{x}(t_k) = [\xi_1(t_k), \dots, \xi_N(t_k)]$ and write (14) into a vector form

$$e_i(t_{k+1}) = f_i(\mathbf{x}(t_k))\mathbf{w}_i^{true} + \eta_i(t_k), \quad (16)$$

with

$$\begin{aligned} f_i(\mathbf{x}(t_k)) &= [f_i^{(1)}(\mathbf{x}(t_k)), f_i^{(2)}(\mathbf{x}(t_k))] \in \mathbb{R}^{1\times 2N},\\ f_i^{(1)}(\mathbf{x}(t_k)) &= [\cos(\xi_i(t_k) - \xi_1(t_k)), \dots, \cos(\xi_i(t_k) - \xi_N(t_k))],\\ f_i^{(2)}(\mathbf{x}(t_k)) &= [\sin(\xi_i(t_k) - \xi_1(t_k)), \dots, \sin(\xi_i(t_k) - \xi_N(t_k))],\\ \mathbf{w}_i^{true} &= [\mathbf{w}_i^{(1)}, \mathbf{w}_i^{(2)}]^T \in \mathbb{R}^{2N\times 1},\\ \mathbf{w}_i^{(1)} &= [w_{i1}^{(1)}, \dots, w_{iN}^{(1)}],\\ \mathbf{w}_i^{(2)} &= [w_{i1}^{(2)}, \dots, w_{iN}^{(2)}], \end{aligned}$$

where $f_i(\mathbf{x}(t_k))$ indicate the transmission functions and $\mathbf{w}_i$ indicate the corresponding transmission weights that represent the topology of the network.

B. Fault Diagnosis Problem Formulation

As stated in Definition 1, if there are no faults occurring in the transmission lines between bus i and other buses, the dynamics of the power networks will evolve according to (16). The expected output for the next sampling time is defined to be

$$\hat{e}_i(t_{k+1}) = f_i(\mathbf{x}(t_k))\mathbf{w}_i^{true}. \tag{17}$$

From (16) and (17), it's easy to find that $e_i(t_{k+1}) - \hat{e}_i(t_{k+1})$ is a stochastic variable with zero mean and variance σ^2. If there are faults occurring in the transmission lines between bus i and other buses, the corresponding transmission weights will change from $\mathbf{w}_i^{true}$ to $\mathbf{w}_i^{fault}$. Similar to the definition of $\mathbf{w}_i^{true}$, $\mathbf{w}_i^{fault} = [\tilde{\mathbf{w}}_i^{(1)}, \tilde{\mathbf{w}}_i^{(2)}]^T$ where $\tilde{\mathbf{w}}_i^{(1)} = [\tilde{w}_{i1}^{(1)}, \ldots, \tilde{w}_{in}^{(1)}]$ and $\tilde{\mathbf{w}}_i^{(2)} = [\tilde{w}_{i1}^{(2)}, \ldots, \tilde{w}_{in}^{(2)}]$. Then we have

$$\tilde{e}_i(t_{k+1}) = f_i(\mathbf{x}(t_k))\mathbf{w}_i^{fault} + \eta_i(t_k), \tag{18}$$

where $\tilde{e}_i$ is the output when there are faults. From (17) and (18), it's easy to find that $\tilde{e}_i(t_{k+1}) - \hat{e}_i(t_{k+1})$ is a stochastic variable with mean $f_i(\mathbf{x}(t_k))(\mathbf{w}_i^{fault} - \mathbf{w}_i^{true})$ and variance σ^2. Denoting $y_i = \tilde{e}_i - \hat{e}_i$, $\mathbf{w}_i = \mathbf{w}_i^{fault} - \mathbf{w}_i^{true}$, we have

$$y_i(t_{k+1}) = f_i(\mathbf{x}(t_k))\mathbf{w}_i + \eta_i(t_k). \tag{19}$$

In the noiseless case, when there are no faults, $\forall i$, y_i and w_i are zeros; when there are faults, certain y_i are nonzeros. So faults can be *detected* when not all y_i are zeros. However, in the noisy case, even when there are no faults, y_i is nonzero most of the time since it is a stochastic variable with zero mean. This can be interpreted in a probabilistic way by Chebyshev's Inequality, i.e. $\mathcal{P}(|e_i(t_{k+1}) - \hat{e}_i(t_{k+1})| \geq k\sigma) \leq \frac{1}{k^2}$ where $k \in \mathbb{R}^+$. When there are no faults, that explains the deviation between the expected and true outputs cannot be much greater than zero with high probability. Thus when $|\tilde{e}_i(t_{k+1}) - \hat{e}_i(t_{k+1})|$ is above a predefined threshold (much greater than σ), the faults can be *isolated* with high probability (e.g. if the threshold is set to $k\sigma = 10\sigma$, the probability is 99%). Based on the above explanations, we will later summarise the fault *detection* and *isolation* procedures in Algorithm 2. The remaining task is to identify the location of the faults or equivalently find the nonzeros entries in w_i, which is known to be the fault *identification* procedure. Assuming that M successive data points are sampled and defining

$$\begin{aligned}
\mathbf{y}_i &\triangleq [y_i(t_1), \ldots, y_i(t_M)]^T \in \mathbb{R}^M, \\
\mathbf{\Phi}_i &\triangleq \begin{bmatrix} f_i^{(1)}(\mathbf{x}(t_0)) & f_i^{(2)}(\mathbf{x}(t_0)) \\ \vdots & \vdots \\ f_i^{(1)}(\mathbf{x}(t_{M-1})) & f_i^{(2)}(\mathbf{x}(t_{M-1})) \end{bmatrix} \\
&= \begin{bmatrix} f_i(\mathbf{x}(t_0)) \\ \vdots \\ f_i(\mathbf{x}(t_{M-1})) \end{bmatrix} \in \mathbb{R}^{M \times 2N}, \\
\boldsymbol{\eta}_i &\triangleq [n_i(t_0), \ldots, n_i(t_{M-1})]^T \in \mathbb{R}^M,
\end{aligned}$$

we can write n independent equations:

$$\mathbf{y}_i = \mathbf{\Phi}_i \mathbf{w}_i + \boldsymbol{\eta}_i, \quad (i = 1, \ldots, N). \tag{20}$$

We want to find $\mathbf{w}_i$ given the output data stored in $\mathbf{y}_i$. To solve for $\mathbf{w}_i$ in (20) corresponds to solving linear regression problems that can be solved using standard least square approaches. Since there would be n independent linear regression problems, we can just consider one single problem and omit the subscripts i in (20) for simplicity of notation. We then write

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\eta}. \tag{21}$$

IV. Algorithm for solving Eq. (21)

We address the linear regression problem under the following assumption.

Assumption 2: A maximum of S transmission lines are faulty, i.e., $\mathbf{w}$ has at most S non-zero entries. In other words, $\mathbf{w}$ is S-sparse or mathematically, $\|\mathbf{w}\|_0 \leq S$. The constant S is assumed unknown to the system administrator.

Remark 1: Assumption 2 is a realistic one for small values of S since in the context of a networked system, it is typically not the case that all the transmission lines are faulty simultaneously. Alternatively, in the case when the faults are ubiquitous, we cannote that buses in power networks are typically sparsely connected thus yielding that the number of faults is relatively smaller than the size of the network N.

However, standard least square approaches cannot be used to detect these faults efficiently. Typically, one can use the pseudoinverse of $\mathbf{\Phi}$, $\mathbf{\Phi}^\dagger$, to get an estimate for $\mathbf{w}$. For affine systems such as (21), the pseudoinverse may be used to construct the solution of minimum Euclidean norm among all solutions. The estimation is generically dense (hence, violating Assumption 2) and one cannot identify which transmission lines are likely to be faulty by identifying the nonzero entries of the estimated $\mathbf{w}_i^{fault} - \mathbf{w}_i^{true}$.

To alleviate the difficulties mentioned above, the linear regression problem (21) can be defined as a compressive sensing, or sparse signal recovery problem [7], [8], with observation vector $\mathbf{y}$, known regressor matrix $\mathbf{\Phi}$, unknown coefficients $\mathbf{w}$, and additive noise $\boldsymbol{\eta}$. In sparse problems, the prior belief is that only a small fraction of the elements appearing in $\mathbf{w}$ are non-negligible. The general aim is to identify the sparsest representation for $\mathbf{w}$.

For $\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\eta}$, the likelihood of the output given $\mathbf{w}$ is

$$\mathcal{P}(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I}) \propto \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2\right]. \tag{22}$$

We define a prior distribution $\mathcal{P}(\mathbf{w})$ as follows and and compute the *posterior distribution* over $\mathbf{w}$ via Bayes' rule:

$$\mathcal{P}(\mathbf{w}) \propto \exp\left[-\frac{1}{2}g(\mathbf{w})\right] = \exp\left[-\frac{1}{2}\sum_j g(w_j)\right], \tag{23}$$

where $g(w_j)$ is an arbitrary function of w_j. We then formulate a *maximum a posteriori* (MAP) estimate:

$$\begin{aligned}
\mathbf{w}_{\mathbf{MAP}} &= \arg\max_{\mathbf{w}} \mathcal{P}(\mathbf{w}|\mathbf{y}) \\
&= \arg\min_{\mathbf{w}} \{\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \sigma^2 g(\mathbf{w})\}, \quad (24)
\end{aligned}$$

where $g(\mathbf{w})$ is defined as a penalty function.

If we define $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_N]^T \in \mathbb{R}_+^N$, we can represent the prior in the following relaxed (variational) form:

$$\mathcal{P}(\mathbf{w}) = \prod_j \mathcal{P}(w_j),\ \mathcal{P}(w_j) = \max_{\gamma_j > 0} \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j), \quad (25)$$

where $\varphi(\gamma_j)$ is a nonnegative function which is treated as a hyperprior with γ_j being its associated hyperparameters. Throughout, we call $\varphi(\gamma_j)$ the "*potential function*". We have also shown that $\log \mathcal{P}(\sqrt{w_j})$ is concave on $(0, \infty)$ [9].

For a fixed $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]$, we define a relaxed prior which is a joint probability distribution over $\mathbf{w}$ and $\boldsymbol{\gamma}$

$$\begin{aligned}\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma}) &= \prod_j \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j) \\ &= \mathcal{P}(\mathbf{w}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}) \leq \mathcal{P}(\mathbf{w}), \quad (26)\end{aligned}$$

where $\mathcal{P}(\mathbf{w}|\boldsymbol{\gamma}) \triangleq \prod_j \mathcal{N}(w_j|0, \gamma_j), \mathcal{P}(\boldsymbol{\gamma}) \triangleq \prod_j \varphi(\gamma_j)$. Since is $\mathcal{P}(\mathbf{y}|\mathbf{w})$ is Gaussian in (22), we can get a relaxed posterior which is also Gaussian

$$\mathcal{P}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}) = \frac{\mathcal{P}(\mathbf{y}|\mathbf{w})\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma})}{\int \mathcal{P}(\mathbf{y}|\mathbf{w})\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma}) d\mathbf{w}} = \mathcal{N}(\mathbf{m}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}), \quad (27)$$

where

$$\begin{aligned}\mathbf{m}_{\mathbf{w}} &= \boldsymbol{\Gamma}\boldsymbol{\Phi}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{y}, \quad (28) \\ \boldsymbol{\Sigma}_{\mathbf{w}} &= \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\boldsymbol{\Phi}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{\Phi}, \quad (29)\end{aligned}$$

with $\boldsymbol{\Gamma} \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$.

Now the key question is how to choose the most appropriate $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1, \ldots, \hat{\gamma}_N]$ to maximise $\prod_j \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j)$ such that $\mathcal{P}(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\gamma}})$ can be a "good" relaxation to $\mathcal{P}(\mathbf{w}|\mathbf{y})$. Using the product rule for probabilities, we can write the full posterior

$$\begin{aligned}\mathcal{P}(\mathbf{w}, \boldsymbol{\gamma}|\mathbf{y}) &\propto \mathcal{P}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}|\mathbf{y}) \\ &= \mathcal{N}(\mathbf{m}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}) \times \frac{\mathcal{P}(\mathbf{y}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma})}{\mathcal{P}(\mathbf{y})}. \quad (30)\end{aligned}$$

Since $\mathcal{P}(\mathbf{y})$ is independent of $\boldsymbol{\gamma}$, the quantity

$$\mathcal{P}(\mathbf{y}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}) = \int \mathcal{P}(\mathbf{y}|\mathbf{w})\mathcal{P}(\mathbf{w}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}) d\mathbf{w}$$

is the prime target for variational methods [10]. This quantity is known as evidence or marginal likelihood. A good way of selecting $\hat{\boldsymbol{\gamma}}$ is to choose it as the minimiser of the sum of the misaligned probability mass, e.g.,

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \arg\min_{\boldsymbol{\gamma} \geq \mathbf{0}} \int \mathcal{P}(\mathbf{y}|\mathbf{w}) \left|\mathcal{P}(\mathbf{w}) - \mathcal{P}(\mathbf{w}; \boldsymbol{\gamma})\right| d\mathbf{w} \\ &= \arg\max_{\boldsymbol{\gamma} \geq \mathbf{0}} \int \mathcal{P}(\mathbf{y}|\mathbf{w}) \prod_j \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j) d\mathbf{w}. \quad (31)\end{aligned}$$

The second equality is a consequence of $\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma}) \leq \mathcal{P}(\mathbf{w})$ (see (26)). The procedure in (31) is referred to as evidence maximisation or type-II maximum likelihood [11]. It means that the marginal likelihood can be maximised by selecting the most probable hyperparameters able to explain the observed data. Once $\hat{\boldsymbol{\gamma}}$ is computed, an estimate of the unknown weights can be obtained by setting $\hat{\mathbf{w}}$ to the posterior mean (28):

$$\hat{\mathbf{w}} = \mathbb{E}(\mathbf{w}|\mathbf{y}; \hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}^T)^{-1}\mathbf{y}. \quad (32)$$

Theorem 1: [9] The optimal hyperparameters $\hat{\boldsymbol{\gamma}}$ in (31) can be achieved by minimising the following cost function

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = &\log\left|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T\right| \\ &+ \mathbf{y}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{y} + \sum_j p(\gamma_j), \quad (33)\end{aligned}$$

where $p(\gamma_j) = -2\log\varphi(\gamma_j)$, and the cost function is nonconvex.

Furthermore, based on duality lemma (see Sec. 4.2 in [12]), we can create a strict upper bounding auxiliary function $\mathcal{L}_{\boldsymbol{\gamma},\mathbf{w}}(\boldsymbol{\gamma}, \mathbf{w})$ of $\mathcal{L}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$ in (31),

$$\begin{aligned}&\mathcal{L}_{\boldsymbol{\gamma},\mathbf{w}}(\boldsymbol{\gamma}, \mathbf{w}) \\ \triangleq\ &\langle\boldsymbol{\gamma}^*, \boldsymbol{\gamma}\rangle - h^*(\boldsymbol{\gamma}^*) + \mathbf{y}^T\left(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T\right)^{-1}\mathbf{y} \\ =\ &\frac{1}{\sigma^2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_j\left(\frac{w_j^2}{\gamma_j} + \gamma_j^*\gamma_j\right) - h^*(\boldsymbol{\gamma}^*). \quad (34) \\ \triangleq\ &\frac{1}{\sigma^2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_j\left(\frac{w_j^2}{\gamma_j} + \gamma_j^*\gamma_j\right). \quad (35)\end{aligned}$$

For a fixed $\boldsymbol{\gamma}^*$, we notice that $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$ is jointly convex in $\mathbf{w}$ and $\boldsymbol{\gamma}$ and can be globally minimised by solving over $\boldsymbol{\gamma}$ and then $\mathbf{w}$. Since $w_j^2/\gamma_j + \gamma_j^*\gamma_j \geq 2w_j\sqrt{\gamma_j^*}$, for any $\mathbf{w}$, $\gamma_j = |w_j|/\sqrt{\gamma_j^*}$ minimises $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$.

The next step is to find a $\hat{\mathbf{w}}$ that minimises $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$. When $\gamma_j = |w_j|/\sqrt{\gamma_j^*}$ is substituted into $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$, $\hat{\mathbf{w}}$ can be obtained by solving the following weighted convex ℓ_1-minimisation problem

$$\begin{aligned}\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}}\{\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + 2\sigma^2\sum_j u_j|w_j|\} \\ &= \arg\min_{\mathbf{w}}\{\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + 2\sigma^2\sum_j\sqrt{\gamma_j^*}|w_j|\}, \quad (36)\end{aligned}$$

where $\sqrt{\gamma_j^*}$ are the weights. We can then set

$$\gamma_j = \frac{|\hat{w}_j|}{\sqrt{\gamma_j^*}}, \forall j, \quad (37)$$

and, as a consequence, $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$ will be minimised for any fixed $\boldsymbol{\gamma}^*$.

Consider again $\mathcal{L}_{\boldsymbol{\gamma},\mathbf{w}}(\boldsymbol{\gamma}, \mathbf{w})$ in (34). For any fixed $\boldsymbol{\gamma}$ and $\mathbf{w}$, the tightest bound can be obtained by minimising over $\boldsymbol{\gamma}^*$. The tightest value of $\boldsymbol{\gamma}^* = \hat{\boldsymbol{\gamma}}^*$ equals the slope at the current $\boldsymbol{\gamma}$ of the function $h(\boldsymbol{\gamma}) \triangleq \log|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T| + \sum_j p(\gamma_j)$. Using basic principles in convex analysis, we then obtain the following analytic form for the optimiser $\boldsymbol{\gamma}^*$:

$$\begin{aligned}\hat{\boldsymbol{\gamma}}^* &= \nabla_{\boldsymbol{\gamma}}\left(\log|\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T| + \sum_j p(\gamma_j)\right) \\ &= \mathrm{diag}\left[\boldsymbol{\Phi}^T\left(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^T\right)^{-1}\boldsymbol{\Phi}\right] + p'(\boldsymbol{\gamma}), \quad (38)\end{aligned}$$

where $p'(\boldsymbol{\gamma}) = [p'(\gamma_1), \ldots, p'(\gamma_N)]^T$.

The algorithm is then based on successive iterations of (36), (37) and (38) until convergence to $\hat{\boldsymbol{\gamma}}$. We then compute the posterior mean and covariance as in (28) and (29)

$$\begin{aligned}\hat{\mathbf{w}} &= \hat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}^T)^{-1}\mathbf{y}, \quad (39) \\ \boldsymbol{\Sigma}_{\hat{\mathbf{w}}} &= \hat{\boldsymbol{\Gamma}} - \hat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}^T(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{\Phi}, \quad (40)\end{aligned}$$

where $\hat{\mathbf{\Gamma}} = \text{diag}[\hat{\gamma}]$. The above described procedure is summarised in Algorithm 1.

Algorithm 1 Reweighted ℓ_1-minimisation on hyperparameter $\boldsymbol{\gamma}$

Data: Successive observations of $\mathbf{y}$ from t_0 to t_M.
Result: Posterior mean for $\mathbf{w}$.
Step 1: Set iteration count k to zero and initialise each $u_j^{(0)} = \sqrt{\gamma_j^*}$, with randomly chosen initial values for γ_j^*, $\forall j$, e.g., with $\gamma_j^* = 1$, $\forall j$.
Step 2: At the k^{th} iteration, solve the reweighted ℓ_1-minimisation problem

$$\hat{\mathbf{w}}^{(k)} = \arg\min_{\mathbf{w}} \{\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + 2\sigma^2 \sum_j u_j^{(k)} |w_j|\}.$$

Step 3: Compute $\gamma_j^{(k)} = \frac{|\hat{w}_j^{(k)}|}{\sqrt{\gamma_j^{*(k)}}}, \forall j$.
Step 4: Update $\hat{\boldsymbol{\gamma}^*}^{(k+1)}$ using (38)

$$\hat{\boldsymbol{\gamma}^*}^{(k+1)} = \text{diag}\left[\mathbf{\Phi}^T \left(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}^{(k)}\mathbf{\Phi}^T\right)^{-1} \mathbf{\Phi}\right] + p'(\boldsymbol{\gamma}^{(k)}).$$

Step 5: Update weights $u_j^{(k+1)}$ for the ℓ_1-minimisation at the next iteration $u_j^{(k+1)} = \sqrt{\hat{\boldsymbol{\gamma}}_j^{*(k+1)}}$.
Step 6: $k \to k+1$ and iterate Steps 2 to 5 until convergence to some $\hat{\boldsymbol{\gamma}}$.
Step 7: Compute $\hat{\mathbf{w}} = \mathbb{E}(\mathbf{w}|\mathbf{y}; \hat{\boldsymbol{\gamma}}) = \hat{\mathbf{\Gamma}}\mathbf{\Phi}^T(\sigma^2\mathbf{I} + \mathbf{\Phi}\hat{\mathbf{\Gamma}}\mathbf{\Phi}^T)^{-1}\mathbf{y}$.

Based on Algorithm 1, we can summarise the fault diagnosis algorithm in Algorithm 2.

Remark 2: If a convex optimisation algorithm is used, there will be no exact zeros during the iterations and, strictly speaking, we will always get a solution with 0-*Sparsity* even when the RIP condition holds. However, some of the estimated weights will be very small compared to other weights, e.g., $\pm 10^{-3}$ compared to 1, i.e. the "energy" some of the estimated weights will be several orders of magnitude lower than the average "energy", e.g., $\|w_j\|_2^2 \ll \|\mathbf{w}\|_2^2$. Thus a threshold needs to be defined *a priori* to prune the "small" weights at each iteration. An important feature of Algorithm 1 is its very cheap algorithmic complexity since the repeated execution scales as $\mathcal{O}(MN\|\mathbf{w}^{(k)}\|_0)$ (see [13], [14]). Since at each iteration certain weights are estimated to be zero, certain dictionary functions spanning the corresponding columns of $\mathbf{\Phi}$ are pruned out for the next iteration.

Algorithm 2 Diagnosis for faults

Initialization:
Set a threshold σ^* as indicated in Section III-B, e.g. $10 \times \sigma$;
Iteration:

for $k = 0, \dots, T$ **do**
  % T is an integer indicating the count of sampling and the number of diagnosis rounds;
  Collect $\xi_i(t_k)$ and $\zeta_i(t_k)$ in (12) and (13)
  for $i = 1, \dots, N$ **do**
    Calculate the output data $e_i(t_{k+1})$ in (14);
    Calculate the expected output $\hat{e}_i(t_{k+1})$ in (17);
    if $|e_i(t_{k+1}) - \hat{e}_i(t_{k+1})| > \sigma^*$ **then**
      Fault is detected for bus i; % {*fault detection procedure*}
      Compute $y_i(t_{k+1})$ in (19);
      if $|y_i(t_{k+1})| > \sigma^*$ **then**
        Isolate bus i; % {*fault isolation procedure*}
      end if
    end if
    Set $M \leftarrow k$;
    Apply Algorithm 1 to identify the faults $\hat{\mathbf{w}}_i$; % {*fault identification procedure*}
  end for
  if $\forall i$, $\|\hat{\mathbf{w}}_i\|_0$ converge to some constant **then**
    Break;
  end if
end for

Output:
An estimate for the faults $\hat{\mathbf{w}}$ in (20), $i = 1, \dots, n$;

V. NUMERICAL STUDY

The effectiveness of our theoretic developments is here illustrated for a randomly generated power network with 20 buses. If all the buses are fully connected, the possible number of transmission lines is 380. Here we assume the number of transmission lines is 79 (the sparsity of the network is around 20%). Its dynamics can be described by nonlinear swing equations as in (10) and (11). $w_{ij}^{(1)}$ and $w_{ij}^{(2)}$ are positive real numbers as showed in Fig. 3a. Let the noise variance $\sigma^2 = 1$. All the coefficients of the model we use are selected similar as those in [2], [15].

Since the sampling frequency is around 50 Hz [2], [15], we assume the sampling interval to be $20ms$. We thus assume that the discretisation step in Section III is performed using a sampling interval $\Delta t = 20ms$. Consider the power networks model in (10) and (11). At time instant $t = 3s$, there are faults occurring in five transmission lines simultaneously. Specifically, the behaviour of the faults can described as follows: $\forall (i,j) \in \{(5,18), (7,2), (11,15), (16,18), (19,9)\}$, $w_{ij}^{(1)}$ and $w_{ij}^{(2)}$ in (6) respectively (which correspond to cos and sin terms) are set to zeros. 5 buses are involved in these transmission lines, i.e. buses 5, 7, 11, 16 and 19. Following the setup in Algorithm 2, we want to detect and isolate these 5 buses. After detection and isolation, the identification procedure will be performed. We consider $\sigma^* = 10\sigma = 10$ to initialise Algorithm 2.

First, we detect and isolate the buses with $|y_i(t_{k+1})| > \sigma^*$. It is shown in Fig. 1 that at time instant $t = 3.02s$ (only one sampling time after the faults occur), buses 5, 11, 16 and 19 can be isolated. After 3 identification rounds, bus 7 is isolated to be faulty at time instant $t = 3.06s$. Next, we identify the faults that occur in the transmission lines connecting the buses isolated, i.e. buses 5, 7, 11, 16 and 19. In Fig. 2, the time trajectory of $\|\hat{\mathbf{w}}_i\|_0$ for $i = 5, 7, 11, 16, 19$ are depicted starting at the time point $t = 3.02s$ when the faults are detected. We set the pruning threshold (mentioned in Remark 2) to 10^{-3}, i.e., $\|w_j\|_2^2/\|\mathbf{w}\|_2^2 < 10^{-3}$. We define a positive integer n^* to indicate the identification rounds which are required to terminate the identification procedure,

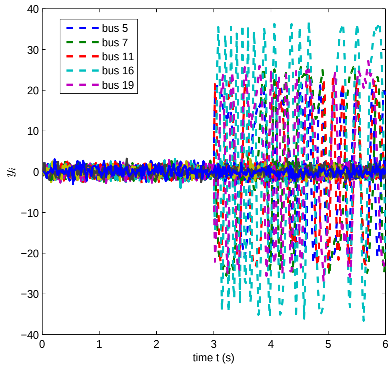


Fig. 1: Time-series of y_i for all buses. The dashed lines are indicating for bus i, $i = 5, 7, 11, 16, 19$.

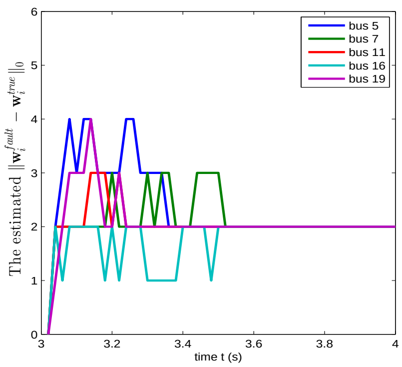
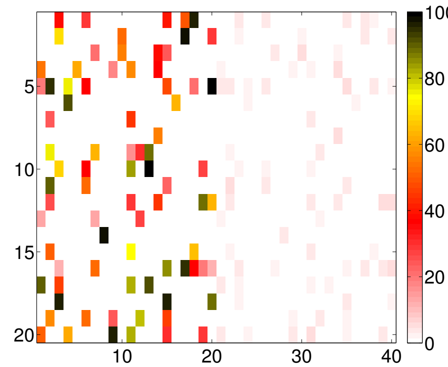
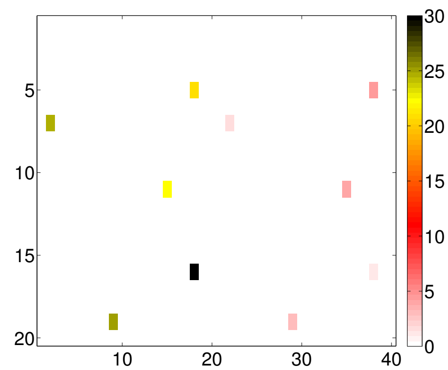


Fig. 2: Time-series of the sparsity of the estimated fault, i.e. $\|\mathbf{w}_i^{fault} - \mathbf{w}_i^{true}\|_0$ for bus $i = 5, 7, 11, 16, 19$.



(a) True weight matrix with around 20% nonzero entries. The left half of the matrix corresponds to the weights for cos terms while the right half is for sin terms.



(b) The absolute error weight matrix, which is defined as $|\mathbf{w}_i^{fault} - \mathbf{w}_i^{true}|$.

e.g. $n^* = 10$. As shown in Fig. 2, at time instant $t = 3.52s$, the sparsity of the estimated fault, i.e. $\|\mathbf{w}_i^{fault} - \mathbf{w}_i^{true}\|_0$ for bus $i = 5, 7, 11, 16, 19$ all become to 2 and keep unchanged hereafter. At time instant $t = 3.72s$, 10 rounds after $t = 3.52s$, we terminate the identification procedure. Finally at time instant $t = 3.72s$, the sparsity for all the estimated faults is stable and we stop the algorithm. In Fig. 3a and Fig. 3b, we illustrate the true weight matrix and the estimated absolute error matrix $|\mathbf{w}_i^{fault} - \mathbf{w}_i^{true}|$. As we can see, all the 5 faults that are occurring in the transmission lines have been identified with high accuracy.

VI. Conclusion

In this paper, we consider the problem of automatic fault diagnosis in large scale power networks described by second-order nonlinear swing equations. This work is in particular focusing on a class of faults that occur in the transmission lines. Later we applied tools from compressive sensing and variational Bayesian inference to detect, isolate and identify the faults.

References

[1] M. Shahidehpour, F. Tinney, and Y. Fu, "Impact of security on power systems operation," *Proceedings of the IEEE*, vol. 93, no. 11, pp. 2013–2025, 2005.

[2] P. Kundur, N. J. Balu, and M. G. Lauby, *Power system stability and control*. McGraw-hill New York, 1994, vol. 4, no. 2.

[3] M.-A. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification," *Automatic Control, IEEE Transactions on*, vol. 34, no. 3, pp. 316–321, 1989.

[4] J. Chen and R. J. Patton, "Robust residual generation using unknown input observers," *Robust model-based fault diagnosis for dynamic systems*, pp. 65–108, 1999.

[5] S. X. Ding, *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer, 2008.

[6] I. Shames, A. M. Teixeira, H. Sandberg, and K. H. Johansson, "Distributed fault detection for interconnected second-order systems," *Automatica*, 2011.

[7] E. Candès and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[8] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] W. Pan, Y. Yuan, J. Gonçalves, and G.-B. Stan, "Bayesian approaches to nonlinear network reconstruction," *submitted*.

[10] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[11] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[12] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[13] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l 1 minimisation," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[14] D. Wipf and S. Nagarajan, "Iterative reweighted l1 and l2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[15] M. Pavella, D. Ernst, and D. Ruiz-Vega, *Transient stability of power systems: a unified approach to assessment and control*. Springer, 2000.