LA-UR-12-21904
Approved for public release; distribution is unlimited.

Title: Model Bank State Estimation for Power Grids Using Importance Sampling

Author(s): Lawrence, Earl C.
Bent, Russell W.
Vander Wiel, Scott A.

Intended for: Technometrics
Web

Model Bank State Estimation for Power Grids Using Importance Sampling

Earl Lawrence*
Statistical Sciences, Los Alamos National Laboratory
and
Russell Bent
Energy and Infrastructure Analysis, Los Alamos National Laboratory
and
Scott Vander Wiel
Statistical Sciences, Los Alamos National Laboratory

June 27, 2013

Abstract

Power grid operators decide where and how much power to generate based on the current topology and demands of the network. The topology can change as safety devices trigger (connecting or disconnecting parts of the network) or as lines go down. Often, the operator cannot observe these events directly, but instead has contemporary measurements and historical information about a subset of the line flows and bus (node) properties. This information can be used in conjunction with a computational model to infer the topology of the network. We present a Bayesian approach to topological inference that considers a bank of possible topologies. The solution provides a probability for each member in the model bank. The approach has two important features. First, we build a statistical approximation, or emulator, to the computational model, which is too computationally expensive to run a large number of times. Second, we use the emulator in an importance sampling scheme to estimate the probabilities. The resulting algorithm is fast enough to use in real time and very accurate. This paper has supplementary online material.

Keywords: electric power, network modeling, simulator, surrogate model, emulator, importance sampling

*

1 Introduction

This paper presents a Bayesian procedure for estimating unobserved power grid components (downed lines, demands, intermittent generation) from a small set of direct observations of some network quantities. Modern electric power grid operations are heavily dependent on understanding the topology of the network and the demands placed upon it. For example, downed power lines may require altering the amount of power generation at various locations across the grid in order to continue meeting projected demand. Grid operators typically have only a partial observation of the state of the network (voltages, flow, consumption, production, presence/absence of some lines), so the power engineering community has devoted a lot of effort to the problem of estimating the remaining network quantities given this partial observation (Singh et al. 2005; Wu and Liu 1989; Singh and Glavitsch 1991; Singh and Alvarado 1995; Clements and Davis 1988; Vinod Kumar et al. 1996; Alves da Silva et al. 1991; Singh et al. 2010).

This state estimation problem can be succinctly stated as follows. Data on line flows, demands, and generations are collected at a sparse subset of the grid's nodes and lines. These data are used to estimate the remaining flows, demands, and generations along with possible downed lines and tripped safety devices. Operators use the state estimates as the basis for a number of decisions such as optimal generation dispatch, inter-area power exchange, and modification of the settings of automatic control devices. Further, these estimators are important for applications like grid restoration and contingency planning for assessing vulnerability of critical infrastructure, situations in which data are often the most uncertain. This paper address the power grid state estimation problem using a rigorous Bayesian formulation.

An important component of state estimation for power grids is the use of a power flow solver. As explained in more detail below, the power flow solver uses certain known quantities (perhaps known only distributionally) to estimate unknown or unobserved quantities. This presents a difficulty because the solver can be computationally slow, especially for larger networks. Any state estimation procedure may need to use it sparingly in order to provide predictions in real time. The methodology presented here is specifically designed to meet this challenge.

Many of these state estimation methods (e.g. Clements and Davis 1988; Wu and Liu 1989) are based on relatively simple statistical analysis that involves using least squares methods and an assumed network topology and then greedily altering the topology in response to large residuals. A critical assumption in all of these approaches is the existence of a single best state estimate that is used as a deterministic input to a decision making process. In practice, the operator is expected to provide a measure of quality control on the state estimates and in general this approach has served the power engineering community well. However, mistakes in this process can contribute to disastrous consequences, such as the 2003 Northeast Blackout which affected 55 million people in the United States and Canada (and led to some harrowing moments in the days immediately preceding the wedding of one of the authors). This event was caused in part by state estimator problems and operator error (U.S. - Canada Power System Outage Task Force 2004; Andersson et al. 2005). Outcomes like this provide motivation to develop a statistically sound approach to predict the network state, particularly topology (the presence or absence of power lines).

A more recent approach in Singh et al. (2010) assumes that a bank of models contains all of the important network topologies and attempts to estimate the most correct model based on the available data. Their paper frames the problem in traditional Bayesian statistical terms, which should naturally lead to a posterior distribution over possible states (versus choosing a single state). Unfortunately, this paper presents a flawed implementation of the Bayesian solution. The paper inappropriately applies the state estimation formula for a hidden Markov model (Eq 8 in the reference) to the estimation of the network state despite lacking the sequential data framework for which hidden Markov models are useful. The misunderstanding is further demonstrated when the paper states that “[o]ver the iterations one model has asymptotic probability equal to one while others have zero probabilities.” Thus, the incorrect estimation scheme robs the Bayesian formulation of its most compelling feature: the distribution over the possible states.

The key contribution of our paper is a Bayesian statistical approach to estimating the probabilities for each member of a bank of possible power grid topologies based on incomplete measurements of the network. This approach borrows ideas (but not the specific statistical approach) from the computer experiments literature (Morris et al. 1993; Kennedy and O'Hagan 2001; Higdon et al. 2005, 2008) to create a computationally efficient approximating surrogate model for use in place of the computationally expensive physics-based power flow solver. The surrogate model is used as part of an importance sampling scheme (Liu 2001) to estimate probabilities for each member of the model bank. Since the surrogate model is precomputed, the resulting scheme is fast enough to be used in real time. Further, it accurately assesses the uncertainty in the resulting state estimate.

The remainder of the paper is organized as follows. Section 2 provides a motivating example from Singh et al. (2010) and provides an overview of power flow, measurements, and the power flow solvers. Section 3 presents the surrogate model and importance sampling methodology for computing model bank probabilities. Section 4 discusses results on the motivating example. Section 5 considers an application to a larger network, an IEEE benchmark network used for testing reliability methodology (Reliability Test System Task Force 1999). Finally, Section 6 provides a discussion of the results and some ideas for the future application of statistical methodology to power grid modeling.

2 Background and Formulation

We will explain the basic problem using the small network from Singh et al. (2010) shown in Figure 1. This network consists of two subnets that are joined between node 9 of network 1 and node 12 of network 2. Each node or *bus* in the network is represented by a thick horizontal line. The triangles attached to some of the buses represent loads that consume power. Several devices (denoted by squares) can trigger automatically in order to disconnect parts of the network or link the two networks together. Power enters the network at three places: the two hashed boxes represent other power grids and the circle with a G represents a generator.

A network operator needs to make decisions on power generation and dispatch based on the network's performance and topology, in particular possible downed lines and tripped safety devices. This topological information is usually not directly available and must be inferred. Indirect information on performance and topology comes from three places. First, data are collected from monitoring devices spread across the network in relatively small numbers. In the case of our small example, power flow is monitored only on the lines

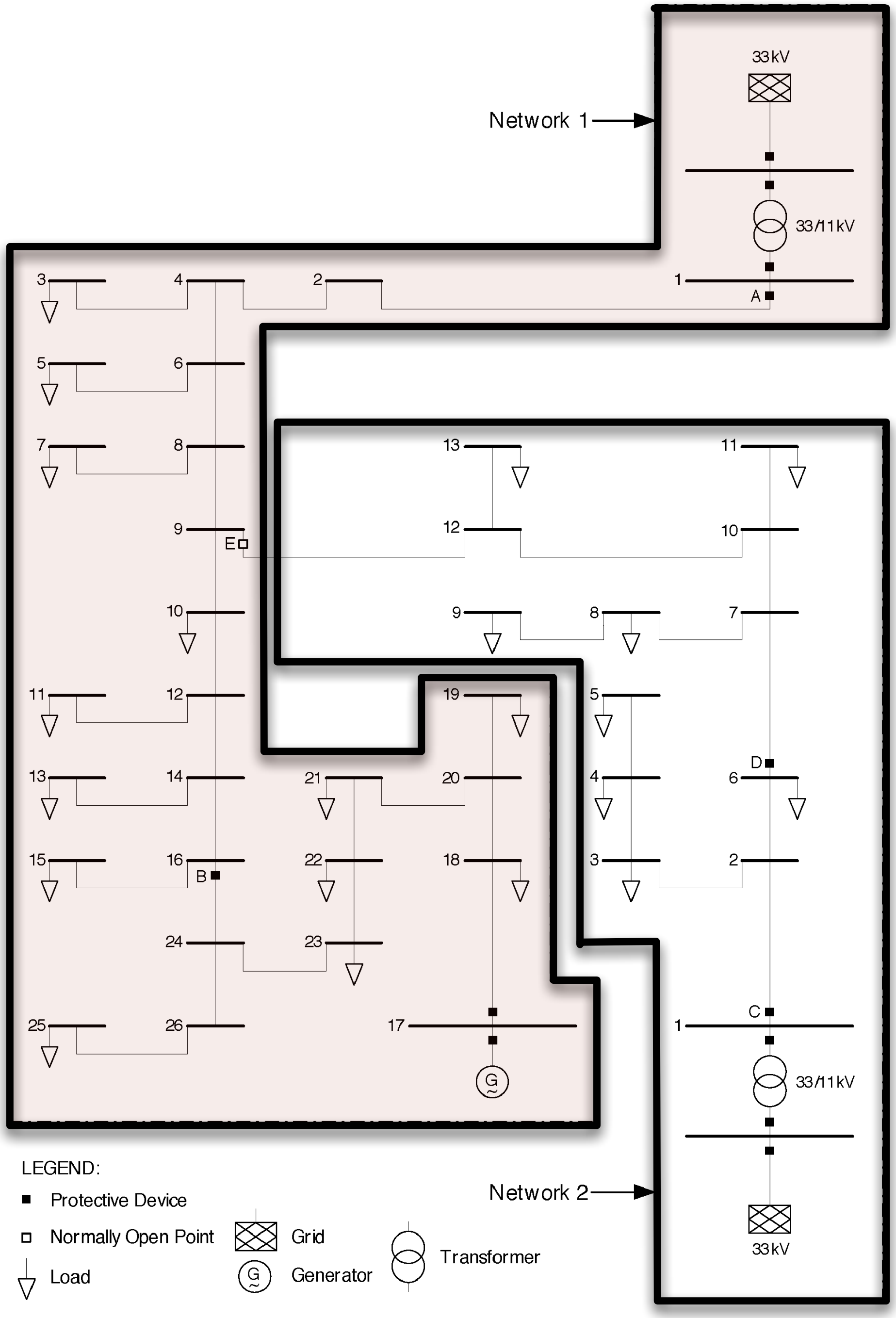

Figure 1: UK Network. Original appears in Singh et al. (2010). The online version of this figure is in color.

between buses 1 and 2 of each subnet. This amounts to monitoring the power flowing into the network from outside power grids. The relatively small number of monitoring points is a notable feature of the data. All of the power engineering community's recent estimation work cited above assumes this limitation. Further, a recent Department of Energy report (Department of Energy 2012) points out the continued sparsity of monitoring devices despite a recent push toward more highly instrumented Smart Grids. The networks are typically quite large and monitoring equipment still very new, so state estimation will continue to operate in this regime for the foreseeable future.

Second, the operator knows historical information about unmonitored loads in the form of prior distributions. The amount of power used by each load is not known exactly, but can be summarized by a distribution. Typically, these distributions are described as Gaussian with percentage errors. For example Singh et al. (2010) follow these conventions with load standard deviations equal to 20% of the mean.

Finally, the operator can use a power flow solver to compute unknown network quantities from known ones based on physical laws. The power flow solver allows the operator to match unknown network quantities with observed data through the power flow equations. We will consider two steady-state versions of these equations, the lossless alternating current (AC) and direct current (DC) power flow models. We will describe these two models in some detail below. In the next section, we will discuss how to build a statistical model around them for comparing with data.

AC Power Flow

As the name implies, an alternating current is sinusoidal and is usually represented with a complex number where the two terms, real and imaginary, are called the real and reactive power. The lossless AC power flow equations (Glover et al. 2008, Ch. 6) describe the real and reactive power flowing on the line between buses i and j:

$$\begin{aligned} P_{i,j} &= |V_i||V_j| \left[G_{i,j} \cos(\theta_i - \theta_j) + B_{i,j} \sin(\theta_i - \theta_j)\right] \\ Q_{i,j} &= |V_i||V_j| \left[G_{i,j} \sin(\theta_i - \theta_j) - B_{i,j} \cos(\theta_i - \theta_j)\right]. \end{aligned}$$

The parameters $G_{i,j}$ and $B_{i,j}$ are known quantities related to line properties called resistance and reactance and may be zero to indicate no line is present between a pair of buses. The

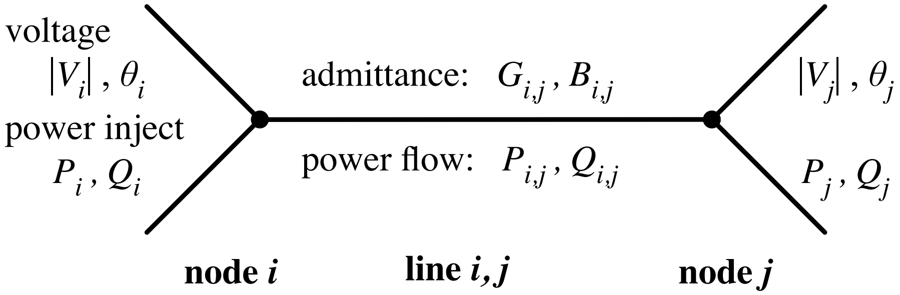

Figure 2: Summary of bus and line related quantities. Admittances and flows are line related parameters. Voltage quantities and power injections are nodal quantities.

parameter θ_i is the voltage angle at each bus i and $|V_i|$ is the voltage magnitude at each bus. The voltage angles and magnitudes are necessary to compute the $P_{i,j}$ and $Q_{i,j}$. The power flow equations are written in terms of these line flows:

$$\begin{aligned} P_i &= \sum_{j=1}^{N} P_{i,j} \\ Q_i &= \sum_{j=1}^{N} Q_{i,j}. \end{aligned}$$

P_i and Q_i are the real and reactive power, respectively, at bus i. Individual buses are generators (net power going out), loads (net power coming in), or neutral. The quantity N is the total number of buses. Figure 2 summaries the quantities associated with lines and buses.

There are some additional constraints on these equations. First, because these equations assume that no power is lost, the sum of the power across all of the buses is zero, $\sum_{i=1}^{N} P_i = 0$ and $\sum_{i=1}^{N} Q_i = 0$. In other words, all of the power generated in the system is used by the system. Second, the voltage angles only matter up to a rotation because they only enter the equations through their differences. Finally, the voltage magnitudes are also not uniquely identified. These restrictions inform the solution methodology and which parameters we attempt to observe.

There are $2N$ AC power flow equations for any network and they are written in terms of $4N$ quantities (again, $G_{i,j}$ and $B_{i,j}$ are known properties of the line). Given values for $2N$ of the variables, we can attempt to solve for the remaining $2N$ (in reality, this is not

Bus Type	Known	Unknown
slack	$\theta_1 = 0$, $\lvert V_1\rvert$	P_1, Q_1
generation	P_i, $\lvert V_i\rvert$	θ_i, Q_i
non-generation	P_i, Q_i	θ_i, $\lvert V_i\rvert$

Table 1: Elements of the AC power flow equations by bus type. (Glover et al. 2008, Ch. 6)

completely straightforward since the equations are nonlinear). For most buses, the real and reactive power are treated as known. For buses with generation capacity, the real power and voltage magnitude are treated as known. One bus, $i = 1$ without loss of generality, is designated the *slack bus*. For the slack bus, the voltage angle and magnitude are assumed to be known, with $\theta_1 = 0$ to remove degeneracies associated with constraints on the set of voltage magnitudes and angles. Thus, each bus has two known quantities and the solver seeks the remaining two. Table 1 summarizes the knowns and unknowns for the solver (Glover et al. 2008, Ch. 6). Iterative solvers, like Newton-Raphson, are typically used, but convergence is not guaranteed and the procedure can be time consuming.

DC Power Flow

Because of the complexity of the AC solution, power engineers often use a simplified set of equations known as the DC power flow equations (Glover et al. 2008, Ch. 6). This approach uses a number of standard simplifying assumptions. The reactive power is often significantly smaller than the real power, so this equation is eliminated. The values of G are also significantly smaller than B because of line properties, so the G terms are eliminated. Differences in the voltage angles are usually small, so the cosine terms can be set to one and the sine terms can be set to their argument. Finally, the voltage magnitudes are all close to unity when the problem is solved for the per-unit system. The resulting equations are given by

$$P_i = \sum_{j=1}^{N} B_{i,j} \left(\theta_i - \theta_j\right), \tag{1}$$

which is linear in the voltage angles. In matrix form we have $P = b\theta$ where the matrix b is constructed from the reactance parameters to replicate (1). The constraint $\sum_{i=1}^{N} P_i = 0$ still holds, so again we use a slack bus and fix the voltage angle of the slack bus, $\theta_1 = 0$.

The power at the remaining buses is treated as known. We can then eliminate the first row and column of the matrix b and solve for the remaining voltage angles. These are used to compute the power at the slack bus, as well as the power flowing on the lines, $B_{i,j}\left(\theta_i - \theta_j\right)$. This simplified model is commonly used, but the quality of the approximation to the AC model is system and application specific (Stott et al. 2009; Overbye et al. 2004). Despite this, the DC model can provide some insight that we will use in what follows.

3 Estimation for Model Bank Probabilities

This section discusses estimating probabilities for each topology in a bank of models using the available data and the power flow solver.

As in Singh et al. (2010), the model bank consists of a set of m network topologies, $\{s_1, \cdots, s_m\}$, deemed interesting or important by the operator. Let Y be a vector of measurements on d different network quantities. Although the set of possible measurements is quite general, we consider only power flowing on a subset of the lines, so that each element of Y is a measurement of $P_{i,j}$ or $Q_{i,j}$ for some pair of buses i and j. Based on these measurements, we want to estimate the probability that each member of the model bank represents the network's current configuration.

Assume that each measurement, Y_ℓ, is independently normally distributed with some mean μ_ℓ and standard deviation given as a proportion ρ_ℓ of the mean (see Evans 2012 for this error structure and values of ρ):

$$Y_\ell \sim \mathrm{N}\left(\mu_\ell, \rho_\ell^2 \mu_\ell^2\right). \tag{2}$$

The mean itself is obtained by solving the power flow equations described earlier for the unknowns in Table 1 (Glover et al. 2008, Ch. 6). The quantities listed as known in Table 1 are required as inputs to solve the equations, but they are typically imprecisely known and summarized with a distribution. Let Z be the vector of nodal quantities used as inputs, which we will generally refer to as injections (injections are loads and generations and these comprise most of the quantities in Z). We assume that each injection follows a Gaussian prior distribution with some known mean ν_i and standard deviation given as a proportion

β_i of the mean,

$$Z_i \sim \mathrm{N}\left(\nu_i, \beta_i^2 \nu_i^2\right). \tag{3}$$

For a given network topology s_k and a given set of injections Z, the mean for measurement Y_ℓ is a deterministic output of the power flow solver, $\mu_\ell(s_k, Z)$. Finally, assume that each model bank member has some prior probability of occurrence given by $\pi(s_k)$.

Our goal is to estimate $\Pr\{s_k|Y=y\}$, therefore the injections, Z, are nuisance parameters that need to be integrated out. There are circumstances in which we would like to estimate the posterior for Z given the observations, but this is left for future work. Consider the integral for the probability of a topology s_k,

$$\Pr\{s_k|Y=y\} \propto \pi(s_k) \int f(y|\mu(s_k, Z))\pi(Z)dZ = \pi(s_k)f(y|s_k), \tag{4}$$

where f is the likelihood described in (2) and $\pi(Z)$ is the prior for Z given in (3). Since Z enters into the distribution for Y nonlinearly through the solver in $\mu(s_k, Z)$, analytic integration is not possible. However, it is simple to sample Z from its prior, so Monte Carlo methods are a good approach. The straightforward approach estimates $f(y|s_k)$ by averaging $f(y|\mu(s_k, Z))$ over independent draws of Z from its prior. Unfortunately, the distribution for the injections induces a distribution on μ whose mass is very spread out compared to the likelihood $f(y|\mu)$. This results in few draws for which $f(y|\mu)$ is significantly greater than zero. An extremely large number of samples of Z could be required to estimate the integral in (4). Calculating the solution to the power flow equations for potentially hundreds of thousands of draws of Z and hundreds of model bank members quickly becomes infeasible. This is true even for the DC model, especially for large networks. We consider a two part improvement to the brute force approach: a statistical approximation to the power flow solver and importance sampling.

Approximating the Power Flow Solver

The well-developed literature on computer experiments (Morris et al. 1993; Kennedy and O'Hagan 2001; Higdon et al. 2005, 2008) is concerned with approximating a deterministic and computationally expensive computer model with a statistical model. The computer model is run at small number of carefully selected input settings, and an interpolating response surface, often a Gaussian process, is fit to the outputs. The response surface can be used to predict the output of the computer model at untried input settings with much less computational effort. This response surface, called an emulator, can be used to maximize a response or estimate input settings that make the computer model match experimental data.

Our problem is similar: we have a computational model, the power flow solver, that takes the injections and the topology as inputs and produces deterministic line flows as outputs. Our goal is to relate the topological input to the observed line flow data, and the injections are merely a nuisance parameter. Our solution is to build an emulator that integrates over the injections. Effectively, we view the power flow solver as a stochastic computational model where the randomness arises from the injections and the topology is the only input. One run of the solver takes a topology as an input, draws a set of injections from their prior, and produces an output. Since we view our simulator as stochastic, we will build a stochastic emulator as well. The stochastic emulator will produce outputs with the same distribution as the solver.

In the DC power flow model, the voltage angles θ are the solution of the linear system $P = b\theta$. In this case, the P represents the nodal injections, the Z of our statistical model. If these are normally distributed, then the solution for θ is also normally distributed. Similarly, the predicted power on the line between any two buses i and j, $P_{i,j} = b_{i,j}(\theta_i - \theta_j)$ is normally distributed. This last quantity plays the role of μ in our statistical model. Thus, in the DC model, the Gaussian distribution on the Z induces a Gaussian distribution on μ. This suggests that line flows for each scenario are well emulated by using a multivariate Gaussian with the correct mean and covariance for each topology. Because the DC model approximates the AC model, we anticipate that predictions from the AC model, particularly the real power, might be approximately Gaussian. To be more general and to account for unexpected behavior in reactive power, we assume that some transformation $h(\mu)$ follows a normal distribution. Thus, our emulator for the power flow solver is a multivariate Gaussian distribution for a transformation of the line flows:

$$\eta(\mu|s_k) \propto |\Sigma_k|^{-1} \exp\left\{-\frac{1}{2}\left[h\left(\mu\right) - \nu_k\right]^T \Sigma_k^{-1} \left[h\left(\mu\right) - \nu_k\right]\right\} \mathcal{J}_h\left(\mu\right),$$

where $\mathcal{J}_h\left(\mu\right)$ is the Jacobian of the transformation, which just involves derivatives of $h(\cdot)$.

We fit the emulator by estimating the parameters of the multivariate Gaussian for each topology and choosing a form for $h(\cdot)$ if necessary. We do this by taking a limited number of draws of Z and computing the corresponding sample of μ for each topology. We use exploratory data analysis to find a normalizing transformation h for the samples of μ and then estimate ν_k and Σ_k from the transformed samples. This fitted distribution can then be used for estimating the model bank probabilities. One important feature of this emulator is that it can be precomputed and the solver need not be run in any monitoring situation.

Importance Sampling

Assume that we have the fitted emulator distributions $\eta(\mu|s_k)$ for each network topology s_k. Assume also that we have some observation vector $Y = y$. Recall from (2) that y given μ has density

$$f(y|\mu) \propto \prod_{\ell=1}^{d} \frac{1}{|\rho_\ell^2 \mu_\ell^2|} \phi\left(\frac{y_\ell - \mu_\ell}{|\rho_\ell \mu_\ell|}\right),$$

where $\phi(\cdot)$ is the density for a standard normal random variable.

We could now sample the μ directly from η, but this sample would still produce many draws for which $f(y|\mu)$ is very close to zero leading to a Monte Carlo estimator that has poor accuracy for a given number of samples. Thus, we turn to importance sampling (*e.g.* Liu 2001), which draws a biased sample and weights it to obtain the correct average. Sampling values of μ near the data point guarantees non-negligible values of $f(y|\mu)$. Thus, let $g(\mu|Y = y)$ be the importance distribution chosen to be Gaussian with mean vector y and diagonal covariance with diagonal entries given by $(\alpha\rho_\ell y_\ell)^2$:

$$g(\mu|y) \propto \prod_{\ell=1}^{d} \frac{1}{|\alpha\rho_\ell y_\ell|} \phi\left(\frac{\mu_\ell - y_\ell}{|\alpha\rho_\ell y_\ell|}\right),$$

where $\alpha > 1$ provides additional variation around the observation. Because the emulator samples directly from the induced distribution on μ, it is much easier to formulate an importance sampler than would be the case had we retained the dependence on the injections. We use M samples from $g(\mu|y)$, $\mu_1, \ldots, \mu_M$, to estimate $f(y|s_k)$ with

$$\tilde{f}(y|s_k) = \frac{1}{M} \sum_{i=1}^{M} f(y|\mu_i) \frac{\eta(\mu_i|s_k)}{g(\mu_i|y)}. \tag{5}$$

Finally, the estimated probability of scenario s_k is $\tilde{\Pr}\{s_k|Y = y\} \propto \pi(s_k)\tilde{f}(y|s_k)$.

To assess the efficiency of the importance sampling algorithm, consider a toy problem that captures the features of the power problem. As in the larger problem, assume that $y \sim N(\mu, (.01\mu)^2)$ and $\mu \sim N(0, 1)$. Consider an observation of $y = 3$ and that we want to estimate the density of y at the value 3 independent of μ. Two approaches are the importance sampling algorithm presented here and straightforward Monte Carlo. We conducted a numerical test that computed the estimate under both algorithms 1000 times and examined the variance of the resulting estimates. The importance sample estimates are based on 1000 samples from the importance distribution. In order to produce estimates with similar variance, the straightforward Monte Carlo method needs about 5000000 draws.

In summary, we propose the following scheme to estimate the probability of each member of the model bank given an observation y.

Offline Precomputation

1. Draw a number of injection realizations from $\pi(Z)$. This number can be as large as desired since this need not be done in real time.

2. For each topology, s_k, use the solver to compute a sample for $\mu|s_k$ from the sample of Z.

3. For each sample of $\mu|s_k$, estimate a distribution $\eta(\mu|s_k)$ composed of a transformation, $h(\cdot)$ for μ and a Gaussian.

Online Real Time Monitoring

1. Draw a number of realizations of μ_{imp} from $g(\mu|y)$ which is $\text{N}\{y, \text{diag}[(\alpha\rho_\ell y_\ell)^2]\}$.

2. For each scenario, compute $\tilde{f}(y|s_k)$ using (5) with the sample μ_{imp}.

3. Multiply the estimates $\tilde{f}(y|s_k)$ by the prior and normalize to obtain model bank probabilities.

4 UK Network

Return now to the network in Figure 1. In Singh et al. (2010), the authors use this network to demonstrate their model bank methodology. We do the same, with a focus on network scenarios associated with network 1. As in Singh et al. (2010) the network is measured at a single point: the line between buses 1 and 2 in network 1. The relevant model bank from Singh et al. (2010) contains five network configurations:

1. **Normal**: normal operation.
2. **B Opens**: protective device B opens, separating the lower half of network 1.
3. **E Closes**: the two networks are connected by device E and network 2 consumes power provided by network 1.
4. **Lose Load 17**: the generator at bus 17 is lost.
5. **Lose Load 18**: the large load at bus 18 is dropped.

To do the calculations, generate a large number of draws for each of the injections (indicated by downward pointing triangles in Figure 1), as prescribed in the Offline Precomputation section of the algorithm summary at the end of Section 3. For each set of injections, solve the power flow equations to obtain the prediction for the flow on the line between buses 1 and 2 under each of the five scenarios. For each of these five samples of μ corresponding to the five topologies in the model bank, estimate the parameters of a Gaussian distribution.

In the DC model, the observations and the fitted Gaussian are univariate. The normal quantile-quantile plots are shown in Figure 1 of the supplementary material. In the DC model, the sample for μ is theoretically Gaussian, so the fit is expectedly good.

The AC model uses a bivariate observation of real and reactive power and we fit a bivariate Gaussian to each of the five samples corresponding to the five topologies. Figure 2 of the supplementary material shows marginal normal quantile-quantile plots for the samples of μ under the AC model. In this case, the Gaussian fit is an approximation, but appears to be very good and no transformation is required.

	Truth					
Estimate		Normal	B Opens	E Closes	Lose 17	Lose 18
	Normal	**0.61**	0.02	0.31	0.03	0.01
	B Opens	0.02	**0.98**	–	–	–
	E Closes	0.33	–	**0.45**	0.22	–
	Lose Load 17	0.03	–	0.23	**0.75**	–
	Lose Load 18	0.01	–	–	–	**0.99**

Table 2: Confusion table for the UK network using the DC solver. For each topology in the column headers, 1000 simulated observations are generated. For each observation, the emulator-based importance sampling method is used to estimate probabilities for each member of the model bank. The probabilities in each column are averaged over the estimates for all the samples and sum to one.

First, consider simulating data from each of the five members of the model bank and computing the probabilities for each sample. For each topology, we generate 1000 simulated observations. For each observation, we apply the emulator-based importance sample method to estimate probabilities for each member of the model bank. Table 2 shows the results using the DC model for both simulation and prediction. The column labels give the topology used to simulate data. The entries in each column are averaged over the 1000 simulated observations and each column sums to one. For example, with an observation simulated from the normal topology, the method assigns an average probability of 0.61 to the normal topology and an average probability of 0.33 to the E Closes scenario. Comfortingly, the diagonal elements are the largest value in each column. Nevertheless, there is still some ambiguity, as seen in the third column for device E closing, linking the two networks. Concluding that a single topology is absolutely correct could lead to grid operator mistakes. Table 3 gives the results obtained when using the AC model to simulate and predict. In this case, the diagonal elements represent the bulk of the probability for each scenario, rising dramatically in some cases, as compared with Table 2. This is likely the result of the added information contained in the measurement of reactive power.

	Truth					
		Normal	B Opens	E Closes	Lose 17	Lose 18
Estimate	Normal	**0.70**	–	0.29	0.01	–
	B Opens	–	**1.00**	–	–	–
	E Closes	0.29	–	**0.70**	0.01	–
	Lose Load 17	0.01	–	0.01	**0.98**	–
	Lose Load 18	–	–	–	–	**1.00**

Table 3: Confusion table for the UK network using the AC solver. See the caption of Table 2 for details.

Figure 3 shows another summary for the DC model results. This plot shows the results of the estimation scheme for a range of possible observations. For this simple network, this plot could directly be used for diagnostic purposes. Suppose the monitoring device returns the observation of real power y_P. The user finds this value on the x-axis of the plot. The vertical value of each of the five lines at this point on the x-axis gives the probability for the corresponding topology. The sum of the values of the five lines at each point on the x-axis is always one. For example, assume we see a measurement of $y_p \approx 1$. The figure indicates that the topologies "E Closes" and "Lose Load 17" have probabilities just under 0.5 and the "Normal" topology has probability of about 0.05. The other two topologies have zero probability. Although the possible observations can take values anywhere on the real line, the range of the figure's x-axis has been chosen to contain the vast majority of possible observations under the prior distributions.

Figure 4 is similar to Figure 3, but corresponds to the AC model. An AC observation is bivariate with both a real and reactive component. Each of the five panels corresponds to a member of the model bank. Again, this plot could be used directly for diagnostic purposes. Suppose the monitoring device returns an observation of real and reactive power (y_P, y_Q). To obtain the probability of the "Normal" topology, the user finds the value of y_P on the x-axis and y_Q on the y-axis of the panel labeled "Normal". The color at this point indicates the probability of this topology. Similarly for the other four topologies. The sum of a particular point across the five panels is always one. Again, the ranges of the axes are set to contain the vast majority of possible observations under the prior distributions.

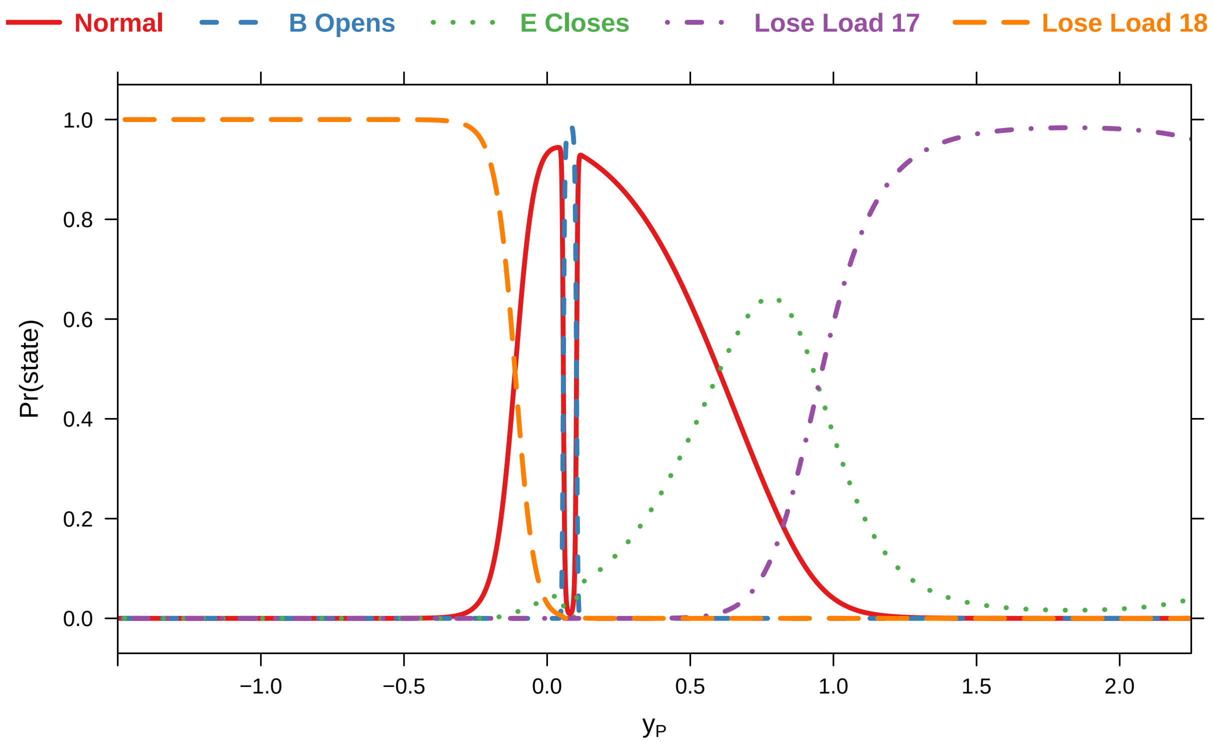

Figure 3: Estimated state probabilities for a range of possible observations, y_P. Predictions are made using the algorithm with the DC model for the UK network. At any observed power, y_P, the height of the five curves must sum to one. The online version of this figure is in color.

Figure 5 shows the maximum probability for any topology over the observation space (the maximum across the five panels in Figure 4 at each point). Previous state estimation methods typically assume the ability to compute the correct answer with certainty. Most of the region does favor some topology with probability near unity, but a significant portion of the high density region of the observation space gives ambiguous results (with maximum probabilities as low as 0.20) demonstrating the importance of considering the posterior probability vector for the entire model bank.

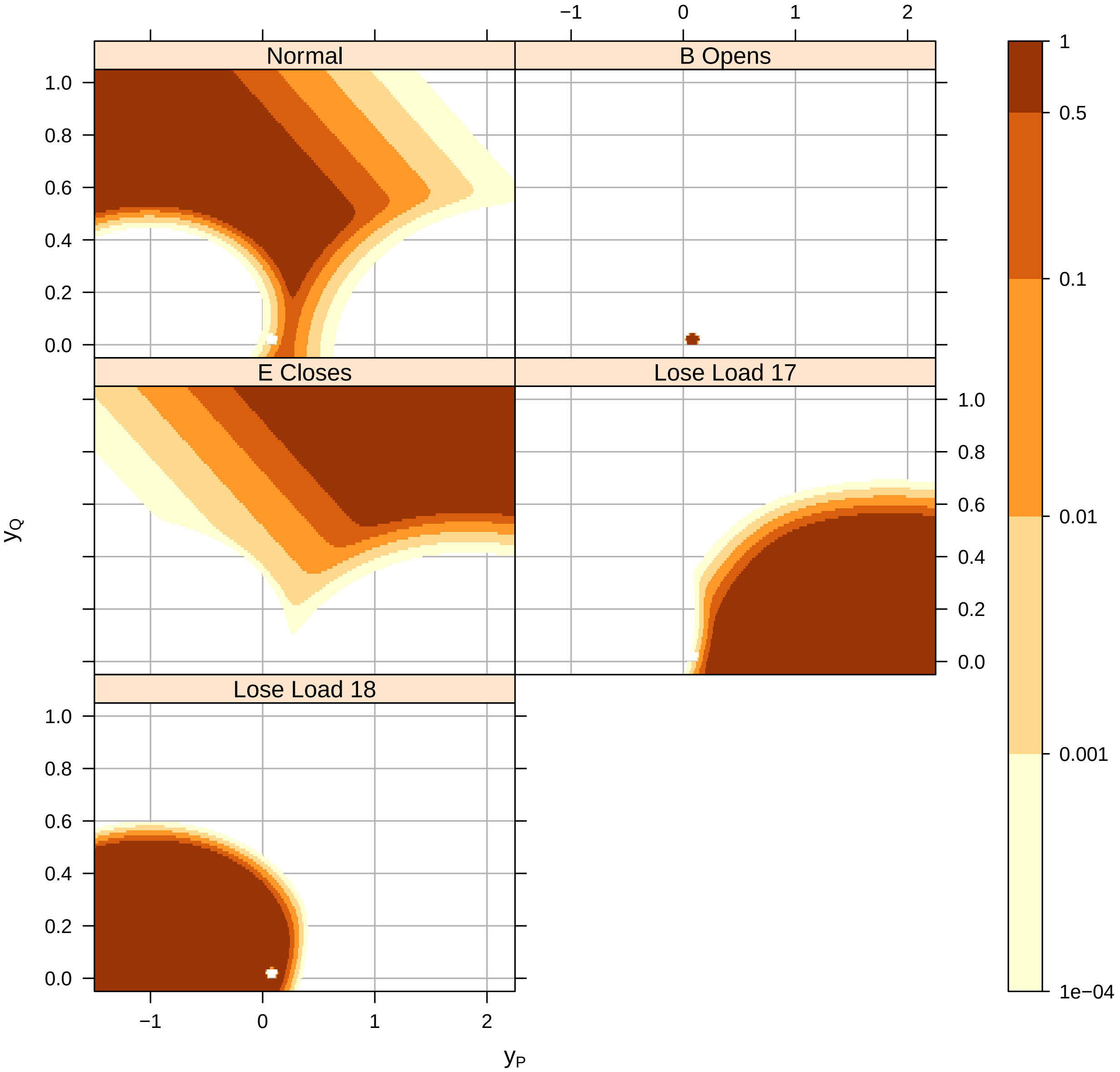

Figure 4: Results of the estimation scheme for a range of possible observations corresponding to using the AC model for the UK network. At each (y_P, y_Q) location the shading indicates the probability of the scenario indicated in the panel title. The sum of a particular point across the five panels is always one. The online version of this figure is in color.

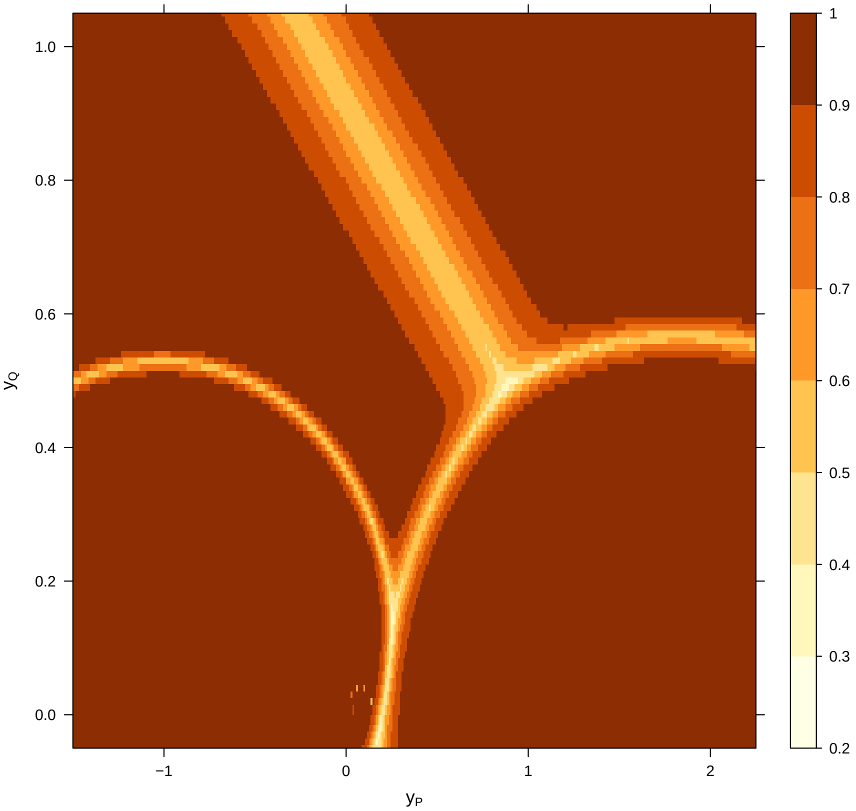

Figure 5: Maximum topology probabilities as a function of the observation of the real y_P and reactive y_Q power at the line between buses 1 and 2. The online version of this figure is in color.

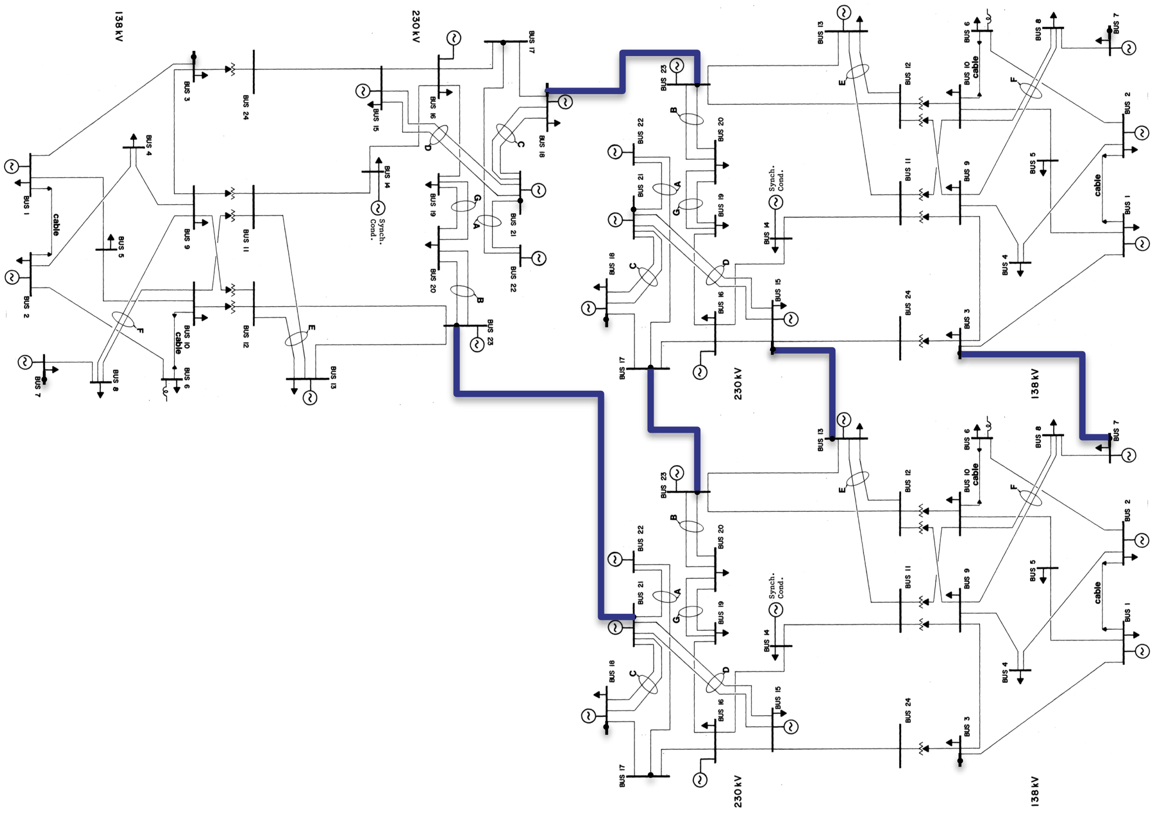

Figure 6: 1996 IEEE Reliability Test System. This network consists of three identical subnets that are connected by five lines indicated in bold. These five lines will serve as our monitoring points. Figure is adapted from Reliability Test System Task Force (1979).

5 Reliability Test System 1996

The 1996 edition of the IEEE Reliability Test System (RTS-96) (Reliability Test System Task Force 1999) is a 72 bus network designed as benchmark for comparing reliability methodologies. As shown in Figure 6, it consists of three identical 24 bus networks connected together. We consider monitoring the five lines that connect that the three identical 24 bus networks. This corresponds to measuring the inter-area exchanges of power between connected networks. Our model bank contains 124 topologies consisting of normal operating conditions, all of the single line losses (except the monitored lines), and a subset of two-line losses (those that share common transmission corridors). This is the set of contingencies defined in Reliability Test System Task Force (1999).

As before and described in the Precomputation part of the algorithm, we draw a large number of samples, 10000, for each injection and solve the power flow equations for each sample and each member of the model bank. In this case, the distribution of real power is well-represented by a Gaussian, but the fit to the reactive power is not as good. We transform the reactive power to improve the fit using a three parameter Gamma distribution function and the inverse standard normal

$$h_s(y_\ell) = \Phi^{-1}\left[F\left(y_\ell - \gamma_{s,\ell}; \alpha_{s,\ell}, \beta_{s,\ell}\right)\right],$$

where $F(\cdot)$ is the Gamma distribution function. This was chosen after some exploration of the data. This transformation is specific to topology and measurement. Again, this transformation is only applied to measurements ℓ that correspond to reactive power measurements; real power measurements appear to be Gaussian without transformation. The transformation parameters are estimated using the Exact Likelihood method discussed in Meeker and Escobar (1994). This method assumes that data are only ever measured and stored up to some precision and so are all actually discrete. The method is helpful in cases where using the density could allow the likelihood to go to the infinity, which can be the case in the three parameter Gamma with its lower threshold parameter. We discretized the data into 1,000,000 evenly spaced bins over the range of each measurement. Multivariate Gaussians are fit to the transformed sample for each model bank member giving 124 multivariate Gaussian fits for each of the DC and AC models. Gaussian quantile-quantile plots for the AC results are shown in Figures 3 and 4 of the supplementary material. The Gaussian appears to be a good fit.

For each of the 124 scenarios, we generate 1000 simulated observations from each topology and compute the classification probabilities for each topology based on each sample. This is done for both the DC and AC models. Note that this is 124000 runs of the importance sampling scheme using precomputed Gaussian fits. Because of the precomputed emulator and the efficient importance sampling, the computation of the probabilities for a single observation under the AC is extremely fast, taking less than 1s. This demonstrates the ability of the method to quickly estimate the probabilities for a larger network in an online setting.

Figure 7 summarizes the results of these calculations for the AC model. There is one point for each of the 124 topologies. The points show the posterior probability of the data-generating topology over the 1000 simulated observations plotted against the largest average probability for an incorrect model. The circles represent results in which the true topology has the highest average probability. For cases in which the true topology does not have the highest average probability, an "x" is plotted. In 59 of the 124 scenarios, the true topology has the highest average probability. In 110 of the scenarios, the true topology is among the top five highest probabilities. All of the x points fall near the 45° line. This suggests that in the scenarios for which the true topology is not given the highest average probability, there is some ambiguity in which the true topology is in a group of topologies all of which have similar probabilities. More monitoring points, or more carefully selected monitoring points, would likely increase these probabilities and will be the subject of future work.

Compared to the DC results, the average probability for the true topology is larger under the AC model in 98 of the 124 cases and the probability of the truth is on average twice as big. Figure 5 in the supplementary materials compares the average probability for the true topology under the DC and AC power flow models. Figure 6 in the supplementary material shows the posterior probability of the data-generating topology over the 1000 simulated observations plotted against the largest average probability for an incorrect model for the DC power flow model (similar to Figure 7 shown here for the AC power flow model). Using the DC model, the true scenario is identified with the highest average probability only 34 times out of 124, with the remaining cases giving results in which the truth has a similar probability to a number of other models. In this case, the AC model is clearly superior for identifying the topology, so it is important that the algorithm is able to produce real-time estimates of the probability using the more realistic model.

6 Discussion

This paper presents a formulation for the power grid model bank problem and a scheme for computing the model bank probabilities. We build a multivariate distribution to emulate the output of the power flow solver and use it in an importance sampling scheme for computing model bank probabilities. The resulting algorithm is extremely fast and accurate. In many ways, this problem is just the tip of the iceberg for statistical modeling in power

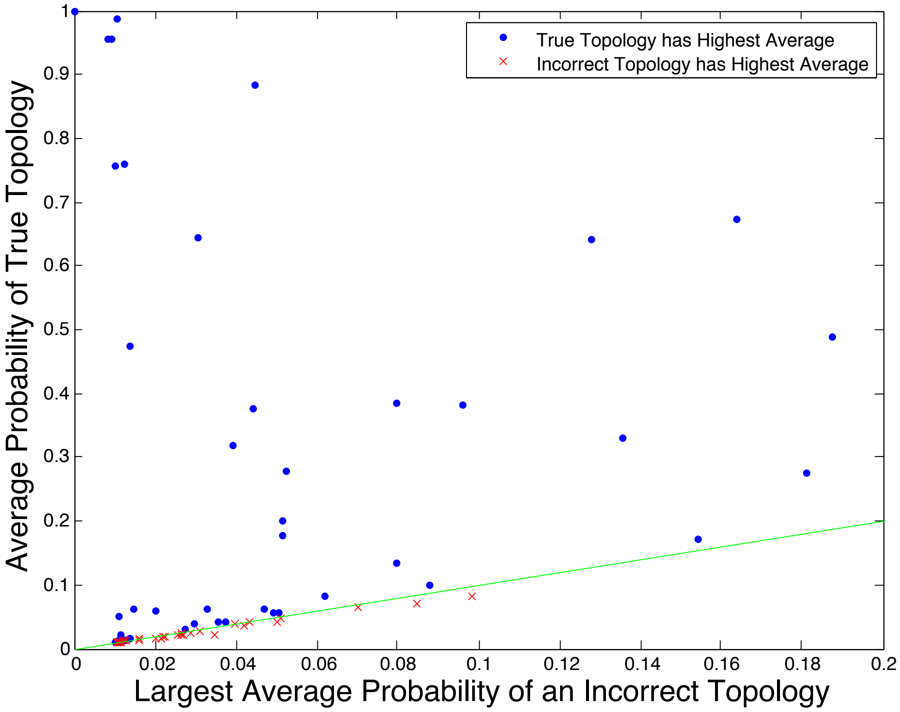

Figure 7: Classification probability for the data generating topology averaged over 1000 simulated observations for each of 124 scenarios plotted against the highest average probability of an incorrect model. These results are based on the AC power flow equations. The online version of this figure is in color.

engineering applications. There are several important extensions to using a model bank and moving beyond it.

The model bank probabilities aid the network operator in making decisions about network function, but we haven't yet considered the posterior distribution of the injections, Z, given the data and a particular member of the model bank. The conditional distributions for the injections, and thus the power on each line, could differ greatly depending on which model bank member is assumed. If the operator is making decisions using the wrong configuration, he could take actions that lead to line overloads and grid failures. Thus, one important future research direction is to accurately infer the posterior distributions of the injections for all of the relevant network configurations. It is possible that a number of configurations have high probability, but very different distributions for injections and line flows. Currently, given knowledge of the network topology and the load profile, an optimization procedure is used to decide how much power each generator should produce. In cases where there is ambiguity about the network topology and large differences in the load profile between possible configurations, new techniques are needed to solve the generator dispatch problem accounting for uncertainty. A third direction is testing whether data arise from a scenario that is outside the model bank.

Another important extension involves choosing monitoring points in the network. In the Smart Grid era, monitoring devices are becoming more readily available. Their utility is maximized by considering how additional pieces of information aid in the ability to distinguish one configuration from another. The measurement points chosen in our example were based on reasonable choices, but obviously leave room for improvement. Perhaps more importantly, as sensors become ubiquitous, the amount of data outpaces the ability to process it. Another outcome of this extension will be the determination of which data are most helpful for state estimation.

In this paper, we discuss a very simple emulation approach: the distribution of the line predictions for each network configuration is approximated as a Gaussian, possibly after transformation. More complex choices could be made as well. First, a simple extension is to use mixtures of Gaussians, which might be useful in the event that the approximation for the AC solver breaks down or when the priors for the injections have more complex forms. Both of these methods effectively integrate out the injections by sampling from their priors, but the injections could be incorporated in a more complex emulator, perhaps based on Gaussian processes, using the injections as inputs to predict flows more directly and precisely.

Finally, the model bank is necessarily limited, and it seems likely that data arising from outside the bank indicate a serious problem requiring quick diagnosis. Thus, a major goal is to move beyond the model bank and use the data to estimate the state of the network and its graph directly. This is a challenging problem due to the combinatorial nature of the space of graphs.

Supplementary Material

Normal quantile-quantile plots and additional RTS 96 results referenced in the main paper (pdf file).

Acknowledgements

This work was supported by the LDRD program at Los Alamos National Laboratory. Thank you to Kary Myers for organizing the Conference on Data Analysis and for helpful comments on the paper. Thank you also to the reviewers, the associate editor, and especially the editor.

References

Alves da Silva, A. P., Quintana, V. H., and Pang, G. K. H. (1991). Solving data acquisition and processing problems in power systems using a pattern analysis approach. *Proceedings of the Institute of Electrical Engineering C*, 138(4):365–376.

Andersson, G., Donalek, P., Farmer, R., Hatziargyriou, N., Kamwa, I., Kundur, P., Martins, N., Paserba, J., Pourbeik, P., Sanchez-Gasca, J., Schulz, R., Stankovic, A., Taylor, C., and Vittal, V. (2005). Causes of the 2003 major grid blackouts in north america and europe, and recommended means to improve system dynamic performance. *IEEE Transactions on Power Systems*, 20(4):1922–1928.

Clements, K. A. and Davis, P. W. (1988). Detection and identification of topology errors in electric power systems. *IEEE Transactions on Power Systems*, 3(4):1748–1753.

Department of Energy (2012). 2010 smart grid system report.

Evans, J. W. (2012). Interface between automation and the substation. In McDonald, J. D., editor, *Electric Power Substations Engineering*, chapter 6. CRC Press.

Glover, J. D., Sarma, M. S., and Overbye, Thomas, J. (2008). *Power System Analysis and Design*. Cengage Learning.

Higdon, D., Gattiker, J. R., Williams, B., and Rightley, M. (2008). Computer model calibration using high dimensional output. *Journal of the American Statistical Association*, 103:570–583.

Higdon, D., Kennedy, M. C., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2005). Combining field data and computer simulations for calibration and prediction. *SIAM Journal of Scientific Computing*, 26(2):448–466.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, 63(3):425–464.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

Meeker, W. Q. and Escobar, L. A. (1994). Maximum likelihood methods for fitting parametric statistical models. In Stanford, J. L. and Vardeman, S. B., editors, *Statistical Methods for Physical Science*, chapter 8, pages 211–244. Academic Press.

Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35(3):243–255.

Overbye, Thomas, J., Cheng, X., and Sun, Y. (2004). A comparison of the ac and dc power flow models for lmp calculations. *Proceedings of the 37th Hawaii International Conference on System Sciences.*

Reliability Test System Task Force (1979). Ieee reliability test system. *IEEE Transactions on Power Apparatus and Systems.*

Reliability Test System Task Force (1999). The IEEE reliability test system - 1996. *IEEE Transactions on Power Systems*, 14(3):1010–1020.

Singh, D., Pandey, J. P., and Chauhan, D. S. (2005). Topology identification, bad data processing and state estimation using fuzzy pattern matching. *IEEE Transactions on Power Systems*, 20(3):1570–1579.

Singh, H. and Alvarado, F. L. (1995). Network topology determination using least absolute value state estimation. *IEEE Transactions on Power Systems*, 10(3):1159–1165.

Singh, N. and Glavitsch, H. (1991). Detection and identification of topological errors in online power system analysis. *IEEE Transactions on Power Systems*, 6(1):324–330.

Singh, R., Manitsas, E., Pal, B. C., and Strbac, G. (2010). A recursive bayesian approach for identification of network configuration changes in distribution system state estimation. *IEEE Transactions on Power Systems*, 25(3):1329–1336.

Stott, B., Jardim, J., and Alsac, O. (2009). Dc power flow revisited. *IEEE Transactions on Power Systems*, 24(3):1290–1300.

U.S. - Canada Power System Outage Task Force (2004). Final report on the August 14, 2003 blackout in the United States and Canada. Technical report, United States Department of Energy.

Vinod Kumar, D. M., Srivastava, S. C., Shah, S., and Mathur, S. (1996). Topology processing and static state estimation using artificial neural networks. *IEE Proceedings Generation, Transmission & Distribution*, 143(1):99–105.

Wu, F. F. and Liu, W.-H. E. (1989). Detection of topology errors by state estimation. *IEEE Transactions on Power Systems*, 4(1):176–183.