SIP (2018), vol. 7, e19, page 1 of 17 © The Authors, 2018.

doi:10.1017/ATSIP.2018.20

ORIGINAL PAPER

Covariance selection quality through detection problem and AUC bounds

NAVID TAFAGHODI KHAJAVI AND ANTHONY KUH

Graphical models are increasingly being used in many complex engineering problems to model the dynamics between states of the graph. These graphs are often very large and approximation models are needed to reduce the computational complexity. This paper considers the problem of quantifying the quality of an approximation model for a graphical model (model selection problem). The model selection often uses a distance measure such as the Kullback–Leibler (KL) divergence between the original distribution and the model distribution to quantify the quality of the model approximation. We extend and broaden the body of research by formulating the model approximation as a detection problem between the original distribution and the model distribution. We focus on Gaussian random vectors and introduce the Correlation Approximation Matrix (CAM) and use the Area Under the Curve (AUC) for the formulated detection problem. The closeness measures such as the KL divergence, the log-likelihood ratio, and the AUC are functions of the eigenvalues of the CAM. Easily computable upper and lower bounds are found for the AUC. The paper concludes by computing these measures for real and synthetic simulation data. Tree approximations and more complex graphical models are considered for approximation models.

Keywords: Statistical model selection, Detection problem, Tree approximation, Area under the curve, Covariance selection

Received 17 July 2018; Revised 25 October 2018

I. INTRODUCTION

Graphical models are useful tools for describing the geometric structure of networks in numerous applications such as energy, social, sensor, biological, and transportation networks [1] that deal with high-dimensional data. Learning from these high-dimensional data requires large computation power which is not always available [2, 3]. The hardware limitation for different applications forces us to compromise between the accuracy of the learning algorithm and its time complexity by using the best possible approximation algorithm given the constrained graph. In other words, the main concern is to compromise between model complexity and its accuracy by choosing a simpler, yet informative model. To address this concern, many approximation algorithms are proposed for model selection and imposing structure given data. For the Gaussian distribution, the covariance selection problem is presented and studied in [4, 5]. This paper goes beyond the seminal work of [6] on model approximation by formulating the model approximation problem as a detection problem. The detection problem allows us to look at the model approximation problem in a broader and more accurate way, by being able to study new measures to assess the approximation quality and to look at the distribution of the sufficient statistic under each hypothesis. Here, we introduce the Correlation Approximation Matrix (CAM) and use the CAM to assess the quality of the approximation by relating the CAM to information divergences (e.g. Kullback–Leibler (KL) divergence) and the Area Under the Curve (AUC). The CAM, AUC, and reverse KL divergence give new qualitative insights into the quality of the model approximation.

The ultimate purpose of the covariance selection problem is to reduce the computational complexity in various applications. One of the special approximation models is the tree approximation model. Tree approximation algorithms are among the algorithms that reduce the number of computations to get quicker approximate solutions to a variety of problems. If a tree model is used, then distributed estimation algorithms such as message passing algorithm [7] and the belief propagation algorithm [8] can easily be applied and are guaranteed to converge to the maximum likelihood solution.

The Chow-Liu algorithm discussed in [9] gives a method for constructing a tree that minimizes the KL divergence between the model and model tree approximation. The Chow-Liu Minimum Spanning Tree (MST) algorithm for Gaussian distributions is to find the optimal tree structure using a KL divergence cost function [4]. The Chow-Liu MST algorithm constructs a weighted graph by computing pairwise mutual information and then utilizes one of the

Department of Electrical Engineering, University of Hawaii, Honolulu, HI 96822, USA

Corresponding author:
Navid Tafaghodi Khajavi
Email: navidt@hawaii.edu

MST algorithms such as the Kruskal algorithm [10] or the Prim algorithm [11]. *How good is the Chow-Liu solution that minimizes the KL divergence? Can we formulate other measures to assess the model approximation?* These questions are becoming more important to answer as we study engineering applications that are modeled by larger and larger graphical models thus requiring simple model approximations. Before addressing these questions (which is the topic of this paper), we discuss other work and an application.

Other research in approximating the correlation matrix and the inverse correlation matrix with a more sparse graph representation while retaining good accuracy include the first order Markov chain approximation [10], penalized likelihood methods such as LASSO [5, 12], and graphical LASSO [13]. The first order Markov chain approximation method uses a regret cost function to output first-order Markov chain structured graph [14] by utilizing a greedy type algorithm. Penalized likelihood methods use an L1-norm penalty term in order to sparsify the graph representation and eliminate some edges. Recently, a tree approximation in a linear, underdetermined model is proposed in [15] where the solution is based on expectation, maximization algorithm combined with the Chow Liu algorithm.

Sparse modeling has many applications in distributed signal processing and machine learning over graphs. One important application is monitoring the electric power grid at the distribution level. The *smart grid* is a promising solution that delivers reliable energy to consumers through the power grid when there are uncertainties such as distributed renewable energy generation sources. Smart grid technologies such as smart meters and communication links are added to the distribution grid in order to obtain the high-dimensional, real-time data and information and overcome uncertainties and unforeseen faults. The future grid will incorporate distributed renewable energy generation such as solar photovoltaics, with these energy sources being intermittent and highly correlated. Here we can model both energy sources and energy users by nodes on a graph with edges representing electric feeder lines. The graphs for distribution networks can be very large with renewable energy sources adding complexity to the graphs. This necessitates the need for model selection.

This paper discusses the quality of the model selection, focusing on the Gaussian case, i.e. covariance selection problem. We ask the following important question: *"given an approximation model, is the model approximation of the covariance matrix for the Gaussian model a good approximation?"* To answer this question, we need to pick a closeness criterion which has to be coherent and general enough to handle a wide variety of problems and also has asymptotic justification [16]. In many applications, the – KL divergence has been proposed as a closeness criterion between the original distribution and its model approximation distribution [4, 9]. Besides that, other closeness measures and divergences are used for the model selection. One example is the use of the reverse KL divergence as the closeness criterion in variational methods to learn the desired approximation structure [17].

In this paper, we bring a different perspective to quantify the quality of the model approximation problem by formulating a general detection problem. This formulation gives statistical insight on how to quantify a selected model. Also, the detection problem formulation leads to the calculation of the log-likelihood ratio test (LLRT) statistic, the KL divergence and the reverse KL divergence as well as the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) where the AUC is used as the accuracy measure for the detection problem. The detection problem formulation is a different approach which gives us a broader view by determining whether a particular model is a good approximation or not. The AUC does not depend on a specific operating point on the ROC and broadly summarizes the entire detection framework. It also effectively combines the detection probability and the false-alarm probability into one measure. The AUC determines the inherent ability of the test to distinguish (in conventional detection problem) or not to distinguish (in model approximation problem) between two hypotheses/models. More specifically, the detection formulation and particularly the AUC gives us additional insight about any approximation since it is a way to formalize the model approximation problem. This fact is the contribution of this paper and leads to qualitative insights by computing AUC and its bounds. For Gaussian data, the LLRT statistic simplifies to an indefinite quadratic form. We define a key quantity which is the CAM. The CAM is the product of the original correlation matrix and the inverse of the model approximation correlation matrix. For Gaussian data, this matrix contains all the information needed to compute the information divergences, the ROC curve and the area under the ROC curve, i.e. the AUC. We also show the relationship between the CAM, the AUC and the Jeffreys divergence [18], the KL divergence and the reverse KL divergence. We present an analytical expression to compute the AUC for a given CAM that can be efficiently evaluated numerically. We then show the relation between the AUC, the KL divergence, the LLRT statistics, and the ROC curve. We also present analytical upper and lower bounds for the AUC which only depend on eigenvalues of the CAM. Throughout the discussion section, we pick the tree approximation model as a well-known subset of all graphical models. The tree approximation is considered since they are widely used in literature and it is much simpler performing inference and estimation on trees rather than graphs that have cycles or loops. We perform simulations over synthetic and real data for several examples to explore and discuss our results. Simulation results indicate that $1-\text{AUC}$ is decreasing exponentially as the number of nodes in the graph increases which is consistent with the analytical results obtained from the AUC upper and lower bounds.

The rest of this paper is organized as follows. In Section II, we give a general framework for the detection problem and the corresponding sufficient test statistic, the log-likelihood ratio test. The LLRT for Gaussian data as well as its distribution under both hypotheses are also presented in this section. The ROC curve and the AUC definition, as

well as an analytical expression for the AUC, are given in Section III. Section IV provides analytical lower and upper bounds for the AUC. The lower bound for the AUC uses the Chernoff bound and is a function of the CAM eigenvalues. The upper bound is obtained by finding a parametric relationship between the AUC and the KL and reverse KL divergences. Then, Section V presents the tree approximation model and provides some simulations over synthetic examples as well as real solar data examples and investigates the quality of the tree approximation based on the numerically evaluated AUC and also its analytical upper and lower bounds. Finally, Section VI summarizes results of this paper and discusses future directions for research.

II. DETECTION PROBLEM FRAMEWORK

In this section, we present a framework to quantify the quality of a model selection. More specifically, we formulate a detection problem to distinguish between the covariance matrix of a multivariate normal distribution and an approximation of the aforementioned covariance matrix based on the given model. A key quantity, the Correlation Approximation Matrix (CAM) is introduced in this section and for Gaussian data, we can calculate the KL divergence and log-likelihood ratios, that all depend on the eigenvalues of the CAM.

A) Model selection problem

We want to approximate a multivariate distribution by the product of lower order component distributions [19]. Let random vector $\underline{X} \in \mathbb{R}^n$, have a distribution with parameter Θ, i.e. $\underline{X} \sim f_{\underline{X}}(\underline{x})$. We want to approximate the random vector $\underline{X}$, with another random vector associated with the desired model.[1] Let the model random vector $\underline{X}_{\mathcal{M}} \in \mathbb{R}^n$ have a distribution with parameter $\Theta_{\mathcal{M}}$, associated with the desired model, i.e. $\underline{X} \sim f_{\underline{X}_{\mathcal{M}}}(\underline{x})$. Also, let $\mathcal{G} = (\mathcal{V}, \mathcal{E}_{\mathcal{M}})$ be the graph representation of the model random vector $\underline{X}_{\mathcal{M}}$ where sets $\mathcal{V}$ and $\mathcal{E}_{\mathcal{M}}$ are the set of all vertices and the set of all edges of the graph representing of $\underline{X}_{\mathcal{M}}$, respectively. Moreover, $\mathcal{E}_{\mathcal{M}} \subseteq \psi$ where ψ is the set of all edges of a complete graph with vertex set $\mathcal{V}$.

Remark. *Covariance selection is presented in [4]. Moreover, tree model as a special case for the model selection is discussed in subsection A.*

B) General detection framework

The model selection is extensively studied in the literature [4]. Minimizing the KL divergence between two distributions or the maximum likelihood criterion are proposed in many state of the art works to quantify the quality of the model approximation. A different way to look at the problem of quantifying the quality of the model approximation is to formulate a detection problem [20]. Given the set of data, the goal of the detection problem is to distinguish between *the null hypothesis* and *the alternative hypothesis*. To set up a detection problem for the model selection, we need to define these two hypotheses as follows

- The null hypothesis, $\mathcal{H}_0$: data are generated using the known/original distribution,
- The alternative hypothesis, $\mathcal{H}_1$: data are generated using the model/approximated distribution.

Given the set up for the null hypothesis and the alternative hypothesis, we need to define a test statistic to quantify the detection problem. The likelihood ratio test (the Neyman–Pearson (NP) Lemma [21]) is the most powerful test statistic where we first define the LLRT as

$$l(\underline{x}) = \log \frac{f_{\underline{X}}(\underline{x}|\mathcal{H}_1)}{f_{\underline{X}}(\underline{x}|\mathcal{H}_0)} = \log \frac{f_{\underline{X}_{\mathcal{M}}}(\underline{x})}{f_{\underline{X}}(\underline{x})}$$

where $f_{\underline{X}}(\underline{x}|\mathcal{H}_0)$ is the random vector $\underline{X}$ distribution under the null hypothesis while $f_{\underline{X}}(\underline{x}|\mathcal{H}_1)$ is the random vector $\underline{X}$ distribution under the alternative hypothesis.

Let $l(\underline{X})$ be the LLRT statistic random variable. Then, we define *the false-alarm probability* and *the detection probability* by comparing the LLRT statistic under each hypothesis with a given threshold, τ, and computing the following probabilities

- The false-alarm probability, $P_0(\tau)$, under the null hypothesis, $\mathcal{H}_0$: $P_0(\tau) = \Pr(l(\underline{X}) \geq \tau|\mathcal{H}_0)$,
- The detection probability, $P_1(\tau)$, under the alternative hypothesis, $\mathcal{H}_1$: $P_1(\tau) = \Pr(l(\underline{X}) \geq \tau|\mathcal{H}_1)$.

The NP Lemma [21] is the most powerful test at a given false-alarm rate (significant level). The most powerful test is defined by setting the false-alarm rate $P_0(\tau) = \bar{P}_0$ and then computing the threshold value $\tau = \tau_0$ such that $\Pr(l(\underline{X}) \geq \tau_0|\mathcal{H}_0) = \bar{P}_0$.

Definition 1. *The KL divergence between two multivariate continuous distributions with probability density functions (PDF) $p_{\underline{X}}(\underline{x})$ and $q_{\underline{X}}(\underline{x})$ is defined as*

$$\mathcal{D}(p_{\underline{X}}(\underline{x})||q_{\underline{X}}(\underline{x})) = \int_{\mathcal{X}} p_{\underline{X}}(\underline{x}) \log \frac{p_{\underline{X}}(\underline{x})}{q_{\underline{X}}(\underline{x})}\, d\underline{x}$$

where $\mathcal{X}$ is the feasible set.

Throughout this paper, we may use other notation such as the KL divergence between two covariance matrices for zero-mean Gaussian distribution case or the KL divergence between two random variables in order to present the KL divergence between two distributions.

Proposition 1. *Expectation of the LLRT statistic under each hypothesis is*

- $\mathrm{E}\left(l(\underline{X})|\mathcal{H}_0\right) = -\mathcal{D}(f_{\underline{X}}(\underline{x})||f_{\underline{X}_{\mathcal{M}}}(\underline{x}))$,
- $\mathrm{E}\left(l(\underline{X})|\mathcal{H}_1\right) = \mathcal{D}(f_{\underline{X}_{\mathcal{M}}}(\underline{x})||f_{\underline{X}}(\underline{x}))$.

Proof: Proof is based on the KL divergence definition. □

[1] Examples of possible models: tree structure, sparse structure and Markov chain.

Remark. *The relationship between the NP lemma and the KL divergence is previously stated in [22] with the similar straightforward calculation, where the LLRT statistic loses power when the wrong distribution is used instead of the true distribution for one of these hypotheses.*

In a regular detection problem framework, the NP decision rule is to accept the hypothesis $\mathcal{H}_1$ if the LLRT statistic, $l(\underline{x})$, exceeds a critical value, and reject it otherwise. Furthermore, the critical value is set based on the rejection probability of the hypothesis $\mathcal{H}_0$, i.e. false-alarm probability. However, we pursue a different goal in the approximation problem scenario. Our goal is to approximate a model distribution with PDF $f_{\underline{X}_\mathcal{M}}(\underline{x})$, as close as possible to the given distribution with PDF $f_{\underline{X}}(\underline{x})$. The closeness criterion is based on the modified detection problem framework where we compute the LLRT statistic and compare it with a threshold. In an ideal case where there is no approximation error, the detection probability must be equal to the false-alarm probability for the optimal detector at all possible thresholds, i.e. the ROC curve [23] that represents best detectors for all threshold values should be a line of slope 1 passing through the origin.

In the next subsection, we assume that the random vector $\underline{X}$ has zero-mean Gaussian distribution. Thus, the covariance matrix of the random vector $\underline{X}$ is the parameter of interest in the model selection, i.e. covariance selection.

C) Multivariate Gaussian distribution

Let random vector $\underline{X} \in \mathbb{R}^n$, have a zero-mean jointly Gaussian distribution with covariance matrix $\mathbf{\Sigma}_{\underline{X}}$, i.e. $\underline{X} \sim \mathcal{N}(\underline{0}, \mathbf{\Sigma}_{\underline{X}})$ where the covariance matrix $\mathbf{\Sigma}_{\underline{X}}$ is positive-definite, $\mathbf{\Sigma}_{\underline{X}} \succ 0$. In this paper, the null hypothesis, $\mathcal{H}_0$, is the hypothesis that the parameter of interest is known and is equal to $\mathbf{\Sigma}_{\underline{X}}$ while the alternative hypothesis, $\mathcal{H}_1$, is the hypothesis that the random vector $\underline{X}$ is replaced by the model random vector $\underline{X}_\mathcal{M}$. In this scenario, the model random vector $\underline{X}_\mathcal{M}$ has a zero-mean jointly Gaussian distribution (the model approximation distribution) with covariance matrix $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}$ i.e. $\underline{X}_\mathcal{M} \sim \mathcal{N}(\underline{0}, \mathbf{\Sigma}_{\underline{X}_\mathcal{M}})$ where the covariance matrix $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}$ is also positive-definite, $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}} \succ 0$. Thus, the LLRT statistic for the jointly Gaussian random vectors ($\underline{X}$ and $\underline{X}_\mathcal{M}$) is simplified as

$$l(\underline{x}) = \log \frac{\mathcal{N}(\underline{0}, \mathbf{\Sigma}_{\underline{X}_\mathcal{M}})}{\mathcal{N}(\underline{0}, \mathbf{\Sigma}_{\underline{X}})} = -c + k(\underline{x}), \tag{1}$$

where $c = -1/2\log(|\mathbf{\Sigma}_{\underline{X}}\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}^{-1}|)$ is a constant and $k(\underline{x}) = \underline{x}^T\mathbf{K}\underline{x}$ where $\mathbf{K} = 1/2(\mathbf{\Sigma}_{\underline{X}}^{-1} - \mathbf{\Sigma}_{\underline{X}_\mathcal{M}}^{-1})$ is an indefinite matrix with both positive and negative eigenvalues.

We define the CAM associated with the covariance selection problem and dissimilarity parameters of the CAM as follows.

Definition 2 (Correlation approximation matrix). *The CAM for the covariance selection problem is defined as* $\mathbf{\Delta} \triangleq \mathbf{\Sigma}_{\underline{X}}\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}^{-1}$ *where* $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}$ *is the model covariance matrix.*

Definition 3 (Dissimilarity parameters for covariance selection problem). *Let* $\alpha_i \triangleq \lambda_i + \lambda_i^{-1} - 2$ *for* $i \in \{1, \ldots, n\}$ *be dissimilarity parameters of the CAM correspond to the covariance selection problem where* $\lambda_i > 0$ *for* $i \in \{1, \ldots, n\}$ *are eigenvalues of the CAM.*

Remark. *The CAM is a positive definite matrix. Moreover, eigenvalues of the CAM contains all information necessary to compute cost functions associated with the model selection.*

Theorem 1 (Covariance Selection [4]). *Given a multivariate Gaussian distribution with covariance matrix* $\mathbf{\Sigma}_{\underline{X}} \succ 0$, $f_{\underline{X}}(\underline{x})$, *and a model* $\mathcal{M}$, *there exists a unique approximate multivariate Gaussian distribution with covariance matrix* $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}} \succ 0$, $f_{\underline{X}_\mathcal{M}}(\underline{x})$, *that minimize the KL divergence,* $\mathcal{D}(f_{\underline{X}}(\underline{x})||f_{\underline{X}_\mathcal{M}}(\underline{x}))$ *and satisfies the covariance selection rules, i.e. the model covariance matrix satisfies the following covariance selection rules*

- $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}(i, i) = \mathbf{\Sigma}_{\underline{X}}(i, i), \quad \forall\, i \in \mathcal{V}$
- $\mathbf{\Sigma}_{\underline{X}_\mathcal{M}}(i, j) = \mathbf{\Sigma}_{\underline{X}}(i, j), \quad \forall\, (i, j) \in \mathcal{E}_\mathcal{M}$
- $\mathbf{\Sigma}^{-1}_{\underline{X}_\mathcal{M}}(i, j) = 0, \quad \forall\, (i, j) \in \mathcal{E}^c_\mathcal{M}$

where the set $\mathcal{E}^c_\mathcal{M} = \psi - \mathcal{E}_\mathcal{M}$ *represents the complement of the set* $\mathcal{E}_\mathcal{M}$.

Remark. *The CAM is defined as* $\mathbf{\Delta} \triangleq \mathbf{\Sigma}_{\underline{X}}\mathbf{\Sigma}^{-1}_{\underline{X}_\mathcal{M}}$. *Thus, the constant c can be written as* $c = -1/2\log(|\mathbf{\Delta}|)$.

Then, for any given covariance matrix and its model covariance matrix that satisfies conditions in Theorem 1, the summation of diagonal coefficients of the CAM is equal to n, i.e. the result in Theorem 1 implies that $tr(\mathbf{\Delta}) = n$. Using this result and the definition of the KL divergence for jointly Gaussian distributions, we have

$$\mathcal{D}(f_{\underline{X}}(\underline{x})||f_{\underline{X}_\mathcal{M}}(\underline{x})) = c + \frac{1}{2}tr(\mathbf{\Delta}) - \frac{n}{2}$$

which results in $c = \mathcal{D}(f_{\underline{X}}(\underline{x})||f_{\underline{X}_\mathcal{M}}(\underline{x}))$.

D) Covariance selection example

Here we choose the tree approximation model as an example. Figure 1 indicates two graphs: (a) the complete graph and (b) its tree approximation model where edges in the graph represent non-zero coefficients in the inverse of the covariance matrix [4].

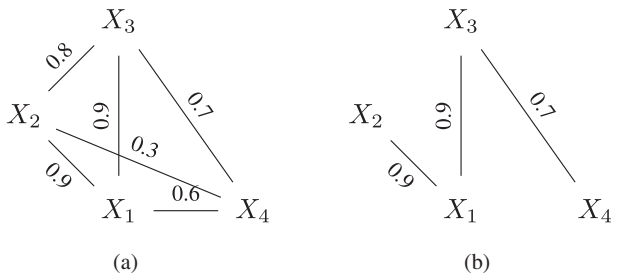

Fig. 1. (a) The complete graph; (b) The tree approximation of the complete graph.

The correlation coefficient between each pair of adjacent nodes has been written on each edge. The correlation coefficient between each pair of nonadjacent nodes is the multiplication of all correlations on the unique path that connects those nodes. The correlation matrix for each graph is

$$\mathbf{\Sigma}_{\underline{X}} = \begin{bmatrix} 1 & 0.9 & 0.9 & 0.6 \\ 0.9 & 1 & 0.8 & 0.3 \\ 0.9 & 0.8 & 1 & 0.7 \\ 0.6 & 0.3 & 0.7 & 1 \end{bmatrix}$$

and

$$\mathbf{\Sigma}_{\underline{X}_{\mathcal{T}}} = \begin{bmatrix} 1 & 0.9 & 0.9 & 0.63 \\ 0.9 & 1 & 0.81 & 0.567 \\ 0.9 & 0.81 & 1 & 0.7 \\ 0.63 & 0.567 & 0.7 & 1 \end{bmatrix}.$$

The CAM for the above example is

$$\mathbf{\Delta} = \begin{bmatrix} 1 & 0 & 0.0412 & -0.0588 \\ 0.0474 & 1 & 0.3042 & -0.5098 \\ 0.0474 & -0.0526 & 1 & 0 \\ 0.9789 & -1.2632 & 0.1421 & 1 \end{bmatrix}.$$

The CAM contains all information about the tree approximation.[2] Here we assume cases that Gaussian random variables have finite, nonzero variances. The value of the KL divergence for this example is $-0.5\log(|\mathbf{\Delta}|) = 0.6218$.

Remark. *Without loss of generality, throughout this paper, we work with normalized correlation matrices, i.e. the diagonal elements of the correlation matrices are normalized to be equal to one.*

E) Distribution of the LLRT statistic

The random vector $\underline{X}$ has Gaussian distribution under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Thus under both hypotheses, the real random variable, $k(\underline{X}) = \underline{X}^T\mathbf{K}\underline{X}$ has a generalized chi-squared distribution, i.e. the random variable, $k(\underline{X})$, is equal to a weighted sum of chi-squared random variables with both positive and negative weights under both hypotheses. Let us define $\underline{W} = \mathbf{\Sigma}_{\underline{X}}^{-1/2}\underline{X}$ under $\mathcal{H}_0$ and $\underline{Z} = \mathbf{\Sigma}_{\underline{X}_{\mathcal{M}}}^{-1/2}\underline{X}$ under $\mathcal{H}_1$, where $\mathbf{\Sigma}_{\underline{X}}^{1/2}$ and $\mathbf{\Sigma}_{\underline{X}_{\mathcal{M}}}^{1/2}$ are the square root of covariance matrices $\mathbf{\Sigma}_{\underline{X}}$ and $\mathbf{\Sigma}_{\underline{X}_{\mathcal{M}}}$, respectively. Then the random vectors $\underline{W} \sim \mathcal{N}(\underline{0}, \mathbf{I})$ and $\underline{Z} \sim \mathcal{N}(\underline{0}, \mathbf{I})$ are zero-mean Gaussian distributions with the same covariance matrices, $\mathbf{I}$, where $\mathbf{I}$ is the identity matrix of dimension n. Note that, the CAM is a positive definite matrix with $\lambda_i > 0$ where $1 \leq i \leq n$. Thus, the random variable $k(\underline{X})$, under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ can be written as:

$$K_0 \triangleq k(\underline{X})|\mathcal{H}_0 = \frac{1}{2}\sum_{i=1}^{n}(1-\lambda_i)W_i^2$$

and

$$K_1 \triangleq k(\underline{X})|\mathcal{H}_1 = \frac{1}{2}\sum_{i=1}^{n}(\lambda_i^{-1}-1)Z_i^2$$

[2]Dissimilarity parameters α_i and eigenvalues of CAM contains all information about the tree approximation.

respectively, where random variables W_i and Z_i, are the i-th element of random vectors $\underline{W}$ and $\underline{Z}$, respectively. Moreover, random variables W_i^2 and Z_i^2, follow the first-order central chi-squared distribution. Note that, similarly random variable $l(\underline{X}) \triangleq -c + k(\underline{X})$ is defined under each hypothesis as

$$L_0 \triangleq l(\underline{X})|\mathcal{H}_0 = -c + K_0$$

and

$$L_1 \triangleq l(\underline{X})|\mathcal{H}_1 = -c + K_1.$$

Remark. *As a simple consequence of the covariance selection theorem, the summation of weights for the generalized chi-squared random variable, the expectation of $k(\underline{X})$, is zero under the hypothesis $\mathcal{H}_0$, i.e. $\mathrm{E}(K_0) = 1/2\sum_{i=1}^{n}(1-\lambda_i) = 0$ [4], and this summation is positive under the hypothesis $\mathcal{H}_1$, i.e. $\mathrm{E}(K_1) = 1/2\sum_{i=1}^{n}(\lambda_i^{-1}-1) \geq 0$.*

III. THE ROC CURVE AND THE AUC COMPUTATION

In this section, we focus on studying the properties of the ROC curve and finding an analytical expression for the AUC which again depends on the eigenvalues of the CAM.

A) The ROC curve

The ROC curve is the parametric curve where the detection probability is plotted versus the false-alarm probability for all thresholds, i.e. each point on the ROC curve represents a pair of $(P_0(\tau), P_1(\tau))$ for a given threshold τ. Set $z = P_0(\tau)$ and $\eta = P_1(\tau)$, the ROC curve is $\eta = h(z)$. If $P_0(\tau)$ has an inverse function, then the ROC curve is $h(z) = P_1(P_0^{-1}(z))$. In general, the ROC curve, $h(z)$, has the following properties [23]

- $h(z)$ is concave and increasing,
- $h'(z)$ is positive and decreasing,
- $\int_0^1 h'(z)\,dz \leq 1$.

Note that, for the ROC curve, the slope of the tangent line at a given threshold, $h'(z)$, gives the likelihood ratio for the value of the test [23].

Remark. *For the ROC curve for our Gaussian random vectors, we have $h'(z)$ is positive, continuous and decreasing in the interval $[0,\ 1]$ with right continuity at 0 and left continuity at 1. Moreover,*

$$\int_0^1 h'(z)\,dz = 1$$

since $h(0) = 0$ and $h(1) = 1$.

Definition 4. *Let $f_{L_0}(l)$ and $f_{L_1}(l)$ be the probability density function of the random variables L_0 and L_1, respectively.*

Lemma 1. *Given the ROC curve, $h(z)$, we can compute following KL divergences*

$$\mathcal{D}(f_{L_1}(l)||f_{L_0}(l)) = -\int_0^1 \log(h'(z))\,dz.$$

and

$$\begin{aligned}\mathcal{D}(f_{L_0}(l)||f_{L_1}(l)) &= -\int_0^1 h'(z)\,\log(h'(z))\,dz \\ &\overset{(*)}{=} -\int_0^1 \log\left(\frac{d\,h^{-1}(\eta)}{d\,\eta}\right) d\eta\end{aligned}$$

where (*) holds if the ROC curve, $\eta = h(z)$, has an inverse function.

Proof: These results are from the Radon–Nikodým theorem [24]. Simple, alternative calculus-based proofs are given in Appendix VI. □

B) Area under the curve

As discussed previously, we examine the ROC with a goal that the model approximation results in the ROC being a line of slope 1 passing through the origin. This is in contrast to the conventional detection problem where we want to distinguish between the two hypotheses and ideally have a ROC that is a unit step function. AUC is defined as the integral of the ROC curve (Fig. 2) and is a measure of accuracy in decision problems.

Definition 5. *The area under the ROC curve (AUC) is defined as*

$$AUC = \int_0^1 h(z)\,d\,z = \int_0^1 P_1(\tau)\,d\,P_0(\tau), \qquad (2)$$

where τ is the detection problem threshold.

Remark. *The AUC is a measure of accuracy for the detection problem and $1/2 \leq \text{AUC} \leq 1$. Note that, in conventional decision problems, the AUC is desired to be as close as possible to 1 while in approximation problem presented here we want the AUC to be close to $1/2$.*

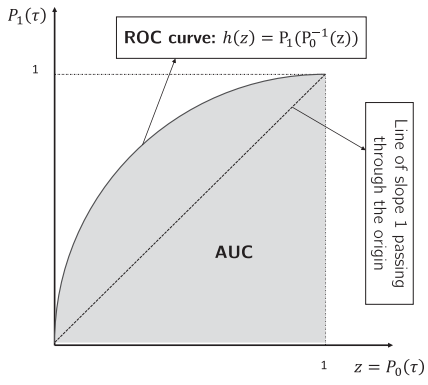

Fig. 2. The ROC curve and the area under the ROC curve. Each point on the ROC curve indicates a detector with given detection and false-alarm probabilities.

Theorem 2 (Statistical property of AUC [25]). *The AUC for the LLRT statistic is*

$$AUC = \Pr(L_1 > L_0).$$

Corollary 1. *From Theorem 2, when PDFs for the LLRT statistic under both hypotheses exist, we can compute the AUC as*

$$AUC = \int_0^\infty (f_{L_1} \star f_{L_0})(l)\,dl, \qquad (3)$$

where $\left(f_{L_1} \star f_{L_0}\right)(l) \triangleq \int_{-\infty}^{\infty} f_{L_1}(\tau)\,f_{L_0}(l+\tau)\,dl$ is the cross-correlation between $f_{L_1}(l)$ and $f_{L_0}(l)$.

Proof: A proof based on the definition of the AUC (2), is given in [26]. □

Let us define the difference LLRT statistic random variable as $L_\Delta \triangleq L_1 - L_0$. Then, we get

$$\begin{aligned}AUC &= \Pr(L_\Delta > 0) \\ &= 1 - F_{L_\Delta}(0)\end{aligned}$$

where $F_{L_\Delta}(l)$ is the cumulative distribution function (CDF) for random variable L_Δ. Note that we define the difference LLRT statistic random variable to simplify the notation and easily show that the AUC only depends on this difference.

The two conditional random variables L_0 and L_1 are independent.[3] Thus, the cross-correlation between the corresponding two distributions is the distribution of the difference LLRT statistic, L_Δ. We can write the random variable L_Δ as

$$\begin{aligned}L_\Delta &= -c + K_1 - (-c + K_0) \\ &= K_1 - K_0.\end{aligned}$$

Replacing the definition for K_0 and K_1, we have

$$L_\Delta = \frac{1}{2}\sum_{i=1}^{n}(\lambda_i^{-1} - 1)Z_i^2 - \frac{1}{2}\sum_{i=1}^{n}(1-\lambda_i)W_i^2.$$

We can rewrite the difference LLRT statistic, L_Δ, in an indefinite quadratic form as

$$L_\Delta = \frac{1}{2}\underline{V}^T(\mathbf{\Lambda} - \mathbf{I})\underline{V}$$

where

$$\underline{V} = \begin{bmatrix}\underline{W} \\ \underline{Z}\end{bmatrix}$$

and

$$\mathbf{\Lambda} = \begin{bmatrix}\lambda_1 & & & & \mathbf{0} \\ & \ddots & & & \\ & & \lambda_n & & \\ & & & \lambda_1^{-1} & \\ \mathbf{0} & & & & \ddots \\ & & & & & \lambda_n^{-1}\end{bmatrix}.$$

[3] By the definition of the detection problem.

C) Analytical expression for AUC

To compute the CDF of random variable L_Δ, we need to evaluate a multi-dimensional integral of jointly Gaussian distributions [27] or we need to approximate this CDF [28]. More efficiently, as discussed in [29] for the real-valued case, the CDF of the random variable L_Δ can be expressed as a single-dimensional integral of a complex function[4] in the following form

$$F_{L_\Delta}(l) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{(l/2)(j\omega+\beta)}}{j\omega+\beta} \times \frac{1}{\sqrt{|\mathbf{I}+(1/2)(\mathbf{\Lambda}-\mathbf{I})(j\omega+\beta)|}} d\omega$$

where $\beta > 0$ is chosen such that matrix $\mathbf{I} + \beta/2(\mathbf{\Lambda} - \mathbf{I})$, is positive definite and simplifies the evaluation of the multivariate Gaussian integral [29].

Special case: When $\mathbf{\Lambda} = \mathbf{I}$, i.e. the given covariance obeys the model structure, then

$$AUC = 1 - F_{L_\Delta}(0) = 1 - \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{j\omega+\beta} = \frac{1}{2}$$

for $\beta > 0$ and is also independent of the value of the parameter β.

Picking an appropriate value for the parameter β,[5] the AUC can be numerically computed by evaluating the following one dimension complex integral

$$AUC = 1 - \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{j\omega+\beta} \times \frac{1}{\sqrt{|\mathbf{I}+1/2(\mathbf{\Lambda}-\mathbf{I})(j\omega+\beta)|}} d\omega.$$

Furthermore, since $\mathbf{\Lambda} \succ 0$, choosing $\beta = 2$ and changing variable as $\nu = \omega/2$, we have

$$AUC = 1 - \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{j\nu+1} \frac{1}{\sqrt{|\mathbf{\Lambda}+j\nu(\mathbf{\Lambda}-\mathbf{I})|}} d\nu. \quad (4)$$

Moreover, $|\mathbf{\Lambda} + j\nu(\mathbf{\Lambda} - \mathbf{I})| = \prod_{i=1}^{p}(1 + \alpha_i\nu^2 - j\alpha_i\nu)$. This equation shows that the AUC only depends on α_i.

Remark. *Since the AUC integral in (4) cannot be evaluated in closed form, it cannot be used directly in obtaining model selection algorithms. Numerical evaluation of the AUC using the one-dimensional complex integral (4) is very efficient and fast compared with the numerical evaluation of a multi-dimensional integral of jointly Gaussian CDF.*

IV. ANALYTICAL BOUNDS FOR THE AUC

Section III derived an analytical expression for the AUC based on zero mean Gaussian distributions. In this section, we find analytical lower and upper bounds for the AUC. These bounds will give us insight on the behavior of the AUC. The lower bound for the AUC depends directly on the eigenvalues of the CAM (using Chernoff bounds) whereas the upper bound depends indirectly on the eigenvalues of the CAM through the KL and reverse KL divergences (using properties of the ROC curve).

A) Generalized asymmetric Laplace distribution

In this subsection, we present the probability density function and moment generating function for the difference LLRT statistic random variable, L_Δ. We will use this result in computing the AUC bound.

The difference LLRT statistic random variable, L_Δ, follows the generalized asymmetric Laplace (GAL) distribution[6] [30]. For a given i where $i \in \{1, \ldots, n\}$, we define a random variable L_{Δ_i} as

$$L_{\Delta_i} = \frac{\lambda_i - 1}{2} W_i^2 - \frac{1 - \lambda_i^{-1}}{2} Z_i^2. \quad (5)$$

Then, difference LLRT statistic random variable, L_Δ, can be written as

$$L_\Delta = \sum_{i=1}^{n} L_{\Delta_i}$$

where L_{Δ_i} are independent and have GAL distributions at position 0 with mean $\alpha_i/2$ and PDF [30]

$$f_{L_{\Delta_i}}(l) = \frac{e^{l/2}}{\pi\sqrt{\alpha_i}} K_0\left(\sqrt{\alpha_i^{-1} + \frac{1}{4}}\,|l|\right), \quad l \neq 0, \quad (6)$$

where $K_0(-)$ is the modified Bessel function of second kind [31]. The moment generating function (MGF) for this distribution is

$$M_{L_{\Delta_i}}(t) = \frac{1}{\sqrt{1 - \alpha_i t - \alpha_i t^2}}$$

for all t that satisfies $1 - \alpha_i t - \alpha_i t^2 > 0$. From (5), the MGF derivation for the GAL distribution is straightforward and is the multiplication of two MGFs for the chi-squared distribution.

The distribution of the difference LLRT statistic random variable, L_Δ, is

$$f_{L_\Delta}(l) = \overset{n}{\underset{i=1}{*}} f_{L_{\Delta_i}}(l)$$

where $*_{i=1}^{n}$ is the notation we use for the convolution of n functions together. Note that, although the distribution of random variables L_{Δ_i} in (6) has a discontinuity at $l = 0$, the distribution of random variable L_Δ is continuous if there are at least two distribution with non-zero parameters, α_i,

[4] This is the transform to the frequency domain for an arbitrary β.

[5] The parameter β is picked such that $\mathbf{I} + \beta/2(\mathbf{\Lambda} - \mathbf{I}) \succ 0$ and $\beta = 2$ always satisfies this condition since $\mathbf{\Lambda} \succ 0$.

[6] Also known as the variance-gamma distribution or the Bessel function distribution.

in the aforementioned convolution. Moreover, the MGF for $f_{L_\Delta}(l)$ can be computed by multiplying MGFs for L_{Δ_i} as

$$M_{L_\Delta}(t) = \prod_{i=1}^{n} M_{L_{\Delta_i}}(t) \tag{7}$$

for all t in the intersection of all domains of $M_{L_{\Delta_i}}(t)$. The smallest of such intersections is $-1 < t < 0$.

B) Lower bound for the AUC (Chernoff bound application)

Given the MGF for the difference LLRT statistic distribution (7), we can apply the Chernoff bound [32] to find a lower bound for the AUC or upper bound for the CDF of the difference LLRT statistic random variable, L_Δ, evaluated at zero).

Proposition 2. *Lower bound for the AUC is*

$$\Pr(L_\Delta > 0) \geq \max\left\{\frac{1}{2}, 1 - e^{-(1/2)\sum_{i=1}^{n}\log(1+(\alpha_i/4))}\right\} \tag{8}$$

Proof: One-half is a trivial lower bound for AUC. To achieve a non-trivial lower bound, we apply Chernoff bound [32] as follows

$$\Pr(L_\Delta < 0) \leq \inf_t M_{L_\Delta}(t).$$

To complete the proof we need to solve the right-hand-side (RHS) optimization problem.

Step 1: First derivatives of $M_{L_\Delta}(t)$ is

$$\frac{d}{dt}M_{L_\Delta}(t) = M_{L_\Delta}(t) \left(\frac{1}{2}\sum_{i=1}^{n}\frac{\lambda_i - 1}{1-(\lambda_i-1)t} + \frac{\lambda_i^{-1}-1}{1-(\lambda_i^{-1}-1)t}\right)$$
$$= M_{L_\Delta}(t)(1+2t)\sum_{i=1}^{n}\frac{\alpha_i}{2(1-\alpha_i t-\alpha_i t^2)}.$$

Clearly, the first derivative is zero for $t = -1/2$ which is in the feasible domain of the MGF for the difference LLRT statistic. Note that, the smallest feasible domain is $-1 < t < 0$.

Step 2: Second derivatives of $M_{L_\Delta}(t)$ is

$$\frac{d^2}{dt^2}M_{L_\Delta}(t) = M_{L_\Delta}(t) \times \left(\frac{1}{4}\sum_{i=1}^{n}\frac{\lambda_i - 1}{1-(\lambda_i-1)t} + \frac{\lambda_i^{-1}-1}{1-(\lambda_i^{-1}-1)t}\right)^2$$
$$+ M_{L_\Delta}(t)\left(\frac{1}{4}\sum_{i=1}^{n}\frac{(\lambda_i-1)^2}{(1-(\lambda_i-1)t)^2} + \frac{(\lambda_i^{-1}-1)^2}{(1-(\lambda_i^{-1}-1)t)^2}\right).$$

Therefore, we conclude that the second derivative is positive and thus the optimal solution to the RHS optimization problem is at $t = -1/2$. Replacing that in the definition of the moment generation function which results in the following bound

$$\Pr(L_\Delta \leq 0) < \prod_{i=1}^{n}\frac{2}{\sqrt{4+\alpha_i}}$$

which can be written as

$$\Pr(L_\Delta > 0) \geq 1 - \prod_{i=1}^{n}\frac{2}{\sqrt{4+\alpha_i}}$$

which completes the proof. □

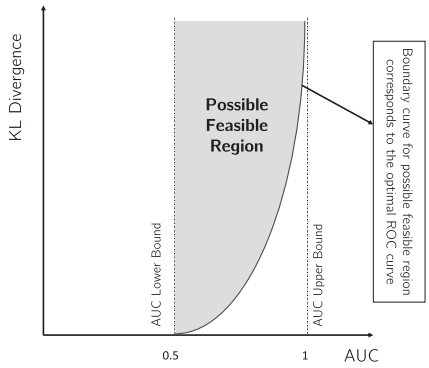

Fig. 3. Possible feasible region for the AUC and the Kl divergence pair for all possible detectors or equivalently all possible ROC curves (the KL divergence is between the LLRT statistics under different hypotheses, i.e. $\mathcal{D}(f_{L_0}(l)||f_{L_1}(l))$ or $\mathcal{D}(f_{L_1}(l)||f_{L_0}(l))$.).

C) Upper bound for the AUC

In this section, we present a parametric upper bound for the AUC, but first, we need to present the following results.

Lemma 2. *Data processing inequality of the KL divergence for the LLRT statistic. We have*

$$\mathcal{D}(f_{L_1}(l)||f_{L_0}(l)) \leq \mathcal{D}(f_{\underline{X}}(\underline{x}|\mathcal{H}_1)||f_{\underline{X}}(\underline{x}|\mathcal{H}_0))$$

and

$$\mathcal{D}(f_{L_0}(l)||f_{L_1}(l)) \leq \mathcal{D}(f_{\underline{X}}(\underline{x}|\mathcal{H}_0)||f_{\underline{X}}(\underline{x}|\mathcal{H}_1)).$$

Proof: This lemma is a special case of the data processing property for the KL divergence [33]. By picking appropriate measurable mapping, here appropriate quadratic function for each equation of the above equations, we conclude the lemma. □

Definition 6 (Feasible Region). *The AUC and the KL divergence pair lie in the feasible region (Fig. 3) for all possible detectors (ROC curves), i.e. no detector with the AUC and the KL divergence pair lie outside the feasible region.*[7]

[7] The definition of the feasible region here is inspired by the joint range of f-divergences [34].

Theorem 3 (Possible feasible region for the AUC and the KL divergence). *Given the ROC curve, the parametric possible feasible region as shown in Fig. 3 can be expressed using the positive parameter $a > 0$ as*

$$\Pr(L_\Delta > 0) = \frac{1}{1-e^{-a}} - \frac{1}{a}$$

and

$$\mathcal{D}_l^* \geq \log(a) + \frac{a}{e^a - 1} - 1 - \log(1 - e^{-a})$$

where

$$\mathcal{D}_l^* = \min\{\mathcal{D}(f_{L_1}(l)||f_{L_0}(l)),\ \mathcal{D}(f_{L_0}(l)||f_{L_1}(l))\}.$$

Proof: Proof is given in the Appendix VI. □

Theorem 3 formulates the relationship between the AUC and the KL divergence. *The results of this theorem is generally true for any LLRT statistic*. Theorem 3 states that for any valid ROC that corresponds to a detection problem, the pair of AUC and KL divergence *must* lie in the possible feasible region (Fig. 3), i.e. outside of this region is infeasible. This possible feasible region results in the general upper bound for AUC.

Since computing, the distribution of the LLRT statistics is not straightforward in most cases, Proposition 3, relaxes the Theorem 3 by bounding the KL divergence between the LLRT statistics using the invariance property of KL divergence for the LLRT statistic (Lemma 2).

Proposition 3. *The parametric upper bound for AUC is*

$$\Pr(L_\Delta > 0) = \frac{1}{1-e^{-a}} - \frac{1}{a}$$

and

$$\mathcal{D}^* \geq \log(a) + \frac{a}{e^a - 1} - 1 - \log(1 - e^{-a})$$

where $a > 0$ is a positive parameter and

$$\mathcal{D}^* = \min\{\mathcal{D}(f_{\underline{X}}(\underline{x}|\mathcal{H}_1)||f_{\underline{X}}(\underline{x}|\mathcal{H}_0)),\ \mathcal{D}(f_{\underline{X}}(\underline{x}|\mathcal{H}_0)||f_{\underline{X}}(\underline{x}|\mathcal{H}_1))\}. \quad (9)$$

Proof: Proof is based on the Lemma 2 and the possible feasible region presented in the Theorem 3. From the Lemma 2, we have

$$\mathcal{D}_l^* \leq \mathcal{D}^*.$$

Then, using the result in the Theorem 3, we get the parametric upper bound. □

D) Asymptotic behavior for AUC bounds

Proposition 4 (Asymptotic behavior of the lower bound). *We have*

$$\Pr(L_\Delta > 0) \geq 1 - e^{-n\left(1 - 1/n\sum_{i=1}^{n}(1+(\alpha_i)/(8))^{-1}\right)}.$$

Proof: Applying the inequality

$$\frac{2x}{2+x} < \log(1+x)$$

for $x > 0$, we achieve the result. □

Proposition 5 (Asymptotic behavior of the upper bound). *The parametric upper bound for AUC has the following asymptotic behavior*

$$\Pr(L_\Delta > 0) \leq 1 - e^{-\mathcal{D}^* - 1}$$

where $\mathcal{D}^*$ is given in (9).

Proof: Proof is as follows.

$$\begin{aligned} -\log(1 - \Pr(L_\Delta > 0)) &= -\log\left(\frac{1}{e^a - 1} + \frac{1}{a}\right) \\ &\leq \log(a) \\ &\leq \mathcal{D}^* + 1. \end{aligned}$$

Applying the exponential function to both sides of the above inequality, we get the upper bound. □

Remark. *The asymptotic lower bound is a function of the number of nodes, n and has an exponential decaying behavior. The asymptotic upper bound also has an exponential decaying behavior with respect to KL divergence.*

Figure 4 shows the possible feasible region and the asymptotic behavior log-scale. As it is shown in this figure, the parametric upper bound can be approximated with a straight line especially for large values of the parameter a (the result in Proposition 5). Also, Fig. 5 shows the possible feasible region and the asymptotic behavior in regular-scale.

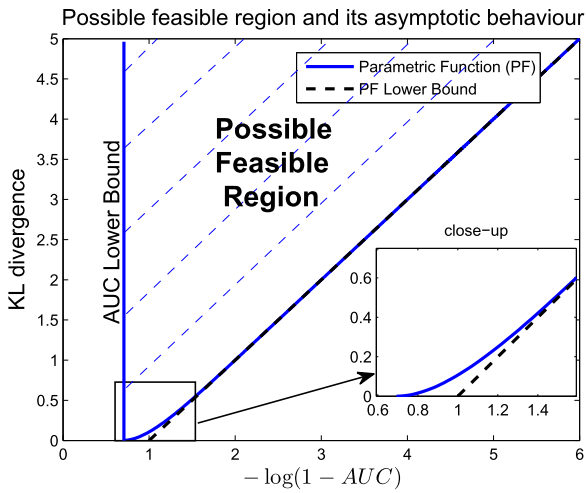

Fig. 4. Log-scale of the possible feasible region and its asymptotic behavior (linear line) for the AUC and the KL divergence pair for all possible detectors or equivalently all possible ROC curves (the KL divergence is between the LLRT statistics under different hypotheses, i.e. $\mathcal{D}(f_{L_1}(l)||f_{L_0}(l))$ or $\mathcal{D}(f_{L_0}(l)||f_{L_1}(l))$.) Close-up part shows the non-linear behavior of the possible feasible region around one.

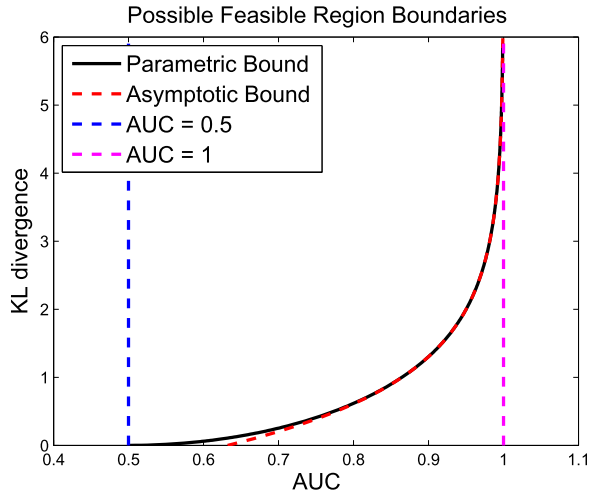

Fig. 5. The possible feasible region boundaries and its asymptotic behavior for the AUC and the KL divergence pair for all possible detectors or equivalently all possible ROC curves (the KL divergence is between the LLRT statistics under different hypotheses, i.e. $\mathcal{D}(f_{L_0}(l)||f_{L_1}(l))$ or $\mathcal{D}(f_{L_1}(l)||f_{L_0}(l))$.).

V. EXAMPLES AND SIMULATION RESULTS

In this section, we consider some examples of covariance matrices for Gaussian random vector $\underline{X}$. We pick the tree structure as the graphical model corresponds to the covariance selection problem. In our simulations, we compare the numerically evaluated AUC and its lower and upper bounds and discuss their asymptotic behavior as the dimension of the graphical model, n, increases.

A) Tree approximation model

The maximum order of the lower order distributions in tree approximation problem is two, i.e. no more than pairs of variables. Let $\underline{X}_{\mathcal{T}} \sim \mathcal{N}(\underline{0}, \mathbf{\Sigma}_{\underline{X}_{\mathcal{T}}})$ have the graph representation $\mathcal{G}_{\mathcal{T}} = (\mathcal{V}, \mathcal{E}_{\mathcal{T}})$ where $\mathcal{E}_{\mathcal{T}} \subseteq \psi$ is a set of edges that represents a tree structure. Let $\underline{X}_r \sim \mathcal{N}(\underline{0}, \mathbf{\Sigma}_{\underline{X}_r})$ have the graph representation $\mathcal{G}_r = (\mathcal{V}, \mathcal{E}_r)$ where $\mathcal{E}_r \subseteq \mathcal{E}_{\mathcal{T}}$ is the set of all edges in the graph of $\underline{X}_r$. The joint PDF for elements of random vector $\underline{X}_r$ can be represented by joint PDFs of two variables and marginal PDFs in the following convenient form

$$f_{\underline{X}_r}(\underline{x}_r) = \prod_{(u,v)\in\mathcal{E}_r} \frac{f_{\underline{X}^u,\underline{X}^v}(\underline{x}^u, \underline{x}^v)}{f_{\underline{X}^u}(\underline{x}^u) f_{\underline{X}^v}(\underline{x}^v)} \prod_{u\in\mathcal{V}} f_{\underline{X}^u}(\underline{x}^u). \quad (10)$$

Using equation (10), we can then easily construct a tree using iterative algorithms (such as the Chow-Liu algorithm [9] combined with the Kruskal [10] algorithm or the Prim [11] algorithm) by adding edges one at a time [35]. Consider the sequence of random vectors $\underline{X}_r$ with $0 \le r \le |\mathcal{E}_{\mathcal{T}}|$, where $\underline{X}_r$ is recursively generated by augmenting a new edge, $(i, j) \in \mathcal{E}_r$, to the graph representation of $\underline{X}_{r-1}$. For the special case of Gaussian distributions, $\mathbf{\Sigma}_{\underline{X}_r}$ has the following recursive formulation [35]

$$\mathbf{\Sigma}_{\underline{X}_r}^{-1} = \mathbf{\Sigma}_{\underline{X}_{r-1}}^{-1} + \mathbf{\Sigma}_{i,j}^{\dagger} - \mathbf{\Sigma}_i^{\dagger} - \mathbf{\Sigma}_j^{\dagger}, \quad \forall\, 0 \le r \le |\mathcal{E}_{\mathcal{T}}|$$

where $\mathbf{\Sigma}_{i,j}^{\dagger} = [\underline{e}_i\ \underline{e}_j]\mathbf{\Sigma}_{i,j}^{-1}[\underline{e}_i\ \underline{e}_j]^T$ and $\mathbf{\Sigma}_i^{\dagger} = \underline{e}_i\mathbf{\Sigma}_i^{-1}\underline{e}_i^T$ where $\underline{e}_i$ is a unitary vector with 1 at the i-th place and $\mathbf{\Sigma}_{i,j}$ and $\mathbf{\Sigma}_i$ are the 2-by-2 and 1-by-1 principle sub-matrices of $\mathbf{\Sigma}_{\underline{X}}$, with initial step $\mathbf{\Sigma}_{\underline{X}_0} = diag(\mathbf{\Sigma}_{\underline{X}})$ where $diag(\mathbf{\Sigma}_{\underline{X}})$ represents a diagonal matrix with diagonal elements of $\mathbf{\Sigma}_{\underline{X}}$.

Remark. *For all* $0 \le r \le |\mathcal{E}_{\mathcal{T}}|$, *we have*

(i) $\mathrm{tr}(\mathbf{\Sigma}_{\underline{X}_r}) = \mathrm{tr}(\mathbf{\Sigma}_{\underline{X}})$
(ii) $\mathrm{tr}(\mathbf{\Sigma}_{\underline{X}}\mathbf{\Sigma}_{\underline{X}_r}^{-1}) = n.$
(iii) $\mathcal{D}(f_{\underline{X}}(\underline{x})||f_{\underline{X}_r}(\underline{x})) = -1/2\log(|\mathbf{\Sigma}_{\underline{X}}\mathbf{\Sigma}_{\underline{X}_r}^{-1}|)$
(iv) $|\mathbf{\Sigma}_{\underline{X}}| \le \ldots \le |\mathbf{\Sigma}_{\underline{X}_r}| \le \ldots \le |\mathbf{\Sigma}_{\underline{X}_0}| = |diag(\mathbf{\Sigma}_{\underline{X}})|$
(v) $H(\underline{X}) \le \ldots \le H(\underline{X}_r) \le \ldots \le H(\underline{X}_0)$

where $H(\underline{X})$ *is differential entropy.*

Tree approximation models are interesting to study since there are algorithms such as Chow-Liu [9] combined by the Kruskal [10] or the Prim's [11] that efficiently compute the model covariance matrix from the graph covariance matrix.

B) Toeplitz example

Here, we assume that the covariance matrix $\mathbf{\Sigma}_{\underline{X}}$ has a Toeplitz structure with ones on the diagonal elements and the correlation coefficient $\rho > -(1/(n-1))$ as off-diagonal elements

$$\mathbf{\Sigma}_{\underline{X}} = \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \ldots & \rho & 1 \end{bmatrix}.$$

For the tree structure model, all possible tree-structured distributions satisfying (10) have the same KL divergence to the original graph, i.e. $\mathcal{D}(f_{\underline{X}}(\underline{x})||f_{\underline{X}_{\mathcal{T}}}(\underline{x}))$ is constant for all possible connected tree approximation model for this example. The reason is that all the weights computed by the Chow-Liu algorithm to construct the weighted graph associated with this problem are the same and are equal to $-1/2\log(1+\rho^2)$, which only depends on the correlation coefficient ρ. In the sequel, we test our results for two tree structured networks: a star network and a chain network.[8]

1) STAR APPROXIMATION

The star covariance matrix is as follows (all the nodes are connected to the first node)[9]

$$\mathbf{\Sigma}_{\underline{X}_{\mathcal{T}}}^{star} = \begin{bmatrix} 1 & \rho & \ldots & \ldots & \rho \\ \rho & \ddots & \rho^2 & \ldots & \rho^2 \\ \vdots & \rho^2 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho^2 \\ \rho & \rho^2 & \ldots & \rho^2 & 1 \end{bmatrix}.$$

[8] A more comprehensive simulation study of this example is provided in [36].
[9] All n possible star networks have the same performance.

For this example, the KL divergence and the Jeffreys divergence can be computed in closed form as

$$\mathcal{D}(\underline{X}||\underline{X}_{star}) = \frac{1}{2}(n-1)\log(1+\rho) - \frac{1}{2}\log(1+(n-1)\rho)$$

and

$$\mathcal{D}_{\mathcal{J}}(\underline{X}, \underline{X}_{star}) = \frac{(n-1)(n-2)\rho^2}{2(1+(n-1)\rho)}$$

respectively, where

$$\mathcal{D}_{\mathcal{J}}(\underline{X}, \underline{X}_{star}) = \mathcal{D}(\underline{X}||\underline{X}_{star}) + \mathcal{D}(\underline{X}_{star}||\underline{X})$$

is the Jeffreys divergence [18]. Moreover, for large values of n, we have that

$$\mathcal{D}(\underline{X}||\underline{X}_{star}) \approx \frac{n}{2}\log(1+\rho)$$

and

$$\mathcal{D}_{\mathcal{J}}(\underline{X}, \underline{X}_{star}) \approx \frac{n}{2}\rho.$$

Figure 6 plots the 1−AUC versus the dimension of the graph, n for different correlation coefficients, $\rho = 0.1$ and $\rho = 0.9$. This figure also indicates the upper bound and the lower bound for the 1−AUC.

2) Chain approximation

The chain covariance matrix is as follows (nodes are connected like a first order Markov chain, 1 to n)

$$\mathbf{\Sigma}^{chain}_{\underline{X}_{\mathcal{T}}} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & \ddots & \ddots & \ddots & \vdots \\ \rho^2 & \ddots & \ddots & \ddots & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^{n-1} & \dots & \rho^2 & \rho & 1 \end{bmatrix}.$$

For this example, the KL divergence and the Jeffreys divergence can be computed in closed form as

$$\mathcal{D}(\underline{X}||\underline{X}_{chain}) = \mathcal{D}(\underline{X}||\underline{X}_{star})$$

and

$$\mathcal{D}_{\mathcal{J}}(\underline{X}, \underline{X}_{chain}) = \frac{\rho^2}{(1+(n-1)\rho)(1-\rho)} \times \left(\frac{n(n-1)}{2} - \frac{n(1-\rho^n)}{1-\rho} + \frac{1-(n+1)\rho^n + n\rho^{n+1}}{(1-\rho)^2}\right)$$

respectively. Moreover, for large values of n we have the following approximation

$$\mathcal{D}_{\mathcal{J}}(\underline{X}, \underline{X}_{chain}) \approx \frac{n}{2}\frac{\rho}{1-\rho}.$$

Figure 7 plots the 1−AUC versus the dimension of the graph, n for different correlation coefficients, $\rho = 0.1$ and $\rho = 0.9$ as well as its upper and lower bounds.

In both Figs 6 and 7, (1 − AUC) and its bounds rapidly go to 0 which means that AUC goes to one as we increase the number of nodes, n, in the graph. More precisely, bounds for 1 − AUC are decaying exponentially as the dimension of the graph, n, increases which is consistent with the theory obtained for analytical bounds. Furthermore, we can conclude from these figures that a smaller ρ results in a better tree approximation, i.e. covariance matrices with smaller correlation coefficients are more like tree structure model. Moreover, comparing the AUC for the star network approximation with the AUC for the chain network approximation we conclude that the star network is a much better approximation than the chain network even though that both approximation networks have the same KL divergences. We can also interpret this fact through the analytical bounds obtained in this paper. The star network is a better approximation than the chain network since the decay rate of 1 − AUC for the star network is less than its decay rate for the chain network.

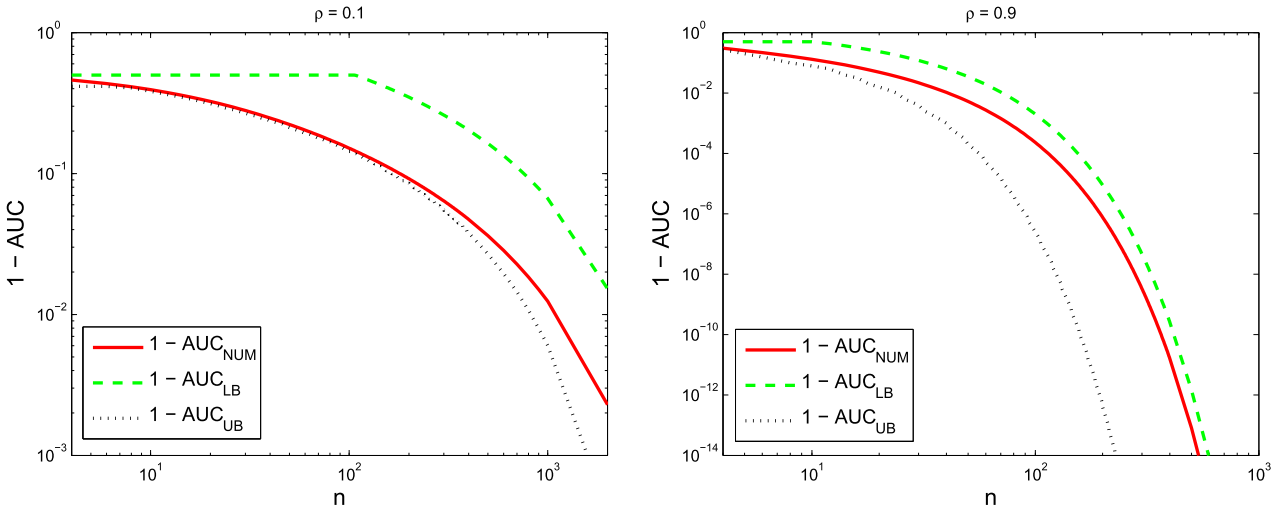

Fig. 6. 1−AUC versus the dimension of the graph, n for Star approximation of the Toeplitz example with $\rho = 0.1$ **(left)** and $\rho = 0.9$ **(right)**. In both figures, the numerically evaluated AUC is compared with its bounds.

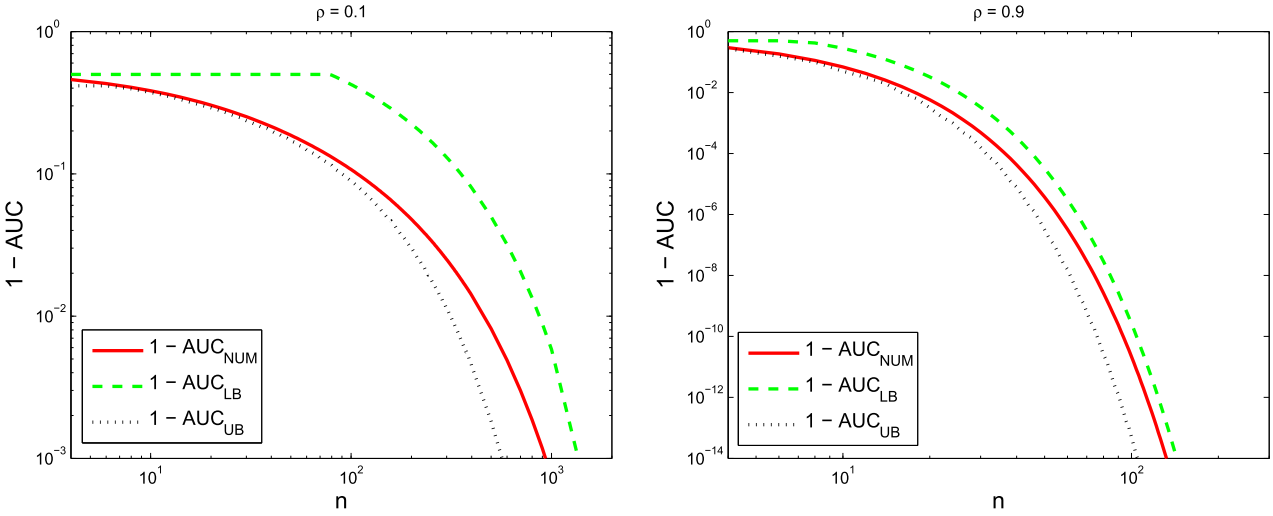

Fig. 7. 1−AUC versus the dimension of the graph, n for Chain approximation of the Toeplitz example with $\rho = 0.1$ **(left)** and $\rho = 0.9$ **(right)**. In both figures, the numerically evaluated AUC is compared with its bounds.

Remark. *The star approximation in the above example has lower AUC than the chain approximation. Practically, it means the correlation between nodes that are not connected in the approximated graphical structure gives a better approximation in the star network than the chain network.*

C) Solar data

In this subsection, we look at real solar data. As discussed in the introduction, part of our motivation for this research is based on looking at distributed state estimation for microgrids with distributed renewable energy sources. Solar radiation data at these energy sources are highly correlated and we look to approximate the distribution of this data with simpler approximations that can be represented by trees. Here we see that when the number of nodes is moderately large (19) (first example) that tree approximations do not work well. However, when we have a small number of sources (6) (second example), then with the proper edge inclusion tree approximations work well.

In this Example, a covariance matrix is calculated based on datasets presented in [37]. Two datasets are used from the National Renewable Energy Laboratory (NREL) website [38]. The first data set is the Oahu solar measurement grid which consists of 19 sensors (17 horizontal sensors and two tilted sensors) and the second one is the NREL solar data for 6 sites near Denver, Colorado. These two data sets are normalized using the standard normalization method and the zenith angle normalization method [37] and then the unbiased estimate of the correlation matrix is computed.[10]

1) The Oahu solar measurement grid dataset

From data obtained from 19 solar sensors at the island of Oahu, we computed the spatial covariance matrix during the summer season at 12:00 PM averaged over a window of 5 min. Then, the AUC and the KL divergence are computed for those tree structures that are generated using Markov Chain Monte-Carlo (MCMC) method. Figure 8 shows the distribution of those tree structures generated using MCMC method versus the KL divergence **(left)** and versus $\log_{10}(1 - \text{AUC})$ **(right)**.[11]

Looking back at Fig. 4, for the very small value of $1 - \text{AUC}$ the relationship between the KL divergence and the boundary of the possible feasible region for $-\log(1 - \text{AUC})$ is linear. This means that if the upper bound is tight then the relationship between the KL divergence and the $-\log(1 - \text{AUC})$ is almost linear. In Fig. 8, the maximum value of 1−AUC for this model is $< 10^{-3}$ which justifies why two distributions in Fig. 8 are scaled/mirrored of each other. Moreover, just by looking at the distribution of tree models in this example, it is obvious that most tree models have similar performance. Only a small portion of the tree models have better performance than the average tree models, but the difference is not that significant.

2) The Colorado dataset

From the solar data obtained from 6 sensors near Denver, Colorado, we computed the spatial covariance matrix during the summer season at 12:00 PM averaged over a window of 5 minutes. Then, the AUC and the KL divergence are computed for all possible tree structures. Figure 9 shows the distribution of all possible tree structures versus the KL divergence **(left)** and versus the AUC **(right)**.

In the Colorado dataset, there are two sensors that are very close to each other compared with the distance between all other pairs of sensors. As a result, if the particular edge between these two sensors is in the approximated tree structure we get a smaller AUC and KL divergence compared with when that particular edge is not in the tree

[10] See [37] for fields definition and other details about the normalization methods for the solar irradiation covariance matrix.

[11] In this example, since the AUC for all generated tree structures is close to one, we plot the distribution of generated trees versus $\log_{10}(1 - \text{AUC})$.

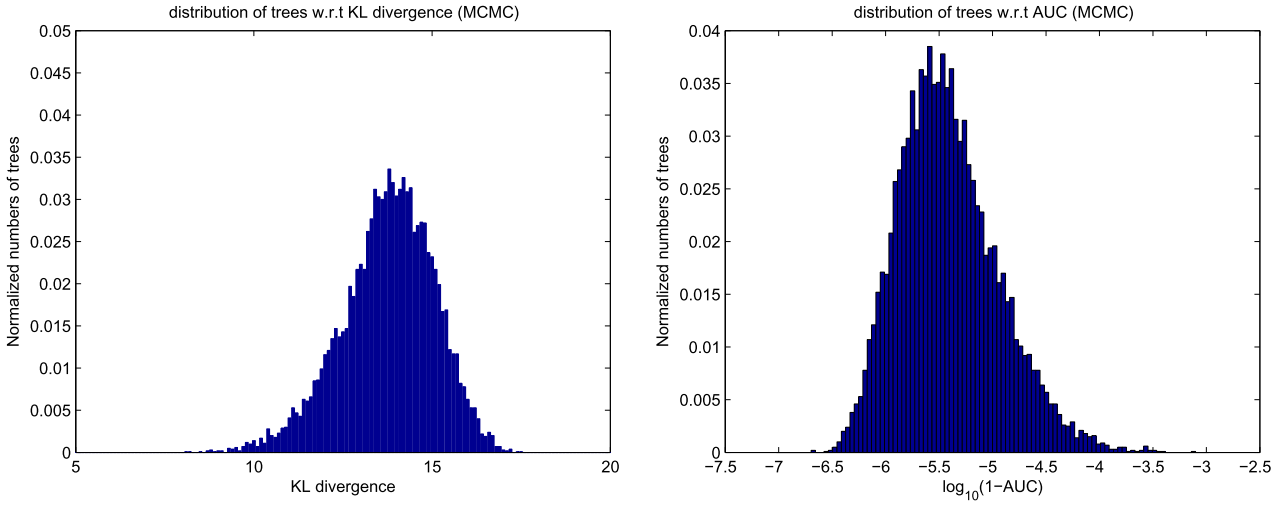

Fig. 8. Left: distribution of the generated trees (Normalized histogram) using MCMC versus the KL divergence and **Right:** distribution of the generated trees (Normalized histogram) using MCMC versus $\log_{10}(1-\text{AUC})$ for the Oahu solar measurement grid dataset in summer season at 12:00 PM.

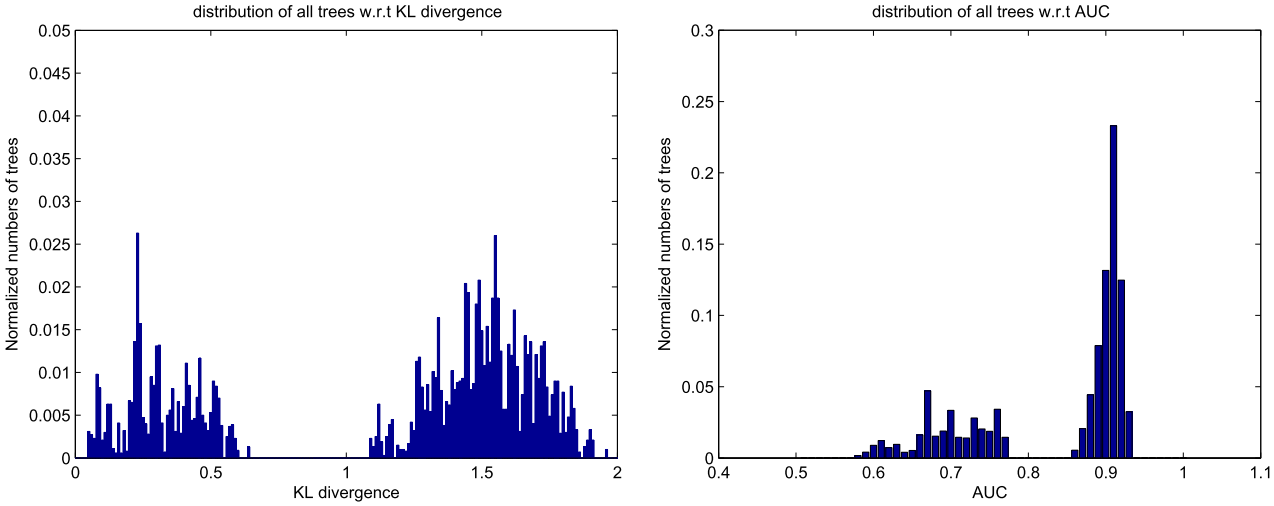

Fig. 9. Left: distribution of all trees (Normalized histogram) versus the KL divergence and **Right:** distribution of all trees (Normalized histogram) versus the AUC for the Colorado dataset in summer season at 12:00 PM.

structure. This explains why the distribution of all trees, in this case, looks like a mixture of two distributions. This result also gives us valuable insight on how to answer the following question, "How to construct informative approximation algorithms for model selection in general." This is an example where almost all trees that contain the particular edge between the two aforementioned sensors are good approximations while the rest of the tree models' give poor performance.

3) Two-dimensional sensor network

In this example, we create a 2D sensor network using a Gaussian kernel [39] as follows

$$\Sigma_{\underline{X}}(i,j) = \left[e^{-(d(i,j)^2)/(2\sigma^2)}\right]$$

where $d(i,j)$ is the Euclidean distance between the i-th sensor and the j-th sensor in the 2D space. All sensors are located randomly in 2D space.[12] We set $\sigma = 1$ and generate a 2D sensor network with 20 sensors. For the 2D sensor network example, Fig. 10 shows the distribution of the generated tree structures using MCMC method versus KL divergence **(left)** and versus $\log_{10}(1-\text{AUC})$ **(right)**. Again we see the mirroring effect in Fig. 10 as we have an almost linear relationship between the KL divergence and $-\log(1-\text{AUC})$. Note that, the covariance matrix generated has one dominant eigenvalue in most cases. Furthermore, Fig. 11 plots 1−AUC as well as its analytical upper bound and lower bound versus the dimension of the graph, n for $\sigma = 1.3$ **(left)** and $\sigma = 1.8$ **(right)**. To generate this figure, we randomly generated 1000 sensor networks and then plot the averaged AUC. As we can see in this figure, the 1−AUC and its bounds decay exponentially which is consistent with the theoretical results of this paper.

[12] Sensors location in each dimension are drawn randomly from a Normal distribution.

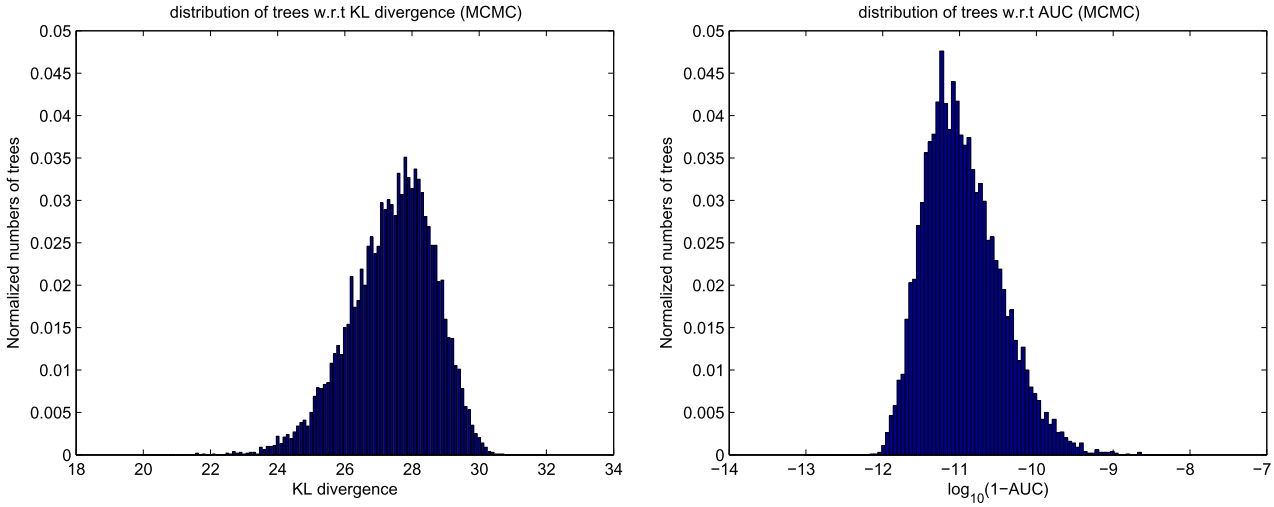

Fig. 10. Left: Distribution of the generated trees (Normalized histogram) using MCMC versus the KL divergence and **Right:** distribution of the generated trees (Normalized histogram) using MCMC versus $\log_{10}(1 - \text{AUC})$ for the 2D sensor network example with 20 sensors and $\sigma = 1$.

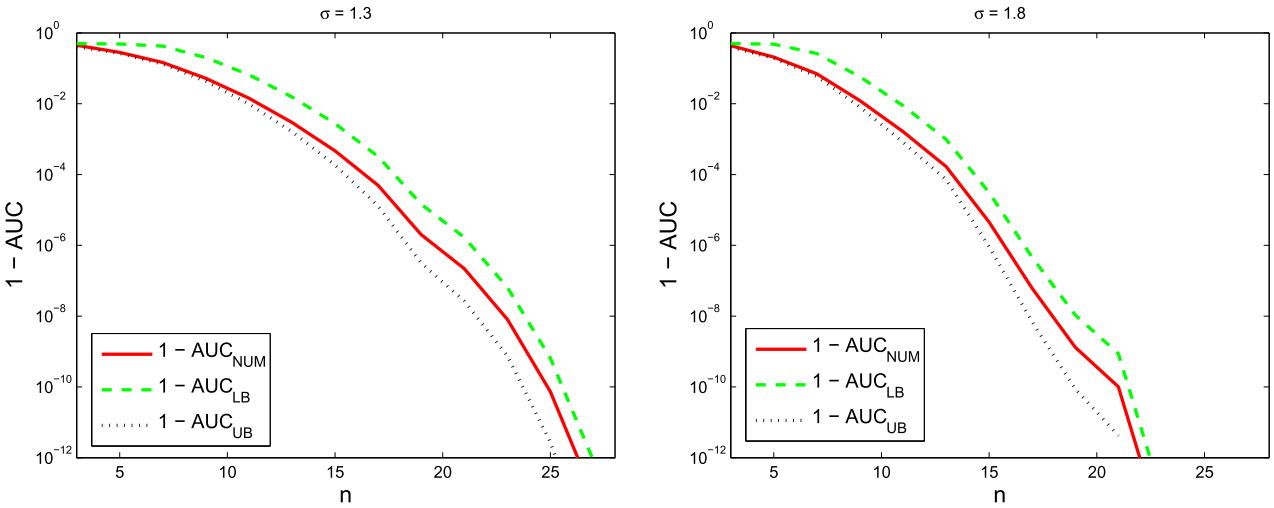

Fig. 11. 1−AUC and its bounds versus the dimension of the graph, n for $\sigma = 1.3$ **(left)** and $\sigma = 1.8$ **(right)**, averaged over 1000 runs of sensor networks generated randomly.

VI. CONCLUSION

In this paper, we formulate a detection problem and investigate the quality of the model selection. More specifically, we consider Gaussian distributions and discuss the covariance selection quality of a given model. We present the CAM and show its relationship with information theory divergences such as the KL divergence, the reverse KL divergence, and the Jeffreys divergence as well as the ROC curve and the area under it, i.e. the AUC, as a measure of accuracy in the detection problem framework. This paper also presents an analytical expression for the AUC that can be efficiently evaluated numerically. AUC analytical lower and upper bounds are also provided. We show that the AUC and the lower bound for the AUC depends on the eigenvalues of the CAM. Upper bounds for the AUC are obtained from finding a parametric relationship between the AUC and the KL/reverse KL divergences. We pick the tree structure as an example of an approximation model and use the Chow-Liu MST algorithm to compute the maximum likelihood tree structure approximation. Then, the quality of the Chow-Liu MST tree algorithm is investigated using the formulated detection problem. Through some examples, we show that in general, the tree approximation is not a good model as the number of nodes in the graphical model increases which is the case in high-dimensional problems such as modeling the electrical distribution grid using smart grid sensor measurements and distributed renewable energy sources. The aforementioned result is also consistent with the analytical results provided in this paper that is 1−AUC decays exponentially as the dimension of the graph increases.

The detection framework presented in this paper can be generalized for non-Gaussian models. The AUC analytical bounds obtained in this paper can also be used in other applications that are using AUC as a relevant criterion. One example is in medicine when the AUC is used for diagnostic tests between positive instance and negative instance [40] where instead of changing the coordinates we can look at the exponent of the AUC bounds. In ongoing work, we are looking at more accurate graphical approximations that involve

non-tree graphs. These approximations use a variation of the CAM which we call the symmetric CAM and simple linear transformations.

ACKNOWLEDGEMENT

Portions of this paper without the theoretical justifications and detailed discussions of the detection problem was presented at the 2016 Information Theory and Application Workshop [26]. Authors would like to thank Prof. Peter Harremoës for his helpful discussions on information divergences and assistance with Theorem 3.

FINANCIAL SUPPORT

This work was supported in part by NSF grant ECCS-1310634, the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370, and the University of Hawaii REIS project.

STATEMENT OF INTEREST

None.

REFERENCES

[1] Shuman, D.I.; Narang, S.K.; Frossard, P.; Ortega, A.; Vandergheynst, P.: The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *Signal Process. Mag., IEEE*, **30** (3) (2013), 83–98.

[2] Koller, D.; Friedman, N.: Probabilistic Graphical Models: Principles; Techniques, The MIT Press, Cambridge, MA, 2009.

[3] Jordan, M.I.: Learning in Graphical Models, vol. **89**, Springer Science & Business Media, 1998.

[4] Dempster, A.P.: Covariance selection. *Biometrics*, **28** (1) (1972), 157–175.

[5] Lauritzen, S.L.: Graphical Models, Oxford University Press, 1996.

[6] Huang, J.Z.; Liu, N.; Pourahmadi, M.; Liu, L.: Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, **93** (1) (2006), 85–98.

[7] Kschischang, F.R.; Frey, B.J.; Loeliger, H.-A.: Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory*, **47** (2) (2001), 498–519.

[8] Loeliger, H.A.; Dauwels, J.; Hu, J.; Korl, S.; Ping, L.; Kschischang, F.R.: The factor graph approach to model-based signal processing. *Proc. IEEE.*, **95** (6) (2007), 1295–1322.

[9] Chow, C.K.; Liu, C.N.: Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, **IT-14** (1968), 462–467.

[10] Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, **7** (1) (1956), 48–50.

[11] Prim, R.C.: Shortest connection networks and some generalizations. *Bell Syst. Tech. J.*, **36** (6) (1957), 1389–1401.

[12] Meinshausen, N.; Buhlmann, P.: Model selection through sparse maximum likelihood estimation. *Ann. Stat.*, **9** (2006), 1436–1464.

[13] Friedman, J.; Hastie, T.; Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9** (3) (2008), 432–441.

[14] Khajavi, N.T.; Kuh, A.: First order markov chain approximation of microgrid renewable generators covariance matrix, in *Proc. of IEEE Int. Symp. on Information Theory, Istanbul, Turkey (ISIT'13)*, July 2013, 1207–1211.

[15] Khajavi, N.T.: Latent tree approximation in linear model, in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE Int. Conf.on*. IEEE, 2017, 5940–5944.

[16] Kadane, J.B.; Lazar, N.A.: Methods and criteria for model selection. *J. Am. Stat. Assoc.*, **99** (465) (2004), 279–290.

[17] MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms, vol. **7**, Citeseer, Cambridge University Press, Cambridge, UK, 2003.

[18] Jeffreys, H.: An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, **186** (1007) (1946), 453–461.

[19] Lewis-II, P.M.: Approximating probability distributions to reduce storage requirements. *Inf. Control*, **2** (3) (1959), 214–225.

[20] Lehmann, E.L.; Romano, J.P.: Testing Statistical Hypotheses, Springer, 2006.

[21] Neyman, J.; Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20** (1928), 263–294.

[22] Eguchi, S.; Copas, J.: Interpreting kullback–leibler divergence with the neyman–pearson lemma. *J. Multivar. Anal.*, **97** (9) (2006), 2034–2040.

[23] Scharf, L.L.: Statistical Signal Processing, vol. **98**, Addison-Wesley Reading, MA, 1991.

[24] Shiryaev, A.N.: Probability, volume 95 of graduate texts in mathematics. 1996.

[25] Hanley, J.A.; McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143** (1) (1982), 29–36.

[26] Khajavi, N.T.; Kuh, A.: The quality of tree approximation from auc bounds. *Information Theory and Applications Workshop*, 2016.

[27] Provost, S.B.; Rudiuk, E.M.: The exact distribution of indefinite quadratic forms in noncentral normal vectors. *Ann. Inst. Stat. Math.*, **48** (2) (1996), 381–394.

[28] Ha, H.T.; Provost, S.B.: An accurate approximation to the distribution of a linear combination of non-central chi-square random variables. *REVSTAT–Stat. J.*, **11** (3) (2013), 231–254.

[29] Al-Naffouri, T.Y.; Hassibi, B.: On the distribution of indefinite quadratic forms in Gaussian random variables, in *Information Theory, 2009. ISIT 2009. IEEE Int. Symp. on*. IEEE, 2009, 1744–1748.

[30] Kotz, S.; Kozubowski, T.; Podgorski, K.: The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance, Springer Science & Business Media (2012).

[31] Abramowitz, M.; Stegun, A.I.: Handbook of mathematical functions. *Applied Mathematics Series*, **55** (1966), 62.

[32] Cover, T.M.; Thomas, J.A.: Elements of Information Theory, John Wiley & Sons, Hoboken, NJ, 2012.

[33] Kullback, S.: Information Theory; Statistics, Courier Corporation, Mineola, NY, 1968.

[34] Harremoës, P.; Vajda, I.: On pairs of f-divergences and their joint range. *arXiv preprint arXiv:1007.0097*, 2010.

[35] Kavcic, A.; Moura, J.M.F.: Matrices with banded inverses: Inversion algorithms and factorization of gauss-markov processes. *IEEE Trans. Inf. Theory.*, **46** (2000), 1495–1509.

[36] Khajavi, N.T.; Kuh, A.: The covariance selection quality for graphs with junction trees through auc bounds, in *Proc. of the Asia-Pacific Signal and Information Processing Association (APSIPA ASC)*, December 2016, 1–5.

[37] Khajavi, N.T.; Kuh, A.; Santhanam, N.P.: Spatial correlations for solar pv generation and its tree approximation analysis, in *Proc. of the Asia-Pacific Signal and Information Processing Association (APSIPA ASC)*, December 2014, 1–5.

[38] National Renewable Energy Laboratory, Measurement and instrumentation data center. [Online]. Available at http://www.nrel.gov/midc/.

[39] Rasmussen, C.E.; Williams, C.K.I.: Gaussian Processes for Machine Learning, The MIT Press, Cambridge, MA, 2006.

[40] Johnson, N.P.: Advantages to transforming the receiver operating characteristic (roc) curve into likelihood ratio co-ordinates. *Stat. Med.*, **23** (14) (2004), 2257–2266.

[41] Flanders, H.: Differentiation under the integral sign. *Am. Math. Mon.*, **80** (6) (1973), 615–627.

[42] Luenberger, D.G.: Optimization by Vector Space Methods, John Wiley and Sons, Inc., 1997.

APPENDIX

Proof of Lemma 1

The calculus-based proof for the special case of continuous PDFs is as follow. We can apply the Leibniz integral rule [41] and compute the derivative of CDFs $P_0(l)$ and $P_1(l)$ as

$$f_{L_0}(l) = -\frac{dP_0(l)}{dl}$$

and

$$f_{L_1}(l) = -\frac{dP_1(l)}{dl}$$

since $f_{L_0}(l)$ and $f_{L_1}(l)$ are continuous functions.[13] We have

$$\begin{aligned}\mathcal{D}\left(f_{L_0}(l)||f_{L_1}(l)\right) &= \int_{-\infty}^{+\infty} \log\frac{f_{L_0}(l)}{f_{L_1}(l)}\, f_{L_0}(l)\, dl \\ &\overset{(a)}{=} -\int_0^1 \log\frac{dP_1}{dP_0}\, dP_0 \\ &\overset{(b)}{=} -\int_0^1 \log h'(z)\, dz\end{aligned}$$

where equality (a) is true since we can replace PDFs $f_{L_0}(l)$ and $f_{L_1}(l)$ using the derivative of their CDFs.

[13] Both $f_{L_0}(l)$ and $f_{L_1}(l)$ are PDFs in generalized Chi-squared distributions class. This means that each of these PDFs are convolution of weighted Chi-squared distributions. Weighted Chi-squared distribution is continuous in its domain thus, convolution of these distributions is continuous in its domain.

Equality (b) is just a change of variable, $z = P_0(l)$, in order to write the integral in terms of the derivative of the ROC curve. Proof for the second part of this lemma is similar to the proof of the first part.

Proof of Theorem 3

Looking back at the properties of the ROC curve, $h(z)$, where $z \in [0, 1]$, the ROC curve has to satisfy the following conditions

- **C1:** $\int_0^1 h'(z)dz = 1$
- **C2:** $h'(z) \geq 0$
- **C3:** $h'(z)$ is decreasing

where $h'(z)$ is the derivative of the ROC curve, $h(z)$. Also for a given ROC curve, $h(z)$, we can compute the AUC as

$$\Pr(L_\Delta > 0) = \int_0^1 h(z)dz.$$

Then, using integration by parts, we can show that

$$1 - \Pr(L_\Delta > 0) = \int_0^1 z\, h'(z)dz.$$

To compute the possible feasible region stated in the Theorem 3, we need to optimize both of following KL divergences, $\mathcal{D}(f_{L_1}(l)||f_{L_0}(l))$ and $\mathcal{D}(f_{L_0}(l)||f_{L_1}(l))$, with respect to the derivative of the ROC curve given a fixed AUC, $\Pr(L_\Delta > 0)$, while conditions, C1, C2, and C3 hold. To solve this optimization, we can use the method of Lagrange multiplier.

First step: Here we minimize $\mathcal{D}(f_{L_1}(l)||f_{L_0}(l))$ with respect to the derivative of the ROC curve given the constraints. Optimization problem is as follow

$$\begin{aligned}&\text{argmin}_{h'(z)} -\int_0^1 \log h'(z)dz \qquad \text{(A.1)}\\ &\text{s. t.} \int_0^1 z\, h'(z)dz = 1 - \Pr(L_\Delta > 0)\\ &\qquad\qquad \text{C1, C2 \& C3.}\end{aligned}$$

To solve this optimization problem, we first write the Lagrangian. We need two coefficients a and b corresponding to conditions in optimization problem (A.1). Then, we can write the Lagrange multiplier as a function of the derivative of the ROC

curve, z, a and b as follow

$$\begin{aligned} &L(h'(z), z, a, b) \\ &\quad = -\int_0^1 \log h'(z) dz \\ &\qquad + a\left(\int_0^1 z h'(z) dz - (1 - \Pr(L_\Delta > 0))\right) \\ &\qquad + b\left(\int_0^1 h'(z) dz - 1\right). \end{aligned}$$

Note that, the Lagrangian, $L(h'(z),\ z,\ a,\ b)$ is a convex functional [42] of $h'(z)$. Thus, we can compute its minimum by taking its derivative with respect to $h'(z)$. Doing so, and applying the Euler-Lagrange equation [42] we get

$$\begin{aligned} \frac{\delta L(h'(z), z, a, b)}{\delta h'(z)} &= \frac{\partial L}{\partial h'} - \frac{d}{dz}\frac{\partial L}{\partial h''} \\ &= \int_0^1 \left(az + b - \frac{1}{h'(z)}\right) dz. \end{aligned}$$

Set $\frac{\delta L(h'(z), z, a, b)}{\delta h'(z)} = 0$, we get

$$h'(z) = \frac{1}{az + b}$$

for all $z \in [0,\ 1]$. From C3, since $h'(z)$ is decreasing, we can conclude that $a > 0$. Moreover, from C1, at optimum we have $\int_0^1 h'(z)dz = 1$ and thus, we can compute one of the coefficients as $b = \frac{a}{e^a - 1}$.

Computing the AUC integral and the KL divergence using the ROC curve, we get the following parametric boundary for the possible feasible region

$$\Pr(L_\Delta > 0) = \frac{1}{1 - e^{-a}} - \frac{1}{a} \qquad \text{(A.2)}$$

and

$$\mathcal{D} = \log(a) + \frac{a}{e^a - 1} - 1 - \log(1 - e^{-a}) \quad \text{(A.3)}$$

where $\mathcal{D} = \mathcal{D}\left(f_{L_1}(l) || f_{L_0}(l)\right)$.

Second step: Here we minimize $\mathcal{D}\left(f_{L_0}(l) || f_{L_1}(l)\right)$. The Lagrange multiplier for this step is similar to the first step but it is more straightforward if we define $g(\eta) = h^{-1}(\eta)$. Note that using integration by parts, we can show that AUC is

$$\Pr(L_\Delta > 0) = \int_0^1 \eta\, g'(\eta) d\eta.$$

Now, we can write the Lagrangian for the optimization problem with respect to $g'(\eta)$. The Lagrangian is convex with respect to $g'(\eta)$, thus taking the derivative and set it equal to zero as follow

$$\frac{\delta L(g'(\eta), \eta, a, b)}{\delta g'(\eta)} = 0$$

we can compute the parametric boundary for the possible feasible region. The parametric boundary, in this case, is the same as a solution in (A.2) and (A.3) with $\mathcal{D} = \mathcal{D}(f_{L_0}(l) || f_{L_1}(l))$. Thus, combining these two steps, for the optimal boundary, we have

$$\mathcal{D}_l^* = \min\{\mathcal{D}(f_{L_1}(l) || f_{L_0}(l)),\ \mathcal{D}(f_{L_0}(l) || f_{L_1}(l))\}.$$

Navid Tafaghodi Khajavi received B.S. degree in Electrical Engineering at Ferdowsi University, an M.S. degree in Electrical Engineering from Shahid Beheshti University, and a Ph.D. degree in Electrical Engineering from University of Hawaii at Manoa. Dr. Tafaghodi Khajavi previously worked at Shahid Beheshti University Cognitive Radio Laboratory and University of Hawaii Big Data Laboratory. He is currently working at Ford Motor Company. Dr. Tafaghodi Khajavi's research is in the area of statistical modeling and machine learning, probabilistic graphical models and big data with applications to real world problems.

Anthony Kuh received his B.S. degree in Electrical Engineering and Computer Science at the University of California, Berkeley in 1979, an M.S. degree in Electrical Engineering from Stanford University in 1980, and a Ph.D. degree in Electrical Engineering from Princeton University in 1987. Dr. Kuh previously worked at AT&T Bell Laboratories and has been on the faculty in Electrical Engineering Department at the University of Hawaii since 1986. He is currently a Professor and previously served as Department Chair. Since January, 2017 Dr. Kuh has been a program director at the National Science Foundation (NSF) in the Electrical Communications, and Cyber Systems Division. Dr. Kuh's research is in the area of neural networks and machine learning, adaptive signal processing, sensor networks, and renewable energy and smart grid applications. Dr. Kuh won an NSF Presidential Young Investigator Award, received a Distinguished Fulbright Scholar's Award working at Imperial College, London, and is an IEEE Fellow. He is currently Vice President of Technical Activities for the Asia Pacific Signal and Information Processing Association (APSIPA), is an associate editor of the APSIPA Transactions on Signal and Information Processing, and served as general co-chair of the APSIPA ASC 2018 held in Honolulu, Hawaii.