Latent Tree Approximation in Linear Model

Navid Tafaghodi Khajavi

Dept. of Electrical Engineering, University of Hawaii, Honolulu, HI 96822
Email: navidt@hawaii.edu

Abstract—We consider the problem of learning underlying tree structure from noisy, mixed data obtained from a linear model. To achieve this, we use the expectation maximization algorithm combined with Chow-Liu minimum spanning tree algorithm. This algorithm is sub-optimal, but has low complexity and is applicable to model selection problems through any linear model.

Index Terms—Tree Approximation, EM Algorithm.

I. INTRODUCTION

This paper considers learning in graphical models where we approximate a graphical model with a tree. The tree approximations are made as it is much simpler performing inference and estimation on trees rather than graphs that have cycles or loops. An example is applying the Gaussian belief propagation (GBP) algorithm [1]. When you have statistical knowledge of the graphical structure (i.e. correlation matrix), then you can use the Kullback Liebler (KL) divergence measure and find the optimal tree using the Chow-Liu algorithm which finds the minimal spanning tree (MST) [2]. There are other methods discussed in literature where instead of minimizing KL divergence, a detection framework is proposed in order to quantify the tree approximation quality [3] and [4]. Moreover, a similar method with a more generalized model is proposed in [5]. This paper considers the case when you do not have direct statistical information about the graph. Here we assume you have an underdetermined and noisy linear model that gives access to samples of data and we develop an algorithm using the EM algorithm and the Chow Liu algorithm to come up with a tree approximation by minimizing the KL divergence.

Given a set of data from a jointly Gaussian random vector, the conventional tree approximation algorithms' goal is to approximate a tree structured distribution for the underlying distribution. A first approach, is the well-known minimum spanning tree (MST) problem solved initially by Chow-Liu [2] which finds the covariance matrix coefficients using the covariance selection method [6]. This algorithm is originally proposed to find the optimal tree structure that connects all the graph nodes and minimizes the KL divergence [7] between the original and the approximate distributions. In contrast, in many applications such as the smart grid application [8], we want to approximate structure (tree or chain) for a sub-graph of system graphical representation from the observation without the knowledge of the latent variable covariance matrix while considering the effects of the whole system, i.e. we want to learn a tree structure for the latent vector covariance matrix in an underdetermined model.

In this paper, we investigate the problem of approximating a tree structured covariance matrix of the latent variable in an underdetermined linear model by constructing an iterative algorithm based on the expectation maximization (EM) algorithm. We also suggest to learn the maximum number of the EM algorithm iterations to avoid overfitting. Simulation results are performed on solar irradiation data from the island of Oahu, Hawaii [9]. As a result we can see the performance improvement using the proposed EM MST algorithm.

The rest of this paper is organized as follows. The tree approximation problem as well as the linear mixing model are stated in section II while the proposed EM MST solution is presented in section III. Section IV presents some simulation on real spatial solar irradiation data and discusses the goodness of proposed solution. Finally, Section V summarizes results of this paper as well as discussion on future works.

Notation: Upper case and lower case letters denote random variables and their realizations, respectively; underlined letters stand for vectors; boldface upper case letters denote matrices; $(\cdot)^T$ and $\mathrm{E}(\cdot)$ stand for transposition and expectation.

II. PROBLEM STATEMENT

We want to approximate a multivariate distribution by the product of lower order component distributions [10]. For the purpose of tree approximation, the maximum order of these lower order distributions is two, i.e. no more than pairs of variables. Let $\underline{X} \sim \mathcal{N}(\underline{0}, \Sigma)$ has a zero-mean Gaussian distribution where Σ is the positive definite covariance matrix and $\underline{X} \in \mathbb{R}^p$ has the graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where sets $\mathcal{V}$ and $\mathcal{E}$ are the set of all vertices and edges of the graph representing of $\underline{X}$. Let $\underline{X}_{\mathcal{T}} \sim \mathcal{N}(\underline{0}, \widetilde{\Sigma})$ has the graph representation $\mathcal{G}_{\mathcal{T}} = (\mathcal{V}, \mathcal{E}_{\mathcal{T}})$ where $\mathcal{E}_{\mathcal{T}} \subseteq \mathcal{E}$ is a set of edges that represents a tree structure and $\widetilde{\Sigma}$ is the positive definite covariance matrix of the tree structure. Distribution of $\underline{X}_{\mathcal{T}}$ (the product of lower order component distributions) is defined as [2]

$$f_{\underline{X}_{\mathcal{T}}}(\underline{x}_{\mathcal{T}}) = \prod_{(u,v) \in \mathcal{E}_{\mathcal{T}}} \frac{f_{\underline{X}^u, \underline{X}^v}(\underline{x}^u, \underline{x}^v)}{f_{\underline{X}^u}(\underline{x}^u) f_{\underline{X}^v}(\underline{x}^v)} \prod_{o \in \mathcal{V}} f_{\underline{X}^o}(\underline{x}^o) \quad (1)$$

where $\underline{X}^u$ and $\underline{x}^u$ are uth element of random vector, $\underline{X}$, and its realization vector, $\underline{x}$, respectively. Let $\underline{Y} \in \mathbb{R}^m$ be the observation vector where $m < p$. Consider the problem of approximating a tree structured distribution of the latent vector, $\underline{X}$, (similar to (1)) in the following linear model[1]:

$$\underline{Y} = \mathbf{H}\underline{X} + \underline{W} \quad (2)$$

where $\mathbf{H}$ is an $m \times p$ characteristic matrix with rank m and $\underline{W}$ is a Gaussian zero-mean random vector with positive definite

---

[1]Without loss of generality, we may assume that $\underline{X}$ and $\underline{Y}$ are zero-mean.

covariance matrix $\mathbf{D}$. Note that, in the linear model (2), the latent vector, $\underline{X}$, and noisy measurements, $\underline{W}$, are independent and Gaussian, thus the observation vector, $\underline{Y}$, has Gaussian distribution. We assume that the covariance matrix of noisy measurements, $\mathbf{D}$, and the characteristic matrix $\mathbf{H}$ are known completely but we may only have *partial prior information*[2] about the latent vector $\underline{X}$.

We seek to approximate the underlying distribution of the latent vector $\underline{X}$ such that it has a tree structure representation graph. In other words, instead of estimating the distribution of the original latent vector $\underline{X}$ and then approximate its tree structure, we seek to directly estimate the distribution of the tree structured latent vector $\underline{X}_{\mathcal{T}}$ using a set of observed data obtained from the linear model (2).

III. TREE APPROXIMATION ALGORITHMS

In the sequel, we first consider the case where the complete knowledge of the distribution of the original latent vector $\underline{X}$, i.e. the covariance matrix Σ of the the zero-mean Gaussian distribution of $\underline{X}$, is known. The optimal tree approximation in this case can be computed using the well-known Chow-Liu tree approximation algorithm [2]. This algorithm can also be applied to the estimated covariance matrix from the set of data observed directly form the latent vector $\underline{X}$. Then, we investigate the case where there is no information about the covariance of the latent variable. In this case the tree approximation algorithm should be performed using the observation data obtained based on the model (2). Since the linear model (2) is under-determined, i.e. $m < p$, there are infinitely many solutions. We will show that by applying the expectation maximization (EM) algorithm, one can approximate a tree structured covariance matrix for the latent vector in this scenario. The problem with this approach is that since the EM algorithm converging to local optima and there are infinitely many solutions, the EM algorithm may converge to a wrong local optima. To deal with this problem, we assumed *partial prior information* is available to help the EM algorithm to converge to the desired optima.

A. Chow-Liu Minimum Spanning Tree for Gaussian distributions

Chow-Liu MST method [2] initially proposed for approximating the joint distribution of discrete variables by product of lower order distributions similar to (1) which involves no more than the pair of variables and it can easily generalized for approximating the joint distribution of Gaussian variables.

Definition 1. *Let $\mathcal{T}^{\star}$ denote the set of all positive definite covariance matrices that their graphical models have tree structures and obey the product rule given in (1), i.e. set of all possible $\widetilde{\Sigma}$'s that obey (1).* ∎

In other words, for all $\widetilde{\Sigma} \in \mathcal{T}^{\star}$, the corresponding zero-mean Gaussian distribution have the same marginals over the graph $\mathcal{G}_{\mathcal{T}}$ as the zero-mean Gaussian distribution with covariance matrix Σ. Note $\mathcal{T}^{\star}$ has finite cardinality of $O(p^{p-1})$.

[2]For instance, very noisy or outdated data directly obtained about $\underline{X}$.

The KL divergence is proposed in [2] to quantify the distance between any distribution and its tree structure approximation. It is shown in [2] that the optimal solution for this problem can be found efficiently using greedy algorithms [11].

Proposition 1. *The approximated tree structure for a zero-mean Gaussian distribution, $f_{\widetilde{\Sigma}}(\underline{x})$ where $\widetilde{\Sigma} \in \mathcal{T}^{\star}$, follows (1) and thus the KL divergence can be simplified as follows:*

$$\mathcal{D}(f_{\Sigma}(\underline{x}) || f_{\widetilde{\Sigma}}(\underline{x})) = -\frac{1}{2}\log\ (|\Sigma\widetilde{\Sigma}^{-1}|).$$

Proof. Derivation is based on the properties of (1) and the KL divergence definition for Gaussian distributions [6]. ∎

1) Chow-Liu MST for a given distribution: Let $f_{\Sigma}(\underline{x}) = \mathcal{N}(\underline{0}, \Sigma)$ be given. Then the Chow-Liu MST distribution is a zero-mean Gaussian distribution with covariance matrix, $\widetilde{\Sigma}^{\star} \in \mathcal{T}^{\star}$, that minimize the KL divergence as follows:

$$\widetilde{\Sigma}^{\star} = \underset{\widetilde{\Sigma} \in \mathcal{T}^{\star}}{\arg\min}\ \mathcal{D}(f_{\Sigma}(\underline{x}) || f_{\widetilde{\Sigma}}(\underline{x})). \qquad (3)$$

where $\mathcal{D}^{\star} \triangleq -\frac{1}{2}\log\ (|\Sigma\widetilde{\Sigma}^{\star-1}|)$ is the minimum and quantifies the distance between the given distribution and its optimal tree approximation. Given the knowledge of Σ, Chow-Liu algorithm can efficiently compute the optimal solution, $\widetilde{\Sigma}^{\star}$. Note that, finding the optimal solution $\widetilde{\Sigma}^{\star}$ requires knowledge of the covariance matrix, Σ.

Assumption: In the rest of this paper, we assume that exact knowledge of the covariance matrix, Σ, is *not available*. We may only have some noisy or mixed observations. The main goal of this paper is to approximate the tree structured covariance matrix, $\widetilde{\Sigma}$, using those noisy or mixed observations.

Definition 2. *Let $\mathcal{T}$ denote the set of all positive definite covariance matrices that have a tree structure.* ∎

Note that, $\mathcal{T}^{\star} \subset \mathcal{T}$ and its cardinality is infinite. In general, without exact knowledge of covariance matrix, Σ, it is not possible to find the optimal solution in (3) over the set $\mathcal{T}$.

B. EM Algorithm to Approximate MST

Consider the linear model presented in (2). Let $\mathcal{Y} : \{\underline{y}_1, \ldots, \underline{y}_R\}$ be a set of R iid observation drawn from this model. Unlike previously mentioned models, one cannot uniquely compute an estimate of the covariance matrix, Σ, since this model is underdetermined. In this section, instead of minimizing an optimization problem similar to (3), we suggest to minimize the KL divergence between the empirical distribution of observation obtained from (2), i.e. set $\mathcal{Y}$, and distribution of random vector $\underline{Y}_{\mathcal{T}}$, where the latent vector $\underline{X}$ in (2) is replaced with the tree structured one, $\underline{X}_{\mathcal{T}}$.

The empirical distribution of observation, $f_{\mathbf{S}_Y}(\underline{y}) \sim \mathcal{N}(\underline{0}, \mathbf{S}_Y)$ is a zero-mean Gaussian distribution where $\mathbf{S}_Y = \frac{1}{R}\sum_{r=1}^{R}(\underline{y}_r - \bar{\underline{y}})(\underline{y}_r - \bar{\underline{y}})^T$ where $\bar{\underline{y}} = \frac{1}{R}\sum_{r=1}^{R}\underline{y}_r$. The random vector $\underline{Y}_{\mathcal{T}}$ distribution is also zero-mean Gaussian distribution with covariance matrix $\mathbf{H}\widetilde{\Sigma}\mathbf{H}^T$, i.e. $f_{\widetilde{\Sigma}}(\underline{y}) \sim \mathcal{N}(\underline{0}, \mathbf{H}\widetilde{\Sigma}\mathbf{H}^T)$. We minimize the KL divergence between these two distributions, in order to estimate the distribution of tree structured

latent variable, $\underline{X}_\mathcal{T}$. The optimization problem is as follows:

$$\widetilde{\boldsymbol{\Sigma}}^* = \underset{\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}}{\arg\min} \ \mathcal{D}\left(f_{\mathbf{S}_Y}(\underline{y}) || f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{y})\right). \tag{4}$$

Unlike previous cases, the Chow-Liu MST algorithm cannot find the optimal solution over set $\mathcal{T}$. Basically, there are infinitely many solutions for the aforementioned optimization. In what follows, we find a reasonable solution by implementing a heuristic algorithm which is a combination of the EM algorithm with the Chow-Liu MST algorithm.

The minimization problem in (4) is identical to the following optimization problem which maximizes the average log-likelihood function where expectation is over the empirical distribution of observation, $f_{\mathbf{S}_Y}(\underline{y})$[3]:

$$\widetilde{\boldsymbol{\Sigma}}^* = \underset{\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}}{\arg\max} \, \mathrm{E}_{\mathbf{S}_Y}\left[\log\left(f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{Y})\right)\right].$$

Given the data set $\mathcal{Y}$, we can write the average log-likelihood maximization as follows:

$$\widetilde{\boldsymbol{\Sigma}}^* = \underset{\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}}{\arg\max} \sum_{\underline{y} \in \mathcal{Y}} \left[\log\left(f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{y})\right)\right].$$

Applying Jensen's inequality [7], we have:

$$\begin{aligned}
\sum_{\underline{y} \in \mathcal{Y}} \left[\log\left(f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{y})\right)\right] &= \sum_{\underline{y} \in \mathcal{Y}} \left[\log\left(\int_{\underline{X}} f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{x}, \underline{y}) d\underline{x}\right)\right] \\
&= \sum_{\underline{y} \in \mathcal{Y}} \left[\log\left(\int_{\underline{X}} g(\underline{x}|\underline{y}) \frac{f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{x}, \underline{y})}{g(\underline{x}|\underline{y})} d\underline{x}\right)\right] \\
&\geq \sum_{\underline{y} \in \mathcal{Y}} \left[\int_{\underline{X}} g(\underline{x}|\underline{y}) \log\left(\frac{f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{x}, \underline{y})}{g(\underline{x}|\underline{y})} d\underline{x}\right)\right] \\
&= \mathcal{F}\left(g(\underline{x}|\underline{y}), \widetilde{\boldsymbol{\Sigma}}\right) \tag{5}
\end{aligned}$$

where $f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{x}, \underline{y})$ is the joint distribution between random vectors $\underline{X}$ and $\underline{Y}$ and obviously depends on the approximation covariance matrix, $\widetilde{\boldsymbol{\Sigma}}$. Also, $g(\underline{x}|\underline{y})$ is an auxiliary conditional distribution and $\mathcal{F}(.,.)$ is a function of the data set, $\mathcal{Y}$, the auxiliary distribution, $g(\underline{x}|\underline{y})$ and the approximation covariance matrix, $\widetilde{\boldsymbol{\Sigma}}$.

Definition 3. *Let $\mathcal{G}$ denote the set of all conditional Gaussian distributions, i.e. all possible $g(\underline{x}|\underline{y})$ distributions.* ∎

Using the EM algorithm, we can iteratively tighten the bound in (5) by iteratively solving for $g(\underline{x}|\underline{y})$ given $\widetilde{\boldsymbol{\Sigma}}$ and vice versa. EM algorithm has two step at each iteration, the E-step and the M-step. The E-step of the $(l+1)$-th iteration of the EM algorithm is:

E-Step: computing $g(\underline{x}|\underline{y}) \in \mathcal{G}$ for all $\underline{y} \in \mathcal{Y}$:

$$\begin{aligned}
g^{l+1}(\underline{x}|\underline{y}) &= \underset{g(\underline{x}|\underline{y}) \in \mathcal{G}}{\arg\max} \ \mathcal{F}(g(\underline{x}|\underline{y}), \widetilde{\boldsymbol{\Sigma}}^l) \\
&= \underset{g(\underline{x}|\underline{y}) \in \mathcal{G}}{\arg\min} \sum_{\underline{y} \in \mathcal{Y}} \mathcal{D}(g(\underline{x}|\underline{y}) || f_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{x}|\underline{y})) \\
&= f_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{x}|\underline{y}), \quad \forall \underline{y} \in \mathcal{Y},
\end{aligned}$$

where the tree structured covariance matrix $\widetilde{\boldsymbol{\Sigma}}^l \in \mathcal{T}$ is the l-th step solution of the EM algorithm. The solution to the E-step is $g^{l+1}(\underline{x}|\underline{y}) = f_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{x}|\underline{y})$ for all $\underline{y} \in \mathcal{Y}$. The M-step of the $(l+1)$-th iteration of the EM algorithm is:

M-Step: tightening the bound in (5) by maximizing the lower bound for all $\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}$:

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}^{l+1} &= \underset{\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}}{\arg\max} \ \mathcal{F}\left(g^{l+1}(\underline{x}|\underline{y}), \widetilde{\boldsymbol{\Sigma}}\right) \\
&= \underset{\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}}{\arg\max} \sum_{\underline{y} \in \mathcal{Y}} \left[\int_{\underline{X}} g^{l+1}(\underline{x}|\underline{y}) \log\left(f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{x}, \underline{y})\right) d\underline{x}\right]
\end{aligned}$$

Plug in the solution of the E-Step into the M-Step optimization problem, we get:

$$\widetilde{\boldsymbol{\Sigma}}^{l+1} = \underset{\widetilde{\boldsymbol{\Sigma}} \in \mathcal{T}}{\arg\max} \sum_{\underline{y} \in \mathcal{Y}} \left[\int_{\underline{X}} f_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{x}|\underline{y}) \log\left(f_{\widetilde{\boldsymbol{\Sigma}}}(\underline{x}, \underline{y})\right) d\underline{x}\right] \tag{6}$$

Optimization problem (6) have to be solved at each iteration of the EM algorithm.

Theorem 1. *Solution to the $(l+1)$-th iteration of the EM algorithm (optimization problem (6)) is given by:*

$$\widetilde{\boldsymbol{\Sigma}}^{l+1} = \texttt{chow-liu}(\boldsymbol{\Omega}^l)$$

where $\boldsymbol{\Omega}^l = \frac{1}{|\mathcal{Y}|} \sum_{\underline{y} \in \mathcal{Y}} [\mathrm{E}_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{X}\underline{X}^T | \underline{Y} = \underline{y})]$ where the expectation is over the conditional distribution, $f_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{x}|\underline{y})$. For the linear model in (2), we compute $\boldsymbol{\Omega}^l$ as

$$\boldsymbol{\Omega}^l = \mathbf{C}^l + \mathbf{C}^l \mathbf{H}^T \mathbf{D}^{-1} \left(\frac{1}{|\mathcal{Y}|} \sum_{\underline{y} \in \mathcal{Y}} \underline{y}\underline{y}^T\right) \mathbf{D}^{-1} \mathbf{H} \mathbf{C}^l$$

where $\mathbf{C}^l = \left((\widetilde{\boldsymbol{\Sigma}}^l)^{-1} + \mathbf{H}^T \mathbf{D}^{-1} \mathbf{H}\right)^{-1}$. [4] ∎

Having the solution of optimization problem (6) in hand, the EM based algorithm is as follow:

EM MST Approximation Algorithm

- Initialization Step [$l = 0$]:
    - $\boldsymbol{\Omega}^0 = \boldsymbol{\Sigma}_0$
- Continue updating [$(l+1)$-th Step]:
    - $\widetilde{\boldsymbol{\Sigma}}^{l+1} = \texttt{chow-liu}(\boldsymbol{\Omega}^l)$
    - $\boldsymbol{\Omega}^{l+1} = \frac{1}{|\mathcal{Y}|} \sum_{\underline{y} \in \mathcal{Y}} [\mathrm{E}_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{X}\underline{X}^T | \underline{Y} = \underline{y})]$
- Stopping criterion: $\mathcal{D}(f_{\widetilde{\boldsymbol{\Sigma}}^l}(\underline{x}) || f_{\widetilde{\boldsymbol{\Sigma}}^{l+1}}(\underline{x})) < \epsilon$ [5]
- Output (approximated tree structure covariance matrix): $\widetilde{\boldsymbol{\Sigma}}^*_{EM} = \widetilde{\boldsymbol{\Sigma}}^{l+1}$ for some l satisfying the stopping criterion

---

[3]Entropy of random variable $\underline{Y}$ with distribution $f_{\mathbf{S}_Y}(\underline{y})$ is eliminated here, since it does not depend on the optimization parameter, $\widetilde{\boldsymbol{\Sigma}}$.

[4]Proof is eliminated due to space limitation.

[5]ϵ is some small number and have been set to avoid over-fitting.

Remark: The symmetric, positive definite matrix, $\mathbf{\Sigma}_0$, is the initialization step in the EM MST approximation algorithm. One can choose any symmetric, positive definite matrix as an initialization step for this algorithm. However, EM based Algorithms converge to local minimum. Thus, having some prior knowledge will help the convergence of the algorithm. We will discuss how to pick the matrix $\mathbf{\Sigma}_0$ in section IV. The covariance matrix, $\widetilde{\mathbf{\Sigma}}_t^*$, is the approximated tree structure solution and is an upper bound for the optimal solution of (4), i.e., $\mathcal{D}(f_{\mathbf{S}_Y}(\underline{y})\|f_{\widetilde{\mathbf{\Sigma}}_{EM}^*}(\underline{y})) \geq \mathcal{D}(f_{\mathbf{S}_Y}(\underline{y})\|f_{\widetilde{\mathbf{\Sigma}}^*}(\underline{y}))$.

Remark: The EM MST algorithm solution converges to one of the infinitely many solutions for the optimization problem in (4). But, the ultimate goal is to approximate a tree structured covariance matrix for the latent variable $\underline{X}$ such that minimizes (3). Since the distribution of $\underline{X}$ is not available, we suggest to minimize (4) which is a heuristic optimization problem based on the available set of observation from the underdetermined system. Running the EM algorithm to find the exact solution for (4) results in finding an overfitted solution for (3). To avoid overfitting, we need to learn the maximum number of iterations for the EM MST algorithm for each particular problem.

IV. SIMULATION RESULTS AND DISCUSSION

In this section, we perform some simulations on real solar irradiation data obtain from [9] to validate our results. We take a year long data for 17 horizontal sensors of the Oahu solar measurement grid sites [12] from $9 : 00$ to $9 : 30$ AM to compute two normalized spatial covariance matrices[6]. We use data of Jan., 1st, 2011 to compute a covariance matrix, $\mathbf{\Sigma}$, for testing the algorithm. To run the algorithm we need some *partial prior information*. In general, we initialized EM MST algorithm with different priors. For solar irradiation problem we could use a forecasting data or very noisy and outdated data or even without any irradiation data and by only knowing the geographical position of sensors, we can come up with an approximation for the covariance matrix only base on the location and use it as *partial prior information*. Here, we use data from Apr., 1st, 2010 to Nov., 30th, 2010 to compute another covariance matrix, $\mathbf{\Sigma}_0$, which can be used as prior to the EM MST algorithm. In the linear model (2), we assume vector $\underline{W}$ is the vector of iid, white Gaussian noise. We generate set $\mathcal{Y}$ with 100 samples with high signal-to-noise-ratio (20 dB) comparing to priors.

Figure 1 indicates the performance of the EM MST approximation algorithm by computing the KL divergence between the approximated tree and the original graph. We randomly generate 1000 different characteristic matrices, $\mathbf{H}$, and plot the average performance. It is clear in the figure that the proposed EM MST approximated tree achieves better performance comparing to the Chow-Liu optimal tree for the prior graph. The only problem is over-fitting as it is stated on this figure. The over-fitting occurs, since we are minimizing the KL divergence computed based on the observation vector

---

[6]See [12] for information about data and preprocessing methods involving computation of spatial covariances
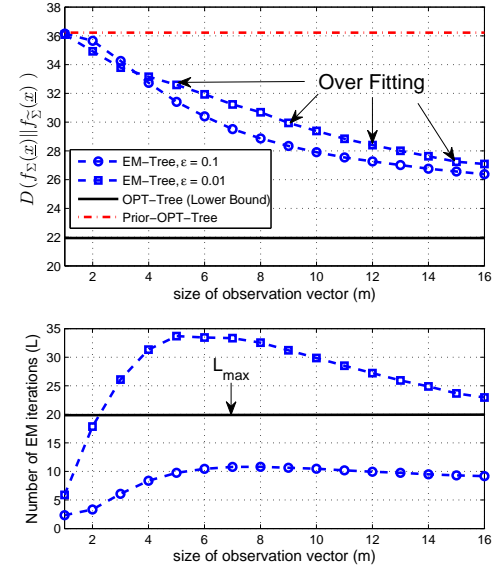

Fig. 1. **Above:** The KL divergence between the approximated tree and the original graph vs. m averaged over 1000 randomly generated characteristic matrices, $\mathbf{H}$, for the EM MST approximated tree with two different stopping criteria, $\epsilon = 0.1$ and $\epsilon = 0.01$, the optimal Chow-liu tree (lower bound) and the optimal tree for prior. **Below:** Average number of EM iterations vs. the number of sensors m.

distribution, $f_{\mathbf{S}_Y}(\underline{y})$, instead of the latent vector distribution, $f_{\mathbf{\Sigma}}(\underline{x})$ which is *unavailable*. To overcome this, we suggest to learn the maximum number of iterations, L_{max}, based on the training data and run the EM algorithm for at most L_{max} iterations. For the set up in this paper we set $L_{max} = 20$.

V. CONCLUSION AND FUTURE DIRECTIONS

We present an iterative algorithm to learn the MST in an underdetermined linear model. The proposed algorithm is a combination of the EM algorithm and the Chow-Liu algorithm. We show that the algorithm needs only to run for a limited number of iterations and each iterations can be computed efficiently using the Chow-Liu algorithm. In future works, we will do a more comprehensive simulation study along with theoretical proof of the EM algorithm convergence. we will also look at the maximum a posteriori solution to the problem as well as learning highly correlated connections in high dimensional data, "Big Data".

REFERENCES

[1] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, June 2007.
[2] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.
[3] N. T. Khajavi and A. Kuh, "The quality of the covariance selection through detection problem and AUC bounds," *arXiv preprint arXiv:1605.05776*, 2016.

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

[5] N. T. Khajavi and A. Kuh, "The goodness of covariance selection problem from AUC bounds," in *Proc. 54th Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2016.

[6] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, March 1972.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[8] N. T. Khajavi and A. Kuh, "First order markov chain approximation of microgrid renewable generators covariance matrix," in *Proc. IEEE International Symposium on Information Theory, Istanbul, Turkey (ISIT' 13)*, July 2013, pp. 1207–1211.

[9] NREL. Solar irradiance website. [Online]. Available: http://www.nrel.gov/midc/, ," .

[10] P. M. Lewis II, "Approximating probability distributions to reduce storage requirements," *Information and control*, vol. 2, no. 3, pp. 214–225, 1959.

[11] Joseph B Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.

[12] N. T. Khajavi, A. Kuh, and N. P. Santhanam, "Spatial correlations for solar PV generation and its tree approximation analysis," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2014, pp. 1–5.