VARIATIONAL BAYES LEARNING OF MULTISCALE GRAPHICAL MODELS

Hang Yu and Justin Dauwels

School of Electrical and Electronics Engineering
Nanyang Technological University, Singapore, 639798

ABSTRACT

Multiscale (multiresolution) graphical models have gained widespread popularity in recent years, since they enjoy rich modeling power as well as efficient inference procedures. Existing approaches to learning multiscale graphical models often leverage the framework of penalized likelihood, and therefore suffer from the issue of regularization selection. In this paper, we propose a novel method to learn multiscale graphical models from the Bayesian perspective. More specifically, the regularization parameters are treated as random variables that follow Gamma distributions. We then derive an efficient variational Bayes algorithm to learn the model, and further demonstrate the advantages of the proposed method through numerical experiments.

Index Terms— Multiscale (multiresolution) models, graphical models, variational Bayes, regularization selection

1. INTRODUCTION

Sparse graphical models allow efficient representation of complex systems with few parameters by learning or imposing a sparse dependency structure. Moreover, the structure can in turn be utilized to derive highly efficient inference algorithms. Thus, graphical models have found applications and permeated the literature in a wide variety of domains, including signal processing [1], image processing and computer vision [2], computational biology and neuroscience [3, 4], geophysics and earth science [5, 6], and sociology [7].

One typically constructs a sparse graphical model by discovering the most important interactions between observed variables [8], as shown in Fig. 1a. We refer to this graphical model as a monoscale graphical model in this context, as opposed to the multiscale models to be introduced. Sparse monoscale graphical models, however, fail to capture the long-range correlations between far-away sites in a spatial domain (e.g., the dependency between nodes X_1 and X_P in Fig. 1a). To overcome this limitation, multiscale graphical models [9] introduce coarser scales to model complex dependency with few parameters. A good example is to model the sea surface temperature using a quadtree (see Fig. 1b), a typical family of multiscale models [10]. We associate the finest scale nodes with the original temperature measuring stations. Under this arrangement, the root node can be regarded as the average global temperature while its children capture the deviations from this mean, and so forth at increasingly finer scales. The long-range statistical dependencies (e.g., between the Arctic Ocean and the Southern Ocean) can be captured easily at coarser scales. Unfortunately, the multi-scale tree model induces blocky artifacts, since some spatially adjacent variables are widely separated in the tree structure. A more satisfying approach is to consider inscale structure [11], which captures short-range correlation in each scale of different resolution as shown in Fig. 1c. Such models are equipped with strong modeling power and efficient inference algorithms [11, 12]

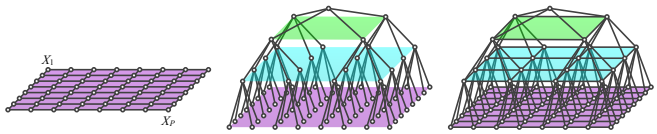

Fig. 1: A graphical models with multiscale structure: (a) a monoscale graphical model; (b) a latent tree (i.e., a quadtree); (c) a multiscale graphical model.

Next, we give a brief review on the learning algorithms of multiscale graphical models. A multiscale graphical model with fixed pyramidal structure (cf. Fig. 1c) is proposed in [11]. One smoothness parameter is introduced to describe the dependence between variables and is inferred via expectation maximization. Although this model is quite useful to estimate a smooth surface given noisy and incomplete observations, the fixed inscale structure seriously limits the modeling power. Choi *et al.* provide a recipe for the problem of fixed structure in [13], learning the inscale structure of all scales via maximum entropy relaxation. Along this line, the authors further develop a multiscale model with sparse inscale conditional covariance [12], in which the inscale structure is inferred individually for each scale, from the coarsest to the finest. Unfortunately, instead of maximizing the likelihood, a pseudo likelihood is maximized comprising of conditional densities of variables in one scale conditioned on other scales. Therefore, the results might be misleading. Another problem with [13, 12] is the selection of regularization parameters (i.e., tolerance parameters in [13, 12]). Theses parameters balance the trade-off between model fitting and the sparsity of the inscale structure, and hence greatly affect the numerical performance and the model interpretation. Standard methods for

regularization selection, such as cross validation, Akaike information criterion, and Bayesian information criterion, are shown to overfit the data for high dimensional problems, leading to dense graphs [14]. As an alternative, stability based methods [15, 16] can reliably infer the graphical model structure by selecting a "stable" graph from bootstrapped sample sets. However, such methods are quite computationally demanding, since the learning algorithm needs to be run on each sample set for every possible combination of regularization parameters. Thus, the computationally complexity grows exponentially with the number of regularization parameters, and these methods are not applicable to the case where each scale in the multiscale graphical model is associated with different regularization parameters [12].

To address the above mentioned problems, we present a novel Bayesian formulation for learning multiscale graphical models. In particular, we focus on the case where variables in all scales are jointly Gaussian distributed. The resulting Gaussian graphical model is directly defined by its precision matrix (inverse covariance matrix), that is, a zero element corresponds to the absence of an edge in the graphical model as well as the conditional independence between two variables. We then impose shrinkage priors on the elements in the precision matrix corresponding to the inscale connections, and further impose Gamma priors on the hyper parameters (i.e., regularization parameters). The parameters of the resulting model is inferred using a variational Bayes (VB) algorithm. Numerical results demonstrate that the proposed approach can select the proper amount of regularization in an automated manner, and therefore provides a good fit with few parameters.

The paper is organized as follows. In Section 2, we briefly introduce multiscale graphical models. We then derive the VB algorithm in Section 3. We present experimental results in Section 4 and offer concluding remarks in Section 5.

2. MULTISCALE GRAPHICAL MODELS

As shown in Fig. 1c, the multiscale graphical models we consider in this paper have fixed tree-structured connections between different scales and sparse inscale structure at each scale [13]. Since all the variables are jointly Gaussian distributed, such connections are characterized by the precision matrix. We assume that the number of scales is M, and we denote the scale with the coarsest resolution as scale 1, and increase the scale index as the resolution increases. Note that variables at scale m are only conditionally dependent on variables at scale $m-1$ and $m+1$. As such, the joint precision matrix of all variables is a block tridiagonal matrix. In the case of the four-scale model in Fig. 1c (M=4), the precision matrix can be partitioned as:

$$K = \begin{bmatrix} K_{[1]} & K_{[1,2]} & & \\ K_{[2,1]} & K_{[2]} & K_{[2,3]} & \\ & K_{[3,2]} & K_{[3]} & K_{[3,4]} \\ & & K_{[4,3]} & K_{[4]} \end{bmatrix}, \tag{1}$$

where $K_{[m]}$ denotes the conditional precision matrix at scale m, and $K_{[m,m+1]}$ represents the tree connections between scale m and $m+1$. Since we can only observe samples of variables at the finest (bottom) scale, the variables at coarser scales are treated as hidden variables. As such, the joint precision matrix can be equivalently partitioned as:

$$K = \begin{bmatrix} K_C & K_{CF} \\ K_{FC} & K_F \end{bmatrix}, \tag{2}$$

where K_C and K_F denote the conditional precision matrices respectively of variables $\boldsymbol{x}_C$ at coarser scales and variables $\boldsymbol{x}_F$ at the finest scale. Note that K_F is $K_{[4]}$ in the above example. Without loss of generality, we assume the mean vector of all variables to be zero. The resulting joint density can be written as:

$$p(\boldsymbol{x}|K) \propto \exp\Big(-\frac{1}{2}\boldsymbol{x}^T K \boldsymbol{x}\Big), \tag{3}$$

where $\boldsymbol{x} = [\boldsymbol{x}_C^T, \boldsymbol{x}_F^T]^T$.

Given N observations of $\boldsymbol{x}_F$ at the finest scale, our objective is to infer the joint precision matrix K. More concretely, we aim to learn the sparse inscale conditional precision matrix $K_{[m]}$ at each scale as well as the parameters of the tree K_T that connects different scales. To this end, we propose a novel Bayesian formulation of the problem as follows. Since sparse inscale conditional precision matrices are favored, we associate the off-diagonal elements K_{ij} of $K_{[m]}$ with Gaussian priors with zero means and precisions λ_{ij}, i.e.,

$$p(K_{ij}|\lambda_{ij}) \propto \sqrt{\lambda_{ij}} \exp\Big(-\frac{1}{2}\lambda_{ij}K_{ij}^2\Big), \tag{4}$$

for all $i > j$ and $(i,j) \in K_I$, where K_I denotes the off-diagonal elements of the inscale conditional precision matrices at all scales. As shown in our numerical experiments, many of the precisions λ_{ij} will take very large values during the learning process, and consequently, the prior can successfully shrink most elements of K_I to zero, thus yielding sparse inscale structure for all scales. We further impose conjugate Gamma hyperprior on the precisions λ_{ij}:

$$p(\lambda_{ij}) = \text{Gamma}(\lambda_{ij}; a, b) \propto \lambda_{ij}^{a-1} \exp(-b\lambda_{ij}). \tag{5}$$

The parameters a and b are set to small values (e.g., 10^{-100}) to obtain a flat non-informative prior. Note that

$$\int p(K_{ij}|\lambda_{ij})p(\lambda_{ij})d\lambda_{ij} = \frac{\Gamma\big(a+\frac{1}{2}\big)}{\Gamma(a)\sqrt{2\pi b}}\left(\frac{1}{1+\frac{1}{2b}K_{ij}^2}\right)^{a+\frac{1}{2}},$$

which is a t distribution. Therefore, we essentially put a t prior on K_{ij}. Such shrinkage prior is often used in the Bayesian framework to promote sparsity [17, 18]. Note that in the literature of learning graphical models [8, 12], Laplace priors are often used since they amount to ℓ_1 norm penalties on the precision matrix and the resulting optimization problem is convex. Although Laplace priors can also be regarded as a scale mixture of Gaussian, the hyperprior on precisions λ_{ij} is the inverse Gamma distribution that is not conjugate to the Gaussian distributions parameterized by precisions [19]. As a result, we employ t prior here since it is more tractable for Bayesian inference.

Altogether, the overall joint model can be expressed as:

$$p(\boldsymbol{x}, K, \Lambda) = \prod_{n=1}^{N} p(\boldsymbol{x}_F^{(n)}, \boldsymbol{x}_C^{(n)}|K) \cdot \prod_{(i,j)\in K_I, i>j} p(K_{ij}|\lambda_{ij})p(\lambda_{ij}), \quad (6)$$

where Λ is a matrix with the same size as K whose (i,j) entry equals λ_{ij} for $(i,j) \in K_I$ and $i > j$, and equals 0 otherwise.

3. VARIATIONAL BAYES INFERENCE

In this section, we devise a VB algorithm to estimate the joint precision matrix K. Specifically, we aim to find a variational distribution $q(\boldsymbol{x}_C, K, \Lambda)$ to approximate the intractable posterior $p(\boldsymbol{x}_C, K, \Lambda|\boldsymbol{x}_F)$ by minimizing the KL divergence between them as measured by $\mathrm{KL}(q|p) = \int q \log(q/p)$. Here, we apply the mean-field approximation, and therefore, the variational distribution can be factorized as:

$$q(\boldsymbol{x}_C, K, \Lambda) = \prod_{n=1}^{N} q(\boldsymbol{x}_C^{(n)}) \prod_{i=1}^{P} \delta(K_{i:P,i} - K^*_{i:P,i}) \cdot \prod_{(i,j)\in K_I, i>j} q(\lambda_{ij}), \quad (7)$$

where $\delta(K_{i:P,i} - K^*_{i:P,i})$ is a delta function which equals 1 when $K_{i:P,i} = K^*_{i:P,i}$ and 0 otherwise, $K_{i:P,i}$ denotes the ith to Pth elements in the ith column of K, and P is the dimension of K. We follow [19, 20] to use delta functions as the variational distributions of elements in K for the sake of convenience. Furthermore, as the algorithm proceeds, many of the precisions λ_{ij} will become very large, and then delta functions can well approximate the true posterior distribution.

The VB update rules can then be derived as follows. For hidden variables $\boldsymbol{x}_C$,

$$q(\boldsymbol{x}_C^{(n)}) \propto \exp\left\{\langle \log p(\boldsymbol{x}_F^{(n)}, \boldsymbol{x}_C^{(n)}|K)\rangle_{\delta(K-K^*)}\right\} \quad (8)$$

$$= p(\boldsymbol{x}_C|\boldsymbol{x}_F; K^*) \quad (9)$$

$$= \mathcal{N}\big(\boldsymbol{x}_C^{(n)}; \mu_{C|F}^{(n)}, \Sigma_{C|F}\big), \quad (10)$$

where the conditional mean $\mu_{C|F}^{(n)} = (K_C^*)^{-1}K_{CF}^*\boldsymbol{x}_F^{(n)}$ and the conditional covariance $\Sigma_{C|F} = (K_C^*)^{-1}$. For precision matrix K, we generate point estimates of $K_{i:P,i}$ for $i = 1, \cdots, P$ sequentially. Equating the corresponding gradient to zero yields:

$$K^*_{i+1:P,i} = -\big\{NS_{ii}[(K^*_{\neg i,\neg i})^{-1}]_{i:P-1,i:P-1} + \mathrm{diag}(\langle\Lambda_{i+1:P,i}\rangle)\big\}^{-1}\big\{NS_{i+1:P,i} + NS_{ii}K^*_{i,1:i-1}[(K^*_{\neg i,\neg i})^{-1}]_{1:i-1,i:P-1}\big\}, \quad (11)$$

$$K^*_{ii} = S_{ii}^{-1} + (K^*_{\neg i,i})^T(K^*_{\neg i,\neg i})^{-1}K^*_{\neg i,i}. \quad (12)$$

In the above expression, $\mathrm{diag}(\boldsymbol{a})$ denotes a matrix with the vector $\boldsymbol{a}$ on the diagonal, $\langle\Lambda_{i+1:P,i}\rangle$ denotes the expectation of $q(\Lambda_{i+1:P,i})$, and $\neg i$ represents all the indices except i. The matrix S can be partitioned in the same way as K in (1), and the submatrices of S can be expressed as in Eq. (13), where S_F is the empirical covariance of variables in the finest scale. Note that we only need to update the elements in K corresponding to the inscale and interscale connections; other elements are set to zero all the time. Furthermore, we can easily prove the positive definiteness of K^* in each iteration following the proof in [19]. In addition, we emphasize that there is no guarantee that the KL divergence is decreased every time we update K^*, due to the non-conjugacy between $p(\boldsymbol{x}|K)$ and $p(K_{ij}|\lambda_{ij})$ [21]. However, as shown in [21], such problem can be fixed by damping. As a result, we employ the damping method with a damping factor $\rho = 0.9$, and find that it successfully ameliorates the monotonicity problem of KL divergence in our experiments. More precisely, given K^* resulting from (11) and (12) as well as the precision matrix in the previous iteration $K^{*(\kappa-1)}$, the precision matrix in the current iteration can be computed as:

$$K^{*(\kappa)} = \rho K^{*(\kappa-1)} + (1-\rho)K^* \quad (14)$$

Finally, for $\lambda_{ij} \in K_I$,

$$q(\lambda_{ij}) \propto p(\lambda_{ij}) \exp\left\{\langle \log p(K_{ij}|\lambda_{ij})\rangle_{\delta(K_{ij}-K^*_{ij})}\right\} \quad (15)$$

$$= \mathrm{Gamma}\Big(\lambda_{ij}; a + \frac{1}{2}, b + (K^*_{ij})^2\Big). \quad (16)$$

The proposed algorithm are summarized in Table 1.

Table 1: VB learning of multiscale graphical models.

Compute empirical covariance of the finest scale S_F. Initialize K^* as a normalized Laplacian matrix of a graph with full inscale structures.

1. Compute $\Sigma_{C|F} = (K_C^*)^{-1}$.
2. Compute the matrix S as given in Eq. (13).
3. Compute $\Sigma = (K^*)^{-1}$. For $i = 1, \cdots, P$, update $K_{i:P,i}$ sequentially as follows:
   (a) Compute $(K^*_{\neg i,\neg i})^{-1} = \Sigma_{\neg i,\neg i} - \Sigma_{\neg i,i}\Sigma_{i,\neg i}/\Sigma_{ii}$.
   (b) Update $K^*_{i+1:P,i}$ using Eq. (11), and hence $K^*_{i,i+1:P} = (K^*_{i+1:P,i})^T$.
   (c) Update K_{ii} using Eq. (12).
   (d) Update the covariance Σ corresponding to the new K^* using Schur complement as follows:
   $$\Sigma_{ii} = S_{ii},$$
   $$\Sigma_{\neg i,i} = -(K^*_{\neg i,\neg i})^{-1}K^*_{\neg i,i}\Sigma_{ii}, \quad \Sigma_{i,\neg i} = \Sigma^T_{\neg i,i},$$
   $$\Sigma_{\neg i,\neg i} = (K^*_{\neg i,\neg i})^{-1} + \Sigma_{\neg i,i}\Sigma_{i,\neg i}/\Sigma_{ii}.$$
4. Compute K^* in the current iteration by damping:
   $$K^{*(\kappa)} = \rho K^{*(\kappa-1)} + (1-\rho)K^*.$$
5. Update $q(\lambda_{ij})$ following (16).

$$S_{[m_1,m_2]} = \begin{cases} [\Sigma_{C|F}]_{[m_1,m_2]} + [\Sigma_{C|F}]_{[m_1,M-1]}K^*_{[M-1,M]}S_FK^*_{[M,M-1]}[\Sigma_{C|F}]_{[M-1,m_2]}, & m_1, m_2 < M \\ -[\Sigma_{C|F}]_{[m_1,M-1]}K^*_{[M-1,M]}S_F, & m_1 < M, m_2 = M \\ S_F, & m_1 = m_2 = M \end{cases}, \quad (13)$$

Table 2: Modeling fractional Brownian motion data.

Criteria	Empirical	MonoGGM	TreeGGM	CO-MultiGGM	VB-MultiGGM
KL div	0	1.1480×10^3	89.0431	13.5270	3.6472
Prm No.	32896	23064	681	1733	1402

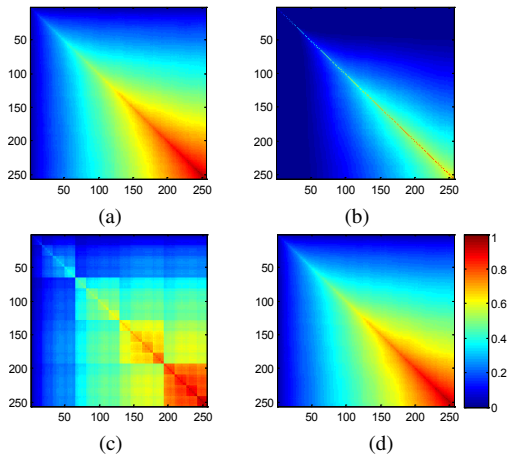

Fig. 2: Fractional Brownian motion data modeling: (a) empirical covariance and covariance resulting from (b) MonoGGM, (c) TreeGGM and (d) VB-MultiGGM.

4. NUMERICAL RESULTS

In this section, we benchmark the proposed method (referred to as VB-MultiGGM) with a model parameterized by the empirical covariance, a monoscale Gaussian graphical model (MonoGGM), a multiscale tree Gaussian graphical model (TreeGGM), and a multiscale Gaussian graphical model whose inscale structure is inferred by ℓ_1 norm penalized convex optimization (CO-MultiGGM). In particular, we follow [12] to learn the inscale structure at each scale individually, from the coarsest to the finest [12]. Moreover, we also utilize the method in [12] to determine the proper amount of regularization. Specifically, different regularization parameters λ_E and λ_N are used respectively to penalize the off-diagonal and diagonal elements in the inscale matrices. For coarser scales, $\lambda_E = \xi/4$ and $\lambda_N = 2\lambda_E$, where ξ is the largest value of the off-diagonal elements in the empirical inscale conditional precision matrix. For the finest scale, $\lambda_E = \xi/2$, and we choose λ_N such that the KL divergence between the empirical and the estimated covariance at the finest scale is minimized. It is shown empirically in [12] that this regularization selection method works well. We compare the four models by means of KL divergence (KL div) and number of parameters (Prm No.).

4.1. Fractional Brownian Motion

We consider fractional Brownian motion in this section, which is known to possess long-range dependence across time [12]. Concretely, the covariance matrix is given by $\Sigma(t_1, t_2) = 0.5(|t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H})$, with Hurst parameter $H = 0.3$ for $t_1, t_2 \in \{1/256, 2/256, \cdots, 1\}$. We next generate 6000 samples from the Gaussian distribution parameterized by this covariance matrix, and learn the five models given the data. The results are summarized in Table 2 and Fig. 2. Obviously, the proposed method outperforms other approaches; it yields a small KL divergence with few parameters. The CO-MultiGGM performs the second best. However, it introduces more parameters, whereas the resulting KL divergence is larger. By comparing these two approaches, we can find that inferring regularization parameters from data in the Bayesian framework helps not only improve the modeling fitting but reduce the number of parameters as well. On the other hand, as shown in Fig. 2b and Fig. 2c, the MonoGGM fails to capture the long-range dependence between variables, while the TreeGGM mistakenly induces the blocky artifacts. In contrast, by learning the inscale structure in an automated manner, the proposed method (cf. Fig. 2d) can both describe the long-range correlation and remove the blocky artifacts. Finally, we can see that the empirical covariance model fits the data the best, but the resulting number of parameters is much larger than that of the rest models. Such model is prohibitive for large-scale problems.

Table 3: Modeling data with polynomially decaying covariance.

Criteria	Empirical	MonoGGM	TreeGGM	CO-MultiGGM	VB-MultiGGM
KL div	0	13.2503	30.0939	11.2937	3.8719
Prm No.	32896	18474	681	1391	1377

4.2. Polynomially Decaying Covariance

We now focus on a set of 256 Gaussian distributed variables allocated spatially on a 16×16 lattice. The corresponding covariance matrix is defined as $\Sigma(s_1, s_2) = \sqrt{d(s_1, s_2)}$, where $d(s_1, s_2)$ is the spatial distance between node s_1 and s_2. Polynomially decaying covariance is typically found in models with long-range correlations [13], in contrast with exponentially decaying covariance associated with short-range correlation. We draw 6000 samples from this model and further learn the five models based on the samples. We list the results in Table 3. Once again, the proposed VB-MultiGGM achieves the best performance, indicating that the VB algorithm can automatically learn the sparse inscale structure so that the resulting multiscale graphical model can well capture the complex interactions between variables with few parameters.

5. CONCLUSION

In this paper, we construct multiscale Gaussian graphical models from a Bayesian perspective and further develop a novel VB algorithm to learn the model. The results show that the proposed method can reliably infer the inscale structure in an automated manner without the need for manually tuning any regularization parameters.

6. REFERENCES

[1] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, P. Li, and F. Kschischang, "The factor graph approach to model-based signal processing," *Proc. IEEE*, vol. 95, no. 6, pp. 1295-1322, 2007.

[2] M. Beal, N. Jojic, H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 828-836, 2003.

[3] H. Yu, J. Dauwels, and X. Wang, "Copula Gaussian Graphical Models with Hidden Variables,"in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 2177-2180, 2012.

[4] J. Dauwels, H. Yu, X. Wang, F. Vialatte, C. Latchoumane, J. Jeong, and A. Cichocki, "Inferring Brain Networks through Graphical Models with Hidden Variables", *Machine Learning and Interpretation in Neuroimaging, Lecture Notes in Computer Science, Springer*, pp. 194-201, 2012.

[5] H. Yu, J. Dauwels, X. Zhang, S. Xu, and W. I. T. Uy, "Copula Gaussian Multiscale Graphical Models with Application to Geophysical Modeling," in *Proc. 15th Int. Conf. Inf. Fusion*, pp. 1741-1748, 2012.

[6] H. Yu, J. Dauwels, and P. Johnathan, "Extreme-Value Graphical Models with Multiple Covariates," accepted by *IEEE Trans. Signal Process.*, 2014.

[7] J. Dauwels, H. Yu, S. Xu, and X. Wang, "Copula Gaussian Graphical Model for Discrete Data", in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 6283-6287, 2013.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432-441, 2008.

[9] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396-1458, Aug, 2002.

[10] A. T. Ihler, S. Kirshner, M. Ghil, A. W. Robertson, and P. Smyth, "Graphical Models for Statistical Inference and Data Assimilation," *Physica D*, vol. 230, pp. 72-87, 2007.

[11] M. J. Choi, V. Chandrasekaran, D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Multiscale stochastic modeling for tractable inference and data assimilation", *Comput. Methods Appl. Mech. Engrg.*, vol. 197, pp. 3492-3515, 2008.

[12] M. J. Choi, V. Chandrasekaran, and A. S. Willsky, "Gaussian Multiresolution Models: Exploring Sparse Markov and Covariance Structure," *IEEE trans. signal process.*, vol. 58, No. 3, pp. 1012-1024, 2010.

[13] M. J. Choi, V. Chandrasekaran, and A. S. Willsky, "Maximum entropy relaxation for mutiscale graphical model selection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1889-1892, 2008.

[14] H. Liu, K. Roeder, and L. Wasserman, "Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models," *Advances in Neural Information Processing Systems*, 2010.

[15] N. Meinshausen, P. Bühlmann, "Stability Selection," *J. Roy. Statist. Soc. B - Stat. Methodol.*, vol. 72, pp. 417-473, 2010.

[16] S. Li, L. Hsu, J. Peng and P. Wang, "Bootstrap inference for network construction with an application to a breast cancer microarray study," *Ann. Appl. Stat.*, vol. 7, no. 1, pp. 391-417, 2013.

[17] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian Methods for Low-Rank Matrix Estimation," *IEEE trans. Signal Process.*, vol. 60, no. 8, pp. 3964-3977, 2012.

[18] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211-244, 2001.

[19] M. Chen, H. Wang, X. Liao, and L. Carin, "Bayesian Learning of Sparse Gaussian Graphical Models", *Technical report*, 2012.

[20] B. M. Marlin, and K. P. Murphy, "Sparse Gaussian Graphical Models with Unknown Block Structure," in *Proc. 26th Int. Conf. Mach. Learn.*, pp. 705-712, 2009.

[21] D. A. Knowles, and T. P. Minka, "Non-conjugate Variational Message Passing for Multinomial and Binary Regression," *Advances Neural Inf. Process. Syst. (NIPS)*, vol. 24, pp. 1701-1709, 2011.