RC24854 (W0909-018) September 2, 2009
Computer Science

IBM Research Report

A MAP Approach to Learning Sparse Gaussian Markov Networks

N. Bani Asadi
Department of Electrical Engineering
Stanford University
Palo Alto, CA
USA

I. Rish, K. Scheinberg, D. Kanevsky, B. Ramabhadran
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

A MAP APPROACH TO LEARNING SPARSE GAUSSIAN MARKOV NETWORKS

N. Bani Asadi

Department of Electrical Engineering,
Stanford University,
Palo Alto, CA

I. Rish, K. Scheinberg, D. Kanevsky, B. Ramabhadran

IBM T. J. Watson Research Center,
1101 Kitchawan Rd., Yorktown Heights, NY

ABSTRACT

Recently proposed l_1-regularized maximum-likelihood optimization methods for learning sparse Markov networks result into convex problems that can be solved optimally and efficiently. However, the accuracy of such methods can be very sensitive to the choice of regularization parameter, and optimal selection of this parameter remains an open problem. Herein, we propose a maximum a posteriori probability (MAP) approach that investigates different priors on the regularization parameter and yields promising empirical results on both synthetic data and real-life application such as brain imaging data (fMRI).

Index Terms— Markov networks, sparse optimization, l_1-regularization, maximum a posteriory probability (MAP), fMRI data analysis

1. INTRODUCTION

In many applications of statistical learning the objective is not simply to construct an accurate predictive model but rather to discover meaningful interactions among the variables. This is particularly important in biological applications such as, for example, reverse-engineering of gene regulatory networks, or reconstruction of brain-activation patterns from functional MRI (fMRI) data. Probabilistic graphical models, such as Markov networks (or Markov Random Fields), provide a principled way of modeling multivariate data distributions that is both predictive and interpretable.

A standard approach to learning Markov network structure is to choose the simplest model, i.e. the sparsest network, that adequately explains the data. Formally, this leads to regularized maximum-likelihood problem with the penalty on the number of parameters, or l_0 norm, a generally intractable problem that was often solved approximately by greedy search [2]. Recently, even better approximation methods were suggested [4, 8, 9, 6, 1] that exploit sparsity-enforcing property of l_1-norm regularization and yield convex optimization problems that can be solved efficiently. However, those approaches are known to be sensitive to the choice of the regularization parameter, i.e. the weight on l_1-penalty, and to the best of our knowledge, selecting the optimal value of this parameter remains an open problem (discussed in the next section)[1].

In this paper, we focus on a maximum a posteriory probability (MAP) approach to selecting regularization parameter λ when learning the structure of a Markov network over Gaussian variables. We advocate using non-uniform prior on λ and present encouraging empirical results on both synthetic and real datasets. Our method compares favorably to existing approaches, often resulting into higher accuracy and a more balanced trade-off between false-positive and false-negative errors.

2. PROBLEM FORMULATION

Let $X = \{X_1, ..., X_p\}$ be a set of p random variables, and let $G = (V, E)$ be an undirected graphical model (Markov network) representing conditional independence structure of the joint distribution $P(X)$. The set of vertices $V = \{1, ..., p\}$ is in the one-to-one correspondence with the set X. The set of edges E contains the edge (i, j) if and only if X_i is conditionally dependent on X_j given all remaining variables; lack of edge between X_i and X_j means that the two variables are conditionally independent given all remaining variables. Let $\mathbf{x} = (x_1, ..., x_p)$ denote a random assignment to X. We will assume a multivariate Gaussian probability density $p(\mathbf{x}) = (2\pi)^{-p/2} \det(C)^{\frac{1}{2}} e^{-\frac{1}{2}\mathbf{x}^T C \mathbf{x}}$, where $C = \Sigma^{-1}$ is the inverse covariance matrix, and the variables are normalized to have zero mean. Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be a set of n i.i.d. samples from this distribution, and let $S = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i$ denote the empirical covariance matrix.

Missing edges in the above graphical model correspond to zero entries in the inverse covariance matrix C, and thus the problem of structure learning for the above probabilistic graphical model is equivalent to the problem of learning the zero-pattern of the inverse-covariance matrix[2]. A common approach is to use l_1-regularization that is known

[1] The difficulty of selecting the regularization parameter has also motivated alternative approaches [3] that avoid the l_1-regularized maximum-likelihood formulation of [6].

[2] Note that the inverse of the *empirical* covariance matrix, even if it exists, does not typically contain exact zeros. Therefore, an explicit sparsity constraint is usually added to the estimation process.

to promote sparse solutions. From the Bayesian point of view, this is equivalent to assuming that the parameters of the inverse covariance matrix $C = \Sigma^{-1}$ are independent random variables C_{ij} following the Laplace distributions $p(C_{ij}) = \frac{\lambda_{ij}}{2} e^{-\lambda_{ij}|C_{ij}-\alpha_{ij}|}$ with zero *location parameters* (means) α_{ij} and equal *scale parameters* $\lambda_{ij} = \lambda$. Then $p(C) = \prod_{i=1}^{p} \prod_{j=1}^{p} p(C_{ij}) = (\lambda/2)^{p^2} e^{-\lambda||C||_1}$, where $||C||_1 = \sum_{ij} |C_{ij}|$ is the (vector) l_1-norm of C.

A common approach to recovering a sparse inverse covariance matrix is to assume a *fixed* parameter λ and find $\arg\max_{C \succ 0} p(C|\mathbf{X})$, where $\mathbf{X}$ is the $n \times p$ data matrix, or equivalently, since $p(C|\mathbf{X}) = P(\mathbf{X}, C)/p(\mathbf{X})$ and $p(\mathbf{X})$ does not include C, to find $\arg\max_{C \succ 0} P(\mathbf{X}, C)$, over positive definite matrices C. This yields the following optimization problem considered in [6, 1, 9]:

$$\max_{C \succ 0} \ln \det(C) - \mathrm{tr}(SC) - \lambda||C||_1 \quad (1)$$

where $\det(A)$ and $\mathrm{tr}(A)$ denote the determinant and the trace (the sum of the diagonal elements) of a matrix A, respectively.

The regularization parameter λ controls the number of non-zero elements (the sparsity) of solutions. The advantage of the above approach is that the problem in eq. 1 is convex, its optimal solution is unique [6] and can be found efficiently using recently proposed methods such as COVSEL [6] or *glasso* [1].

However, there is a known issue with the above approach. As we (and others) observe empirically, the accuracy of the Markov network reconstruction can be very sensitive to the choice of the regularization parameter λ, and there is no known method for optimal selection of λ. The two most commonly used approaches are (1) cross-validation and (2) theoretical derivations. However, λ selected by cross-validation, i.e. the estimate of the *prediction-oracle solution* that maximizes the test data likelihood (i.e. minimizes the predictive risk) is typically too small and yields high false-positive rate[3]. Existing theoretical derivations for λ have their own drawbacks: although [4] show that consistent estimate of the structure is possible, they admit that "such asymptotic considerations give little advice on how to choose a specific penalty parameter for a given problem". As a proxy for "best" λ, [4] in their neighborhood-selection approach provide λ that allows a consistent recovery of sparse structure of the *covariance* rather than the *inverse covariance* matrix, i.e. the recovery of marginal independencies between i-th and j-th variables, rather than conditional independencies given the rest of the variables. Similar approach is used in [6] to derive λ for the optimization problem in eq. 1. In practice, however, this may result into very high values of λ that effectively ignore almost all dependencies (i.e., have high false-negative rate); this was acknowledged by the authors of [6] in their empirical section, and also confirmed by our experiments. Thus, proper choice of regularization parameter for the purpose of model selection remains an open problem.

[3]Moreover, [4] proved that cross-validated λ does *not* lead to consistent model selection for the Lasso problem, since it tends to include too many noisy connections between the variables. This is not necessarily surprising, given that cross-validation selects λ that is best for prediction that might be quite different from the model-selection goal, since it is well-known that multiple probabilistic models having quite different structures may yield very similar distributions.

3. OUR APPROACH

Herein, we investigate a Bayesian approach, treating λ as a random variable with a prior density $p(\lambda)$, rather than as a fixed parameter. Our goal is to find a joint maximum-aposteriory probability (MAP) estimate of λ and C by solving $\max_{C \succ 0, \lambda} \ln p(C, \lambda|\mathbf{X})$, or, equivalently, $\max_{C \succ 0, \lambda} \ln p(\mathbf{X}, C, \lambda)$, where $p(X, C, \lambda) = p(X|C)p(C|\lambda)p(\lambda) =$

$= \prod_i^n [(2\pi)^{-p/2} \det(C)^{\frac{1}{2}} e^{-\frac{1}{2}\mathbf{x}_i^T C \mathbf{x}_i}] (\lambda/2)^{p^2} e^{-\lambda||C||_1} p(\lambda).$

This results into the following problem instead of eq. 1:

$$\max_{\lambda, C \succ 0} \frac{n}{2}[\ln \det(C) - \mathrm{tr}(SC)] + p^2 \ln \frac{\lambda}{2} - \lambda||C||_1 + \ln p(\lambda).$$

We considered two types of priors: a uniform (flat) and an exponential prior $p(\lambda)$. The *uniform (flat) prior* puts equal weight on all values of $\lambda \in [0, \Lambda]$ (assuming sufficiently high Λ), and thus effectively ignores $p(\lambda)$; this prior was used in *Regularized Likelihood* method discussed in the next section. The *exponential prior* assumes that $p(\lambda) = be^{-b\lambda}$, yielding

$$\max_{\lambda, C \succ 0} \frac{n}{2}[\ln \det(C) - \mathrm{tr}(SC)] + p^2 \ln \frac{\lambda}{2} - \lambda||C||_1 - b\lambda. \quad (2)$$

Currently, we estimate b as $||S_r^{-1}||_1/(p^2 - 1)$, where $S_r = S + \epsilon I$ is the empirical covariance matrix, slightly regularized with small $\epsilon = 10^{-3}$ on the diagonal to obtain an invertible matrix when S is not invertible. The intuition behind such estimate is that $b = 1/E(\lambda)$, and we approximate $E(\lambda)$ by the solution to the above optimization problem with C fixed to its empirical estimate S_r^{-1}.

We also performed a limited amount of experiments with the Gaussian prior (see section 4.2).

3.1. Solving Optimization Problem

Exponential Prior. Let us consider the optimization problem in eq. 2. The objective function is concave in C for any fixed λ but is not concave in C and λ jointly. Hence we are looking for a local optimum. We use alternating maximization method, which, for each given fixed value of λ, solves the following problem:

$$\phi(\lambda) = \max_C \frac{n}{2} \ln \det(C) - \frac{n}{2} \mathrm{tr}(SC) - \lambda \|C\|_1 \quad (3)$$

This problem has a unique maximizer $C(\lambda)$ for any value of λ [6]. We now consider the following optimization problem

$$\max_\lambda \psi(\lambda) = \max_\lambda \phi(\lambda) + p^2 \ln \lambda - b\lambda. \quad (4)$$

Clearly, the optimal solution to this problem is also optimal for problem (2). The following simple optimization scheme for problem (4) is applied:

0. Initialize λ^1;
1. find $C(\lambda^k), \phi(\lambda^k)$ and $\psi(\lambda^k)$;
2. If $|p^2/\lambda - \|C(\lambda^k)\|_1 - b| < \epsilon$ go to step 5.
3. $\lambda^{k+1} = p^2/(\|C(\lambda^k)\|_1 + b)$;
4. find $C(\lambda^{k+1})$ and $\psi(\lambda^{k+1})$;
   if $\psi(\lambda^{k+1}) > \psi(\lambda^k)$ go to step 3.
   else $\lambda^{k+1} = (\lambda^k + \lambda^{k+1})/2$. Go to step 4.
5. end

This scheme uses line search along the direction of the derivatives and will converge to the local maximum (if one exists) as long as some sufficient increase condition (such as Armijo rule [5]) is applied in Step 4. Step 1 can be performed by any convex optimization method designed to solve problem (3). In our experiments we used glasso software [1].

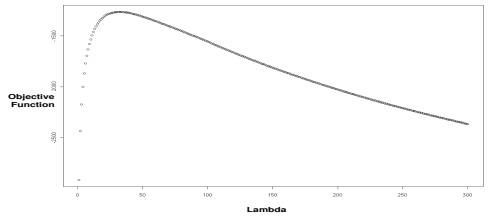

Fig. 1. Typical shape of the objective function $\psi(\lambda)$.

In Figure 1 we show a typical shape of function $\psi(\lambda)$. An analysis of behavior of the elements that compose $\psi(\lambda)$ can show that for any positive b this function goes to $-\infty$ as λ goes to ∞. Hence a maximum of $\psi(\lambda)$ exists for any positive b. The value of λ for which this maximum is achieved increases with b.

Regularized Likelihood with Flat Prior. When $b = 0$, the problem (2) is equivalent to assuming the flat prior on λ. If $n << p$ then the term $p^2 \ln \lambda$ may dominate the total sum and $\psi(\lambda)$ may be unbounded from above. Our experiment show that typically for $n > 3p$ the maximum of $\psi(\lambda)$ is finite.

To be able to handle the cases when $n < 3p$, but only a flat prior is assumed, we propose the following modified optimization procedure. Let

$$\phi(\lambda) = \max_C \frac{n}{2} \ln \det(C) - \frac{n}{2} \mathrm{tr}(SC) - \lambda \|C\|_{1,0},$$

where $\|C\|_{1,0}$ is a sum of the absolute values of all *off-diagonal* elements of C and let $C(\lambda)$ be the solution to the above convex optimization problem. As λ grows the maximum eigenvalue of $C(\lambda)$ no longer converges to zero. In fact one can show that the diagonal elements of $C(\lambda)$ will converge to the inverse of diagonal of the empirical covariance matrix S. Now we consider the following regularized version of the maximum log-likelihood problem

$$\max_\lambda \psi(\lambda) = \max_\lambda \phi(\lambda) + p^2 \ln \lambda - \lambda \sum_i |C_{ii}|. \qquad (5)$$

As in the case of positive b we can show here that a finite maximum always exists. The advantage of this formulation, referred to as *Regularized Likelihood*, is that it does not depend on the choice of b and the regularization term arises naturally from the optimization algorithm. The drawback of this approach is that it no longer can be interpreted as a joint likelihood optimization problem. A procedure, very similar to the algorithm described in (5) can be applied to this regularized approach. The computational results in the next section show that this approach produces good empirical results.

4. EXPERIMENTAL RESULTS

4.1. Synthetic Data

In order to test the structure reconstruction accuracy, we generated two "ground-truth" random inverse-covariance matrices: a very sparse one, with only 4% (off-diagonal) non-zero elements, and a relatively dense one, with 52% (off-diagonal) non-zero elements. We then sampled $n = 30, 50, 500, 1000$ instances from the corresponding multivariate Gaussian distribution over $p = 100$ variables. We used two methods for Bayesian learning of λ discussed in the previous section: (1) *Regularized Likelihood* and (2) *Exponential Prior*. We compare the structure-learning performance as well as the prediction performance of the Bayesian λ with the two other alternatives: (1) λ selected by cross-validation using the prediction error and (2) theoretically derived λ from[6].

We evaluate the structure-learning performance by counting the number of True and False edges that has been discovered. In the following tables **FN** denotes the false-negative error (False Negatives / Positives) and **FP** denotes the false-positive error (False Positive / Negatives). Positives is the number of non-zero off-diagonal elements of the original inverse covariance matrix and Negatives is the number of zero elements of the original inverse covariance matrix. **SE** (Structure Error) is the overall structure reconstruction error computed as (FN*Positives + FP*Negatives)/(Positives+Negatives). **PE** (Prediction error) is the average (over all variables and over all test cases) squared error of prediction on a held-out test data of size 100. λ_r denotes the Bayesian λ that is learned by the regularized likelihood, λ_p denotes the Bayesian λ that is learned by the joint likelihood with the exponential prior on λ, λ_b denotes the λ suggested by Banerjee et. al, and λ_c denotes the λ that is learned by 5-fold cross validation.

As shown in Table 1, cross-validated λ yields much higher false-positive error when compared to other methods (which is consistent with [4]), and the worst total structure-reconstruction error among all competitors on very sparse problems. In particular, it learns a nearly-complete graph when $n > p$. When the original matrix is dense, cross-validates λ yields better results, but effectively it is almost equivalent to setting $\lambda = 0$ (non-regularized maximum likelihood). On the other hand, λ suggested by Banerjee et al yields models that are too sparse, and performs poorly both in

Table 1. Results for $p = 100$ variables.

	Original Density = 4% 356 Positives and 9544 Negatives				Original Density = 52% 5102 Positives and 4798 Negatives			
	FN	**FP**	**SE**	**PE**	**FN**	**FP**	**SE**	**PE**
n=30								
	λ_r = 190, λ_p = 34, λ_b =621, λ_c=2				λ_r = 500, λ_p=32, λ_b=1120, λ_c=0.4			
λ_r	**0.94**	**0.006**	**0.04**	**4.5**	**0.992**	**0.01**	**0.52**	**10.2**
λ_p	**0.7**	**0.05**	**0.07**	**1.88**	**0.87**	**0.12**	**0.51**	**1.4**
λ_b	0.995	0	0.04	6.8	0.9997	0.0004	0.52	14.6
λ_c	0.17	0.3	0.3	1.3	0.5	0.4	0.45	0.54
n=50								
	λ_r = 210, λ_p = 51, λ_b =664, λ_c=1				λ_r = 500, λ_p=47, λ_b=2209, λ_c=0.4			
λ_r	**0.87**	**0.01**	**0.04**	**1.4**	**0.98**	**0.03**	**0.52**	**6.1**
λ_p	**0.66**	**0.05**	**0.07**	**1.8**	**0.87**	**0.12**	**0.51**	**1.36**
λ_b	0.995	0	0.04	6.8	0.999	0.002	0.52	13.5
λ_c	0.03	0.53	0.51	1.16	0.4	0.5	0.45	0.37
n=500								
	λ_r = 55, λ_p = 63, λ_b =2517, λ_c=0.1				λ_r = 23, λ_p=22, λ_b=5438, λ_c=0.1			
λ_r	**0.05**	**0.1**	**0.1**	**0.70**	**0.56**	**0.2**	**0.30**	**0.28**
λ_p	**0.1**	**0.1**	**0.1**	**0.76**	**0.54**	**0.29**	**0.42**	**0.29**
λ_b	0.9	0.01	0.04	3.7	0.98	0.03	0.52	7.4
λ_c	0	0.98	0.95	0.71	0.01	0.93	0.46	0.15
n=1000								
	λ_r = 27, λ_p = 26, λ_b =3401, λ_c=0.1				λ_r = 14, λ_p=14, λ_b=7115, λ_c=0.1			
λ_r	**0**	**0.2**	**0.19**	**0.60**	**0.3**	**0.38**	**0.34**	**0.19**
λ_p	**0**	**0.24**	**0.23**	**0.61**	**0.3**	**0.4**	**0.35**	**0.19**
λ_b	0.84	0.02	0.05	2.95	0.97	0.05	0.52	5.4
λ_c	0	0.99	0.95	0.65	0.01	0.95	0.47	0.13

prediction and model recovery, missing almost all edges with false-negative error often above 90%. Bayesian approach fits between those extremes: it produces intermediate values of λ that yield a much better balance between the two types of errors (see also a typical ROC curve in Figure 2). We note that λ_r that is picked by the regularized likelihood tends to learn a sparser model compared to λ_p that is picked by the joint likelihood with exponential prior on λ. λ_p performs better when $n < p$ using the informative prior.But as the number of observations grow both behave very similarly and they outperform all other methods in terms of model recovery. They out perform cross-validation method in terms of prediction when the original model is sparse and $n > p$.

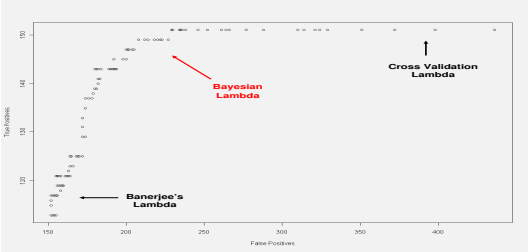

Fig. 2. ROC curve for varying λ, with Bayesian λ between the two extremes: cross-validated and Banerjee's theoretical λ.

4.2. Real-life dataset: brain imaging (fMRI)

We used the fMRI data from the 2007 Pittsburgh Brain Activity Interpretation Competition (PBAIC)[7], where the fMRI data were recorded while subjects were playing a videogame, and the task was to predict several real-valued response variables[4]. Since the "ground truth" network structure is unavailable in real-life scenario (and must be discovered), we only evaluated the predictive ability of our Markov network models. In Table 2 we show the average results for 3 subjects, where the dataset for each subject contained $n = 704$ samples (measurements over time) and approximately $p = 33,000$ variables (voxels). On this dataset, we also experimented with Gaussian vs exponential prior on λ; Gaussian prior appears to yield slightly more accurate results that match the performance of the state-of-art sparse regression method, Elastic Net (EN); both clearly outperform linear regression. Matching state-of-art predictive performance supports our confidence in the Markov network model quality, while the sparse structure we learn can provide scientific insights into brain activation processes (further discussion of which is out of scope of this paper).

Table 2. Results on fMRI data (PBAIC 2007): correlation between the predicted and actual response, averaged over 3 subjects. All methods ran on a subset of preselected 200 voxels (variables) most-correlated with the response. 'OLS' - ordinary least-squares (linear) regression 'EN' - Elastic Net sparse regression, SMN (prior) - our sparse Markov Network learner with a particular prior.

Response	**SMN (exp)**	**SMN (gauss)**	OLS	EN
3 ('Body')	0.44	0.47	0.41	0.49
15('Instructions')	0.52	0.68	0.69	0.69
22('VRfixation')	0.77	0.79	0.78	0.80
24('Velocity')	0.61	0.63	0.59	0.65

5. REFERENCES

[1] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2007.

[2] D. Heckerman. A tutorial on learning Bayesian networks, Tech. Report MSR-TR-95-06. *Microsoft Research*, 1995.

[3] Y. Lin, D.D. Lee, Y. Kim, and B. Taskar. Learning Markov Network Structure via Sparse Ensemble-of-Trees Models. *submitted*, 2008.

[4] N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[5] J. Nocedal and S.J. Wright. *Numerical Optimization, Second Edition*. Springer, 2006.

[6] O.Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, March 2008.

[7] Pittsburgh EBC Group. PBAIC Homepage: http://www.ebc.pitt.edu/2007/competition.html, 2007.

[8] M. Wainwright, P. Ravikumar, and J. Lafferty. High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression. In *NIPS 19*, pages 1465–1472. 2007.

[9] M. Yuan and Y. Lin. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94(1):19–35, 2007.

[4]We experimented with several response variables such as *Instructions*(whether a person is listening to audio instructions), *Body* (looking at virtual person), *VRfixation* (in VR world vs fixation) and *Velocity* (subject moving but not interacting with VR objects) - see [7] for more details.