RC25074 (W1011-028) November 4, 2010
Mathematics

IBM Research Report

Sparse Markov Net Learning with Priors on Regularization Parameters

Katya Scheinberg
IEOR Department
Columbia University
New York, NY

Irina Rish
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Narges Bani Asadi
Department of Electrical Engineering
Stanford University
Palo Alto, CA

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

Sparse Markov Net Learning with Priors on Regularization Parameters

Katya Scheinberg
IEOR Department
Columbia University
New York, NY

Irina Rish
Computational Biology Department
IBM T. J. Watson Research Center
Yorktown Heights, NY

Narges Bani Asadi
Department of Electrical Engineering
Stanford University
Palo Alto, CA

Abstract

In this paper, we consider the problem of structure recovery in Markov Network over Gaussian variables, that is equivalent to finding the zero-pattern of the sparse inverse covariance matrix. Recently proposed l_1-regularized optimization methods result into convex problems that can be solved optimally and efficiently. However, the accuracy such methods can be quite sensitive to the choice of regularization parameter, and optimal selection of this parameter remains an open problem. Herein, we adopt a Bayesian approach, treating the regularization parameter(s) as random variable(s) with some prior, and using MAP optimization to find both the inverse covariance matrix and the unknown regularization parameters. Our general formulation allows a vector of regularization parameters and is well-suited for learning structured graphs such as scale-free networks where the sparsity of nodes varies significantly. We present promising empirical results on both synthetic and real-life datasets, demonstrating that our approach achieves a better balance between the false-positive and false-negative errors than commonly used approaches.

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 (Heckerman 1995). Recently, even better approximation methods were suggested (Meinshausen & Buhlmann 2006; Wainwright, Ravikumar, & Lafferty 2007; Yuan & Lin 2007; O.Banerjee, El Ghaoui, & d'Aspremont 2008; Friedman, Hastie, & Tibshirani 2007; Duchi, Gould, & Koller 2008) 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. Indeed, 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[1]. On the other hand, theoretically derived λ (see (O.Banerjee, El Ghaoui, & d'Aspremont 2008)) has asymptotic guarantee of correct recovery of the *connectivity components* (rather than edges), which correspond to *marginal* rather than conditional independencies, i.e. to the entries in covariance rather then the inverse covariance matrix. Although such approach is asymptotically consistent, for finite number of samples it tends to miss many edges, resulting into high false-negative error rates.

In this paper, we propose a Bayesian approach to regularization parameter selection, that also generalizes to the case of vector-λ, allowing to choose, if necessary, a different sparsity level for different nodes in the network. (This work extends our approach to scalar-λ selection proposed in (Asadi *et al.* 2009); for completeness sake, we will also summarize here the results from (Asadi *et al.* 2009)). More specifically, the regularization parameter controlling the sparsity of solution are considered to be random variable with particular priors, and the objective is to find a MAP solution (Θ, Λ), where Θ is the set of model parameters and Λ is the set of regularization parameters. Our algorithm is based on alternating optimization over Θ and Λ, respectively. Empirical results demonstrate that our approach compares favorably to previous approaches, achieving a better balance between the false-positive and false-negative errors.

Note that our general formulation is well-suited for learn-

[1]This is actually not surprising as it is well known that cross-validated λ for the *prediction* objective can be a bad choice for the *structure recovery/model selection* in l_1-regularized setting (e.g., see (Meinshausen & Buhlmann 2006) for examples when λ selected by cross-validation leads to provably inconsistent structure recovery).

ing structured networks with potentially very different node degrees (and thus different sparsity of the columns in the inverse covariance matrix). One common practical example of such networks are networks with heavy-tail (power-law) degree distributions, also called scale-free networks. Examples of such networks include social networks, protein interaction networks, Internet, world wide web, correlation networks between active brain areas in fMRI studies (V.M. Eguiluz and D.R. Chialvo and G.A. Cecchi and M. Baliki and A.V. Apkarian 2005), and many other real-life networks (see (Barabasi & Albert 1999) for a survey). Empirical results on both random and structured (power-law) networks demonstrate that our approach compares favorably to previous approaches, achieving a better balance between the false-positive and false-negative errors.

Our Approach

Let $X = \{X_1, ..., X_p\}$ be a set of p random variables, and let $G = (V, E)$ be a Markov network (a Markov Random Field, or MRF) representing the conditional independence structure of the joint distribution $P(X)$. The set of vertices $V = \{1, ..., p\}$ is in a one-to-one correspondence with the set of variables in X. The edge set E contains an edge (i, j) if and only if X_i is conditionally dependent on X_j given all remaining variables; the lack of edge between X_i and X_j means that the two variables are conditionally independent given all remaining variables (Lauritzen 1996).

We will assume a multivariate Gaussian probability density function over $X = \{X_1, ..., X_p\}$:

$$p(\mathbf{x}) = (2\pi)^{-p/2} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}}\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)} \quad (1)$$

where μ is the mean and $\mathbf{\Sigma}$ is the covariance matrix of the distribution, respectively, and $\mathbf{x}^T$ denotes the transpose of the column-vector $\mathbf{x}$. Without loss of generality we will assume that the data are normalized to have zero mean ($\mu = \mathbf{0}$), and we only need to estimate the parameter Σ (or Σ^{-1}). Since $det(\Sigma)^{-1} = det(\Sigma^{-1})$, we can now rewrite eq. 1, assuming $C = \Sigma^{-1}$ and $\mu = 0$:

$$p(\mathbf{x}) = (2\pi)^{-p/2} \det(C)^{\frac{1}{2}} e^{-\frac{1}{2}\mathbf{x}^T C \mathbf{x}}. \quad (2)$$

Missing edges in the above graphical model correspond to zero entries in the inverse covariance matrix $C = \Sigma^{-1}$, 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. Note that the inverse of the maximum-likelihood estimate of the covariance matrix Σ (i.e. the empirical covariance matrix $A = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i$ where $\mathbf{x}_i$ is the i-th sample, $i = 1, ..., n$), even if it exists, does not typically contain any elements that are exactly zero. Therefore an explicit sparsity-enforcing constraint needs to be added to the estimation process.

A common approach is to include as penalty the l_1-norm of C, which is equivalent to imposing a Laplace prior on C in maximum-likelihood framework (O.Banerjee, El Ghaoui, & d'Aspremont 2008; Friedman, Hastie, & Tibshirani 2007; Yuan & Lin 2007; Duchi, Gould, & Koller 2008). Formally, the entries C_{ij} of the inverse covariance matrix C are assumed to be independent random variables, each following a Laplace distribution $p(C_{ij}) = \frac{\lambda_{ij}}{2} e^{-\lambda_{ij}|C_{ij}-\alpha_{ij}|}$ with zero location parameter (mean) α_{ij} and common scale parameter $\lambda_{ij} = \lambda$, yielding $p(C) = \prod_{i=1}^{p}\prod_{j=1}^{p} p(C_{ij}) = (\lambda/2)^{p^2} e^{-\lambda||C||_1}$, where $||C||_1 = \sum_{ij}|C_{ij}|$ is the (vector) l_1-norm of C. Then the objective is to find the maximum-likelihood solution $\arg\max_{C\succ 0} p(C|\mathbf{X})$, 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 (O.Banerjee, El Ghaoui, & d'Aspremont 2008; Friedman, Hastie, & Tibshirani 2007; Yuan & Lin 2007; Duchi, Gould, & Koller 2008):

$$\max_{C\succ 0} \ln\det(C) - \mathrm{tr}(AC) - \lambda||C||_1 \quad (3)$$

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

Herein, we make a more general assumption about $p(C)$, allowing different rows in C to have different parameters λ_i, i.e., $p(C_{ij}) = \frac{\lambda_i}{2} e^{-\lambda_i|C_{ij}|}$ This reflects our desire to model structured networks with potentially very different node degrees (i.e., row densities in C). This yields $p(C) = \prod_{i=1}^{p}\prod_{j=1}^{p} \frac{\lambda_i}{2} e^{-\lambda_i|C_{ij}|} = = \prod_{i=1}^{p} \frac{\lambda_i^p}{2^p} e^{-\lambda_i \sum_{j=1}^{p}|C_{ij}|}$.

Moreover, we will take Bayesian approach and assume that parameters λ_i are also random variables following some joint distribution $p(\{\lambda_i\})$. Given a dataset X of n samples (rows) of vector $\mathbf{X}$, the joint log-likelihood can be then written as

$$\ln L(X, C, \{\lambda_i\}) = \ln\{p(X|C)p(C|\{\lambda_i\})p(\{\lambda_i\})\} =$$

$$const + \frac{n}{2}\ln\det(C) - \frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i^T C \mathbf{x}_i + p\sum_{i}^{p}\ln\frac{\lambda_i}{2} -$$
$$-\sum_{i}^{p}\lambda_i\sum_{j=1}^{p}|C_{ij}| + \ln p(\{\lambda_i\}),$$

where $const$ does not depend on C or $\{\lambda_i\}$.

We can also rewrite $\sum_{i=1}^{n}\mathbf{x}_i^T C \mathbf{x}_i = n\mathrm{tr}(AC)$ where tr denotes the trace of a matrix, and $A = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{x}_i$ is the empirical covariance matrix.

We will use the maximum a posteriori probability (MAP) approach that requires maximization of the above joint log-likelihood, rewritten as

$$\max_{C\succ 0,\{\lambda_i\}} \frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)] - \sum_{i}^{p}\lambda_i\sum_{j=1}^{p}|C_{ij}| +$$
$$+p\sum_{i}^{p}\ln\lambda_i + \ln p(\{\lambda_i\}),$$

where $C \succ 0$ constraint ensures the solution C (inverse covariance matrix) is positive definite.

We considered independent λ_i with several types of priors $p(\lambda_i)$: (1) uniform (flat), (2) exponential and (3) Gaussian. The *uniform (flat) prior* puts equal weight on all values of

$\lambda_i \in [0, \Lambda_i]$ (assuming sufficiently high Λ_i), and thus effectively ignores $p(\lambda_i)$; this prior was used in *Regularized Likelihood* method discussed in the next section.

The *exponential priors* $p(\lambda_i) = b_i e^{-b_i \lambda_i}$ yield:

$$\max_{C \succ 0, \lambda \in \mathbf{R}^p} \frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)] - \sum_i^p \lambda_i \sum_{j=1}^p |C_{ij}| + p\sum_i^p \ln\lambda_i - \sum_i^p b_i\lambda_i. \quad (4)$$

Rather than taking a more expensive, fully Bayesian approach here and integrating out C in order to obtain the maximum-likelihood type II estimate of b_i, we will use an approximate estimate $b_i = ||A(i)_r^{-1}||_1/p$, where $A_r = A + \epsilon I$ is the empirical covariance matrix[2], and $A_r(i)$ denotes its i-th row. In other words, b_i is estimated as an average l_1-norm per element of i-th row.

Finally, we also considered the truncated (to exclude negative values of λ) unit-variance Gaussian prior which replaces $\sum_i^p b_i\lambda_i$ in the equation above with $\sum_i^p (\lambda_i - b_i)^2/2$.

Fixed-Point Method for λ Selection

We shall now address the optimization problem arising in selection of parameter λ as discussed in Section .

Scalar λ

We consider the following optimization problems:

$$\max_{C,\lambda} \frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)] - \lambda\|C\|_S + p^2\ln\lambda - \theta(\lambda), \quad (5)$$

where $\theta(\lambda)$ is some given function of λ derived from the particular prior $p(\lambda)$. By $\|C\|_S$ we denote the sum of absolute values of the elements of the matrix $S \cdot C$, where $\cdot$ denotes the element-wise product, where S is a given $p \times p$ matrix with nonnegative entries.

Let $f(C) = \frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)]$, and let us consider the following function:

$$\phi(\lambda) = \max_C f(C) - \lambda\|C\|_S.$$

The function $\frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)] - \lambda\|C\|_S$ is strictly concave and, hence, has a unique maximizer $C(\lambda)$ for any value of λ. From general theory of convex optimization in we know that $\phi(\lambda)$ is a differentiable convex function whose derivative for any given λ equals $\|C(\lambda)\|_S$. The proof of this simple fact can be found in the Appendix.

Lemma 1 *$\phi(\lambda)$ is a differentiable convex function whose derivative for any given λ equals $-\|C(\lambda)\|_S$.*

Now let us consider the following optimization problem

$$\max_\lambda \psi(\lambda) = \max_\lambda \phi(\lambda) + p^2\ln\lambda - \theta(\lambda). \quad (6)$$

Clearly, the optimal solution to this problem is also optimal for problem (5). To find $\phi(\lambda)$ one needs to solve the sparse inverse covariance selection problem with a fixed value of λ. This can be done by applying one of the techniques recently proposed for this method, such as, (O.Banerjee, El Ghaoui, & d'Aspremont 2008; Friedman, Hastie, & Tibshirani 2007; Duchi, Gould, & Koller 2008; Rot 2008). Herein, we used the *glasso* method proposed in (Friedman, Hastie, & Tibshirani 2007).

Notice that $\psi(\lambda)$ is a sum of a convex and a concave functions, hence is neither convex nor concave and may have multiple local optima. In our experiments with $\theta(\lambda) = b\lambda$ we observed that the maximum was unique in most of the cases. In the rare case when it appeared to be not unique, it was not clear if such was the true nature of $\psi(l)$ or a result of inaccuracies in the solution of the convex subproblems.[3]

We will describe the optimization scheme to solve problem 6 in the next section, focusing on the exponential prior. A similar analysis and an update rule can be derived for the Gaussian prior on λ.

Vector λ

Now let us consider a similar problem to (5), but with λ - a vector of weights for the S-norm of the columns of C. Hence we will now consider a vector norm $\|C_i\|_{S_i}$ which is the same the matrix norm we discussed before, applied to columns (or rows) of C and S.

$$\max_{C,\lambda} \quad \frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)] - \sum_{i=1}^p \lambda_i\|C_i\|_{S_i} + p\sum_{i=1}^p \ln\lambda_i - \sum_{i=1}^p b_i\lambda_i$$

As before, let $f(C) = \frac{n}{2}[\ln\det(C) - \mathrm{tr}(AC)]$, and

$$\phi(\lambda) = \max_C f(C) - \lambda\|C\|_S.$$

Notice that, for any fixed $\bar\lambda$, $\sum_i \bar\lambda_i\|C_i\|_{S_i} = \|C\|_{\lambda\cdot S}$, where $\lambda \cdot S$ is a matrix whose i-th column equals $\lambda_i S_{\bar i}$ for all i. This implies that for any fixed and given λ and $\bar\lambda$ function $\phi(\theta) = \phi(\lambda + \theta\bar\lambda)$ reduces to the case of the univariate $\phi(\lambda)$ described in the previous section. This implies, for instance, that the multivariate function $\phi(\lambda)$ is convex in any direction, hence is convex in general. Also from the analysis in

[2] slightly regularized with small $\epsilon = 10^{-3}$ on the diagonal to obtain an invertible matrix when A is not invertible.

[3] Let us consider exponential prior. The derivative of $\psi(\lambda)$ is $p^2/\lambda - \|C(\lambda)\|_S - b$. Here are some observations which help explain why a unique stationary point is typical for the case when $\theta(\lambda) = b\lambda$. We are considering all point for which $\psi(\lambda) = 0$. Multiplying the expression for the gradient by nonnegative λ we have that

$$\lambda\psi'(\lambda) = p^2 - \lambda\|C(\lambda)\|_S - b\lambda$$

If the quantity $\lambda\|C(\lambda)\|_S$ increases as λ grows (which is expected since the decrease of λ usually slows down with the growth of λ) then the right hand side of the last equality is a decreasing function of λ. Hence the equality to zero can only be achieved for a single value of λ which would imply unique maximum. We are not yet aware of any theoretical result that guarantees that $\lambda\|C(\lambda)\|_S$ increases monotonically with λ but we have consistently observed it in the experiments. Note also that for sufficiently small λ the quantity $\lambda\psi'(\lambda)$ is positive, while for sufficiently large λ it becomes negative. Hence an existence of at least one local maximum is always guaranteed.

the previous section it is easy to see that the i-th element of the gradient of $\phi(\lambda)$ equals $-\|C_i(\lambda)\|_{S_i}$.

Now we consider $\psi(\lambda) = \phi(\lambda) + p\sum_{i=1}^{p} \ln \lambda_i - \sum_{i=1}^{p} b_i\lambda_i$. This function is again neither concave nor convex. Its gradient is

$$(\nabla\psi(\lambda))_i = -\|C_i(\lambda)\|_{S_i} + p/\lambda_i - b_i, \quad i = 1, \ldots, p,$$

where $C(\lambda)$ is, again, the maximizer of $f(C) - \lambda\|C\|_S$ for a given λ. Hence for λ^* which maximizes $\psi(\lambda)$ we have

$$\|C_i(\lambda^*)\|_{S_i} + b_i = p/\lambda_i^*, \quad i = 1, \ldots, p,$$

or, equivalently,

$$\lambda_i^* = \frac{p}{\|C_i(\lambda^*)\|_{S_i} + b_i} \quad , i = 1, \ldots, p.$$

Hence λ^* is a fixed point of the following operator $T(\lambda) = p/(\|C(\lambda)\|_S + b)$, where by $p/(\|C(\lambda)\|_S + b)$ we mean a p-dimensional vector with entries $p/(\|C_i(\lambda)\|_{S_i} + b_i)$. To solve this problem we consider applying the following fixed point algorithm

0. Initialize λ^1;
1. find $C(\lambda^k)$ and $\phi(\lambda^k)$;
2. If $\sum_i (p/\lambda_i - \|C_i(\lambda^k)\|_{S_i} - b_i)^2 < \epsilon$ go to step 4.
3. $\lambda_i^{k+1} = p/(\|C_i(\lambda^k)\|_{S_i} + b_i)$; go to step 1.
4. end

Note that in Step 1 we perform a standard inverse covariance selection optimization problem with fixed λ such as is done in the previous section.

In our experiments the fixed point algorithm presented above converged in every experiment. While we do not have theoretical guarantees of the convergence of the algorithm, we will present a modification of the algorithm which invokes a line search algorithm in case the fixed point iteration fails to provide sufficient improvement in the objective function $\psi(\lambda)$.

We apply the following optimization algorithm.

0. Initialize λ^1;
1. find $C(\lambda^k)$ and $\phi(\lambda^k)$;
2. If $\sum_i (p/\lambda_i - \|C_i(\lambda^k)\|_{S_i} - b_i)^2 < \epsilon$ go to step 5.
3. $\lambda_i^{k+1} = p/(\|C_i(\lambda^k)\|_{S_i} + b_i)$; (7)
4. find $C_i(\lambda^{k+1})$ and $\psi(\lambda^{k+1})$;
   if $\psi(\lambda^{k+1}) > \psi(\lambda^k)$ $k = k + 1$, go to step 3.
   else $\lambda^{k+1} = (\lambda^k + \lambda^{k+1})/2$. Go to step 4.
5. end

The proposed algorithm performs a line search along the direction d defined by $d_i = p/(\|C_i(\lambda)\|_{S_i} + b_i) - \lambda_i$, while the gradient of $\psi(\lambda)$ equals g such that $g_i = p/\lambda_i - \|C_i(\lambda)\|_{S_i} - b_i$. If we consider the inner product, we have $d^T g =$

$$\sum_i (p^2/(\lambda_i(\|C_i(\lambda)\|_{S_i} + b_i)) + \lambda_i(\|C_i(\lambda)\|_{S_i} + b_i) - 2p) =$$

$$\frac{1}{p}\sum_i (p/\lambda_i(\|C_i(\lambda)\|_{S_i} + b_i) + \lambda_i(\|C_i(\lambda)\|_{S_i} + b_i)/p - 2) \geq 0.$$

Hence, unless $p = \lambda_i\|C_i(\lambda)\|_{S_i}$ for all i, then we know the direction d makes and obtuse angle with the gradient and, thus, is an ascent direction. In the case when $p = \lambda_i\|C_i(\lambda)\|_{S_i}$ for all i, then the gradient of $\psi(\lambda)$ is zero and the algorithm have converged to a local stationary point. In fact we can show that

$$d^T g/\|d\|\|g\| \geq const > 0,$$

for all cases when $\|d\|\|g\| > 0$, which means that the cosine of angle between the gradient and the direction d remains bounded away from zero, which will in turn imply that sufficient ascent can always be achieved by a line search along direction d. Indeed, from $\|d\|\|g\| \leq \frac{\|d\|^2 + \|g\|^2}{2}$ we have $d^T g/\|d\|\|g\| \geq$

$$2\frac{\sum_{i=1}^{p}(p^2/(\lambda_i(\|C_i(\lambda)\|_{S_i} + b_i) + \lambda_i(\|C_i(\lambda)\|_{S_i} + b_i) - 2p)}{\sum_{i=1}^{p}((\frac{p}{\lambda_i} - (\|C_i(\lambda)\|_{S_i} + b_i))^2 + (\frac{p}{(\|C_i(\lambda)\|_{S_i} + b_i)} - \lambda_i)^2)} \geq$$

$$\sum_{i=1}^{p} \lambda_i(\|C_i(\lambda)\|_{S_i} + b_i) \geq const > 0.$$

The last inequality comes from the facts that $(\|C_i(\lambda)\|_{S_i} + b_i) > b_i)$ and that $\lambda_i \geq \delta > 0$ for all $i = 1, \ldots, p$.

The advantage of the Algorithm (7) is that, while theoretically convergent to the optimum solution, it only resorts to line search if the initial fixed point iteration fails. Hence, in practice, no extra work is necessary to apply this algorithm. In our experiment the number of fixed point iterations was small compared to the dimension p and the algorithm worked very efficiently. The work of each iteration is essentially the same as the work taken by a single solve of the inverse covariance problem, but since the consecutive solves are related, one can successfully utilize warm starts.

Flat Prior: Specific Case

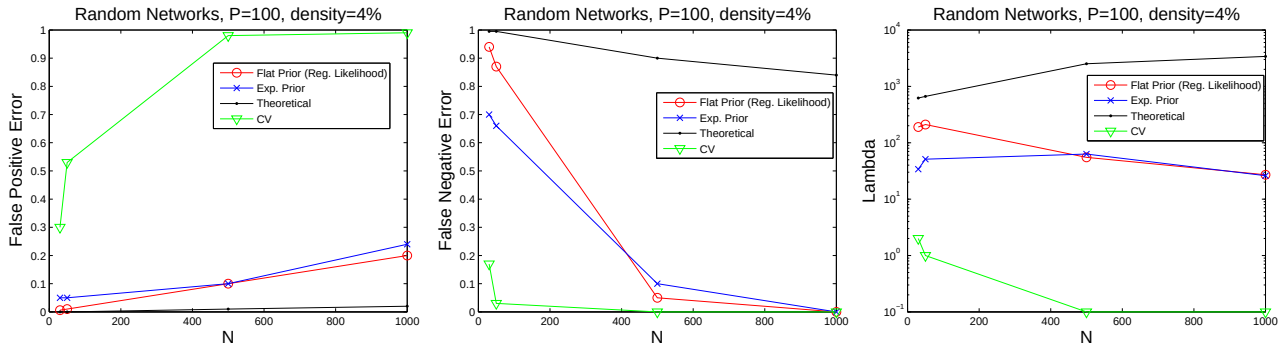
Assuming the flat prior on λ is equivalent to setting $b = 0$ in the exponential-prior formulation. However, when $n << p$, the term $p^2 \ln \lambda$ may dominate the total sum and $\psi(\lambda)$ may be unbounded from above. In order to handle the flat prior case, we propose the following modified 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

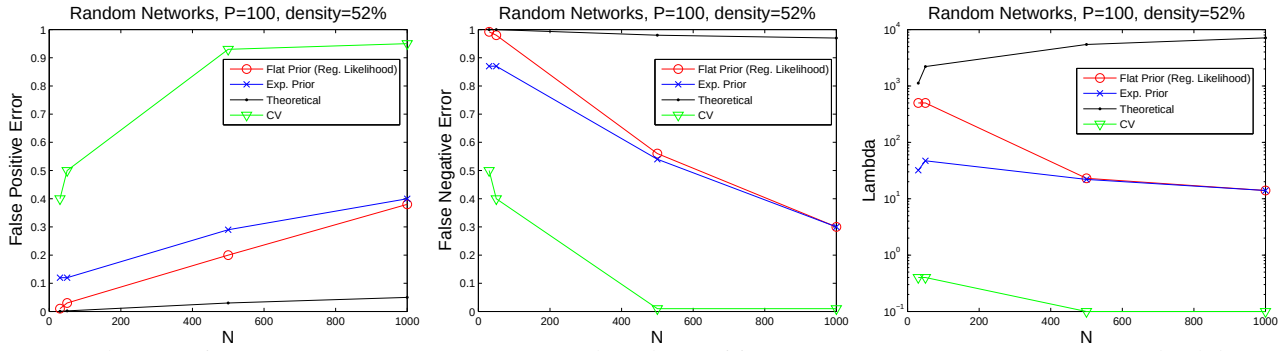
$$\max_\lambda \psi(\lambda) = \max_\lambda \phi(\lambda) + p^2 \ln \lambda - \lambda \sum_i |C_{ii}|. \qquad (8)$$

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



(a) False Negatives (b) False Positives (c) lambda

Figure 1: Results on very sparse random networks (4% density).



(a) False Negatives (b) False Positives (c) lambda

Figure 2: Results on dense random networks (52% density).

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 above can be applied to this regularized approach. The computational results in the next section show that this approach produces good empirical results.

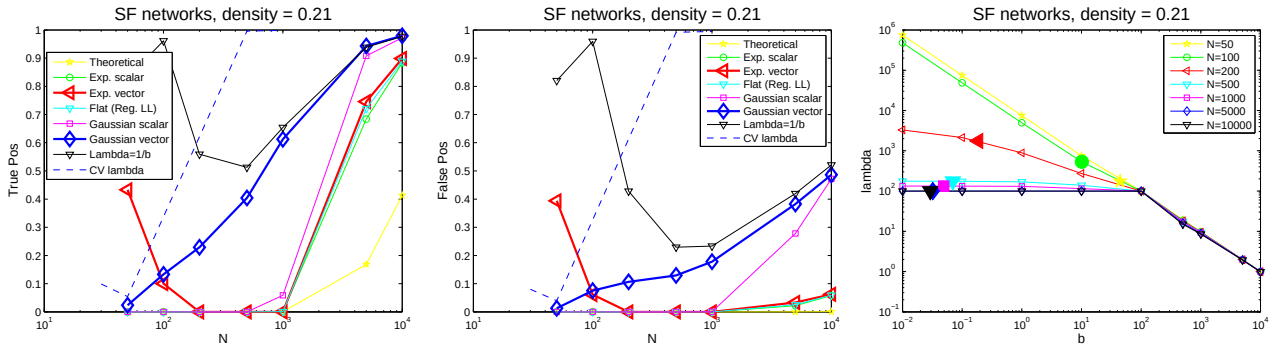
Empirical Evaluation

In order to test structure-reconstruction accuracy, we performed experiments on several types of synthetic problems. (Note that, unlike prediction of an observed variable, structure reconstruction accuracy is harder to test on "real" data since (1) the "true" structure my not be available and (2) known links in "real" networks (e.g., known gene networks) may not necessarily correspond to links in the underlying Markov net.)

In all our experiments, we used *glasso* (Friedman, Hastie, & Tibshirani 2007) method to solve the sparse inverse covariance selection problem with a fixed value of λ (a subproblem in our alternating minimization scheme); we also used *glasso* (Friedman, Hastie, & Tibshirani 2007) when choosing λ via cross-validation (described below).

First, we experimented with uniform random matrices. 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 compared 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 (O.Banerjee, El Ghaoui, & d'Aspremont 2008). Figures 1 and 2 show the results on a sparse (4% link density) and a dense (52% link density) random matrices, respectively. We can clearly see that: (1) cross-validated λ (green) overfits dramatically, producing almost complete matrix (almost 100% false-positive rate); (2) theoretically derived λ (Banerjee *et al.* 2006) (shown in black) is too conservative: it misses almost all edges (has a very high false-negative rate); (3) prior-based approaches - flat prior (red) and exponential prior (blue) yield much more balanced trade-off between the two typos of errors.

We also experimented with semi-realistic, "structured" random networks that followed a power-law degree distribution over the variables. The networks were generated using the preferential attachment (Barabasi-Albert)



(a) TP (b) FP (c) lambda vs b

Figure 3: Results on scale-free networks (21% density).

model (Barabasi & Albert 1999) [4], that produces "scale-free" (power-law) networks containing (relatively few) very highly connected "hubs" besides a large number of sparsely connected nodes. Although such networks are sparse in terms of total number of edges, their power-law structure is a natural candidate for using vector rather than scalar sparsity parameter.

We generated power-law networks with density $5\%, 21\%$ and 31%, measured by the $\%$ of non-zero off-diagonal entries. For each density level, we generated 5 different power-law networks over $p = 100$ variables, that defined the structure of the "ground-truth" inverse covariance matrix, and for each of them, we generated 5 matrices with randomly generated covariances corresponding to the non-diagonal non-zero entries (bounded by 0.1 in order to ensure the resulting matrix is positive-definite)[5] We then sampled $n = 50, 100, 200, 500, 1000, 10000$ instances from the corresponding multivariate Gaussian distribution over $p = 100$ variables.

We evaluated the following methods of selecting λ when reconstructing the sparse MRF structure from data: (1) *theoretical* λ - theoretically derived λ in (Banerjee *et al.* 2006); (2) *cross-validation* λ is selected as one giving best average prediction (i.e. minimizing the sum-squared prediction error) over the network nodes; (3) "exponential scalar" and (4) "exponential vector" correspond to parameter selected using our method with exponential prior, its scalar and vector versions, respectively; (5) Gaussian scalar and (6) Gaussian vector are defined similarly for Gaussian prior; (7) *fixed* $\lambda = 1/b$ simply assigns to λ the mean of the exponential distribution estimated directly from the data as mentioned in section ; finally (8) *flat ("regularized likelihood")* corresponding to the flat-prior version, along with the other scalar priors. We report the *off-diagonal* true positive (TP) and false positive (FP) rates.

Figure 3 summarizes the results on scale-free networks with density 21%, comparing vector-lambda approach to the scalar approach and to a wide variety of other methods mentioned above (we observed similar type of results for other densities). We observed that:

1. similarly to the above experiments with random networks, cross-validation (CV) for prediction often selects nearly-zero λ, and thus is similar to unregularized ML estimate, selecting too many edges and having very high false-positive rate;
2. theoretical (Banerjee's) λ is another extreme: its edge selection is too conservative in order to bound the false-positive rate of the covariance matrix entries asymptotically, and thus its true positive rate is close to zero unless the number of samples becomes very large;
3. *our approaches are in between the two extremes, for both exponential and Gaussian priors.*
4. *vector-λ approaches seem preferable in the relatively low-sample regime (especially Gaussian-vector-method), since their scalar counterparts tend to be too conservative and yield TP=0 in that regime*;
5. regularized likelihood behaves very similar to scalar-exponential method, but does not require parameter tuning;
6. simply setting $\lambda = 1/b$, i.e. to the mean of the exponential distribution, does not seem to work well as its FP rate is very high in both small-sample and large-sample regimes, only going somewhat doing in the mid-sample regime.
7. *Figure 3(c) shows that λ is not very sensitive to the choice of b for a wide range of bs when N gets sufficiently high.*

In summary, empirical results suggest that Bayesian approach to selecting the regularization parameter may provide an attractive alternative to both cross-validates and theoretical λ selection in various empirical settings, since it achieves a better balance between the false-positive and false-negative errors than the two commonly used approaches (cross-validation and theoretical). However, consistency analysis of the proposed approach remains a direction for future work, as well as empirical evaluation of this approach on real-life data.

Real-life dataset: brain imaging (fMRI)

Finally, we present some initial results on real-life data. We used the fMRI data from the 2007 Pittsburgh Brain Activity Interpretation Competition (PBAIC)(Pittsburgh EBC Group 2007), where the fMRI data were recorded while subjects were playing a videogame, and the task was to predict sev-

[4] We used the open-source Matlab code available at http://www.mathworks.com/matlabcentral/fileexchange/11947.

[5] The variance over the results was quite small.

Table 1: 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

eral real-valued response variables. 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 (Pittsburgh EBC Group 2007) for more details. 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 1 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).

Appendix

Proof of Lemma 1:

Proof. Consider $\frac{\phi(\lambda+d\lambda)-\phi(\lambda)}{d\lambda} =$

$$= \frac{f(\lambda+d\lambda) - (\lambda+d\lambda)\|C(\lambda+d\lambda)\|_S - f(\lambda) + \lambda\|C(\lambda)\|_S}{d\lambda} =$$

$$\frac{f(\lambda+d\lambda) - \lambda\|C(\lambda+d\lambda)\|_S - f(\lambda) + \lambda\|C(\lambda)\|_S}{d\lambda} - \|C(\lambda+d\lambda)\|_S$$

We will now show that

$$\lim_{d\lambda\to 0} \frac{f(\lambda+d\lambda) - \lambda\|C(\lambda+d\lambda)\|_S - f(\lambda) + \lambda\|C(\lambda)\|_S}{d\lambda} = 0 \qquad (9)$$

Let us consider only $d\lambda > 0$ for a moment. Assume that

$$\limsup_{d\lambda\to 0} \frac{f(\lambda+d\lambda) - \lambda\|C(\lambda+d\lambda)\|_S - f(\lambda) + \lambda\|C(\lambda)\|_S}{d\lambda} \geq 2\epsilon > 0.$$

This means that there is an infinite sequence $d\lambda_k \to +0$ such that

$$f(\lambda+d\lambda_k) - \lambda\|C(\lambda+d\lambda_k)\|_S \geq f(\lambda) - \lambda\|C(\lambda)\|_S + \epsilon d\lambda_k.$$

Since $\epsilon d\lambda_k > 0$ this means that for some small enough $d\lambda_k$ $C(\lambda + d\lambda_k)$ is a better solution that $C(\lambda)$ for the given λ. Since by assumption $C(\lambda)$ is the maximizer, then we have reached a contradiction and the above $\limsup$ equals to zero.

Now assume

$$\liminf_{d\lambda\to 0} \frac{f(\lambda+d\lambda) - \lambda\|C(\lambda+d\lambda)\|_S - f(\lambda) + \lambda\|C(\lambda)\|_S}{d\lambda} \leq 2\epsilon < 0.$$

Again we have a sequence $d\lambda_k \to +0$ for which

$$f(\lambda+d\lambda_k) - \lambda\|C(\lambda+d\lambda_k)\|_S \leq f(\lambda) - \lambda\|C(\lambda)\|_S + \epsilon d\lambda_k.$$

or

$$f(\lambda+d\lambda_k) - (\lambda+d\lambda_k)\|C(\lambda+d\lambda_k)\|_S \leq f(\lambda) - (\lambda+d\lambda_k)\|C(\lambda)\|_S + \epsilon d\lambda_k + d\lambda(\|C(\lambda)\|_S - \|C(\lambda+d\lambda_k)\|_S).$$

Since $\|C(\lambda)\|_S - \|C(\lambda+d\lambda_k)\|_S) \to 0$ as $d\lambda_k \to 0$, then for large enough k we have

$$f(\lambda+d\lambda_k) - (\lambda+d\lambda_k)\|C(\lambda+d\lambda_k)\|_S < f(\lambda) - (\lambda+d\lambda_k)\|C(\lambda)\|_S,$$

which contradicts the fact that $C(\lambda + d\lambda_k)$ is the optimal solution for $\lambda + d\lambda_k$. The proof can be repeated almost identically for $d\lambda < 0$, hence we have shown (9).

It is now trivial to conclude that the derivative

$$\begin{aligned} \phi'(\lambda) &= \lim_{d\lambda\to 0} \frac{\phi(\lambda+d\lambda) - \phi(\lambda)}{d\lambda} \\ &= \lim_{d\lambda\to 0} -\|C(\lambda+d\lambda)\|_S = -\|C(\lambda)\|_S \end{aligned}$$

The convexity follows from the simple fact that as λ increases $\|C(\lambda)\|_S$ has to decrease, hence the derivative of $\phi(\lambda)$ increases.

■

References

Asadi, N. B.; Rish, I.; Scheinberg, K.; Kanevsky, D.; and Ramabhadran., B. 2009. A MAP Approach to Learning Sparse Gaussian Markov Networks. In *ICASSP*.

Banerjee, O.; Ghaoui, L. E.; d'Aspremont, A.; and Natsoulis, G. 2006. Convex optimization techniques for fitting sparse Gaussian graphical models. In *ICML*. 89–96.

Barabasi, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Duchi, J.; Gould, S.; and Koller, D. 2008. Projected subgradient methods for learning sparse gaussians. In *Proc. of UAI-08*.

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

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

Lauritzen, S. 1996. *Graphical Models*. Oxford University Press.

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

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

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

2008. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* (2):494–515.

V.M. Eguiluz and D.R. Chialvo and G.A. Cecchi and M. Baliki and A.V. Apkarian. 2005. Scale-free functional brain networks. *Physical Review Letters* 94:018102.

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

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