"Recovering the Structure of Sparse Markov Networks from High-Dimensional Data"

Narges Bani Asadi
Irina Rish
Dimitri Kanevsky
Katya Scheinberg
Bhuvana Ramabhadran

IBM TJ Watson Research Center

September 19, 2008

Outline

- Guassian Markov Network

Outline

- Guassian Markov Network
- Learning Sparse Gaussian Networks

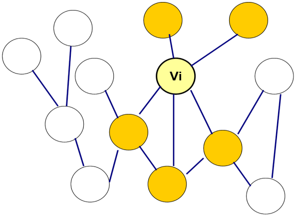
Outline

- Guassian Markov Network
- Learning Sparse Gaussian Networks
- Learning the Sparsity of the Sparse Gaussian Networks

Outline

- Guassian Markov Network
- Learning Sparse Gaussian Networks
- Learning the Sparsity of the Sparse Gaussian Networks
- Results

Markov Networks



$X = \{X_1, ..., X_p\}$

$G = (V, E)$

$P(\mathbf{X}) = \frac{1}{\mathbf{Z}} \prod_k \Phi_k(\mathbf{X})$

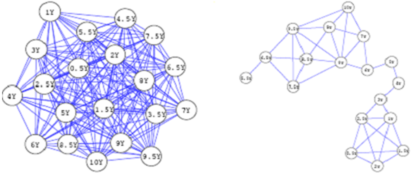
Lack of edge : conditional independence

Gaussian Markov Networks

- $P(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$
- Zeros in Σ: marginal independence
- Zeros in Σ^{-1} : conditional independence : Sparsity
- Sparsity
  - Interpretation
  - Prediction

Maximum Likelihood Estimation of the Inverse Covariance

- $\Sigma^{-1} = \arg\max_{C} \left(\log\det C - Tr(CS)\right)$
  $\Sigma^{-1} = S^{-1}$



[borrowed from A. dAspremont presentation]

- How do we make Σ^{-1} Sparse : Penalize the likelihood

Penalized Likelihood Estimation

$$\Sigma^{-1} = \arg\max_{C} \left(\log\det C - Tr(CS) - \lambda Card(C)\right)$$

$Card(C)$ =number of nonzero elements of C

- $\lambda = 2/(N+1)$ for AIC
- $\lambda = \log(N+1)/(N+1)$ for BIC
- This is a NP-Hard combinatorial problem

L-1 Regularized Likelihood Estimation

- $\Sigma^{-1} = \arg\max_{C} \log(P(X))$
  Subject to $||C||_1 \leq s$

- This is qual to solve

  $$\Sigma^{-1} = \arg\max_{C} logdetC - tr(SC) - \lambda||C||_1$$

- Convex problem with unique solution for a given λ

The Role and Choice of the Sparsity parameter: λ

- λ decides the amount of sparsity
- What is the criteria to pick the best λ?
  - Model structure recovery
  - Prediction power on test data

Learning the Sparsity Parameter λ

- Cross Validation on Training Data
  - To maximize prediction power, ignoring the structure
  - Over fit to Model, almost No Sparsity!

- Method suggested by Banerjee et. al:

$$\lambda(\alpha) = (\max \sigma_i \sigma_j) \frac{t_{n-2}(\alpha/p^2)}{\sqrt{N-2+t_{n-2}^2(\alpha/p^2)}}$$

  - Constructs back sparse Σ not the Σ^{-1}
  - Learns a Too Sparse model!
  - Weak prediction

Being Bayesian about λ

- λ as a random variable: learn its distribution
- Maximize the joint log likelihood

$$\Sigma^{-1}, \widehat{\lambda} = \arg\max_{C,\lambda} \log(P(X))$$

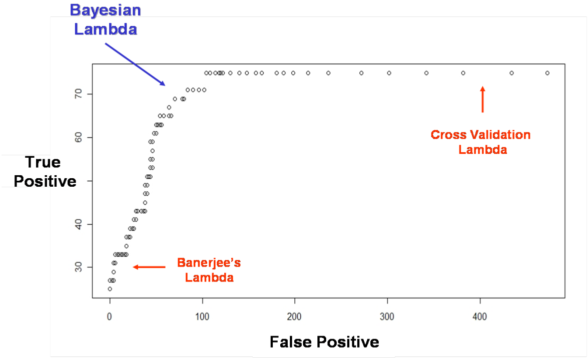
$$P(X) = P(X|C)P(C|\lambda)P(\lambda)$$

$$\log P(X) =$$
$$\log\det C - Tr(CS) - \lambda||C||_1 + P^2\log(\lambda/2) + \log P(\lambda)$$

- The choice of P(λ)

The Bayesian λ

ROC Curve

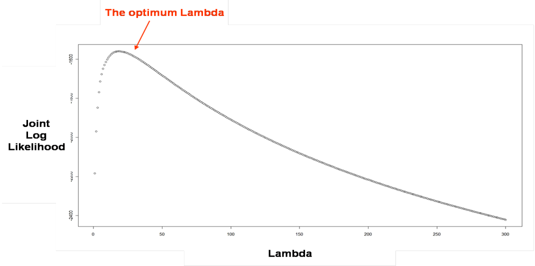


The Optimization Method

- Alternating minimization with line search
  - Estimate Σ^{-1} for an initial λ
  - Update λ in the direction of the gradient: $P^2/\lambda - ||C||_1$
  - Iterate until convergence

The Joint likelihood with Flat prior on λ

- When $N \gg P$ we get a global maximum



- But unbounded for $N \leq P$
  - Add regularization to the objective function
  - Assume a non-flat prior for λ

The Regularized likelihood with Flat Prior on λ

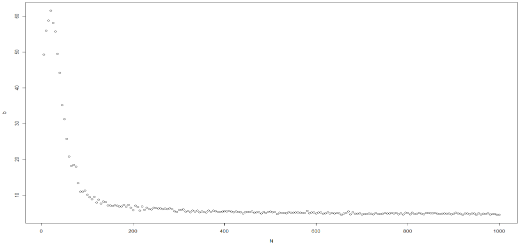
- Do not penalize the diagonal elements in the estimation of the Σ^{-1}
- Update λ as before

$$\log P(X) = \log \det C - Tr(CS) - \lambda ||C||_1 + P^2 \log(\lambda/2) + \log P(\lambda)$$

The Joint likelihood with Exponential Prior on λ

$$P(\lambda) = -b\exp(-b\lambda) \quad E(\lambda) = 1/b$$

- How to learn b from data?
- b indicates density of the empirical inverse covariance
- approximate b with $||S^{-1}||_1/P^2$ multiply it by P/N



b for P=50 as a function of N

Results with Original Density = 4% and with the Prior on λ

P	N	λ	TP	FP	Prediction Error
100	30	λ=34	108	510	1.5
		λ_b=603	2	0	3.2
		λ_c=2	294	3097	0.5
100	50	λ=50	122	516	1.4
		λ_b=693	2	0	3.2
		λ_c=1	350	5087	0.6
100	500	λ=62	328	1236	0.35
		λ_b=2510	46	136	3.2
		λ_c=0.1	356	9390	0.32
100	1000	λ=26	356	2388	0.28
		λ_b=3415	60	204	3.2
		λ_c=0.1	356	9421	0.34

Results with Original Density = 52% and with the Prior on λ

P	N	λ	TP	FP	Prediction Error
100	30	λ=32	690	616	1.9
		λ_b=1120	2	2	6.5
		λ_c=0.4	2630	2157	0.62
100	50	λ=47	694	86	1.9
		λ_b=2209	6	12	6.47
		λ_c=0.4	3225	2555	0.42
100	500	λ=24	2286	1376	0.38
		λ_b=5710	116	158	5.9
		λ_c=0.1	5089	4480	0.17
100	1000	λ=14	3465	1957	0.20
		λ_b=7691	186	240	4.3
		λ_c=0.1	5102	4623	0.15

Results with Original Density = 4% and Regularized Likelihood

P	N	λ	TP	FP	Prediction Error
100	30	λ=190	22	60	3.2
		λ_b=603	2	0	3.2
		λ_c=2	294	2976	0.48
100	50	λ=208	148	156	1.4
		λ_b=693	2	0	3.2
		λ_c=1	350	4979	0.6
100	500	λ=55	336	1132	0.33
		λ_b=2510	46	136	3.2
		λ_c=0.1	356	9390	0.32
100	1000	λ=27	356	2174	0.28
		λ_b=3415	60	204	3.2
		λ_c=0.1	356	9421	0.34

Results with Original Density = 52% and Regularized Likelihood

P	N	λ	TP	FP	Prediction Error
100	30	λ=500	44	72	6.4
		λ_b=1120	2	2	6.5
		λ_c=0.4	2630	2157	0.62
100	50	λ=500	120	156	4.2
		λ_b=2209	6	12	6.3
		λ_c=0.4	3225	2555	0.42
100	500	λ=24	2183	1304	0.35
		λ_b=5710	116	158	5.9
		λ_c=0.1	5085	4480	0.17
100	1000	λ=14	3430	1899	0.20
		λ_b=7691	186	240	4.3
		λ_c=0.1	5102	4625	0.15

The Results on fMRI data

- Test on 2007 PBAIC competition
- Filtering with correlation
- comparison with the Elastic Net Results on correlation of the estimated response with the true response on the test data
  - On the 24th response c=0.82 with 127 pre-selected voxels
    Elastic Net: c=0.69 with 300 pre-selected voxels
  - On the 15th response c=0.74 with 126 pre-selected voxels
    Elastic Net: c=0.66 with 300 pre-selected voxels

On-going Work

- Different λ for each variable in the network
- Learning appropriate prior for λ
- Alternative penalties for the likelihood
- Application of non-convex optimization methods : EBW