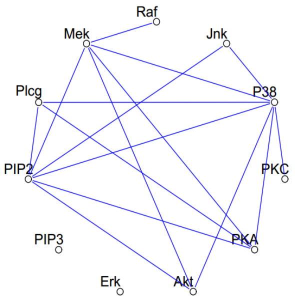
Gaussian Graphical Models

Jiang Guo

Januray 9th, 2013

Contents

Undirected Graphical Model



- Markov Property
  - pairwise

    $$X_u \perp\!\!\!\perp X_v | X_{V \setminus \{u,v\}} \text{ if } \{u,v\} \notin E$$

  - local
  - global
- Conditional Independence
- Partial Correlation

Gaussian Graphical Model

Multivariate Gaussian

$$X \sim \mathcal{N}_d(\xi, \Sigma)$$

- If Σ is positive definite, distribution has density on $\mathcal{R}^d$

$$f(x|\xi,\Sigma) = (2\pi)^{-d/2}(det\Omega)^{1/2}e^{-(x-\xi)^T\Omega(x-\xi)/2}$$

  where $\Omega = \Sigma^{-1}$ is the Precision matrix of the distribution.
- Marginal distribution: $X_s \sim \mathcal{N}_{d'}(\xi_s, \Sigma_{s,s})$
- Conditional distribution:

$$X_1|X_2 \sim \mathcal{N}(\xi_{1|2}, \Sigma_{1|2})$$

where: $\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2)$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Gaussian Graphical Model

Multivariate Gaussian

- Sample Covariance Matrix

$$\tilde{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \xi)(x_i - \xi)^T$$

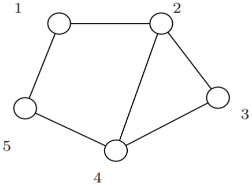
- Precision Matrix

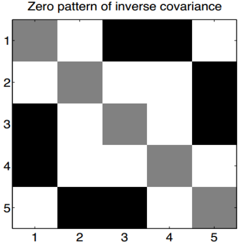
$$\Omega = \Sigma^{-1}$$

In high dimensional settings where $p \gg n$, the $\tilde{\Sigma}$ is not invertible (semi-positive definite).

Gaussian Graphical Model

- Every Multivariate Gaussian distribution can be represented by a pairwise Gaussian Markov Random Field (GMRF)
- GMRF: Undirected graph $G = (V, E)$ with
  - vertex set $V = \{1, ..., p\}$ corresponding to random variables
  - edge set $E = \{(i, j) \in V | i \neq j, \Omega_{ij} \neq 0\}$





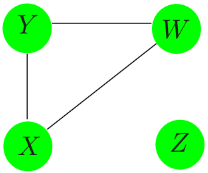
- Goal: Estimate sparse graph structuregiven $n \ll p$ iid observations.

Precision matrix estimation

Graph recovery

- also known as "Graph structure learning/estimation"
- For each pair of nodes (variables), decide whether there should be an edge
- Edge(α, β) not exists $\Leftrightarrow \alpha \perp\!\!\!\perp \beta | V \backslash \{\alpha, \beta\} \Leftrightarrow \Omega_{\alpha,\beta} = 0$
- Precision matrix is sparse

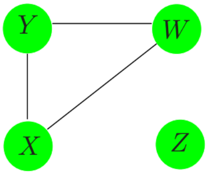
Hence, it turns out to be a non-zero patterns detection problem.



$$\begin{array}{ccccc} & x & y & z & w \\ x & 0 & 1 & 0 & 1 \\ y & 1 & 0 & 0 & 1 \\ z & 0 & 0 & 0 & 0 \\ w & 1 & 1 & 0 & 0 \end{array}$$

Precision matrix estimation

Parameter Estimation



$$\begin{array}{ccccc} & x & y & z & w \\ x & 0 & ? & 0 & ? \\ y & ? & 0 & 0 & ? \\ z & 0 & 0 & 0 & 0 \\ w & ? & ? & 0 & 0 \end{array}$$

Sparsity

The word "sparsity" has (at least) four related meanings in NLP! (Noah Smith et al.)

1. Data sparsity: N is too small to obtain a good estimate for $\mathbf{w}$. (usually bad)
2. "Probability" sparsity: most of events receive *zero* probability
3. Sparsity in the dual: Associated with SVM and other kernel-based methods.
4. Model sparsity: Most dimensions of $\mathbf{f}$ is not needed for a good h_w; those dimensions of $\mathbf{w}$ can be zero, leading to a sparse $\mathbf{w}$ (model)

We focus on sense #4.

Sparsity

Linear regression

$$f(\vec{x}) = w_0 + \sum_{i=1}^{d} w_i * x_i$$

- **sparsity** means some of the w_i(s) are zero

1. problem 1: why do we need sparse solution?
   - feature/variable selection
   - better interprete the data
   - shrinkage the size of model
   - computatioal savings
   - discourage overfitting
2. problem 2: how to achieve sparse solution?
   - solutions to come...

Sparsity

Ordinary Least Square

- Objective function to minimize:

$$\mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2$$

Sparsity

Ridge: L2 norm Regularization

$$\min \mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2 + \frac{\lambda}{2} \|\vec{w}\|_2^2$$

- equaivalent form (constrained optimization):

$$\min \mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2$$

subject to $\|\vec{w}\|_2^2 \leqslant C$

- Corresponds to zero-mean Gaussian prior $\vec{w} \sim \mathcal{N}(0, \sigma^2)$, i.e. $p(w_i) \propto exp(-\lambda \|w\|_2^2)$

Sparsity

Lasso: L1 norm Regularization

$$\min \mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2 + \frac{\lambda}{2} \|\vec{w}\|_1$$

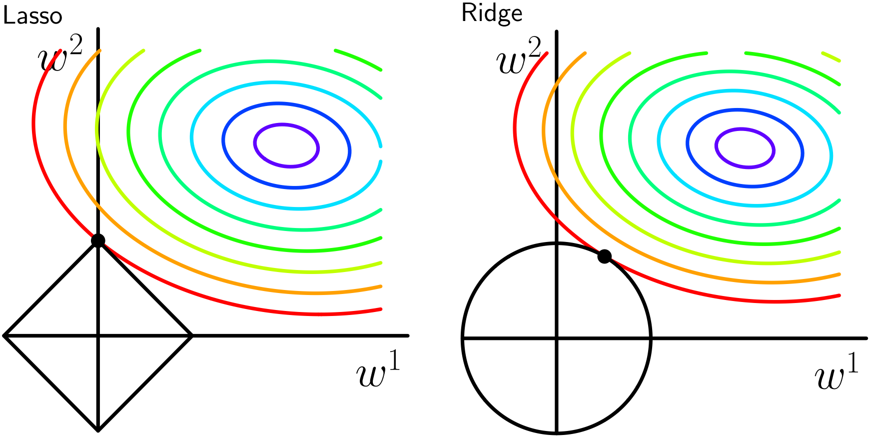
- equaivalent form (constrained optimization):

$$\min \mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2$$

$$\text{subject to } \|\vec{w}\|_1 \leqslant C$$

- Corresponds to zero-mean Laplace prior $\vec{w} \sim Laplace(0, b)$, i.e. $p(w_i) \propto exp(-\lambda |w_i|)$
- sparse solution

Why lasso sparse? an intuitive interpretation



Algorithms for the Lasso

- Subgradient methods
- interior-point methods (Boyd et al 2007)
- Least Angle RegreSsion(LARS) (Efron et al 2004), computes entire path of solutions. State of the Art until 2008.
- Pathwise Coordinate Descent (Friedman, Hastie et al 2007)
- Proximal Gradient (project gradient)

Pathwise Coordinate Descent for the Lasso

- Coordinate descent: optimize one parameter (coordinate) at a time.
- How? suppose we had only one predictor, problem is to minimize

$$\sum_i (\gamma_i - x_i\beta)^2 + \lambda|\beta|$$

- Solution is the soft-thresholded estimate

$$sign(\hat{\beta})(|\hat{\beta}| - \lambda)_+$$

where $\hat{\beta}$ is the least square solution. and

$$sign(z)(|z| - \lambda)_+ = \left\{ \begin{array}{lll} z - \lambda & \text{if} & z > 0 \text{ and } \lambda < |z| \\ z + \lambda & \text{if} & z < 0 \text{ and } \lambda < |z| \\ 0 & \text{if} & \lambda > |z| \end{array} \right\}$$

Pathwise Coordinate Descent for the Lasso

- With multiple predictors, cycle through each predictor in turn. We compute residuals $\gamma_i = y_i - \sum_{j \neq k} x_{ij}\hat{\beta}_k$ and apply univariate soft-thresholding, pretending our data is (x_{ij}, r_i)
- Start with large value for λ (high sparsity) and slowly decrease it.
- Exploits current estimation as warm start, leading to a more stable solution.

Variable selection

Lasso: L_1 regularization

$$\mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2 + \frac{\lambda}{2} \|\vec{w}\|_1$$

Variable selection

Lasso: L_1 regularization

$$\mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2 + \frac{\lambda}{2} \|\vec{w}\|_1$$

Subset selection: L_0 regularization

$$\mathcal{L} = \sum_{i=1}^{N} |y_i - f(x_i)|^2 + \frac{\lambda}{2} \|\vec{w}\|_0$$

- Greedy forward
- Greedy backward
- Greedy forward-backward (Tong Zhang, 2009 and 2011)

Greedy forward-backward algorithm

Input: $\mathbf{f}_1, \ldots, \mathbf{f}_d, \mathbf{y} \in R^n$ and $\epsilon > 0$
Output: $F^{(k)}$ and $\mathbf{w}^{(k)}$
let $F^{(0)} = \emptyset$ and $\mathbf{w}^{(0)} = 0$
let $k = 0$
while true
  let $k = k + 1$
  // forward step
  let $i^{(k)} = \arg\min_i \min_\alpha R(\mathbf{w}^{(k-1)} + \alpha \mathbf{e}_i)$
  let $F^{(k)} = \{i^{(k)}\} \cup F^{(k-1)}$
  let $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$
  let $\delta^{(k)} = R(\mathbf{w}^{(k-1)}) - R(\mathbf{w}^{(k)})$
  if $(\delta^{(k)} \leq \epsilon)$
    $k = k - 1$
    break
  endif
  // backward step (can be performed after each few forward steps)
  while true
    let $j^{(k)} = \arg\min_{j \in F^{(k)}} R(\mathbf{w}^{(k)} - \mathbf{w}_j^{(k)} \mathbf{e}_j)$
    let $\delta' = R(\mathbf{w}^{(k)} - \mathbf{w}_{j^{(k)}}^{(k)} \mathbf{e}_{j^{(k)}}) - R(\mathbf{w}^{(k)})$
    if $(\delta' > 0.5\delta^{(k)})$ **break**
    let $k = k - 1$
    let $F^{(k)} = F^{(k+1)} - \{j^{(k+1)}\}$
    let $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$
  end
end

Forward Step

Backward Step

Graphical Lasso

- Recall the Gaussian graphical model(multi-variate gaussian)

$$f(x|\xi,\Sigma) = (2\pi)^{-d/2}(det\Omega)^{1/2}e^{-(x-\xi)^T\Omega(x-\xi)/2}$$

- log-likelihood

$$\log\det\Omega - trace(\hat{\Sigma}\Omega)$$

Graphical lasso (Friedman et al 2007)

- Maximize the L_1 penalized log-likelihood:

$$\log\det\Omega - trace(\hat{\Sigma}\Omega) - \lambda\|\Omega\|_1$$

- Coordinate descent

CLIME

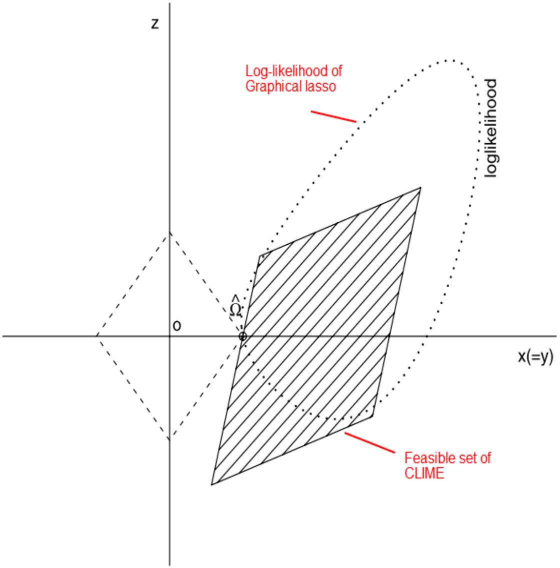
- Constrained L_1 Minimization approach to sparse precision matrix Estimation. (CAI et al 2011)
- CLIME estimator

$$\min \|\Omega\|_1 \text{ subject to:}$$
$$\|\hat{\Sigma}\Omega - I\|_\infty \leqslant \lambda_n, \Omega \in \mathcal{R}^{p \times p}$$

- Solution $\hat{\Omega}$ have to be symmetrized
- Equivalent to solving the p optimization problems:

$$\min \|\beta\|_1 \text{ subject to: } \|\hat{\Sigma}\beta - e_i\|_\infty \leqslant \lambda_n$$

CLIME


z
Log-likelihood of
Graphical lasso
loglikelihood
$\hat{\Omega}$
o
x(=y)
Feasible set of
CLIME

Gaussian Graphical Model and Column-by-Column Regression

- Consider the conditional distribution

$$X_1|X_2 \sim \mathcal{N}(\xi_{1|2}, \Sigma_{1|2})$$

$$\text{where: } \xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \xi_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

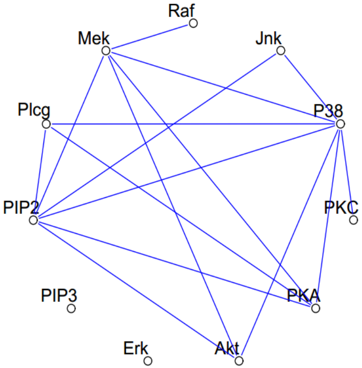
- By standardizing the data matrix, i.e. $\xi_1 = \xi_2 = 0$

$$X_1|X_2 \sim \mathcal{N}(\Sigma_{12}\Sigma_{22}^{-1}X_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Column-by-Column Regression

$$X_1 = \alpha_1^T X_2 + \epsilon_1 \text{ where } \epsilon_1 \sim \mathcal{N}(0, \sigma_1)$$

Column-by-Column Regression



- node-by-node neighbor detection
- Easy to implement
- Easy to parallelize, scalable to large scale data

TIGER

A generalized framework for supervised learning

$$\hat{\beta} = \arg\min_{\beta} L(\beta; X, Y) + \Omega(\beta)$$

- A Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models (Han et al 2012)
- SQRT-Lasso

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^d} \{ \frac{1}{\sqrt{n}} \|y - X\beta\|_2 + \lambda \|\beta\|_1 \}$$

- Tuning-insensitive (good point)
- State-of-the-art

Greedy methods

- High-dimensional (Gaussian) Graphical Model Estimation Using Greedy Methods (Pradeep et al 2012)
- Rather than exploiting lasso-like methods to achieve sparse solution, we apply greedy methods to do variable selection
- **Global greedy**: treat each element of the Precision Matrix as a Variable
- **Local greedy**: Column-by-column fashion
- (Potentially) state-of-the-art

Global Greedy

- Estimate graph structure through a series of forward and backward stagewise steps
- Forward Step: Choose "best" new edge and add to current estimate, as long as decrease in loss δ exceeds stopping criterion.
- Backward Step: Choose "weakest" current edge and remove if increase in loss is $<$ product of backward step factor and decrease in loss due to previous forward step ($\nu\delta$).
- "Best" and "Weakest" determined by sample-based Gaussian MLE

$$\mathcal{L}(\Omega) = \log \det \Omega - trace(\hat{\Sigma}\Omega)$$

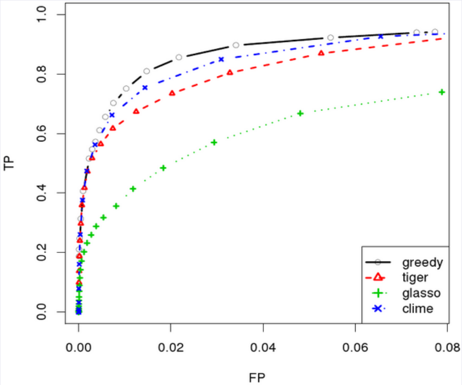
Local Greedy

- Estimates each node's neighborhood in parallel using a series of forward and backward steps
- Forward Step: Choose "best" new edge and add to current estimate, as long as decrease in loss δ exceeds stopping criterion.
- Backward Step: Choose "weakest" current edge and remove if increase in loss is $<$ product of backward step factor and decrease in loss due to previous forward step ($\nu\delta$).

- "Best" and "Weakest" determined by least-square loss

Measure methods

Graph recovery: ROC Curves

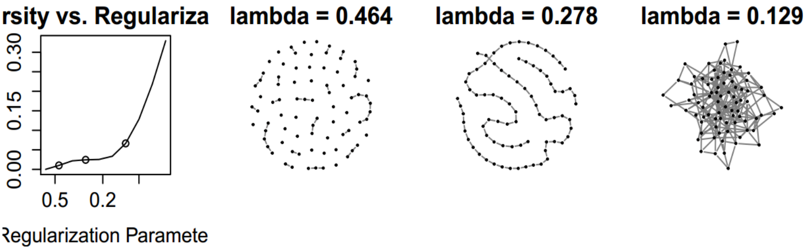


- TP: True positive rate
- FP: False positve rate

Graph recovery: ROC Curves

Decrease the regularization parameter gradually to achieve the solution path

An illustration:



Measure methods

Parameter estimation: norm error

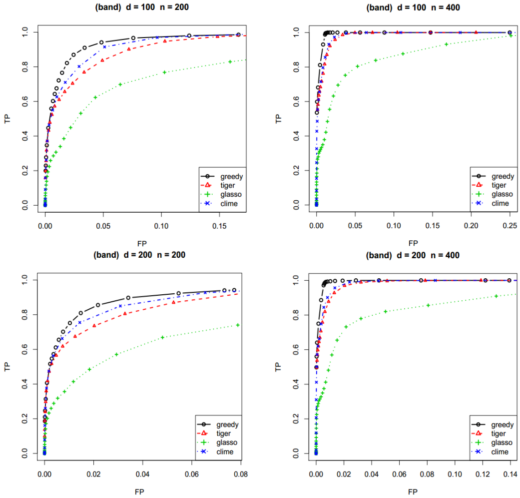
- Frobenius norm error: $\|\hat{\Omega} - \Omega\|_F$

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}^2|}$$

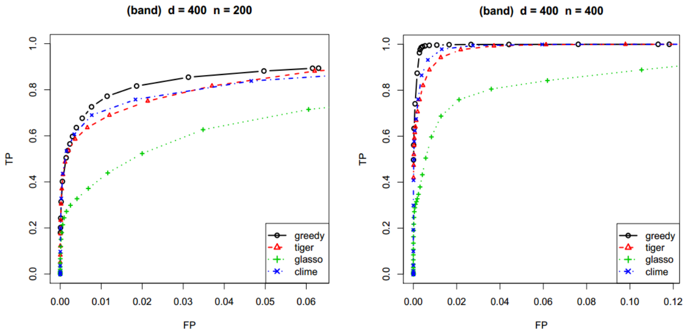
- Spectrum norm error: $\|\hat{\Omega} - \Omega\|_2$

$$\|A\|_2 = \sqrt{\lambda_{max}(A * A)} = \sigma_{max}(A)$$

Performance: Greedy, TIGER, Clime, Glasso



Performance: Greedy, TIGER, Clime, Glasso



non-gaussian scenario

Question?

In a general graph, whether a relationship exists between **conditional independence** and the structure of the **precision matrix**?

non-gaussian scenario

Question?

In a general graph, whether a relationship exists between **conditional independence** and the structure of the **precision matrix**?

Remain unsolved. Recently, there are some progresses:

- High dimensional semiparametric Gaussian copula graphical models (Han et al. 2012)
- The nonparanormal: Semiparametric estimation of high dimensional undirected graphs (Han et al. 2009)

non-gaussian scenario

Question?

In a general graph, whether a relationship exists between **conditional independence** and the structure of the **precision matrix**?

Remain unsolved. Recently, there are some progresses:

- High dimensional semiparametric Gaussian copula graphical models (Han et al. 2012)
- The nonparanormal: Semiparametric estimation of high dimensional undirected graphs (Han et al. 2009)

- Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses (*NIPS 2012 Outstanding Student Paper Awards*)

Applications

- Biomatics

Applications

- Biomatics
- Social media

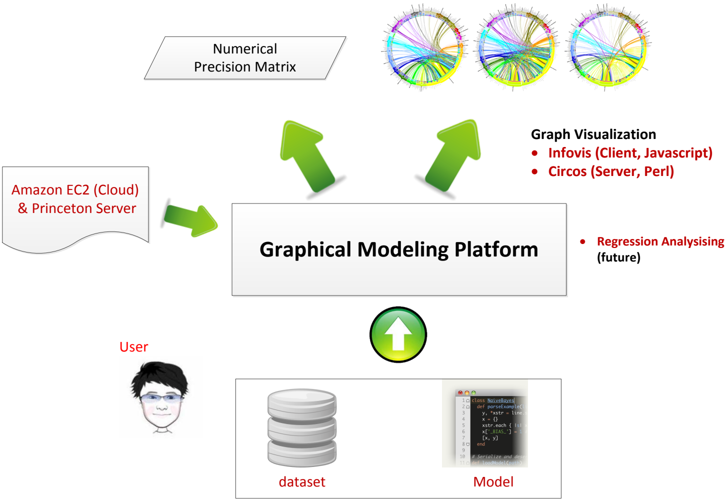
Applications

- Biomatics
- Social media
- NLP?

Applications

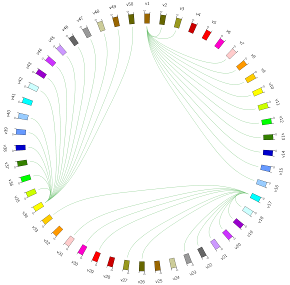
- Biomatics
- Social media
- NLP?

Project: a graphical modeling platform

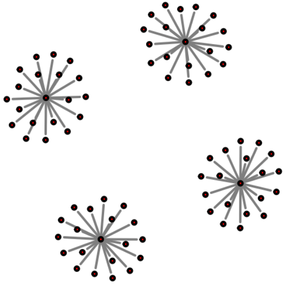


Project: a graphical modeling platform

- My visualization(1):



- R Visualization:



Arguable points

- It is student who should push supervisor, rather than the superviser push student
- Work extremely hard, blindly trust your supervisor
- Keep quality first, quantity will follow
- Science is about problems, not equations.
- Being focused.
- Think less, do more.
- Open mind.