

Network-based Penalized Regression with Application to eQTL Analysis

Wei Pan¹

(joint work with Benhuai Xie¹, Xiaotong Shen²)

¹Division of Biostatistics, School of Public Health

²School of Statistics

University of Minnesota

Beijing, China

June 2008

Outline

- Introduction: eQTL problem
- Review: Existing penalized methods
- New method
- Numerical Results: simulated and real data
- Discussion

Introduction

- Problem: expression quantitative trait loci (eQTL) mapping

$$Y_g = X\beta_g + \epsilon_g, \quad E(\epsilon_g) = 0, \quad (1)$$

for $g = 1, \dots, G$.

X : DNA markers; obs (Y_1, \dots, Y_G, X) .

Q: which markers are associated with Y_g ?

\implies variable selection or ...

- Typical approaches:
Gene-by-gene, separately,
- BUT, genes are related...
e.g. as described by a network:

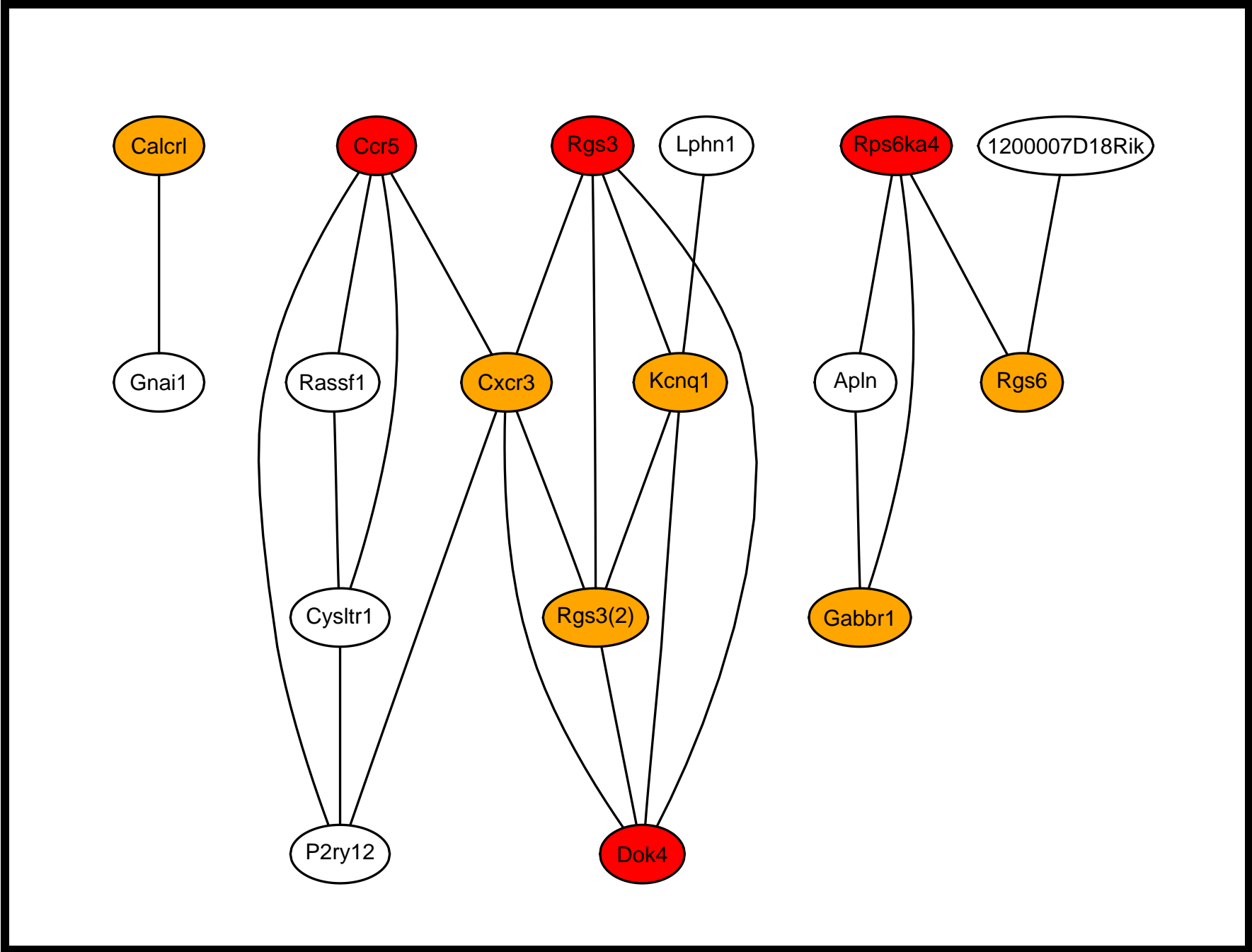


Figure 1:

$\implies Y'_g$ s are correlated, and more likely to be co-regulated!

- Network assumption/prior: if two genes $g \sim h$ in a network, then $|\beta_g| \approx |\beta_h|$.
- Goal: utilize the above assumption/prior.
- How?
- Reformulate the original multiple regressions to a single regression:

$$Y_c = (Y'_1, \dots, Y'_G)',$$

$$X_c = \text{diag}(X, \dots, X),$$

$$\beta = (\beta'_1, \dots, \beta'_G)',$$

$$Y = X\beta + \epsilon, \quad E(\epsilon) = 0, \quad (2)$$

- Q: how to incorporate the network prior into a single regression?

From now on, only need to consider a single regression.

- Bayesian or penalization: PLSE

$$\hat{\beta} = \arg \min_{\beta} L(\beta) + p_{\lambda}(\beta),$$

- Lasso (Tibshirani 1996):

$$p_{\lambda}(\beta) = \lambda \sum_{k=1}^p |\beta_k|.$$

Feature: variable selection; since $\hat{\beta}_k = 0$.

But ...

- network-based penalty of Li and Li (2008):

$$p_{\lambda}(\beta) = \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i \sim j} \left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_j}{\sqrt{d_j}} \right)^2, \quad (3)$$

d_i : degrees of node i ;

Feature: two λ 's and two terms for diff purposes ...

New Method

- New penalty: need to specify λ and w_i 's,

$$p_\lambda(\beta; \gamma, w) = \lambda 2^{1/\gamma'} \sum_{i \sim j} \left(\frac{|\beta_i|^\gamma}{w_i} + \frac{|\beta_j|^\gamma}{w_j} \right)^{1/\gamma} \quad (4)$$

- γ :
a larger γ smooths more;
 $\gamma = \infty$ maximally forces $\hat{\beta}_i = \hat{\beta}_j$ if $i \sim j$!
- w_i : smooth what?
 - 1) $w_i = d_i^{(\gamma+1)/2}$: smooth $|\beta_i|/\sqrt{d_i}$, as in Li and Li;
 - 2) $w_i = d_i$: smooth $|\beta_i|$Some theory under simplified cases.
- Feature: each term is an L_γ norm
 \implies **grouped** variable selection!
 \implies tend to realize $\hat{\beta}_i = \hat{\beta}_j = 0$ if $i \sim j$!

see Yuan and Lin 2006; Zhao et al 2007

Corollary 1 *Assume that $X'X = I$. For any edge $i \sim j$, a sufficient condition for $\hat{\beta}_i = \hat{\beta}_j = 0$ is*

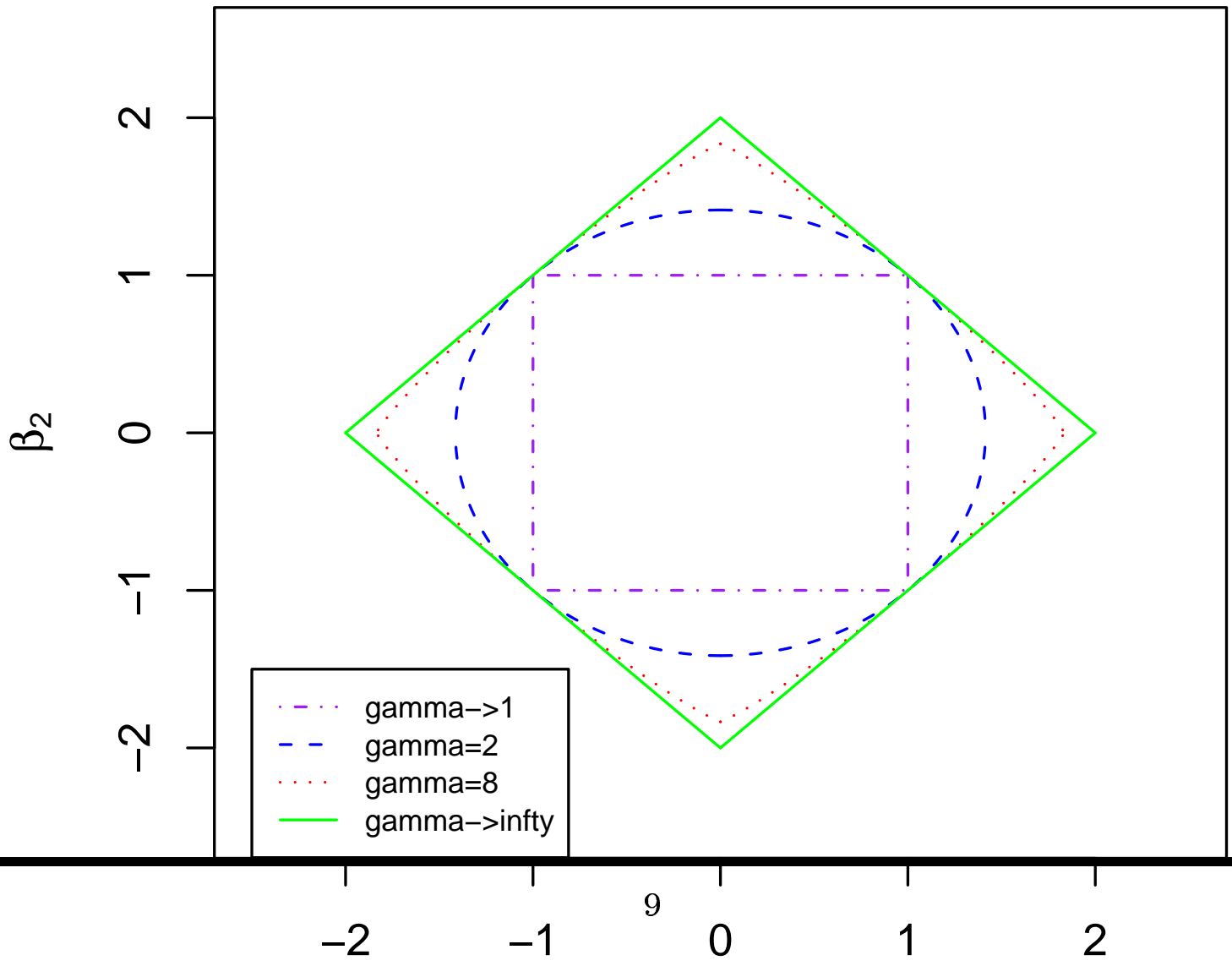
$$\|(\tilde{\beta}_i, \tilde{\beta}_j)\|_{\gamma'}^{(1/w_i, 1/w_j)} \leq \lambda 2^{1/\gamma'}, \quad (5)$$

and a necessary condition is

$$\|(\tilde{\beta}_i, \tilde{\beta}_j)\|_{\gamma'}^{(1/w_i, 1/w_j)} \leq \lambda 2^{1/\gamma'} + d_i + d_j - 2, \quad (6)$$

where $(\tilde{\beta}_i, \tilde{\beta}_j)$ are LSEs.

\implies effects of γ : a larger γ , stronger grouped variable selection!



- Other theoretical results (under simplified conditions): shrinkage effects, grouping effects ...
- Computational algorithm:
Generalized boosted lasso (GBL) (Zhao and Yu 2004);
provided *approximate* solution paths.
- Use CV to choose tuning parameter λ .

Results

- Real data: mouse data
see Fig 1
- Simulation: mimicking the mouse data
 q_0 : # false positives;
 q_1 : # true positives.

Case	Gene	Rps6ka4		120007D18Rik		Apln	
		q_0	q_1	q_0	q_1	q_0	q_1
	True	0	2	0	1	0	2
1	Lars	9.58	1.98	7.48	0.97	10.23	1.98
	Net	7.06	2.00	7.59	0.99	8.62	2.00
	<i>P</i>	< .0001	.1583	.6999	.1583	.0003	.1583
2	Lars	8.23	1.52	6.62	0.77	8.94	1.66
	Net	6.79	1.93	6.98	0.88	8.31	1.91
	<i>P</i>	< .0001	< .0001	.2603	.0040	.1468	< .0001
3	Lars	6.35	1.06	5.02	0.45	6.86	1.05
	Net	6.69	1.55	6.44	0.68	7.50	1.61
	<i>P</i>	.2527	< .0001	.0002	< .0001	.0943	< .0001
4	Lars	4.99	0.70	4.10	0.21	5.23	0.67
	Net	5.81	1.11	5.25	0.37	6.06	1.08
	<i>P</i>	.0035	< .0001	.0002	< .0001	.0131	< .0001

Discussion

- Development of more efficient and accurate algorithms (with X Shen);
- Extending to SVM (with a student);
- Relaxing the smoothness assumption (with a student):
New assumption: neighboring genes are more likely to participate or not participate at the same time; no assumption on the smoothness of regression coefficients.
Extending MRF approaches in DE gene analysis (Wei and Li 2007; Wei and Pan 2008)
.....

Acknowledgement: This research was supported by NIH.

You can download our papers from
<http://www.biostat.umn.edu/rrs.php>

Thank you!