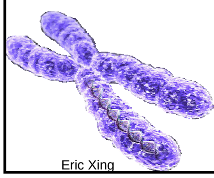
Advanced Machine Learning

Applications: Computational Biology

Eric Xing

Lecture 19, August 14, 2008



Reading: see class homepage

Biological Data Analysis

- Dynamic, noisy, heterogeneous, high-dimensional data

- "High-resolution" inference
- Parsimonious
- Scalability
- Stability
- Sample complexity
- Confidence bound

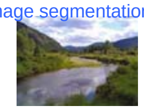
Structured Prediction Problem

- Unstructured prediction

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

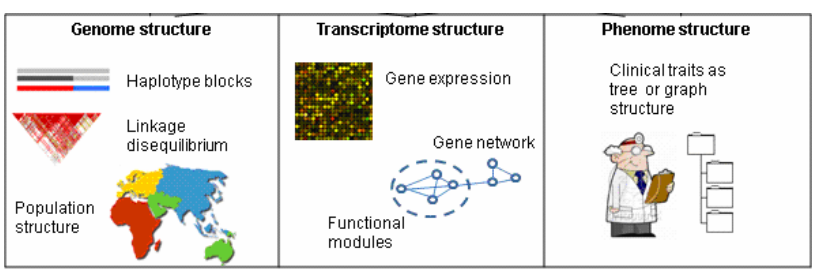
- Structured prediction
  - Part of speech tagging

    $\mathbf{x} =$ "Do you want sugar in it?" $\Rightarrow$ $\mathbf{y} =$ <verb pron verb noun prep pron>
  - Image segmentation



$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots \\ y_{21} & y_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix}$$

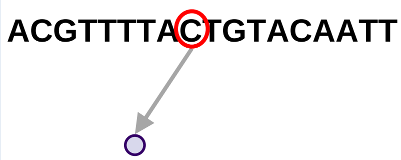
Genome and Phenome Structures



Genome-Phenome Association

Traditional view

causal SNP

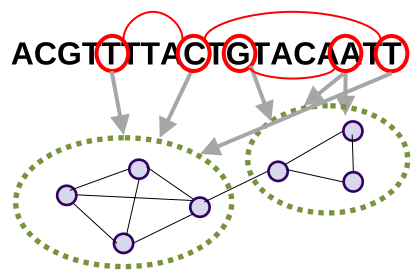


a univariate phenotype:

i.e., disease/control

Modern view

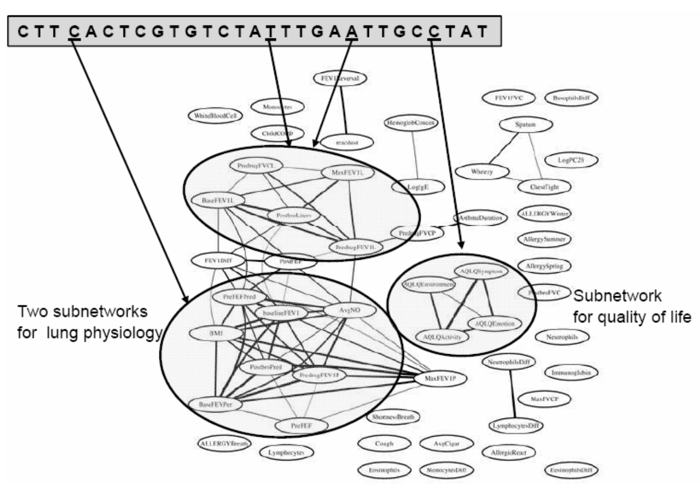
causal SNP networks



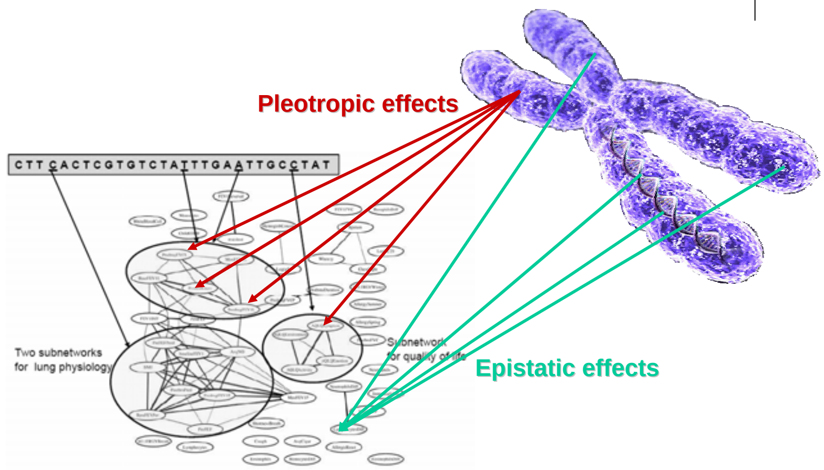
Multivariate complex syndrome (e.g., asthma):

- age at onset,
- presence/absence of eosinophilic inflammation,
- history of eczema
- genome-wide expression profile
- …

The Asthma Phenotype Network



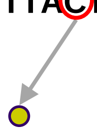
The Asthma Phenotype Network



Inferring Genome-Phenome Association

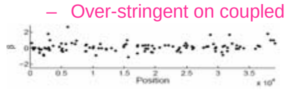
Traditional view

ACGTTTTACTGTACAATT

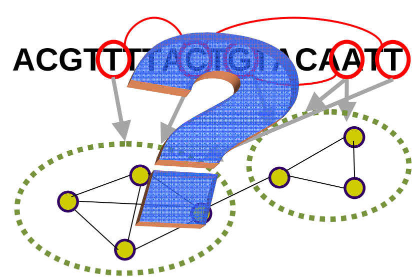


- **Pair-wise association tests?**
  - Ignore SNP dependencies
  - Many many FPs

- **Regression?**
  - Over-stringent on coupled SNPs



Modern view



- **Structured regularized regression**
  - ✓ explicitly capture structures
  - ✓ efficient
  - ✓ sparse (parsimonious)
  - ✓ provable guarantees

Association Mapping

	Phenotype (BMI)	Genotype
Individual 1	2.5	. . C . . . . . T . . C . . . . T . . . . . C . . . . . A . . C . . . . T . . .
Individual 2	4.8	. . G . . . . . A . . G . . . . A . . . . . C . . . . . T . . C . . . . T . . .
⋮		
Individual N	4.7	. . G . . . . . T . . C . . . . T . . . . . G . . . . . T . . G . . . . T . . .

Benign SNPs **Causal SNP**

Association Mapping as Regression

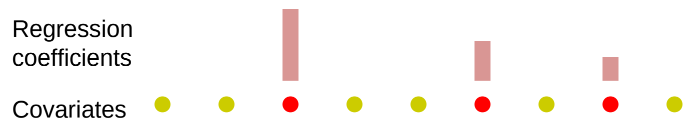
	Phenotype (BMI)	Genotype
Individual 1	2.5	. . 0 . . . . . 1 . . 0 . . . . 0 . . .
Individual 2	4.8	. . 1 . . . . . 1 . . 1 . . . . 1 . . .
⋮		
Individual N	4.7	. . 2 . . . . . 2 . . 1 . . . . 0 . . .

$$\mathbf{y}_i = \sum_{j=1}^{J} x_{ij}\beta_j$$

SNPs with large $|\beta_j|$ are relevant

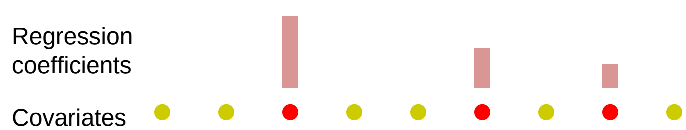
Sparse Regression

- Lasso (Tibshirani, 1996) : Learn a sparse regression model

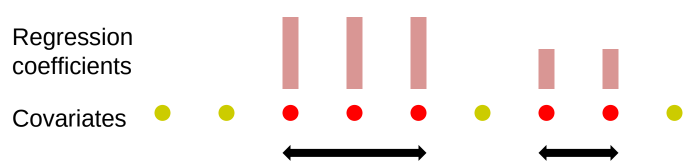


Sparse Regression

- Lasso (Tibshirani, 1996) : Learn a sparse regression model

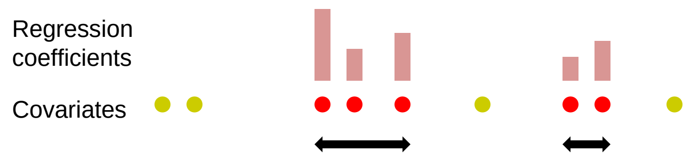


- Fused lasso (Tibshirani et al., 2005) : Fuse adjacent coefficient values, assuming covariates are ordered



Sparse Regression

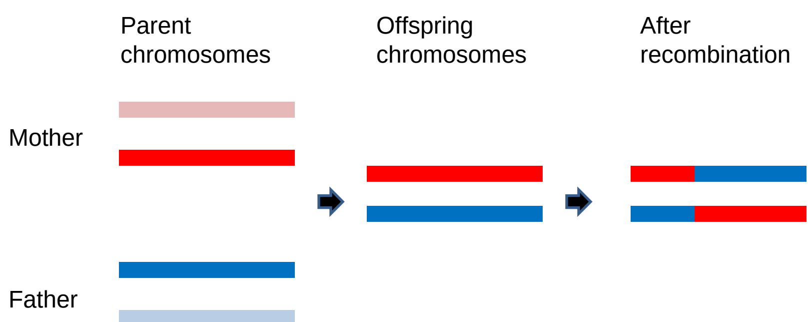
- Block-regularized regression



  - The block boundaries are determined probabilistically.

- Motivated by **association mapping** problem in computational biology
  - The block structure in genome arising from a non-random recombination process

Recombination

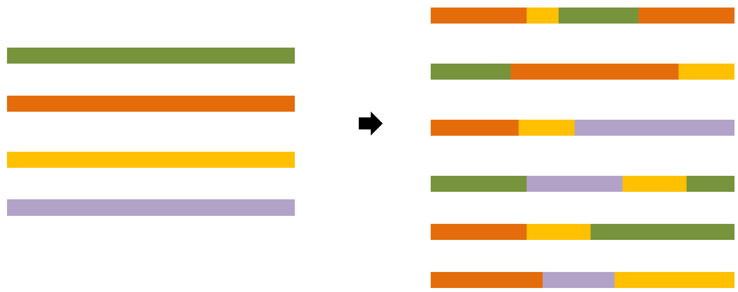


Recombination rate ρ : frequency of recombination per unit distance on chromosome (often, per kb)

After Many Generations with Recombination ...

Ancestor chromosomes

Descendent chromosomes



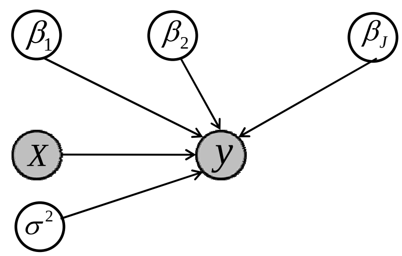
After Many Generations with Recombination ...

Ancestor chromosomes

Descendent chromosomes

Bayesian Variable Selection

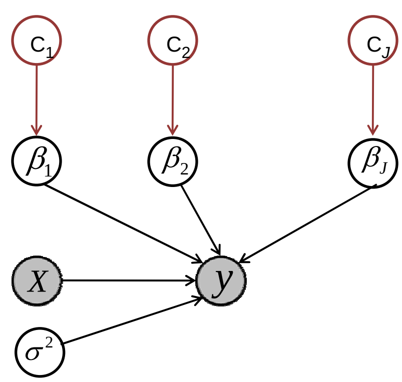
(George and McCulloch, 1993, Ishwaran and Rao, 2005)



$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Bayesian Variable Selection

(George and McCulloch, 1993, Ishwaran and Rao, 2005)



If C_j = 0 (irrelevant), $\beta_j = 0$

If C_j = 1 (relevant), use Laplacian prior

$$\beta_j | c_j \sim \frac{1}{2(2\lambda\sigma^2)} \exp\left(-\frac{|\beta_j|}{2\lambda\sigma 2}\right)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Bayesian Variable Selection

(George and McCulloch, 1993, Ishwaran and Rao, 2005)

Bernoulli prior on C_i 's

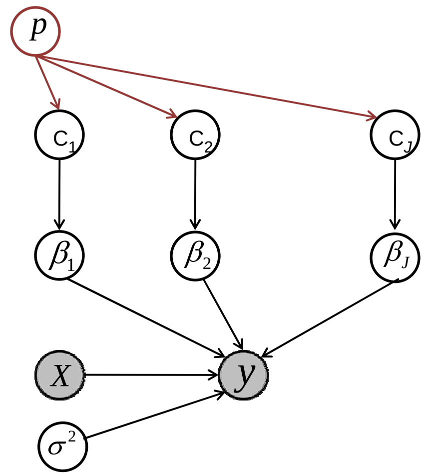
$$c_j \sim \text{Bernoulli}(p)$$

If $C_j = 0$ (irrelevant), $\beta_j = 0$

If $C_j = 1$ (relevant), use Laplacian prior

$$\beta_j | c_j \sim \frac{1}{2(2\lambda\sigma^2)} \exp\left(-\frac{|\beta_j|}{2\lambda\sigma 2}\right)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$



Markov Chain Prior

$$P(\mathbf{c}) = P(c_1) \prod_{j=2}^{J} P(c_j | c_{j-1})$$

Markov Chain Prior

$$P(\mathbf{c}) = P(c_1) \prod_{j=2}^{J} P(c_j | c_{j-1})$$

- $c_j = c_{j-1}$ if
  1) the distance between the two SNPs is small, or
  2) the recombination rate between the two SNPs is small

Markov Chain Prior

$$P(\mathbf{c}) = P(c_1) \prod_{j=2}^{J} P(c_j | c_{j-1})$$

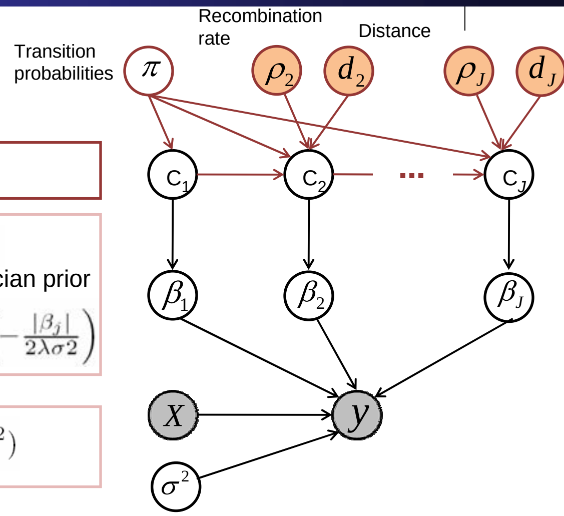
Poisson process

$$P(c_j|c_{j-1}) = \exp(-d_j \rho_j)\ \delta(c_j, c_{j-1}) + (1 - \exp(-d_j \rho_j))\ \Pi_{c_{j-1}, c_j}$$

- ρ_j : Recombination rate at *j*th SNP
- d_j : Distance between *j*th and (*j*-1)th SNP
- Π : Transition probability matrix

$$\begin{pmatrix} \pi_0 & 1-\pi_0 \\ 1-\pi_1 & \pi_1 \end{pmatrix}$$

Block-regularized Regression with Markov Chain Prior



Markov chain prior on C_j 's

If C_j = 0 (irrelevant), $\beta_j = 0$
If C_j = 1 (relevant), use Laplacian prior

$$\beta_j | c_j \sim \frac{1}{2(2\lambda\sigma^2)} \exp\left(-\frac{|\beta_j|}{2\lambda\sigma 2}\right)$$

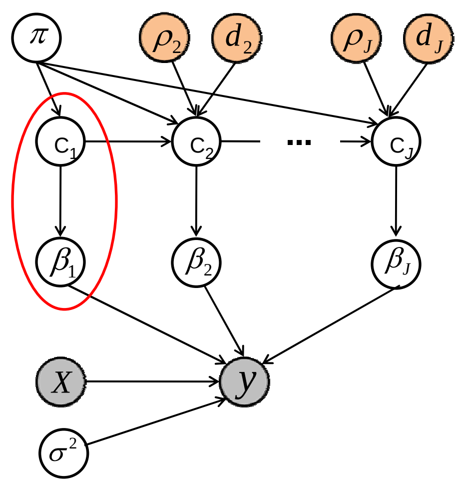
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Learning with MCMC

- In each iteration
  - Sample (C_j, β_j)'s

$$p(\beta_j, c_j | \boldsymbol{\beta}_{-j}, \mathbf{c}_{-j}, \mathbf{y}, \mathbf{X}, \sigma^2)$$
$$= p(\beta_j | \boldsymbol{\beta}_{-j}, \mathbf{c}, \mathbf{y}, \mathbf{X}, \sigma^2)$$
$$\cdot P(c_j | \boldsymbol{\beta}_{-j}, \mathbf{c}_{-j}, \mathbf{y}, \mathbf{X}, \sigma^2)$$

  - Sample π
  - Sample σ^2

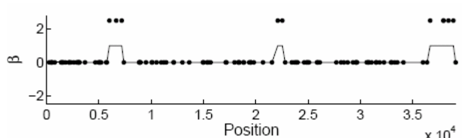


Experiments

- Simulation study
  - Comparison with
    - Bayesian variable selection with independent Bernoulli prior
    - Lasso
    - Ridge regression
  - Simulate covariates from *ms* (Hudson, 2002)
  - Estimate recombination rates using *PHASE* (Li and Stephens, 2004)
  - 10 relevant SNPs out of 100-250 SNPs
  - 180 individuals
  - MCMC sampling for 5000 iterations after 2000 burn-in
- Mouse dataset

Simulations

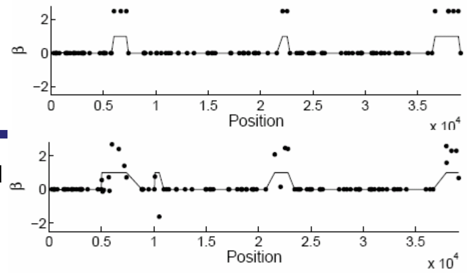
True Model



Simulations

True Model

Block-regularized regression

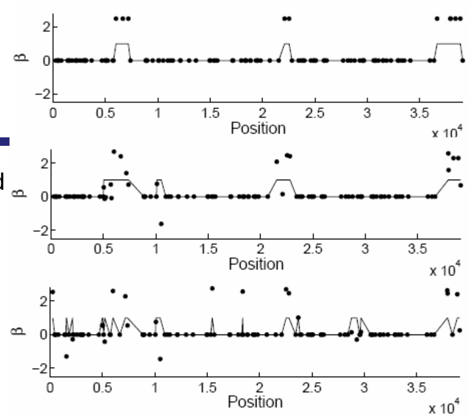


Simulations

True Model

Block-regularized regression

Independent Bernoulli prior



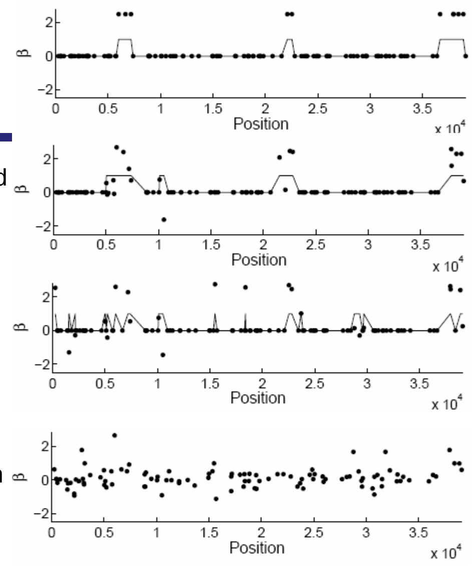
Simulations

True Model

Block-regularized regression

Independent Bernoulli prior

Ridge regression



Simulations

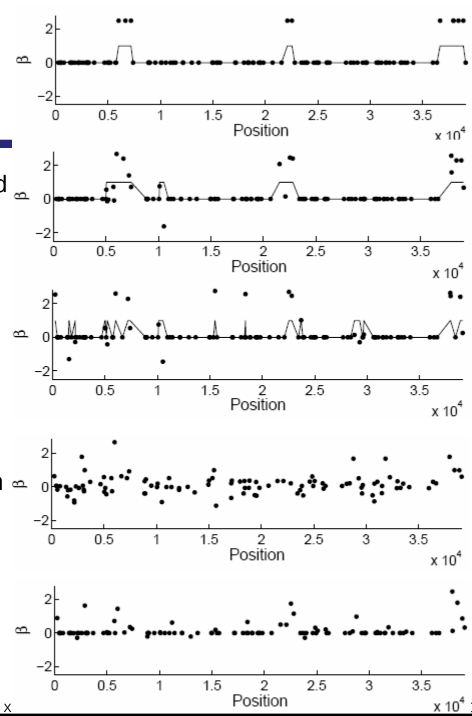
True Model

Block-regularized regression

Independent Bernoulli prior

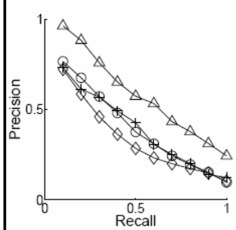
Ridge regression

Lasso



Precision and Recall

- Block
- Bernoulli
- Ridge
- Lasso

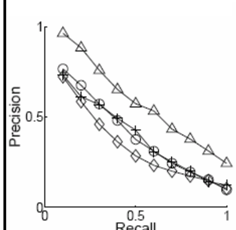


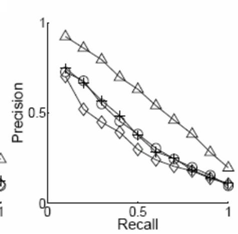
ρ = 0.05/kb

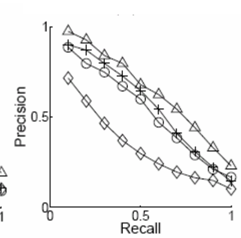
Low Recombination Rate High

Precision and Recall

- Block
- Bernoulli
- Ridge
- Lasso



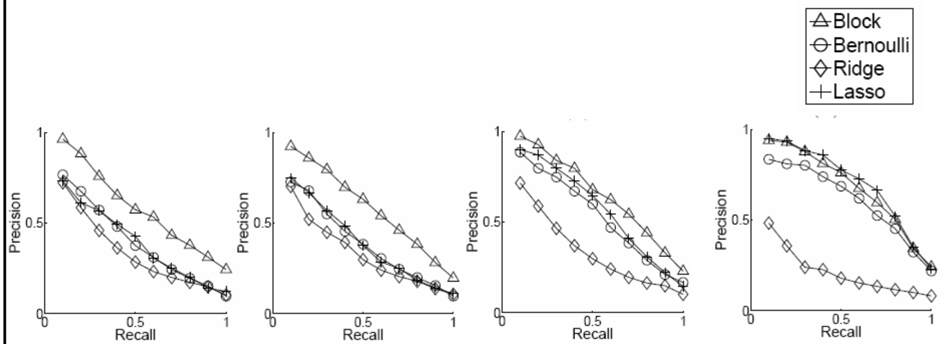




ρ = 0.05/kb ρ = 0.1/kb ρ = 0.5/kb ρ = 1.0/kb

Low Recombination Rate High

Precision and Recall

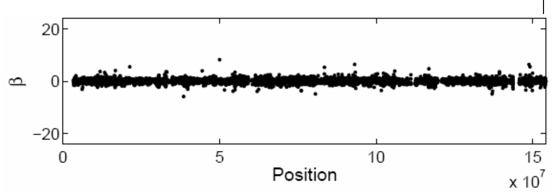


ρ = 0.05/kb ρ = 0.1/kb ρ = 0.5/kb ρ = 1.0/kb

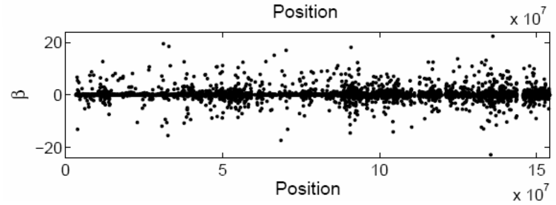
Low Recombination Rate High

Mouse Data (BROAD institute)

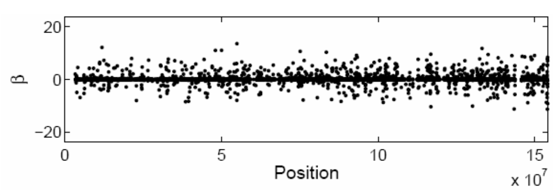
Block-regularized regression



Independent Bernoulli prior

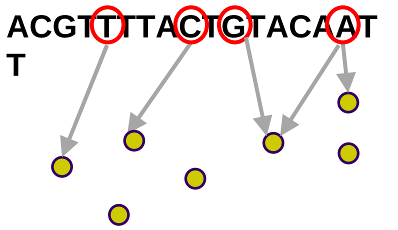


Lasso

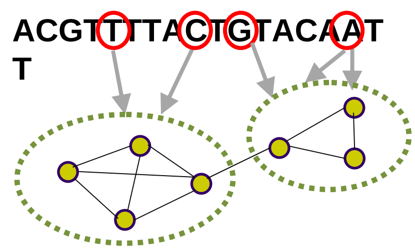


Multiple-trait Association

Simple (univariate) traits



Complex (multivariate) traits

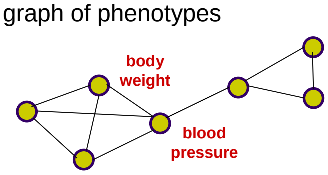


$$\hat{\mathbf{B}}^{\text{lasso}} = \text{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \lambda \sum_{k,j} |\beta_{kj}| \quad \Rightarrow$$

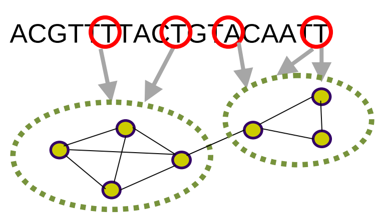
Need additional constraints on **B** apart from sparsity

Multiple-trait Association

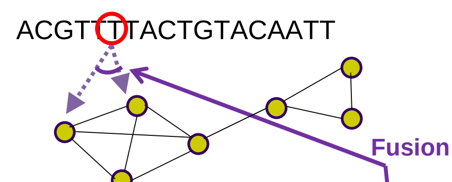
Step 1: Thresholded correlation graph of phenotypes



Overall effect of the fusion penalty



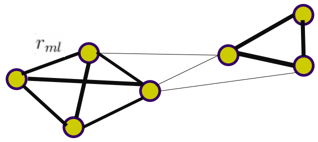
Step 2: Graph-constrained fused lasso



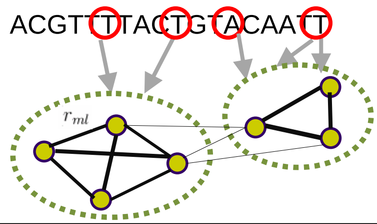
$$\hat{\mathbf{B}}^{\text{GC}} = \text{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)$$
$$+ \lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l)\in E} \sum_j |\beta_{jm} - \text{sign}(r_{ml})\beta_{jl}|$$

Multiple-trait Association

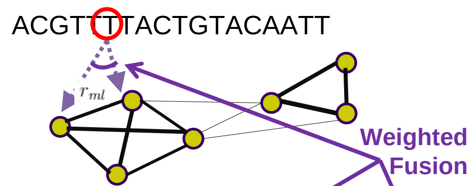
Step 1: Thresholded correlation graph of phenotypes with **weights**



Overall effect of the weighted fusion penalty



Step 2: Graph-weighted fused lasso



$$\hat{\mathbf{B}}^{\mathrm{GW}} = \operatorname{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l)\in E} f(r_{ml}) \sum_j |\beta_{jm} - \mathrm{sign}(r_{ml})\beta_{jl}|$$

Convex Optimization

- Quadratic programming formulation
  - Graph-constrained fused lasso

$$\hat{\mathbf{B}}^{\mathrm{GC}} = \operatorname{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)$$

  - Gra

$$\text{s. t.} \quad \sum_k \sum_j |\beta_{jk}| \le s_1 \text{ and } \sum_{(m,l)\in E} \sum_j |\beta_{jm} - \mathrm{sign}(r_{ml})\beta_{jl}| \le s_2$$

$$\hat{\mathbf{B}}^{\mathrm{GW}} = \operatorname{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)$$

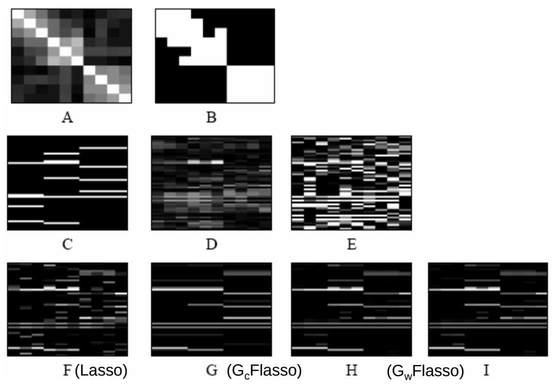
$$\text{s. t.} \quad \sum_k \sum_j |\beta_{jk}| \le s_1 \text{ and } \sum_{(m,l)\in E} f(r_{ml}) \sum_j |\beta_{jm} - \mathrm{sign}(r_{ml})\beta_{jl}| \le s_2$$

- Many ... g convex optimization problems can be used

Simulated Data

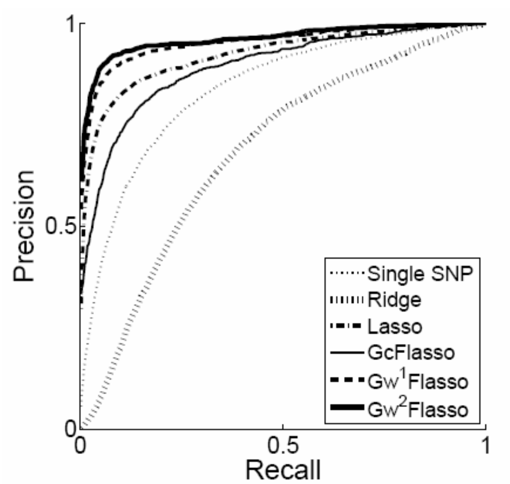
Phenotype Correlation Structure

Regression Coefficients



A: The correlation matrix of phenotypes, B: the edges of the phenotype correlation graph obtained at threshold 0.3 are shown as white pixels, C: The true regression coefficients used in simulation. Rows correspond to SNPs and columns to phenotypes. D: -log(p-value). Absolute values of the estimated regression coefficients are shown for E: ridge regression, F: lasso, G: G$_c$Flasso, H: G_1^wFlasso, and I: G_2^wFlasso.

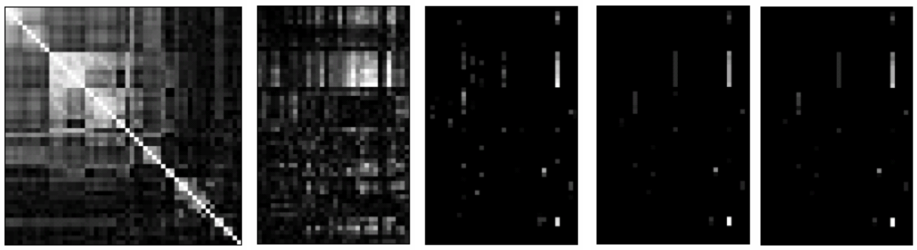
ROC on Simulated Data



The same trend is observed under:

- Different sample sizes,
- Association strength
- Edge weight cut-offs

Asthma Multiple-trait Association

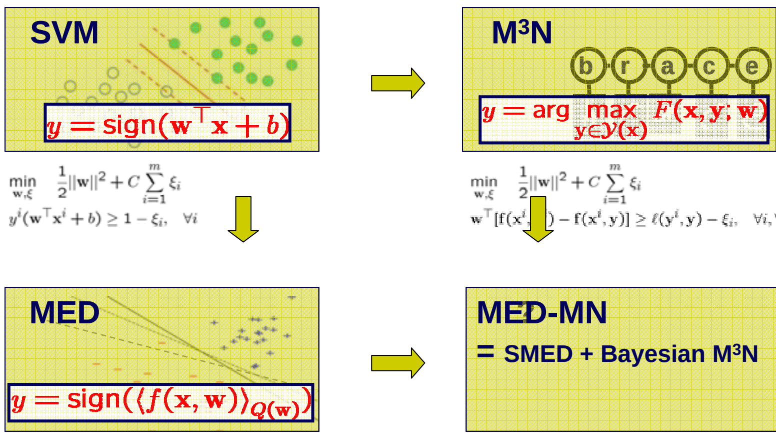


Phenotype Correlation Structure | Single-marker Single-trait test | Lasso | Graph-constrained Fused lasso | Graph-weighted Fused lasso

Summary

- Likelihood-based Structured Input or Structured Output
  - Block-regularized regression makes use of the prior knowledge on the block structure such as distance and recombination rate between adjacent SNPs.
  - Graph-guided fused lasso framework incorporates correlation information among phenotypes to detect pleiotropic effect of genotypic variations.
- Future Work
  - Combine structural information in both genome and phenome in a single statistical method

Margin-Based Discriminative Learning Paradigms



SVM:

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$y^i(\mathbf{w}^\top\mathbf{x}^i + b) \ge 1 - \xi_i, \quad \forall i$$

M3N:

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$\mathbf{w}^\top[\mathbf{f}(\mathbf{x}^i,\mathbf{y}^i) - \mathbf{f}(\mathbf{x}^i,\mathbf{y})] \ge \ell(\mathbf{y}^i,\mathbf{y}) - \xi_i, \quad \forall i, \forall \mathbf{y} \ne \mathbf{y}^i$$

MED:

$$\min_{Q} \ KL(Q\|Q_0)$$
$$y^i\langle f(\mathbf{x}^i)\rangle_Q \ge \xi_i, \quad \forall i$$

Maximum Entropy Discrimination Markov Networks

- Structured MaxEnt Discrimination (SMED):

$$\text{P1}: \quad \min_{p(\mathbf{w}),\xi} \ \boxed{KL(p(\mathbf{w})\|p_0(\mathbf{w})) + U(\xi)}$$
$$\text{s.t.} \quad p(\mathbf{w}) \in \mathcal{F}_1, \ \xi_i \ge 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

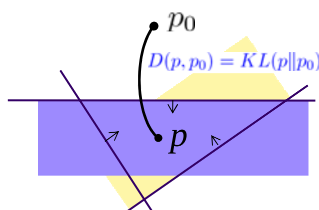
- Feasible subspace of weight distribution:

$$\mathcal{F}_1 = \left\{p(\mathbf{w}) : \boxed{\int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\,d\mathbf{w} \ge -\xi_i}, \ \forall i, \forall \mathbf{y} \ne \mathbf{y}^i\right\},$$

expected margin constraints.

- Average from distribution of M3Ns

$$h_1(\mathbf{x}; p(\mathbf{w})) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w})F(\mathbf{x},\mathbf{y};\mathbf{w})d\mathbf{w}$$



Solution to MaxEnDNet

- Theorem:
  - Posterior Distribution:

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\{\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta \ell_i(\mathbf{y})]\}$$

  - Dual Optimization Problem:

$$\begin{aligned} \mathrm{D1}: \quad & \max_{\alpha} \ -\log Z(\alpha) - U^\star(\alpha) \\ & \text{s.t.} \ \ \alpha_i(\mathbf{y}) \geq 0, \ \forall i, \ \forall \mathbf{y}, \end{aligned}$$

$U^\star(\cdot)$ is the conjugate of the $U(\cdot)$, i.e., $U^\star(\alpha) = \sup_{\xi}\left(\sum_{i,y} \alpha_i(y)\xi_i - U(\xi)\right)$

Gaussian MaxEnDNet (reduction to M³N)

- Theorem
  - Assume $F(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}), U(\xi) = C\sum_i \xi_i$, and $p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0,I)$
    - Posterior distribution: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, I), \ \text{where} \ \mu_{\mathbf{w}} = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})$
    - Dual optimization:

$$\begin{aligned} & \max_{\alpha} \ \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\|\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})\|^2 \\ & \text{s.t.} \ \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C; \ \alpha_i(\mathbf{y}) \geq 0, \ \forall i, \ \forall \mathbf{y}, \end{aligned}$$

    - Predictive rule:

$$h_1(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x},\mathbf{y};\mathbf{w})\, d\mathbf{w} = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \mu_{\mathbf{w}}^\top \mathbf{f}(\mathbf{x},\mathbf{y})$$

- Thus, MaxEnDNet subsumes M³Ns and admits all the merits of max-margin learning
- Furthermore, MaxEnDNet has at least three advantages …

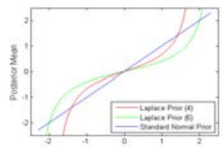
Three Advantages

- An averaging Model: PAC-Bayesian prediction error guarantee

$$\Pr_Q(M(h,\mathbf{x},\mathbf{y}) \le 0) \le \Pr_{\mathcal{D}}(M(h,\mathbf{x},\mathbf{y}) \le \gamma) + O\left(\sqrt{\frac{\gamma^{-2}KL(p\|p_0)\ln(N|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\right).$$

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M³N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M³N)

$$\forall k,\ \langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}$$



η_k

- Integration of Generative and Discriminative principles
  - Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)

Key Challenges

- Extremely high dimensionality and low data volume
  - d ~ 1M
  - N ~ 1K
  - Sample complexity with bounded error?
- Sparsity bias of the model
  - Often <100 features out of the !M are relevant
  - Regularization schemes to enforce sparsity
- Structures and hidden variables
  - Inputs and outputs often bear intricate structures (e.g., chain or graphical dependencies)
  - How to capture other latent structures between unobserved variables
- Generalizability and scalability
  - Move efficient convex opt solver and Bayesian inference algorithms
- Provable theoretical guarantees
  - Consistency and sparsistency
  - Stability, convergence rate, etc.