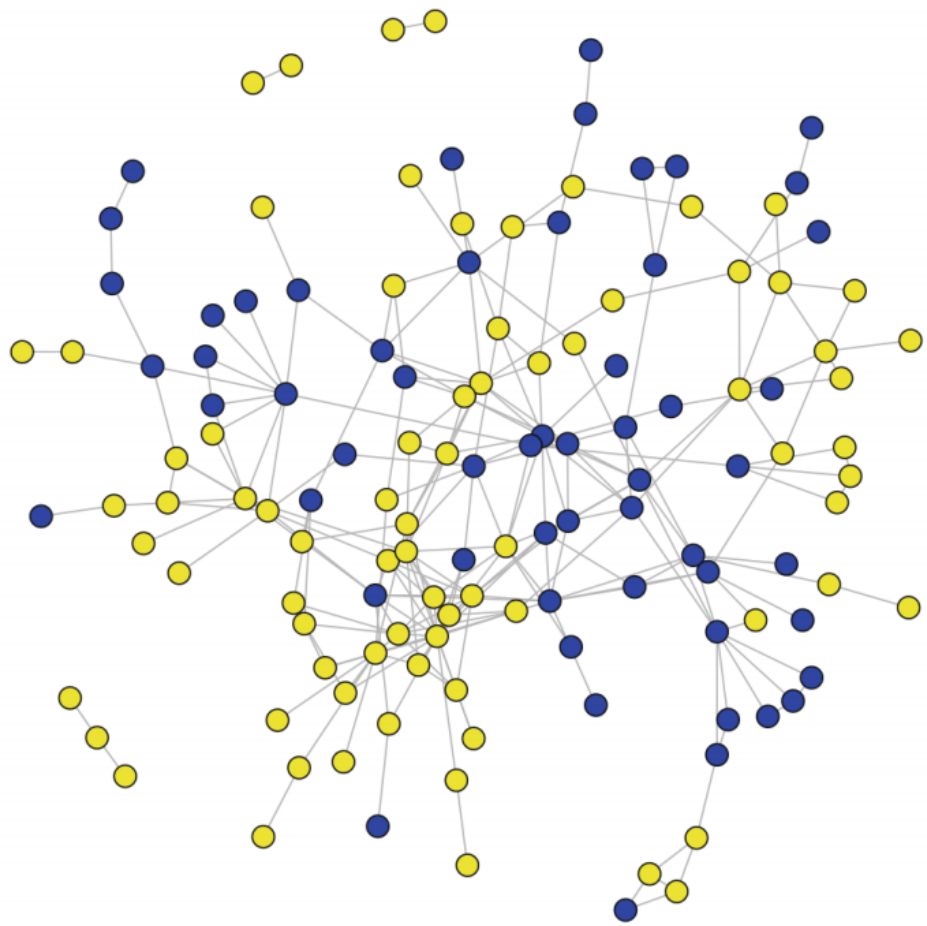
Vertex Property Prediction

Given graph structure and vertex attributes (x), predict target attribute of interest

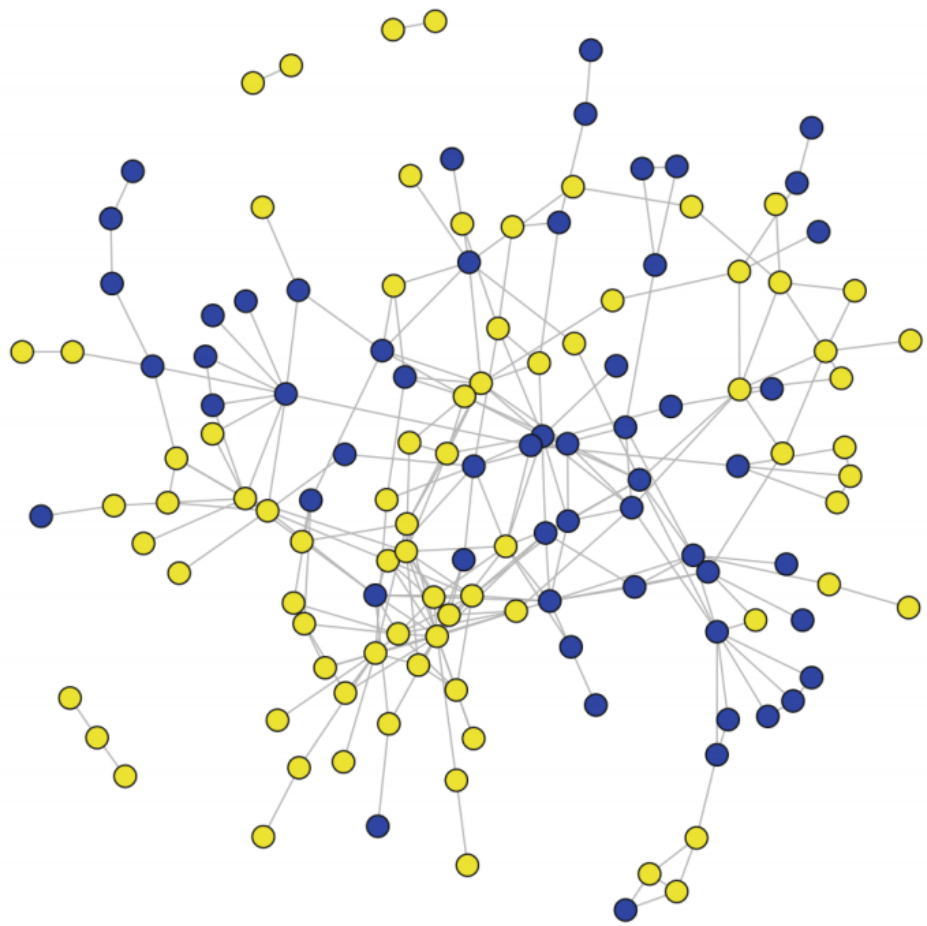
Example: protein function from a ppi network (guilt-by-association)



Vertex Property Prediction

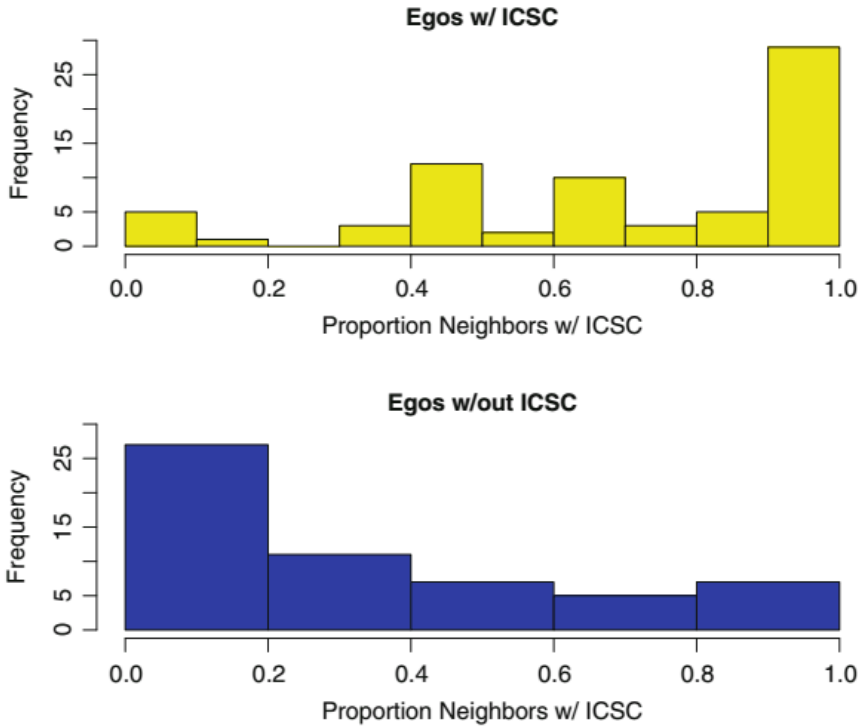
Simplest approach: use neighbors

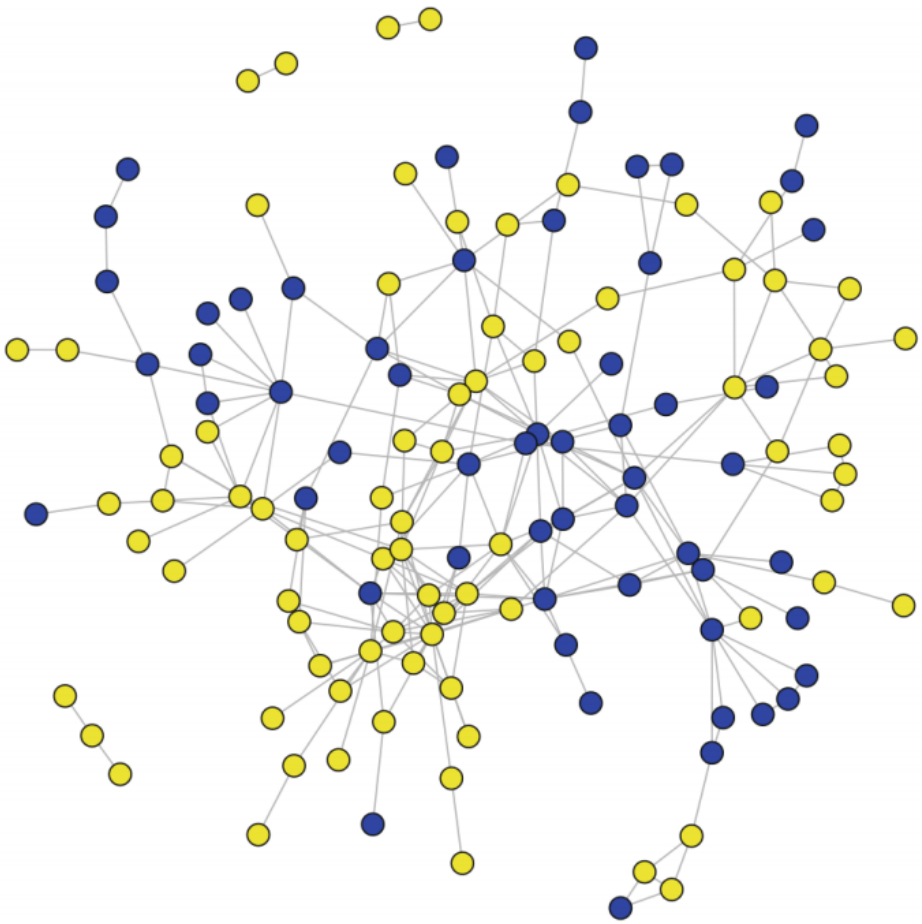
Use target values of graph neighbors:
- average for continuous
- majority class for categorical



Vertex Property Prediction

Simplest approach: use neighbors

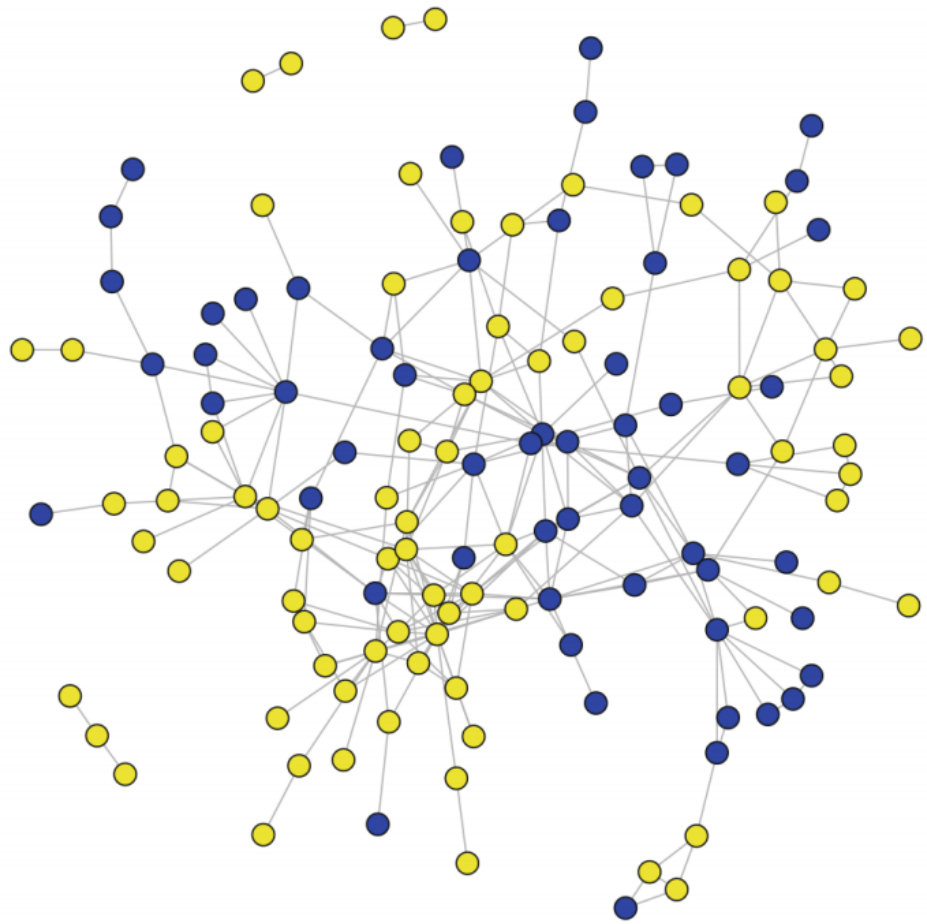




Vertex Property Prediction

Regression approach

Build a function $f(x)$ that is *smooth* over the graph



Smoothing Splines and the Beaver Dam Eye Study

- Great history of using smoothing spline (SS) models for analyzing Beaver Dam Eye Study (BDES) data [Wahba et al. 1998a, b,1999, 2000, 2002, 2006]
- In particular, smoothing spline ANOVA (SS-ANOVA) model of pigmentary abnormalities (PA)

  [Ann. Statistics 28 (2000)]

SS-ANOVA for Bernoulli

Binary outcome, disease/no disease

Predictors (bmi, sysbp...)

- Estimate o

$$p(x) = Pr\{y = 1|x\}$$

$$f(x) = \log \frac{p(x)}{1 - p(x)}$$

- *f* is nonparametric (or semiparametric)
- *f* has ANOVA-like decomposition

SS-ANOVA

- Model for pigmentary abnormalities (PA), female BDES subjects [Ann. Statistics 28 (2000)]

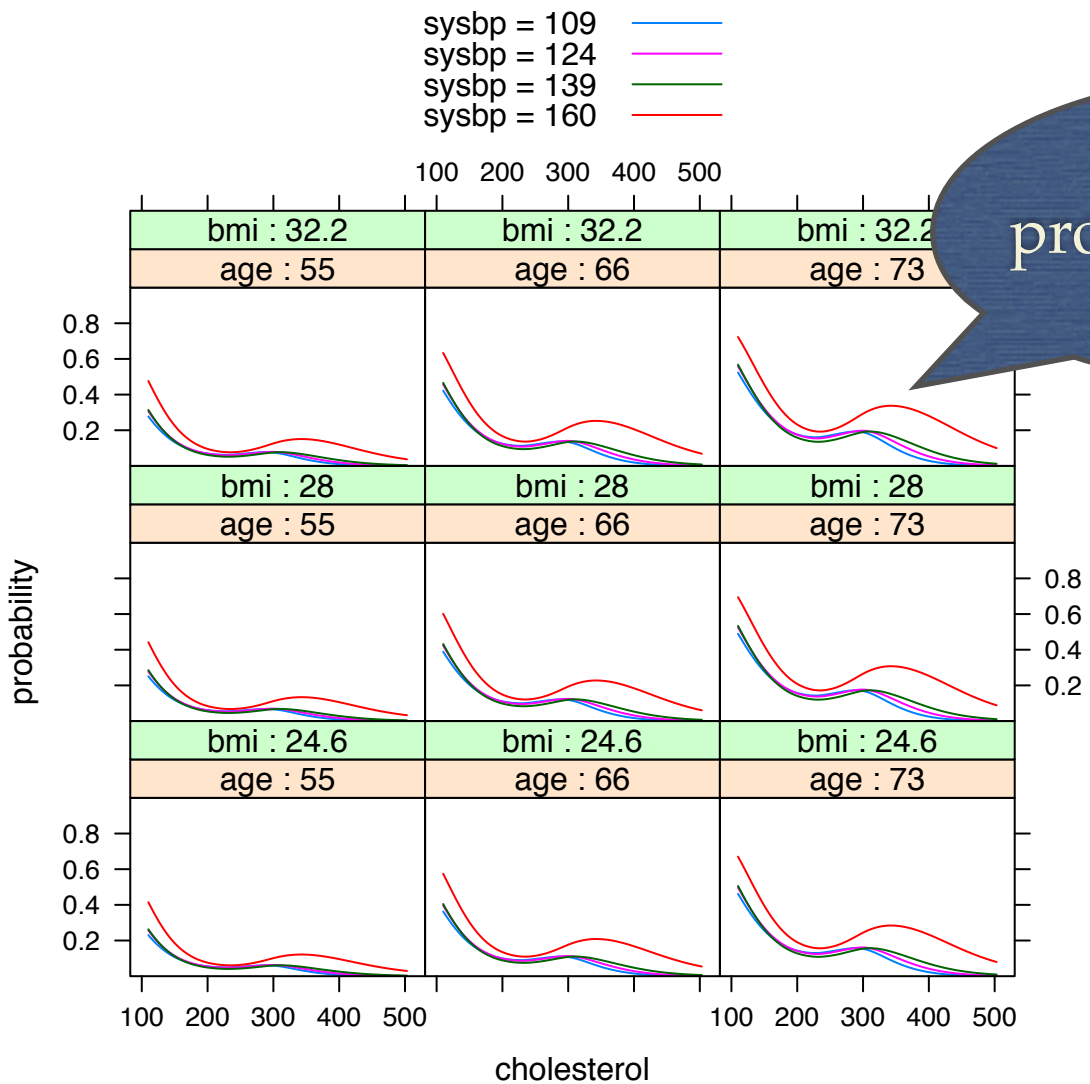
$$f(t) = \mu + f_1(\text{sysbp}) + f_2(\text{chol}) + f_{12}(\text{sysbp}, \text{chol}) + d_{\text{age}} \cdot \text{age} + d_{\text{bmi}} \cdot \text{bmi} + d_{\text{horm}} \cdot I_1(\text{horm}) + d_{\text{hist}} \cdot I_2(\text{hist}) + d_{\text{smoke}} \cdot I_3(\text{smoke}),$$

hormone replacement yes/no

history of heavy drinking

history of smoking

SS-ANOVA



nonlinear protective effect of cholesterol

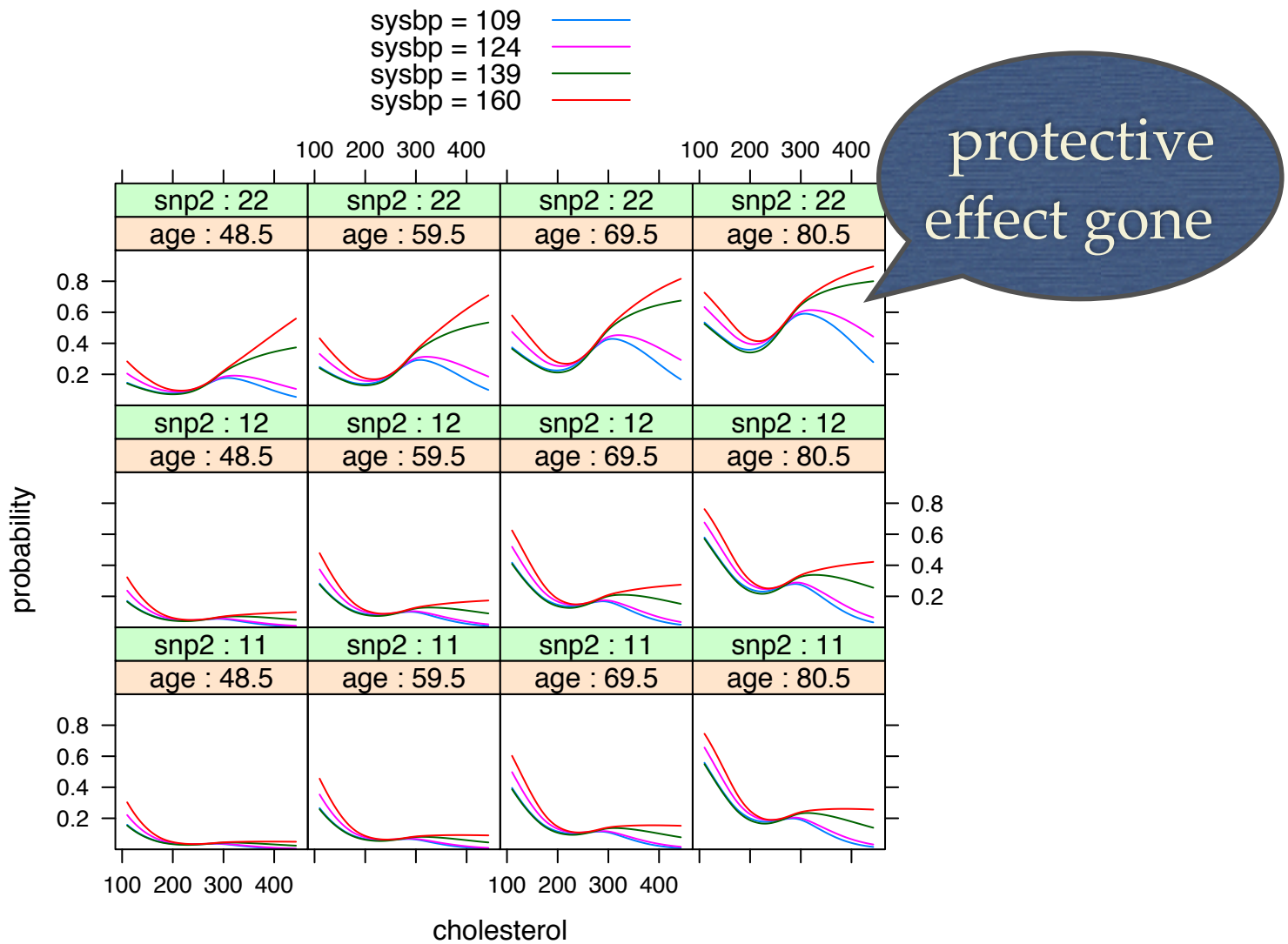
New Data Sources

1. Familial relationships were ascertained for BDES subjects
   - pedigrees later constructed for subset of subjects

New Data Sources

2. Results linking variation in specific genetic regions and AMD (age-related macular degeneration)
   - in particular, CFH and LOC387715 (ARMS2) genes
   - genetic marker data for specific SNPs, including these two gene regions, generated for subjects in pedigree data

SS-ANOVA (w/ ARMS2)



SS-ANOVA

- Goal: Extend SS-ANOVA model with genetic covariates and pedigree information
  - pedigrees are trickier
  - method: define a pedigree dissimilarity, incorporate to SS-ANOVA model

SS-ANOVA and Pedigrees

- Main idea: make use of ANOVA-like decomposition
- Term for each data type: environmental covariates (as in the original SS-ANOVA model), genetic markers, pedigree data
  - SS-ANOVA can give relative importance of each component

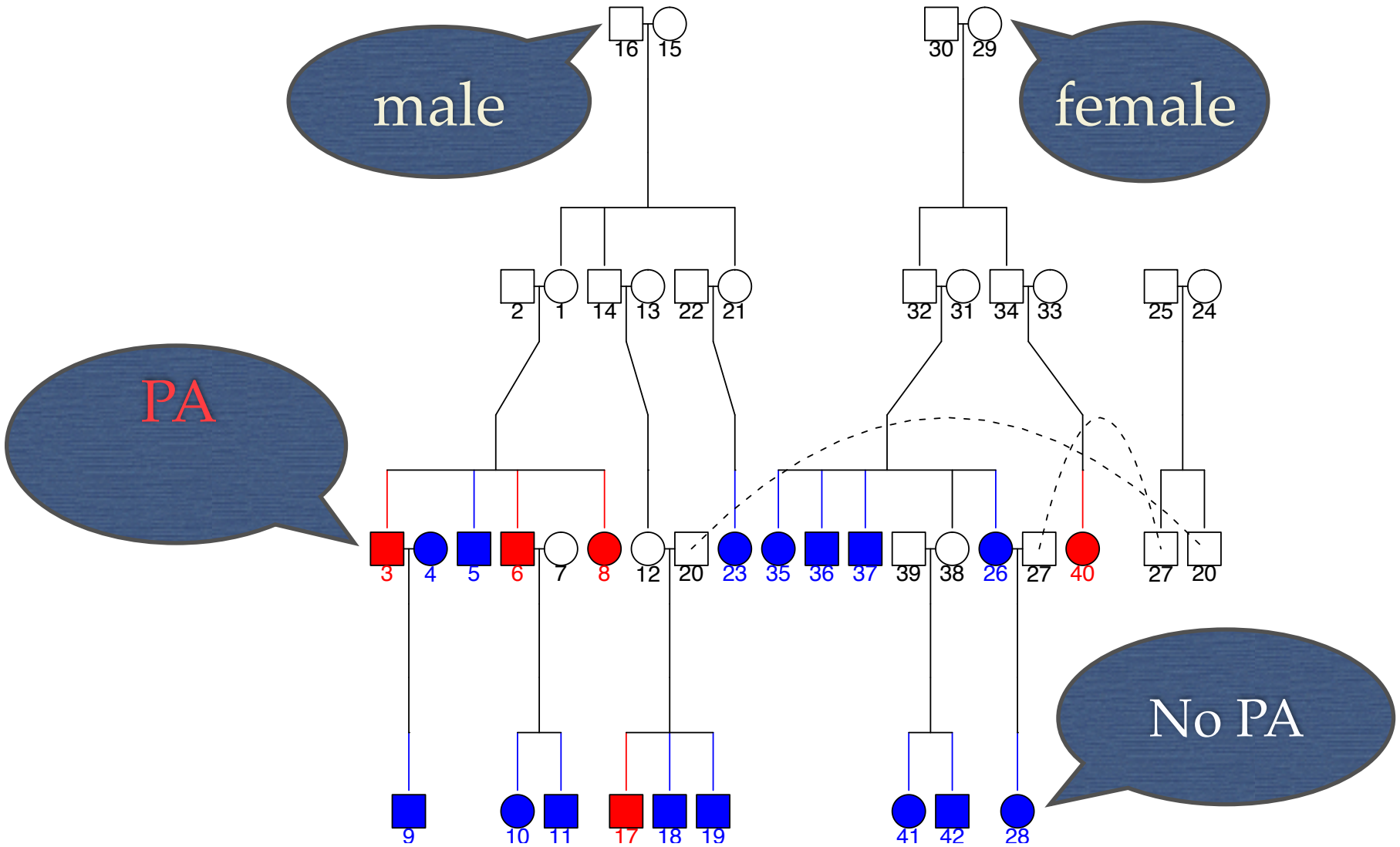
Graph-Based

- The method presented is quite general,
  - can incorporate data where relationships are represented by a graph
- Big Assumption:
  1. graph does not have determinant effect on outcome,
  2. instead, it is one of multiple, comparable, model components that affect outcome

Outline

1. Pedigrees and dissimilarities
2. SS-ANOVA models
3. Kernel representations of pedigree data
4. Case study: Beaver Dam Eye Study

Pedigrees


male
female
PA
No PA
16 15
30 29
2 1 14 13 22 21
32 31 34 33
25 24
3 4 5 6 7 8 12 20 23 35 36 37 39 38 26 27 40 27 20
9 10 11 17 18 19 41 42 28

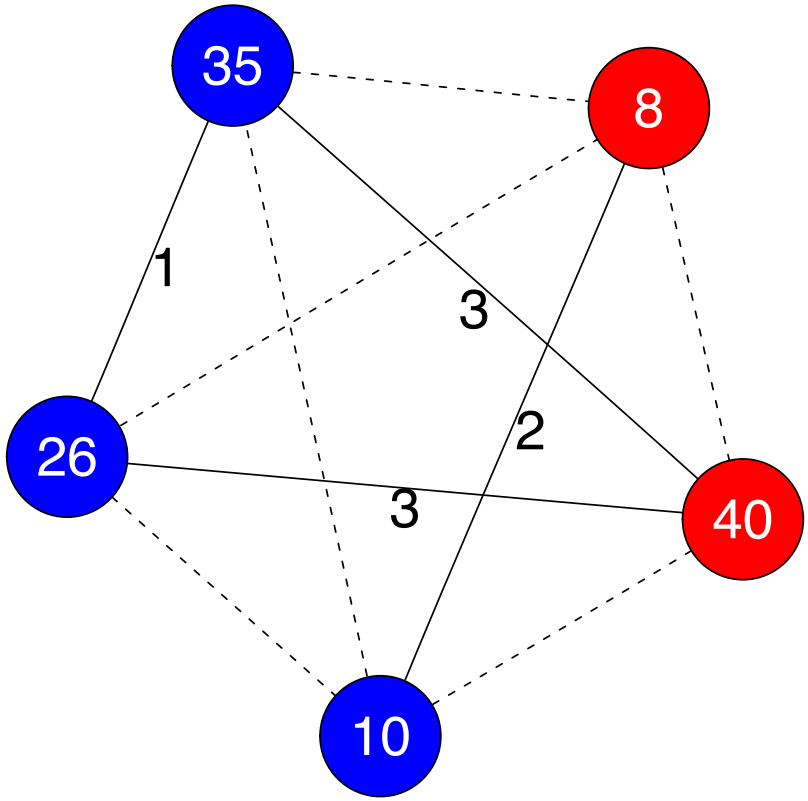
Pedigree Dissimilarity

- Use Malecot's kinship coefficient (φ):
  - for subjects *i* and *j*: the probability that randomly chosen alleles, one from each subject, are *identical by descent*
  - e.g. parent-offspring: 1/4
  - e.g. siblings: 1/4
- Pedigree dissimilarity: $$d_{ij} = -\log_2(2\varphi_{ij})$$

Relationship Graph

- In studies like BDES, not all members of pedigree are subjects
- Instead of full pedigree we have a *relationship graph*

Relationship Graph



Relationship	Distance
sibs	1
avuncular	2
first-cousins	3
unrelated	∞

We will extend the SS-ANOVA model with an encoding of this relationship graph

SS-ANOVA

- Recall: estimate log odds ratio

$$p(x) = Pr\{y = 1|x\}$$

$$f(x) = \log \frac{p(x)}{1 - p(x)}$$

- *f* is nonparametric (or semiparametric)
- *f* has ANOVA-like decomposition

SS-ANOVA

- Estimate is solution of penalized likelihood prob[illegible] reproduci[illegible]rt spa[illegible] of the [illegible]

parametric part

nonparametric part

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1, \mathcal{H}_0 \perp \mathcal{H}_1$$

- Parametric part specified by finite set of functions

$$\mathrm{span}(\phi_1, \ldots, \phi_M) = \mathcal{H}_0$$

SS-ANOVA

- $\mathcal{H}_1$ is RKHS with associated kernel *k*:

$$g \in \mathcal{H}_1 \Rightarrow \langle k(x,\cdot), g\rangle_{\mathcal{H}_1} = g(x)$$

$$k(x_i, x_j) = \langle k(x_i,\cdot), k(x_j,\cdot)\rangle_{\mathcal{H}_1}$$

- e.g. Gaussian kernel:

tunable hyper-parameter

$$k(x_i, x_j) = \exp\left\{-\gamma\|x_i - x_j\|^2\right\}$$

SS-ANOVA

- Estimate is the solution of a penalized likelihood problem

# of obs.

Log-odds ratio

penalty

$$\min_{f \in \mathcal{H}} I_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \left[-y_i f(x_i) + \log(1 + e^{f(x_i)}) \right] + \lambda \| P_1 f \|^2_{\mathcal{H}_1},$$

negative log-likelihood

Tunable hyper-parameter

SS-ANOVA

- By Kimeldorf and Wahba representer theorem, minimizer of penalized likelihood problem has finite representation

$$f(\cdot) = \sum_{i=1}^{m} d_j \phi_j(\cdot) + \sum_{i=1}^{n} c_i k(x_i, \cdot).$$

parametric part

nonparametric part

SS-ANOVA

- Minimizer is solution to

kernel matrix

$$\min_{c\in\mathbb{R}^n, d\in\mathbb{R}^m} \sum_{i=1}^{n} \left[-y_i f_i + \log(1+e^{f_i})\right] + n\lambda c^T K c,$$

$$f = Td + Kc$$

$$T_{ij} = \phi_j(x_i)$$

$$K_{ij} = k(x_i, x_j)$$

SS-ANOVA

- For example,

cubic splines

nonparametric part

$$f(t) = \mu + f_1(\text{sysbp}) + f_2(\text{chol}) + f_{12}(\text{sysbp}, \text{chol}) +$$
$$\boxed{\begin{aligned} d_{\text{age}} \cdot \text{age} + d_{\text{bmi}} \cdot \text{bmi} + d_{\text{horm}} \cdot I_1(\text{horm}) + \\ d_{\text{hist}} \cdot I_2(\text{hist}) + d_{\text{smoke}} \cdot I_3(\text{smoke}), \end{aligned}}$$

Parametric part

SS-ANOVA

- In SS-ANOVA model, $\mathcal{H}_1$ is assumed to be direct sum of multiple RKHS so:

$$g(x) = \sum g_\alpha(x_\alpha) + \sum_{\alpha<\beta} g_{\alpha\beta}(x_\alpha, x_\beta) + \cdots$$

main effects

$$g_\alpha \in \mathcal{H}_\alpha$$

second-order interactions

SS-ANOVA

- For example,

main effects

second-order interaction

$$f(t) = \mu + f_1(\text{sysbp}) + f_2(\text{chol}) + f_{12}(\text{sysbp}, \text{chol}) + d_{\text{age}} \cdot \text{age} + d_{\text{bmi}} \cdot \text{bmi} + d_{\text{horm}} \cdot I_1(\text{horm}) + d_{\text{hist}} \cdot I_2(\text{hist}) + d_{\text{smoke}} \cdot I_3(\text{smoke}),$$

SS-ANOVA

- In SS-ANOVA model, $\mathcal{H}_1$ is assumed to be direct sum of multiple RKHS so:

$$g(x) = \sum_{\alpha} g_\alpha(x_\alpha) + \sum_{\alpha<\beta} g_{\alpha\beta}(x_\alpha, x_\beta) + \cdots$$

$$g_\alpha \in \mathcal{H}_\alpha$$

- So, we can write [illegible] term as:

tunable hyper-parameters

$$\lambda \|g\|^2_{\mathcal{H}_1} = \lambda \left[\sum_{\alpha} \theta_\alpha^{-1} \|g_\alpha\|^2_{\mathcal{H}_\alpha} + \sum_{\alpha<\beta} \theta_{\alpha\beta}^{-1} \|g_{\alpha\beta}\|^2_{\mathcal{H}_{\alpha\beta}} + \cdots \right]$$

SS-ANOVA

- A kernel $k_\alpha(\cdot,\cdot)$ function is associated with each component $\mathcal{H}_\alpha$

- For penalty with coefficients θ , the kernel for $\mathcal{H}_1$ is then

$$k(\cdot,\cdot) = \sum_\alpha \theta_\theta k_\alpha(\cdot,\cdot) + \sum_{\alpha\beta} \theta_{\alpha\beta} k_{\alpha\beta}(\cdot,\cdot) + \cdots$$

SS-ANOVA

- Coefficients θ may be interpreted as relative importance of each model component
- Hyper-parameters (λ, θ) tuned with GACV [Xiang and Wahba '96]
  - approximation of Kullback-Leibler divergence between estimate and unknown "true" function

Extending SS-ANOVA

- Add a main effect term to decomposition of $\mathcal{H}_1$ that encodes pedigree data:

$$f(t_i) = \mu + g_1(t_i) + g_2(t_i) + h(z(t_i)),$$

Extending SS-ANOVA

- Add a main effect term to decomposition of $\mathcal{H}_1$ that encodes pedigree data:

$$f(t_i) = \mu + g_1(t_i) + g_2(t_i) + h(z(t_i)),$$

- Requires properly defining a kernel matrix
- θ gives relative importance of model components

Regularized Kernel Estimation

- Given pedigree dissimilarity data
- Estimate a kernel matrix *K* that induces distances:

dissimilarity ~ distance

$$d_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$$

- Add *K* to SS-ANOVA model
- Additionally, get an embedding in "pedigree" Euclidean space

Regularized Kernel Estimation

- Regularization: Minimize the rank of *K*?
  - Can help with estimation
  - Definitely helps with visualization
- Get a convex relaxation to rank minimization by minimizing trace of K.

Regularized Kernel Estimation

- Given, *N* objects and *some* dissimilarity information $d_{ij} \in \Omega$ where $|\Omega| < \binom{N}{2}$
- Solve:

regularization parameter

$$\begin{aligned} \min \quad & \textstyle\sum_{ij \in \Omega} |d_{ij}^2 - \hat{d}_{ij}^2(K)| + \lambda \mathrm{trace}(K) \\ \text{s.t.} \quad & K \succeq 0 \end{aligned}$$

- where $\hat{d}_{ij}^2(K) = K_{ii} + K_{jj} - 2K_{ij}$

Regularized Kernel Estimation

- It is a convex optimization problem
  - In particular, a linear semidefinite program
- There are interior-point methods (& code) of polynomial complexity to solve exactly
- DSDP5 [Benson & Yu '00], CSDP [Borchers '99], SDPT3 [Toh et al. '99]

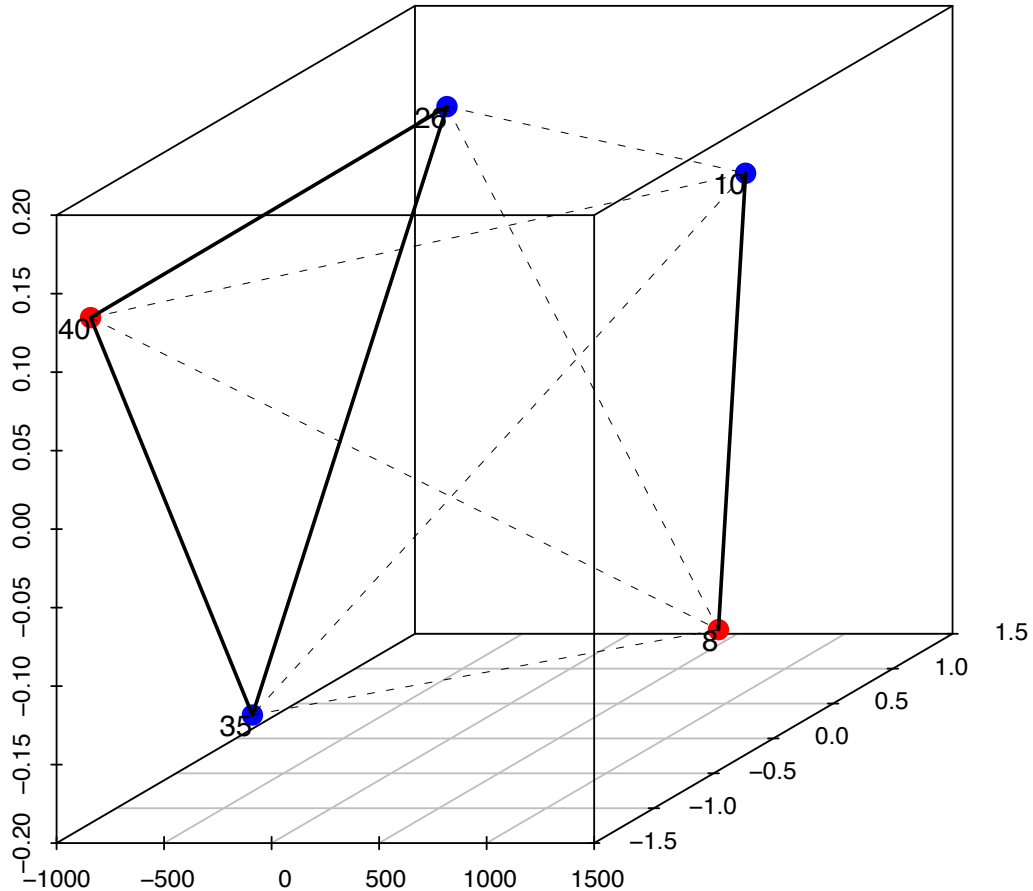
Regularized Kernel

- Also, can get embedding in "pedigree" space:
  - since solution to RKE *K* is positive semidefinite, we can write

$$K = XX^T$$

  - using *r* leading eigenvalues and eigenvectors of *K*
  - then, *X* is an *r*-dimensional embedding in "pedigree" space

Regularized Kernel Estimation



Interpretation: embedding gives pedigree *pseudo*-attributes over which a smooth function can be estimated, using, e.g. Gaussian kernel

Another Method

- Graph kernel: given pedigree dissimilarity, use

$$K_{ij} = \exp\{-\gamma d_{ij}^2\}$$

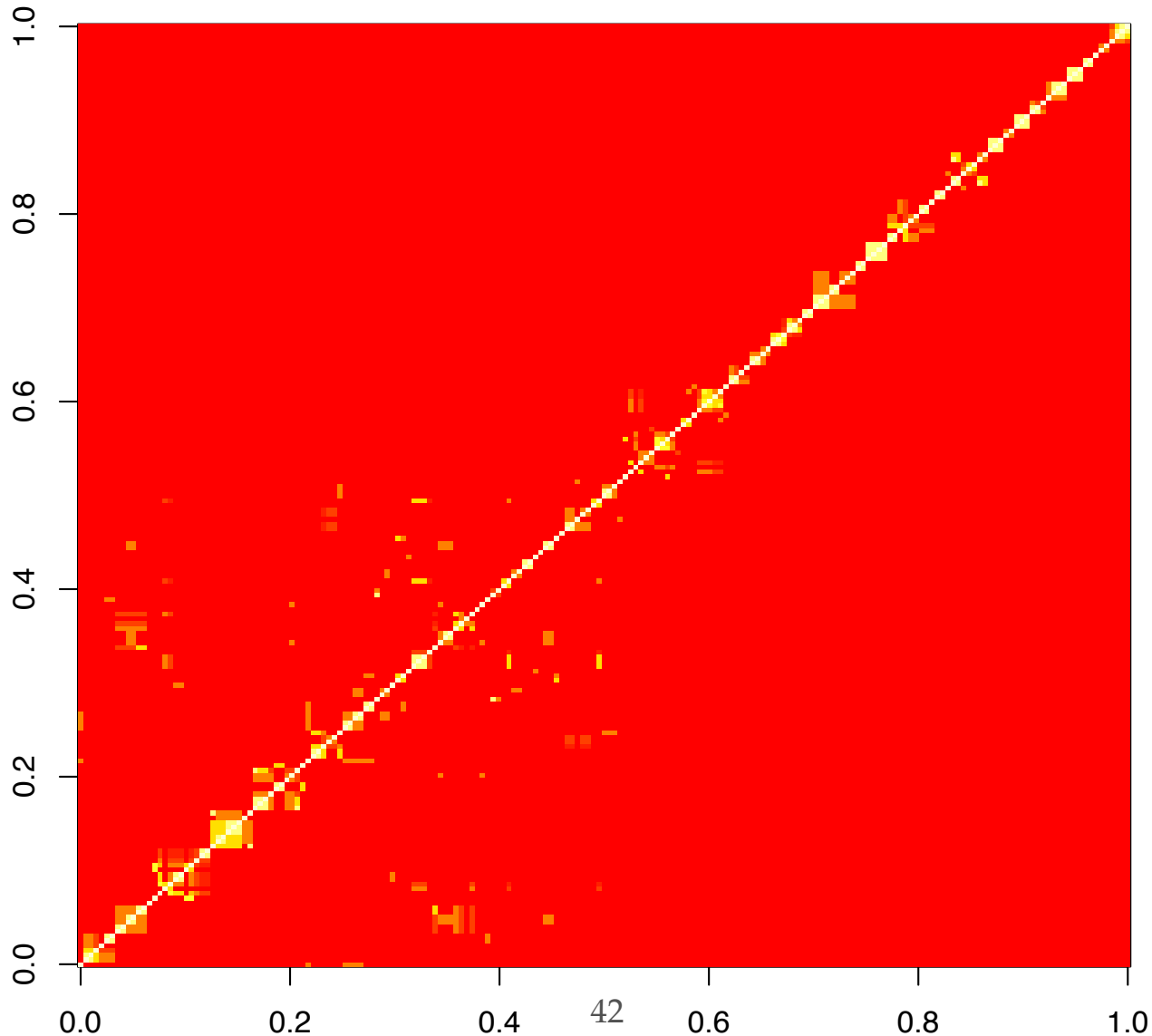
- Not necessarily positive semi-definite
- Project to space of psd matrices by truncating eigendecomposition

Diffusion Effect

- Gaussian kernels over pedigree graphs or RKE embedding result in kernels that are:
  1. very sparse: very few relationships in pedigree graphs
  2. very diffuse: using pedigree dissimilarity, there is rapid decay as relationship dissimilarity increases

Diffusion Effect

Gaussian kernel over graph, $\gamma = 0.1$



Diffusion Effect

- Would prefer kernels that:
  - depend only on distances, rotationally invariant
  - parameterized decay as relationship dissimilarity increases
- Solution: Matérn kernel family [Matérn ’86, Stein ‘99]

Matérn Kernels

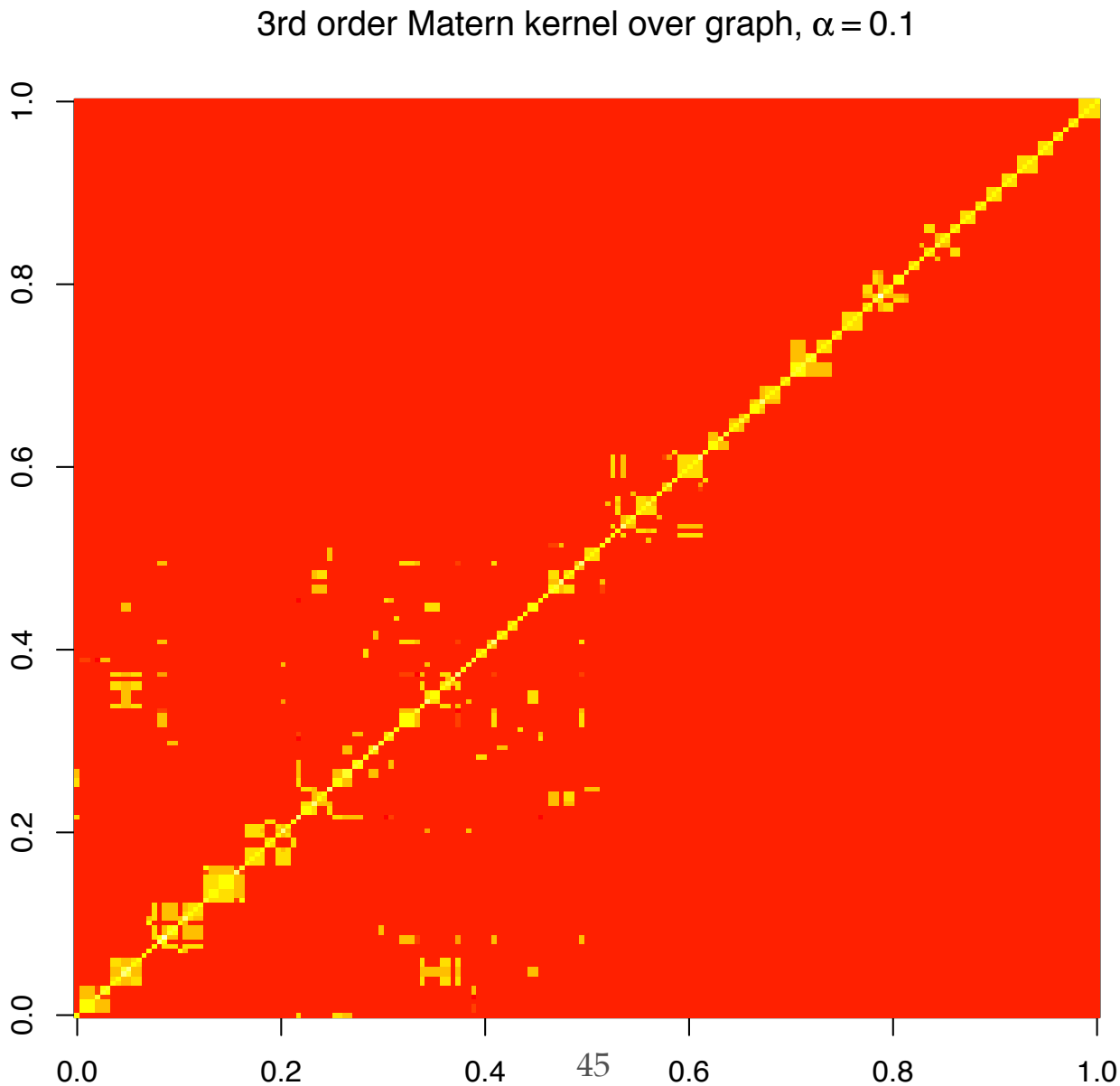
- General form Matérn kernel:

$$k_\nu(i,j) = \exp\{-\alpha d_{ij}\}\pi_\nu(\alpha, d_{ij}),$$

- π_ν is a polynomial, can control exponential decay, α is a scale parameter
- In experiments, we use 3rd order Matérn kernel:

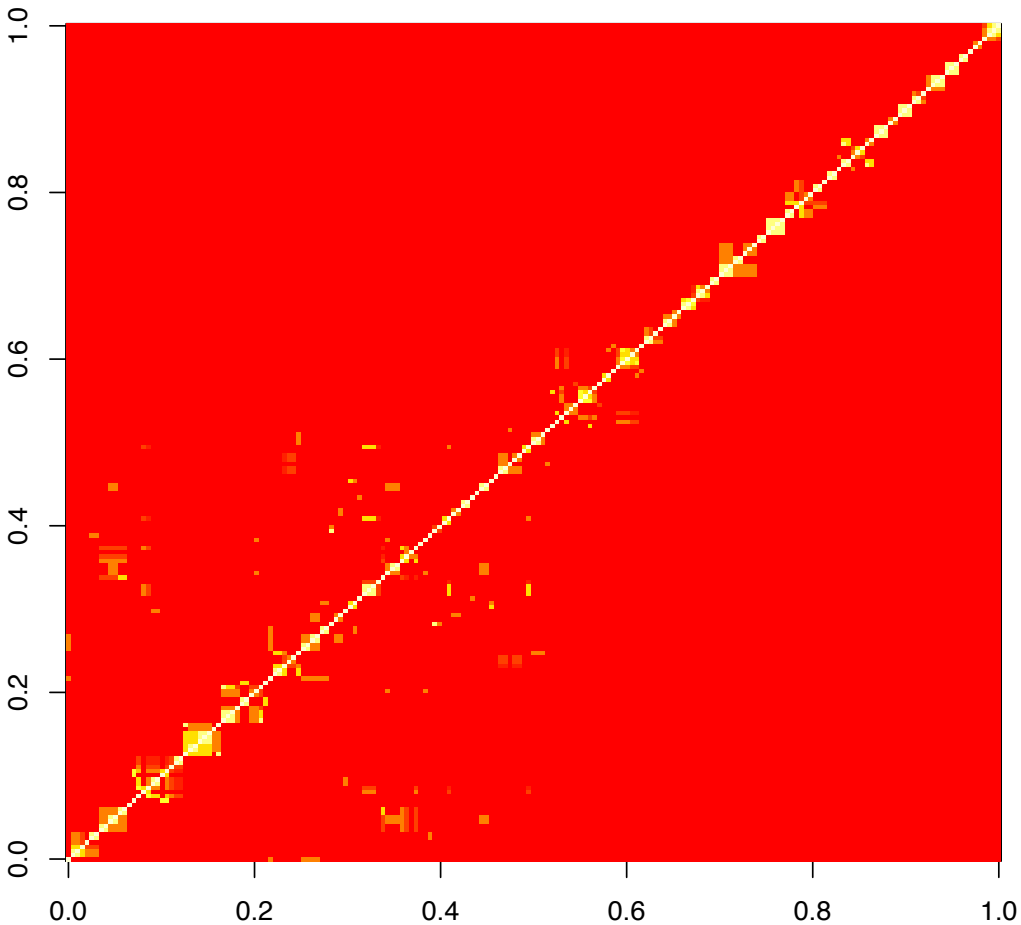
$$k_3(i,j) = \frac{1}{\alpha^7}\exp\{-\alpha\tau\}[15 + 15\alpha\tau + 6\alpha^2\tau^2 + \alpha^3\tau^3],$$

Matérn Kernel

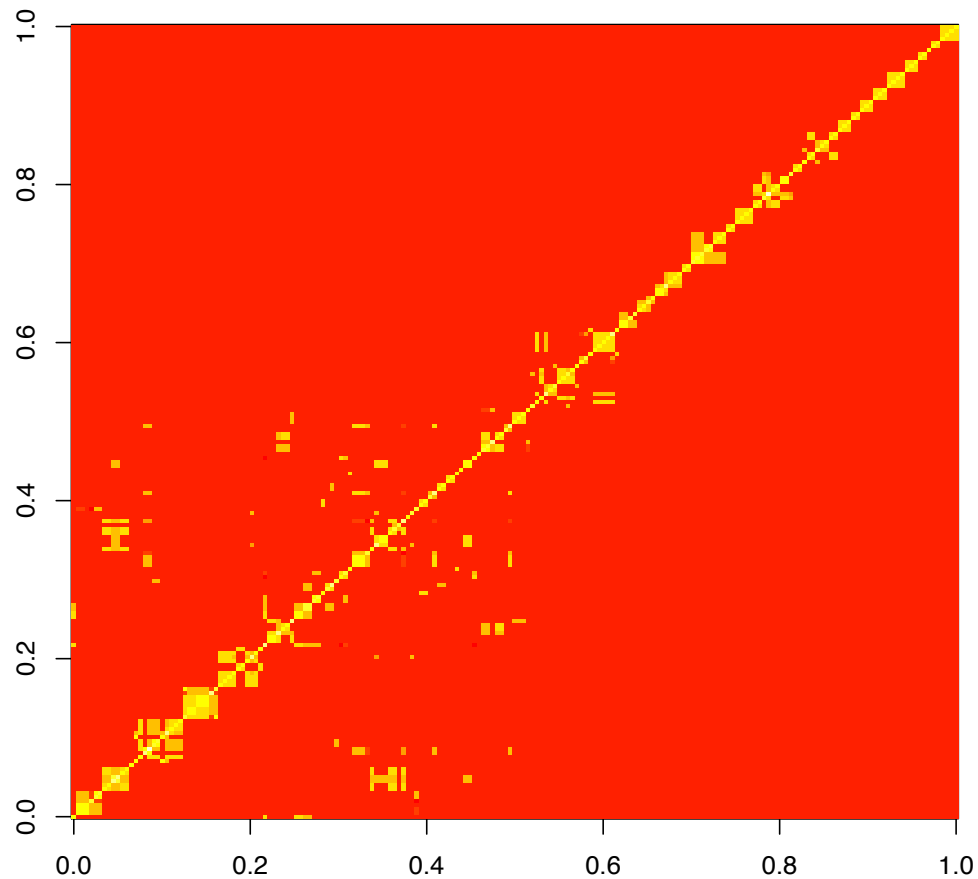


Diffusion Effect

Gaussian kernel over graph, $\gamma = 0.1$

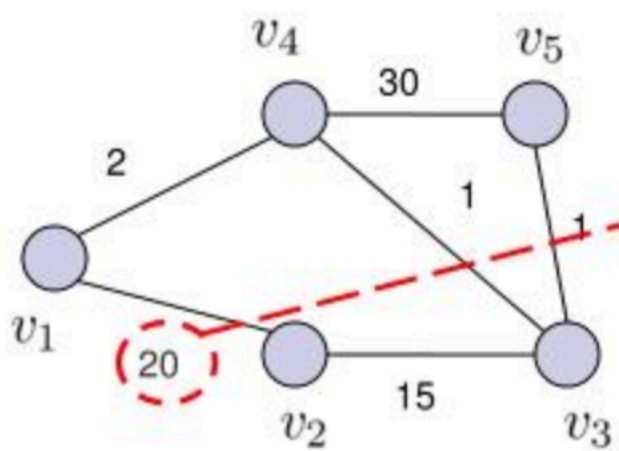


3rd order Matern kernel over graph, $\alpha = 0.1$



Another kernel alternative

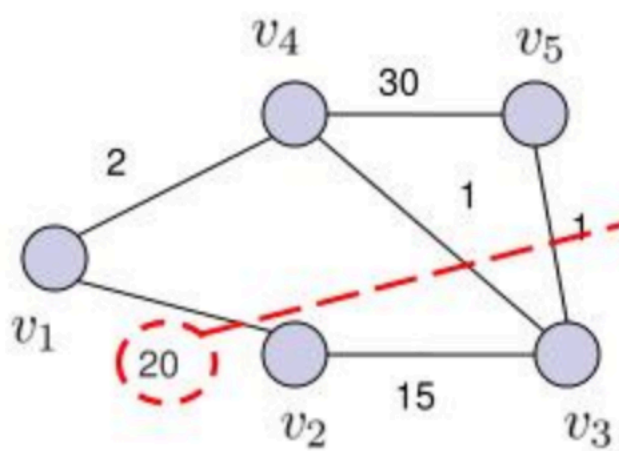
- Another commonly used kernel for graphs is based on the graph *Laplacian L=D-A* with *D* diagonal degree matrix and adj (or weight) matrix A



$$L = \begin{pmatrix} 22 & -20 & 0 & -2 & 0 \\ -20 & 35 & -15 & 0 & 0 \\ 0 & -15 & 17 & -1 & -1 \\ -2 & 0 & -1 & 33 & -30 \\ 0 & 0 & -1 & -30 & 31 \end{pmatrix}$$

Another kernel alternative

- We can use L as kernel matrix
  - Alternatively, the first few eigenvalues/ eigenvectors of L



$$L = \begin{pmatrix} 22 & -20 & 0 & -2 & 0 \\ -20 & 35 & -15 & 0 & 0 \\ 0 & -15 & 17 & -1 & -1 \\ -2 & 0 & -1 & 33 & -30 \\ 0 & 0 & -1 & -30 & 31 \end{pmatrix}$$

SS-ANOVA

- Recall SS-ANOVA setting:

$$\min_{c\in\mathbb{R}^n, d\in\mathbb{R}^m} \sum_{i=1}^{n} \left[-y_i f_i + \log(1+e^{f_i})\right] + n\lambda c^T K c,$$

- (Weighted) Laplacian as K makes penalty correspond to

$$\sum_{i\sim j} w_{ij}(f_i - f_j)^2$$

Case Study: BDES

- Cohort: female subjects in BDES I, with full ascertained data for: both genetic markers, pedigree, and covariates in model
- In pedigrees containing two or more of these subjects (n=684)

Case Study: BDES

- Use two SNPs:
  1. near complement factor H (CFH)
  2. near ARMS2
- Each SNP has three levels: 11, 12, 22
- Linear terms for levels 12 and 22 added to model (level 11 folded into intercept)

Case Study: BDES

- Full model:

$$
\begin{aligned}
f(t) = \mu + d_{\mathrm{SNP1,1}} \cdot I(X_1 = 12) + d_{\mathrm{SNP1,2}} \cdot I(X_1 = 22) + \\
d_{\mathrm{SNP2,1}} \cdot I(X_2 = 12) + d_{\mathrm{SNP2,2}} \cdot I(X_2 = 22) + \\
f_1(\mathrm{sysbp}) + f_2(\mathrm{chol}) + f_{12}(\mathrm{sysbp}, \mathrm{chol}) + \\
d_{\mathrm{age}} \cdot \mathrm{age} + d_{\mathrm{bmi}} \cdot \mathrm{bmi} + d_{\mathrm{horm}} \cdot I_1(\mathrm{horm}) + \\
d_{\mathrm{hist}} \cdot I_2(\mathrm{hist}) + d_{\mathrm{smoke}} \cdot I_3(\mathrm{smoke}) + \\
h(z(t))
\end{aligned}
$$

Marker data

environmental covariates

pedigree data

Case Study: BDES (Models)

- Indicate models by components:
  1. **S**: genetic markers
  2. **C**: environmental covariates
  3. **P**: pedigree data
- For example:
  - **S-only**: model containing only marker data
  - **S+P:** model contains marker and pedigree data
  - **S+C+P:** full model

Case Study: BDES

1. GAUSSIAN: Gaussian kernel over relationship graph
2. MATERN: 3rd-order Matérn kernel over relationship graph
3. RKE/GAUSSIAN: Gaussian kernel over RKE embedding
4. RKE/MATERN: 3rd-order Matérn kernel over RKE embedding

Case Study: BDES

- Parameters to tune:
  - regularization parameters in penalized likelihood and RKE
  - ANOVA decomposition coefficients
  - kernel hyper-parameters
- All tuning done by minimizing GACV criterion

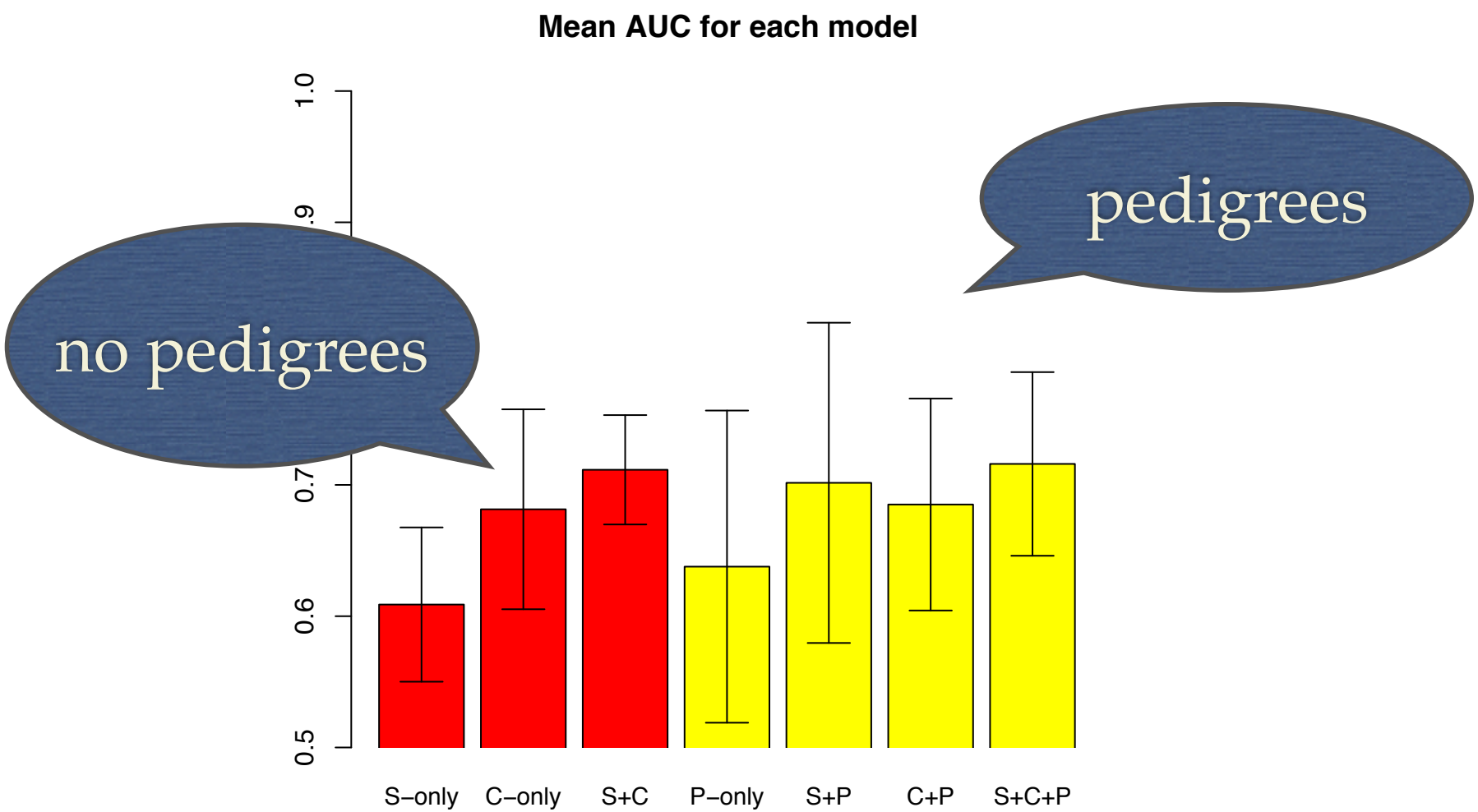
Case Study: BDES

- Penalized likelihood problem solved by quasi-Newton method (gss R package) [Gu '07]
- Tuning: gss finds ANOVA decomposition parameters using a quasi-Newton method
  - remaining parameters tuned by grid search
  - Rmpi library used for parallel grid search
- RKE problem solved by CSDP library [Borchers '99] (R interface)

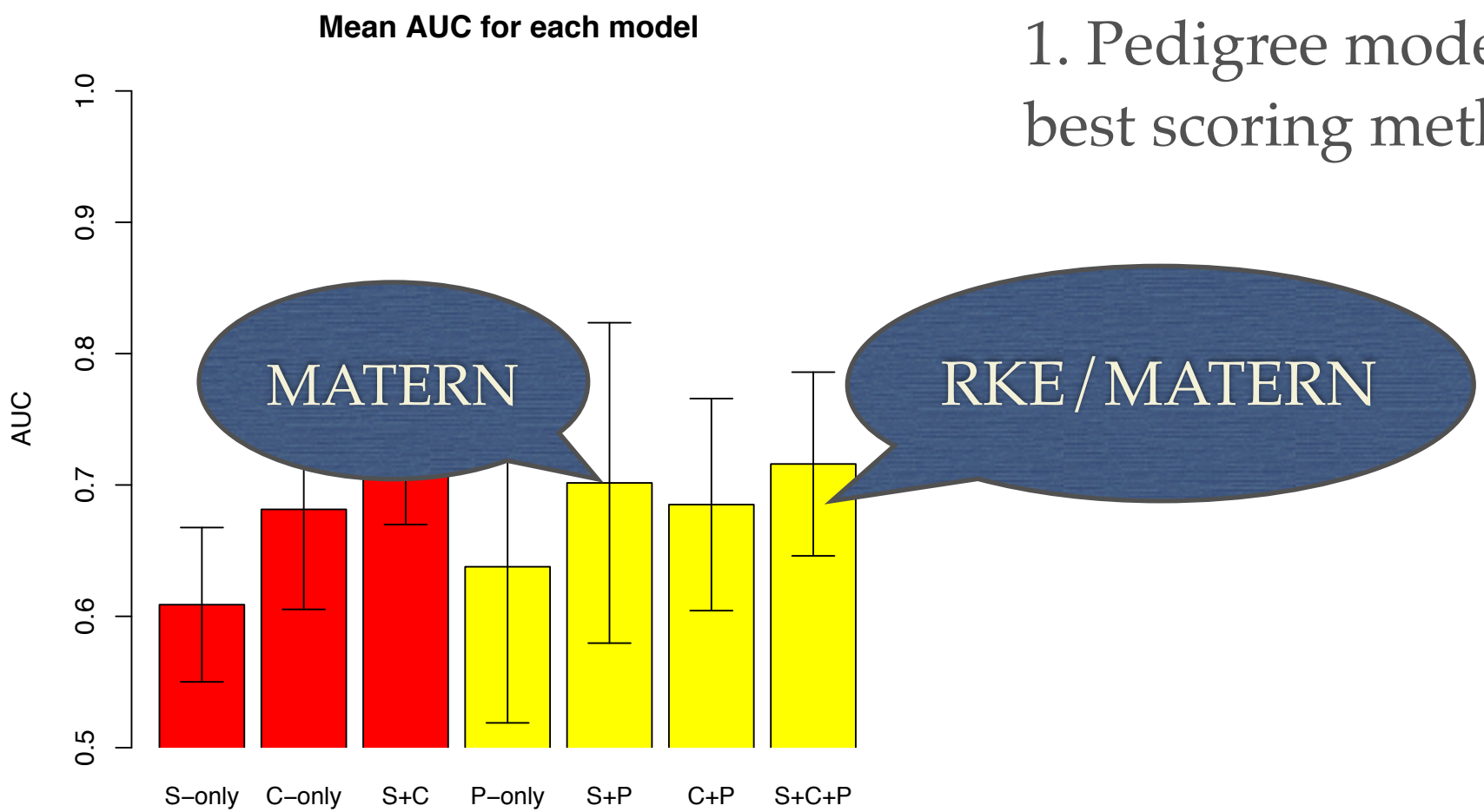
Case Study: BDES

- Prediction performance measured by area under ROC curve (AUC)
- Estimated by 10-fold cross validation

Case Study: BDES

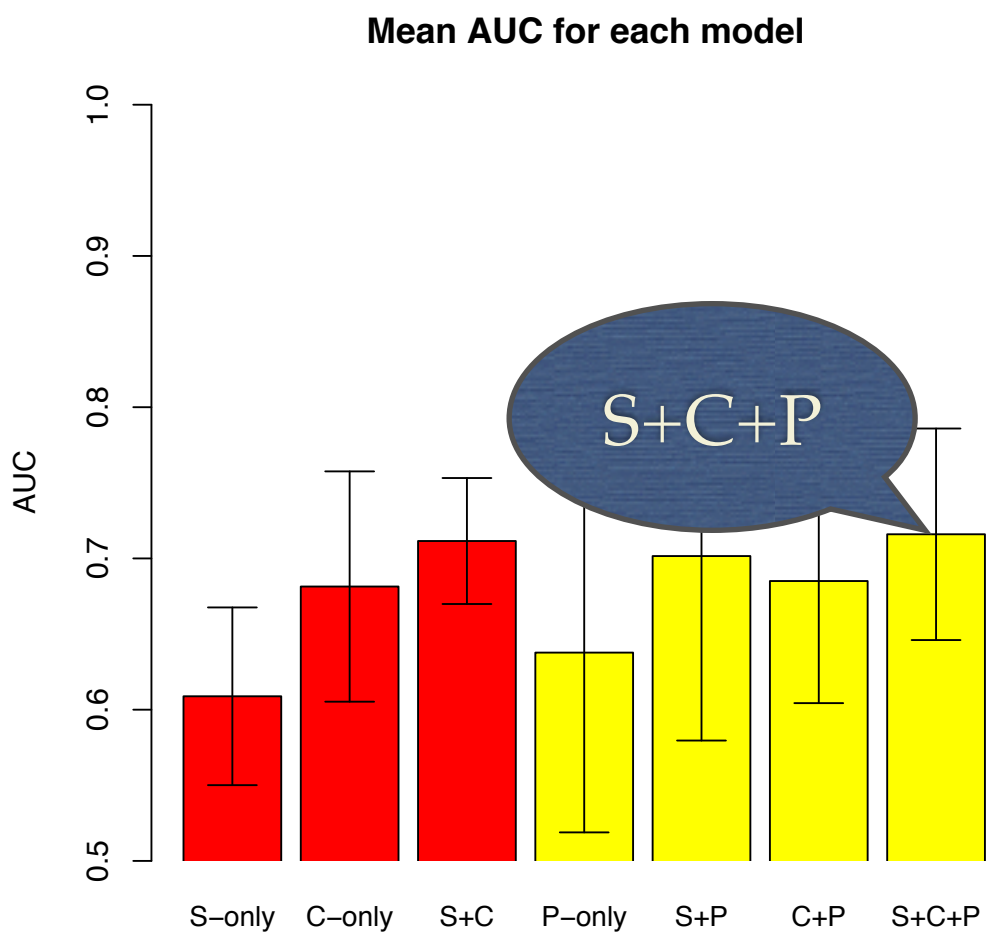


Case Study: BDES



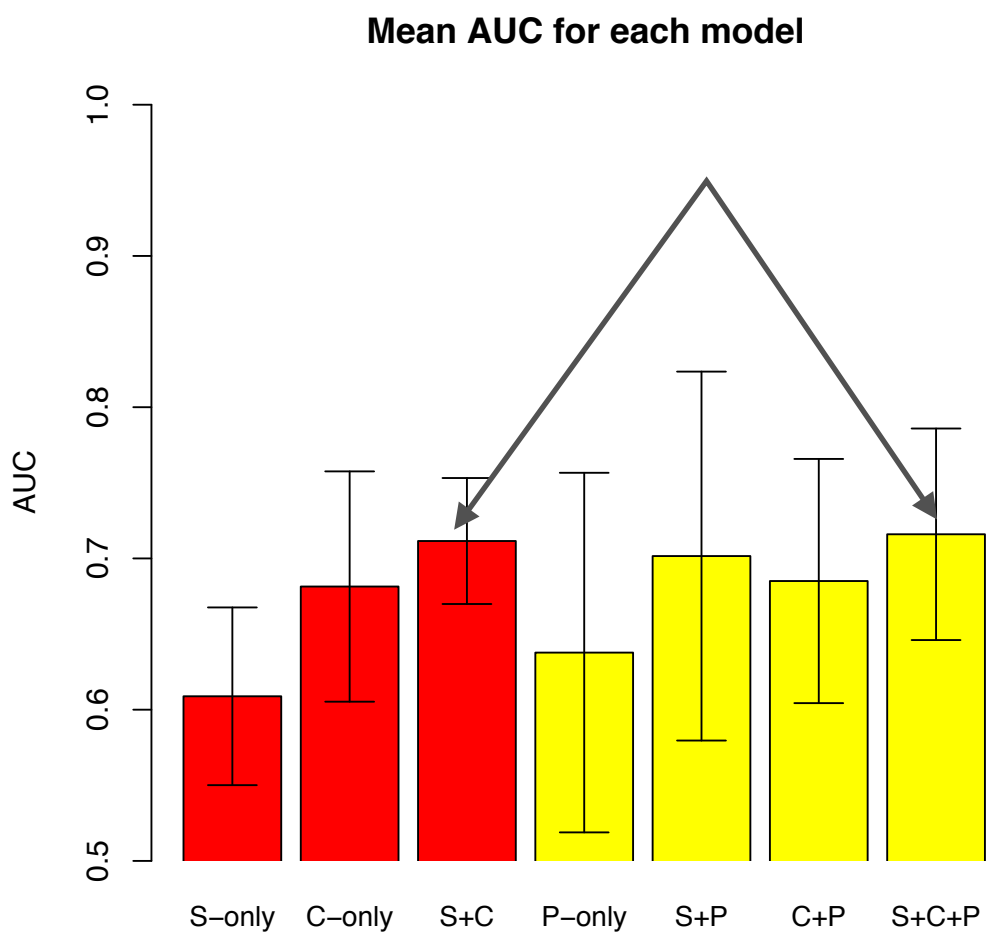
1. Pedigree models refer to best scoring method

Case Study: BDES



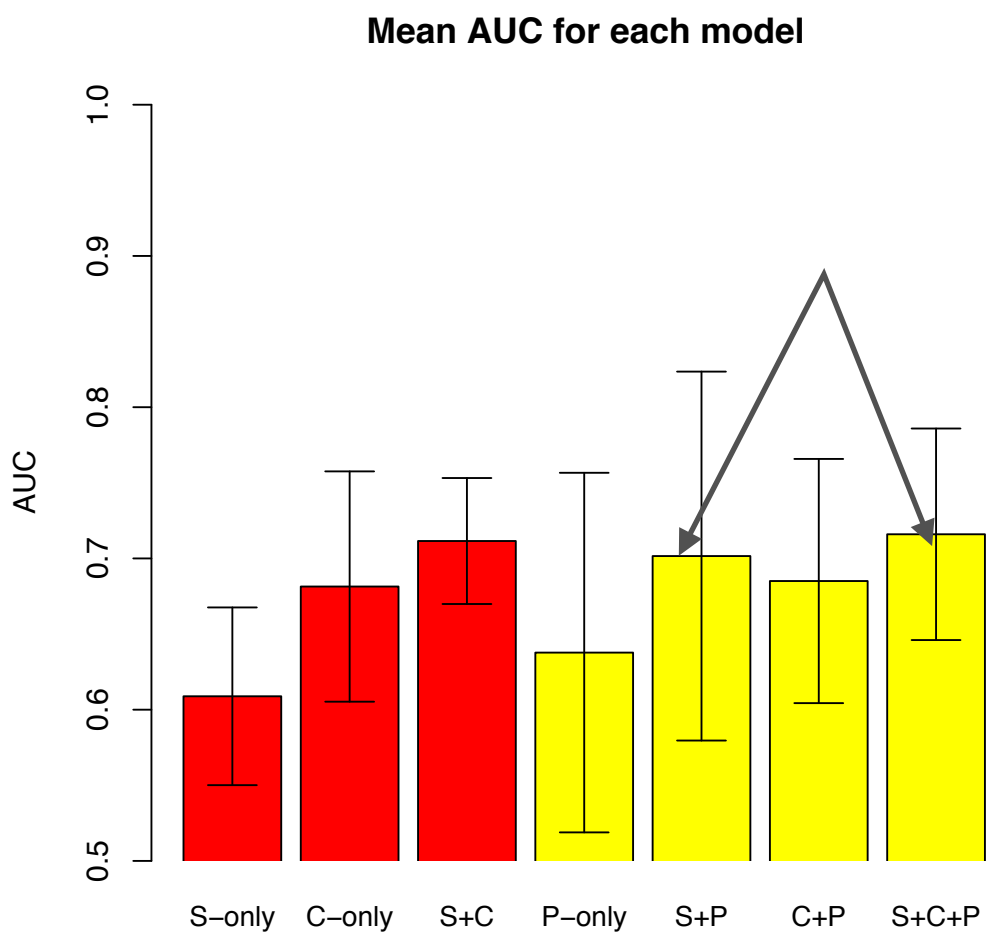
1. Pedigree models refer to best scoring method
2. Best scoring model is **S+C+P**

Case Study: BDES



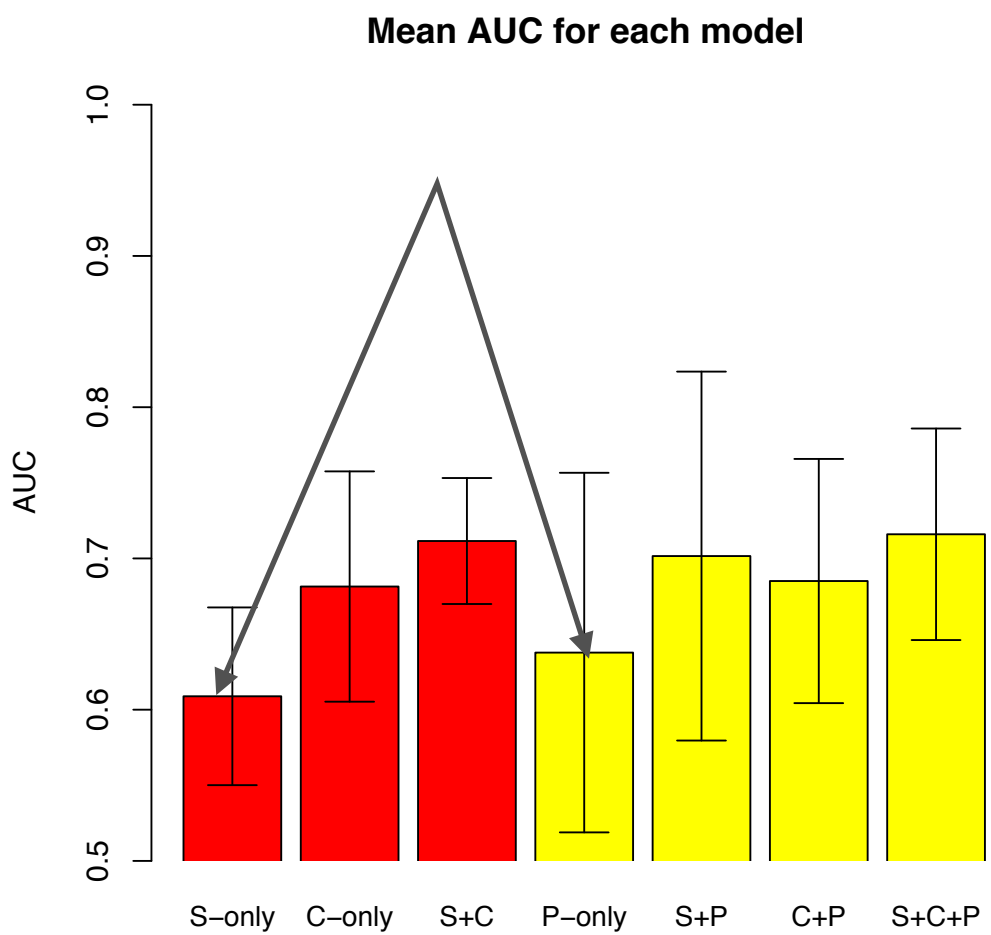
1. Pedigree models refer to best scoring method
2. Best scoring model is **S+C+P**
   a. **S+C**, similar

Case Study: BDES



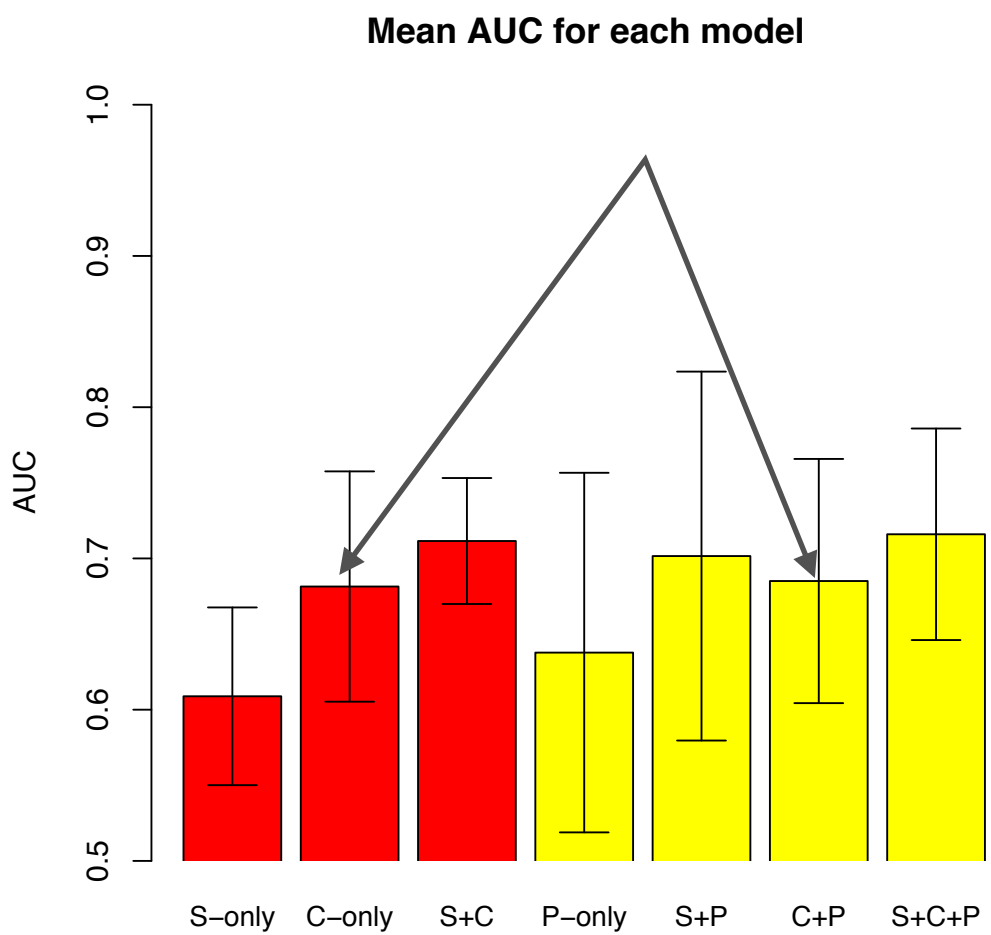
1. Pedigree models refer to best scoring method
2. Best scoring model is **S+C+P**
   a. **S+C**, similar
   b. **S+P**, similar

Case Study: BDES



1. Pedigree models refer to best scoring method
2. Best scoring model is **S+C+P**
   a. **S+C**, similar
   b. **S+P**, similar
3. **P-only ≈ S-only**

Case Study: BDES



1. Pedigree models refer to best scoring method
2. Best scoring model is **S+C+P**
   a. **S+C**, similar
   b. **S+P**, similar
3. **P-only ≈ S-only**
4. **S+P > S-only**
5. **C+P ≈ C-only**

Case Study: BDES

- Paper also has simulation results where **S+C+P > S+C**
- Simulates setting where markers (**S**) do not model entire genetic influence in disease risk
- then, pedigree data (**P**) models remaining genetic influence

Extensions

1. More complex models for marker data
   - multiple markers per gene, interactions

Extensions

1. More complex models for marker data
2. Alternative pedigree dissimilarity measures
   - e.g., dissimilarity for spouses encoding *some* notion of environmental sharing
3. Multiple pedigree dissimilarity measures

Extensions

1. More complex models for marker data
2. Alternative pedigree dissimilarity measures
3. Multiple pedigree dissimilarity measures
4. Interactions between components: **S*P + C**

Extensions

1. More complex models for marker data
2. Alternative pedigree dissimilarity measures
3. Multiple pedigree dissimilarity measures
4. Interactions between components: **S*P + C**
5. Further understanding of diffusion effect:
   - depends on dissimilarity measure
   - when is Matérn better than Gaussian?

Summary

- Extended existing nonparametric disease risk model with relationship data
  - using general methodology encoding relationships in graphs
- Assumption: graph relationships is one of multiple, comparable, model components affecting outcome

Dissertation Summary

- Taking data relationships into account is challenging
  - Graph structure in relationships makes analysis viable

Dissertation Summary

- Taking data relationships into account is challenging
  - Graph structure in relationships makes analysis viable
- Solve interesting problems with well-known optimization methods

Dissertation Summary

- Taking data relationships into account is challenging
  - Graph structure in relationships makes analysis viable
- Solve interesting problems with well-known optimization methods
- Apply to real data: genomic and epidemiological applications, decision-making applications

Dissertation Summary

- Taking data relationships into account is challenging
  - Graph structure in relationships makes analysis viable
- Solve interesting problems with well-known optimization methods
- Apply to real data: genomic and epidemiological applications, decision-making applications
- Produce publicly-available software

Thanks!