The Collaborative Center for Statistics in Science



Search for the Smallest Random Forest

all to see that $(\partial/\partial\beta)\pi(\beta; k, c) = c$

$$og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$$

$$+ \sum_i \frac{\partial}{\partial \beta}log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\}]$$

$$= [1 - \gamma(0; y_{ii}, 1) - \gamma(0; y_{ij}, 1)] + \gamma(0; y_{ij}, 1) + \gamma(0; y_{ij}, 1) + \gamma(0; y_{ij}, 1) + \gamma(0; y_{ij}, 1)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0} = \sum_$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Heping Zhang
Yale University
Joint Work with Minghui Wang

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}| e_{ij} = 0\} P \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{e_{i}\}] \end{aligned}$$

Outline



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

ble to see that
$$(\partial/\partial \beta)\pi(\beta; k, c) = c$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ & = [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- Background
- Goal
- Key idea
- Method
- Simulation
- Application

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|e_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{e_{ij}\}]$

Background



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ a$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F]$$

=
$$\prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{ij} | c_{ij} = 0\} F]$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) = c$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 -$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Random forests have emerged as one of the most commonly used nonparametric statistical methods in many scientific areas, particularly in analysis of high throughput genomic data.

$$\begin{aligned} d_i|y_i| &= \frac{e^{-i\beta}}{P\{y_i\}} \prod_j [P\{y_{ij}|e_{ij} = 0\}P\\ &= \frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{e_{ij}\}] \end{aligned}$$

-P(M) Π(B) N(B) P(C) Background



ole to see that $(\partial/\partial \beta)\pi(\beta; h, c) =$ $\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial B}\log(P\{y_i\})$

 $\frac{\partial}{\partial \mathbf{B}} \log[\pi(\mathbf{B}; y_{ij}, 0) P\{dd|M_{ij}\}$

A general practice in using random $P(y) = \prod_{[P(y)]_{ij}=0}^{P(y)} A$ forests is to generate a sufficiently large number of trees, although it is subjective as to how large is sufficient.

Furthermore, random forests are viewed as a "black-box" because of its sheer size.

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1 - \frac{1}{2}]$$
 renience, we drop the two irrelevants
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$
$$= \sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$$

 $AA - P\{dd, AA\} - P\{AA\}[P\{DD]$





$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Goal



3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P\}$$

$$= \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; h, c) = c$$

$$\begin{aligned} \log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\ &+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)] \end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

renience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA$$
 $= P{dd, AA} - P{AA}[P{DE}$

Explore whether it is possible to find a common ground between a forest and a single tree

- retain the easy interpretability of the tree-based methods
- avoid the problems that the tree-based methods suffer from.

Does a forest have to be large, or how small can a forest be?

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P\}$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Key idea



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e)$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P_i] \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P_i] \end{split}$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, c) =$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{0}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \gamma(\boldsymbol{0}; \ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Shrink the forest with two objectives

- maintain a similar (or even better) level of prediction accuracy
- reduce the number of the trees in the forest to a manageable level



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_{j} |P\{y_{ij}|c_{ij}=0\}P$$

$$= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta;y_{ij},0)P\{z_{ij}=0\}P$$

$$= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta;y_{ij},0)P\{z_{ij}=0\}P$$



3; k, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k$ K - 1, $\gamma(\beta, 0, \epsilon) = 0$, and $\gamma(\beta, K, \epsilon)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}]$$

$$= \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) = c$$

he null hypothesis that $\beta = 0$, we have

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd]M_{ij}\} - \text{by similarity}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Three measures are considered to determine the importance of a tree in a forest

- by prediction
- by restricted similarity

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]^{p}$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | c_{ij} = 0\} I \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P\{...] \end{split}$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

he to see that (θ/θβ)π(β; λ. ∮— "by prediction" method

- focuses on the prediction
- A tree can be removed if its removal from the forest has the minimal impact on the overall prediction accuracy.

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\} P P\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\} P P\{M_{i}|y_{i}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{a_{i}, a_{j}, a_{j}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{a_{i}, a_{j}\} = \frac{P\{y_{ij} = k|c_{ij} = a_{j}\}}{P\{y_{ij} = k|c_{ij} = a_{j}\}} \end{aligned}$$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij}| c_{ij} = 0\}]$$

$$= \prod_{j} [\pi(\beta; y_{ij}, 0)]P\{$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=\epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA$$
 = $P{dd, AA} - P{AA}[P{DD}$

"by prediction" method

- For tree T in forest F, calculate the prediction accuracy of forest $F_{(-T)}$ that excludes T.
- $-\Delta_{l-T}$ represents the difference in prediction accuracy between F and $F_{(-T)}$.
 - The tree with the smallest $\Delta_{(-T)}$ is the least important one and hence subject to removal.

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, i)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

be to see that (θ/θβ)π(β; λ.) = "by similarity" method

is based on the similarity between two trees.

$$\begin{aligned} \frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\} \\ = [1 - \gamma(0; y_{ii}, 1) - \gamma(0; \end{cases} \end{aligned}$$

be null hypothesis that β = 0, we h — A tree can be removed if it is "similar" to other trees in the forest.

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0} = \sum_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DE]$$

$= \frac{P\{M_i\}}{P\{y_i\}} \prod_{i} [\pi(\beta; y_{ij}, 0)P\{c_{(\beta; k_i, c)} = P\{y_{ij} = k_i c_{ij} = k_i c_{i$

Method



3;
$$k$$
, ϵ) = $P\{y_{ij} = k | \epsilon_{ij} = \epsilon\} = \gamma(\beta; k - 1, \gamma(\beta, 0, \epsilon) = 0, \text{ and } \gamma(\beta, K, 0, \epsilon) = 0\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} I$$
$$= \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

"by similarity" method

ole to see that $(\partial/\partial \beta)\pi(\beta; k, c) =$ $\operatorname{og}(P\{M_i|y_i\}) = -\frac{\partial}{\partial B} \operatorname{log}(P\{y_i\})$ $+\sum \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{\theta}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{\theta}) - \gamma(\boldsymbol{\theta}; \ \boldsymbol{\theta})] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{\beta}] = \sum_i [1 - \frac{1}{\beta}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA\} - P\{dd, AA\} - P\{AA\}[P\{DD]$$

- The correlation of the predicted outcomes by two trees gives rise to a similarity between the two trees.
- For tree T, the average of its similarities with all trees in $F_{(-T)}$, denoted by \mathcal{O}_T , reflects the overall similarity between T and $F_{(-T)}$.
 - The tree with the highest ρ_{τ} is the most similar to the trees in $F_{(-T)}$ and hence subject to removal.

$$\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij}=0\}P] P\{M_{i}\} - \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta;y_{ij},0)P\{c_{ij},a_{i$$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0$, and $\gamma(\beta, K, 0, e) = 0$

$$P\{y_i\} = \prod_j [P\{y_{ij} | a_{ij} = 0\} F$$
$$= \prod_j [\pi(\beta; y_{ij}, 0) P\{$$

by restricted similarity" method

 $+\sum_{i}\frac{\partial}{\partial\beta}\log[\pi(\beta;y_i)]$

— is based on the weighted similarity between two trees.

- he null hypothesis that $\beta = 0$, we h
 - A tree can be removed if it is "similar" to other trees in the forest.

$$\begin{split} \frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \eta_{ij}, 1)] - \gamma(0; \eta_{ij}, 1) - \gamma(0; \eta_{ij}, 1) - \gamma(0; \eta_{ij}, 1) - \gamma(0; \eta_{ij}, 1)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DE$$

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\} P^{-1}$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0) P\{c_i]$



$$P\{y_{i}\} = \frac{1}{f}$$
3; $k, \epsilon = P\{y_{ij} = k | \epsilon_{ij} = \epsilon \} = \gamma(\beta; k | K - 1, \gamma(\beta, 0, \epsilon) = 0, \text{ and } \gamma(\beta, K, E) = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} E] = \prod_{j} [\pi(\beta; y_{ij}, 0) P^{T}]$

"by restricted similarity" method

- ole to see that $(\partial/\partial \beta)\pi(\beta; k, \epsilon) =$ Evaluate the pairwise similarity of two trees in $\operatorname{og}(P\{M_i|y_i\}) = -\frac{\partial}{\partial \mathbf{g}} \operatorname{log}(P\{y_i\})$ $+\sum_{i}\frac{\partial}{\partial\beta}\log[\pi(\beta;y_i)]$
 - Select the pair of trees being most similar.

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] + \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] \end{split}$$

 $\mathcal{O}_{\text{BB}} \log[\pi(\beta; \chi_0, 0)] P\{dd[M_0]\}$ — Calculate \mathcal{O}_T for the two trees and the one with higher ρ_{τ} is subject to removal.

forest *F*, according to their predicted outcomes.

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_{j} [1 + \frac{\partial}{\partial \beta} \log P\{y_j\}]|_{\beta=0}$$

Distribute the weight of T to all other trees in $F_{(-T)}$, proportional to the pairwise similarity in \mathcal{P}_T .

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA - P\{dd, AA\} - P\{AA\}[P\{DE]$$

$$\begin{aligned} \{M_i|y_i\} &= \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P P\{M_i|y_i\} = \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta;y_{ij},0)P\{c_{ij},c_{ij} = 0\}P P\{M_i,c_{ij} = 0\}P\{M_i\} \prod_j [\pi(\beta;y_{ij},0)P\{c_{ij},c_{ij} = 0\}P P\{M_i,c_{ij} = 0\}P\{M_i\} \prod_j [\pi(\beta;y_{ij},c_{ij} = 0)P\{M_i,c_{ij} = 0\}P P\{M_i,c_{ij} = 0\}P\{M_i,c_{ij} = 0\}P P\{M_i,c_{ij} = 0\}P$$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K)$

$$P\{y_i\} = \prod_j [P\{y_{ij} | c_{ij} = 0\} P$$

Select the optimal size sub-forest

ple to see that $(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$ $og(P\{M_i|y_i\}) = -\frac{\partial}{\partial B}log(P\{y_i\})$

$$+ \sum_{i} \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]$$

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \, y_{ij}, \, \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{\theta}; \, y_{ij}, \, \boldsymbol{1}) - \gamma(\boldsymbol{\theta}; \, \boldsymbol{\theta})] \end{split}$$

$$\frac{\partial}{\partial \beta} {\log P\{y_i\}|}_{\beta=0} = \sum_j [1$$

- Let h(i), $i=1,...N_f$ -1, denote the performance trajectory of a sub-forest of *i* trees
 - N_f is the size of the original random forest.
- If we have only one realization of h(i), we select the optimal size sub-forest by maximizing h(i) over $i=1,...N_f-1.$
 - If we have multiple realizations of h(i), we select the optimal size sub-forest by using the 1-se rule.

The size of this smallest sub-forest is called the critical point of the performance trajectory.

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

$$\begin{split} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij}=0\} P P\{M_{i}|y_{i}\}] \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{c(\beta; k_{i}, c)}\}] \end{split}$$



3; k, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0$, and $\gamma(\beta, K, K)$

$$P\{y_i\} = \prod_j [P\{y_{ij} | q_j = 0\}]$$

$$= \prod_j [\pi(\beta; y_{ij}, 0)] P\{q_j = 0\}$$

ble to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

$$og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$$

$$+ \sum_i \frac{\partial}{\partial \beta}log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma] = \sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$$

$$|AA\rangle = P\{dd, |AA\rangle = P\{AA\}[P\{DD]$$

Simulation Designs

- For each data set, we generated 500 observations, each of which has one response variable and 30 predictors from Bernoulli distribution with success probability of 0.5.
- $+\sum_{\partial\beta}^{\partial} e^{i\pi(\beta)}$ Chose ν of the 30 variables to determine the response variable.

$$y = \begin{cases} 1, & \text{if } \sum_{i=1}^{\nu} X_i / \nu + \sigma > 0.5, \\ 0 & \text{Otherwise.} \end{cases}$$

- ullet Where σ is a random variable following the normal distribution with mean zero and variance.
- Considered two choices for ν (5 and 10) and two choices of σ (0.1 and 0.3).

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|z_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{z_i\}]$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta)$; k
 $K = 1$, $\gamma(\beta, 0, e) = 0$, and $\gamma(\beta, K, 0)$

$$\begin{split} P\{y_i\} &= \prod_j [P\{y_{ij} | a_{ij} = 0\} P] \\ &= \prod_j [\pi(\beta; y_{ij}, 0) P\{\beta\}] \end{split}$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=$$

he null hypothesis that
$$\beta = 0$$
, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{log}[\pi(\boldsymbol{\beta}; \ \mathbf{y}_{ij}, \ \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{\theta}; \ \mathbf{y}_{ij}, \ \boldsymbol{1}) - \gamma(\boldsymbol{\theta}; \ \end{split}$$

$$\frac{\partial}{\partial \mathbf{\beta}} \log P\{y_i\}|_{\mathbf{\beta}=0} = \sum_i [1$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA\} - P\{dd, AA\} - P\{AA\}[P\{DD]$$

To perform an unbiased comparison of the three tree removal measures, we simulated three independent data sets

- The training set is used to train the initial random forest
- The execution set is used to delete trees from the initial forest to produce sub-forests
- The evaluation set is used to evaluate the prediction performance of the sub-forests

The generation and use of these three data sets constituted one run of simulation, and we replicated 100 times.



B; k, ϵ) = $P\{y_{ij} = k | \epsilon_{ij} = \epsilon$) = $\gamma(\beta; I$ K - 1, $\gamma(\beta, 0, \epsilon) = 0$, and $\gamma(\beta, K)$

- Randomly selected one run of simulation and presented the stepwise change in the prediction performance in Figure 1.
- The "by prediction" method is preferable he null hypothesis that $\beta = 0$, we h
 - $= [1 \gamma(0; y_{ij}, 1) \gamma(0;$
 - $\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 \frac{\partial}{\partial \beta} \log P\{y_i\}]_{\beta=0}$
 - It can identify a critical point during the tree removal process in which the performance of the sub-forest deteriorates very rapidly.
- The performance of the sub-forests may begin to improve before the critical point.

$$AA\} = P\{dd,\,AA\} = P\{AA\}[P\{DD,\,AA\}] = P\{AA\}[P\{DD,\,$$



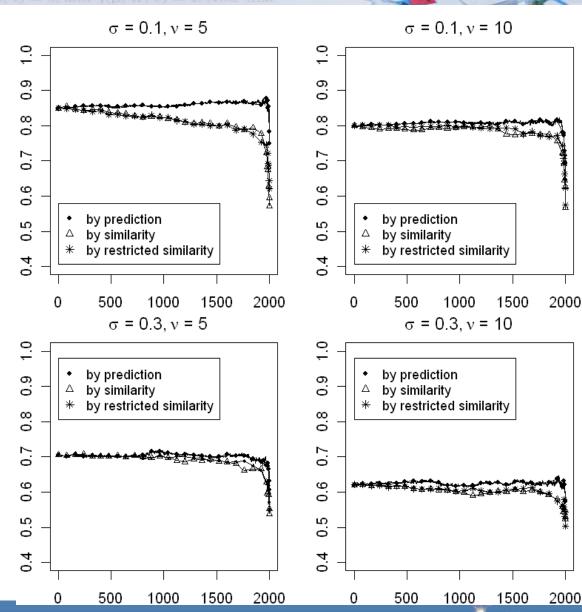
3;
$$k$$
, c) = $P\{y_{\theta} = k | c_{\theta} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, c) = 0, \text{ and } \gamma(\beta, K, c) = 0, \text{ and } \gamma(\beta, K, c)$

$$P\{y_i\} = \prod_i [P\{y_{ij} | c_{ij} = 0\}]$$

Prediction
performance of
sub-forests
produced from
different datasets
and methods

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1 - \frac{1}{2}]$$
 senience, we drop the two irrelevants
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

$$= \sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$$
 we coefficient of linkage disequilibrity,
$$AA\} - P\{dd, AA\} - P\{AA\}[P\{DD]$$



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e) = \gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, P\{y_{i}\} = \prod P\{y_{i} | g_{i} = 0\} \}$

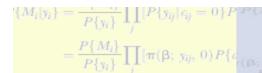
$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} I = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{,$$

- prediction performance using the results in five randomly selected runs.
- Although the variation of the trajectories is notable, the sizes of the optimal subforests are within a reasonable range (11-36) for the "by prediction" method.

 $=\sum_{j}\frac{1-\gamma(y_{ij})-\gamma}{P\{M_{ij}}$

e coemcient of tillkage tilsequinor

$$AA\} - P\{dd, AA\} - P\{AA\}[P\{DD]$$

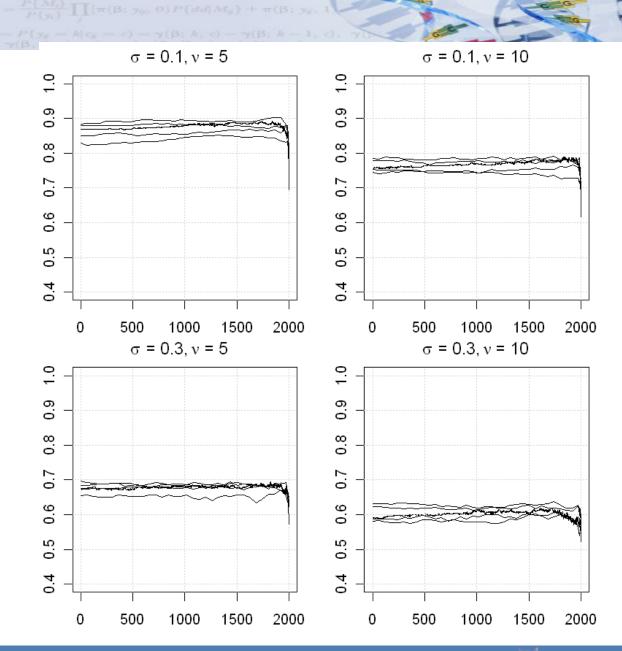


3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e) = \gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, P\{y_i\} = \prod_{j} [P\{y_{ij} | e_{ij} = 0\} F$

Performance
trajectory of the
"by prediction"
method using
the results in
five randomly
selected runs for
four data sets.

$$= \sum_{j} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$$

$$= \sum_{j} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$$
e coefficient of linkage disequilibrity, $AA\} - P\{dd, AA\} - P\{AA\}[P\{DE\}]$





$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}PP\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{j} = 0\}P\{dd\{M_{ij}\} + P\{y_{ij}|c_{j} = 0\}P\{dd\{M_{ij}\} + P\{y_{$$

3;
$$k$$
, e) = $P\{y_{\theta} = k | e_{\theta} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ an$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | q_{ij} = 0\} I \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P\{..., q_{ij}, q_{ij$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevar

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

The medians of the numbers of trees in the optimal sub-forests in 100 replications.

| σ | v | | |
|----------|------------|------------|--|
| | 5 | 10 | |
| 0.1 | 20(13, 29) | 31(20, 40) | |
| 0.3 | 22(15, 32) | 18(11, 37) | |



$$\begin{aligned} \{M_i | y_i\} &= \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij} | c_{ij} = 0\} P^2 \\ &= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0) P\{a\}] \end{aligned}$$



3;
$$k, c$$
) = $P\{y_{ij} = k | c_{ij} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, K) = 0\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} I = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{.$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) =$$

$$\begin{aligned} \log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\ &+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)] \end{aligned}$$

be null hypothesis that $\beta = 0$, we have

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

senience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA \} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

In practice, we generally have one data set only.

May not have the execution and evaluation data sets as in previous simulation.

How do we select the optimal sub-forest with only one data set?

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_j |\pi(\beta; y_{ij}, 0)P\{c_{ij}\}|$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) =$$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd] \underbrace{\delta}_{ij}\}$$

$$= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Considered four bootstrap-based approaches and examined them in simulated data sets.

We have the "golden" standard to be compared with in the simulated data set.





3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, 0, e) = 0\}$

$$P\{y_i\} = \prod_j [P\{y_{ij} | c_{ij} = 0\} P$$

$$=\prod_{j}[\pi(\beta;\,y_{ij}, \bigcirc)F$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) =$$

$$\begin{split} \log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\ &+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)] \end{split}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; \, y_{ij}, \, \boldsymbol{0}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{0}; \, y_{ij}, \, \boldsymbol{1}) - \gamma(\boldsymbol{0}; \, \boldsymbol{0})] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA\} - P\{dd, AA\} - P\{AA\}[P\{DD]$$

After constructing an initial forest using the whole data set as the training data set

- use one bootstrap data set for execution and the out-of-bag (oob) samples for evaluation.
- 一 use the oob samples for both execution and evaluation.
 - $\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_{i=1}^{|I|}$ use the bootstrap samples for both execution and evaluation.
- re-draw bootstrap samples for execution and redraw bootstrap samples for evaluation.

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|e_{ij} = 0\}P^{ij}$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{e_{ij}\}]$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0$, and $\gamma(\beta, K, 0, e) = 0$

$$P\{y_i\} = \prod_i [P\{y_{ij} | c_{ij} = 0\} P$$

$$= \prod_{j} [\pi(\beta; y_{ij}, \underline{\bullet}) P\{$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}\log(P\{y_i\})$$

 $+\sum \frac{\partial}{\partial \alpha}\log[\pi(\beta; y_i)]$

Figure 3 compares the performance of the four bootstrap-based approaches in the four simulation data sets.

he null hypothesis that $\beta = 0$, we

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\}]$$

$$= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)]$$

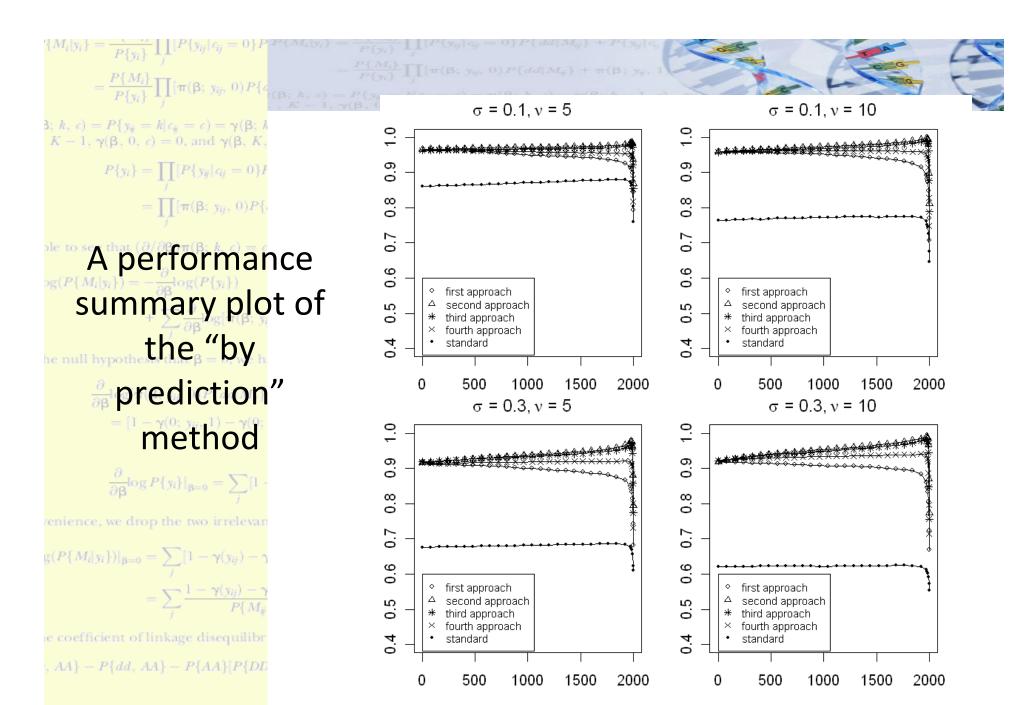
The comparison is based on the average performance in 100 runs.

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0} = \sum_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$



The Collaborative Center for Statistics in Science



$$\{M_i|y_i\} = \frac{\Box}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$



3;
$$k, c) = P\{y_{ij} = k | c_{ij} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P\{y_i\} = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{y_j\}] = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{y_j\}]$$

bootstrap-based approaches may not enull hypothesis that $\beta=0$, we have

forest, the similarity among the trajectories is most relevant, because it could lead to the same or similar sub-

e coefficient of linkage disequilibr

$$|AA| = P\{dd, |AA| = P\{AA\}[P\{DD]$$

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P]^{2} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}] \end{aligned}$$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] P\{y_i\} = \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{y_i\}$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) =$$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\}]$$

$$= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

enience, we drop the two irrelevan

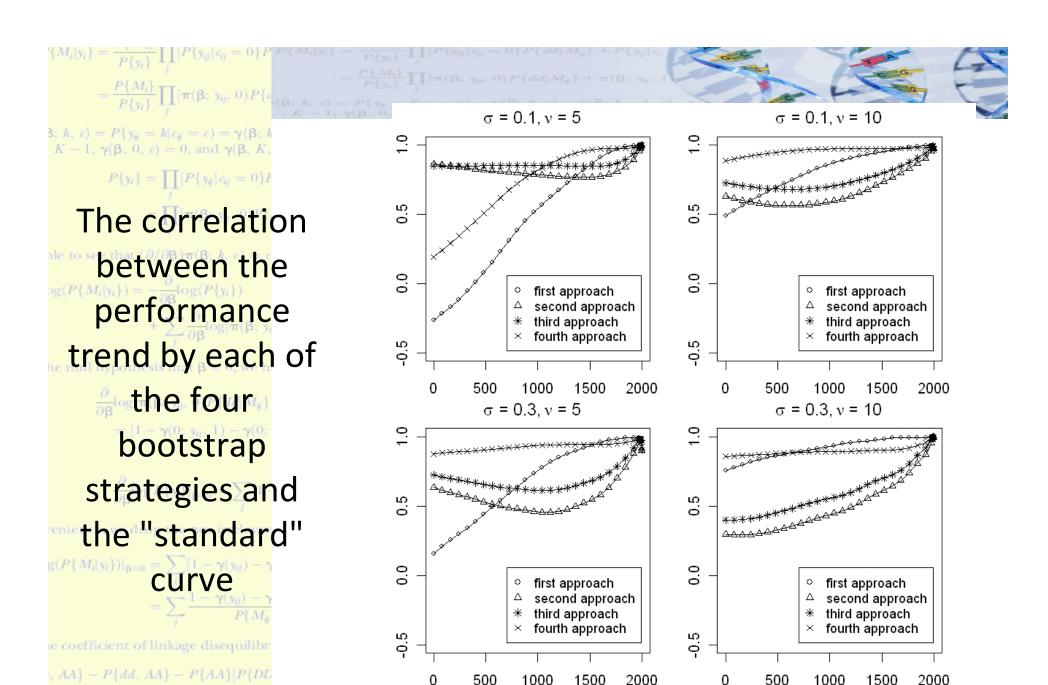
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

In Figure 4, we examined the correlation between the original (the "golden" standard) trajectory and each of the four bootstrap approaches.





$$\begin{split} \{M_{i}|y_{i}\} &= \frac{\Box}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P^{TP}] \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta;|y_{ij},|0)P\{c_{j},|\beta\}] \end{split}$$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, i)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

Using the bootstrap samples for

execution and the oob samples for evaluation is an effective sample-reuse

approach to selecting the optimal sub-

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Application



3;
$$k$$
, ϵ) = $P\{y_{ij} = k | \epsilon_{ij} = \epsilon\} = \gamma(\beta; k - 1, \gamma(\beta, 0, \epsilon) = 0, \text{ and } \gamma(\beta, K, \epsilon) = 0, \text{ and } \gamma(\beta, K, \epsilon) = 0, \text{ and } \gamma(\beta, K, \epsilon$

$$P\{y_i\} = \prod_{j} [P\{y_j | a_j = 0\}]$$

$$= \prod_{j} [\pi(\beta; y_j) P Dataset$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial\beta}\log(P\{y_i\})$$

$$+ \sum_i \frac{\partial}{\partial\beta}\log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\}]$$

- the microarray data set of a cohort of 295 young patients with breast cancer, containing expression profiles from 70 previously selected genes.
- previously studied by van de Vijver *et al*.

The responses of all patients are defined by whether the patients remained disease-free five years after their initial diagnoses or not.

AA = $P\{dd, AA\}$ = $P\{AA\}[P\{DE]$

$$\begin{split} \{M_i|y_i\} &= \frac{-1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij}=0\}P] \\ &= \frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta;y_{ij},0)P\{c_{ij}\}] \end{split}$$

$= \frac{1}{P\{y_i\}} \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] P} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{c_{ij}, c_{ij} = 0\}} - \frac{1}{P\{y_$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ and } \gamma($

$$P\{y_i\} = \prod_{j} [P\{y_{ij}|c_{ij} = 0\}]I$$

the to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

$$g(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta} log(P\{y_i\})$$

$$\pm \sum_{i} \frac{\partial}{\partial \beta} log[\pi(\beta; \cdot)]$$

The "by prediction" measure

- $\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial B}\log(P\{y_i\})$ The original data set to construct an initial forest
 - $+\sum_{i}\frac{\partial}{\partial \beta}\log[\pi(\beta;y]]$ A bootstrap data set for execution
 - The oob samples for evaluation.

he null hypothesis that $\beta = 0$, we h The procedure is replicated for a total of 100 times.

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- $\frac{1}{1-\gamma(0)(y_0,1)-\gamma(0)}$ The oob error rate is used to compare the performance of the initial random forest and the optimal sub-forest.
- The sizes of the optimal sub-forests fall in a relatively narrow range, of which the 1st quartile, the median, and the 3rd quartile are 13, 26 and 61, respectively.

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

$= \frac{1}{P\{y_i\}} \prod_{j} \frac{|P\{y_{ij}|c_{ij}=0\}P}{|P\{y_i\}|} \prod_{j} \frac{|\pi(\beta;y_{ij},0)P\{c_{ij}\}}{|P\{y_i\}|} \prod_{j} \frac{|\pi(\beta;y_{ij},0)P\{c_{ij}\}}{|P\{y_{ij}\}|} \prod_{j} \frac{|\pi(\beta;y_{ij},0)P\{c_{ij}\}}{|P\{y_{ij}\}$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, E)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

The smallest optimal sub-forest in the 100 repetitions with the size of 7 is $+\sum_{j}\frac{\partial}{\partial\beta}\log[\pi(\beta;y)]$ selected. he null hypothesis that $\beta = 0$, we

As a benchmark, we used the 70-gene classifier proposed by Vijver, et al.

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1]$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

$-\frac{P\{y_i\}}{P\{y_i\}}\prod_{j} P\{y_{ij}|a_{ij}=0\}P$ $-\frac{P\{M_i\}}{P\{y_i\}}\prod_{j} [\pi(\beta;y_{ij},0)P\{a_{ij}=0\}P\}$ $-\frac{P\{M_i\}}{P\{y_i\}}\prod_{j} [\pi(\beta;y_{ij},0)P\{a_{ij}=0\}P\}$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, E)$

$$P\{y_i\} = \prod_i [P\{y_{ij}|c_{ij} = 0\} P\{y_{ij}|c_{ij} = 0\}]$$

Table 2 presents the misclassification

ole to see that $(\partial/\partial \beta)\pi(\beta; k, c)$ rates based on the oob samples. $og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \mathbf{R}}log(P\{y_i\})$

$$+ \sum_{j} \frac{\partial}{\partial \beta} log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2} - \frac{1}{2}]$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$\{AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- The initial forest and the optimal sub-forest achieve almost the same level of performance accuracy.
- The 70-gene classifier has an out-of-bag error rate which is much higher than those of the forests.

Comparison of prediction performance of the initial random forest, the optimal sub-forest, and a previously established 70-gene classifier

| | Method | Error rate | True | Good | Poor |
|---|---|------------|-----------|------|------|
| | $\operatorname{sat} (\partial/\partial \beta)\pi(\beta; h, \epsilon) = \epsilon$ $\Delta = -\frac{\partial}{\partial \beta} \log(P\{y_i\})$ | | Predicted | | |
| | Random Forest | 26.0% | Good | 141 | 17 |
| he null hyp $\frac{\partial}{\partial \beta}$ l | othesis that $\beta = 0$, we h $ \log[\pi(\beta; y_{ij}, 0) P\{dd M_{ij}\} $ | | Poor | 53 | 58 |
| F | Sub-forest | 26.0% | Good | 146 | 22 |
| | $\frac{1}{\beta} \log P\{y_i\} _{\beta=0} = \sum_{j} [1 - \frac{1}{\beta}]$ e drop the two irrelevan | | Poor | 48 | 53 |
| $g(P\{M_i y_i\})$ | 70-gene Classifier | 35.3% | Good | 103 | 4 |
| e coefficier | $P\{M_{ij}\}$ at of linkage disequilibr | | Poor | 91 | 71 |

Application



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e)$

Main motivation

ble to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}\log(P\{y_i\})$$

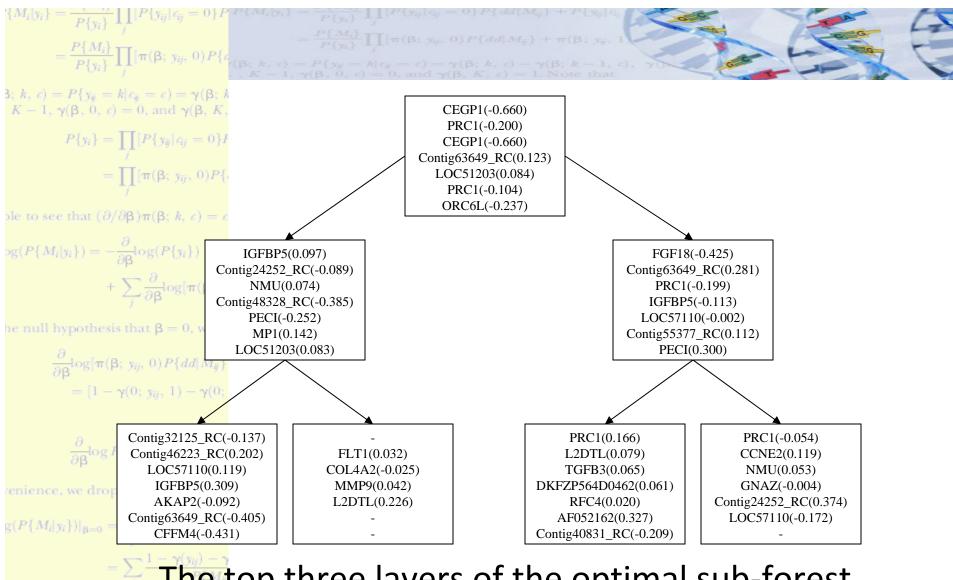
$$+ \sum_j \frac{\partial}{\partial \beta}\log[\pi(\beta; y_i)]$$

— seek the smallest possible forest to enable us to examine the forest.

Figure 5 displays the most critical part (the top three layers) of the optimal subforest consisting of the seven trees.

The selected genes are quite diverse and $\sum_{j=0}^{\lfloor (P\{M_{i}|y_{i}\})\rfloor_{\beta=0}} = \sum_{j=1}^{\lfloor (1-\gamma(y_{ij})-\gamma)\rfloor} unique.$ $= \sum_{j=1}^{\lfloor (1-\gamma(y_{ij})-\gamma)\rfloor} \frac{1-\gamma(y_{ij})-\gamma}{P\{M_{ij}\}}$

$$AA$$
 = $P{dd, AA} - P{AA}[P{DD}$



The top three layers of the optimal sub-forest consisting of seven trees

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Conclusion



3;
$$k, c) = P\{y_{ij} = k | c_{ij} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, -1, \gamma(\beta)) = 0, \text{ and } \gamma(\beta, K, -1, \gamma(\beta)) = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] = \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{x_{ij} = 0\}$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; h, e) = e$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial\beta}\log(P\{y_i\})$$

$$+ \sum_j \frac{\partial}{\partial\beta}\log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} \frac{\partial}{\partial \beta} &\log[\pi(\beta;\,y_{ij},\,0)P\{dd|M_{ij}\} \\ &= [1-\gamma(0;\,y_{ij},\,1)-\gamma(0;\, \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 -$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

It is possible to construct a highly accurate random forest consisting of a manageable number of trees.

- the size of the optimal sub-forest is in the range of tens
- some sub-forests can even over-perform the original forest in terms of prediction accuracy

The key advantage

the ability to examine and present the forests.

The limitation

- future samples and studies are needed to evaluate the performance of the forest-based classifiers.

The Collaborative Center for Statistics in Science



$$P\{M_{i}|y_{i}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P\{dd|M_{ij}\} + P\{y_{ij}|c_{ij}\}]$$
$$= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{dd|M_{ij}\} + \pi(\beta; y_{ij}, 1)]$$

ole to see that $(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$

og
$$(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta} log(P\{y_i\})$$

 $+ \sum_i \frac{\partial}{\partial \beta} log[\pi(\beta; y_i)]$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA$$
 = $P{dd, AA} - P{AA}[P{DE}$

Thank You!