Convex hierarchical models

Robert Tibshirani
Stanford University

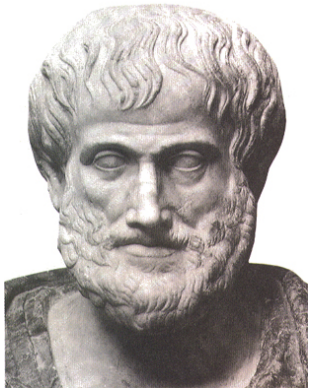
November 16, 2012

Topics

1. **Supervised learning**: extension of lasso to incorporate pairwise interactions in a hierarchical fashion. R package "hierNet"
   [PhD thesis of Jacob Bien; joint with Jon Taylor]
2. **Hypothesis testing:** application of same framework to large scale hypothesis testing of interactions. R package to come.
   [Joint with Jacob Bien and Noah Simon]

Talk Outline

Linear regression setting: Why Focus on Interactions?



"The whole is not the same as the sum of its parts."
– Aristotle

- In other words, sometimes an additive model is not enough!
- Biology: When gene A and gene B are expressed together, this is highly predictive of disease.

Interaction Models

Fitting regression models with interactions is challenging.

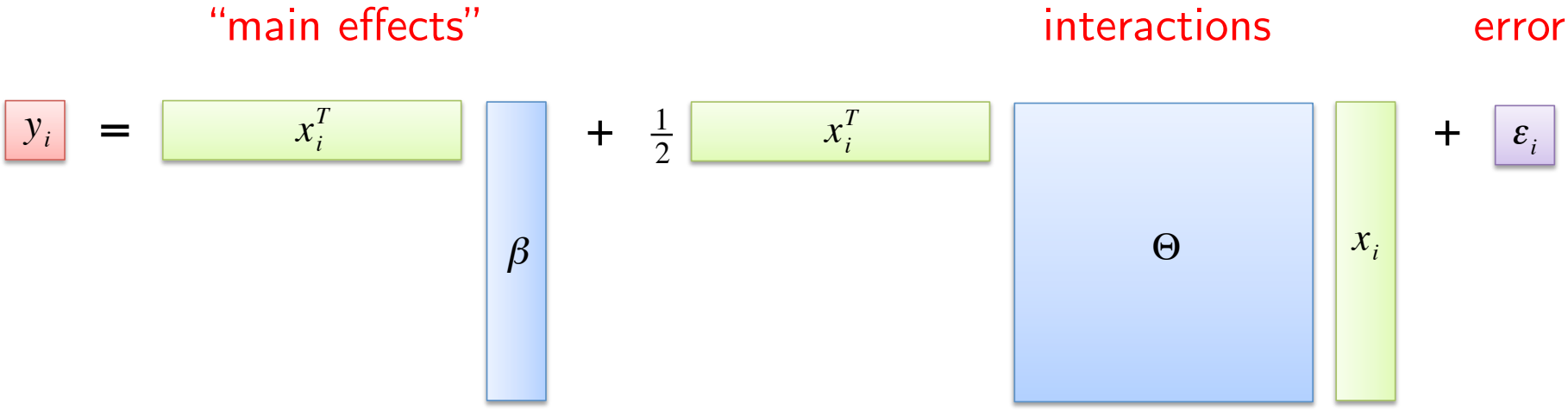
- p variables $\rightarrow \binom{p}{k}$ k-order interactions
- Which interactions are important? (Variable selection)
- Risk of overfitting

Often mentioned problem by biologists

First-Order Interaction Model

We focus in this talk on a first-order interactions model

$$Y = \sum_{j=1}^{p} \beta_j X_j + \sum_{j<k} \Theta_{jk} X_j X_k + \epsilon.$$

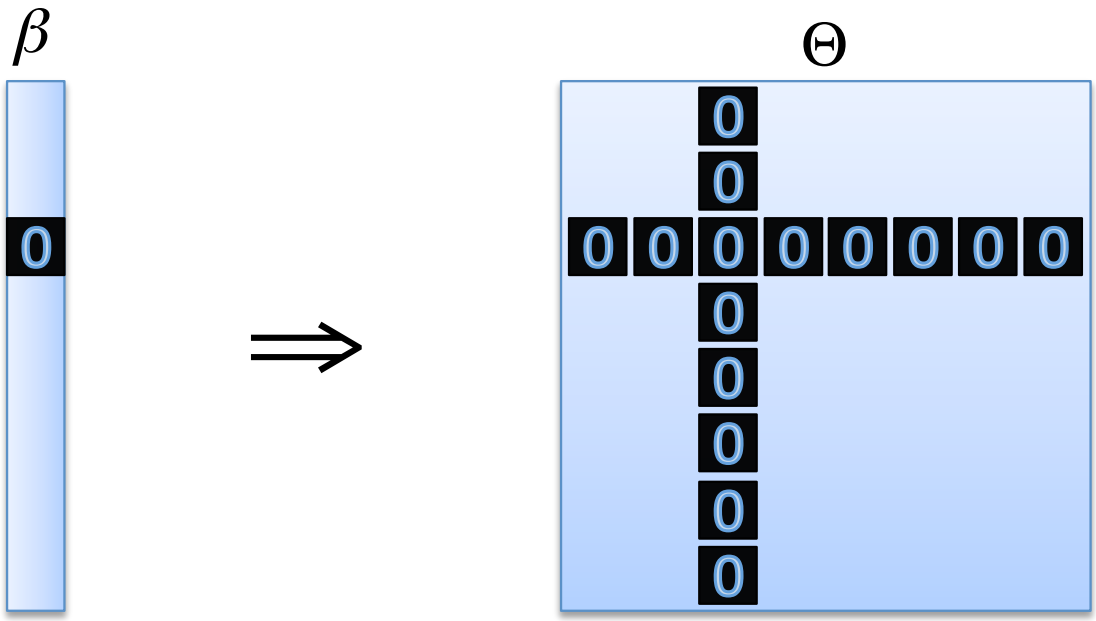


- X_j can be continuous or discrete.
- $\Theta \in \mathbb{R}^{p \times p}$ is a symmetric matrix.
- $\Theta_{jj} = 0$ for this talk.

Strong Hierarchy

Only include $X_j X_k$ if X_j and X_k are in model.

$$\beta_j = 0 \implies \begin{cases} \Theta_{j1} = \cdots = \Theta_{jp} = 0 \\ \Theta_{1j} = \cdots = \Theta_{pj} = 0 \end{cases}$$



- i.e. $\Theta_{jk} \neq 0 \implies \beta_j \neq 0$ AND $\beta_k \neq 0$
- **Weak hierarchy:** $\Theta_{jk} \neq 0 \implies \beta_j \neq 0$. Θ_{jk} not symmetric.

Why demand hierarchy?

> *"One general principle that can be used in such cases is that* **large component main effects are more likely to lead to appreciable interactions than small components.** *Also, the interactions corresponding to larger main effects may be in some sense of more practical importance."* *–D.R. Cox (International Statistical Review, 1984)*

Our method is based on this idea.

In Praise of Hierarchy

Potential advantages:

1. Statistical efficiency (Cox)
2. Computational efficiency
3. Cost effectiveness

Generic Approaches to Hierarchy

- Model selection **in two steps**.
  1. Choose main effects.
  2. Choose from allowed interactions.
- Model selection on all parameters **jointly**. Then force in main effects violating hierarchy.
- Our method chooses interactions and main effects together seamlessly in a single optimization while enforcing hierarchy restrictions

  Main effects guide interaction search and vice versa.

The Lasso

Observe n predictor-response pairs $(x_i,\ y_i)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Forming $X \in \mathbb{R}^{n \times p}$, with standardized columns, the **Lasso** is an estimator defined by the following optimization problem (Tibshirani 1996, Chen et al. 1998):

$$\underset{\beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - \beta_0 1 - X\beta\|^2 + \lambda\|\beta\|_1$$

- Penalty $\Longrightarrow$ sparsity (feature selection)
- Fit versus parsimony tradeoff
- Convex problem (good for computation and theory)
- Easier to analyze than a stepwise procedure

Talk Outline

The "All-Pairs Lasso"

A straightforward, non-hierarchical approach would be the following:

The "All-Pairs Lasso"

$$\underset{\beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^p,\ \Theta \in \mathbb{R}^{p \times p}}{\text{minimize}} \quad q(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1 \quad \text{s.t.} \quad \Theta = \Theta^T,$$

where

$$q(\beta_0, \beta, \Theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta - \frac{1}{2} x_i^T \Theta x_i)^2$$

Notes:

- Treats interaction and main effect variables identically
- Standardizing... multiple possibilities.

Our Proposal

Hierarchical Interactions Lasso

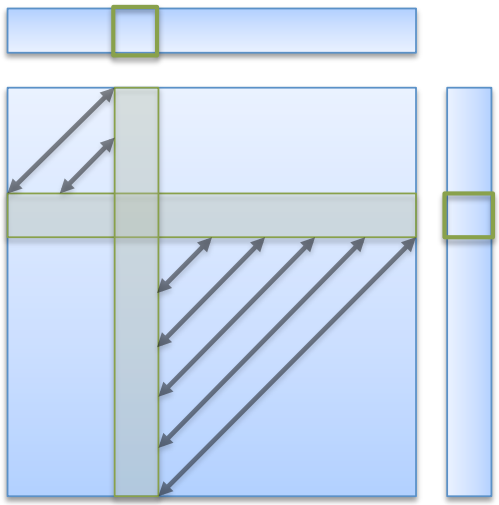
$$\underset{\beta_0 \in \mathbb{R},\ \beta^+, \beta^- \in \mathbb{R}^p,\ \Theta \in \mathbb{R}^{p \times p}}{\text{minimize}} \quad q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1$$

$$\text{s.t.} \quad \Theta = \Theta^T,$$

$$\left.\begin{array}{r} \|\Theta_j\|_1 \le \beta_j^+ + \beta_j^- \\ \beta_j^+ \ge 0, \beta_j^- \ge 0 \end{array}\right\} \text{for } j = 1, \ldots, p.$$

Just like the "All-Pairs Lasso" except that it

- writes main effects, β_j, as $\beta_j^+ - \beta_j^-$
- has additional constraints $\|\Theta_j\|_1 \le \beta_j^+ + \beta_j^-$



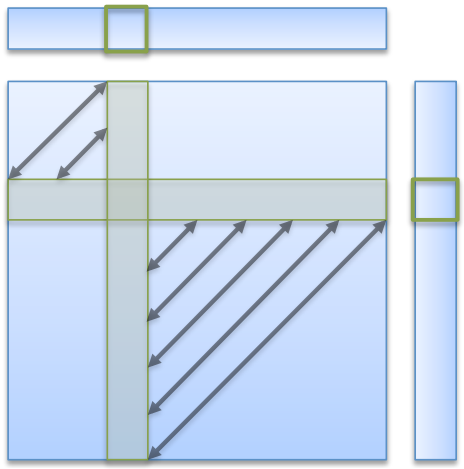
Motivation

Recall the All-Pairs Lasso:

$$\underset{\beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^p,\ \Theta \in \mathbb{R}^{p \times p}}{\text{minimize}} \quad q(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1 \quad \text{s.t.} \quad \Theta = \Theta^T,$$

Idea: Suppose we added constraints

$$\|\Theta_j\|_1 \leq |\beta_j| \text{ for } j = 1, \ldots, p.$$



Notes:

(+) $\Theta_{jk} \neq 0 \implies \beta_j \neq 0, \beta_k \neq 0$ (Strong hierarchy!).

(−) Not convex.

Our Proposal Again

The Hierarchical Lasso is a **convex relaxation** of the nonconvex version.

Hierarchical Interactions Lasso

$$\underset{\beta_0\in\mathbb{R},\ \beta^+,\beta^-\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{minimize}} \quad q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1$$

$$\text{s.t.} \quad \Theta = \Theta^T,$$

$$\left.\begin{array}{r} \|\Theta_j\|_1 \le \beta_j^+ + \beta_j^- \\ \beta_j^+ \ge 0, \beta_j^- \ge 0 \end{array}\right\} \text{for } j = 1, \ldots, p.$$

Replaced nonconvex $\|\Theta_j\|_1 \le |\beta_j|$ by convex $\|\Theta_j\|_1 \le \beta_j^+ + \beta_j^-$.

An Illustrative Example: Olive Oil

The Data:

- $n = 572$ olive oil samples
- $p = 8$ fatty acid concentrations
- binary response

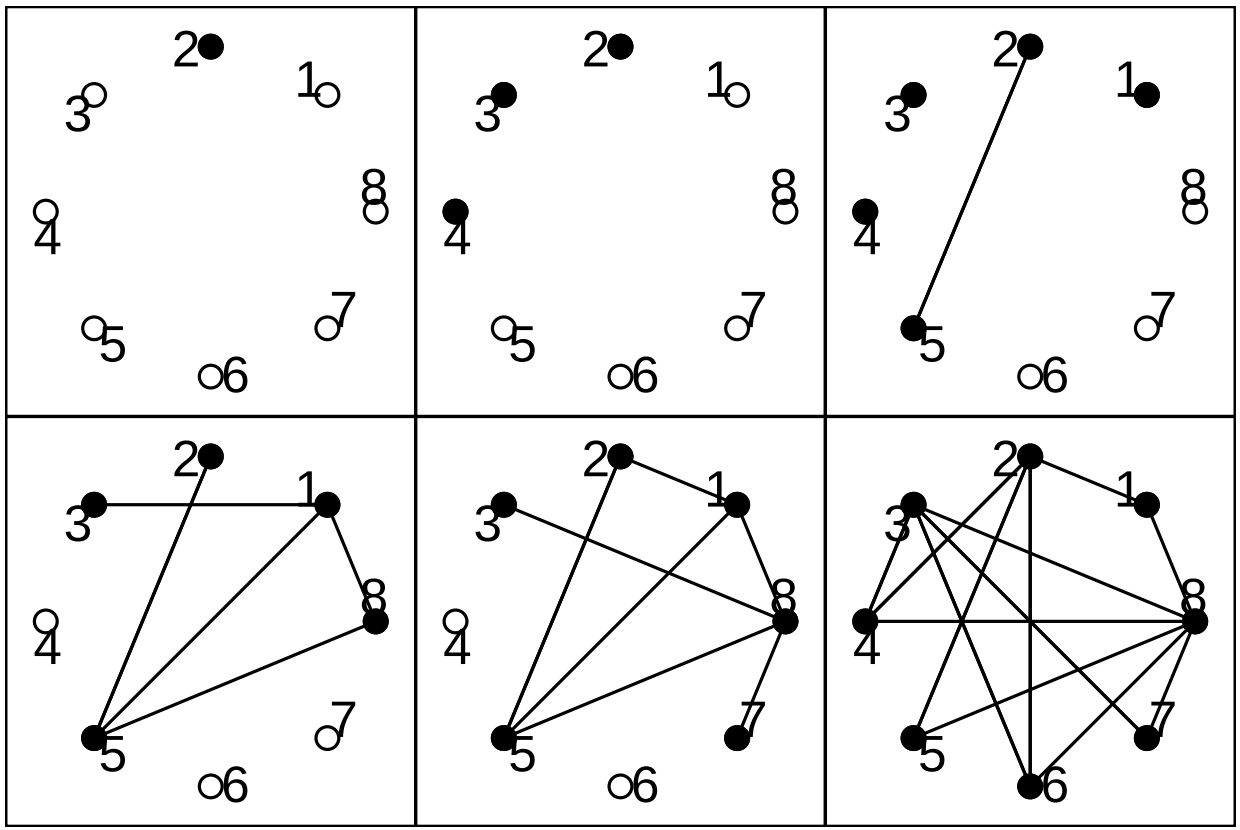
$$Y = \begin{cases} 1 & \text{olive oil is from Southern Apulia} \\ 0 & \text{otherwise} \end{cases}$$

The Methods:

- Main effects Lasso (MEL) with logistic loss: $X_1, \ldots, X_p$.
- All-Pairs Lasso (APL) with logistic loss: $X_1, \ldots, X_p, X_1X_2, \ldots, X_{p-1}X_p$.
- Hierarchical Lasso (HL) with logistic loss: Interactions with strong hierarchy.

An Illustrative Example: Olive Oil

Hierarchical Lasso (six values of λ)



An Illustrative Example: Olive Oil

Two notions of sparsity

Parameter sparsity

- Number of nonzero parameters in model
- Statistician's definition of "sparsity"
- Often related to degrees of freedom

Practical sparsity

- Number of raw variables, X_j, needed to make predictions
- Data collector's concern
- Related to cost, time, effort

Hierarchy favors models that "reuse" measured variables.

An Illustrative Example: Olive Oil

Two notions of sparsity

Parameter sparsity

- Number of nonzero parameters in model
- Statistician's definition of "sparsity"
- Often related to degrees of freedom

Practical sparsity

- Number of raw variables, X_j, needed to make predictions
- Data collector's concern
- Related to cost, time, effort

Hierarchy favors models that "reuse" measured variables.

An Illustrative Example: Olive Oil

Two notions of sparsity

Parameter sparsity

- Number of nonzero parameters in model
- Statistician's definition of "sparsity"
- Often related to degrees of freedom

Practical sparsity

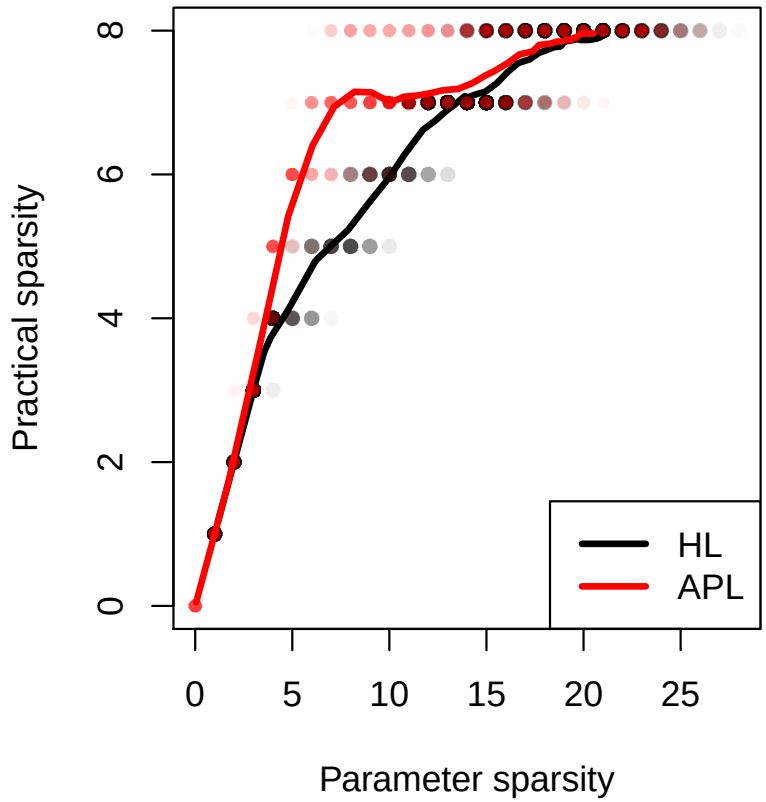
- Number of raw variables, X_j, needed to make predictions
- Data collector's concern
- Related to cost, time, effort

Hierarchy favors models that "reuse" measured variables.

An Illustrative Example: Olive Oil

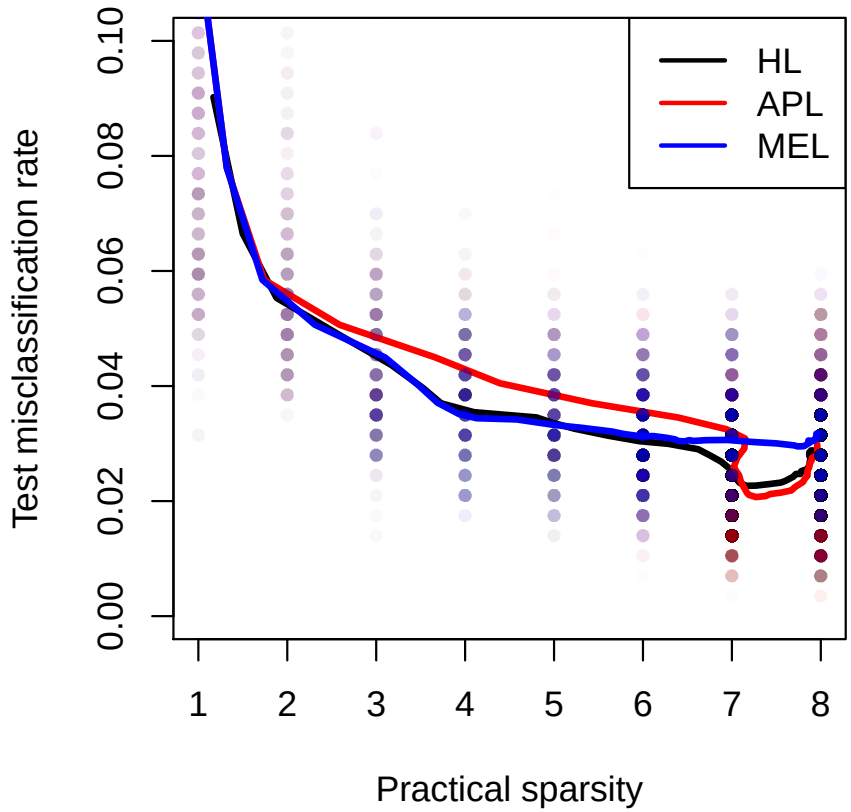
Two notions of sparsity

Lines show average over 100 random train-test splits.



An Illustrative Example: Olive Oil

Prediction error



Talk Outline

Effect of the constraint

The KKT conditions for our problem give a simple understanding of the effect of the constraint.

Brief background: Let $r^{(-j)} = y - \hat{y}^{(-j)}$ be the partial residual, and assume that $\|x_j\|^2 = 1$. The solution is a fixed point of the following:

Linear regression

$$\hat{\beta}_j = x_j^T r^{(-j)}$$

Lasso

$$\hat{\beta}_j = \mathcal{S}(x_j^T r^{(-j)},\ \lambda)$$

[i.e., soft-threshold by λ]

Effect of the constraint

The KKT conditions for our problem give a simple understanding of the effect of the constraint.

All-Pairs Lasso

$$\hat{\beta}_j = \mathcal{S}(x_j^T r^{(-j)},\ \lambda)$$

$$\hat{\Theta}_{jk} = \frac{\mathcal{S}[(x_j * x_k)^T r^{(-jk)},\ \lambda]}{\|x_j * x_k\|^2}$$

Hierarchical Lasso

$$\hat{\beta}_j = \mathcal{S}(x_j^T r^{(-j)},\ \lambda\ -\ \hat{\boldsymbol{\alpha}}_{\mathbf{j}})$$

$$\hat{\Theta}_{jk} = \frac{\mathcal{S}[(x_j * x_k)^T r^{(-jk)},\ \lambda\ +\ \hat{\boldsymbol{\alpha}}_{\mathbf{j}}\ +\ \hat{\boldsymbol{\alpha}}_{\mathbf{k}}]}{\|x_j * x_k\|^2}$$

where $\hat{\alpha}_j \geq 0$ and $\|\widehat{\Theta}_j\|_1 < \hat{\beta}_j^+ + \hat{\beta}_j^- \implies \hat{\alpha}_j = 0$.

Is our estimator hierarchical?

$$\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ + \hat{\beta}_j^- > 0 \stackrel{?}{\implies} \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0$$

Intuition:
For hierarchy to be violated, $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ must perfectly cancel, i.e. $\hat{\beta}_j^+ = \hat{\beta}_j^- > 0$.
Analogous to getting an exact zero in linear regression.

Theorem

Suppose y is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$. If $(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta})$ is a solution,[a] then strong hierarchy holds with probability 1, i.e.

$$\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0 \text{ AND } \hat{\beta}_k^+ - \hat{\beta}_k^- \neq 0.$$

[a] to our problem with an Elastic Net (Zou & Hastie 2005) penalty added to the objective

Is our estimator hierarchical?

$$\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ + \hat{\beta}_j^- > 0 \stackrel{?}{\implies} \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0$$

Intuition:
For hierarchy to be violated, $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ must perfectly cancel, i.e. $\hat{\beta}_j^+ = \hat{\beta}_j^- > 0$.
Analogous to getting an exact zero in linear regression.

Theorem

Suppose y is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$. If $(\hat{\beta}^+, \hat{\beta}^-, \widehat{\Theta})$ is a solution,[a] then strong hierarchy holds with probability 1, i.e.

$$\widehat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j^+ - \hat{\beta}_j^- \neq 0 \text{ AND } \hat{\beta}_k^+ - \hat{\beta}_k^- \neq 0.$$

[a] to our problem with an Elastic Net (Zou & Hastie 2005) penalty added to the objective

Our method

- Solve our convex problem along a grid of λ values.
- λ chosen by cross-validation.

Degrees of Freedom – a teaser

- *Lasso*: Degrees of freedom is # of nonzero parameters
- *HierNet:* Degrees of freedom is bounded above by total # of parameters minus # of main effects "forced in" by hierarchy constraint.

Talk Outline

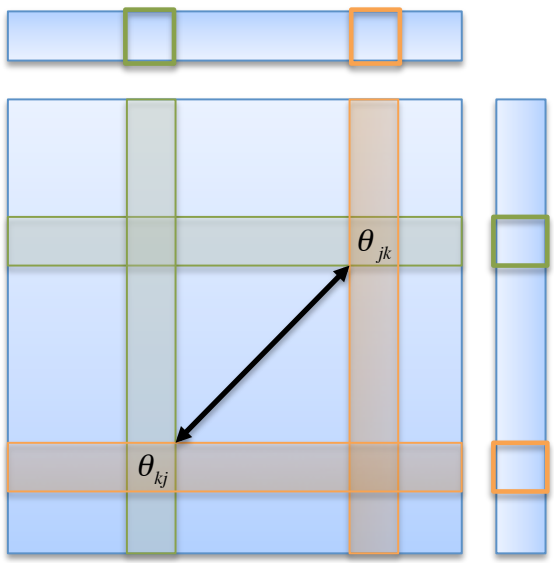
Computation

Why it is challenging

- Blockwise coordinate descent requires that the nondifferentiable part of the objective be separable (Tseng 2001).
- By symmetry, each hierarchy constraint is on both row and column.

$$\|\Theta_j\|_1 \leq \beta_j^+ + \beta_j^- \qquad \Theta = \Theta^T$$

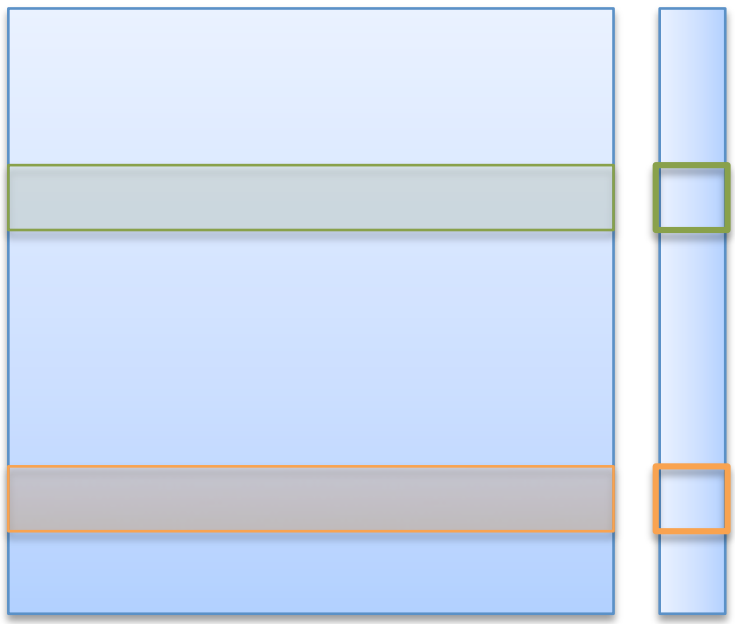
We have overlapping ℓ_1 constraints:



Computation

Notice that without the symmetry constraint, the constraints decouple:

$$\|\Theta_j\|_1 \leq \beta_j^+ + \beta_j^- \qquad \cancel{\Theta = \Theta^T}$$



In this case, optimization becomes much easier (e.g. blockwise coordinate descent works).

Our approach

- generalized gradient with ADMM to enforce symmetry
- faster coordinate descent/ADMM procedure under development

Simulation Study

30 simulations...

- $n = 100$ and $p = 30$ (435 pairs)
- β has 10 nonzeros.
- Θ has 20 nonzeros (i.e. 10 nonzero interactions).
- Signal-to-noise: ≈ 1.5 for main effects and ≈ 1 for interactions

Four scenarios:

1. Truth is hierarchical. **[main effects + interactions]**
2. Truth is not hierarchical. **[main effects + interactions]**
3. Truth has no main effects. **[interactions]**
4. Truth has no interactions. **[main effects]**

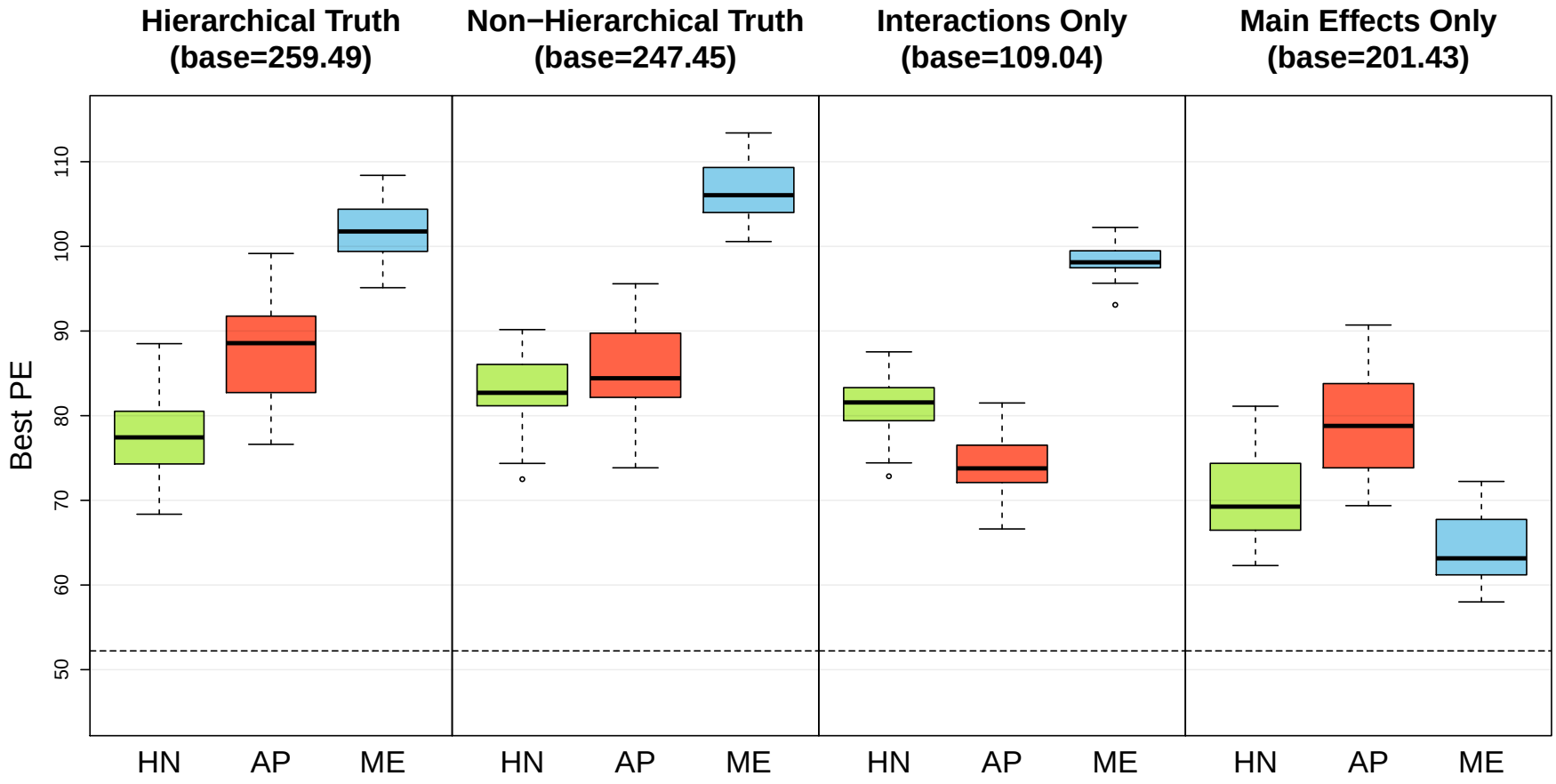
Three methods:

	Lasso
Hierarchical	*Our method*
All-Pairs	
Main Effects	

Simulation Results

Prediction Error

HN: HierNet
AP: All-Pairs Lasso
ME: Main Effects Lasso



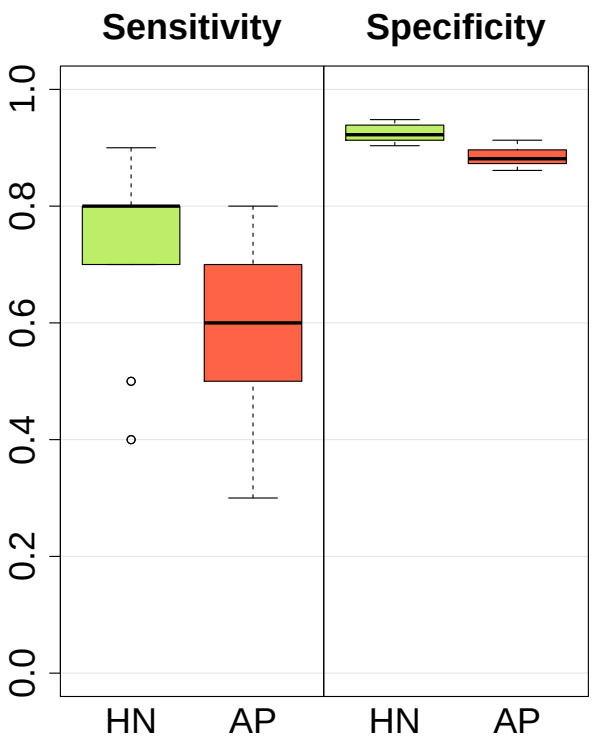
(based on 30 simulations, dashed line is Bayes error rate)

Simulation Results

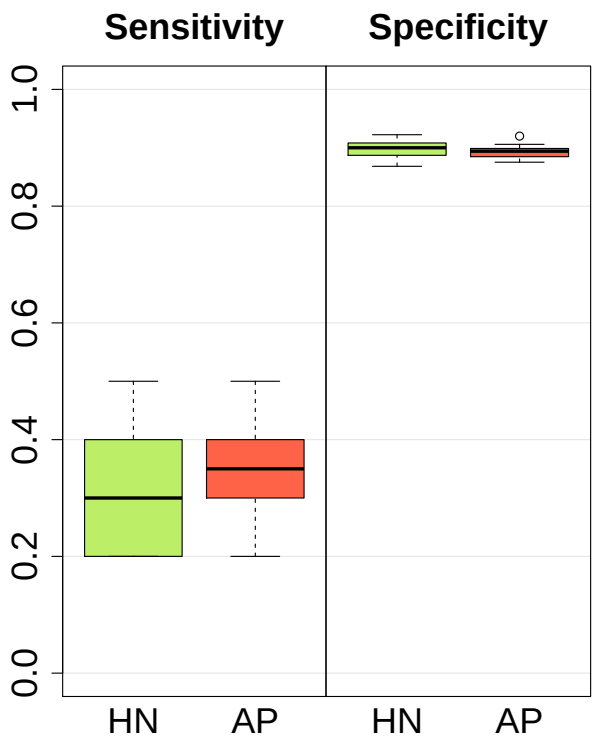
Ability to detect interactions

HN: HierNet
AP: All-Pairs Lasso

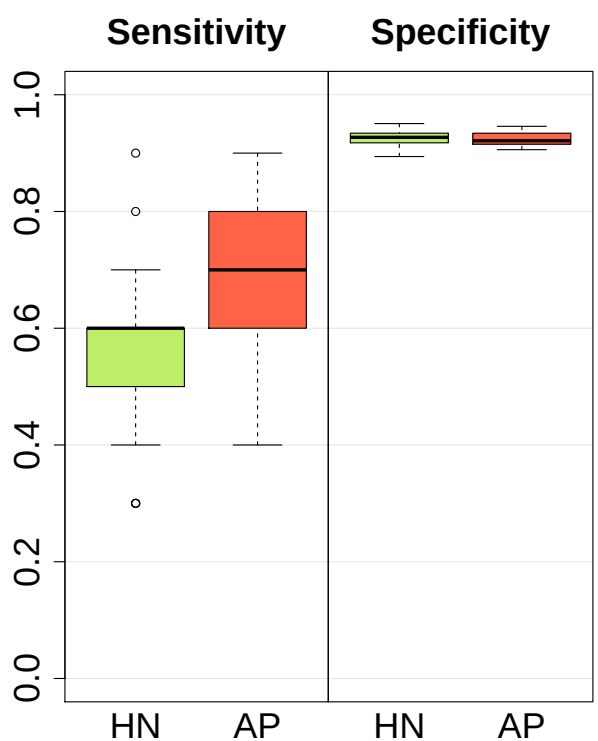
Hierarchical Truth



Non-Hierarchical Truth



Interactions Only



Talk Outline

Large scale testing of interactions

- We consider the standard two-class problem with $y_i = 1$ or 2 and p features $\{x_{i1}, x_{i2}, \ldots x_{ip}\}$ measured on each of $i = 1, 2, \ldots N$ observations.
- we are interested in testing all of the pairwise interactions of all features j, k in predicting the outcome y.
- Define the main effect contrasts by

$$w_j = \frac{\bar{x}_{j,1} - \bar{x}_{j,2}}{s_j \sqrt{1/n_1 + 1/n_2}}$$

- Define the interaction contrast z_{jk} to the standardized difference of the correlation between features j and k between the two groups
- Simon and Tibshirani (2012) show that z_{jk} is a reasonable measure of the interaction of j and k in a logistic regression model of y

Approaches to testing of interactions

- a standard test would order the interactions by their absolute value $|z_{jk}|$.
- we instead want to exploit a (weak) hierarchical assumption: the jkth interaction is more likely to be large if either of the main effects j or k is large
- could do a two stage screening approach. Instead we propose a seamless one stage method

Convex hierarchical test

- Consider minimizing the function

$$L_{\lambda_1,\lambda_2}(\{\beta_j^+\},\{\beta_j^-\},\{\theta_{jk}\}) = \frac{1}{2}\sum_{j=1}^{p}(w_j-(\beta_j^+-\beta_j^-))^2+\frac{1}{2}\sum_{j=1}^{p}\sum_{j\neq k}(z_{jk}-\theta_{jk})^2+ \lambda_1\sum_{j=1}^{p}[\beta_j^++\beta_j^-]+\lambda_2\sum_{j=1}^{p}\sum_{k\neq j}|\theta_{jk}| \quad (1)$$

  subject to the constraint that $\beta_j^{\pm}\geq 0$ (here we think of the main effect as $\hat{\beta}_j=\hat{\beta}_j^+-\hat{\beta}_j^-$).
- We add the weak hierarchy constraints $\sum_k|\theta_{jk}|\leq\beta_j^++\beta_j^-$.
- We also set $\lambda_1=\lambda_2=\lambda$ for simplicity

The resulting test

- Our idea is to fit a path of models (parameterized by λ) and then define the test statistic for the jkth interaction to be $\hat{\lambda}'_{jk}$, the largest λ for which either $\hat{\theta}_{jk}$ or $\hat{\theta}_{kj}$ is nonzero.
- In the same way, for each main effect j we compute $\hat{\lambda}_j$, the largest λ for which either $\hat{\beta}_j^+$ or $\hat{\beta}_j^-$ is non-zero.
- Note that without hierarchy constraint, $\hat{\lambda}'_{jk} = |\hat{z}_{jk}|$

Properties

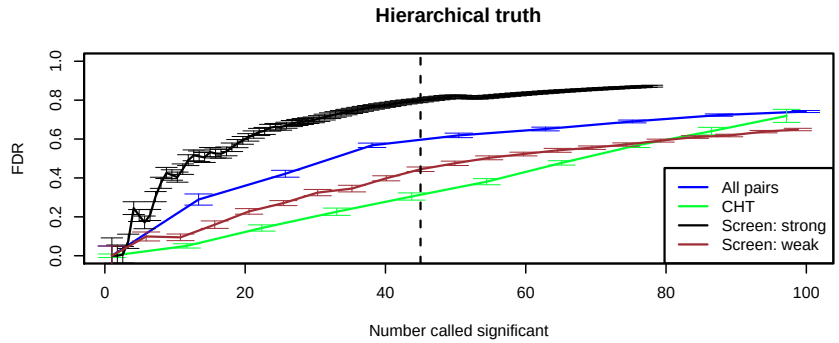
- Once $\hat{\lambda}_{jk}$ becomes non-zero (as λ decreases) it stays non-zero. Hence the test "makes sense"
- After MANY pages of algebra, we show that the main effect and interaction test statistics have the following closed-form expressions:

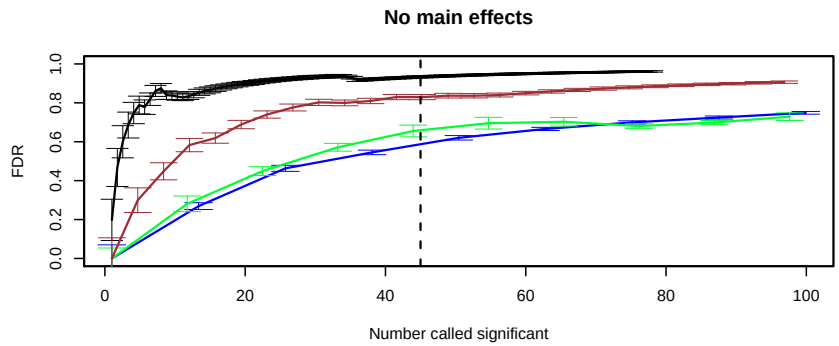
$$\hat{\lambda}_j = \max\left\{|w_j|, \frac{|w_j| + \|z_{j\cdot}\|_\infty}{2}\right\}$$

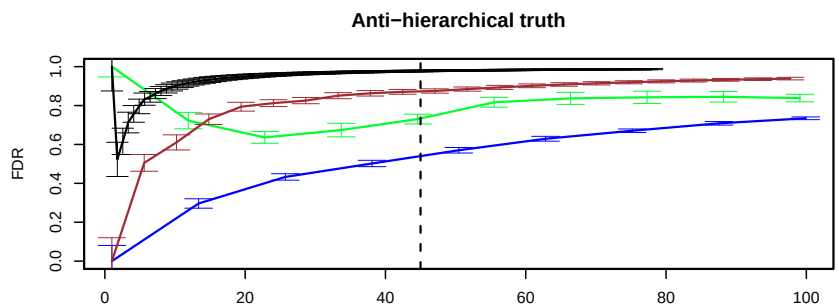
$$\hat{\lambda}_{jk} = \min\left\{|z_{jk}|, \frac{|z_{jk}|}{2} + \frac{\left[|w_j| - \sum_{l:|z_{jl}|>|z_{jk}|}\left(|z_{jl}| - |z_{jk}|\right)\right]_+}{2}\right\}, \tag{2}$$

where $z_{j\cdot} = \{z_{jk} : k \neq j\} \in \mathbb{R}^{p-1}$ is the vector of interaction contrasts involving the jth variable

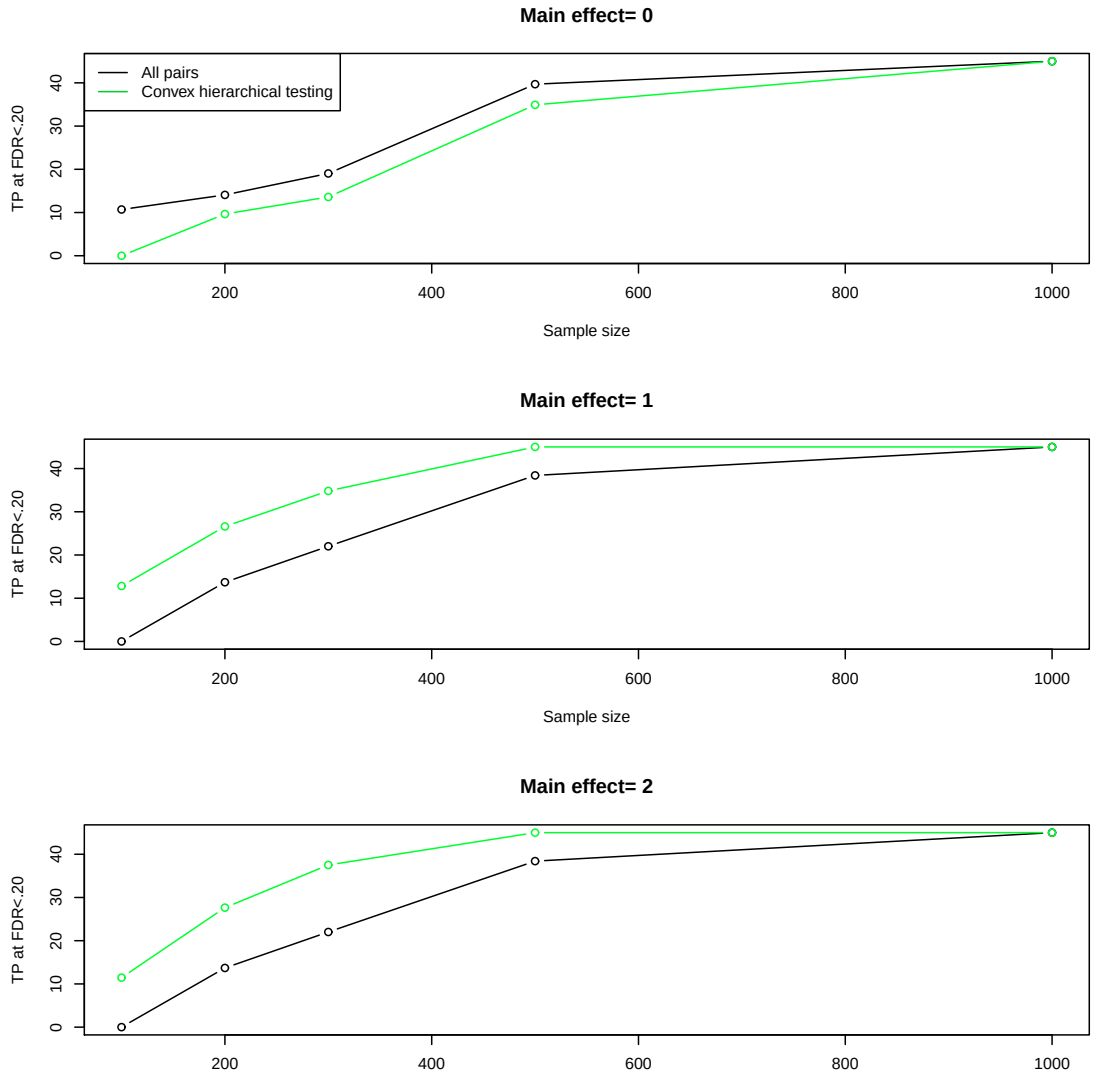
Simulations of False Discovery rates







Average number of True Positives (non-null interactions)



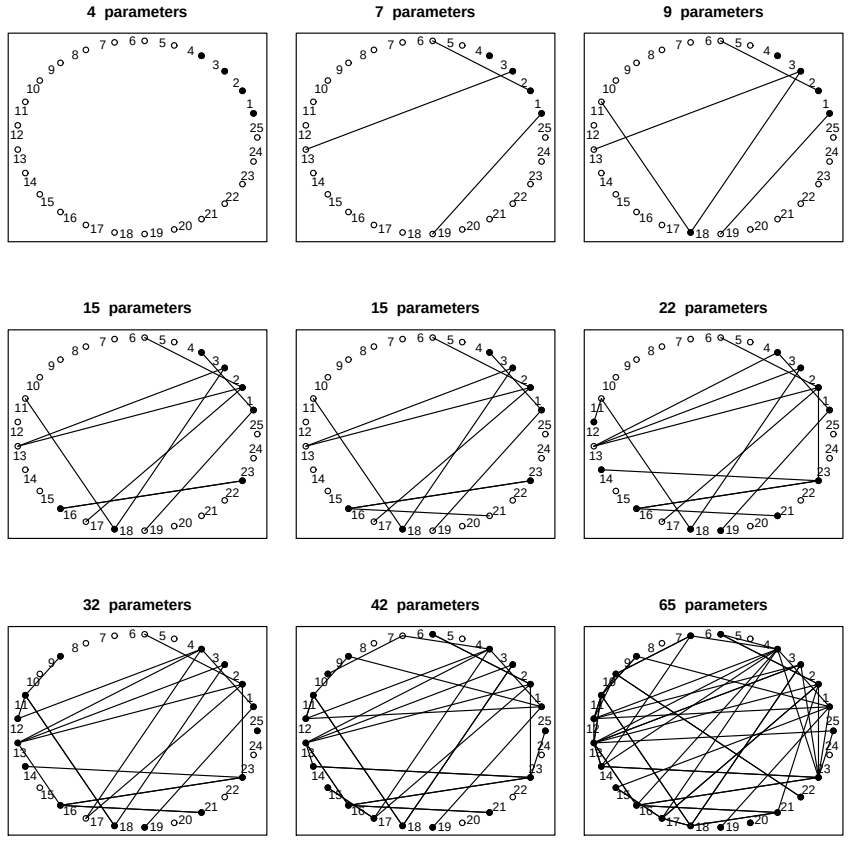
Hypertension data

Predictor number	Name	Predictor number	Name
1	Reached menopause?	14	PTPN1i4INV
2	insulin t=-10	15	Cyp11B2x1INV
3	insulin t=60	16	PTPN1x9INV
4	insulin t=120	17	ADRB3W1R
5	HUT2SNP5	18	KLKQ3E
6	HUT2SNP7	19	AGT2R1A1166C
7	BADG16R	20	AVPR2G12E
8	AVPR2A1629G	21	MLRI2V
9	AGT2R2C1333T	22	AGTG6A
10	PPARG12	23	Cyp11B2-5paINV
11	CD36x2aINV	24	PTPN1i1
12	MLRi6INV	25	PTPN1i4
13	Cyp11B2i4INV		

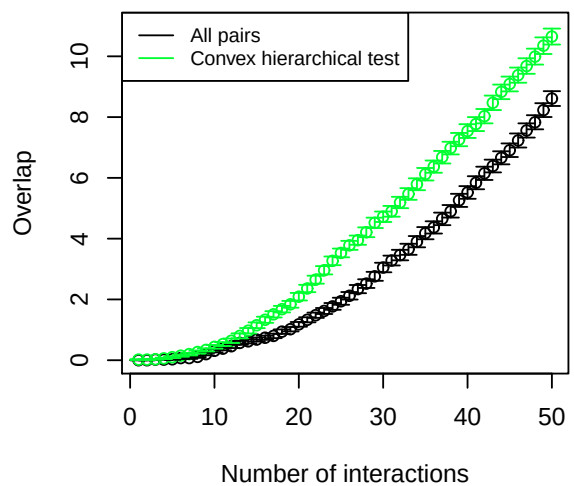
Top ten interactions found

All pairs	
Predictor 1	Predictor 2
Cyp11B2x1INV	AGTG6A
PPARG12	ADRB3W1R
PPARG12	CD36x2aINV
MLRi6INV	Cyp11B2x1INV
Cyp11B2i4INV	AGTG6A
Cyp11B2x1INV	AVPR2G12E
AGTG6A	PTPN1i1
Affected status	KLKQ3E
Reached menopause?	HUT2SNP5
AVPR2A1629G	PPARG12
Convex hierarchical test	
Reached menopause?	HUT2SNP5
insulin t=-10	MLRi6INV
Affected status	insulin t=60
Affected status	KLKQ3E
Reached menopause?	MLRi6INV
Reached menopause?	PTPN1x9INV
insulin t=-10	ADRB3W1R
Reached menopause?	AGTG6A

Wheel plot


4 parameters
7 parameters
9 parameters
15 parameters
15 parameters
22 parameters
32 parameters
42 parameters
65 parameters

Prop of ints found in both random halves of the data



More...

- Estimation of FDR
- Optimality?

Thanks for listening!

Chen, S. S., Donoho, D. L. & Saunders, M. A. (1998), 'Atomic decomposition by basis pursuit', *SIAM Journal on Scientific Computing* pp. 33–61.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B* **58**(1), 267–288.

Tseng, P. (2001), 'Convergence of block coordinate descent method for nondifferentiable maximization', *J. Opt. Theory and Applications* **109**(3), 474–494.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B.* **67**(2), 301–320.

References

Chen, S. S., Donoho, D. L. & Saunders, M. A. (1998), 'Atomic decomposition by basis pursuit', *SIAM Journal on Scientific Computing* pp. 33–61.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B* **58**(1), 267–288.

Tseng, P. (2001), 'Convergence of block coordinate descent method for nondifferentiable maximization', *J. Opt. Theory and Applications* **109**(3), 474–494.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B.* **67**(2), 301–320.