An Inferential Perspective on Data Depth

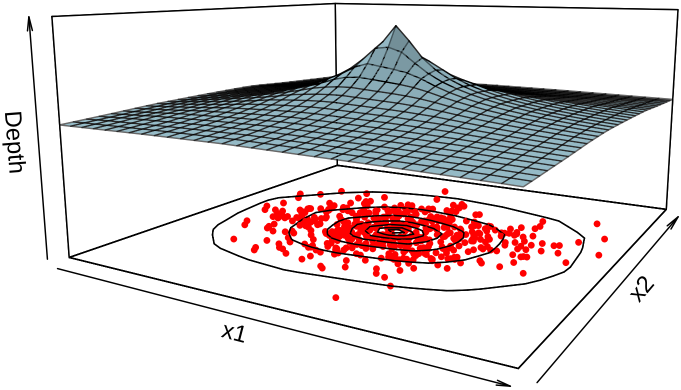
Subhabrata Majumdar

under supervision of
Snigdhansu Chatterjee
University of Minnesota, School of Statistics

May 18, 2017

What is depth?

Example: 500 points from $\mathcal{N}_2((0,0)^T, \text{diag}(2,1))$



A scalar measure of how much inside a point is with respect to a data cloud

Table of contents

1. Motivation
2. Signed Peripherality Functions
3. Nonconvex Penalized Multitask Regression using Depth-based Penalty
4. Generalized model discovery using statistical evaluation maps
5. Selecting important SNPs from multi-SNP mixed models on Twin Studies GWAS data

Outline

1. **Motivation**
2. Signed Peripherality Functions
3. Nonconvex Penalized Multitask Regression using Depth-based Penalty
4. Generalized model discovery using statistical evaluation maps
5. Selecting important SNPs from multi-SNP mixed models on Twin Studies GWAS data

Formal definition of depth

For any multivariate distribution $F = F_{\mathbf{X}}$, the depth of a point $\mathbf{x} \in \mathbb{R}^p$, say $D(\mathbf{x}, F_{\mathbf{X}})$ is any real-valued function that provides a 'center outward ordering' of $\mathbf{x}$ with respect to F (Zuo and Serfling, 2000).

Desirable properties (Liu, 1990)

(P1) *Affine invariance*: $D(A\mathbf{x} + \mathbf{b}, F_{A\mathbf{X}+\mathbf{b}}) = D(\mathbf{x}, F_{\mathbf{X}})$

(P2) *Maximality at center*: $D(\theta, F_{\mathbf{X}}) = \sup_{\mathbf{x} \in \mathbb{R}^p} D(\mathbf{x}, F_{\mathbf{X}})$ for $F_{\mathbf{X}}$ with center of symmetry θ, the *deepest point* of $F_{\mathbf{X}}$.

(P3) *Monotonicity w.r.t. deepest point*: $D(\mathbf{x}; F_{\mathbf{X}}) \leq D(\theta + a(\mathbf{x} - \theta), F_{\mathbf{X}})$

(P4) *Vanishing at infinity*: $D(\mathbf{x}; F_{\mathbf{X}}) \to \mathbf{0}$ as $\|\mathbf{x}\| \to \infty$.

Examples and utility

- **Halfspace depth** (HD) (Tukey, 1975) is the minimum probability of all halfspaces containing a point.

$$HD(\mathbf{x}, F) = \inf_{\mathbf{u} \in \mathbb{R}^p; \mathbf{u} \neq \mathbf{0}} P(\mathbf{u}^T \mathbf{X} \geq \mathbf{u}^T \mathbf{x})$$

- **Projection depth** (PD) (Zuo, 2003) is based on an outlyingness function:

$$O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T \mathbf{x} - m(\mathbf{u}^T \mathbf{X})|}{s(\mathbf{u}^T \mathbf{X})}; \quad PD(\mathbf{x}, F) = \frac{1}{1 + O(\mathbf{x}, F)}$$

- Used extensively for classification, robust estimation of outlyingness, L-estimation of location and scale, hypothesis testing;

Objective

Although the nonparametric concept of data depth has gained visibility and has seen many applications in recent years, its utility in achieving traditional parametric inferential goals is largely unexplored. In this proposal we develop different approaches to address this.

- Signed Peripherality Functions: robust location and scale inference;
- Nonconvex Penalized Multitask Regression using Depth-based Penalty;
- **Generalized Model Discovery using Statistical Evaluation Maps: applications in Indian Monsoon, fMRI and Minnesota Twin Studies data**.

Outline

1. Motivation
2. **Signed Peripherality Functions**
3. Nonconvex Penalized Multitask Regression using Depth-based Penalty
4. Generalized model discovery using statistical evaluation maps
5. Selecting important SNPs from multi-SNP mixed models on Twin Studies GWAS data

Signed peripherality maps

Spatial signs (Locantore et al., 1999):

$$\mathbf{S}(\mathbf{x}) = \begin{cases} \mathbf{x}\|\mathbf{x}\|^{-1} & \text{if } \mathbf{x} \neq \mathbf{0} \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0} \end{cases}$$

Fix a depth function $D(\mathbf{x}, F) = D(\mathbf{x}, [\mathbf{X}])$. Define the signed peripherality as

$$\kappa(D(\mathbf{x}, [\mathbf{X}]).\mathbf{S}(\mathbf{x})$$

where $\kappa : [0, \infty) \mapsto [0, \infty)$ is a bounded monotone transformation.

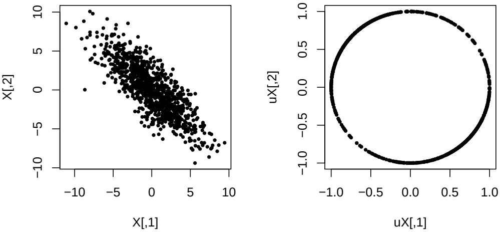
Two cases-

(1) κ **is increasing:** Robust location inference. Improves on existing estimators for robust location estimates and high-dimensional testing;

(1) κ **is decreasing:** Robust scale inference.

Spatial sign-based PCA (Locantore et al., 1999)

- Say $\mathbf{x}$ follows an elliptic distribution with mean μ, covariance matrix $\mathbf{\Sigma}$.
- Sign covariance matrix (SCM): $\mathbf{\Sigma}_S = \mathbb{E}\mathbf{S}(\mathbf{X} - \mu)\mathbf{S}(\mathbf{X} - \mu)^T$
- SCM has same eigenvectors as $\mathbf{\Sigma}$. PCA using SCM is robust, but not efficient.



Spatial ranks

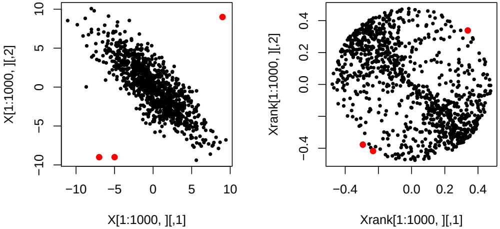
- Transform the original observation

$$\tilde{\mathbf{x}} = D^{-}(\mathbf{x}, [\mathbf{X}])\mathbf{S}(\mathbf{x} - \mu)$$

  where $D^{-}(\mathbf{x}, [\mathbf{X}])$ is the *inverse depth* of **x**: any bounded nonnegative-valued monotone transformation on the depth function.

  This is the *Spatial Rank* of **x**.

- Depth Covariance Matrix (DCM) $\tilde{\mathbf{\Sigma}} := \mathbb{V}(\tilde{\mathbf{X}})$. Has more information than spatial signs, so more efficient.



Results using spatial ranks

- Recovery of population eigenvectors and eigenvalues using DCM: asymptotic and robustness properties of estimates;
- Adaptations in Sufficient Dimension Reduction and functional PCA;
- Simulations and real data applications.

Outline

1. Motivation
2. Signed Peripherality Functions
3. **Nonconvex Penalized Multitask Regression using Depth-based Penalty**
4. Generalized model discovery using statistical evaluation maps
5. Selecting important SNPs from multi-SNP mixed models on Twin Studies GWAS data

Penalized multitask regression

Consider the multitask linear regression model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times q}$ is the matrix of responses, and $\mathbf{E}$ is $n \times q$ the noise matrix: each row of which is drawn from $\mathcal{N}_q(\mathbf{0}_q, \boldsymbol{\Sigma})$ for a $q \times q$ positive definite matrix $\boldsymbol{\Sigma}$.

We are interested in sparse estimates of the coefficient matrix $\mathbf{B}$ through solving penalized regression problems of the form

$$\min_{\mathbf{B}} \text{Tr}\{(\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB})\} + P_\lambda(\mathbf{B}). \tag{1}$$

Our estimator

We incorporate measures of data depth as a row-level penalty function. Specifically, we estimate the coefficient matrix **B** by solving the following constrained optimization problem:

$$\hat{\mathbf{B}} = \underset{\mathbf{B}}{\text{argmin}} \left[\text{Tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{B})^T (\mathbf{Y} - \mathbf{X}\mathbf{B})\} + \lambda \sum_{j=1}^{p} D^-(\mathbf{b}_j, F) \right]$$

where $D^-(\mathbf{x}, F)$ is an inverse depth function.

The idea

- Regularization penalties can be interpreted as distance from the origin.
- We want to use penalties that are ‘distance from a distribution centered at the origin’.
- The distribution F is fixed at the start of the modelling process, and can represent a prior belief on how the different responses are related among themselves.
- Two advantages: (1) penalty attains nonconvex shape by inverting the depth function, (2) has a natural bayesian interpretation.
- For now we assume F to be spherically symmetric - plausible from a frequentist perspective.

Simplifications

- As F to be spherically symmetric, D^- becomes a function of the row-norm $r_j = \|\mathbf{b}_j\|_2$: $D^-(\mathbf{b}_j, F) = p_F(r_j)$.
- Use the first order Taylor approximation around a 'close enough' point r_j^* instead of $p_F(r_j)$. This is local linear approximation (Zou and Li, 2008):

$$p_F(r_j) \simeq p_F(r_j^*) + p'_F(r_j^*)(r_j - r_j^*)$$

Thus the modified solution is:

$$\hat{\mathbf{B}}^{(1)} = \underset{\mathbf{B}}{\text{argmin}} \left[\text{Tr}\{(\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB})\} + \lambda \sum_{j=1}^{p} p'_F(r_j^*) r_j \right]$$

The close enough matrix $\mathbf{B}^*$ to start from can be the least squares estimate. We call this *Local Approximation by Row Norm* (LARN).

Overview of work

- Derived theoretical properties: oracle property, near-minimax optimal performance, post-estimation thresholding to recover within row support;
- A block coordinate descent algorithm to compute solution;
- Simulation and data example.

Outline

1. Motivation
2. Signed Peripherality Functions
3. Nonconvex Penalized Multitask Regression using Depth-based Penalty
4. **Generalized model discovery using statistical evaluation maps**
5. Selecting important SNPs from multi-SNP mixed models on Twin Studies GWAS data

The idea

- In a parametric modelling setup, any candidate model is a subset of the parameter space;
- We have a collection of models, and want to classify them as ‘good’ or ‘bad’ based on if they match with a baseline model with respect to a predefined criterion;
- We shall compare sampling distributions of (potentially transformed) parameter estimates of a candidate model with that of a baseline model using a generic quantity called the ***e*-value**.

Implementation: a fast variable selection ‘algorithm’

Linear regression: $Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$. Assume that there is a true parameter vector $\boldsymbol{\beta}_0$, with non-zero index set $\mathcal{S}_0$.

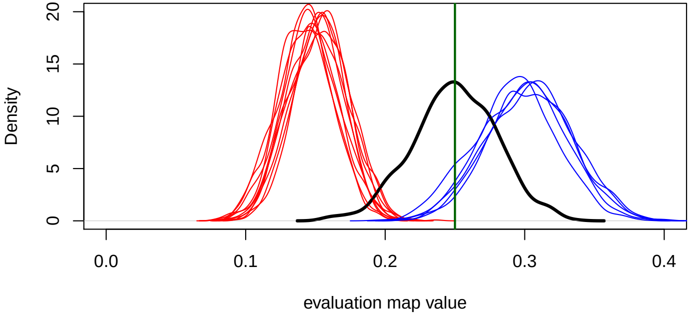
1. Get $\hat{\boldsymbol{\beta}}$. Obtain its bootstrap distribution: $[\hat{\boldsymbol{\beta}}]$;
2. Replace the j-th coefficient with 0, name it $\hat{\boldsymbol{\beta}}_{-j}$. Do the same for its bootstrap distribution, say $[\hat{\boldsymbol{\beta}}_{-j}]$. Repeat for all j;
3. e-value of j-th covariate = mean depth of $\hat{\boldsymbol{\beta}}_{-j}$ with respect to $[\hat{\boldsymbol{\beta}}]$, i.e. $\mathbb{E}D(\hat{\boldsymbol{\beta}}_{-j}, [\hat{\boldsymbol{\beta}}])$;
4. Select covariates with e-value less than mean depth of full model:

$$\hat{\mathcal{S}}_0 = \left\{ j : \mathbb{E}D(\hat{\boldsymbol{\beta}}_{-j}, [\hat{\boldsymbol{\beta}}]) < \mathbb{E}D(\hat{\boldsymbol{\beta}}, [\hat{\boldsymbol{\beta}}]) \right\}$$

Then $\mathbb{P}(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \rightarrow 1$ as $n \rightarrow \infty$.

```
   DroppedVar        Cn
1        - x2 0.2356008
2        - x3 0.2428004
3        - x4 0.2448785
4        - x1 0.2473548
5        - x5 0.2486610
6       - x20 0.2503475
7      <none> 0.2505000
8        - x9 0.2522873
9       - x21 0.2538186
10      - x22 0.2547132
11      - x14 0.2548410
12      - x17 0.2554293
13      - x13 0.2559990
14      - x10 0.2564211
15      - x24 0.2566334
16      - x19 0.2568725
17      - x25 0.2573902
18       - x8 0.2578656
19      - x16 0.2588032
20      - x12 0.2590218
21       - x6 0.2595048
22      - x23 0.2598039
23      - x15 0.2605307
24      - x11 0.2606763
25      - x18 0.2610460
26       - x7 0.2613168
```

Why it works

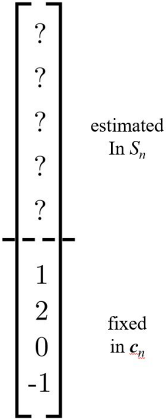


1. Start with full model;
2. Drop an essential predictor ⇒ Model becomes wrong ⇒ *e*-value shifts to left;
3. Drop a non-essential predictor ⇒ Model still correct but nested within full model ⇒ *e*-value shifts to right.

The framework

Definition of a model

$\boldsymbol{\theta}_n =$



Given traingular sequences of observable data and unknown parameters:

$$\{(\mathcal{B}_n, \boldsymbol{\theta}_n), n \in \mathbb{N}\}$$

$$\mathcal{B}_n = \{B_{n1}, \ldots, B_{nk_n}\}, \quad \boldsymbol{\theta}_n \in \boldsymbol{\Theta}_n \subseteq \mathbb{R}^{p_n}$$

a candidate model $\mathcal{M}_n$ is specified by:

(a) The set of indices $\mathcal{S}_n \subseteq \{1, 2, \ldots, p_n\}$ where parameter values are estimated from the data;

(b) An ordered vector of known constants $\mathbf{c}_n = (c_{nj} : j \neq \mathcal{S}_n)$ for parameters not indexed by $\mathcal{S}_n$.

Denote the model space of $\mathcal{M}_n$ by $\boldsymbol{\Theta}_{mn}$.

The preferred model

Among all such possible models, we designate one of them as the *preferred model*: say $\mathcal{M}_{*n} = (\mathcal{S}_{*n}, \mathbf{c}_{*n})$.

This is the baseline model we shall compare all models with.

Example

For variable selection, the model with all covariates is the preferred model.

For hypothesis testing, the null model is preferred model.

Designate an element of the preferred model space $\boldsymbol{\Theta}_{*n}$ as the *preferred parameter vector* θ_{0n}. This is generally informative of the data generating process.

Method of estimation

For any candidate model $\mathcal{M}_n$, we consider estimating functionals $\Psi_{sni}(.)$ that admit an unique minimizer $\boldsymbol{\theta}_{mn} \in \boldsymbol{\Theta}_{mn}$:

$$\Psi_{sn}(\boldsymbol{\theta}) = \mathbb{E} \sum_{i=1}^{k_n} \Psi_{sni}(\boldsymbol{\theta}, B_{ni}); \quad \boldsymbol{\theta}_{mn} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{mn}}{\text{argmin}} \, \Psi_{sn}(\boldsymbol{\theta})$$

The estimate corresponding to model $\mathcal{M}_n$ is obtained by minimizing its sample version:

$$\hat{\boldsymbol{\theta}}_{mn} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{mn}}{\text{argmin}} \sum_{i=1}^{k_n} \Psi_{sni}(\boldsymbol{\theta}, B_{ni})$$

We assume there exist $a_{sn} \uparrow \infty$ such that $[a_{sn}(\hat{\boldsymbol{\theta}}_{sn} - \boldsymbol{\theta}_{sn})]$ 'converges' to a distribution as $n \rightarrow \infty$. Here $\boldsymbol{\theta}_{sn}$ is $\boldsymbol{\theta}_{mn}$ is $\mathcal{S}_n$ indices, and same for $\hat{\boldsymbol{\theta}}_{sn}$.

A common transformation

Now we map parameters from the parameter space to a common multivariate space using a known smooth function:

$$\mathbf{G}_{mn} : \mathbf{\Theta}_n \mapsto \mathbb{R}^{d_n}; \quad d_n = o(\min_s \{a_{sn}, a_{*n}\})$$

This is for easy comparison of different kinds of modelling methods.

Example

Want to compare the following models built on data $\{(y_i, X_{1i}, X_{2i}) : i = 1, \ldots, n\}$:

(1) Linear regression - $Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), \sigma > 0$;
(2) Semiparametric regression - $Y_i = X_{1i}\beta_1 + g(X_{2i}) + \epsilon_i$ for unknown g;
(3) Single index model - $Y_i = h(X_{1i}\beta_1 + X_{2i}\beta_2) + \epsilon_i$ for unknown h;

Comparing all model in terms of prediction errors has a clearer interpretation.

‘Good’ and ‘bad’ models

For $\mathbf{g} \in \mathbb{R}^{d_n}$ and $\mathcal{G}'_n \subseteq \mathbb{R}^{d_n}$ define the following:

$$d(\mathbf{g}, \mathcal{G}'_n) := \inf_{\mathbf{g}' \in \mathcal{G}'_n} \|\mathbf{g} - \mathbf{g}'\|$$

Then

(a) For two sequences of models, say $\{\mathcal{M}_{1n}\}$ and $\{\mathcal{M}_{2n}\}$, we say $\{\mathcal{M}_{1n}\}$ *is nested within* $\{\mathcal{M}_{2n}\}$ if, for all sequences $\{\mathbf{g}_{1n} : \mathbf{g}_{1n} \in \mathcal{G}_{1n}\}$ we have

$$\lim_{n \rightarrow \infty} d(\mathbf{g}_{1n}, \mathcal{G}_{2n}) = 0$$

with $\mathcal{G}_{1n} := \{\mathbf{G}_{m1n}(\boldsymbol{\theta}(\mathcal{M}_{1n})) : \boldsymbol{\theta}(\mathcal{M}_{1n}) \in \boldsymbol{\Theta}_{m1n}\}$ etc.

(b) A sequence of models $\{\mathcal{M}_n\}$ is called *adequate* if the model $\mathcal{M}_{0n}$ corresponding to the singleton set $\boldsymbol{\Theta}_{0n} = \{\boldsymbol{\theta}_{0n}\}$, is nested within $\mathcal{M}_n$.

(c) A model that is not adequate is an *inadequate model*.

Motivation

Covers obvious cases:

$$\begin{array}{lcl} (1,2,3,0) & - & \text{preferred parameter vector} \\ (*,*,*,0) & - & \text{adequate model} \\ (*,*,*,*) & - & \text{full model} \\ (*,*,0,*) & - & \text{inadequate model} \end{array}$$

Covers limiting cases:, e.g. $(*,*,*,\delta_n), \delta_n = o(1)$ will be an adequate model in our framework.

Such data generating models, e.g.

$$Y_{ni} = X_{1i}\beta_{01} + X_{2i}\delta_n + \epsilon; \quad \beta_{01} \in \mathbb{R}, \delta_n = o(1)$$

for linear regression, frequently arise from prior choices in bayesian variable selection techniques.

Statistical evaluation maps and *e*-values

We now introduce an *evaluation function*:

$$E_n : \mathbb{R}^{d_n} \times \tilde{\mathbb{R}}^{d_n} \mapsto [0, \infty)$$

to quantify the relative position of $\hat{\mathbf{G}}_{mn} \equiv \mathbf{G}_{mn}(\hat{\theta}_{mn})$ with respect to the preferred model estimate distribution. Denote this by

$$E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}])$$

***e*-value** is simply a functional of the distribution of this random evaluation function. Denote this by $e_n(\mathcal{M}_n)$.

We shall now elaborate on the following choice of the *e*-value:

$$e_n(\mathcal{M}_n) = \mathbb{E} E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}])$$

Results

Result: model adequacy and *e*-values

Theorem

Under some regularity conditions, as $n \to \infty$:

1. *For the preferred model $e_n(\mathcal{M}_{*n}) \to e_* < \infty$;*
2. *For any adequate model, $|e_n(\mathcal{M}_n) - e_n(\mathcal{M}_{*n})| \to 0$;*
3. *For any inadequate model, $e_n(\mathcal{M}_n) \to 0$.*

Bootstrap estimation of *e*-values

We shall use bootstrap to generate multiple copies of the preferred model estimate and thus approximate $[\hat{\mathbf{G}}_{*n}]$.
Under standard regularity conditions (Chatterjee and Bose, 2005), we shall calculate the bootstrap estimate $\hat{\boldsymbol{\theta}}_{rmn}$ by solving

$$\hat{\boldsymbol{\theta}}_{rmn} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{mn}} \sum_{i=1}^{k_n} \mathbb{W}_{rsni} \Psi_{sni}(\boldsymbol{\theta}, B_i)$$

where $\mathbb{W}_{rsni}$ are i.i.d. weights chosen independently from the data satisfying:

$$\begin{aligned} \mathbb{E}\mathbb{W}_{rsn1} &= 1 \\ \mathbb{V}\mathbb{W}_{rsn1} &= \tau_{sn}^2 \uparrow \infty \\ \tau_{sn}^2 &= o(a_{sn}^2), \\ \mathbb{E}\mathbb{W}_{rsn1}\mathbb{W}_{rsn2} &= O(k_n^{-1}), \\ \mathbb{E}\mathbb{W}_{rsn1}^2\mathbb{W}_{rsn2}^2 &\to 1, \\ \mathbb{E}\mathbb{W}_{rsn1}^4 &< \infty. \end{aligned}$$

This is the *Generalized bootstrap*.

Result: bootstrap estimation of *e*-values

Theorem

Suppose

$$\hat{e}_n(\mathcal{M}_n) = \mathbb{E}_r E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1 * n}])$$

based on two independent sets of bootstrap samples.
Then for the above bootstrap scheme, as $n \rightarrow \infty$:

1. *For any adequate model, $|\hat{e}_n(\mathcal{M}_n) - \hat{e}_n(\mathcal{M}_{*n})| \overset{P_n}{\rightarrow} o_P(1)$;*
2. *For any inadequate model, $\hat{e}_n(\mathcal{M}_n) \overset{P_n}{\rightarrow} o_P(1)$.*

where P_n is probability conditional on the data.

Fast bootstrap approximation of coefficient estimates

The bootstrap sample is related to the actual estimate through score vectors and hessian matrices:

$$\hat{\boldsymbol{\theta}}_{rsn} = \hat{\boldsymbol{\theta}}_{sn} - \frac{\tau_{sn}}{a_{sn}} \left[\sum_{i=1}^{k_n} \Psi''_{sni}(\hat{\boldsymbol{\theta}}_{sn}, B_i) \right]^{-1/2} \sum_{i=1}^{k_n} W_{rsni} \Psi'_{sni}(\hat{\boldsymbol{\theta}}_{sn}, B_i) + \mathbf{R}_n$$

with $\mathbb{E}_r \|\mathbf{R}_n\|^2 = o_P(1)$ and $W_{rsni} := (\mathbb{W}_{rsni} - 1)/\tau_{sn}$.

This means we can compute $\hat{\boldsymbol{\theta}}_{rsn}$ just by generating Monte-Carlo samples and reusing other model objects. This makes the bootstrap procedure very fast.

A plug-in model estimate

Start with the preferred model estimate $\hat{\boldsymbol{\theta}}_{*n} = (\hat{\theta}_{*n1}, \ldots, \hat{\theta}_{*np_n})^T$. For any model $\mathcal{M}_n$, define $\hat{\boldsymbol{\theta}}_{mn}$ as:

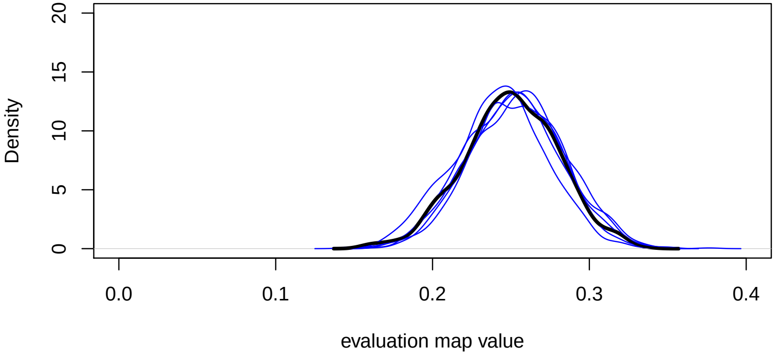
$$\hat{\boldsymbol{\theta}}_{mnj} = \begin{cases} \text{Estimated } \hat{\theta}_{*nj} & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases}$$

We shall work with these plugin estimates. This saves time while still ensuring consistency of all model estimates.

Same for bootstrap versions:

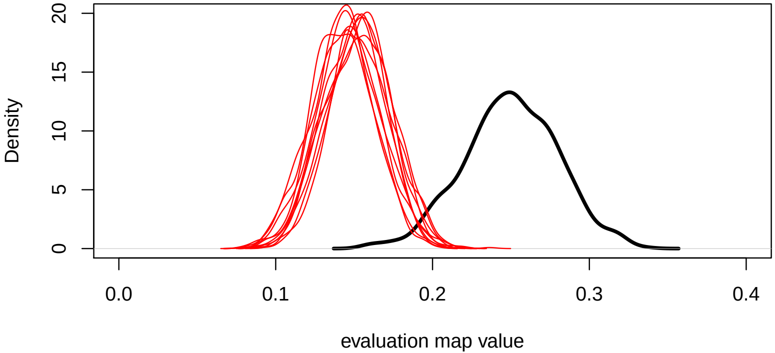
$$\hat{\boldsymbol{\theta}}_{rmnj} = \begin{cases} \text{Estimated } \hat{\theta}_{r*nj} & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases}$$

A generic model selection recipe

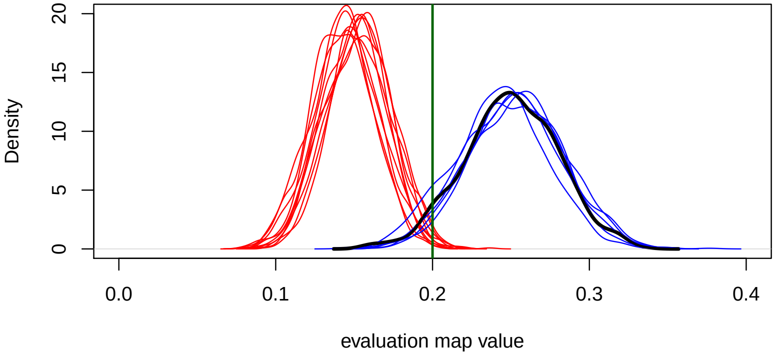


For large enough *n*, *e*-values for all adequate models will be close to that of the preferred model.

A generic model selection recipe



But *e*-values of all inadequate models will be very small.



Thus we can choose an appropriate threshold ϵ_n such that any model with *e*-value below that threshold is inadequate, but *e*-value above the threshold implies the model is adequate.

Feature selection with data depth

Now take fixed $p \equiv p_n$, and drop subscripts in $\mathcal{M}_n, \mathcal{M}_{*n}$ etc. Take $E_n = D$, a depth function. Then

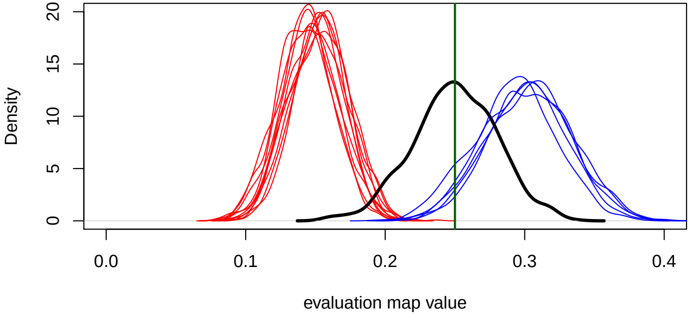
Theorem

For two nested models: $\mathcal{M}_1$ nested within $\mathcal{M}_2$, we have $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$ for large enough n.

Notice now that all adequate models are by definiton nested within the preferred model. So $e_n(\mathcal{M}_*) < e_n(\mathcal{M})$ for any adequate model. Thus for large enough n, we can take threshold $\epsilon_n = e_n(\mathcal{M}_*)$ now.

$$\begin{aligned}
(1, 2, 3, 0) \quad & - \quad \text{preferred parameter vector} \\
(*, *, *, 0) \quad & - \quad \text{adequate model} \\
(*, *, *, *) \quad & - \quad \text{full model}
\end{aligned}$$

One-step variable selection



1. Start with full model;
2. Drop an essential predictor ⇒ Model becomes inadequate ⇒ *e*-value shifts to left;
3. Drop a non-essential predictor ⇒ Model still adequate but nested within full model ⇒ *e*-value shifts to right.

Numerical studies and data applications

Simulation: linear mixed models

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon} \in \mathbb{R}^{n_i}$$
$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{V}_i); \quad \mathbf{V}_i = \sigma^2 \mathbf{I}_{n_i} + \mathbf{Z}_i \Delta \mathbf{Z}_i^T$$

- m subjects, n_i observations per subject, $n = m \times n_i$ total observations;
- $p = 9, \boldsymbol{\beta} = (1, 1, 0, 0, 0, 0, 0, 0, 0)^T$;
- Elements of $\mathbf{X}_1, \ldots, \mathbf{X}_m$ chosen from Unif$(-2, 2)$, random effect design matrix $\mathbf{Z}_i$ is first 4 columns of $\mathbf{X}_i$;
- $$\Delta = \begin{pmatrix} 9 & & & \\ 4.8 & 4 & & \\ 0.6 & 1 & 1 & \\ 0 & 0 & 0 & 0 \end{pmatrix}$$
- Two settings: (i) $m = 30, n_i = 5$, (ii) $m = 60, n_i = 10$;
- We use i.i.d. draws of Gamma(1,1) as bootstrap weights $W_i + 1$.

Simulation results

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		$n_i = 5, m = 30$			$n_i = 10, m = 60$		
e-value based	$\tau_n/\sqrt{n} = 1$	57.4	0.0	5.24	43.8	0.0	4.03
	2	30.4	0.0	3.32	12.3	0.0	2.42
	3	15.6	0.0	2.54	3.2	0.0	2.10
	4	7.3	0.0	2.24	1.0	0.0	2.03
	5	3.0	0.0	2.09	0.7	0.0	2.02
	6	1.7	0.0	2.05	0.3	0.0	2.01
	7	1.0	0.0	2.03	0.0	0.0	2.00
	8	0.7	0.0	2.02	0.0	0.0	2.00
	9	0.0	0.0	2.00	0.0	0.0	2.00
	10	0.0	0.0	2.00	0.0	0.0	2.00
Peng and Lu (2012)	BIC	21.5	9.9	2.26	1.5	1.9	2.10
	AIC	17	11.0	2.43	1.5	3.3	2.20
	GCV	20.5	6	2.30	1.5	3	2.18
	$\sqrt{\log n/n}$	21	15.6	2.67	1.5	4.1	2.26

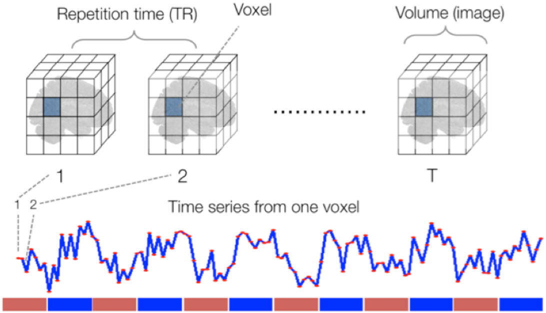
Comparison between our method and that proposed by Peng and Lu (2012) through average false positive percentage, false negative percentage and model size

Simulation results

Method	$\tau_n/\sqrt{n}$	Setting 1	Setting 2
e-value based	1	2	16
	2	36	67
	3	60	91
	4	80	97
	5	91	98
	6	95	99
	7	97	100
	7	98	100
	8	100	100
	10	100	100
Bondell et al. (2010)		73	83
Peng and Lu (2012)		49	86
Fan and Li (2012)		90	100

Comparison of our method and three sparsity-based methods of mixed effect model selection through accuracy of selecting correct fixed effects

Data analysis: spatio-temporal dependency in fMRI visual task data



Dimensions of 3D voxel array $64 \times 64 \times 33$

- Data collected from 19 subjects, each of which were shown two types of pictures at certain time points within a test period (Wakeman and Henson, 2015).
- Each subject went through 9 runs. In each run 210 images of their brain were recorded in 2-second intervals.
- We use the data from a single run on subject 1, and perform a voxelwise analysis to find out the effect of time lags and activation of neighboring voxels on the activation of each voxel.

Model 1: temporal model

$$y_i(t) = x_{ia}(t)\beta_{ia} + x_{ib}(t)\beta_{ib} + \sum_{l=1}^{q} t^{l-1}\gamma_{il} + \sum_{K=1}^{5} y_i(t-k)\delta_{i,t-k} + \epsilon_i(t)$$

- Voxel index i, time index t;
- x_{ia}, x_{ib} are stimulus values for the two types of tasks: calculated through a deterministic equation;
- $t^{l-1}\gamma_{il}$ are the polynomial drift terms quantifying background noise;
- Time dependency quantifies by the autoregressive terms $y_i(t-k)\delta_{i,t-k}$;
- We set an AR(5) structure and quadratic drift ($q = 2$) and fit linear models for each voxel. Following this we run e-value based variable selection on all terms.
- Less than 0.1% voxels were selected for each predictor, and in no particular pattern: indicating lack of temporal dependence.

Model 2: Spatial model

$$
\begin{aligned}
y_i(t) &= x_{ia}(t)\beta_{ia} + x_{ib}(t)\beta_{ib} + \sum_{l=1}^{q} t^{l-1}\gamma_{il} + \sum_{n \in N_i} y_n(t)\delta_{i,n} + \epsilon_i(t) \\
&= \tilde{\mathbf{x}}_i(t)^T \boldsymbol{\theta}_i + \epsilon_i(t)
\end{aligned}
$$

- Want to quantify effect of immediate neighbors: N_i the set of neighbors of i-th voxel;

- Edge or corner voxels excluded: so 26 neighbors for a voxel, thus total 30 predictors.

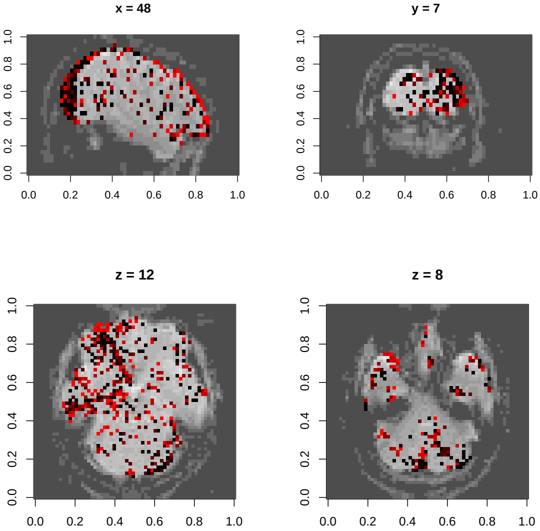
- We estimate the set of non-zero coefficients in $\boldsymbol{\theta}_i$ using our method. Suppose this set is R_i, and its subsets corresponding to neighbor and non-neighbor (i.e. stimuli and drift) terms are S_i and T_i, respectively.

Model 2: Spatial model

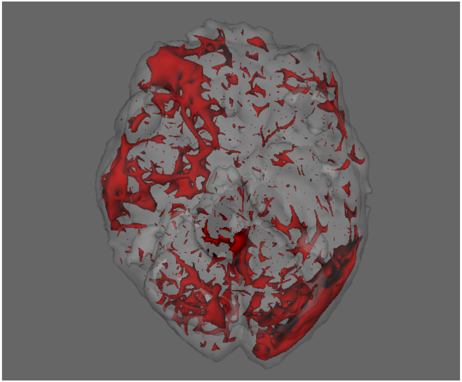
To quantify the effect of neighbors we now calculate the corresponding F-statistic:

$$F_i = \frac{(\sum_{n \in S_i} \tilde{x}_{i,n} \hat{\theta}_{i,n})^2}{(y_i(t) - \sum_{n \in T_i} \tilde{x}_{i,n} \hat{\theta}_{i,n})^2} \frac{|n - T_i|}{|S_i|}$$

and obtain its p-value, i.e. $P(F_i \geq F_{|S_i|,|n-T_i|})$.



Plot of significant *p*-values at $\alpha = 0.05$ at specified cross-sections



A smoothed surface obtained from the *p*-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas as well as Cerebellum

Outline

1. Motivation

2. Signed Peripherality Functions

3. Nonconvex Penalized Multitask Regression using Depth-based Penalty

4. Generalized model discovery using statistical evaluation maps

5. **Selecting important SNPs from multi-SNP mixed models on Twin Studies GWAS data**

Motivation

Genome-Wide Association Studies (GWAS) based on families are used in behavioral genetics to control for environmental variation, thus requiring smaller sample size to detect Single Nucleotide Polymorphisms (SNP) responsible behind traits like alcoholism and drug addiction, and also to quantify gene-environment interaction.

Two challenges:

1. SNPs highly correlated, weak signals of individual SNPs;
2. Need to use mixed models to account for within-family dependence.

Statistical model

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i \boldsymbol{\beta}_g + \mathbf{C}_i \boldsymbol{\beta}_c + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{V}_i); \quad \mathbf{V}_i = \sigma_a^2 \boldsymbol{\Phi}_i + \sigma_c^2 \mathbf{1}\mathbf{1}^T + \sigma_e^2 \mathbf{I}_{n_i}$$

- Total m families, with the i-th pedigree containing n_i individuals;
- $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$ are the quantitative trait values for individuals in i-th pedigree, $\mathbf{G}_i \in \mathbb{R}^{n_i \times p_s}$ containing their genotypes for a bunch of SNPs, $\mathbf{C}_i \in \mathbb{R}^{n_i \times p}$ contain the data on individual-specific covariates;
- Three variance components correspond to polygenic effect due to other SNPs, shared environment effect and individual-specific effects. This is called the **ACE model**.

The relationship matrix Φ_i

$$\mathbf{\Phi}_{MZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1 \\ 1/2 & 1/2 & 1 & 1 \end{bmatrix},$$

$$\mathbf{\Phi}_{DZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{bmatrix},$$

$$\mathbf{\Phi}_{Adopted} = \mathbf{I}_4$$

$\mathbf{\Phi}_i$ depends on the type of the *i*-th family:

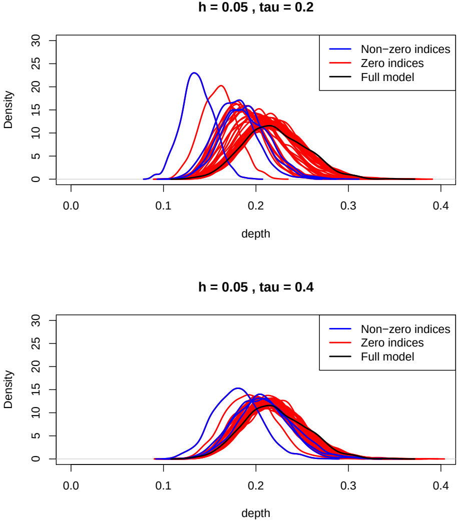
MZ = family with identical or monozygous twins,
DZ = family with identical or dizygous twins.

Objective

Want to detect the non-zero entries of $\boldsymbol{\beta}_g$ in the above model.

State-of-the-art is to perform single-SNP analysis and then correct for multiple testing. This loses power. We want to use the ***e*-values** method to improve that.

Problem: weak signal of individual SNPs



Using means as *e*-values makes the procedure very conservative.

But it may still be possible to use a tail quantile to distinguish between distributions.

A new *e*-value

We now take as *e*-value the *q*-th quantile of the evaluation map distribution, for some fixed $q \in (0, 1)$. Under slightly stronger assumptions than before, all results go through for a general evaluation map in the population and bootstrap worlds.

1. Estimate the full model coefficient, say $\hat{\boldsymbol{\beta}}_g \equiv \hat{\boldsymbol{\beta}}$ (by R package `regress`)
2. Obtain its bootstrap distribution: $[\hat{\boldsymbol{\beta}}]$;
3. Replace the *j*-th coefficient with 0, name it $\hat{\boldsymbol{\beta}}_{-j}$. Do the same for its bootstrap distribution, say $[\hat{\boldsymbol{\beta}}_{-j}]$. Repeat for all *j*;
4. *e*-value of *j*-th covariate = tail probability of the *q*-th quantile of $[E(\hat{\boldsymbol{\beta}}_{-j}, [\hat{\boldsymbol{\beta}}])]$ with respect to $[E(\hat{\boldsymbol{\beta}}, [\hat{\boldsymbol{\beta}}])]$;
5. Select *j*-th covariate if its *e*-value is less than *tq*, for some $0 < t < 1$.

Simulation setup

- 250 pedigrees, each of size 4: consisting of parents and MZ twins;
- $\alpha = 0$, no environmental covariates;
- 50 SNPs in correlated blocks of 6,4,6,4 and 30: MAF of SNPs in the blocks 0.2, 0.4, 0.4, 0.25 and 0.25;
- $\sigma_a^2 = 4, \sigma_c^2 = 1, \sigma_e^2 = 1$;
- First SNP of first 4 blocks are causal: each having heritability (a measure of magnitude of non-zero effect) $h/6\%$;
- Full setup replicated 1000 times.

- Methods compared:
  mBIC2 - Variant of BIC that control false discovery rate at 0.05;
  RFGLS - Fast method of fitting single-SNP ACE models. Do Benjamini-Hochberg correction on *p*-values to control FDR at 0.05.

Simulation results

6x Heritability	mBIC2	RFGLS +BH	quantile *e*-values				
			q	$t = 0.8$	$t = 0.7$	$t = 0.6$	$t = 0.5$
$h = 10$	0.79/0.99	0.95/0.92	0.9	0.95/0.97	0.95/0.97	0.95/0.98	**0.94**/**0.98**
			0.5	0.96/0.97	0.96/0.98	0.95/0.98	0.94/0.98
			0.2	0.96/0.94	0.96/0.97	0.95/0.97	0.95/0.98
$h = 5$	0.41/0.99	0.62/0.97	0.9	0.72/0.95	0.7/0.96	0.69/0.96	**0.66**/**0.97**
			0.5	0.78/0.94	0.75/0.94	0.72/0.95	0.71/0.96
			0.2	0.83/0.91	0.78/0.94	0.75/0.95	0.73/0.95
$h = 2$	0.11/0.99	0.14/0.99	0.9	0.26/0.97	0.24/0.97	0.23/0.98	**0.21**/**0.98**
			0.5	0.34/0.95	0.28/0.96	0.27/0.97	0.26/0.97
			0.2	0.46/0.91	0.34/0.95	0.3/0.96	0.27/0.96
$h = 1$	0.05/0.99	0.04/0.99	0.9	0.12/0.98	0.1/0.98	0.09/0.99	**0.08**/**0.99**
			0.5	0.16/0.96	0.13/0.97	0.12/0.97	0.11/0.98
			0.2	0.25/0.93	0.16/0.96	0.13/0.97	0.13/0.97
$h = 0$	–/0.99	–/0.99	0.9	–/0.99	–/0.99	–/0.99	**–/0.99**
			0.5	–/0.98	–/0.98	–/0.99	–/0.99
			0.2	–/0.94	–/0.98	–/0.98	–/0.99

Average true positive/ true negative detection proportions over 1000 replications

Simulation results

6x	mBIC2	RFGLS	quantile *e*-values				
Heritability		+BH	q	$t = 0.8$	$t = 0.7$	$t = 0.6$	$t = 0.5$
$h = 10$	0.84/0.99	0.96/0.99	0.9	0.96/0.97	0.96/0.97	0.95/0.98	**0.94/0.98**
			0.5	0.96/0.97	0.96/0.97	0.95/0.98	0.95/0.98
			0.2	0.97/0.95	0.96/0.97	0.96/0.97	0.95/0.98
$h = 5$	0.48/0.99	0.64/0.99	0.9	0.73/0.95	0.71/0.95	0.7/0.96	**0.67/0.97**
			0.5	0.79/0.93	0.76/0.94	0.73/0.95	0.72/0.95
			0.2	0.85/0.91	0.79/0.93	0.76/0.94	0.74/0.95
$h = 2$	0.16/0.99	0.16/0.99	0.9	0.29/0.96	0.27/0.97	0.25/0.98	**0.23/0.98**
			0.5	0.37/0.95	0.31/0.96	0.3/0.96	0.29/0.97
			0.2	0.53/0.91	0.38/0.95	0.33/0.95	0.3/0.96
$h = 1$	0.08/0.99	0.05/0.99	0.9	0.15/0.97	0.13/0.98	0.12/0.98	**0.10/0.99**
			0.5	0.2/0.96	0.17/0.97	0.15/0.97	0.13/0.98
			0.2	0.35/0.93	0.21/0.96	0.17/0.97	0.16/0.97
$h = 0$	–/0.98	–/0.99	0.9	–/0.97	–/0.98	–/0.98	**–/0.99**
			0.5	–/0.95	–/0.97	–/0.97	–/0.98
			0.2	–/0.90	–/0.95	–/0.97	–/0.97

Average true positive/ true negative detection proportions over 1000 replications: true positive = can detect any SNP in the same block

Analyzing the Minnesota Twin Studies data

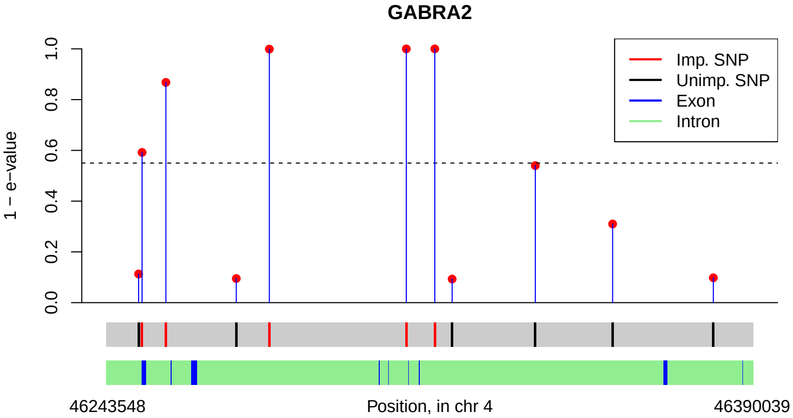
- Analyze data on families with MZ twins: 682 families;
- Response variable: amount of alcohol consumption;
- Look at models specific to well-studied genes for alcoholism: GABRA2, ADH1B, ADH1C, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT;
- Group together ADH genes as individual genes have very small number of SNPs. Also do SLC6A4+DRD2 together as they are known to interact with each other.
- We take $q = 0.9, t = 0.5$ to increase specificity, i.e. detection of true negatives.

Number of detected SNPs

Gene	Total/detected SNP
GABRA2	11/5
ADH	21/5
SLC6A3	18/4
SLC6A4	5/0
OPRM1	46/29
CYP2E1	9/5
DRD2	17/0
ALDH2	5/5
COMT	15/9
SLC6A4 + DRD2	22/0

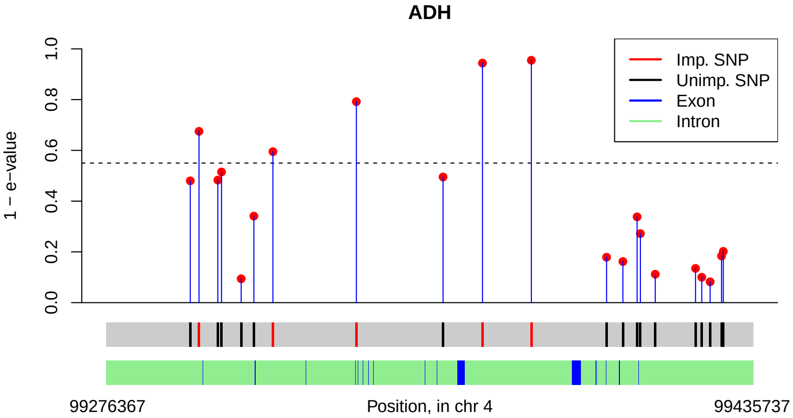
Table of analyzed genes and detected SNPs in them

GABRA2



Detects rs1808851 and rs279856, which have very high correlation with the well-known rs279858. This was missed by a previous analysis (Irons, 2012).

ADH genes



Detects rs13103626 and rs10516430 at positions 99317251 and 99337881 in chr 4: close to the well-known rs1229984 at position 99318162.

Conclusion and future work

In this thesis, I have proposed several methods of using depths or depth-like quantities in traditional statistical inference, e.g. PCA, sparse regression, model selection.

Future works include:

1. Formulation of depth in a general Hilbert space;
2. Exploring different uses of the *e*-values framework: for example in multiple testing;
3. Extension of the signed rank and depth-based penalization framework.

References

H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077, 2010.

S. Chatterjee and A. Bose. Generalized bootstrap for estimating equations. *Ann. Statist.*, 33:414–436, 2005.

Y. Fan and R. Li. Variable selection in linear mixed effect models. *Ann. Statist.*, 40(4):2043–2068, 2012.

D. E. Irons. *Characterizing specific genetic and environmental influences on alcohol use*. PhD thesis, University of Minnesota, sep 2012.

R.Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18:405–414, 1990.

N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, and K.L. Cohen. Robost principal components of functional data. *TEST*, 8:1–73, 1999.

H. Peng and Y. Lu. Model selection in linear mixed effect models. *J. Multivariate Anal.*, 109:109–129, 2012.

J.W. Tukey. Mathematics and picturing data. In R.D. James, editor, *Proceedings of the International Congress on Mathematics*, volume 2, pages 523–531, 1975.

D. G. Wakeman and R. N. Henson. A multi-subject, multi-modal human neuroimaging dataset. *Scientif. Data*, 2: article 15001, 2015. DOI: 10.1038/sdata.2015.1.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36: 1509–1533, 2008.

Y. Zuo. Projection-based depth functions and associated medians. *Ann. Statist.*, 31:1460–1490, 2003.

Y. Zuo and R. Serfling. General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482, 2000.

THANK YOU!