Selective inference: a conditional perspective

Xiaoying Tian Harris
Joint work with Jonathan Taylor

August 22, 2016

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model $=$ lm(y $\sim$ X1 $+$ X2 $+$ X3 $+$ X4)

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model = lm(y $\sim$ X1 + X2 + X3 + X4)
  model = lm(y $\sim$ X1 + X2 + X4)

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$
- model = lm(y $\sim$ X1 + X2 + X3 + X4)
  model = lm(y $\sim$ X1 + X2 + X4)
  model = lm(y $\sim$ X1 + X3 + X4)

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model = lm(y $\sim$ X1 + X2 + X3 + X4)
  model = lm(y $\sim$ X1 + X2 + X4)
  model = lm(y $\sim$ X1 + X3 + X4)
- Inference after model selection
  1. Use data to select a set of variables E
  2. Normal z-test to get p-values

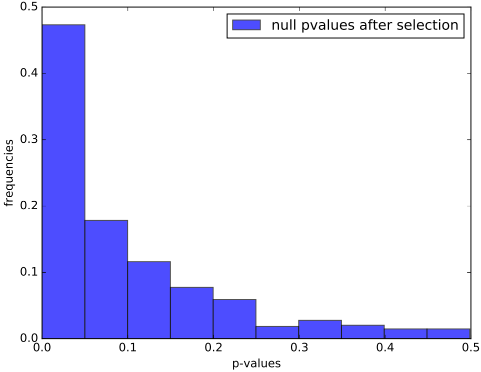
Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model = lm(y $\sim$ X1 + X2 + X3 + X4)
  model = lm(y $\sim$ X1 + X2 + X4)
  model = lm(y $\sim$ X1 + X3 + X4)
- Inference after model selection
  1. Use data to select a set of variables E
  2. Normal z-test to get p-values
- Problem: inflated significance
  1. Normal z-tests need adjustment
  2. Selection is biased towards "significance"

Inflated Significance

Setup:

- $X \in \mathbb{R}^{100\times 200}$ has i.i.d normal entries
- $y = X\beta + \epsilon$, $\epsilon \sim N(0, I)$
- $\beta = (\underbrace{5, \ldots, 5}_{10}, 0, \ldots, 0)$
- LASSO, nonzero coefficient set E
- z-test, null pvalues for $i \in E$, $i \notin \{1, \ldots, 10\}$



Post-selection inference

- PoSI approach:
  1. Reduce to simultaneous inference
  2. Protects against any selection procedure
  3. Conservative and computationally expensive

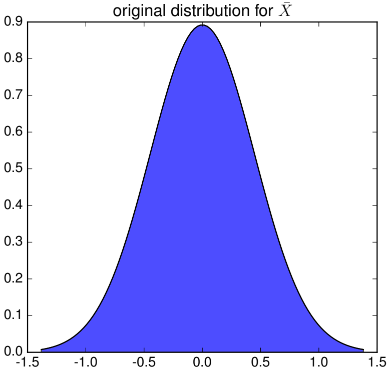
Post-selection inference

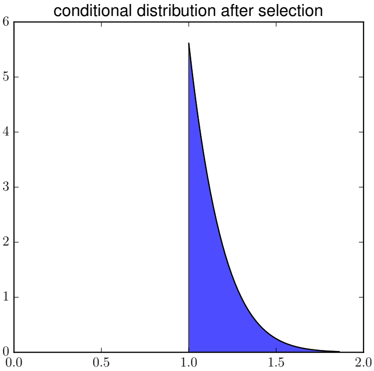
- PoSI approach:
  1. Reduce to simultaneous inference
  2. Protects against any selection procedure
  3. Conservative and computationally expensive
- Selective inference approach:
  1. Conditional approach
  2. Specific to particular selection procedures
  3. More powerful tests

Conditional approach: example

Consider the selection for "big effects":

- $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$
- Select for "big effects", $\overline{X} > 1$
- Observation: $\overline{X}_{obs} = 1.1$, with $n = 5$
- Normal z-test v.s. selective test for $H_0 : \mu = 0$.

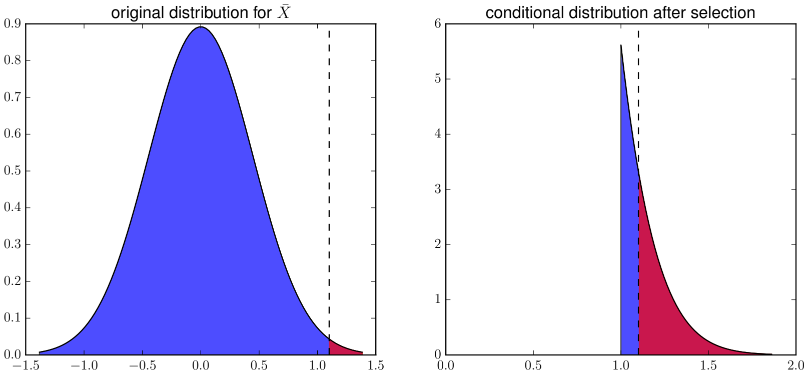




Conditional approach: example

Consider the selection for “big effects”:

- $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$
- Select for “big effects”, $\overline{X} > 1$
- Observation: $\overline{X}_{obs} = 1.1$, with $n = 5$
- Normal z-test v.s. selective test for $H_0 : \mu = 0$.



Moral of selective inference

Conditional approach:

- Selection, e.g. $\overline{X} > 1$.
- Conditional distribution after selection, e.g. $N(\mu, \frac{1}{n})$, truncated at 1.
- Target of inference may (or may not) depend on the selection.

  1. Not dependent: e.g. $H_0 : \mu = 0$.
  2. Dependent: e.g. two-sample problem, inference for variables selected by LASSO

Moral of selective inference

Conditional approach:

- Selection, e.g. $\overline{X} > 1$.
- Conditional distribution after selection, e.g. $N(\mu, \frac{1}{n})$, truncated at 1.
- Target of inference may (or may not) depend on the selection.

  1. Not dependent: e.g. $H_0 : \mu = 0$.
  2. Dependent: e.g. two-sample problem, inference for variables selected by LASSO

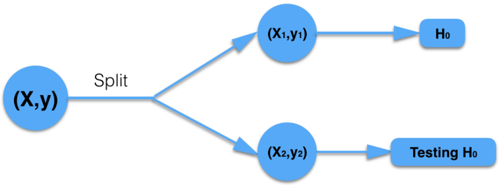
- Random hypothesis?

Random hypothesis

- Replication studies

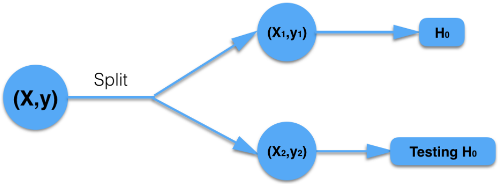
Random hypothesis

- Replication studies
- Data splitting: observe data (X, y), with X fixed, entries of y are independent (given X)



Random hypothesis

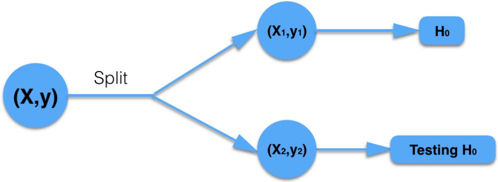
- Replication studies
- Data splitting: observe data (X, y), with X fixed, entries of y are independent (given X)



Random hypothesis selected by the data

Random hypothesis

- Replication studies
- Data splitting: observe data (X, y), with X fixed, entries of y are independent (given X)



Random hypothesis selected by the data

- Data splitting as a conditional approach:

$$\mathcal{L}(y_2) = \mathcal{L}(y_2 | H_0 \text{ selected by } y_1).$$

Selective inference: a conditional approach

- Data splitting as a conditional approach:

$$\mathcal{L}(y_2) = \mathcal{L}(y_2 | H_0 \text{ selected by } y_1).$$

- Inference based on the conditional law:

$$\mathcal{L}(y | H_0 \text{ selected by } y^*), \qquad y^* = y^*(y, \omega),$$

where ω is some randomization independent of y.

Selective inference: a conditional approach

- Data splitting as a conditional approach:

$$\mathcal{L}(y_2) = \mathcal{L}(y_2 | H_0 \text{ selected by } y_1).$$

- Inference based on the conditional law:

$$\mathcal{L}(y | H_0 \text{ selected by } y^*), \qquad y^* = y^*(y, \omega),$$

  where ω is some randomization independent of y.

- Examples of y^*:
  1. $y^* = y$, ω is void
  2. $y^* = y_1$, where ω is a random split
  3. $y^* = y + \omega$, where $\omega \sim N(0, \gamma^2)$, additive noise

Different y^*

	$y^* = y$	$y^* = y_1$	$y^* = y + \omega$	randomized LASSO
y	Lee et al. (2013), Taylor et al.(2014)	Data splitting, Fithian et al.(2014)	T. & Taylor (2015)	T. & Taylor (2015)

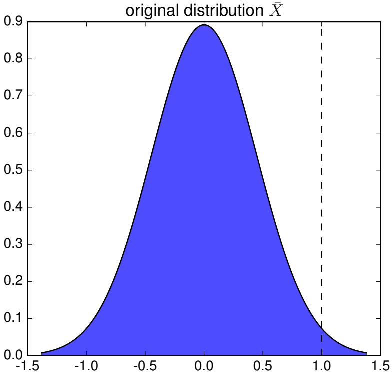
Different y^*

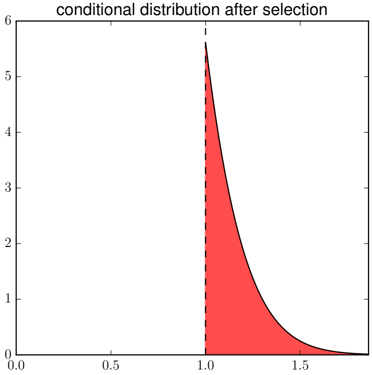
	$y^* = y$	$y^* = y_1$	$y^* = y + \omega$	randomized LASSO
y	Lee et al. (2013), Taylor et al.(2014)	Data splitting, Fithian et al.(2014)	T. & Taylor (2015)	T. & Taylor (2015)

- Randomization transfers the properties of unselective distributions to selective counterparts.
- Much more powerful tests.

Selective v.s. unselective distributions

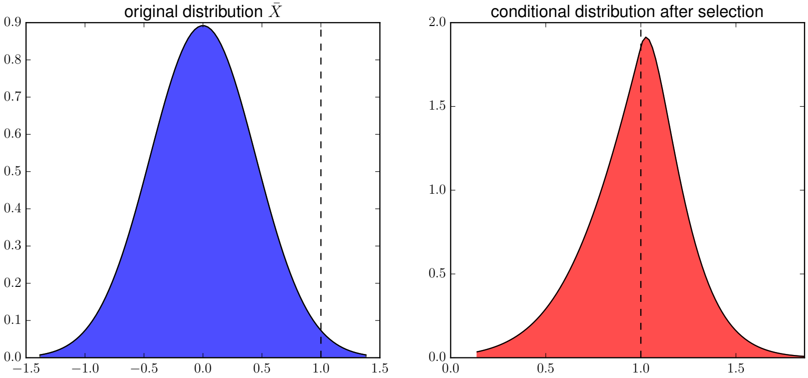
Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, $n = 5$.
Selection: $\overline{X} > 1$.





Selective v.s. unselective distributions

Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, $n = 5$.
Selection: $\overline{X} + \color{red}{\omega} > 1$, where $\omega \sim \text{Laplace}\,(0.15)$
Explicit formulas for the densities of the selective distribution.



The selective distribution is much better behaved after randomization

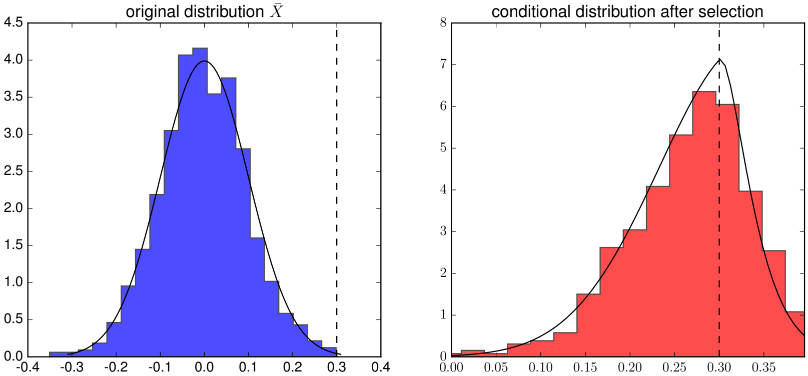
Selective v.s. unselective distributions: weak convergence

Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, $n = 100$.
Selection: $\overline{X} + \omega > 0.3$, $\omega \sim$ Laplace(0.03)

Selective v.s. unselective distributions: weak convergence

Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim}$ Laplace $\left(0, \frac{1}{\sqrt{2}}\right)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, $n = 100$.

Selection: $\overline{X} + \omega > 0.3$, $\omega \sim$ Laplace(0.03)



Selective central limit theorem

- Suppose $X_i \overset{i.i.d}{\sim} \mathbb{F}$, $X_i \in \mathbb{R}^k$.
- Linearizable statistics: $T = \frac{1}{n}\sum_{i=1}^n \xi_i(X_i) + o_p(n^{-\frac{1}{2}})$, with ξ_i being measurable to X_i's.
- Suppose $\xi_i(X_i) \in \mathbb{R}^p$, with mean $\mu \in \mathbb{R}^p$ and variance $\Sigma \in \mathbb{R}^{p\times p}$.

Theorem (Selective CLT, T. and Taylor (2015))

If model selection is made with $T^* = T^*(T, \omega)$, where the selection satisfies some regularity conditions, then

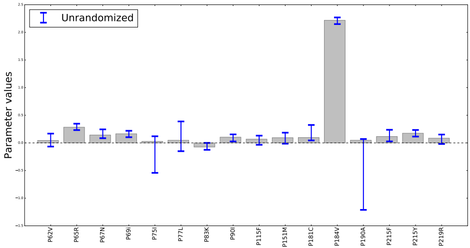
$$\mathcal{L}(T \mid H_0 \text{ selected by } T^*) \Rightarrow \mathcal{L}(N(\mu, \Sigma) \mid H_0 \text{ selected by } T^*),$$

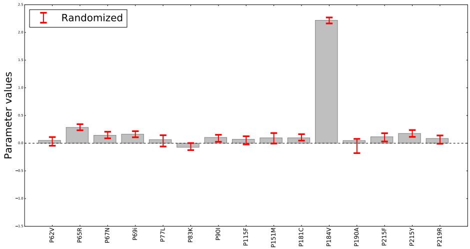
if T has moment generating function in a neighbourhood of the origin.

Power comparison

HIVDB http://hivdb.stanford.edu/

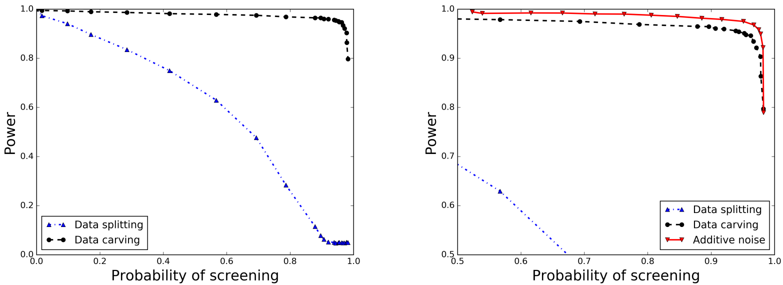
Unrandomized $y^* = y$, randomized $y^* = y + \omega$, $\omega \sim N(0, 0.1\sigma^2)$.





Tradeoff between power and model selection

- Setup $y = X\beta + \epsilon$, $n = 100, p = 200$, $\epsilon \sim N(0, I)$, $\beta = (\underbrace{7, \ldots, 7}_{7}, 0, \ldots, 0)$. X is equicorrelated with $\rho = 0.3$.
- Use randomized y^* to fit Lasso, active set E:
  1. Data splitting / Data carving: $y^* = y_1$ random subset of y,
  2. Additive randomization: $y^* = y + \omega$, $\omega \sim N(0, \gamma^2 I)$.



Data carving picture credit Fithian et al. (2014).

A general randomization approach

- Limitations of some randomization schems:
  1. Data splitting / Data carving: non-independent data structure.
  2. Additive noise: discrete data
- Randomized convex program

Randomized convex program: an example

Randomized Lasso:

$$\hat{\beta}(y,\omega) = \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \omega^T\beta,$$

with λ fixed. A choice of λ, see Negahban et al. (2010).
Choice of the distribution for ω,

- $\omega \sim \text{Laplace}(\gamma)$, γ controls the amount of randomization
- $\omega = 0 \Rightarrow$ Lasso

Randomized convex program: an example

Randomized Lasso:

$$\hat{\beta}(y,\omega) = \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 + \omega^T \beta,$$

with λ fixed. A choice of λ, see Negahban et al. (2010).
Choice of the distribution for ω,

- $\omega \sim$ Laplace(γ), γ controls the amount of randomization
- $\omega = 0 \Rightarrow$ Lasso

Advantages:

- Can replace squared-error loss function with any loss.
- Simplicity of sampling.

The conditional distribution to sample

Target of inference based on $\hat{\beta}(y, \omega)$,

$$\mathcal{L}(y \mid \hat{\beta}(y, \omega) \in A).$$

- $A \subseteq \mathbb{R}^p$, where only coordinates in E can be nonzero.
- A can be the quadrant determined by the signs of $\hat{\beta}_{obs}$, Lee et al. (2013) with $\omega = 0$.

$$\hat{\beta}(y, \omega) \in A \iff (y, \omega) \in B$$

The conditional distribution to sample

Target of inference based on $\hat{\beta}(y, \omega)$,

$$\mathcal{L}(y \mid \hat{\beta}(y, \omega) \in A).$$

- $A \subseteq \mathbb{R}^p$, where only coordinates in E can be nonzero.
- A can be the quadrant determined by the signs of $\hat{\beta}_{obs}$, Lee et al. (2013) with $\omega = 0$.

$$\underbrace{\hat{\beta}(y, \omega) \in A}_{\text{simple}} \iff \underbrace{(y, \omega) \in B}_{\text{difficult}}$$

Change of variables

Summary:

- The conditional law is

$$\mathcal{L}(y \mid \hat{\beta}(y,\omega) \in A) = \mathcal{L}(y \mid (y,\omega) \in B)$$

  with B being a complicated set...

- Map

$$(y,\omega) \mapsto (y, \hat{\beta}(y,\omega))$$

Change of variables

Summary:

- The conditional law is

$$\mathcal{L}(y \mid \hat{\beta}(y,\omega) \in A) = \mathcal{L}(y \mid (y,\omega) \in B)$$

  with B being a complicated set...

- Map

$$(y,\omega) \mapsto (y, \hat{\beta}(y,\omega))$$

  Inverse true?

Change of variables: continued

- No! (y, ω) cannot be reconstructed from $(y, \hat{\beta})$. Lasso is a mix of hard and softthresholding.
- Subgradient of ℓ_1 penalty carries "information" about the inactive variables.

KKT condition and the subgradient

- Equalities
$$-X^T(y - X\hat{\beta}) + \hat{z} + \omega = 0.$$
Simple case, when $X = I$,
$$\begin{cases} y_E - \hat{\beta}_E + \hat{z}_E + \omega_E & = 0 \\ y_{-E} + \hat{z}_{-E} + \omega_{-E} & = 0 \end{cases}$$

KKT condition and the subgradient

- Equalities
$$-X^T(y - X\hat{\beta}) + \hat{z} + \omega = 0.$$
Simple case, when $X = I$,
$$\begin{cases} y_E - \hat{\beta}_E + \hat{z}_E + \omega_E & = 0 \\ y_{-E} + \hat{z}_{-E} + \omega_{-E} & = 0 \end{cases}$$
- Inequalities:
$$\hat{z}_E \cdot \hat{\beta}_E > 0 \qquad |\hat{z}_{-E}| < \lambda$$

KKT condition and the subgradient

- Equalities

$$-X^T(y - X\hat{\beta}) + \hat{z} + \omega = 0.$$

  Simple case, when $X = I$,

$$\begin{cases} y_E - \hat{\beta}_E + \hat{z}_E + \omega_E & = 0 \\ y_{-E} + \hat{z}_{-E} + \omega_{-E} & = 0 \end{cases}$$

- Inequalities:

$$\hat{z}_E \cdot \hat{\beta}_E > 0 \qquad |\hat{z}_{-E}| < \lambda$$

- Reconstruction Ψ:

$$\Psi : (y, \hat{\beta}, \hat{z}) \mapsto (y, X^T(y - X\hat{\beta}) - \hat{z}) = (y, \omega)$$

KKT condition and the subgradient

- Equalities
$$-X^T(y - X\hat{\beta}) + \hat{z} + \omega = 0.$$
Simple case, when $X = I$,
$$\begin{cases} y_E - \hat{\beta}_E + \hat{z}_E + \omega_E & = 0 \\ y_{-E} + \hat{z}_{-E} + \omega_{-E} & = 0 \end{cases}$$

- Inequalities:
$$\hat{z}_E \cdot \hat{\beta}_E > 0 \qquad |\hat{z}_{-E}| < \lambda$$

- Reconstruction Ψ:
$$\Psi : (y, \hat{\beta}, \hat{z}) \mapsto (y, X^T(y - X\hat{\beta}) - \hat{z}) = (y, \omega)$$

- Conditional law
$$(y, \hat{\beta}, \hat{z}) \mid \hat{z}_E \cdot \hat{\beta}_E > 0 \qquad |\hat{z}_{-E}| < \lambda$$

Summary

- Conditional approach
- Randomized selection procedure is more powerful
- Sampling the selective distribution (to be continued)
- Reference: http://arxiv.org/abs/1507.06739

Summary

- Conditional approach
- Randomized selection procedure is more powerful
- Sampling the selective distribution (to be continued)
- Reference: http://arxiv.org/abs/1507.06739

Thank you!

Random hypothesis: revisited

In high dimensional statistics, the consistency of the estimators depends on the rate (Negahban et al. 2010),

$$\sigma\sqrt{\frac{\log p}{n}}$$

- Cross validation: $y^* = y_1 \in \mathbb{R}^{n_1}$,

$$n \to n_1$$

- Additive randomization: $y^* = y + \omega$, $\sigma^* = \sqrt{1+\gamma}\sigma$

Fithian, W., Sun, D. & Taylor, J. (2014), 'Optimal inference after model selection', *arXiv:1410.2597 [math, stat]* . arXiv: 1410.2597.
URL: *http://arxiv.org/abs/1410.2597*

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2013), 'Exact post-selection inference with the lasso', *arXiv preprint arXiv:1311.6238* .

Negahban, S., Ravikumar, P., Wainwright, M. J. & Yu, B. (2010), 'A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers', *arXiv:1010.2731* .
URL: *http://arxiv.org/abs/1010.2731*