LAD Fused Lasso Signal Approximation

Xiaoli Gao
Oakland University

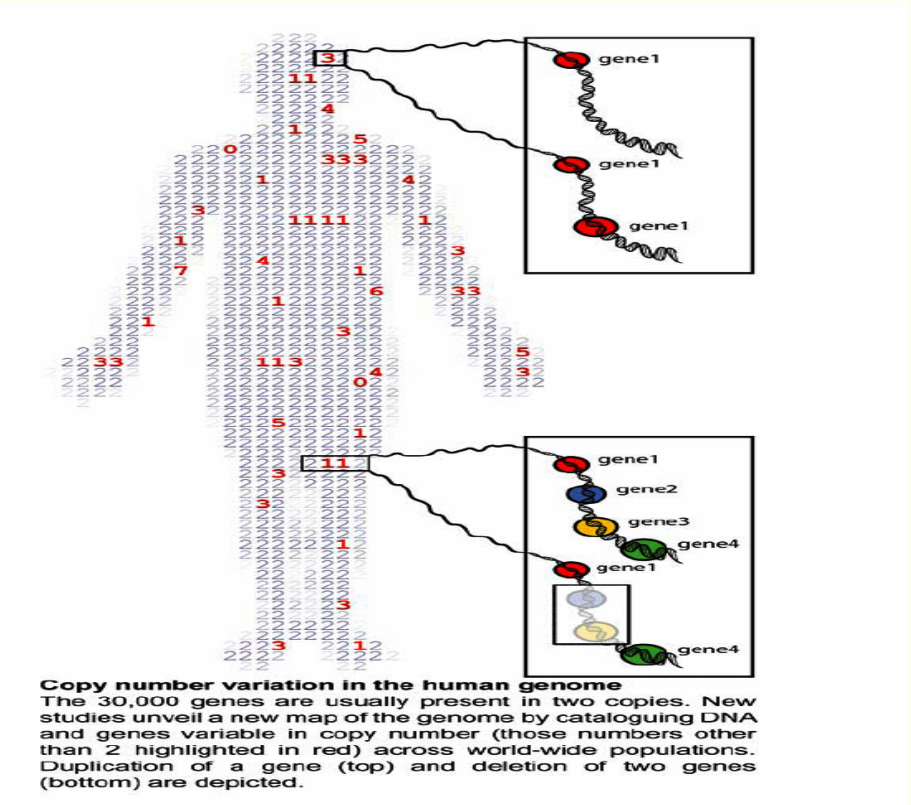
International Workshop on Perspectives on High-dimensional Data Analysis, Toronto

June 10, 2011

Outline

- Background: DNA copy number variation
- LAD Fused lasso signal approximation
- Asymptotic properties of LAD-FLSA
- Computation and numerical studies
- Concluding remarks

DNA copy number variation in the human genome



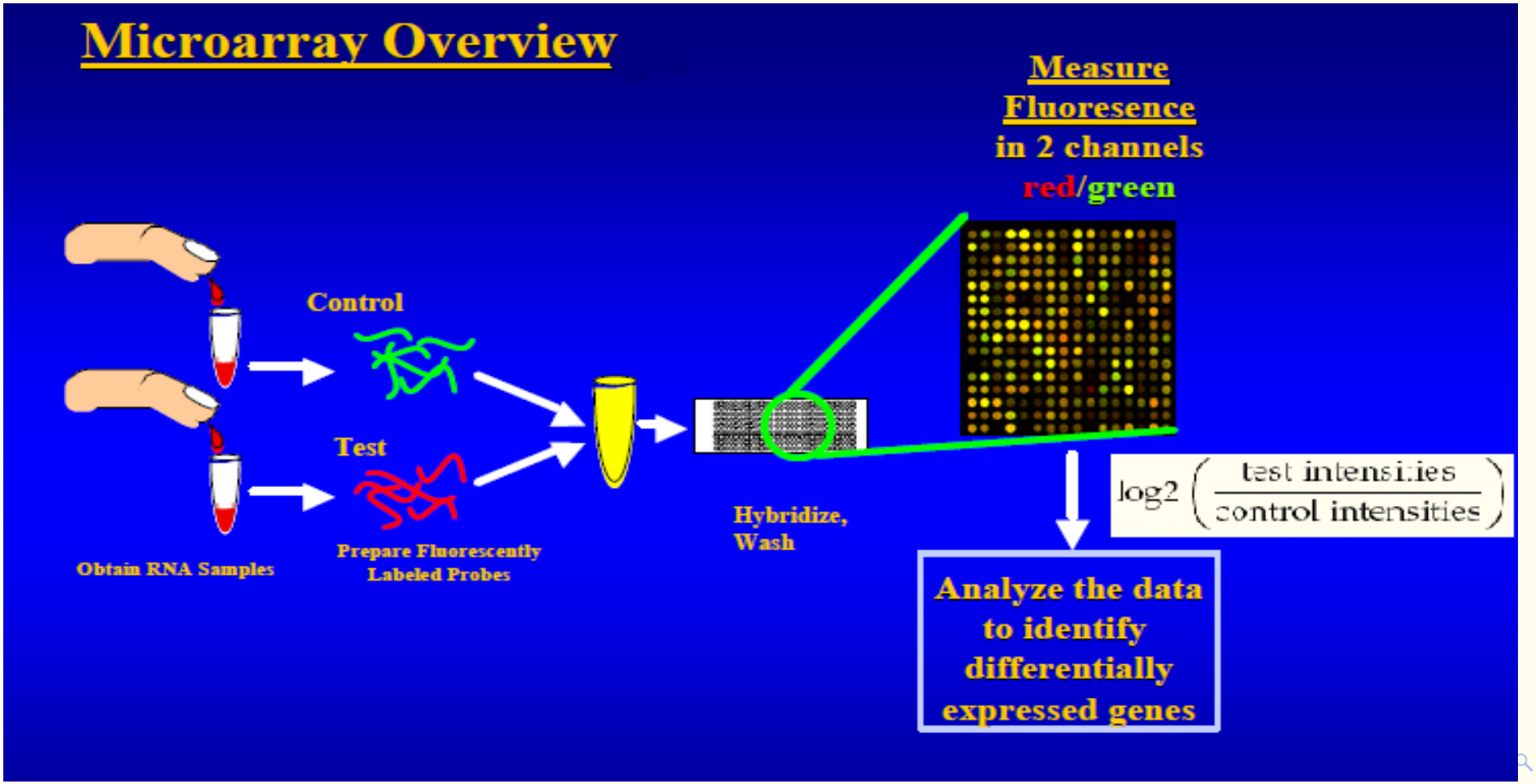
Copy number variation in the human genome
The 30,000 genes are usually present in two copies. New studies unveil a new map of the genome by cataloguing DNA and genes variable in copy number (those numbers other than 2 highlighted in red) across world-wide populations. Duplication of a gene (top) and deletion of two genes (bottom) are depicted.

Copy number variation (CNVs)

Deletions, insertions, duplications and complex multi-site variants are collectively termed as CNV. Most CNVs are neutral but some of them are functional and influence phenotypic differences between humans.

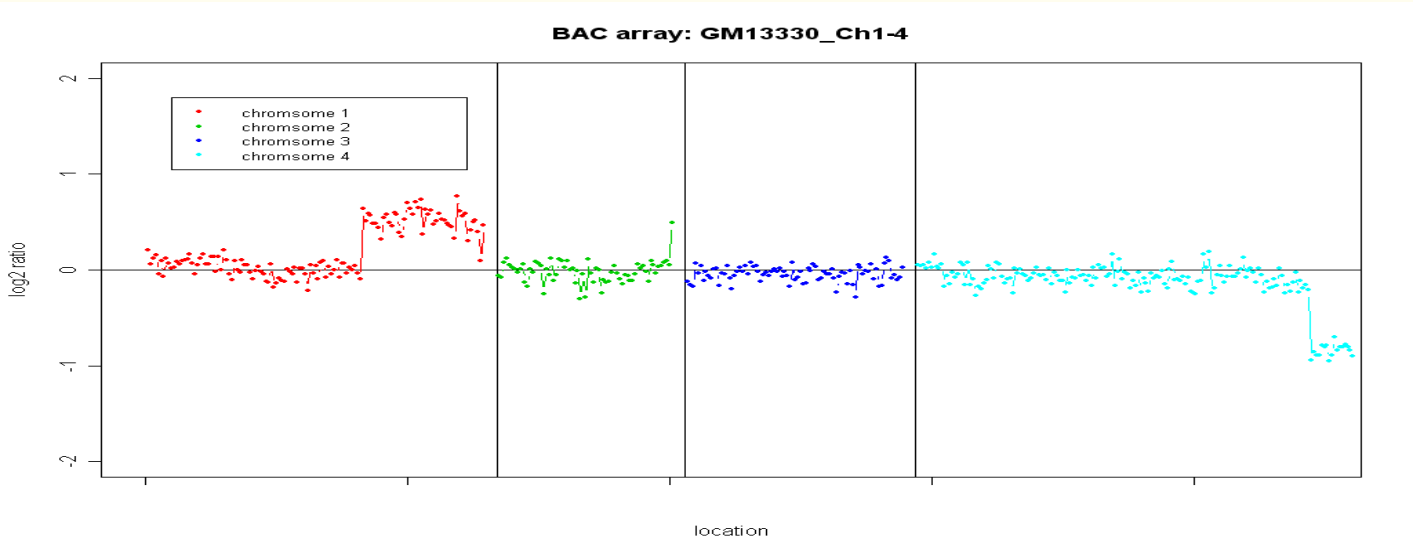
- CNVs influence gene expression, phenotypic variation by disrupting genes and altering gene dosage.
- CNVs improve our ability of survive (e.g. mutations in the CCR5gene protect against AIDS)
- CNVs confer risk to complex diseases.
- The contribution of CNV to many common diseases (e.g. heart disease, cancer, diabetes, and psychiatric disorders like schizophrenia and bipolar disorder) is largely unknown.

Steps in using a microarray



CNVs data: BAC array

Figure: Sample data of Copy number



CNV data features

- High dimensionality
- Sparsity
- Spatial local smoothness (along the chromosome)

Statistical Methods

- Formulate CNV detection into a regression problem
  - Select copy number amplification/deletion regions correctly
  - Recover the underlying relative intensities and detect all the true copy number variations.

Signal approximation model

Linear regression: $y_i = \mu_i^0 + \varepsilon_i, i = 1, \cdots, n.$

- μ_i^0 is the true signal at i
- ε_i is the noise
- y_i is the realization of hidden signal and noise at i

True model

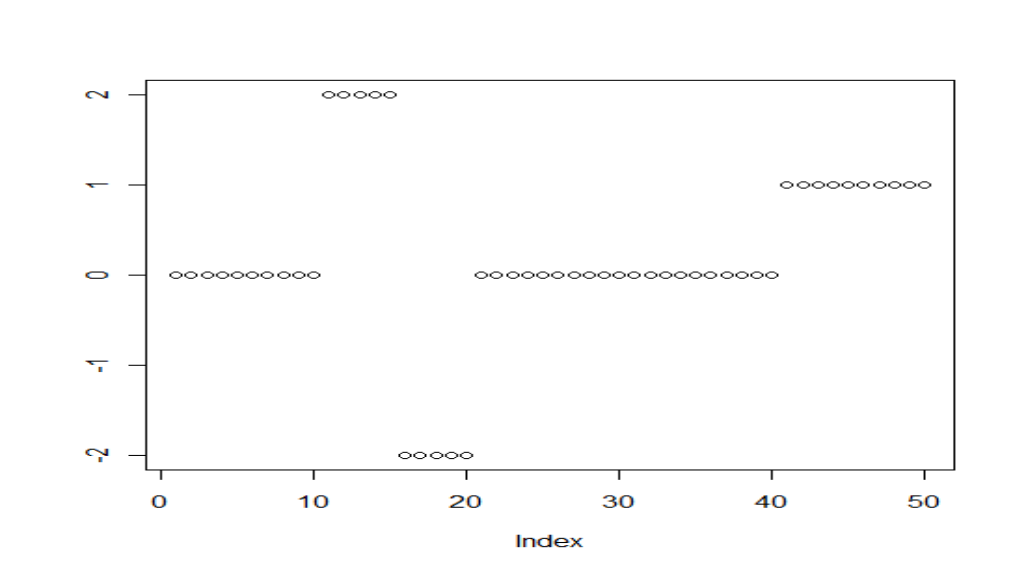
- Blocky:

$$\mu_i^0 = \sum_{j=1}^{J_0} v_j^0 \, I(i \in \mathcal{B}_j^0),$$

  where $\mathcal{B}_j^0$, $1 \leq j \leq J_0$ is the block partition and J_0 is the number of blocks

- Sparse: many of μ_i^0's are zero $\Longleftrightarrow v_j^0 = 0$, for $j \notin \mathcal{K}^0$, $\mathcal{K}^0$ is the nonzero block set.

A toy example



- $\mu' = (\mathbf{0}'_{10}, \mathbf{2}'_5, -\mathbf{2}'_5, \mathbf{0}'_{20}, \mathbf{1}'_{10})$
- $v' = (0, 2, -2, 0, 1)$
- $J_0 = 5$, $\mathcal{K}^0 = \{2, 3, 5\}$, $\mathcal{T}^0 = \{11, 16, 21, 41\}$

Recover the hidden signals

- How to estimate $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ when n increases?
- Penalized objective function:

$$\text{Loss}(\boldsymbol{\mu}; \mathbf{y}) + P_\lambda(\boldsymbol{\mu})$$

Fused Lasso penalty (Tibshirani et al. 2005)

- How to obtain a "blocky" and "sparse" estimate?
  - The lasso penalty: penalizing the ℓ_1 norm of the signals $\|\boldsymbol{\mu}\|_1 \equiv \sum_{i=1}^{n} |\mu_i|$ to enforce a *sparse* solution
  - The total variation penalty: penalizing $\|\boldsymbol{\mu}\|_{\mathrm{TV}} \equiv \sum_{i=1}^{n} |\mu_i - \mu_{i-1}|$ to enforce a *blocky* solution
- The combination of these two penalties results in the fused lasso (FL) penalty (Tibshirani et al., 2005)
- The signal approximation approach using fused lasso is denoted as "FLSA"

Previous penalized methods on signal approximation

We only list a few studies on the theoretical investigation:

- LS signal approximation using total variation penalty (LS-FSA, Boysen et al. 2009, Harchaoui and Lévy-Leduc 2010,$\cdots$)

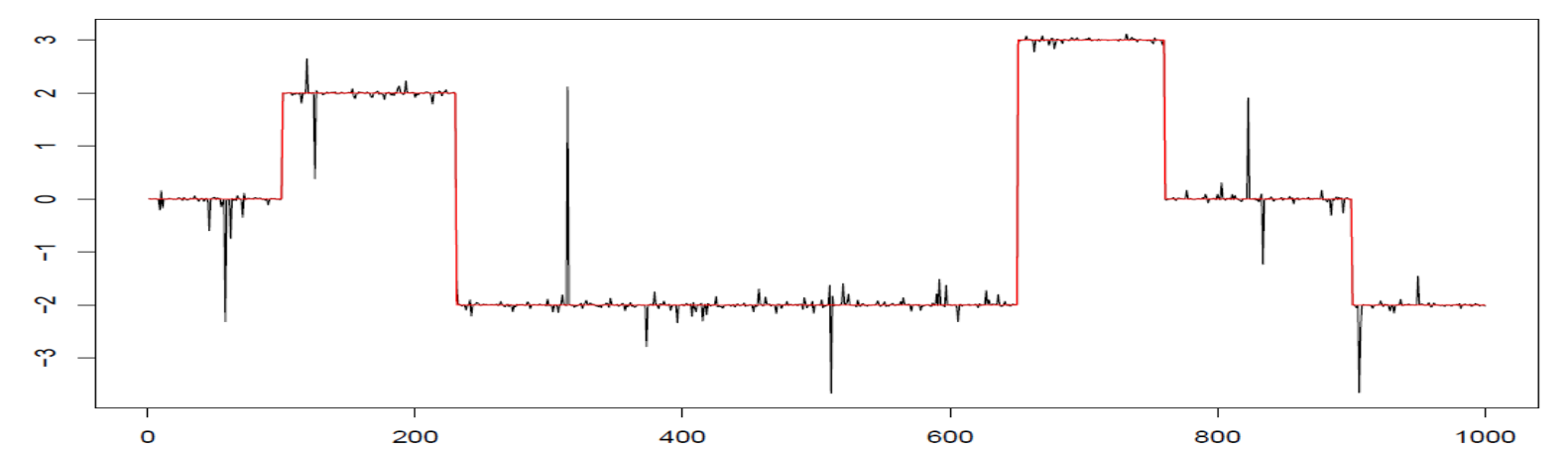
$$\widehat{\boldsymbol{\mu}}_n^{\ell_2}(\lambda_n) = \arg\min\{\sum_{i=1}^{n}(y_i-\mu_i)^2 + \lambda\sum_{i=2}^{n}|\mu_i-\mu_{i-1}|\}$$

- LS signal approximation using fused lasso penalty (LS-FLSA, Rinaldo 2009,$\cdots$)

$$\widehat{\boldsymbol{\mu}}_n^{\ell_2}(\lambda_{1n},\lambda_{2n}) = \arg\min\{\sum_{i=1}^{n}(y_i-\mu_i)^2 + \lambda_{1n}\sum_{i=1}^{n}|\mu_i| + \lambda_{2n}\sum_{i=2}^{n}|\mu_i-\mu_{i-1}|\}$$

LS solution can be invalid

- Previous work are studied under ℓ_2 loss
- When data is contaminated or the normal assumption is violated, ℓ_1 loss becomes a good alternative



LAD fused lasso signal approximation

- LAD-FLSA solution (Gao and Huang 2010):

$$\widehat{\boldsymbol{\mu}}_n^{\ell_1}(\lambda_{1n},\lambda_{2n}) = \arg\min\{\sum_{i=1}^{n}|y_i-\mu_i| + \lambda_{1n}\sum_{i=1}^{n}|\mu_i| + \lambda_{2n}\sum_{i=2}^{n}|\mu_i-\mu_{i-1}|\}$$

- LAD-FSA solution:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{F}}(\lambda_{2n}) = \widehat{\boldsymbol{\mu}}_n^{\mathrm{FL}}(0,\lambda_{2n}) = \arg\min\left\{\sum_{i=1}^{n}|y_i-\mu_i| + \lambda_{2n}\sum_{i=2}^{n}|\mu_i-\mu_{i-1}|\right\}.$$

- LAD has robust properties when the data set is contaminated by some outliers
- Computation is easy since all three terms are ℓ_1 norm

Statistical questions and properties

Questions: Under both the *block* and the *sparsity* assumptions, for large n,

(1) how close $\widehat{\mu}_n$ can be to the true model μ^0 asymptotically?

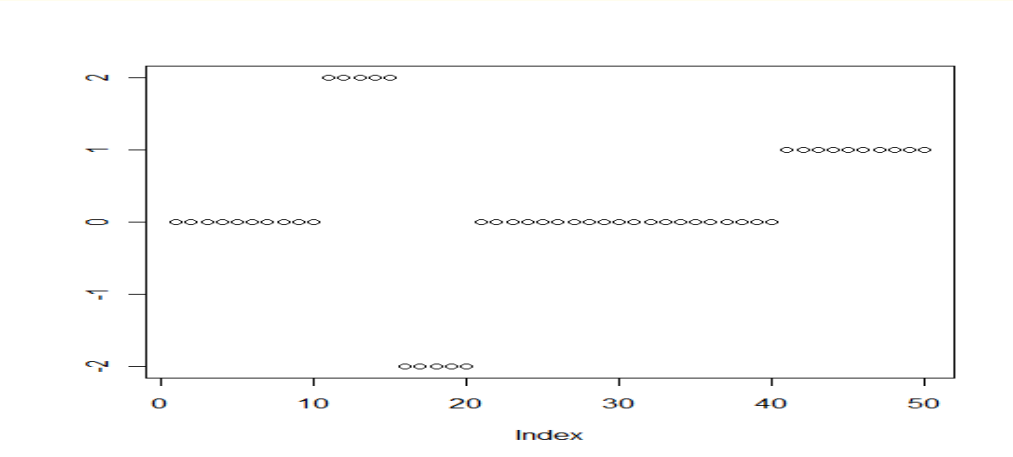
(2) how accurately $\widehat{\mu}_n$ can recover the true nonzero blocks with a large probability?

(3) what is the complexity of LAD-FLSA as a modeling procedure for (λ_1, λ_2)?

In general, the theoretical studies on LAD regression is much harder than LS regression

Some notations

- $b^0_{\min} = \min_{1 \le j \le J_0} b^0_j$, the smallest block size;
- $a_n = \min_{i \in \mathcal{J}^0} |\mu^0_i - \mu^0_{i-1}|$, the smallest jump;
- $\rho_n = \min_{j \in \mathcal{K}^0} |\nu^0_j|$, the smallest nonzero signal intensity.
- e.g. $b^0_{\min} = 5$, $a_n = 1$, $\rho_n = 1$.



Error assumption A1

(A1) Random errors ε_i's are independent and identically distributed with median 0, and have a density f that is continuous and positive in a neighborhood of 0.

Assumption A2

(A2) There exists a constant $M_1 > 0$ such that the true jump size $J_0 < M_1 \Lambda_n$, where $\Lambda_n = \max\{16n/(\lambda_{2n}^2 - 2n^2\lambda_{1n}^2), n/(\lambda_{2n} - n\lambda_{1n})\} + 1$ for $\lambda_{2n}^2 > 2n^2\lambda_{1n}^2$.

Estimation consistency

Theorem

Suppose (A1) and (A2) hold. Then there exists a constant $0 < c < 1$ *such that*

$$\mathbf{P}\left(\|\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^0\|_n \geq \gamma_n\right) \leq \Lambda_n \exp\{\Lambda_n \log n - (1-c)^2 (f(0)/8) n\gamma_n^2\} + (8/f(0))(\Lambda_n/(n\gamma_n^2))^{1/2},$$

where Λ_n *is defined in (A2) and* $\gamma_n = 2/(c\sqrt{f(0)})[\lambda_{1n} + 2\lambda_{2n} + ((M_1+1)\Lambda_n/n)^{1/2}]$, *where* $\|\mathbf{x}\|_n = \left(\sum_{i=1}^n x_i^2/n\right)^{1/2}$. *Furthermore, if we choose* λ_{1n} *and* λ_{2n} *such that* $\lambda_{1n} + 2\lambda_{2n} = (c\sqrt{f(0)}/2)\gamma_n - ((M_1+1)\Lambda_n/n)^{1/2}$, *then*

$$\mathbf{P}\left(\|\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^0\|_n \geq \gamma_n\right) \leq \Lambda_n n^{\{1-M_3 f(0)(1-c)^2\}\Lambda_n} + O\left(1/\sqrt{\log n}\right),$$

where $\gamma_n = (8M\Lambda_n(\log n)/n)^{1/2}$.

Comments on the estimation results

- If the number of jumps is bounded, the convergence rate in terms of $\|\cdot\|_n$ norm can be compared to the optimal rate $n^{-1/2}$ (Yao and Yu, 1989)
- Furthermore, if the blocks partition is done correctly with a large probability, then $\widehat{\boldsymbol{\nu}}_n \to \boldsymbol{\nu}^0$ (*in terms of* ℓ_2 *norm*) at rate $((\log n)/b^0_{\min})^{-1/2}$.

Definition

$\widehat{\mu}_n$ is *jump selection consistent* if

$$\lim_{n\to\infty} \mathbf{P}\left(\{\widehat{J} = J_0\} \bigcap \{\cap_{1\le j\le J_0}\{\widehat{\mathcal{B}}_j = \mathcal{B}_j^0\}\}\right) = 1.$$

Definition

$\widehat{\mu}_n$ is *jump sign consistent* if

$$\lim_{n\to\infty} \mathbf{P}\left(\{\widehat{\mathcal{J}} = \mathcal{J}^0\} \bigcap \{\mathrm{sgn}(\widehat{\mu}_i - \widehat{\mu}_{i-1}) = \mathrm{sgn}(\mu_i^0 - \mu_{i-1}^0), \forall i \in \mathcal{J}^0\}\right) = 1.$$

Definition

$\widehat{\mu}_n$ is *block selection consistent* if

$$\lim_{n\to\infty} \mathbf{P}\left(\{\widehat{\mathcal{J}} = \mathcal{J}^0\} \bigcap \{\widehat{\mathcal{K}} = \mathcal{K}^0\}\right) = 1.$$

Definition

$\widehat{\mu}_n$ is *block sign consistent* if

$$\lim_{n\to\infty} \mathbf{P}\left(\{\widehat{\mathcal{J}} = \mathcal{J}^0\} \bigcap \{\widehat{\mathcal{K}} = \mathcal{K}^0\} \bigcap \{\mathrm{sgn}(\widehat{\nu}_j) = \mathrm{sgn}(\nu_j^0), \forall j \in \mathcal{K}^0\}\right) = 1.$$

Assumptions B

(B1) (a) $\lambda_{2n} \to \infty$;
(b) there exists a $\delta > 0$, such that $\lambda_{2n}(\log(n - J_0))^{-1/2} > (1+\delta)/2$.

(B2) (a) $(b^0_{\min})^{1/2} a_n \to \infty$;
(b) there exists $\delta > 0$, such that
$(b^0_{\min}/\log(J_0))^{1/2} a_n > 3(1+\delta)/(\sqrt{2} f(0))$ for sufficiently large n.

(B3) $\lambda_{2n} < (f(0)/3) b^0_{\min} a_n$ for sufficiently large n.

Jump selection consistency of LAD-FSA solution

A LAD-FSA solution is

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{F}}(\lambda_{2n}) = \widehat{\boldsymbol{\mu}}_n^{\mathrm{FL}}(0, \lambda_{2n}) = \arg\min \left\{ \sum_{i=1}^{n} |y_i - \mu_i| + \lambda_{2n} \sum_{i=2}^{n} |\mu_i - \mu_{i-1}| \right\}.$$

Theorem

A LAD-FSA solution $\widehat{\boldsymbol{\mu}}_n^{F}(\lambda_{2n})$ is jump sign consistent under (A1) and (B1-B3).

Assumptions C

(C1): (a) $\lambda_{1n}(b^0_{\min})^{1/2} \to \infty$ when $n \to \infty$;
(b) there exists $\delta > 0$, such that
$\lambda_{1n}(b^0_{\min}/\log(J_0 - K_0))^{1/2} > 4\sqrt{2}(1+\delta)$.

(C2): $\lambda_{2n}/b^0_{\min} < \lambda_{1n}/8$ when n is large enough.

(C3): (a) $\rho_n(b^0_{\min})^{1/2} \to \infty$ when $n \to \infty$;
(b) there exists $\delta > 0$ such that
$\rho_n(b^0_{\min}/\log(K_0))^{1/2} > 2\sqrt{2}(1+\delta)/f(0)$.

(C4): $\lambda_{2n}/b^0_{\min} < f(0)\rho_n/3$ when n is large enough.

(C5): $\lambda_{1n} < f(0)\rho_n/2$ when n is large enough.

Note: some conditions can be redundant. (C2) & (C5) $\Longrightarrow$ (C4).
(C5) & (C1-a) $\Longrightarrow$ (C3-a).

Sign consistency of LAD-FLSA solution

Theorem

Under (A1), (B1-B3) and (C1-C5), a LAD-FLSA estimator is block sign consistent.

Computation

- All ℓ_1 norms facilitate the computation for fixed (λ_1, λ_2)
- Using previous results to provide reasonable ranges of λ_1 and λ_2
- For example: choose $0 < \lambda_1 < 0.5$ with an increment of 0.01; choose $(n/\log(n))^{1/2} < \lambda_{2n} < n^{1/2}$ with an increment of 0.1.

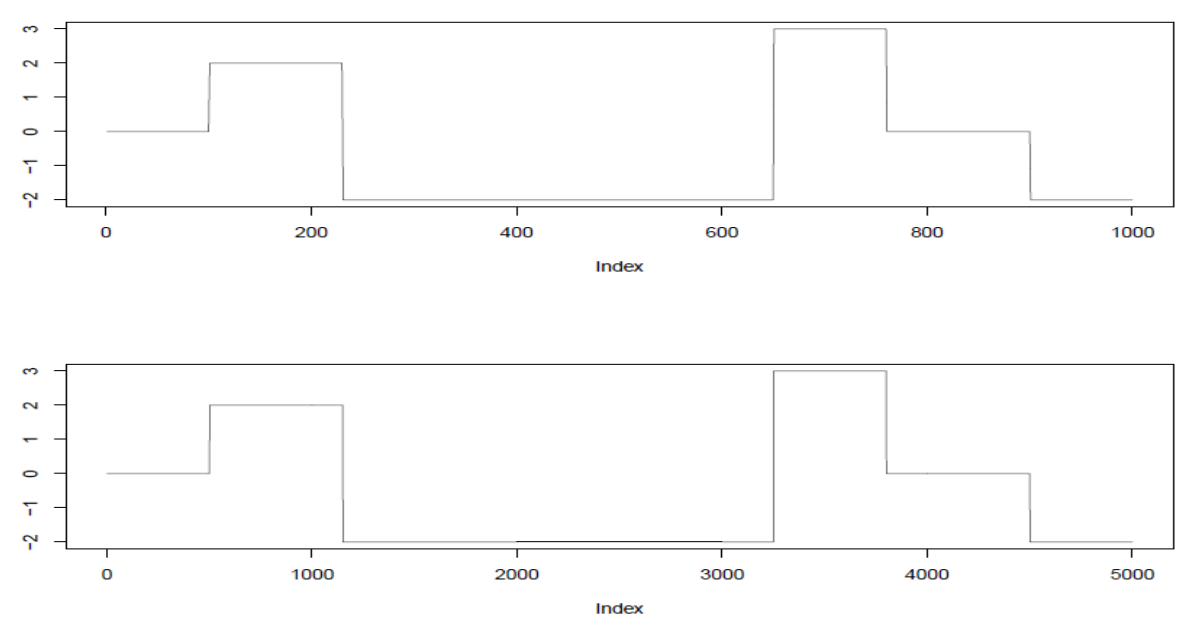
Tuning parameter selection

- $D(\mathcal{M}_{\lambda_1,\lambda_2}) = \sum_{i=1}^{n} \partial \mathrm{E}[\widehat{\mu}_i(\mathbf{y};\lambda_1,\lambda_2)]/\partial y_i$ (Gao and Fang, 2011)
- $\mathbf{E}[|\widehat{\mathcal{K}}(\lambda_1,\lambda_2)|] = D(\mathcal{M}_{\lambda_1,\lambda_2})$
- BIC : $\sum_{i=1}^{n} |y_i - \widehat{\mu}_i(\lambda_1,\lambda_2)| + |\widehat{\mathcal{K}}(\lambda_1,\lambda_2)| \log(n)/2$

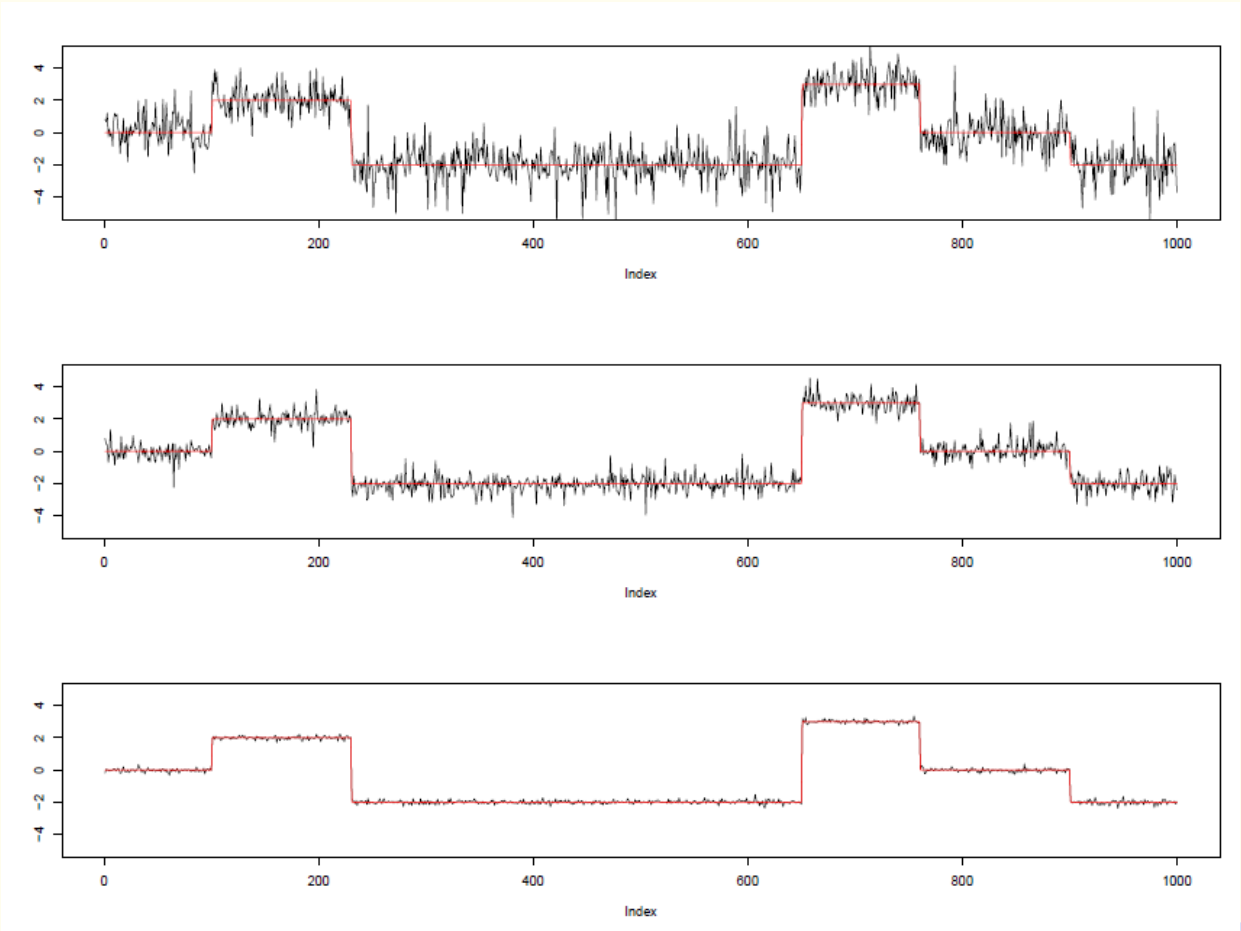
Simulation set-up

- $\boldsymbol{\mu}^0 = (\mathbf{0}'_{p_{1n}} \mathbf{2}'_{p_{2n}} - \mathbf{2}'_{p_{3n}} \mathbf{3}'_{p_{4n}} \mathbf{0}'_{p_{5n}} \mathbf{2}'_{p_{6n}})'$
- normal/double exponential distribution/Cauchy distribution
- Weak/mild/strong noises
- $n = 1000$ or 5000

True model


-2
-1
0
1
2
3
0
200
400
600
800
1000
Index
-2
-1
0
1
2
3
0
1000
2000
3000
4000
5000
Index

Sample data sets



Evaluate the performances

- OFR+6(CFR): the ratio of recovering $\widehat{\mu}^0$ correctly or plus at most six additional false positives (correctly fitted ratio).
- JUMP: the average number (standard deviation) of the number of jumps.
-

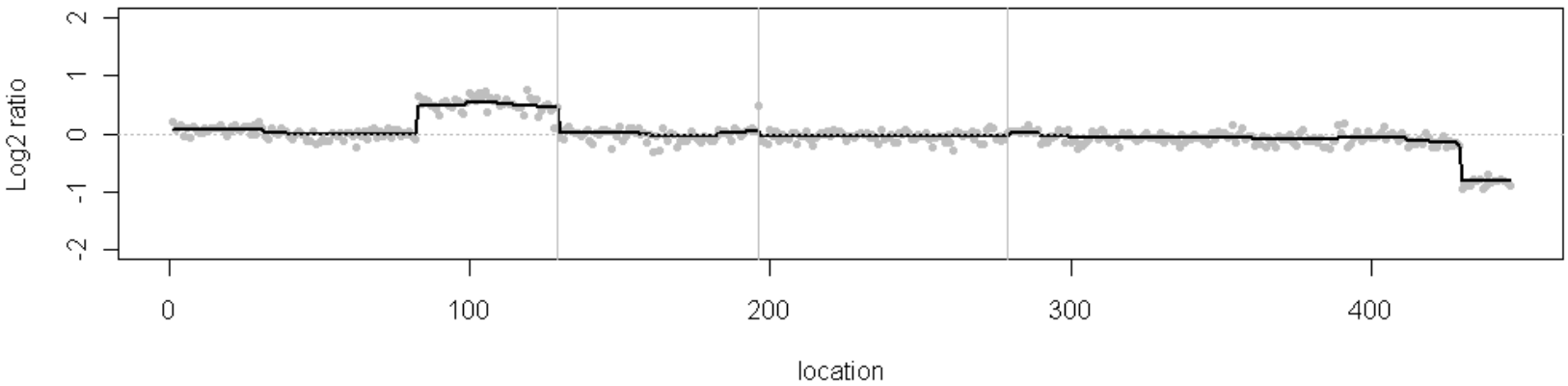
$$\mathrm{LARE}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^0) = \frac{\sum_{i=1}^n |\widehat{\mu}_i - \mu_i^0|}{\sum_{i=1}^n |\mu_i|}. \tag{1}$$

Simulation results

			$n = 1000$			$n = 500$		
ϵ_i	σ	**Model**	LARE[1]	OFR+6(CFR)[2]	JUMP[3]	LARE	OFR+6(CFR)	JUMP
Normal	1.0	LAD-FLSA	0.197	89%(17%)	7.12(1.32)	0.217	81%(10%)	6.95(1.22)
		LS-FLSA	0.035	18%(3%)	7.82(1.47)	0.048	16%(5%)	7.54(1.43)
	0.5	LAD-FLSA	0.098	97%(32%)	5.59 (0.75)	0.109	95%(19%)	5.70(0.89)
		LS-FLSA	0.016	48%(13%)	5.68(0.74)	0.029	55%(10%)	5.64(0.79)
	0.1	LAD-FLSA	0.019	100%(93%)	5(0)	0.021	100%(92%)	5(0)
		LS-FLSA	0.013	100%(93%)	5(0)	0.026	100%(93%)	5(0)
Double Exp.	1.0	LAD-FLSA	0.154	88% (22%)	7.42(1.54)	0.183	93%(27%)	6.88(1.08)
		LS-FLSA	0.031	12%(0%)	7.42(1.42)	0.044	19%(2%)	7.07(1.44)
	0.5	LAD-FLSA	0.077	97%(34%)	5.95(0.90)	0.091	99%(43%)	5.71(0.84)
		LS-FLSA	0.016	57%(12%)	5.73(0.78)	0.029	57%(14%)	5.55(0.66)
	0.1	LAD-FLSA	0.015	100%(97%)	5(0)	0.018	100%(95%)	5.01(0.1)
		LS-FLSA	0.013	100%(97%)	5(0)	0.026	100%(95%)	5.01(0.1)
Cauchy	1.0	LAD-FLSA	0.048	87%(56%)	6.12(1.07)	0.051	82%(50%)	5.95(0.89)
		LS-FLSA	0.239	17%(4%)	16.37(5.38)	0.175	19%(8%)	10.27(3.10)
	0.5	LAD-FLSA	0.028	99%(70%)	5.56(0.86)	0.027	91%(73%)	5.62(0.72)
		LS-FLSA	0.120	39%(17%)	10.67(3.62)	0.081	49%(30%)	7.89(2.16)
	0.1	LAD-FLSA	0.007	95%(92%)	5.18(0.46)	0.005	99%(95%)	5.16(0.37)
		LS-FLSA	0.029	94%(78%)	6.30(1.34)	0.033	85%(76%)	5.63(1.07)

BAC array data analysis

A CGH array consisting of 2400 bacterial artificial chromosome (BAC) clones. Sample CGH copy number data on chromosomes 1–4 from cell line GM 13330 is analyzed.



Numerical studies on DF

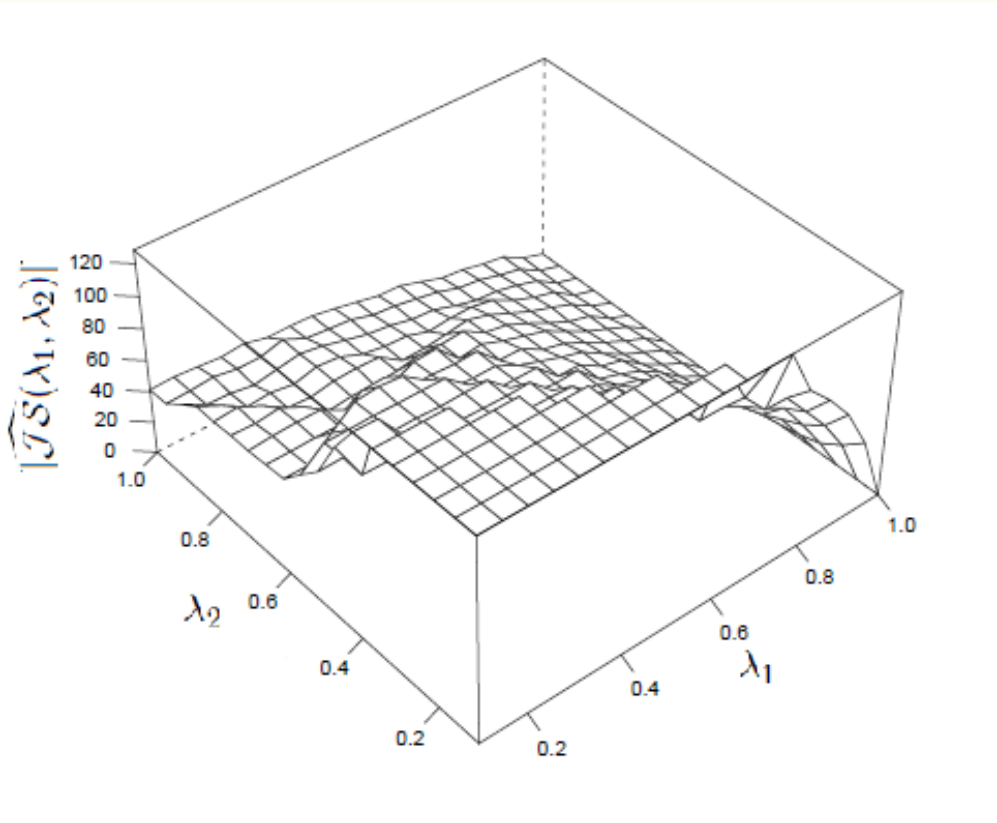
- Hypothetical data from chromosome 1 with 129 locations of GM 13330

$$y_i^0 = y_i + \varepsilon_i^0, \quad i = 1, \cdots, 129, \quad \varepsilon_i^0 \overset{\text{i.i.d}}{\sim} N(0, 0.1\sigma^*),$$

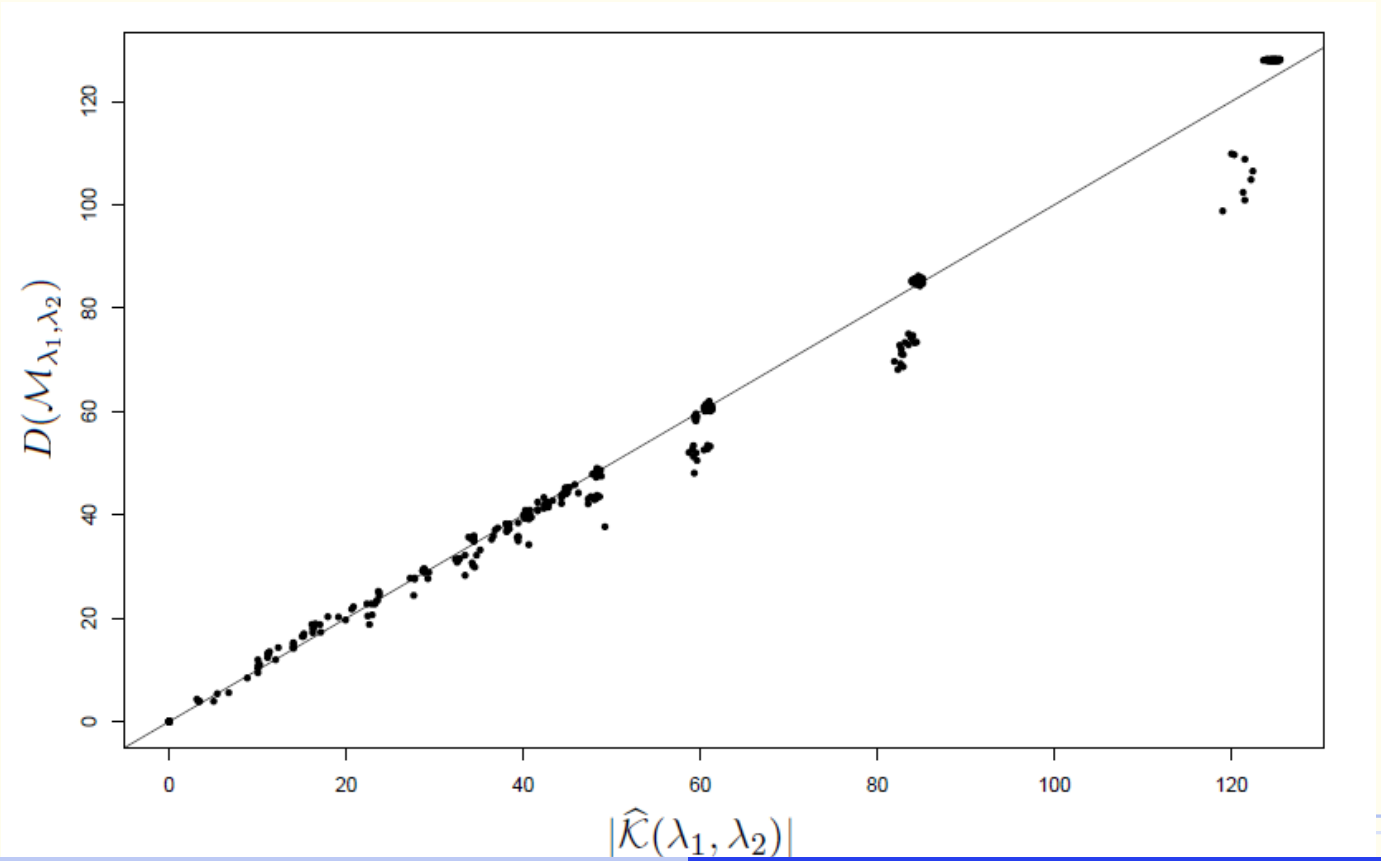
  σ^* is the standard deviations of $y - 1, \cdots, y_{129}$

- $\widehat{\text{GDF}}(\lambda_1, \lambda_2) = |\widehat{\mathcal{K}}(\lambda_1, \lambda_2)|$
- $\text{GDF}(\lambda_1, \lambda_2)$:sum of sensitivities (Algorithm 1 in Ye (1998))
- 500 Monte Carlo simulations

DF changes with tuning parameters



Unbiased estimator of DF



Summary

- (Robust properties) A LAD-FLSA estimate can be more efficient than an LS-FLSA when the data have heavy noises or the data are contaminated by outliers
- (Estimation) A LAD-FLSA estimator can be estimation consistent. It almost reaches an almost optimal rate if the block size is fixed
- (Variable selection) A LAD-FLSA estimator can be sign consistent under some sufficient conditions

References

- Boysen, L., Kempe, A., Liebscher, V., Munk, A. ad Wittich, O.L. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. AOS, **37**, 157–183.
- Gao, X. and Fang, Y. (2011). Generalized degrees of freedom under the L_1 loss function. JSPI, **141**, 677–686 .
- Gao, X.L. and Huang, J. (2010). A robust penalized method for the analysis of noisy DNA copy number data, BMC Genomics, 11:517.
- Gao, X.L. and Huang, J. (2011). Estimation and Selection Properties of the LAD Fused Lasso Signal Approximator. In submission
- Harchaoui, Z. and Lévy-leduc, C. (2010). Multiple change-point estimation with a total variation penalty. JASA, **105**, 1480–1493.

References

- Rinaldo, A. (2009). Properties and refinements of the fused lasso. AOS, **37**, 2922–2952.
- Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. and Albertson D. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. Nature Genetics, **29**, 263–264.
- Tibshirani, R., Saunders, M., Rosset, S., zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. JRSS-B, **67**, 91–108.
- Yao, Y., and Au, S. T. (1989). Least-squares estimation of a step function. Sankhya-A, **51**, 370–381.

Thank Ejaz for the invitation. Thanks to all organizers.