Fused MCP with applications in signal processing

Xianshi Yu

Joint work with Bing-yi Jing, Guangren Yang and Cun-Hui Zhang

Department of Mathematics, HKUST

Jan 2015

Outline

linear regression model

- Given $(y_i, x_{i1}, ..., x_{ip})$, $i = 1, ..., n$, assume

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n.$$

- In matrix form,

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & a_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

Least square estimate (LSE)

- LSE:

$$\hat{\beta}^{lse} = \arg\min_{\beta}\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{XY}$$

- LSE is ill-posed if $p > n$
- In some situations, we can impose some assumptions on β

Lasso

- Lasso (Tibshirani (1996)):

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{2}\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- When $X^T X = I$, then

$$\hat{\beta}^{lasso} = sgn(\hat{\beta}^{lse}) \left(|\hat{\beta}^{lse}| - \lambda \right)^{+}$$

- It works for $p > n$ as well.
- L_1 penalty imposes sparsity on β
- It does shrinkage and variable selection simultaneously.

Fused Lasso

- In some situations, the features have an inherent order.
- Examples:
  - protein mass spectroscopy data
  - gene expression data
- Fused Lasso (Tibshirani (2005)) minimizes:

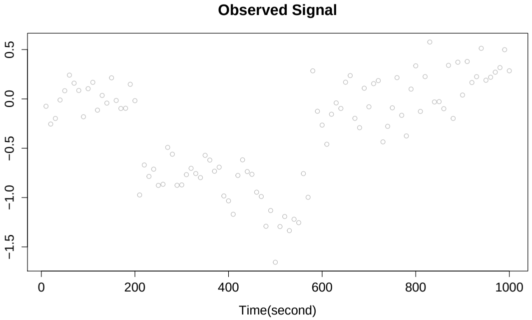
$$\frac{1}{2}\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

- It encourages both sparsity and local constancy in β

Outline

Signal Denoising Problem

- In signal processing scenario, $y_i = \beta_i + \epsilon_i$

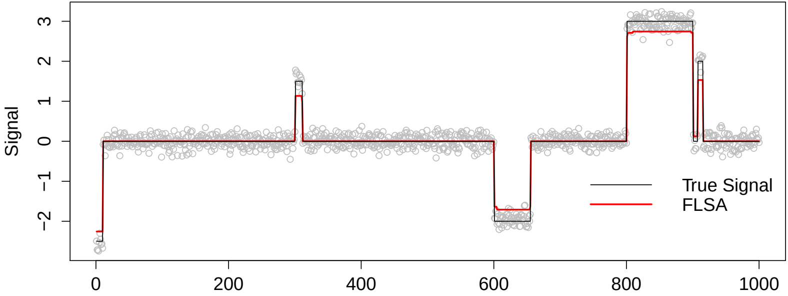


- Fusion penalty $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$ can be applied to denoise the corrupted signal

FLSA

Friedman(2007) introduced FLSA:

Minimize $\frac{1}{2}\|\mathbf{Y}-\beta\|_2^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=2}^{p}|\beta_j-\beta_{j-1}|$.



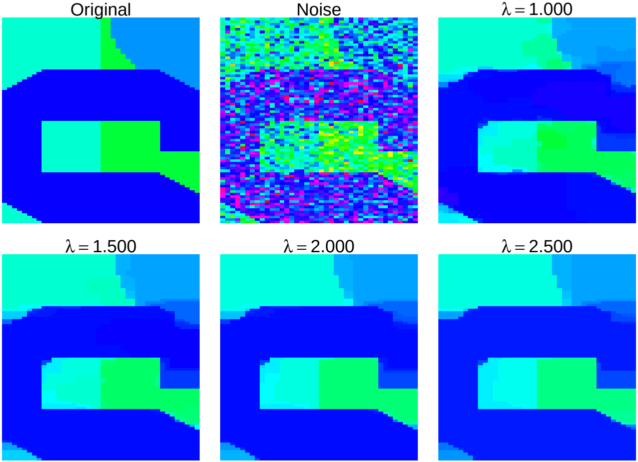
FLSA

FLSA is the discrete analogue of *total variation denoising* established by Rudin *et al.* (1992).

$$\min_u \int_\Omega |\nabla u| \, du \qquad \text{subject to} \qquad \|u - y\|^2 = \sigma^2$$

2-d FLSA

$$\text{Minimize } \frac{1}{2}\|\mathbf{Y}-\beta\|_F^2+\lambda\sum\left|\beta_{i,j}-\beta_{i,j-1}\right|+\lambda\sum\left|\beta_{i,j}-\beta_{i-1,j}\right|$$



Comments on FLSA

- FLSA captures the profile of the signals
- Contrast in signal is shrinked
- Jump points in 1-d recovered signal is not clear
- Edges in recovered image is not clear

Outline

Mcp penalty

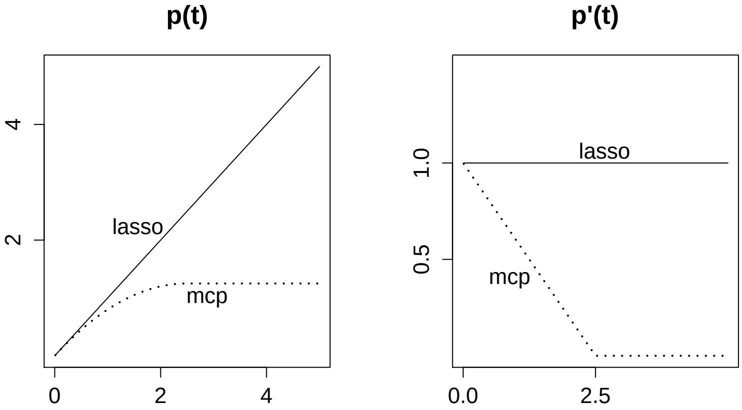
Mcp (Zhang (2010)):

$$\hat{\beta}^{mcp} = \arg\min_{\beta} \frac{1}{2}\{\sum_i (y_i - \sum_j x_{ij}\beta_j)^2\} + \sum_{j=1}^{p} \rho(|\beta_j|\,;\lambda)$$

where $\rho(t;\lambda) = \lambda \int_0^t (1 - x/(\gamma\lambda))_+ dx$

Mcp penalty

Comparison of Mcp and Lasso



Mcp penalty

- $\rho(t;\lambda)$ is nonconvex
- Mcp introduces sparsity
- Mcp dose not shrink β when it is large
- Mcp is nearly unbiased
- Computation: PLUS—-a path algorithm

Fused Mcp

- 1-d Fused Mcp minimizes

$$\frac{1}{2}\sum_i (y_i - \beta_j)^2 + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1) + \sum_{j=2}^{p} \rho(|\beta_j - \beta_{j-1}|; \lambda_2)$$

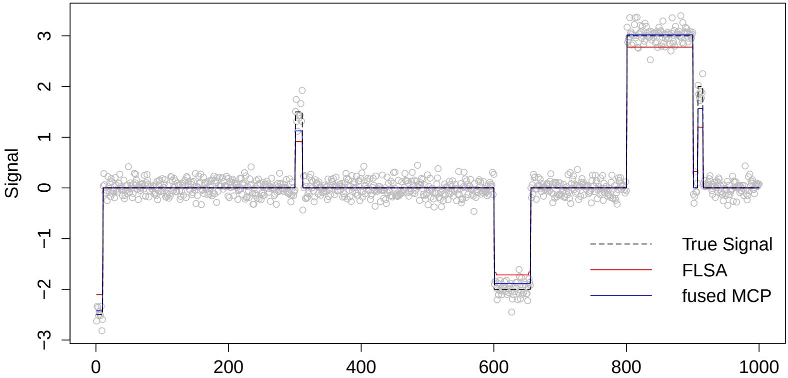
- Specifically, when $\lambda_1 = 0$, it is called fusion MCP.

$$\text{Minimize} \quad \frac{1}{2}\sum_i (y_i - \beta_j)^2 + \sum_{j=2}^{p} \rho(|\beta_j - \beta_{j-1}|; \lambda)$$

- 2-d Fused Mcp minimizes

$$\frac{1}{2}\sum_i (y_i - \beta_j)^2 + \sum \rho(|\beta_{i,j} - \beta_{i,j-1}|; \lambda) + \sum \rho(|\beta_{i,j} - \beta_{i-1,j}|; \lambda)$$

Fused MCP vs. FLSA

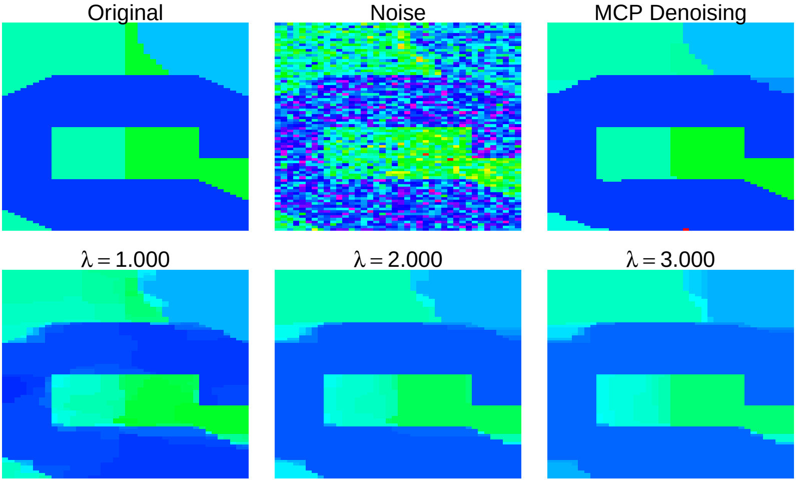


Fused MCP vs. FLSA

Fused MCP has better performance in jump point and jump size detection than FLSA.

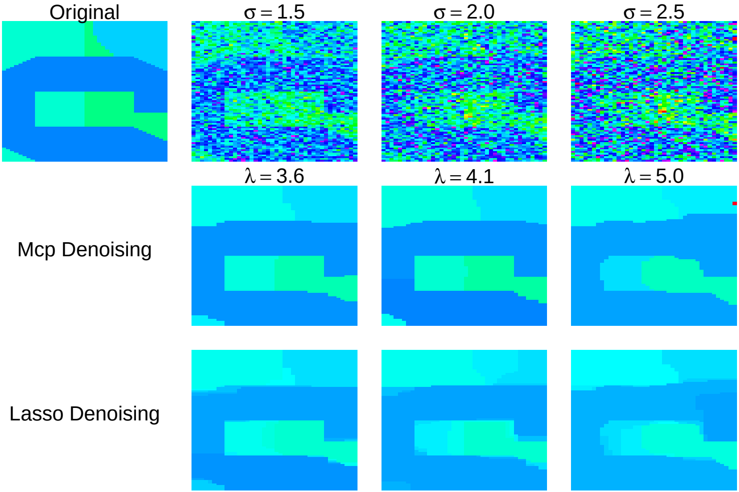
2-d fused MCP vs. 2-d FLSA

- When the noise is small



2-d fused MCP vs. 2-d FLSA

- When the magnitude of noise increases



2-d fused MCP vs. 2-d FLSA

2-d fused Mcp keeps the contrast of the colors and finds the edges in the images effectively. It largely improves 2-d FLSA.

Outline

Algorithms

fused MCP($\lambda_1 > 0$)	adjusted MM Algorithm
fusion MCP($\lambda_1 = 0b$)	PLUS Algorithm
2-d fused MCP	adjusted MM Algorithm

fusion MCP

$$\min_{\beta} \frac{1}{2}\|\mathbf{Y} - \beta\|_2^2 + \sum_{j=2}^{p} \rho(\left|\beta_j - \beta_{j-1}\right|; \lambda)$$

can be transformed to the objective function of an MCP penalized regression problem:

$$\min_{\eta} \frac{1}{2}\|\mathbf{A} - \mathbf{B}\eta\|_2^2 + \sum_{j=1}^{p-1} \rho(\left|\eta_j\right|; \lambda).$$

Thus, PLUS algorithm can be applied.

fused MCP

The objective function becomes more complex, PLUS algorithm is not applicable any more.

We design an adjusted majorization-minimization (MM) algorithm to solve this problem.

MM Algorithm for fused MCP

MM Algorithm

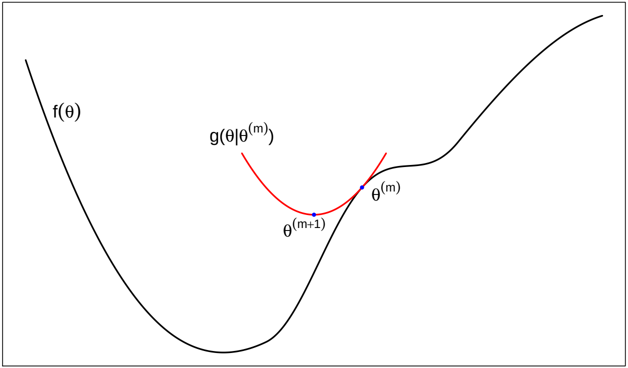


Figure: Illustration of one step in the MM algorithm

MM Algorithm for fused MCP

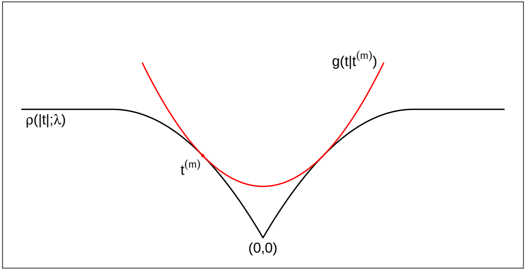
- $g(\theta \mid \theta^{(m)}) \;\geq\; f(\theta)$ for all θ
- $g(\theta^{(m)} \mid \theta^{(m)}) \;=\; f(\theta^{(m)})$
- $g(\theta \mid \theta^{(m)})$ is devised to be easy to solve
- The optimum θ of $f(\theta)$ is found by minimizing $g(\theta \mid \theta^{(m)})$ iteratively

MM Algorithm for fused MCP

The objective function is

$$\frac{1}{2}\sum_{i}(y_i-\beta_j)^2+\sum_{j=1}^{p}\rho(|\beta_j|;\lambda_1)+\sum_{j=2}^{p}\rho(|\beta_j-\beta_{j-1}|;\lambda_2).$$

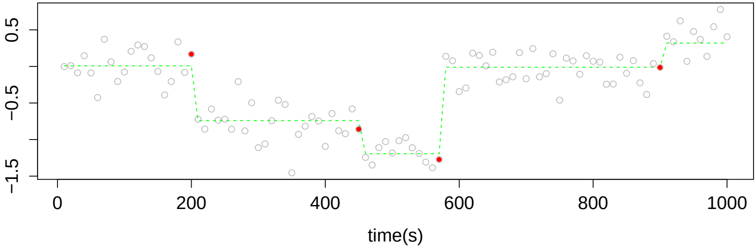
The term $\rho(|t|;\lambda_1)$ is majorized by $g(t \mid t^{(m)})$ and g is quadratic.



Outline

Oracle Estimation

If the jump points are known, they will partition the signal into several segments. A reasonable way to denoise the signal, knowing this partition, would be to average the observed signal in every segment.



The oracle estimation is denoted as $\beta^{\mathcal{O}}$.

Oracle Property of fusion MCP

Theorem

Suppose $\boldsymbol{Y}_{p\times 1} = \beta^o + \varepsilon$, $A^o = \left\{ j | \beta^o_j \neq \beta^o_{j+1} \right\}$ *and* $\tilde{\boldsymbol{H}}$ *is a matrix that depends only on p.* $c_{min}(\boldsymbol{M})$ *denotes the smallest eigenvalue of matrix* $\boldsymbol{M}$*. let*

$$\hat{A} \equiv \left\{ j | \hat{\beta}_j \neq \hat{\beta}_{j+1} \right\},$$
$$\tilde{\boldsymbol{H}}_{\mathcal{O}} \equiv (\tilde{\boldsymbol{H}}_j, j \in A^o)_{p\times|A^o|},$$
$$(\omega^o_j, j \in A^o) \equiv \textit{the diagnal elements of } (\tilde{\boldsymbol{H}}^T_{\mathcal{O}} \tilde{\boldsymbol{H}}_{\mathcal{O}})^{-1}.$$

...

Theorem

If

$$\sup_{\|u\|_2=1} P(u^T\varepsilon > \sigma t) \le e^{-t^2/2} \qquad \forall t > 0, \tag{1}$$

$$\gamma > \frac{1}{c_{min}(\tilde{\boldsymbol{H}}^T \tilde{\boldsymbol{H}})}, \tag{2}$$

$$P(\lambda_l \le \lambda \le \lambda_u) = 1 \text{ and } \lambda_u\gamma \le \eta_\star = \min_{j\in A^o}\left\{\left|\beta_j^o - \beta_{j+1}^o\right|\right\}, \tag{3}$$

then

$$P(\hat{A} \ne A^o) \le P(\hat{\beta} \ne \beta^{\mathcal{O}}) \le \pi_1(\lambda_l) + \pi_2(\lambda_u), \tag{4}$$

where $\pi_1(\lambda) = 2\sum_{j\notin A^o} \exp\left\{-\frac{\lambda^2}{2\sigma^2\|\tilde{\boldsymbol{H}}_j\|_2^2}\right\}$,

$$\pi_2(\lambda) = \sum_{j\in A^o} \exp\left\{-\frac{(\lambda\gamma - \left|\beta_j^o - \beta_{j+1}^o\right|)^2}{2\omega_j^o\sigma^2}\right\}.$$

Outline

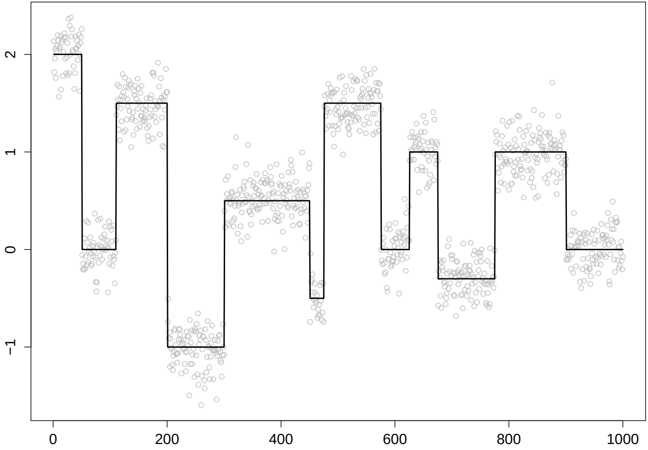
Simulations

We compare the estimation and selection accuracy of fusion MCP and FLSA with $\lambda_1 = 0$. Two types of shapes of the signals are considered.

For each case, we repeat fusion MCP and FLSA for 500 times.

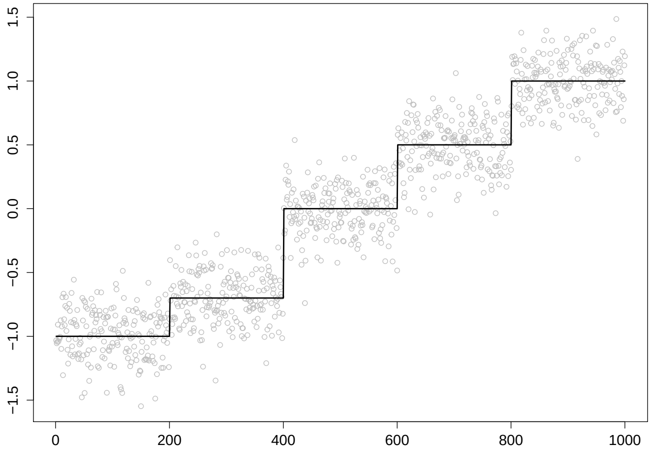
Two Cases

- Case 1



Two Cases

- Case 2



Estimation and Selection Accuracy

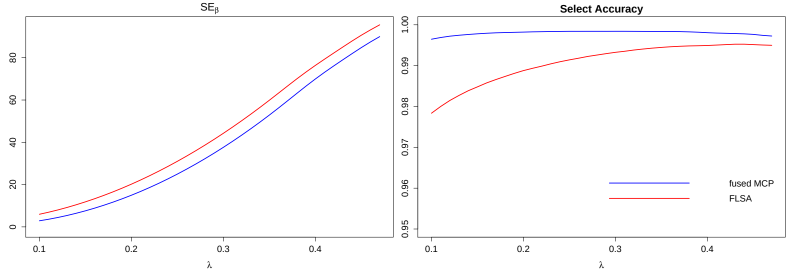


Figure: Case 1. $SE_\beta = \|\hat{\beta} - \beta^o\|^2$.

Estimation and Selection Accuracy

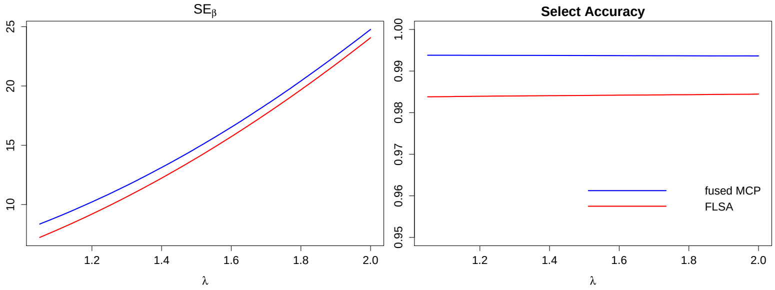


Figure: Case 2

Outline

Summary

- Fused Mcp dose not penalize large jumps
- It has a better performance in jump point(edge) detection
- It keeps the jump size(contrast) of the signal
- In signal processing scenario, both 1-d and 2-d fused Mcp problem can be solved efficiently

References

- Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1(2)**, 302-332.
- Hunter, D. R., and Li, R. (2005). Variable selection using MM algorithms. *Annals of statistics*, **33(4)**, 1617.
- Rudin, L. I., Osher, S. and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, **60(1)**, 259-268.
- Tibshirani, R. (1996). Regression shringkage and selection via the lasso, *J. Royal. Statist. Soc. B.* **58**, 267-288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), **67(1)**, 91-108.
- Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, **39(3)**, 1335-1371.
- Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 894-942.