In Pursuit of Ideal Model Selection for high-Dimensional Linear Regression

PhD thesis by Arash Owrang

Opposition by Prof. Esa Ollila
Department of Signal Processing and Acoustics
Aalto University, Finland

March 8th, 2018, KTH

Thesis outline

- Model selection is a fundamental problem in data analysis as it determines the success/accuracy of what we can learn from data.
- Thesis considers the case that the dimension of the parameters space, N, is much larger than the number of measurements, m.
- $N \gg m$ is a regime opposite to conventional (asymptotic) statistical settings.
- In the linear model, this implies that # of regressors exceeds the # of observations.

Contributions

1. An extended Fisher Information Criterion (EFIC) is proposed to improve model selection in high-dim. linear model
2. COM-Lasso estimator is developed for model selection when multiple measurement vectors are available.
3. Normalized Fused Lasso (NFL) is proposed for change point detection.

Model selection in high-dimensional linear model

- Measurement $\mathbf{y} \in \mathbb{R}^m$
- Regressor matrix $\mathbf{A} = (\mathbf{a}_1 \ \cdots \ \mathbf{a}_N) \in \mathbb{R}^{m \times N}$
- An index set $\mathcal{I} = \{i_1, \ldots, i_k\}$, $1 \leq i_1 < i_2 < \cdots < i_k \leq N$
- Set of indices $\mathcal{J} = \bigcup_{k=1}^{K} \{\mathcal{I} \,|\, |\mathcal{I}| = k\}$ up to cardinality $K \ll m$.
- High-dimensionality: $N = m^d, \quad d > 1.$

Model selection problem: Consider a set of competing hypothesis

$$\mathcal{H}_{\mathcal{I}} : \mathbf{y} = \mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}} + \sigma \boldsymbol{\epsilon}, \qquad \{\epsilon_i\}_{i=1}^{m} \overset{iid}{\sim} \mathcal{N}(0,1).$$

$$\text{or } \mathcal{H}_{\mathcal{I}} : \mathbf{y} \sim \mathcal{N}_m(\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}, \sigma^2 \mathbf{I})$$

where $\sigma >$ (error scale) and $\mathbf{x}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ (signal vector) are unknown. The task is to identify $\mathcal{S} \in \mathcal{J}$ or $\mathcal{H}_{\mathcal{S}}$, under the assumption that $\mathbf{y} \sim \mathcal{H}_{\mathcal{S}}$ for some $\mathcal{S} \in \mathcal{J}$.

Model selection via information criterions

- Parameter vector $\underline{\boldsymbol{\theta}}_{\mathcal{I}} = (\mathbf{x}_{\mathcal{I}}, \sigma) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}_+$.
- Under $\mathcal{H}_{\mathcal{I}}$: $\mathbf{y} \sim \mathcal{N}_m(\mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}, \sigma^2\mathbf{I})$, the MLE-s are

$$\hat{\sigma}^2 = \frac{1}{m}\|\Pi_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2, \quad \hat{\mathbf{x}}_{\mathcal{I}} = \mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{x}_{\mathcal{I}}$$

  where $\Pi_{\mathcal{I}} = \mathbf{A}_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}^{\dagger}$ and $\Pi_{\mathcal{I}}^{\perp} = \mathbf{I} - \Pi_{\mathcal{I}}$ denote the orthogonal projector
- The $-2\times$ log-likelihood function of $\mathbf{y} \sim \mathcal{H}_{\mathcal{I}}$:

$$-2\ln p(\mathbf{y}; \underline{\hat{\boldsymbol{\theta}}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) = m\ln\|\Pi_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 + \text{const.}$$

- General form of information criterions:

$$\begin{aligned}\hat{\mathcal{I}} &= \arg\min_{\mathcal{I}\in\mathcal{J}} \{-2\ln p(\mathbf{y}; \underline{\hat{\boldsymbol{\theta}}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) + \eta(\mathcal{I})\} \\ &= \arg\min_{\mathcal{I}\in\mathcal{J}} \{m\ln\|\Pi_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 + \eta(\mathcal{I})\}\end{aligned}$$

  where penalty term $\eta(\mathcal{I})$ penalizes for overfitting ($\eta(\mathcal{I}) \uparrow$ as $|\mathcal{I}| \uparrow$).

Table: The choice of penalty term for a few model selection criteria

Akaike IC	AIC	$\eta(\mathcal{I}) = 2(\lvert\mathcal{I}\rvert + 1)$
Bayesian IC	BIC	$\eta(\mathcal{I}) = (\lvert\mathcal{I}\rvert + 1)\ln m$
Risk IC	RIC	$\eta(\mathcal{I}) = (\lvert\mathcal{I}\rvert + 1)\ln N$
Fisher IC	FIC	$\eta(\mathcal{I}) = \ln\det \mathbf{F}(\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{I}})$

- $\mathbf{F}(\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{I}})$ is the Fisher information matrix evaluated at the MLE $\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{I}}$.
- In Chapter 4 it is shown that

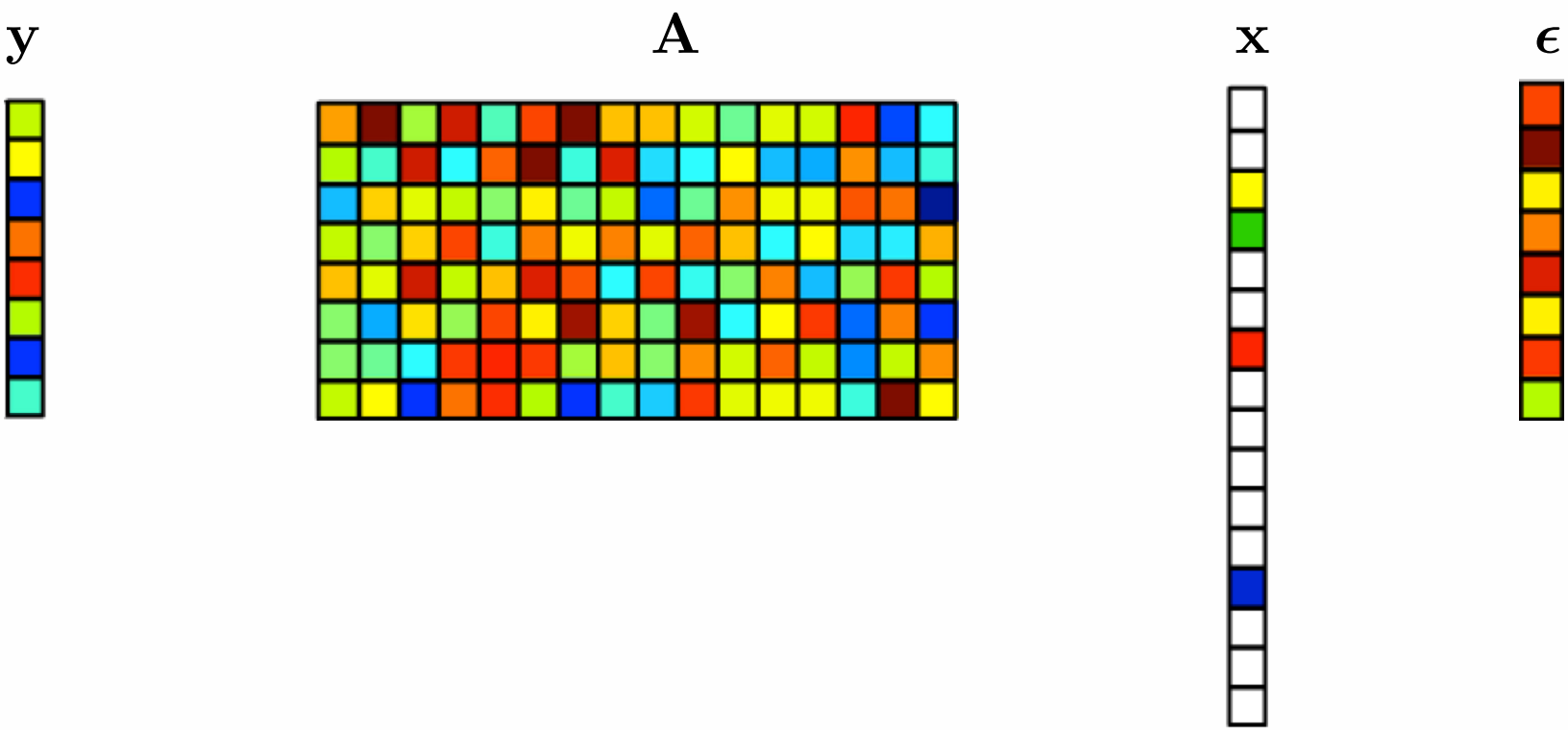
$$\ln\det \mathbf{F}(\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{I}}) = c + \ln\det(\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}) + (\lvert\mathcal{I}\rvert + 2)\{\ln \|\Pi_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 - \ln m\}$$

- BIC and FIC consistent in selecting the true model as $m \to \infty$.
- BIC based on approximation: $\det \mathbf{F}(\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{I}}) \approx m^{|\mathcal{I}|+1}$ for large m.

Model selection via sparse linear regression

$$\underset{m \times 1}{\mathbf{y}} = \underset{m \times N}{\mathbf{A}} \; \underset{N \times 1}{\mathbf{x}} + \underset{m \times 1}{\boldsymbol{\epsilon}}$$

- Known regressor matrix $\mathbf{A}$, unknown *sparse signal* $\mathbf{x}$, noise $\boldsymbol{\epsilon}$
- $\mathcal{S} = \mathrm{supp}(\mathbf{x}) \leq K \ll m$ and $m < N$



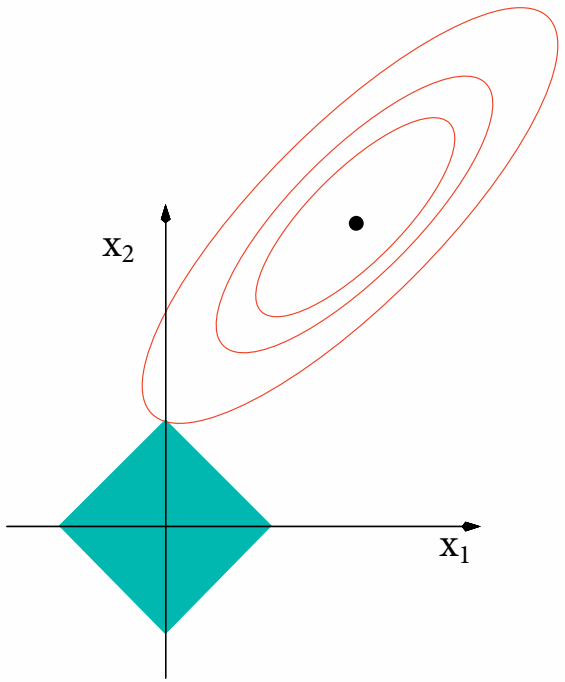
Lasso [Tibshirani, 1996]

- Lasso estimator solves

$$\hat{\mathbf{x}}(\lambda) = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

$$\hat{\mathbf{x}}(t) = \arg\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq t$$

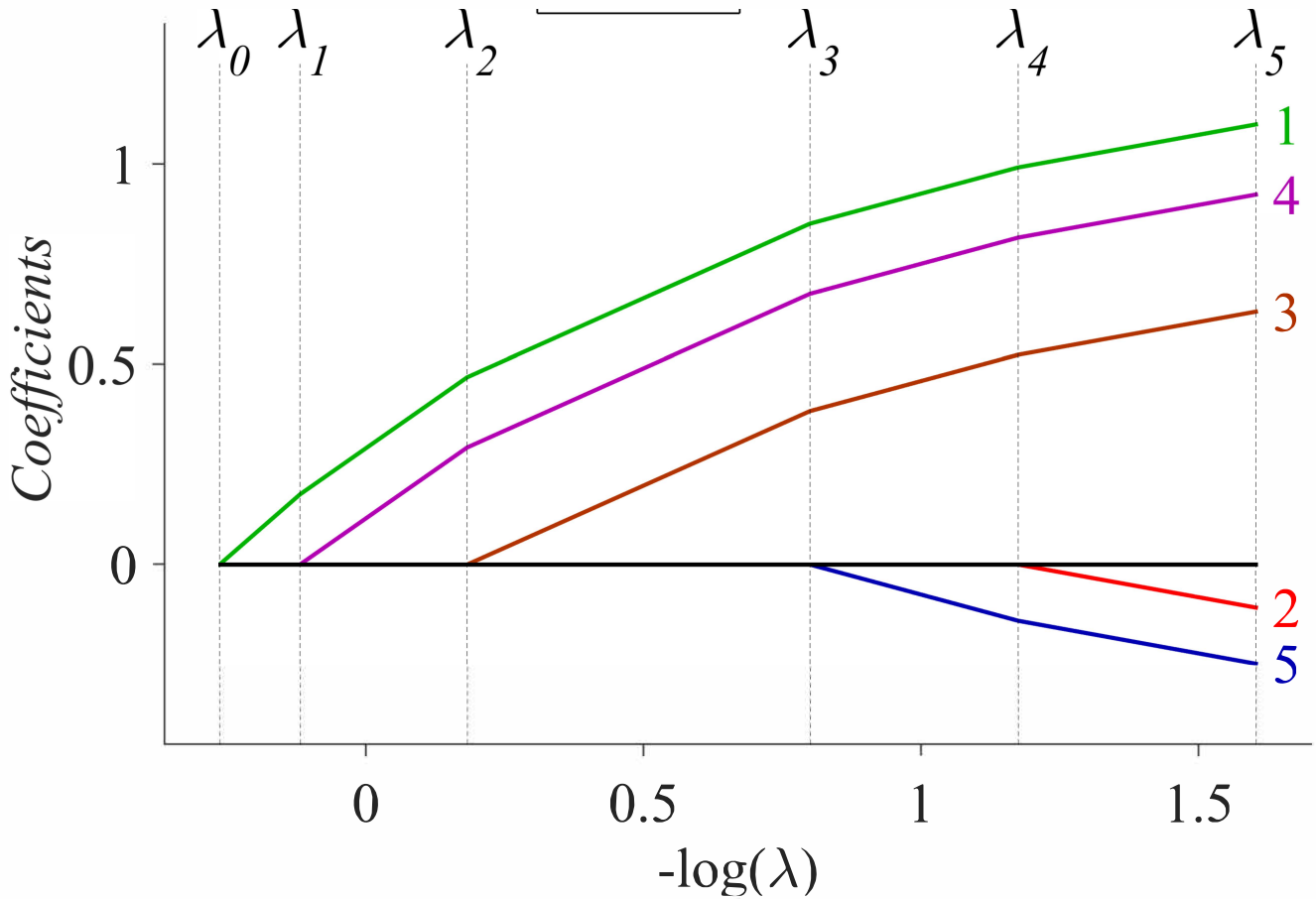
  where $\lambda > 0$ is the penalty parameter (1-to-1 with t).



- λ controls trade-off between the two terms (data fidelity vs sparsity).
  - Lasso provides a modern alternative to model selection: it performs model selection and parameter estimation simultaneously.
  - How many variables Lasso picks (how sparse is $\hat{\mathbf{x}}(\lambda)$) depends on λ.

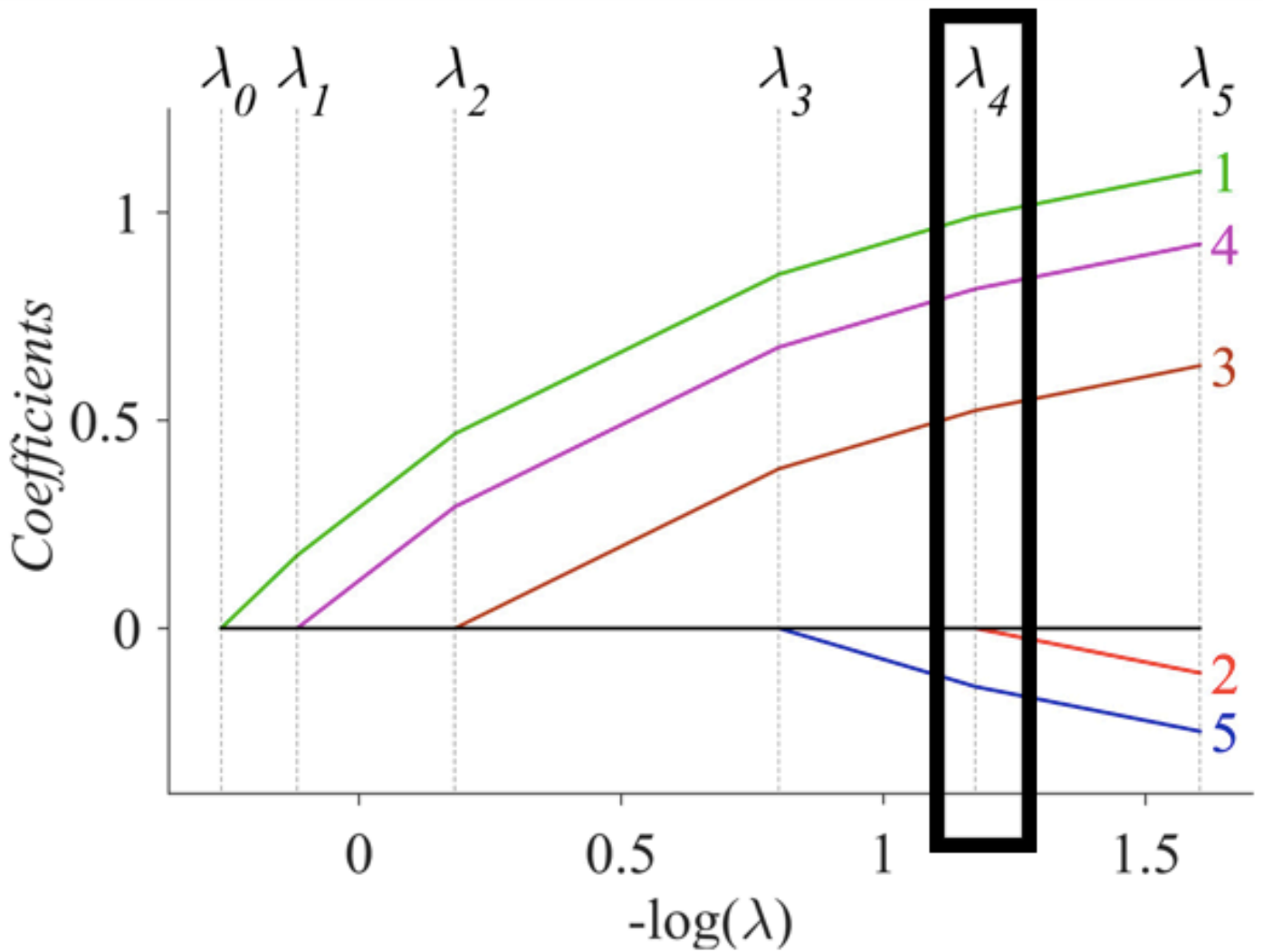
Least Angle Regression Algorithm (LARS)

- LARS [Efron et al., 2004] finds the pivotal penalty parameter values λ_k, where a new variable enters/leaves the active set.
- $\hat{\mathbf{x}}(\lambda)$ as a fnc λ is *piece-wise linear* in each coefficient.



Least Angle Regression Algorithm (LARS)

- LARS [Efron et al., 2004] finds the pivotal penalty parameter values λ_k, where a new variable enters/leaves the active set.
- EFIC can be used to choose the Lasso estimator on the solution path



Extended BIC

- In BIC expression, one assumed uniform prior for $\underline{\boldsymbol{\theta}}_{\mathcal{I}}$.
- In $N \gg m$ case, it is sensible to assign a larger prior for sparse models

$$p(\underline{\boldsymbol{\theta}}_{\mathcal{I}}) = \Pr(|\mathcal{I}| = k) \propto \binom{N}{k}^{-c}$$

  where $c > 0$ is a positive tuning constant.
- This gives *extended BIC (EBIC)* [Chen and Chen, 2008] criterion:

$$\mathrm{EBIC}(\mathcal{I}) = \mathrm{BIC}(\mathcal{I}) + 2c \ln \binom{N}{|\mathcal{I}|}$$

- Pitfalls of EBIC in high-dimensions or high-SNR:
  1. poor approximation of $\det \mathbf{F}(\underline{\hat{\boldsymbol{\theta}}}_{\mathcal{I}})$ by $m^{|\mathcal{I}|+1}$
  2. too conservative choise for tuning constant c $(> 1 - 1/(2d))$.

Extended FIC (EFIC)

- Recalling $N = m^d$, the authors use the approximation:

$$\ln \binom{N}{|\mathcal{I}|} \approx d|\mathcal{I}| \ln m$$

- This and previous eq. for $\ln \det \mathbf{F}(\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{I}})$ yields the proposed EFIC:

$$\begin{aligned}\text{EFIC}(\mathcal{I}) =& (m - |\mathcal{I}| - 2) \ln \|\Pi_{\mathcal{I}}^{\perp} \mathbf{y}\|_2^2 \\ &+ \ln \det(\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}) + (1 + 2cd)|\mathcal{I}| \ln m\end{aligned}$$

- With little manipulation, one may write it as

$$\text{EFIC}(\mathcal{I}) = \text{BIC}(\mathcal{I}) + 2c \cdot \gamma_{\text{RIC}}(\mathcal{I}) + \ln \det(\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}) - (|\mathcal{I}| + 2) \ln \|\Pi_{\mathcal{I}}^{\perp} \mathbf{y}\|_2^2$$

- My interpretation:
  - $-(|\mathcal{I}| + 2) \ln \|\Pi_{\mathcal{I}}^{\perp} \mathbf{y}\|_2^2$ corrects for the bias in $\hat{\sigma}^2 = \frac{1}{m} \|\Pi_{\mathcal{I}}^{\perp} \mathbf{y}\|_2^2$. Namely, for large $|\mathcal{I}|$, $\hat{\sigma}^2 \to 0$ as $|\mathcal{I}|/m \to 1$.
  - $c > 0$ is the degree of belief in RIC penalty.

EFIC vs BIC

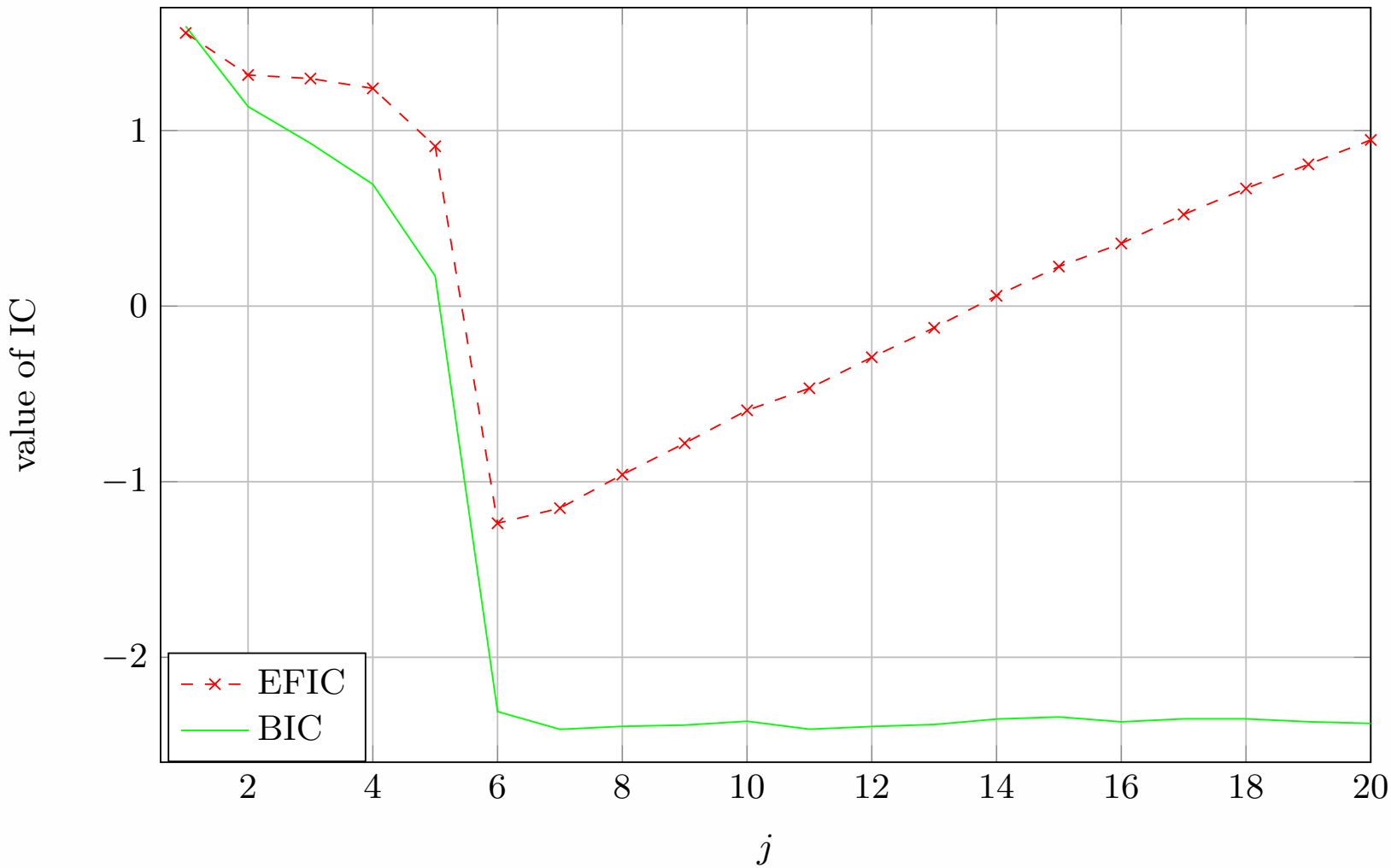


Figure 3.4: The comparison between the behavior of extended FIC and BIC versus the selection of indices of the models provided by the solution path of Lasso. The setting is $\sigma^2 = 10^{-1}$, $|\mathcal{S}| = 5$, $m = 100$ and $N = \lceil m^d \rceil$, for $d = 1.3$. Label six corresponds to the true model.

EFIC vs EBIC

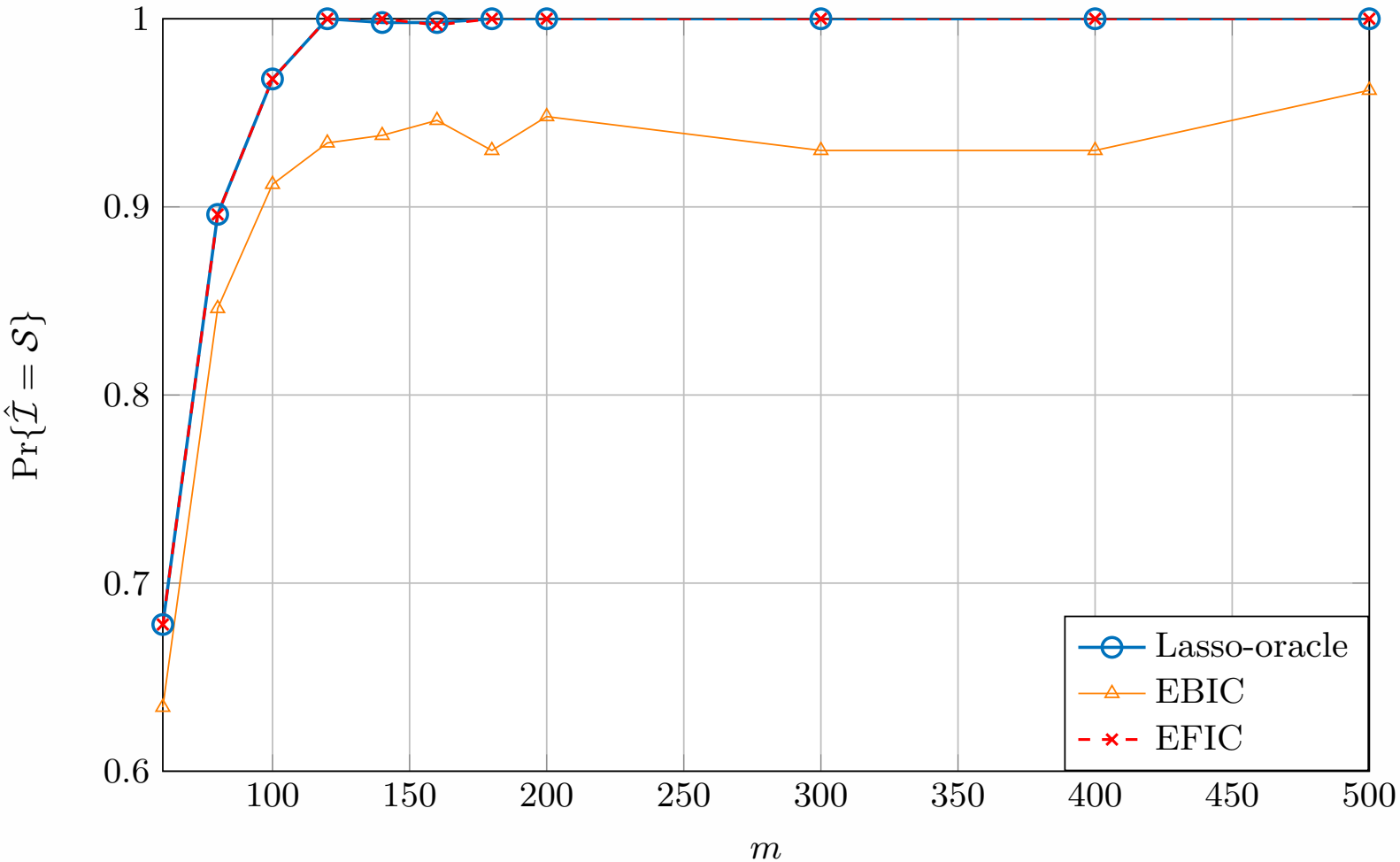


Figure 3.5: The empirical probability of $\{\hat{\mathcal{I}} = \mathcal{S}\}$ versus m when $\mathbf{A}$ has an uncorrelated structure. Here, $\mu = 0$, $\sigma^2 = 10^{-0.3}$, $|\mathcal{S}| = 5$ and $N = \lceil m^d \rceil$ for $d = 1.3$.

Computation of EFIC

- It is not computationally feasible to go through set $\{\mathcal{H}_{\mathcal{I}} : \mathcal{I} \in \mathcal{J}\}$ of competing hypothesis ($|\mathcal{J}| = O(N^K)$).
- Instead authors consider only K hypothesis with index sets

$$\mathcal{I}_1 \subset \mathcal{I}_2 \subset \cdots \subset \mathcal{I}_K.$$

The index sets are found from Lasso path at pivotal values $\lambda_1 > \lambda_2 > \cdots > \lambda_K$ (computed by LARS algorithm):

$$\mathcal{I}_k = \text{supp}(\hat{\mathbf{x}}(\lambda_k)), \quad k = 1, \ldots, K$$

where $\lambda_1 = \|\mathbf{A}^T\mathbf{y}\|_\infty \Rightarrow \hat{\mathbf{x}}(\lambda_1) = \mathbf{0}$ and $\mathcal{I}_1 = \{\emptyset\}$ and generally:

$$|\mathcal{I}_1| = 0, |\mathcal{I}_2| = 1, \ldots, |\mathcal{I}_K| = K - 1$$

(given no predictor leaves the active set in $\lambda \in (\lambda_0, \lambda_K]$.

Small correlations between predictors

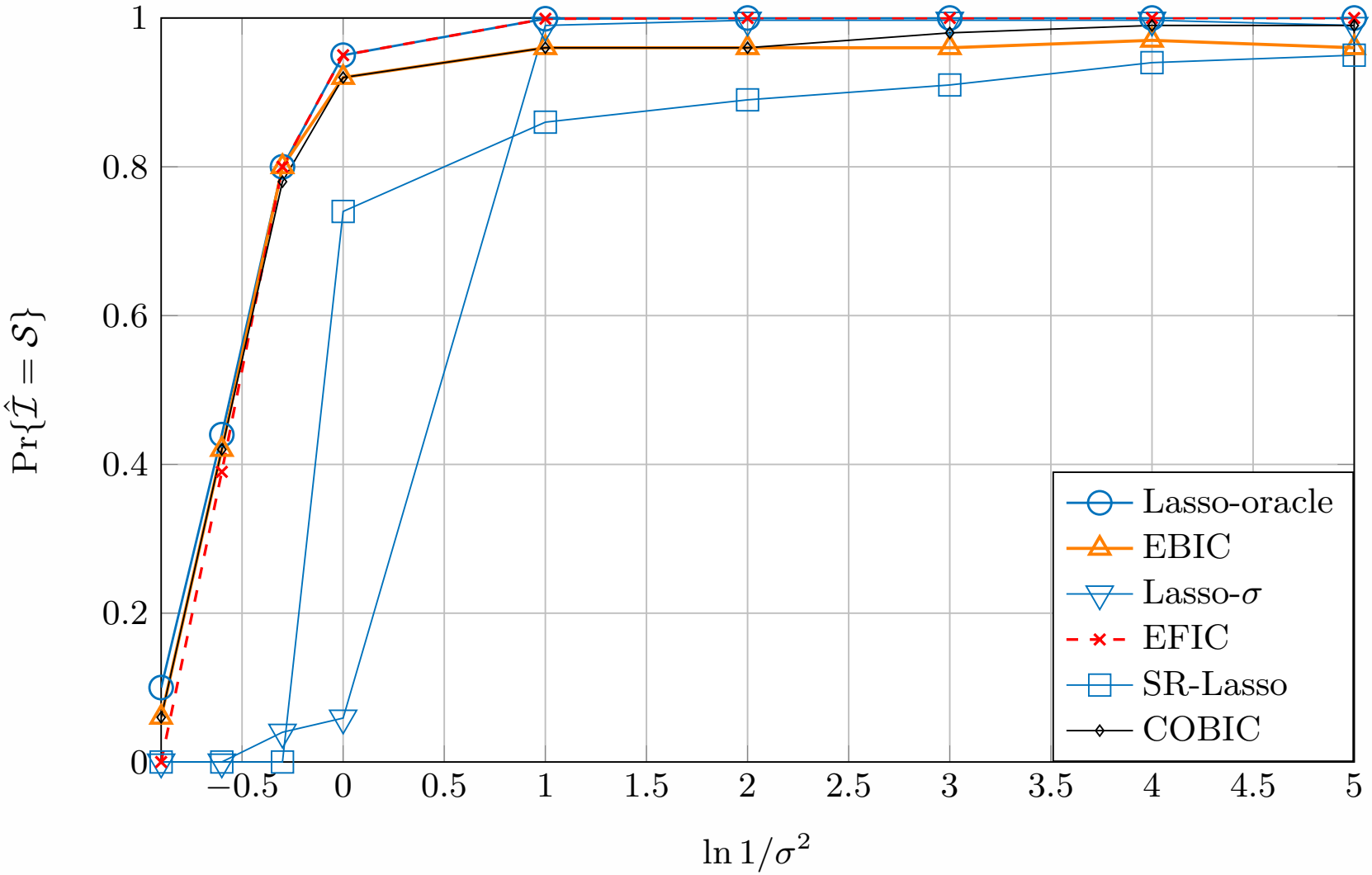


Figure 4.6: The empirical probability of $\{\hat{\mathcal{I}} = \mathcal{S}\}$ versus $\ln(1/\sigma^2)$ when $\mathbf{A}$ has a correlated structure. Here, $\mu = 0.25$, $m = 200$, $|\mathcal{S}| = 5$ and $N = \lceil m^d \rceil$ for $d = 1.3$.

Theoretical contributions

The authors consider the cases:

1. $\sigma \to 0$
2. $m \to \infty$

Restricted eigenvalue property: The normalized matrix $\tilde{\mathbf{A}}$ satisfies the restricted eigenvalue property if any restricted sub-matrix $\tilde{\mathbf{A}}_{\mathcal{I}}^T \tilde{\mathbf{A}}_{\mathcal{I}}$ obeys

$$\min_{|\mathcal{I}| \leq 2K} \Lambda_{\min}(\tilde{\mathbf{A}}_{\mathcal{I}}^T \tilde{\mathbf{A}}_{\mathcal{I}}) \geq \frac{C_{\min}}{\ln m},$$

for some constant $C_{\min} > 0$. Here, $\Lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of the corresponding matrix.

Theorem 4.2.1. *Let m be the fixed number of measurements and assume that $N = m^d$. Then, under the restricted eigenvalue property, the estimate of (4.9), $\hat{\mathcal{I}}$, obeys $\hat{\mathcal{I}} = \mathcal{S}$ with a probability approaching one as $\sigma \to 0$.*

Theorem 4.2.2. *Suppose that the matrix* $\mathbf{A} \in \mathbb{R}^{m \times N}$, *with* $N = m^d$, *satisfies the restricted eigenvalue property. Moreover, assume that the columns of* $\mathbf{A}$ *fulfill*

$$\|\mathbf{a}_i\|_2^2 = \Omega(m^a) \tag{4.15}$$

for some constant $a > 0$. *Then, the EFIC's estimate obeys* $\hat{\mathcal{I}} = \mathcal{S}$ *with probability one as* $m \to \infty$, *if* c *is chosen such that*

$$c > 1 - \frac{a}{2d} + \frac{1}{d}.$$

- The authors propose to use

$$c = 1 - \frac{a}{2d} + \frac{2}{d}$$

  where $d = \ln N / \ln m > 1$ as $N = m^d$.
- The parameter a computed in practise using norms of $\|\mathbf{a}_i\|$?

Model

L complex-valued measurement vectors:

$$\mathbf{y}(t) = \mathbf{A}\mathbf{x}(t) + \boldsymbol{\epsilon}(t), \quad t = 1, \ldots, L$$

Assumptions:

1. each $\mathbf{x}(t) \in \mathbb{C}^m$ is K-sparse with *common* support $\mathcal{S} = \mathrm{supp}(\mathbf{x}(t))$, $t = 1, \ldots, L$.
2. signal $\mathbf{x}_{\mathcal{S}}(t)$ is random, with $[\mathbf{x}_{\mathcal{S}}(t)]_j \overset{iid}{\sim} \mathcal{N}(0, p_{j,j})$, $j \in \mathcal{S}$.
3. noise $\boldsymbol{\epsilon}(t)$ is random, with $\boldsymbol{\epsilon}(t) \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma})$.
4. unknown noise covariance matrix $\boldsymbol{\Sigma}$ can linearly parametrized such that

   $$\mathrm{vec}(\boldsymbol{\Sigma}) = \mathbf{Q}\mathbf{h}$$

   for some known matrix $\mathbf{Q} \in \mathbb{C}^{m^2 \times \kappa}$ ($[\mathbf{Q}]_{i,j} \in \{0, 1, \pm j\}$) and $\mathbf{h} \in \mathbb{R}^{\kappa}$, where $\kappa \leq m^2 - |\mathcal{S}|$.

Under the Asssumptions 1-4, it holds that

$$\mathbf{y}(t) \sim \mathcal{N}_m(\mathbf{0}, \mathbf{R})$$

where

- $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \mathbf{\Sigma}$ (pos. def. $m \times m$ matrix)
- $\mathbf{P} = \mathrm{diag}(p_{1,1}, \ldots, p_{N,N})$ s.t. $p_{i,i} = 0$ for $i \in \mathcal{S}^c$.

  $\Rightarrow \mathbf{p} = \mathrm{vec}(\mathbf{P}) \in \mathbb{R}_+^{N^2}$ is K-sparse.

Consequently $\mathbf{r} = \mathrm{vec}(\mathbf{R})$ becomes

$$\mathbf{r} = (\overline{\mathbf{A}} \otimes \mathbf{A})\mathbf{p} + \mathbf{Q}\mathbf{h}$$

COM-Lasso idea:

- use covariance matching (COMET) [Ottersten et al., 1998] principle to estimate $\mathbf{p}$ and $\mathbf{h}$.
- Utilize the fact that $\mathbf{p}$ is K-sparse and non-negative (non-neg. Lasso).

COM-Lasso method

- Map $\mathbf{r}$ in $\mathbb{C}^{m^2}$ to $\mathbf{f}$ in $\mathbb{R}^{m^2}$ (Hermitian symmetry reduces the unknowns):

$$\mathbf{f} = \mathbf{T}\mathbf{r} = \mathbf{T}\{(\overline{\mathbf{A}} \otimes \mathbf{A})\mathbf{p} + \mathbf{Q}\mathbf{h}\}$$

  where $\mathbf{f}$ contains the m^2 real-valued unknowns of $\mathbf{r}$

- Estimate is $\hat{\mathbf{f}} = \mathbf{T}\hat{\mathbf{r}}$, where

$$\hat{\mathbf{r}} = \mathrm{vec}(\hat{\mathbf{R}}), \quad \hat{\mathbf{R}} = \frac{1}{L}\sum_{t=1}^{L} \mathbf{y}(t)\mathbf{y}(t)^H$$

- Since $\hat{\mathbf{R}}$ is a Wishart matrix, one has that

$$\mathrm{cov}(\mathrm{vec}(\hat{\mathbf{R}})) = \frac{1}{L}(\mathbf{R}^\top \otimes \mathbf{R})$$

$$\mathbf{\Gamma} = L \cdot \mathrm{cov}(\hat{\mathbf{f}}) = \mathbf{T}(\mathbf{R}^\top \otimes \mathbf{R})\mathbf{T}^H$$

  (and estimate $\hat{\mathbf{\Gamma}} = \mathbf{T}(\hat{\mathbf{R}}^\top \otimes \hat{\mathbf{R}})\mathbf{T}^H$)

COM-Lasso method

- Map $\mathbf{r}$ in $\mathbb{C}^{m^2}$ to $\mathbf{f}$ in $\mathbb{R}^{m^2}$ (Hermitian symmetry reduces the unknowns):

$$\mathbf{f} = \mathbf{T}\mathbf{r} = \mathbf{T}\{(\overline{\mathbf{A}} \otimes \mathbf{A})\mathbf{p} + \mathbf{Q}\mathbf{h}\}$$

  where $\mathbf{f}$ contains the m^2 real-valued unknowns of $\mathbf{r}$

- Estimate is $\hat{\mathbf{f}} = \mathbf{T}\hat{\mathbf{r}}$, where

$$\hat{\mathbf{r}} = \mathrm{vec}(\hat{\mathbf{R}}), \quad \hat{\mathbf{R}} = \frac{1}{L}\sum_{t=1}^{L} \mathbf{y}(t)\mathbf{y}(t)^H$$

- Since $\hat{\mathbf{R}}$ is a Wishart matrix, one has that

$$\mathrm{cov}(\mathrm{vec}(\hat{\mathbf{R}})) = \frac{1}{L}(\mathbf{R}^\top \otimes \mathbf{R})$$
$$\mathbf{\Gamma} = L \cdot \mathrm{cov}(\hat{\mathbf{f}}) = \mathbf{T}(\mathbf{R}^\top \otimes \mathbf{R})\mathbf{T}^H$$

  (and estimate $\hat{\mathbf{\Gamma}} = \mathbf{T}(\hat{\mathbf{R}}^\top \otimes \hat{\mathbf{R}})\mathbf{T}^H$)

- The COMET principle finds $\mathbf{p}$ and $\mathbf{h}$ as minimizers of

$$\begin{aligned}\eta(\mathbf{p},\mathbf{h}) &= (\mathbf{f}-\hat{\mathbf{f}})^\top \hat{\mathbf{\Gamma}}^{-1}(\mathbf{f}-\hat{\mathbf{f}}) \\ &= \left\|\hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}\big(\hat{\mathbf{r}}-(\overline{\mathbf{A}}\otimes\mathbf{A})\mathbf{p}-\mathbf{Q}\mathbf{h}\big)\right\|_2^2\end{aligned}$$

- Minimizing $\eta(\mathbf{p},\mathbf{h})$ for fixed $\mathbf{p}$ yields the (conditional) minimizer

$$\hat{\mathbf{h}} = \hat{\mathbf{h}}(\mathbf{p}) = (\hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}\mathbf{Q})^\dagger \hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}(\hat{\mathbf{r}}-(\overline{\mathbf{A}}\otimes\mathbf{A})\mathbf{p})$$

- Then authors then solve $\hat{\mathbf{p}}$ as minimizer of

$$\eta_{\min}(\mathbf{p}) = \eta(\mathbf{p},\hat{\mathbf{h}}(\mathbf{p})) = \|\mathbf{z}-\mathbf{\Phi}\mathbf{p}\|_2^2$$

where $\mathbf{z}$ and $\mathbf{\Phi}$ are functions of $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{r}}$ (and known matrices $\mathbf{T}$, $\mathbf{A}$ and $\mathbf{Q}$) and $\mathbf{p}$ is K-sparse and non-negative.

$\Rightarrow$ find $\hat{\mathbf{p}}$ by non-neg. Lasso, where EFIC is derived for model selection.

\# of hypothesis is narrowed down by inspecting only pivotal values at non. neg. Lasso path using the modified LARS algorithm.

- The COMET principle finds $\mathbf{p}$ and $\mathbf{h}$ as minimizers of

$$\begin{aligned}\eta(\mathbf{p},\mathbf{h}) &= (\mathbf{f}-\hat{\mathbf{f}})^\top \hat{\mathbf{\Gamma}}^{-1}(\mathbf{f}-\hat{\mathbf{f}}) \\ &= \left\|\hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}\left(\hat{\mathbf{r}}-(\overline{\mathbf{A}}\otimes\mathbf{A})\mathbf{p}-\mathbf{Q}\mathbf{h}\right)\right\|_2^2\end{aligned}$$

- Minimizing $\eta(\mathbf{p},\mathbf{h})$ for fixed $\mathbf{p}$ yields the (conditional) minimizer

$$\hat{\mathbf{h}} = \hat{\mathbf{h}}(\mathbf{p}) = (\hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}\mathbf{Q})^\dagger \hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}(\hat{\mathbf{r}}-(\overline{\mathbf{A}}\otimes\mathbf{A})\mathbf{p})$$

- Then authors then solve $\hat{\mathbf{p}}$ as minimizer of

$$\eta_{\min}(\mathbf{p}) = \eta(\mathbf{p},\hat{\mathbf{h}}(\mathbf{p})) = \|\mathbf{z}-\mathbf{\Phi}\mathbf{p}\|_2^2$$

  where $\mathbf{z}$ and $\mathbf{\Phi}$ are functions of $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{r}}$ (and known matrices $\mathbf{T}$, $\mathbf{A}$ and $\mathbf{Q}$) and $\mathbf{p}$ is K-sparse and non-negative.

$\Rightarrow$ find $\hat{\mathbf{p}}$ by non-neg. Lasso, where EFIC is derived for model selection.

# of hypothesis is narrowed down by inspecting only pivotal values at non. neg. Lasso path using the modified LARS algorithm.

- The COMET principle finds $\mathbf{p}$ and $\mathbf{h}$ as minimizers of

$$\begin{aligned}\eta(\mathbf{p},\mathbf{h}) &= (\mathbf{f}-\hat{\mathbf{f}})^\top \hat{\mathbf{\Gamma}}^{-1}(\mathbf{f}-\hat{\mathbf{f}}) \\ &= \left\|\hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}\big(\hat{\mathbf{r}}-(\overline{\mathbf{A}}\otimes\mathbf{A})\mathbf{p}-\mathbf{Q}\mathbf{h}\big)\right\|_2^2\end{aligned}$$

- Minimizing $\eta(\mathbf{p},\mathbf{h})$ for fixed $\mathbf{p}$ yields the (conditional) minimizer

$$\hat{\mathbf{h}} = \hat{\mathbf{h}}(\mathbf{p}) = (\hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}\mathbf{Q})^\dagger \hat{\mathbf{\Gamma}}^{-1/2}\mathbf{T}(\hat{\mathbf{r}}-(\overline{\mathbf{A}}\otimes\mathbf{A})\mathbf{p})$$

- Then authors then solve $\hat{\mathbf{p}}$ as minimizer of

$$\eta_{\min}(\mathbf{p}) = \eta(\mathbf{p},\hat{\mathbf{h}}(\mathbf{p})) = \|\mathbf{z}-\mathbf{\Phi}\mathbf{p}\|_2^2$$

  where $\mathbf{z}$ and $\mathbf{\Phi}$ are functions of $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{r}}$ (and known matrices $\mathbf{T}$, $\mathbf{A}$ and $\mathbf{Q}$) and $\mathbf{p}$ is K-sparse and non-negative.

$\Rightarrow$ find $\hat{\mathbf{p}}$ by non-neg. Lasso, where EFIC is derived for model selection.

  # of hypothesis is narrowed down by inspecting only pivotal values at non. neg. Lasso path using the modified LARS algorithm.

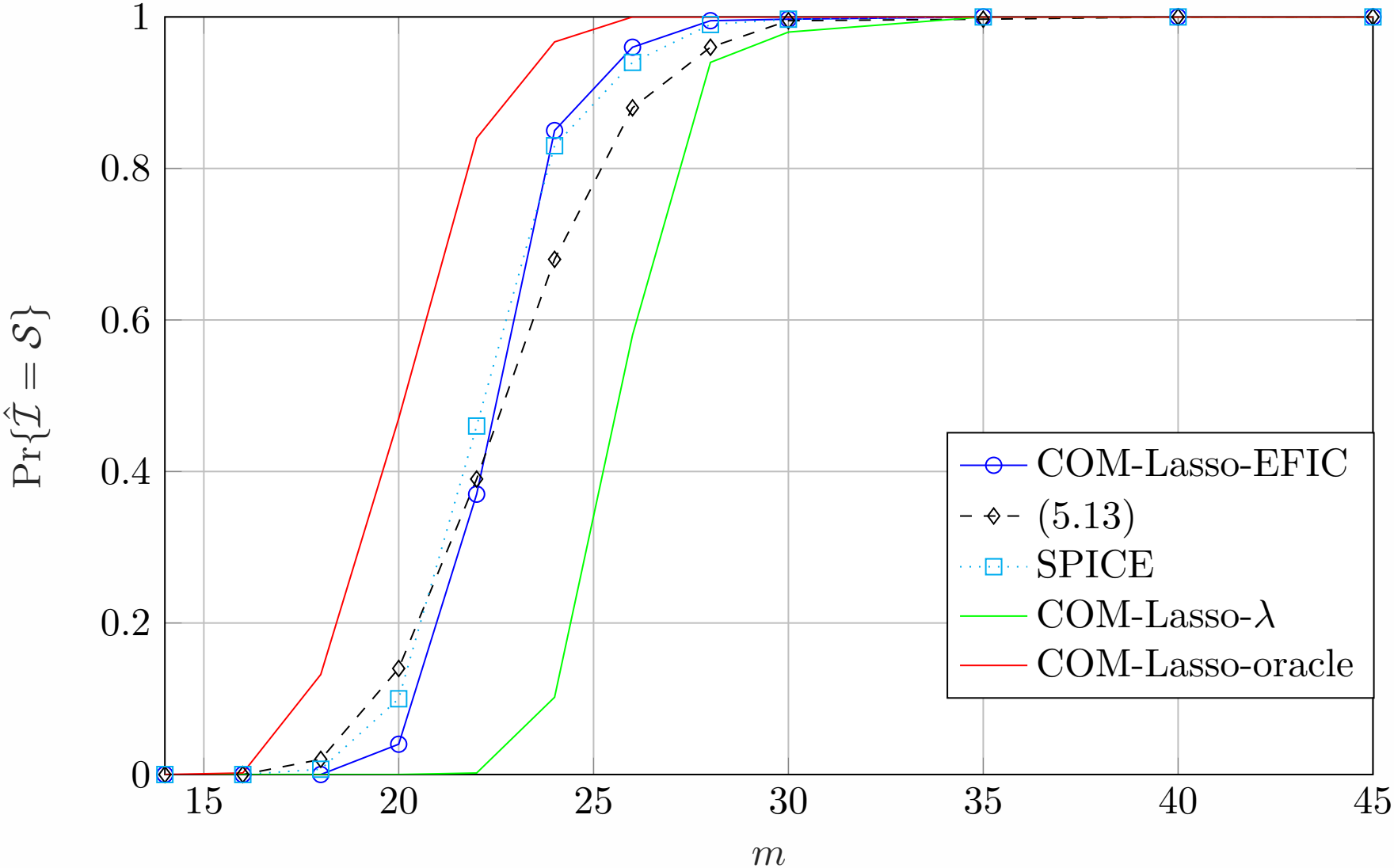


Figure 5.1: The empirical probability of $\{\hat{\mathcal{I}} = \mathcal{S}\}$ versus m when $\mathbf{A}$ has an uncorrelated structure, i.e. $\mu = 0$, and $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. Here, $|\mathcal{S}| = 20$, $\sigma^2 = 10$, $N = \lceil (m^2-1)^d \rceil$ for $d = 1.2$ and $L = 4m \ln m$.

When NFL  is not NFL

- If the desired signal is piecewise constant over neighboring values, then Fused Lasso [Tibshirani et al, 2005] can be used to encourage *smoothness* of the estimates.
- Noisy measurements $y(t)$ of the *piecewise constant* signal $m^*(t)$:

$$y(t) = m^*(t) + \sigma\epsilon(t)$$

  where $m^*(t)$ has change points at K locations $s_1 < s_2 < \ldots < s_K$ and the signal remains constant atleast for two consecutive samples.
- The authors show that FL is inconsistent in detecting the true change points.
- On the contrary, the proposed normalized fused Lasso (NFL) is consistent (when $\sigma \to 0$) in detecting change points.

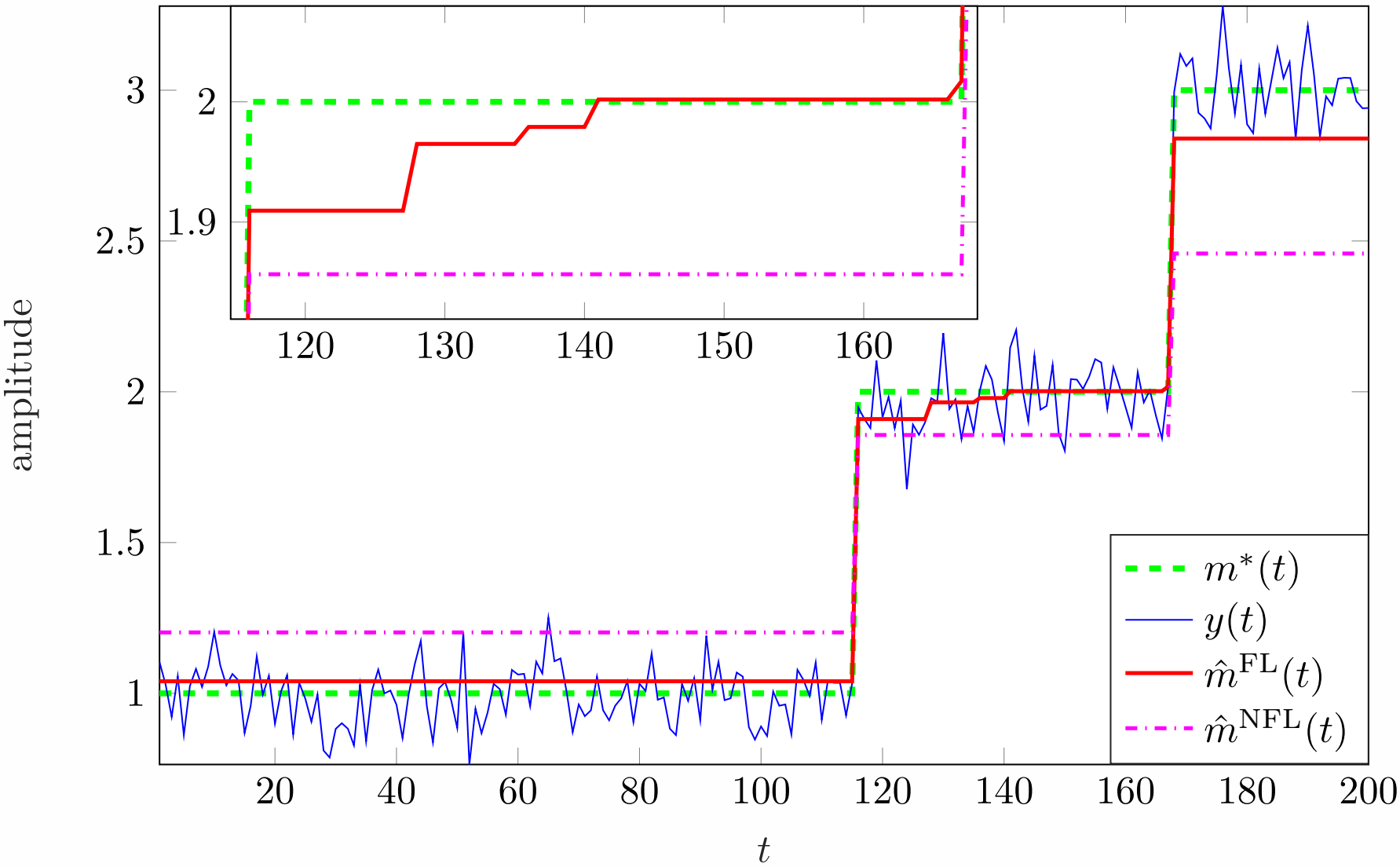


Figure 6.1: The solution of FL, $\hat{m}^{\mathrm{FL}}(t)$, is cluttered with small steps when $\sigma = 0.1$. The small box in the left top corner magnifies the intermediate level of $\hat{m}^{\mathrm{FL}}(t)$ and $\hat{m}^{\mathrm{NFL}}(t)$ ($y(t)$ is eliminated for the sake of visibility).

Fused Lasso

- FL solves

$$\hat{\mathbf{m}}^{\mathsf{FL}} = \arg\min_{\mathbf{m}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y}-\mathbf{m}\|_2^2 + \lambda \underbrace{\sum_{t=2}^{N} |m(t)-m(t-1)|}_{=\|\mathbf{D}\mathbf{m}\|_1}$$

  for some penalty parameter $\lambda > 0$.

- An alternative (Lasso-type) formulation of FL is [Rojas and Wahlberg, 2014]:

$$\hat{\mathbf{x}}^{\mathsf{FL}} = \arg\min_{\mathbf{x}\in\mathbb{R}^{N-1}} \left\{ \frac{1}{2}\|\tilde{\mathbf{y}}-\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \right\}$$

  where $\mathbf{A} \in \mathbb{R}^{N\times(N-1)}$ verifies $a_{i,j} = \frac{j}{N} - 1$, $i \leq j$,and $a_{i,j} = \frac{j}{N}$ otherwise, and $\tilde{\mathbf{y}}$ is mean centered version of $\mathbf{y}$.

- Solutions are related by $\hat{\mathbf{x}}^{\mathsf{FL}} = \mathbf{D}\hat{\mathbf{m}}^{\mathsf{FL}}$.

Normalized fused Lasso

- The proposed NFL solves

$$\hat{\mathbf{x}}^{\text{NFL}} = \underset{\mathbf{x}\in\mathbb{R}^{N-1}}{\arg\min}\left\{\frac{1}{2}\|\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1\right\}$$

where $\tilde{\mathbf{A}}$ is *normalized version* of $\mathbf{A}$ having unit norm columns.

Theorem 6.2.1. *Assume that for a particular realization of* $\tilde{\boldsymbol{\epsilon}}$ *there is a* $\lambda_p > 0$ *such that*

$$\|\sigma\tilde{\mathbf{A}}_{\mathcal{S}^c}^T\boldsymbol{\Pi}_{\mathcal{S}}^{\perp}\tilde{\boldsymbol{\epsilon}} + \lambda_p\tilde{\mathbf{A}}_{\mathcal{S}^c}^T\tilde{\mathbf{A}}_{\mathcal{S}}^{\dagger^T}\operatorname{sgn}(\tilde{\mathbf{x}}_{\mathcal{S}}^*)\|_\infty < \lambda_p, \tag{6.12}$$

$$\min_{i\in\mathcal{S}}|\tilde{x}_i^*| > \|\sigma\tilde{\mathbf{A}}_{\mathcal{S}}^{\dagger}\tilde{\boldsymbol{\epsilon}} - \lambda_p\left(\tilde{\mathbf{A}}_{\mathcal{S}}^T\tilde{\mathbf{A}}_{\mathcal{S}}\right)^{-1}\operatorname{sgn}(\tilde{\mathbf{x}}_{\mathcal{S}}^*)\|_\infty, \tag{6.13}$$

where the matrix $\tilde{\mathbf{A}}_{\mathcal{S}}^{\dagger} = (\tilde{\mathbf{A}}_{\mathcal{S}}^T\tilde{\mathbf{A}}_{\mathcal{S}})^{-1}\tilde{\mathbf{A}}_{\mathcal{S}}^T$ *is the Moore-Penrose pseudo-inverse of* $\tilde{\mathbf{A}}_{\mathcal{S}}$ *and* $\boldsymbol{\Pi}_{\mathcal{S}}^{\perp}$ *denotes the orthogonal projection matrix defined as* $\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} = \mathbf{I} - \tilde{\mathbf{A}}_{\mathcal{S}}\tilde{\mathbf{A}}_{\mathcal{S}}^{\dagger}$. *Then,* $\hat{\mathbf{x}}^{\text{NFL}}$, *obtained by solving (6.9) with* $\lambda = \lambda_p$, *satisfies* $\operatorname{supp}(\hat{\mathbf{x}}^{\text{NFL}}) = \mathcal{S}$ *and* $\operatorname{sgn}(\hat{\mathbf{x}}_{\mathcal{S}}^{\text{NFL}}) = \operatorname{sgn}(\tilde{\mathbf{x}}_{\mathcal{S}}^*)$.

Contributions

1. A. Owrang and M. Jansson, "A model selection criterion for high-dimensional linear regression," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3436-3446, Jul. 2018.
2. A. Owrang and M. Jansson "Weighted Covariance Matching Based Square Root Lasso," ICASSP'15
3. A. Owrang, M. Malek-Mohammadi, A. Proutiere, and M. Jansson, "Consistent Change Point Detection for Piecewise Constant Signals With Normalized Fused Lasso," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 799-803, Jun. 2017.
4. A. Owrang and M. Jansson, "Model Selection With Covariance Matching Based Non-negative Lasso," To be submitted.
5. M. Malek-Mohammadi, M. Jansson, A. Owrang, A. Koochakzadeh and M. Babaie-Zadeh, "DOA Estimation in Partially Correlated Noise Using Low-rank/Sparse Matrix Decomposition," SAM'14.

References

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Björn Ottersten, Peter Stoica, and Richard Roy. Covariance matching estimation techniques for array signal processing applications. *Digital Signal Processing*, 8(3): 185–210, 1998.

Cristian R Rojas and Bo Wahlberg. On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc., Ser. B*, 58:267–288, 1996.