ℓ_1 regularization in the high-dimensional setting: Thresholds for sparsity recovery and model selection

Martin Wainwright

Department of Electrical Engineering and Computer Science

Department of Statistics

UC Berkeley, California, USA

wainwrig@{stat,eecs}.berkeley.edu

Graph model selection based on joint work with:

John Lafferty	Carnegie Mellon University, USA
Pradeep Ravikumar	Carnegie Mellon University, USA

Introduction

- sparsity recovery: how to recover a "suitably sparse" signal β^* from noisy observations?

- broad range of applications:
  - subset selection in regression
  - signal denoising and constructive approximation
  - graphical model selection

- natural optimization-theoretic formulation via ℓ_0 "norm":

$$\|\beta^*\|_0 \quad := \quad \text{card}\,\{i \;\mid\; \beta^*_i \neq 0\}\,.$$

- ℓ_0 problems NP-hard in general $\Longrightarrow$ need for computationally tractable relaxations

Subset selection in regression

- consider the standard linear regression model

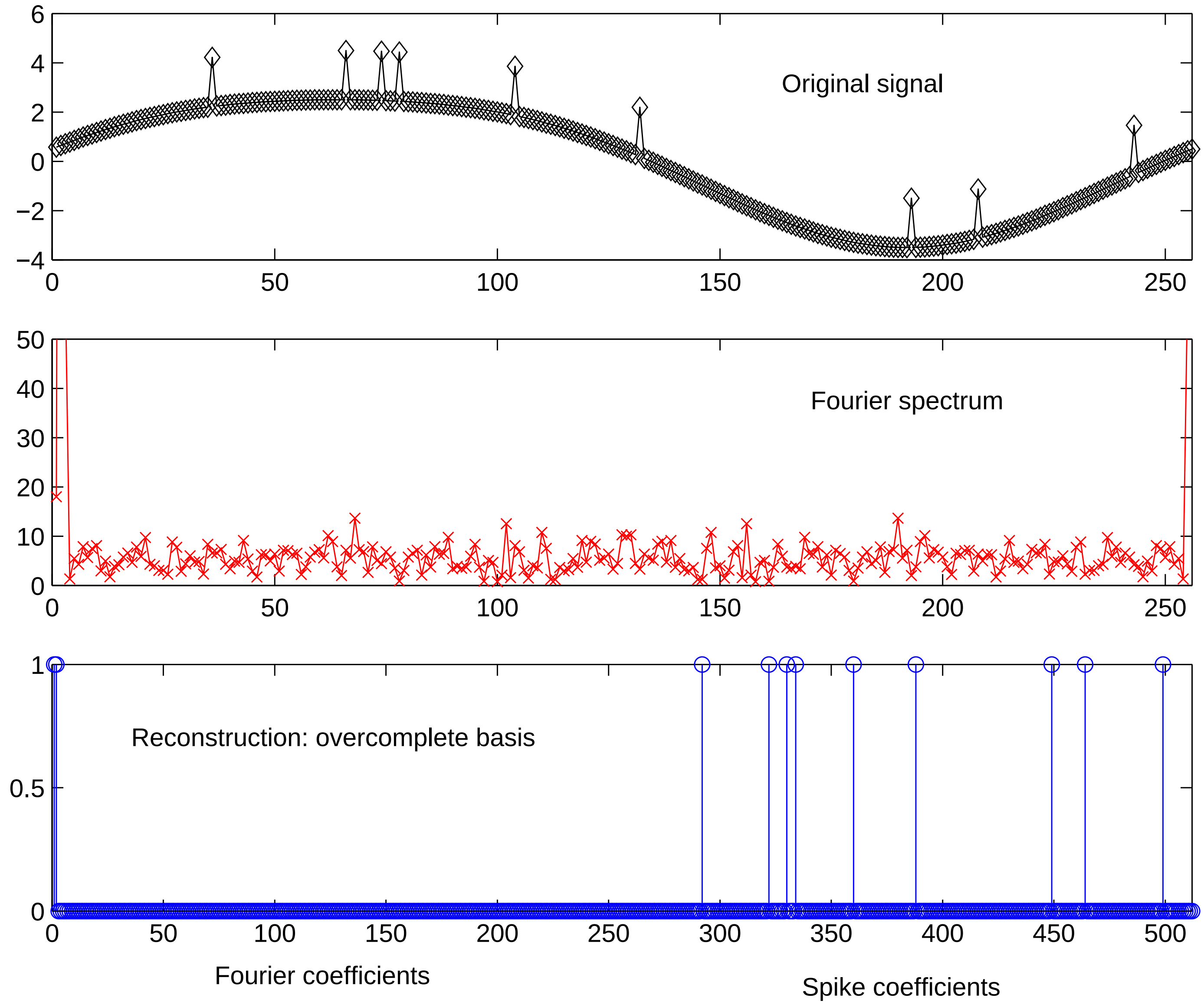
$$y_k = x_k^T \beta^* + w_k$$

where
- (x_k, y_k) are observed data
- observation noise $w_k \sim N(0, \sigma^2)$
- $\beta^* \in \mathbb{R}^p$ is the regression vector

- vector $x \in \mathbb{R}^p$ may include a large number of irrelevant variables (e.g., bioinformatics, sparse representations in signal processing)

- **subset selection:** how to choose the relevant subset S of indices for β^*?

Illustration: Reconstruction in overcomplete bases



Graphical model selection

- given samples $z^k = \begin{bmatrix} z_1^k & z_2^k & \cdots & z_p^k \end{bmatrix}$ of an m-dimensional random vector

- say that we want to fit a Markov random field to this data

- there are $p = \binom{m}{2}$ possible edges to include/exclude

- **graphical model selection:** how to choose the appropriate subset S of edges to include?

- classical model selection criteria (AIC, BIC): typically involve some form of ℓ_0 "norm" penalty

Sparsity recovery with ℓ_1 relaxations

Noiseless setting: *Linear programming* (Chen et al., 1998)

Given perfect observations $y_k = x_k^T \beta^*$ for $k = 1, \ldots, n$.

ℓ_0 problem (L_0)	ℓ_1 relaxation (L_1)
$\min_{\beta \in \mathbb{R}^p} \; \|\beta\|_0$ s.t. $x_k^T \beta = y_k, \quad k = 1, \ldots, n$	$\min_{\beta \in \mathbb{R}^p} \; \|\beta\|_1$ s.t. $x_k^T \beta = y_k, \quad k = 1, \ldots, n$

Noisy setting: *Quadratic programming* (Tibshirani, 1996)

Given noisy observations $y_k = x_k^T \beta^* + w_k$ where $w_k \sim N(0, \sigma^2)$.

ℓ_0 problem (Q_0)	ℓ_1 relaxation (Q_1)
$\min_{\beta \in \mathbb{R}^p} \sum_{k=1}^{n} (y_k - x_k^T \beta)^2 + \lambda \|\beta\|_0$	$\min_{\beta \in \mathbb{R}^p} \sum_{k=1}^{n} (y_k - x_k^T \beta)^2 + \lambda \|\beta\|_1$

Partial overview of previous work

- pioneering work on basis pursuit (relaxation L_1)
  (Chen, Donoho & Saunders, 1998)
- characterization of success for basis pursuit
  (e.g., Candes/Tao, Donoho, Elad, Goyal, Tropp ....)
- use/analysis of ℓ_1-constrained quadratic programming (Lasso)
  (e.g., Tibshirani, 1996; Knight & Fu, 2000...)
- use of Lasso for Gaussian graphical model selection
  (Meinshausen & Buhlmann, 2005; Zhao & Yu, 2006)
- noiseless setting: analysis of random Gaussian ensembles (Candes & Tao, 2005; Donoho, 2005)

Problem formulation

- given fixed but unknown vector $\beta^* \in \mathbb{R}^p$, define its *support set*

$$S = \{i \in \{1, \ldots, p\} \mid \beta_i^* \neq 0\}$$

and $s = |S|$.

- hence p is the *ambient dimension* of the problem (typically $p \gg s$)

- given n observations of the form

$$y_k = x_k^T \beta^* + w_k$$

Question: For which sequences $(n, p(n), s(n))$ is it possible/impossible to recover the support set S using the Lasso?

Assumptions on random Gaussian ensembles

- vector observation $Y = X\beta^* + W$ with random design matrix

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}, \qquad x_k \sim N(0, \Sigma)$$

1. **Dependency condition:** There exist constants $C_{min} > 0$ and $C_{max} < +\infty$ such that the min./max. eigenvalues satisfy

$$C_{min} \;\leq\; \Lambda_{min}(\Sigma_{SS}), \qquad \textit{and} \qquad \Lambda_{max}(\Sigma_{SS}) \;\leq\; C_{max}.$$

2. **Mutual incoherence:** There exists an $\delta \in (0, 1]$ such that

$$\|\Sigma_{S^c S}(\Sigma_{SS})^{-1}\|_\infty \;\;\leq\;\; 1 - \delta.$$

Illustrative examples

1. Uniform Gaussian ensemble $\Sigma = I$.

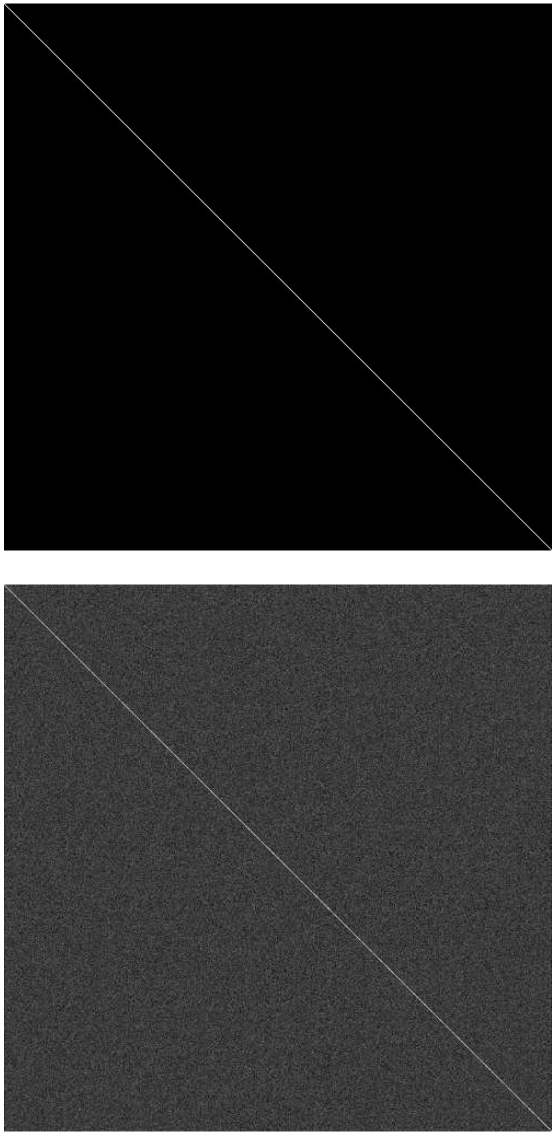
2. Toeplitz ensembles $\Sigma = \text{toep}\begin{bmatrix} 1 & \mu & \mu^2 & \cdots & \mu^{p-1} \end{bmatrix}$.

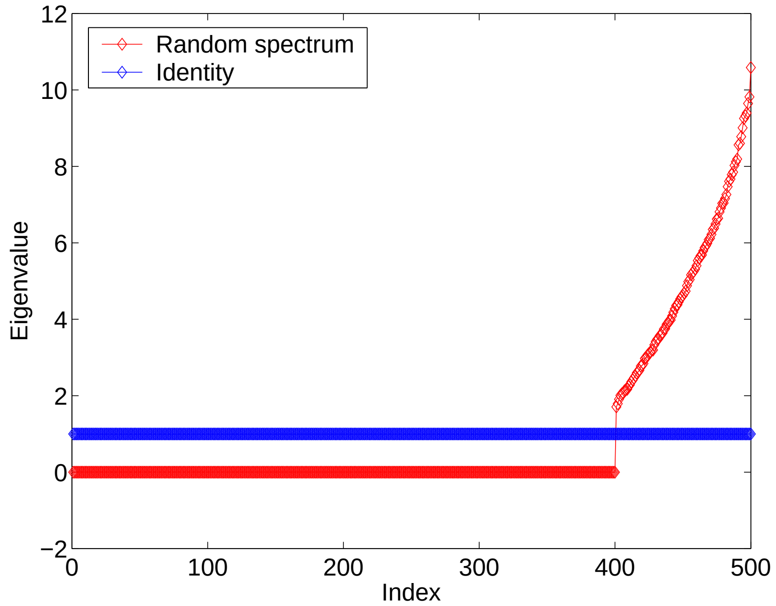
3. Bounded correlation models $|\Sigma_{ij}| \leq \frac{1}{2s-1}$.

4. Diagonally dominant matrices

Key remark: Depending on n and p, the random matrix $X^T X$ can have eigenvalues far away from those of Σ.

Covariance Σ versus random matrix





Thresholds for linear regression

Consider the sparse linear regression model

$$y_k = x_k^T \beta^* + w_k, \qquad k = 1, \ldots, n$$

where

- $\beta^* \in \mathbb{R}^p$ and $\|\beta^*\|_0 = s$.
- observation noise $w_k \sim N(0, \sigma^2)$
- random design vectors $x_k \sim N(0, \Sigma)$

Theorem: Successful recovery with the Lasso has threshold

$$n = \Theta\left(s \log(p - s) + s + 1\right).$$

I.e., there are constants $\theta_\ell \leq 1 \leq \theta_u$ such that for all $\epsilon > 0$:

(a) if $n > 2\,(\theta_u + \epsilon)\, s \log(p - s) + s + 1$, then $\mathbb{P}[\text{Success}] \rightarrow 1$ as $n \rightarrow +\infty$.

(a) conversely, if $n < 2\,(\theta_\ell - \epsilon)\, s \log(p - s) + s + 1$, then $\mathbb{P}[\text{Success}] \rightarrow 0$ as $n \rightarrow +\infty$.

Some corollaries

Linear underdetermined scaling:

- suppose that $n = \beta p$ for some $\beta \in (0, 1)$. Then w.h.p the Lasso recovers any sparsity pattern with $s = O(\frac{p}{\log p})$.
- sharp contrast with earlier results in the *noiseless setting*, where $s = \gamma p$ can be recovered (Donoho, 2005; Candes & Tao, 2005)

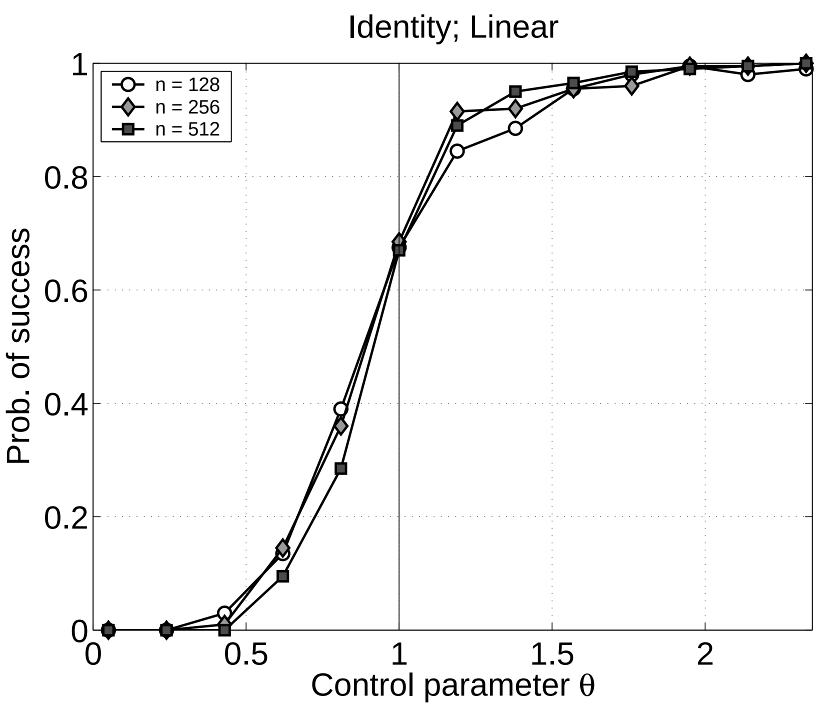
Exponential scaling: (Meinshausen & Buhlmann, Zhao & Yu, 2006)

Suppose that

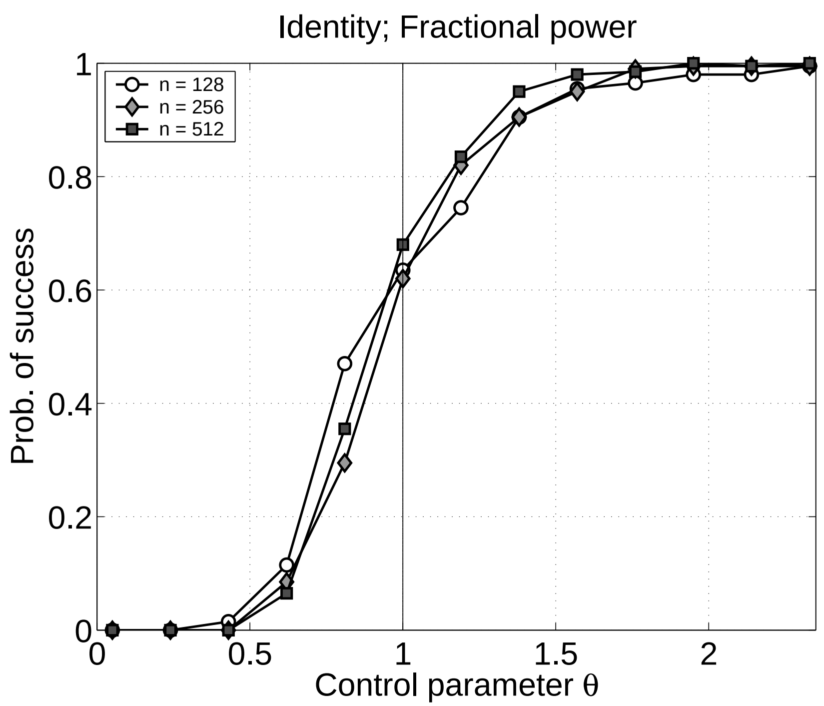
$$s = O(n^{c_1}) \qquad \text{and} \qquad p = O\left(\exp(n^{c_2})\right)$$

where $c_1 + c_2 < 1$. Then the Lasso recovers w.h.p. in recovering the sparsity pattern.

Illustration: Uniform Gaussian ensemble

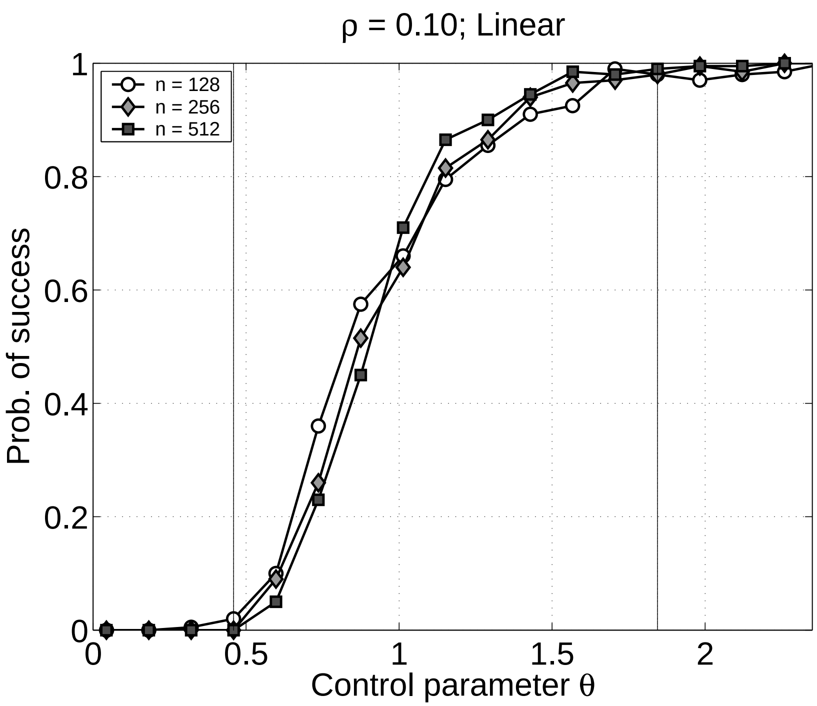


(a)

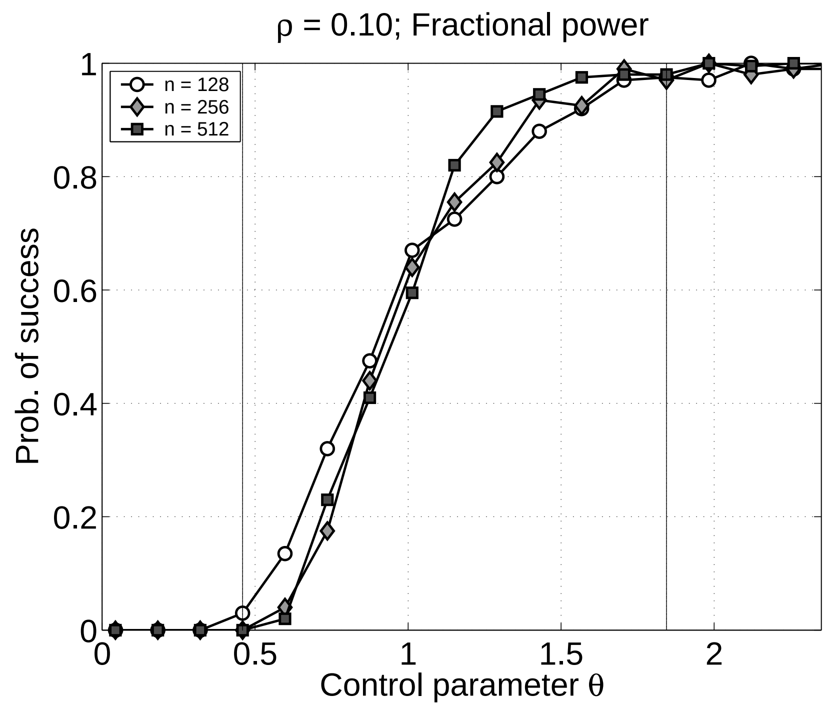


(b)

Illustration: Toeplitz Gaussian ensemble



(a)



(b)

Graphical model selection

- given an *unknown graph* $G = (V, E)$, consider the Markov random field

$$p(z; \beta) \quad \propto \quad \exp \left\{ \sum_{(s,t) \in E} \beta_{st} z_s z_t \right\}.$$

- conditioned on $(z_2, \ldots, z_m)$, the variable Z_1 has distribution

$$p_1(z; \beta) \;:=\; \mathbb{P}(Z_1 = 1 \mid z_2, \ldots, z_m) \quad = \quad \frac{1}{1 + \exp\left(\sum_{\mathbf{t} \in \mathcal{N}(\mathbf{1})} \beta_{1t} z_t \right)}.$$

- **Strategy:** perform logistic regression of node Z_1 on the rest to determine neighborhood structure $\mathcal{N}(1)$

- perform analogous regressions to determine neighborhood structures $\mathcal{N}(i), i \in V$ for the full graph

Method and notation

Method: Given samples $(z_1^k, z_2^k, \ldots, z_m^k)$:

1. For each node $i \in V$, perform ℓ_1 regularized logistic regression of z_i on the remaining variables $z_{\setminus i}$:

$$\widehat{\beta}^i \quad := \quad \arg\min_{\beta} \frac{1}{n} \sum_{k=1}^{n} \left[\log\left(1 + \beta^i \cdot z_{\setminus i}^k\right) - z_i^k \left(z_{\setminus i}^k\right) \cdot \beta^i \right] + \lambda_n \|\beta^i\|_1.$$

2. Estimate the local neighborhood $\widehat{\mathcal{N}}(i)$ as the support (non-negative entries) of the regression vector $\widehat{\beta}^i$.

Notation:

- define Fisher information matrix (at node i):

$$Q_i^* \quad = \quad \mathbb{E}\left[p_i(Z;\beta)\ (1 - p_i(Z;\beta)\ Z Z^T \right].$$

- focusing on a fixed node i, let Q_{SS}^* denote the submatrix associated with the support of $\mathcal{N}(i)$.

Assumptions

Dependency condition: There exist constants $C_{min} > 0$ and $C_{max} < +\infty$ such that

$$C_{min} \leq \Lambda_{min}(Q^*_{SS}), \qquad \textit{and} \qquad \Lambda_{max}(Q^*_{SS}) \leq C_{max}.$$

Incoherence There exists an $\delta \in (0, 1]$ such that

$$\|Q^*_{S^cS}(Q^*_{SS})^{-1}\|_\infty \quad \leq \quad 1 - \delta.$$

Growth rates: The growth rates of the number of observations n, the graph size p, and the maximum node degree $d_{\max}$ satisfy

$$\frac{n}{d_{\max}^5} - 6d_{\max}\log(d_{\max}) - 2\log(p) \quad \rightarrow \quad +\infty.$$

Model selection via regression

Method: Given samples $(z_1^k, z_2^k, \ldots, z_m^k)$:

1. For each node $i \in V$, perform ℓ_1 regularized logistic regression of Z_i on the remaining variables.
2. Estimate the local neighborhood $\widehat{\mathcal{N}}(i)$ as the support (non-negative entries) of the regression vector.
3. Combine the neighborhood estimates in a consistent manner (AND, or OR rule).

Theorem Suppose that the triple $(n, p, d_{\max})$ and the regularization parameter λ_n satisfy the conditions:
(a) $n\lambda_n^2 - 2\log(p) \to +\infty$, and (b) $d_{\max}\lambda_n \to 0$.

Then $\mathbb{P}[\hat{\mathcal{N}}_n(i) = \mathcal{N}(i),\ \forall\, i \in V_n] \to 1$ as $n \to +\infty$.

Summary and future directions

- for ℓ_1-regularized linear regression, established sharp thresholds for sparsity recovery
identity ensemble:	results are sharp
more general ensembles:	results can be sharpened
- established sufficient conditions for consistent model selection via logistic regression

Open questions:

- methods can be extended to more general families of graphical models
- can mutual incoherence be eliminated/weakened?
- what are fundamental information-theoretic limits of recovery?