Structure Learning in Infrastructure Networks

Rajasekhar Anguluri

Department of Computer Science and Electrical Engineering

University of Maryland, Baltimore County (UMBC)

rajangul@umbc.edu

Summer Workshop, July 15 – 18, 2024

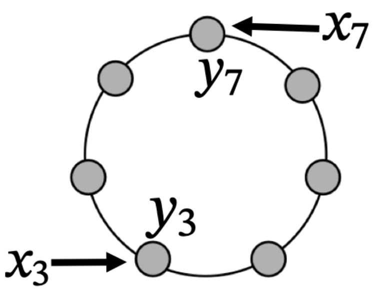
IIT, Bombay

slides template inspiration: Joel Tropp, Caltech

https://rajanguluri.github.io

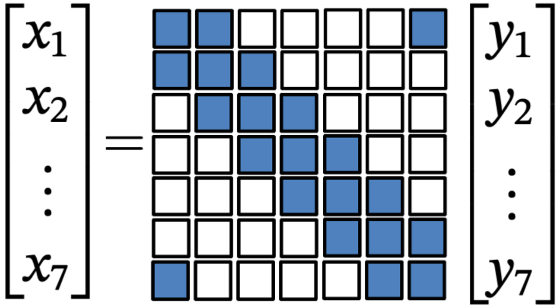
Structure Learning Problems: Recap

Network Structure = Laplacian's Sparsity Pattern



infrastructure network

sparsity (zero & non-zero) of L captures network connections



x nodal injections

L network Laplacian

y node potentials

- ***measurables:*** p-dim vectors x and y
- ***full coverage:*** access x or/and y
- ***partial coverage:*** sub-vectors of x or/and y

- ***linear model:***

$$\mathrm{Vec}(X) = H(Y)\,\mathrm{Ve}(L) + \mathrm{Vec}(E) \quad \text{full coverage}$$

- ***covariance models:***

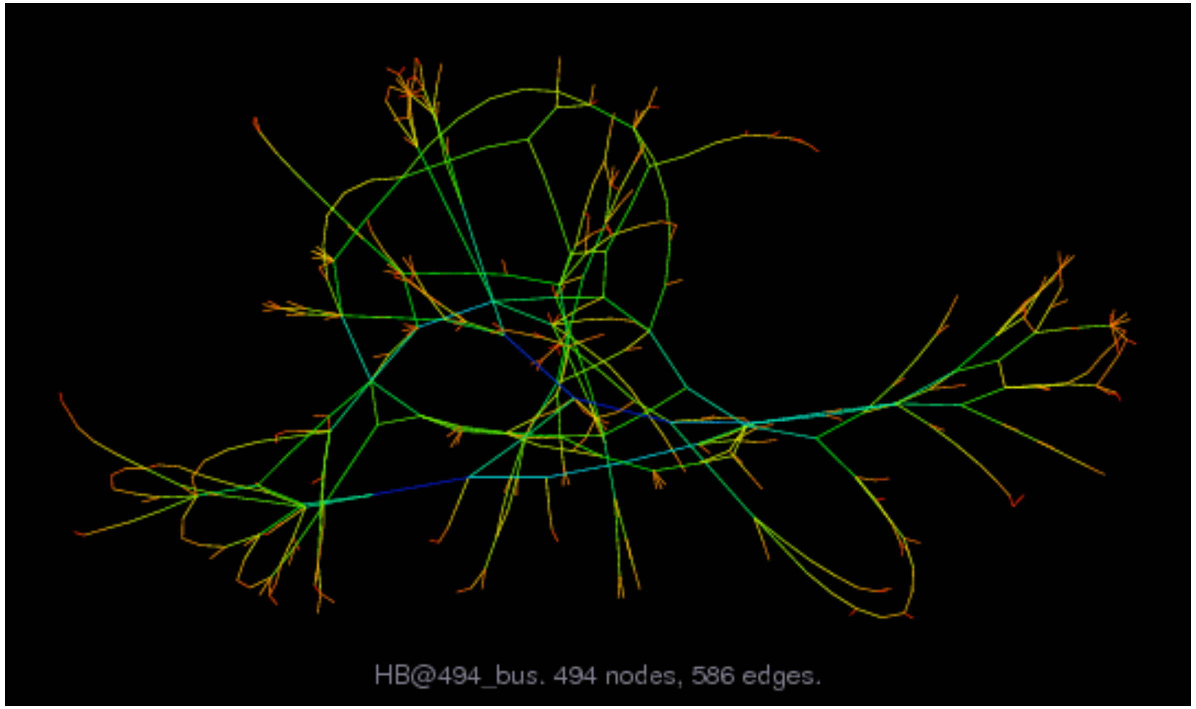
$$\Omega = L\Omega_x L \qquad \text{full coverage}$$

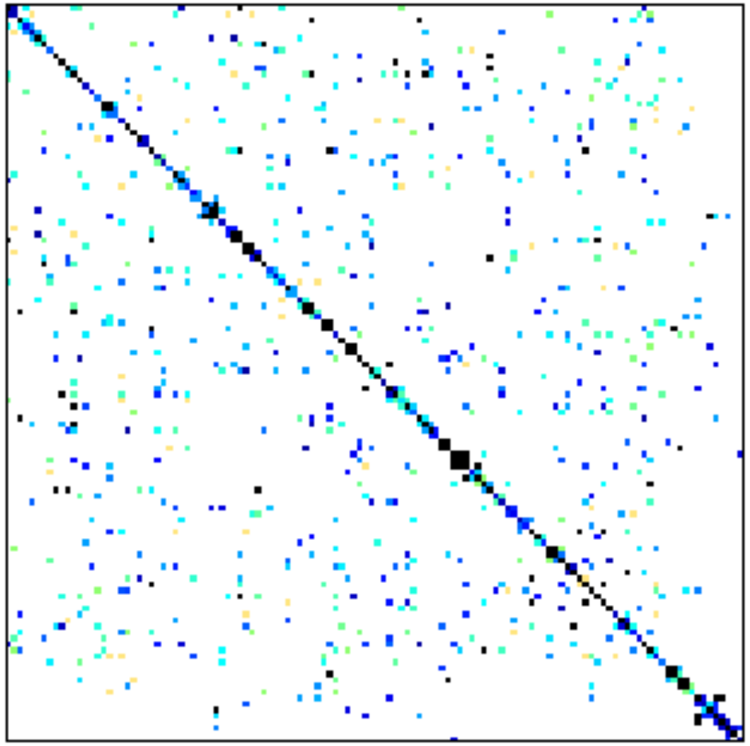
$$\Omega_{OO} = K_{OO} - K_{OH}K_{HH}^{-1}K_{HO} \qquad \text{partial}$$

- ***Estimation:***

1. estimate the vector $\mathrm{Ve}(L)$ from data
2. estimate matrices Ω *and* Ω_{OO} from data

Infrastructure Networks have Sparse Edges

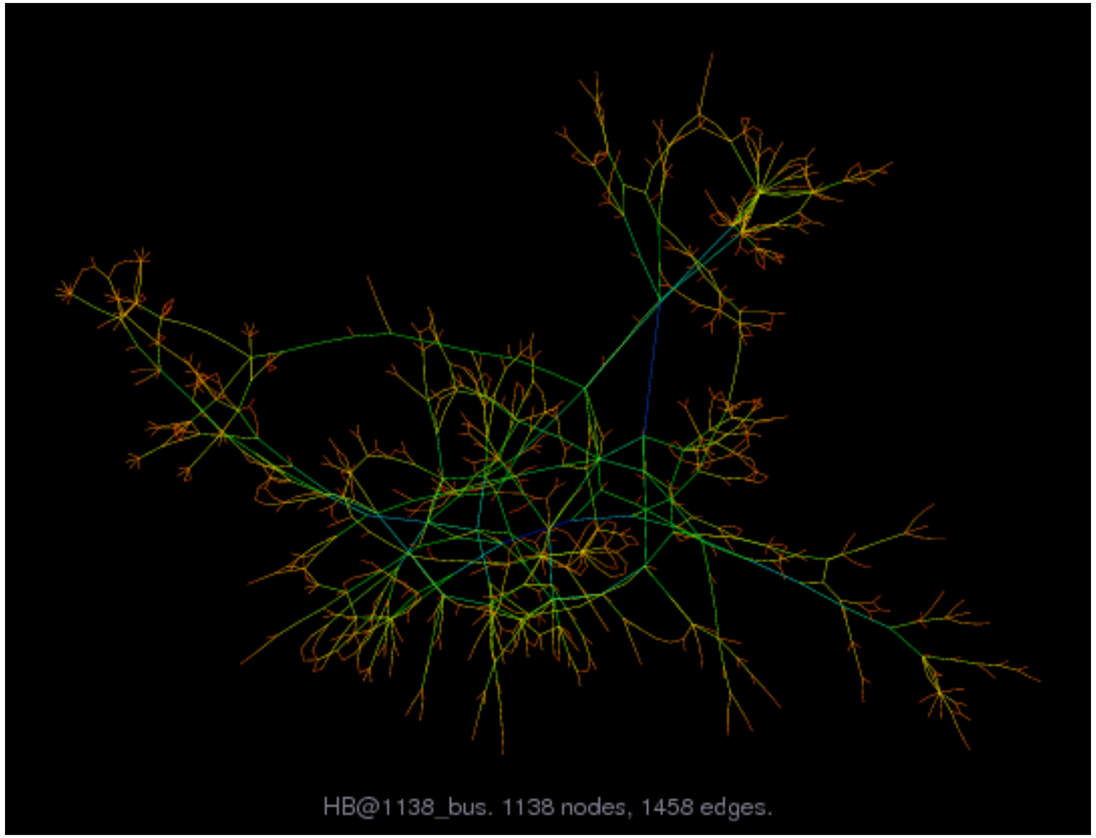


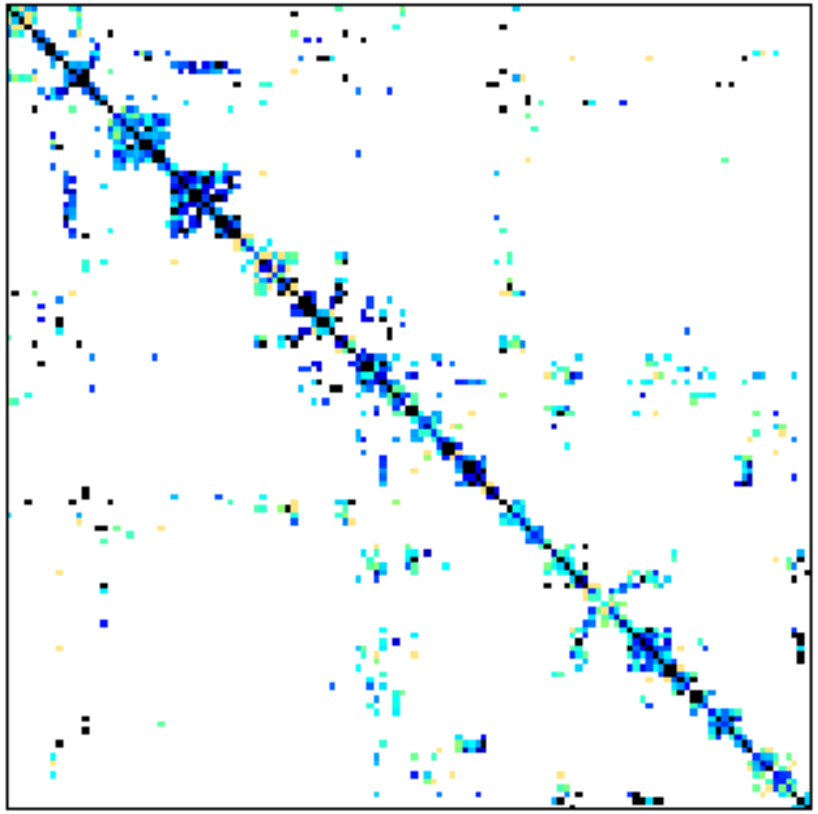


Visualization for 494 bus power network: (right) sparsity of the Laplacian matrix (colors represent the intensity of the weights) and (left) graph pattern

see here for other visualizations https://networkrepository.com/power-US-Grid.php

Infrastructure Networks have Sparse Edges





Visualization for 1132 bus power network: (right) sparsity of the Laplacian matrix (colors represent the intensity of the weights) and (left) graph pattern

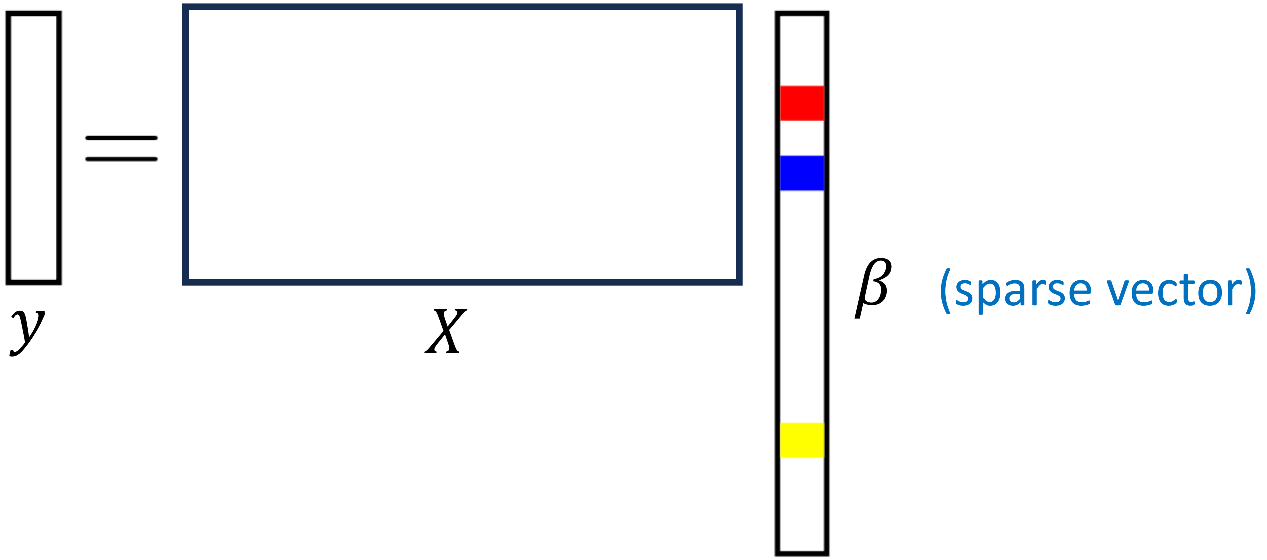
Sparse Estimation: Overview

- **Goals:** introduce basic concepts in sparse models; the role of convexity in developing an optimization method
  - **Goal 1:** sparse linear regression problem
  - **Goal 2:** sparse inverse covariance estimation problem
  - **Goal 3:** <u>a</u>lternating <u>d</u>irection <u>m</u>ethod of <u>m</u>ultipliers (ADMM)

Sparse Estimation: Overview

- **Goals:** introduce basic concepts in sparse models; the role of convexity in both analysis and optimization
  - **Goal 1:** sparse linear regression problem
  - **Goal 2:** sparse inverse covariance estimation problem
  - **Goal 3:** <u>a</u>lternating <u>d</u>irection <u>m</u>ethod of <u>m</u>ultipliers (ADMM)

Sparse Linear Regression: Basic Problem



Find $\beta \in \mathbb{R}^p$ such that $y = X\beta$

- $X = [x_1, \dots, x_p] \in \mathbb{R}^{n \times p}$ is a full rank matrix with $p \gg n$ (high-dimensions)
- A vector is s-sparse if has at most s non-zero entries

Linear Systems of Equations

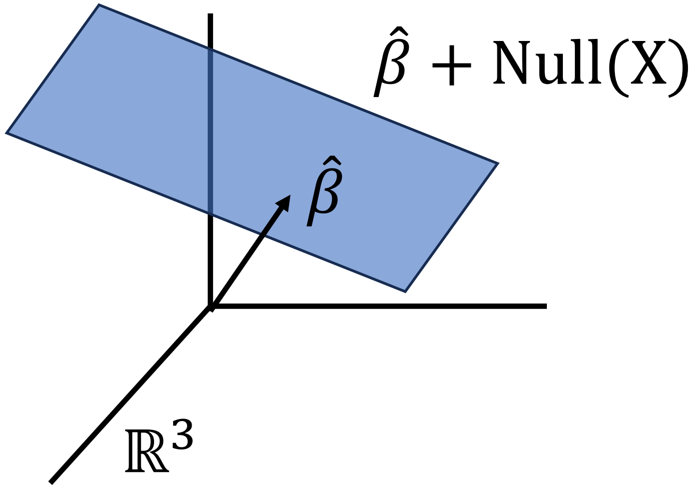
Thm: Consider a linear system $y = X\beta$.

- Existence: a solution β exists if and only if $y \in \text{range}(X)$
- Uniqueness: Let β_0 satisfy $y = X\beta$. The "infinite" solutions set: $\beta_0 + \text{Null(X)}$

Min-norm sol. Let $X \in \mathbb{R}^{n \times p}$ have full row-rank

$$\min \ ||\beta||_2^2 \quad \text{s.t.} \quad y = X\beta$$

$$\hat{\beta} = X^T(XX^T)^{-1}y$$



Norms: Finite-dimensional vectors

Def: A norm $||\beta||$: $\mathbb{R}^p \to [0, \infty)$ is a non-negative function with

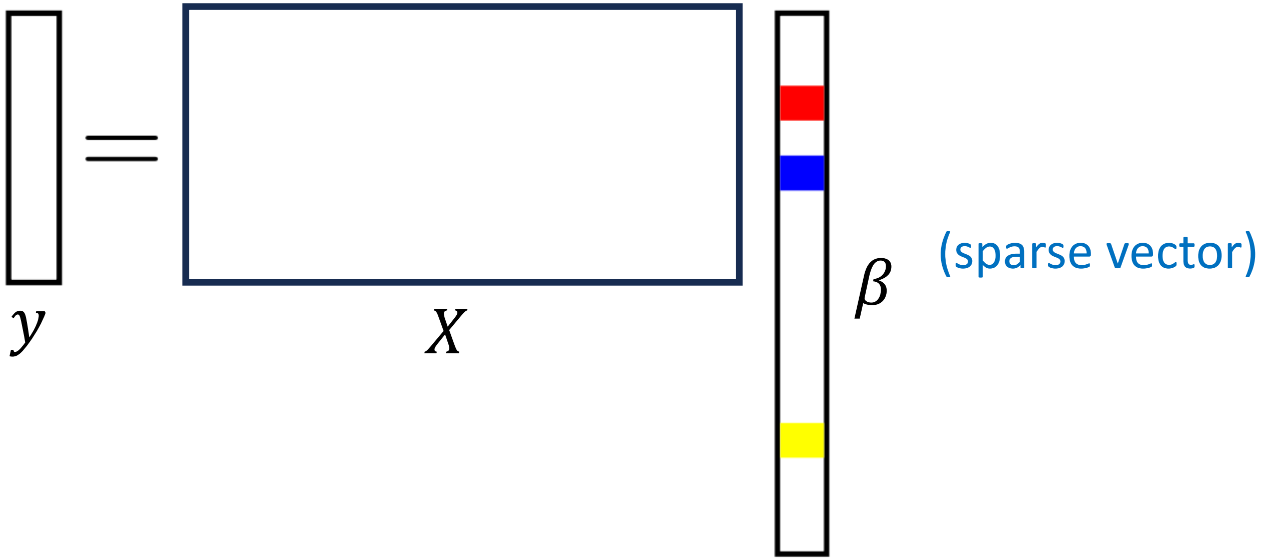
$||\alpha\beta|| = |\alpha|\ ||\beta||$ (positive scaling)

$||\beta_1 + \beta_2|| = ||\beta_1||_2 + ||\beta_2||_2$ (triangle inequality)

$||\beta|| = 0 \iff \beta = 0$ (non-degeneracy)

- ℓ_2 - (Euclidean): $||\beta||_2 = (\beta_1^2 + \cdots + \beta_p^2)^{1/2}$
- ℓ_1 - (Manhattan): $||\beta||_1 = |\beta_1| + \cdots |\beta_p|$
- ℓ_0 - (counting): $||\beta||_0 = |\mathrm{supp}(\beta)|$ (pseudo norm; fails scaling !)

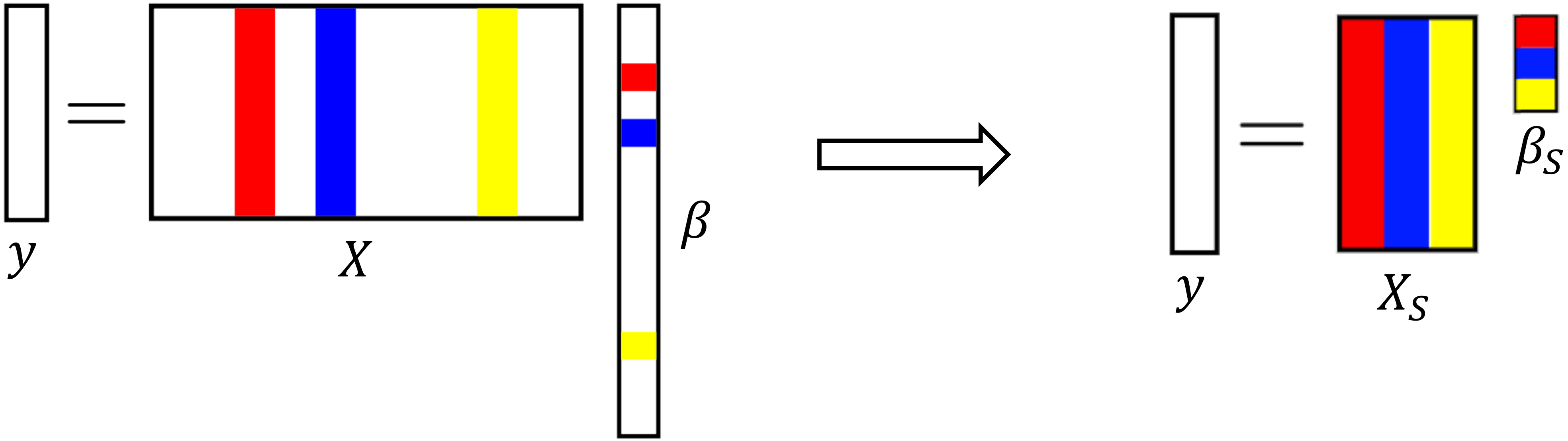
Sparse Linear Regression: ℓ_0- Minimization



find a sparse $\beta \in \mathbb{R}^p$ by solving the ℓ_0- norm minimization:

$$(\mathrm{P}_0) \qquad \begin{aligned} &\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \|\beta\|_0 \\ &\text{subject to } y = X\beta \end{aligned}$$

Sparse Linear Regression: ℓ_0- Minimization



- ℓ_0 - minimization is a combinatorial problem
- ℓ_0 - minimization is equivalent to column selection
- exhaustive search in exponential in s (for a $s-$ sparse vector)

Sparse Linear Regression: ℓ_0- Minimization

Exercise: suppose that the sparse solution to (P0) contains $s \leq p$ non-zero entries. Show that the exhaustive search algorithm should check at least $\sum_{j=1}^{s-1} {}^{p}C_j$ subsets.

$$\sum_{j=1}^{p} \binom{p}{j} = 2^p$$

$$2^{512} = 13407807929942597099574024998205846127479365820592393377723561443721 76403007354697680187429816690342769003185818648605085375388281194656 9946433649006084096.$$

Greedy Search or Convex Relaxation?

- **original problem:** computationally infeasible

$$(\mathrm{P}_0) \qquad \underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|\beta\|_0 \quad \text{subject to } y = X\beta$$

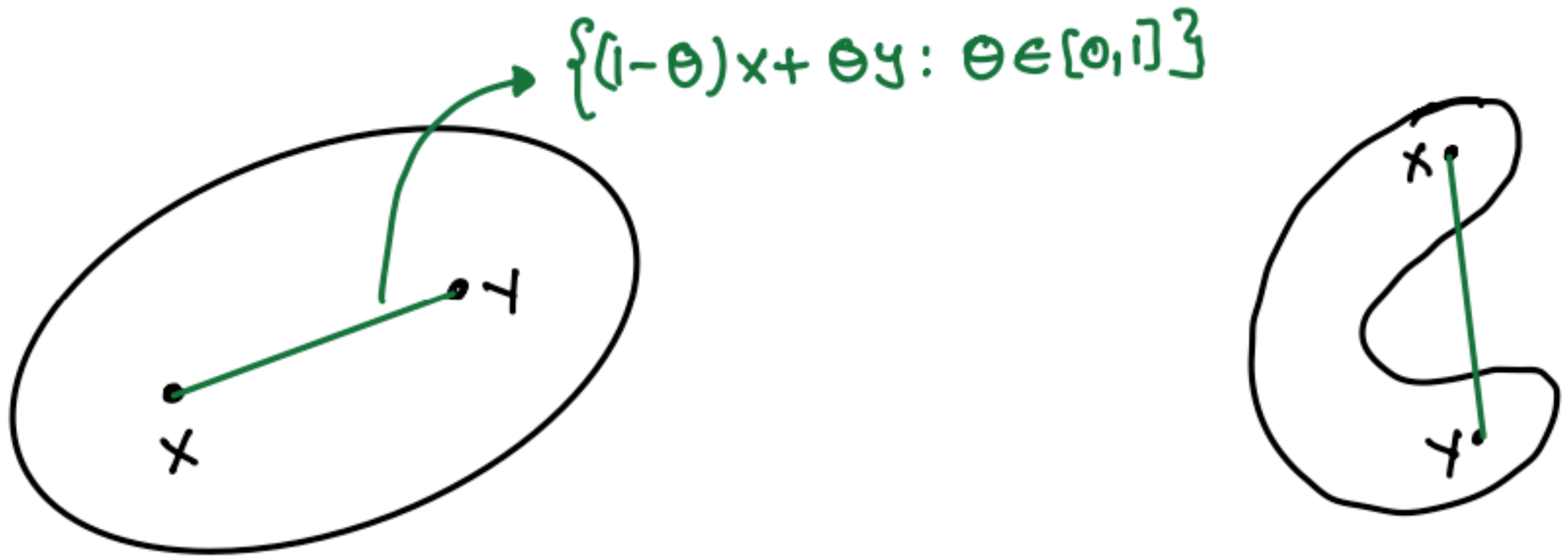
- **greedy approach:** search over subsets in an “intelligent” way

- **convex approach:** replace ℓ_0 - norm with ℓ_1 - norm: (why?)

$$(\mathrm{P}_1) \qquad \underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|\beta\|_1 \quad \text{subject to } y = X\beta$$

Convex Sets

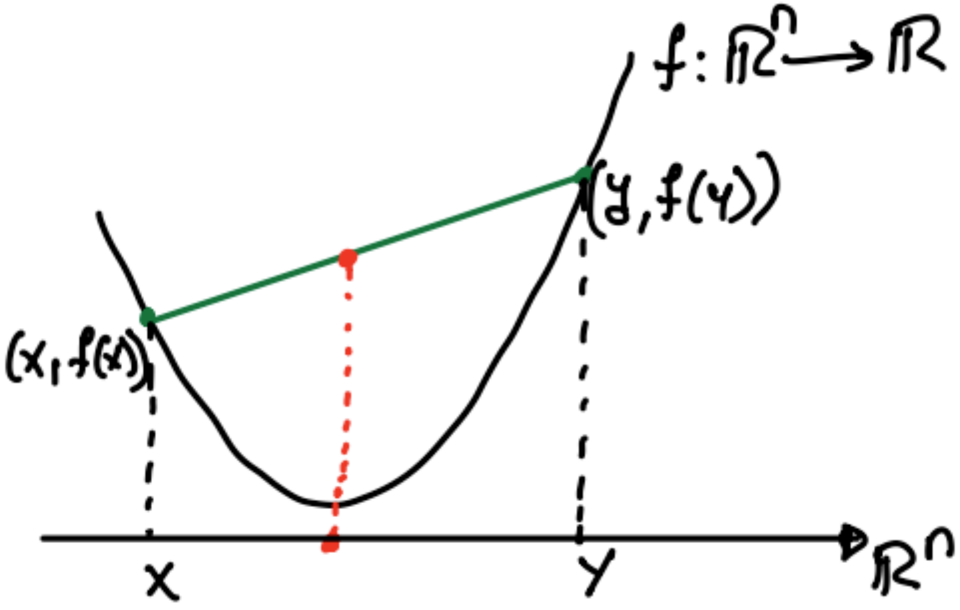
A set $K \subseteq \mathbb{R}^p$ (or $\mathbb{R}^{p\times p}$) is convex if, for all $x, y \in K$, the line segment connecting x and y is in K



Convex Function

A function $f: K \to \mathbb{R}$ is convex if its curve lies below any chord joining two of its points

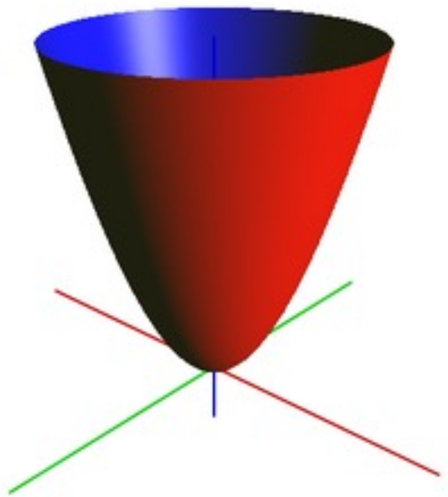
$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$



Convex Function

A function $f: K \to \mathbb{R}$ is convex iff when restricted to any line that intersects its domain is convex; that is,

$$g(t) = f(x + tv) \text{ is convex, } dom(g) = \{t | x + tv \in dom(f)\}$$
$$\text{for all } x \in dom(f) \text{ and } v \in \mathbb{R}^n$$



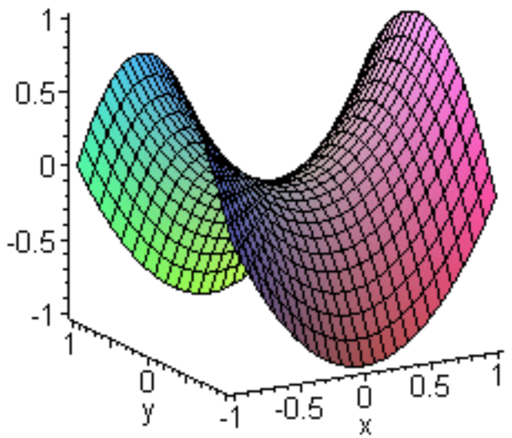


figure source: web

Convex Optimization Problem

- minimization: minimize a "convex" function over a convex set
- maximization: maximize a "concave" function over a convex set

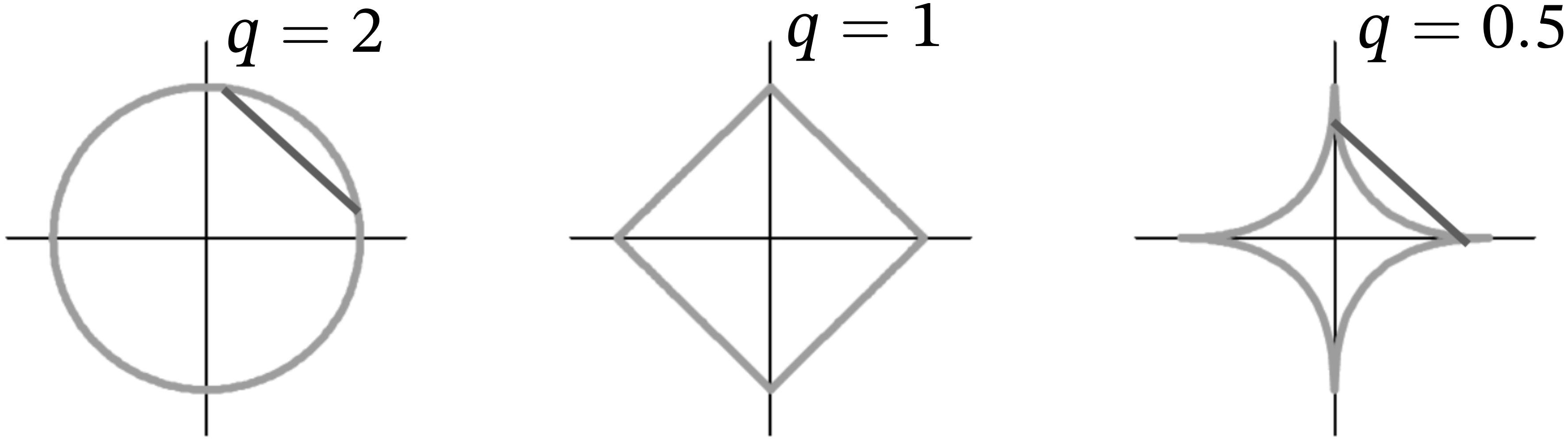
advantages:

- local minima are global minima
- polynomial time-algorithms with convergence
- beautiful theory: linear algebra, matrices, analysis, probability

ℓ_1- norm: "right" convex relaxation of ℓ_0

$$(P_0) \quad \begin{array}{l} \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \|\beta\|_0 \\ \text{subject to } y = X\beta \end{array}$$

$$(P_p) \quad \begin{array}{l} \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \|\beta\|_q^q \\ \text{subject to } y = X\beta \end{array}$$

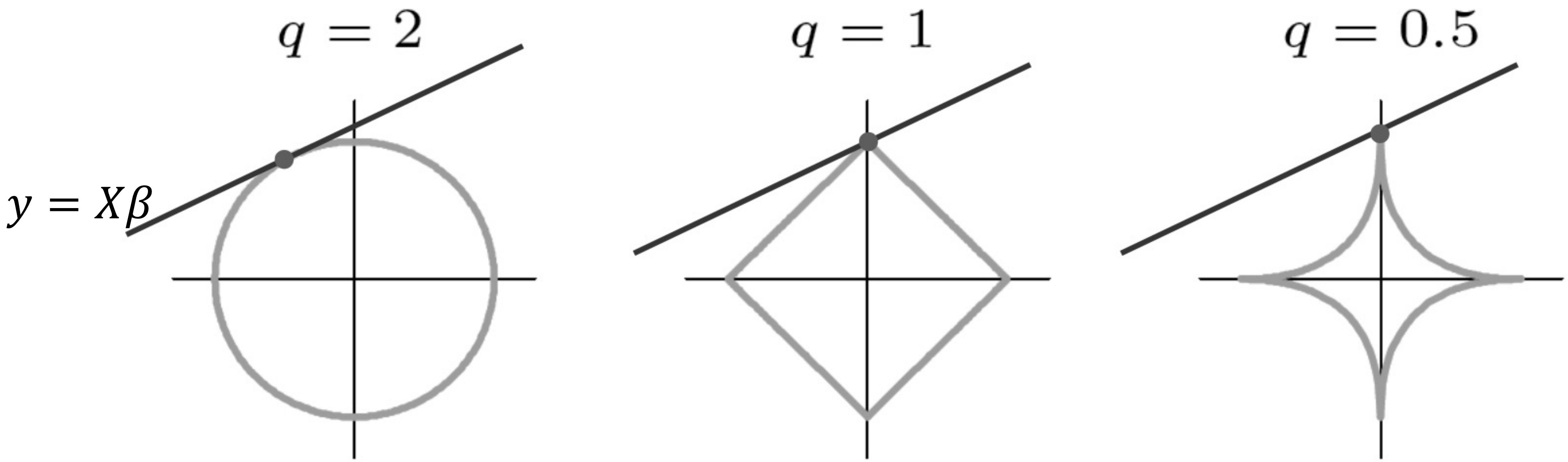


norm balls $||\beta||_q^q \leq 1$: (convex to non-convex; smooth to sharp corners)

ℓ_1- norm: "right" convex relaxation of ℓ_0

$$(P_0) \quad \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \; \|\beta\|_0 \quad \text{subject to } y = X\beta$$

$$(P_p) \quad \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \; \|\beta\|_q^q \quad \text{subject to } y = X\beta$$



contours $||\beta||_q^q = 1$ touching the linear subspace $\{\beta : y = X\beta\}$

Statistical Learning vs Compressed Sensing

- heuristic and intuitive explanation of ℓ_1 relaxation can be rigorously analyzed by one either (i) statistical learning or (ii) compressed sensing
- compressed sensing:

  - design matrix X is random and user choice

  - non-asymptotic analysis with focus on correct support recovery (sparsity pattern)
- statistical learning:

  - design matrix X is non-random

  - asymptotic and non-asymptotic analysis

  - prediction error; estimation consistency; and model selection error (sparsity pattern)

  - wide applications (generalized linear models, graphical models, Bayesian networks, etc.,)

ℓ_1 - Minimization Problems: Noisy Setting

- exercise: show all three forms are equivalent and convex.

$$\begin{aligned} &\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|\beta\|_1 \\ &\text{subject to } \ \|y - X\beta\|_2 \leq \varepsilon \end{aligned}$$

(noisy basis pursuit)

$$\begin{aligned} &\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|y - X\beta\|_2^2 \\ &\text{subject to } \ \|\beta\|_1 \leq t \end{aligned}$$

(no fancy name!)

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

LASSO

- regularization parameter λ is determined empirically often

ℓ_1 - Minimization Problems: Noisy Setting

- exercise: show all three forms are equivalent and convex.

$$\boxed{\begin{aligned} &\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|\beta\|_1 \\ &\text{subject to } \|y - X\beta\|_2 \le \varepsilon \end{aligned}}$$

(noisy basis pursuit)

$$\boxed{\begin{aligned} &\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|y - X\beta\|_2^2 \\ &\text{subject to } \|\beta\|_1 \le t \end{aligned}}$$

(no fancy name!)

$$\boxed{\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{loss function}} + \underbrace{\lambda\|\beta\|_1}_{\text{regularizer}}}$$

LASSO

A Toy Numerical Example (OLS vs LASSO)

- $y = X\beta + \epsilon$ (with $\epsilon_i \sim N(0, \sigma_e^2)$ and $X_{ij} \sim N(0,1)$

$$\beta = \begin{bmatrix} 0 \\ 2 \\ 0 \\ -3 \\ 0 \end{bmatrix} \quad \left[\begin{array}{cc} \text{OLS} & \text{LASSO} \\ \hline 0.2523 & 0 \\ 1.7341 & 2.0933 \\ 2.2651 & 0 \\ -3.8986 & -2.4351 \\ 0.1073 & -0.3054 \end{array}\right] \quad \left[\begin{array}{cc} \text{OLS} & \text{LASSO} \\ \hline 0.0505 & 0 \\ 1.6480 & 1.5579 \\ 0.4530 & 0 \\ -3.0418 & -2.1472 \\ 0.0215 & 0 \end{array}\right] \quad \left[\begin{array}{cc} \text{OLS} & \text{LASSO} \\ \hline 0.0090 & 0 \\ 1.9952 & 1.5579 \\ 0.0049 & 0 \\ -3.0418 & -2.5383 \\ 0.0267 & 0 \end{array}\right]$$

$n = 6; \sigma_e^2 = 0.5$ $\quad$ $n = 6; \sigma_e^2 = 0.1$ $\quad$ $n = 20; \sigma_e^2 = 0.1$

$\lambda = 0.1890$ $\quad$ $\lambda = 0.9625$ $\quad$ $\lambda = 0.5874$

OLS means ordinary least squares

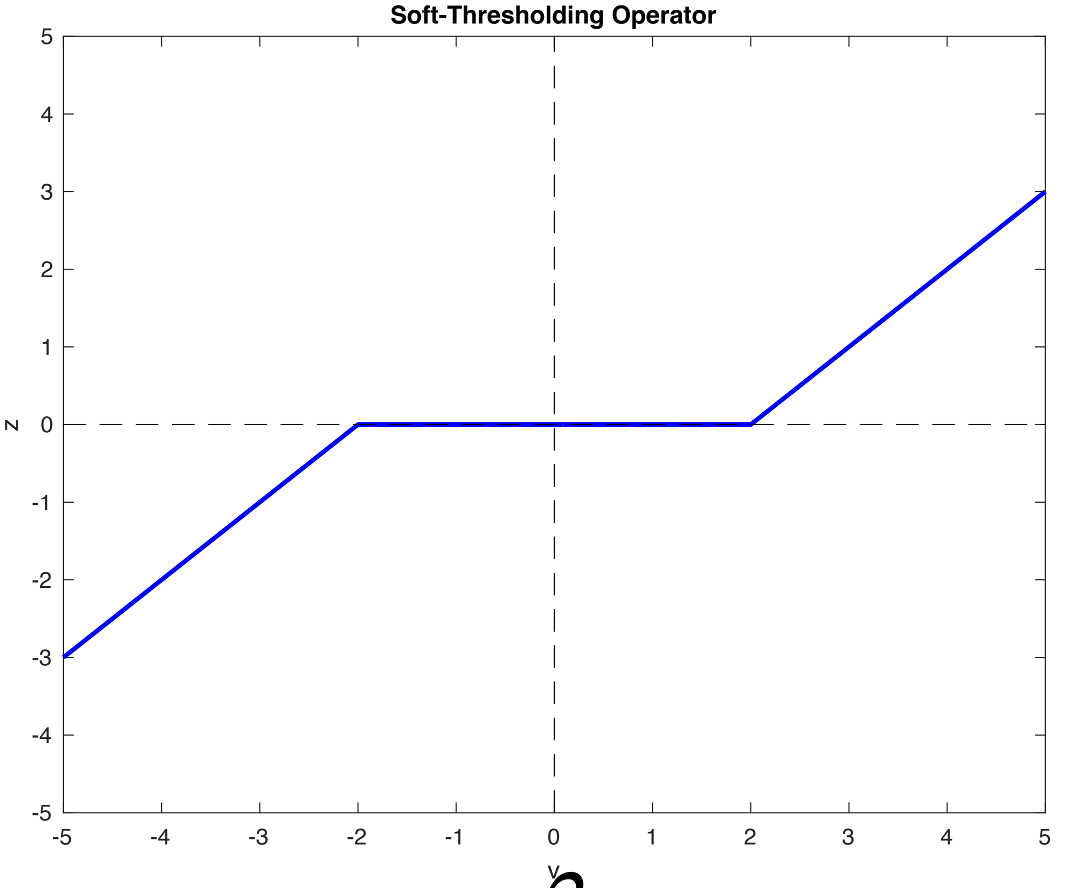
LASSO: Insights

- least absolute shrinkage and selection operator (LASSO)
- *statistics:* popularized by R.Tibshirani in 1990s (dates to 1970)
- *signal processing:* popularized by D. Donoho in 1990s.
- solution requires iterative techniques (e.g., ADMM)
- for $X^T X = I$, LASSO admits closed form solution (see first two papers in the references)

Soft-thresholding Operator

- soft-thresholding or shrinkage operator (see supplement as well)

$S_4(\beta)$



β

Sparse Estimation: Overview

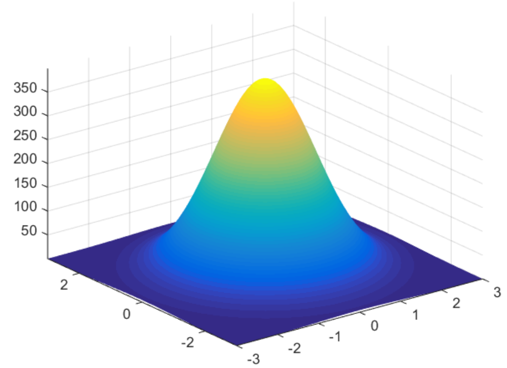
- **Goals:** introduce basic concepts in sparse models; the role of convexity in both analysis and optimization
  - **Goal 1:** sparse linear regression problem
  - **Goal 2:** sparse inverse covariance estimation problem
  - **Goal 3:** alternating direction method of multipliers (ADMM)

Maximum Likelihood Estimate of $\Omega = \Sigma^{-1}$

- let $y \sim \mathcal{N}(0, \Omega^{-1})$, where Ω is the $p \times p$ inverse covariance matrix

- probability density function:

$$f(y) = \frac{1}{\sqrt{(2\pi)^p \det\left(\Omega^{-1}\right)}} \exp\left(-y^T \Omega y/2\right)$$



- unconstrained MLE of Ω based on i.i.d $y_1, \ldots, y_K$

$$\max_{\Omega \succ 0} \underbrace{f(y_1) f(y_2) \ldots f(y_K)}_{\ell(S_K;\Omega)}$$

Exercise problems

- $(K \geq p)$ MLE: $\widehat{\Omega} = S_K^{-1}$
- $(K < p)$ MLE does not exist

MLE: Exercise Problems

- show that the MLE for Σ is not a convex optimization problem:

$$\hat{\Sigma} = \max_{\Sigma \succ 0} \; -[\log(\det \Sigma) + \mathrm{Tr}(S_K \Sigma^{-1})]$$

where $S_K = \frac{1}{K}\sum_{k=1}^{K} y_k y_k^T$; and $\succ 0$ means positive definiteness of matrix

- variable change: let $\Omega = \Sigma^{-1}$. Show the MLE for Ω is a convex optimization:

$$\hat{\Omega} = \max_{\Omega \succ 0} [\log(\det \Omega) - \mathrm{Tr}(S_K \Omega)]$$

- hint: use the convexity of the restricted line segment method

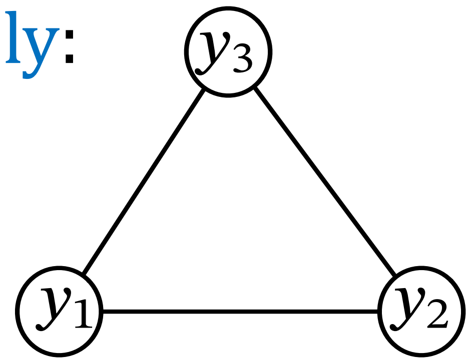
Gaussian Graphical Models: Quick Review

- inverse covariance matrices can be sparse:

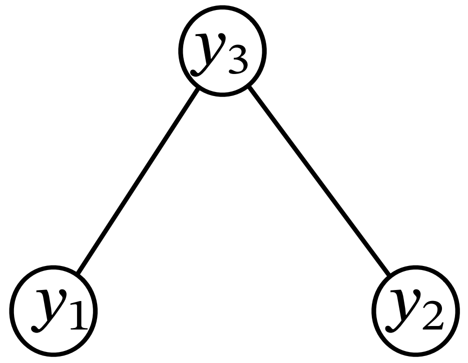
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \sim \mathcal{N}(0, \Sigma) \qquad \Sigma = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad \Omega = \Sigma^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

- $\Sigma_{12}^{-1} = 0$ means y_1 and y_2 are independent conditioned on y_3

- graphically:



covariance or independence (dense typically)



partial correlation or conditional indep (could be sparse)

Sparse Inverse Covariance Matrix Estimation

- $\ell_1 -$ MLE for inverse covariance estimation:

$$\widehat{\Omega} = \max_{\Omega \succ 0} \left[\log(\det \Omega) - \mathrm{Tr}(S_K \Omega)\right] - \lambda ||\Omega||_1$$

- $||\Omega||_1$ is the $\ell_1 -$ norm on the off-diagonal entries

- for $\lambda > 0$, the ℓ_1 −MLE problem is convex (so a unique solution)

- $\ell_1 -$ MLE as a minimization problem (we use this often than the max)

$$\widehat{\Omega} = \min_{\Omega \succ 0} \left[\mathrm{Tr}(S_K \Omega) - \log(\det \Omega)\right] + \lambda ||\Omega||_1$$

Sparse Inverse Covariance with Hidden Nodes

- for the observed inverse covariance matrix we have

$$\Omega_{OO} \;=\; K_{OO} - K_{OH}K_{HH}^{-1}K_{HO} \triangleq \Theta - \bar{L}$$

- let the sample covariance : $\bar{S}_K = \frac{1}{K}\sum_{k=1}^{K} y_{O,k}y_{O,k}^T$ and consider the MLE

$$\widehat{\Omega} = \min_{\Theta,\bar{L}\succ 0} \left[\mathrm{Tr}(\bar{S}_K(\Theta - \bar{L})) - \log(\det(\Theta - \bar{L}))\right] + \lambda||\Theta||_1 + \alpha\mathrm{Tr}(\bar{L})$$

subject to $\Theta - \bar{L} \succ 0; \bar{L} \succcurlyeq 0$

- the MLE is jointly convex in $(\Theta, \bar{L})$; and $\alpha, \lambda > 0$ is user defined; $\mathrm{Tr}(\bar{L})$ is the sum of singular values
- the ADMM method is described in the handwritten notes supplement
- we skip how to decompose $K_{OO} = S + M$; (for context see slides for day 2) and details are in [2]

[1] R. Anguluri (2023) Grid topology identification with hidden nodes via structured norm minimization, IEEE CSS Letters, 6: 1244-1249

Beyond Simple Sparse Models

$$\text{Loss}(\beta; \text{data}) + \text{Regularizer}(\beta)$$

other losses
(e.g., likelihoods)

- generalized linear models (exponential family noise)
- Gaussians and Ising models (Markov random fields)
- Principal component and factor analysis

structure beyond
naïve sparsity

- elastic Net
- fused Lasso
- block l1-lq norms (group Lasso)
- non-convex penalties

Sparse Estimation: Overview

- **Goals:** introduce basic concepts in sparse models; the role of convexity in both analysis and optimization

  - **Goal 1:** sparse linear regression problem

  - **Goal 2:** sparse inverse covariance estimation problem

  - **Goal 3:** alternating direction method of multipliers (ADMM)

Alternating Direction Method of Multipliers

- decomposes complex into simpler problems
- suitable for large-scale and distributed optimization
- easy to handle non-differentiable functions (e.g., $\ell_1 -$ norm)
- robust convergence properties
- old (1976) but gold technique

ADMM: General Recipe

- general problem form (with f, g convex):

$$\begin{aligned} \text{minimize}_{x,z} \quad & f(x) + g(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

- $L_\rho(x, y, z) = f(x) + g(z) + v^T(Ax + Bz - c) + (\rho/2)||Ax + Bz - c||_2^2$

- ADMM:

$$x^{k+1} := \text{argmin}_x L_\rho\left(x, z^k, y^k\right) \quad \text{// x- minimization}$$

$$z^{k+1} := \text{argmin}_z L_\rho\left(x^{k+1}, z, y^k\right) \quad \text{// z- minimization}$$

$$v^{k+1} := v^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right) \quad \text{// multiplier update}$$

Source: https://stanford.edu/~boyd/admm.html

ADMM for LASSO

- LASSO problem:

$$\text{minimize} \quad (1/2)\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- ADMM form:

$$\text{minimize} \quad (1/2)\|y - X\beta\|_2^2 + \lambda\|z\|_1$$

$$\text{subject} \quad \beta - z = 0$$

- ADMM (scaled):

$$\beta^{k+1} := \left(X^T X + \rho I\right)^{-1} \left(X^T y + \rho(z^k - v^k)\right)$$ // x- minimization

$$z^{k+1} := S(\beta^{k+1} + v^k, \lambda/\rho)$$ // element-wise soft thresholding

$$v^{k+1} := v^k + \left(\beta^{k+1} - z^{k+1}\right)$$ // multiplier update

Source: https://stanford.edu/~boyd/admm.html

ADMM for Sparse Inverse Covariance Matrix

- MLE (minimization) problem:

$$\text{minimize} \quad \mathrm{Tr}(S\Omega) - \log\det(\Omega) + \lambda\|\Omega\|_1$$

- ADMM form:

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}(S\Omega) - \log\det(\Omega) + \lambda\|Z\|_1 \\ \text{subject to} & \Omega - Z = 0 \end{array}$$

- ADMM (scaled):

$$\Omega^{k+1} := \underset{\Omega}{\mathrm{argmin}} \left(\mathrm{Tr}(S\Omega) - \log\det\Omega + (\rho/2)\left\|\Omega - Z^k + U^k\right\|_F^2\right)$$ // X- minimization

$$Z^{k+1} := S\left(\Omega^{k+1} + U^k, \lambda/\rho\right)$$ // soft thresholding

$$U^{k+1} := U^k + (\Omega^{k+1} - Z^{k+1})$$ // multiplier update

Source: https://stanford.edu/~boyd/admm.html

To learn more…

Contact: rangulur@asu.edu https://rajanguluri.github.io (lecture notes: coming soon)

Books:

- I. Rish and G. Grabarnik (2014). Sparse modeling: theory, algorithms, and applications, CRC press
- J. Suzuki (2021). Sparse estimation with math and Python, Springer.
- F. Bach et.al. (2012). Optimization with sparsity-inducing penalties, Now Publishers (free online).
- T Hastie, R. Tibshirani, and M. Wainwright (2015). Statistical learning with sparsity, CRC press (free online)
- M. Nagahara (2020). Sparsity methods for systems and control, Now Publishers (free online)
- J. Wright and Y. Ma (2022). High-dimensional analysis with low-dimensional models, Cambridge Press (free online)

Papers:

- N. Gauraha (2018). Introduction to the lasso: A convex optimization approach for high-dimensional problems, Resonance 23 (4), 439-464
- N. Gauraha (2016). Constraints and conditions: the Lasso oracle-inequalities, arXiv:1603.06177 (2016).
- M. Drton and M.H. Maathuis (2017). Structure learning in graphical modeling, Annual Review of Statistics and Its App., 4(1), 365-393
- Y. Zhao and X. Huo (2018). A survey of numerical algorithms that can solve the Lasso problems, Wiley Interdisciplinary Reviews: Computational Statistics, 15(4), e1602 (free online)
- K. Scheinberg and S. Ma (2012). Optimization methods for sparse inverse covariance selection. In Optimization for Machine Learning, MIT press.