Graphical Model Selection for Big Data over Networks

Alexander Jung

Department of Computer Science
Aalto University, School of Science

20.12.2016

Outline

Being a Statistician Then

"You haven't told me yet," said Lady Nuttal, "what it is your fiance does for a living."

"He's a statistician," replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that statisticians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

...

taken from Kendall and Stuart

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012

The Data Deluge

"We're drowning in information and starving for knowledge."
- Rutherford D. Rogers.

Data, Data Everywhere

- Atacama Large Millimeter Array yields hundreds of TB (10^{12} bytes)/year
- CERN experiment generates data at rate of PB (10^{15} bytes)/second
- annual internet traffic crushed ZB (10^{21} bytes) borderline
- Big Data challenge: how to process data at internet scale?

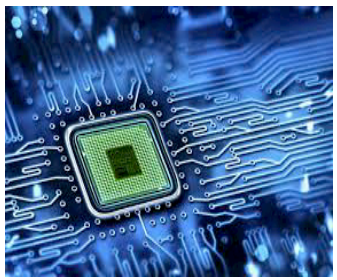
The Three V's of Big Data

- Volume: deal with Zettabyte ($= 10^{21}$ bytes) scale data
- Velocity: CERN experiment generates Petabytes ($= 10^{15}$ bytes) per second
- Variety: heterogeneous data (text, audio, video, graphs, ...)

Big Data over Networks

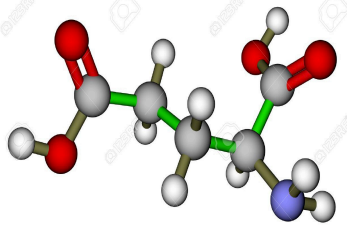
often the datasets have an intrinsic network structure

chip design



internet

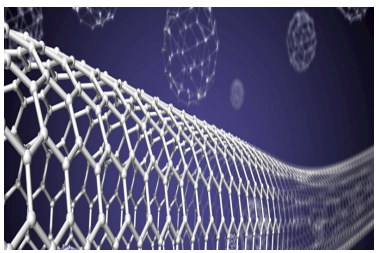
bioinformatics



social networks

universe

material science



cf. L. Lovász, "Large Networks and Graph Limits"

Big Data over Networks

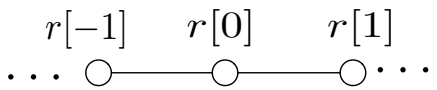
- observe dataset $\mathcal{D} = \{z_1, \ldots, z_p\}$ with data points z_i
- particular data point z_i might be audio, video or text data
- data points are structured by some notion of "similarity"
- z_i, z_j similar if they belong to same user account
- represent data point z_i by node $i \in \mathcal{V}$ of graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- edge (i, j) connects similar data points z_i and z_j

Graph Signals for Supervised Learning

- consider supervised machine learning from dataset $\mathcal{D}$
- data point z_i associated with a label $r[i]$ (e.g., persons preference for buying red shoes)
- entire labelling is a graph signal $r[\cdot] : \mathcal{V} \to \mathbb{R}$
- graph signal $r[\cdot]$ maps node $i \in \mathcal{V}$ to its label $r[i]$
- graph signal processing (GSP) provides efficient tools for handling large-scale graph signals

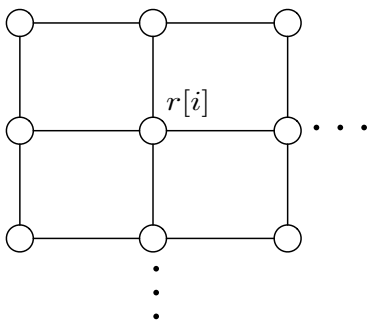
GSP generalizes DSP

- view discrete-time signals as graph signals over chain graph



label $r[i]$ might correspond to presence of "clipping" at time i

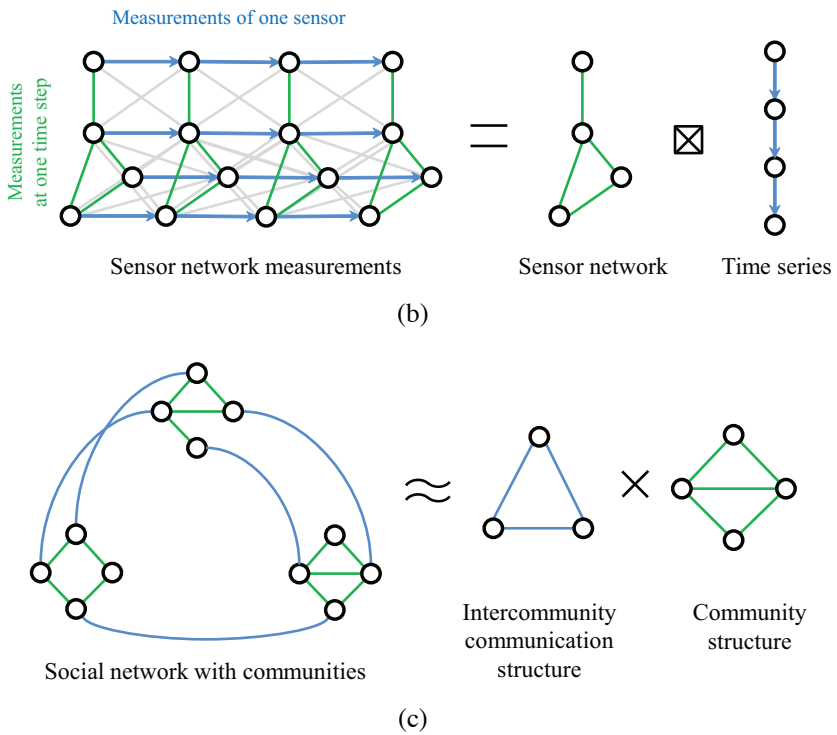
- (greyscale) images are signals over grid graph



label $r[i]$ might be presence of certain object at location i

Fast Algorithms on Graphs

- GSP theory yields fast algorithms for large-scale graphs
- generalizes FFT from chain graph to general graphs
- based on product graph structure



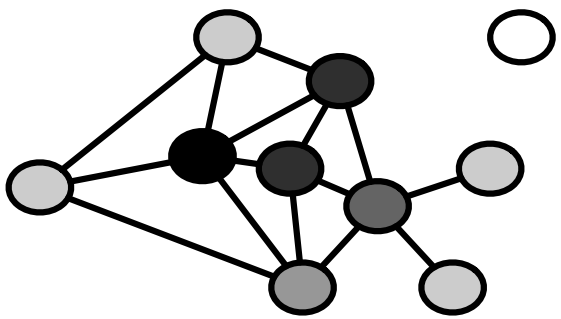
Graph Models: Perfect Match for 3 Vs of Big Data

- graph models lead to message passing algorithms
- message passing algorithms are perfectly scalable
- copes with volume (distributed computing) and velocity (parallel computing) of big data
- "ship computation to data" and not vice-versa!
- graph models also allow to process heterogeneous data

Semi-Supervised Learning for Big Data over Networks

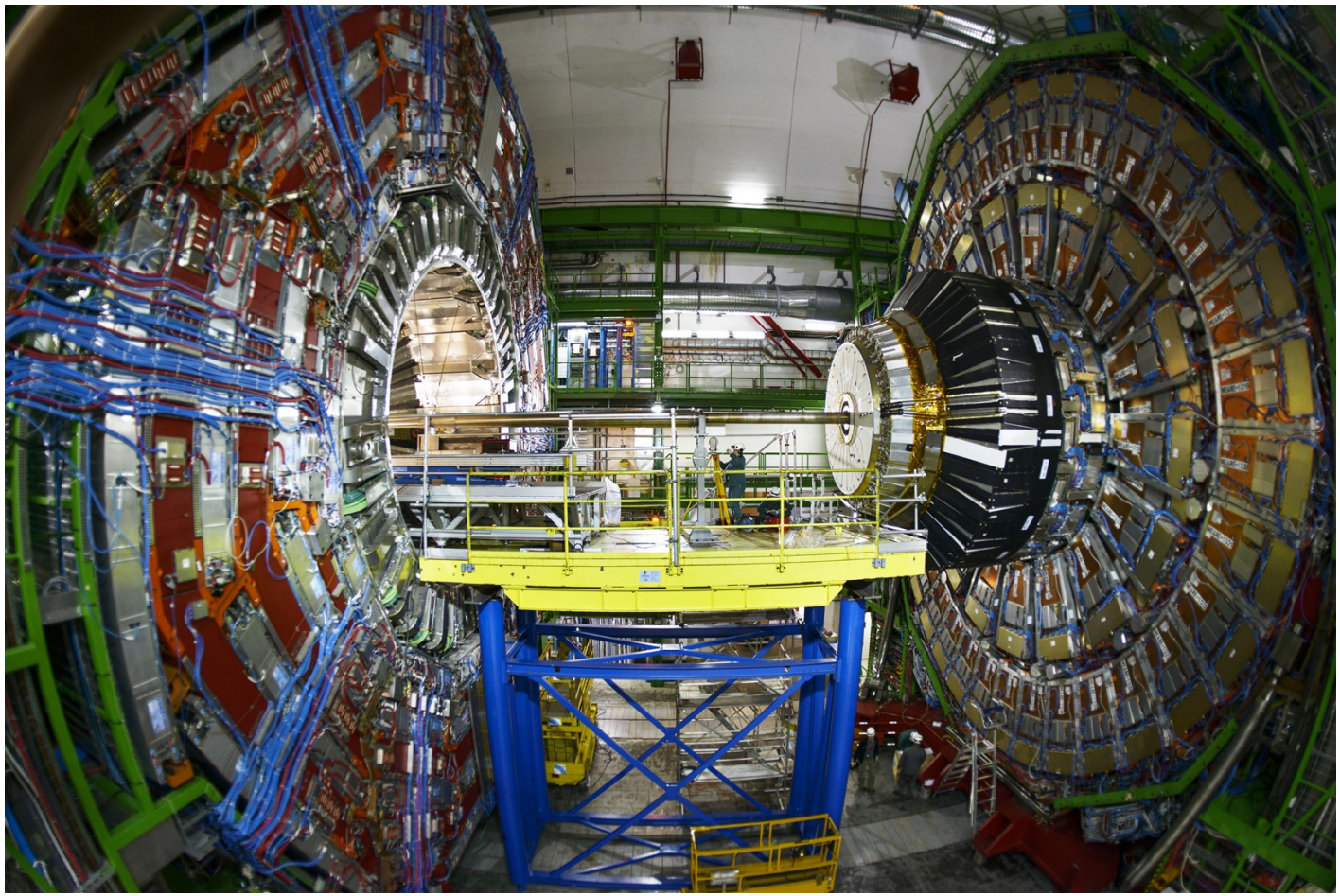
- consider graph signal $r[i]$ representing labeled dataset $\mathcal{D}$
- observe labels only at sampling set $\mathcal{S} \subseteq \mathcal{V}$
- acquiring labels is costly
- how to recover remaining unobserved labels $r[i]$ for $i \in \mathcal{V} \setminus \mathcal{S}$
- central smoothness hypothesis of supervised learning

close-by data points in high-density regions have similar labels



Aquiring Labels (Sampling) in Marine Biology

Aquiring Labels (Sampling) in Particle Physics



Aquiring Labels (Sampling) in Pharmacology

Key Problems

given a graph signal representation of the learning problem:

- how many labels (samples) do we need?
- which nodes should we sample ?
- what are efficient learning algorithms?

all this presupposes that we know the graph structure!

Outline

Conditional Independence Graph

- interpret ith data point as realization of random variable z_i
- associate z_i with node $i \in \mathcal{V}$ of undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- connect $i, j \in \mathcal{V}$ by undirected edge if z_i and z_j conditionally independent given $\{z_l\}_{l \in \mathcal{V} \setminus \{i,j\}}$
- $\mathcal{G}$ is the conditional independence graph (CIG) of $\{z_i\}_{i=1}^{p}$

Modelling Assumptions

- assume the z_j to be jointly Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C} \succ \mathbf{0}$
- edges of CIG $\mathcal{G}$ characterized by non-zero entries of $\mathbf{K} = \mathbf{C}^{-1}$:

$$\{i, j\} \in \mathcal{E} \Longleftrightarrow K_{i,j} \neq 0$$

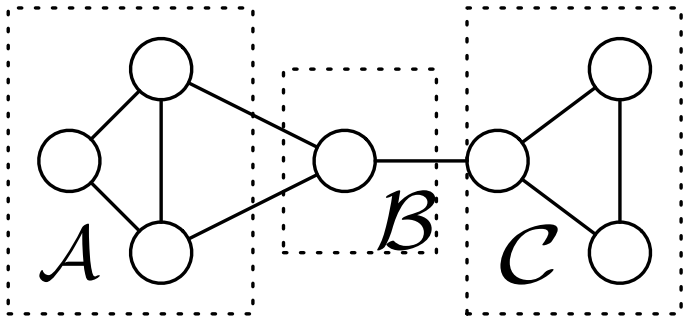
- define neighbourhood $\mathcal{N}(i) := \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$ of node i
- we assume CIG to be sparse, i.e.,

$$|\mathcal{N}(i)| \leq s_{\max}$$

  for some fixed sparsity $s_{\max}$

The Global Markov Property

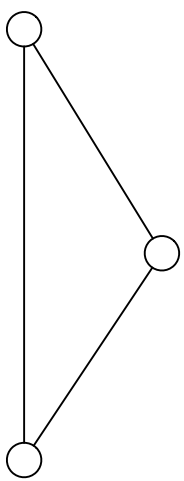
- CIG $\mathcal{G}$ allows to read off conditional independence relations



- node set $\mathcal{B}$ separates $\mathcal{A}$ from $\mathcal{C}$.
- global Markov property: $\{z_i\}_{i\in\mathcal{A}}$ conditionally independent of $\{z_i\}_{i\in\mathcal{C}}$ given $\{z_i\}_{i\in\mathcal{B}}$
- Bayes optimal estimator for z_i depends only on $\{z_j\}_{j\in\mathcal{N}(i)}$
- for sparse graphs, i.e., $|\mathcal{N}(i)| \ll p$, Bayes optimal estimator can be implemented efficiently!

The Training Data

- for each data point z_i, observe samples $x_i[1], \ldots, x_i[N]$
- stack samples into vector $\mathbf{x}[t] = (x_1[t], \ldots, x_p[t])^T$
- each vector sample $\mathbf{x}[t]$ multivariate normal $\sim \mathcal{N}(\mathbf{0}, \mathbf{C}[t])$
- graphical model selection (GMS): learn $\mathcal{G}$ from $\mathbf{x}[t]$



$$x_1[1], \ldots, x_1[N]$$

$$x_2[1], \ldots, x_2[N]$$

$$x_3[1], \ldots, x_3[N]$$

Outline

Outline

The Idea of Neighbourhood Regression

- consider the problem of finding neighbourhood $\mathcal{N}(i)$
- once we found all $\{\mathcal{N}(i)\}_{i\in\mathcal{V}}$, we have the CIG (trivial!)
- neighbourhood $\mathcal{N}(i)$ pops up in regression model for z_i:

$$z_i = \sum_{j\in\mathcal{N}(i)} a_j z_j + e_i$$

- regression coeffs $a_j = -K_{i,j}/K_{i,i}$
- error term e_i uncorrelated with $\{z_j\}_{j\in\mathcal{V}\setminus\{i\}}$
- neighbourhood $\mathcal{N}(i)$ solves sparse regression problem

$$\mathcal{N}(i) = \operatorname*{arg\,min}_{|\operatorname{supp}(\mathbf{a})|\leq s_{\max}} \mathrm{E}\{(z_i - \sum_j a_j z_j)^2\}$$

Learning based on Sparse Regression

- neighbourhood $\mathcal{N}(i)$ characterized by

$$\mathcal{N}(i) = \underset{|\,\mathrm{supp}(\mathbf{a})| \leq s_{\max}}{\arg\min} \; \mathrm{E}\{(z_i - \sum_j a_j z_j)^2\}$$

- not implementable since cannot evaluate expectation E
- consider i.i.d. samples $\mathbf{x}[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$
- replace E with sample average

$$\mathcal{N}(i) \approx \underset{|\,\mathrm{supp}(\mathbf{a})| \leq s_{\max}}{\arg\min} \; (1/N) \sum_{t=1}^{N} (x_i[t] - \sum_j a_j x_j[t])^2$$

- involves search over $\binom{p}{s_{\max}}$ subsets (more on this later!)

The Test Statistic of Sparse Regression

- for any index set $\mathcal{I} = \{i_1, \ldots, i_{s_{\max}}\} \subseteq \mathcal{V}$, define statistic

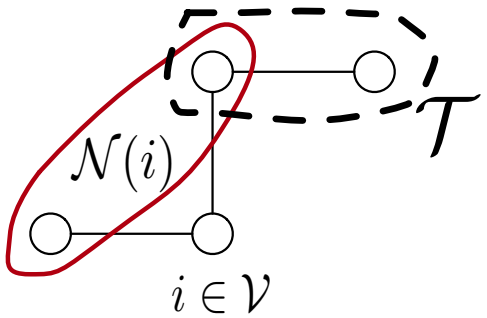
$$Z(\mathcal{I}) := \underset{|\operatorname{supp}(\mathbf{a})| \subseteq \mathcal{I}}{\arg\min} (1/N) \sum_{t=1}^{N} (x_i[t] - \sum_j a_j x_j[t])^2$$
$$= (1/N) \|\mathbf{P}_{\mathcal{I}}^{\perp} \mathbf{x}_i\|_2^2$$

- with the vector $\mathbf{x}_i = (x_i[1], \ldots, x_i[N])^T$
- projection $\mathbf{P}_{\mathcal{I}}^{\perp}$ on orthogonal complement of $\operatorname{span}\{\mathbf{x}_j\}_{j \in \mathcal{I}}$

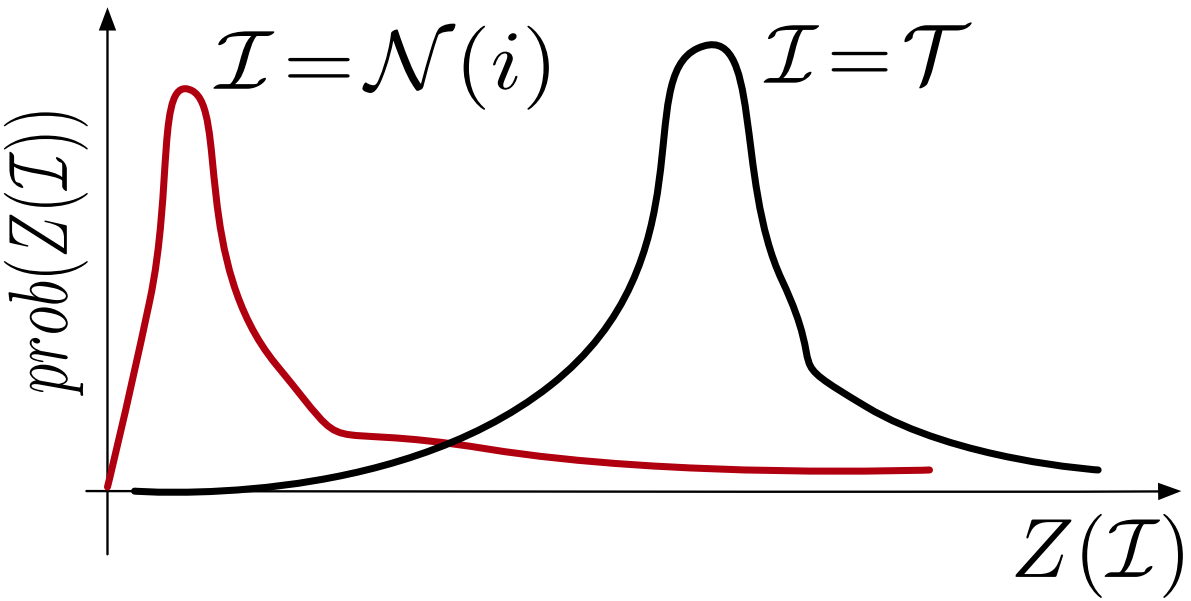
$$\mathbf{P}_{\mathcal{I}}^{\perp} := \mathbf{I} - \mathbf{X}_{\mathcal{I}} \left(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}}\right)^{-1} \mathbf{X}_{\mathcal{I}}^T$$

Pairwise Error

- consider index set $\mathcal{T}$ with $|\mathcal{T}| = s_{\max}$ and $\mathcal{N}(i) \neq \mathcal{T}$



- with high probability $Z(\mathcal{N}(i)) < Z(\mathcal{T})$ for any $\mathcal{T} \neq \mathcal{N}(i)$



Pairwise Error

- assume samples $\mathbf{x}[t]$ i.i.d. with $\sim \mathcal{N}(\mathbf{0}, \mathbf{C})$
- edges of CIG correspond to non-zero locations in $\mathbf{K} = \mathbf{C}^{-1}$
- define (inverse) condition number $\kappa := \lambda_{\min}(\mathbf{C})/\lambda_{\max}(\mathbf{C})$
- strength of edge (i, j) quantified by minimum partial correlation $\rho_{\min} := \min_{(i,j)\in\mathcal{E}} |K_{i,j}|/\sqrt{K_{i,i}K_{j,j}}$
- consider index set $\mathcal{T} \neq \mathcal{N}(i)$ with $|\mathcal{N}(i) \setminus \mathcal{T}| = d$
- probability of confusing $\mathcal{T}$ with $\mathcal{N}(i)$ upper bounded as

$$\text{Prob}\{Z(\mathcal{N}(i)) \geq Z(\mathcal{T})\} \leq 4 \exp\left(-(N - s_{\max})\frac{d\rho_{\min}^2\kappa}{64(d\rho_{\min}^2\kappa + 8)}\right)$$

Sample Complexity of GMS

- combine bound on $\mathrm{Prob}\{Z(\mathcal{N}(i)) \geq Z(\mathcal{T})\}$ with union bound over all $\mathcal{T} \neq \mathcal{N}(i)$ and another union bound over $i \in \mathcal{V}$

- accurate GMS, i.e., $\mathrm{Prob}\{\widehat{\mathcal{G}} \neq \mathcal{G}\} \to 0$ for sample size

$$N \geq s_{\max} + c_1 \max\left\{ \log \binom{p - s_{\max}}{s_{\max}}, \frac{\log(p - s_{\max})}{\kappa \rho^2_{\min}} \right\}$$

- allows for high-dimensional regime: large p, small N

How Low Can We Go?

- so far: sufficient condition (upper bound) on sample size
- what is the fundamental lower bound on sample size N?
- it can be shown that for sample size

$$N \leq c_2 \frac{\log(p - s_{\max})}{\kappa \rho_{\min}^2}$$

  ANY GMS fails with non-negligible probability

- thus, sample complexity of GMS is

$$N \propto \frac{\log(p - s_{\max})}{\kappa \rho_{\min}^2}$$

- sample complexity driven by minimum partial correlation $\rho_{\min}$

Outline

Non-IID Samples

- samples $\mathbf{x}[f] \sim \mathcal{N}(\mathbf{0}, \mathbf{C}[f])$ still independent but varying $\mathbf{C}[f]$
- model useful for two particular process classes:
  - $\mathbf{x}[f]$ are Fourier coeffs of stationary process $\mathbf{y}[t]$:

$$\mathbf{x}[f] = \sum_{t=1}^{N} \mathbf{y}[t] \exp(-(2\pi/N)(t-1)(f-1))$$

  - local cosine basis coeffs of locally stationary process $\mathbf{y}[t]$:

$$\mathbf{x}[f] = \sum_{t=1}^{N} \mathbf{y}[t] g^{(f)}[t]$$

Sample Complexity for Non-IID Samples

- RECALL sample complexity for iid case:

$$N \propto \frac{\log(p - s_{\max})}{\kappa \rho_{\min}^2}$$

- replace $\rho_{\min}$ with minimum average partial correlation

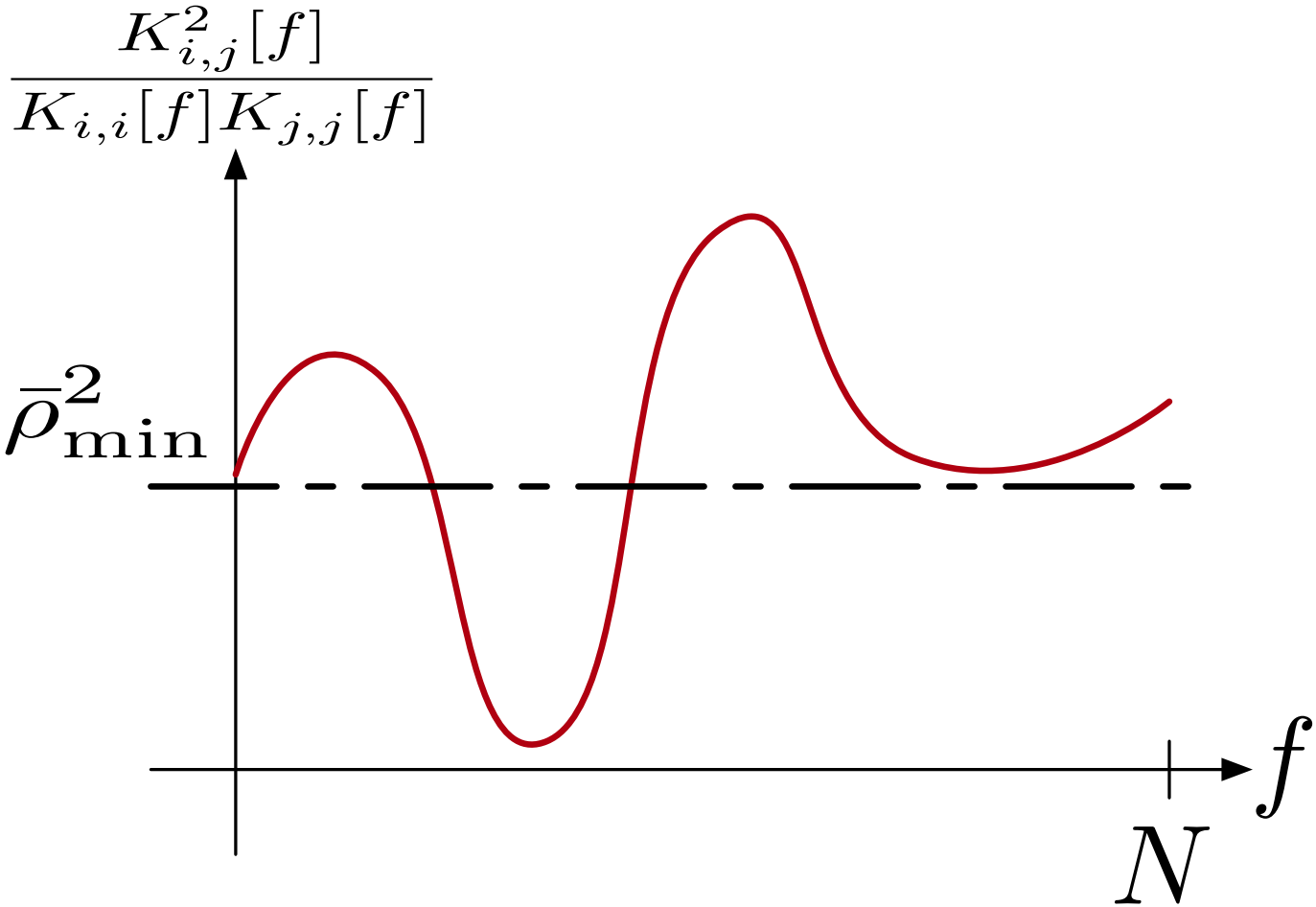
$$\bar{\rho}_{\min} := \min_{(i,j) \in \mathcal{E}} \sqrt{(1/N) \sum_{f=1}^{N} K_{i,j}^2[f] / (K_{i,i}[f] K_{j,j}[f])}$$

with $\mathbf{K}[f] := \mathbf{C}^{-1}[f]$

- sample complexity for non-iid case:

$$N \propto \frac{\log(p - s_{\max})}{\kappa \bar{\rho}_{\min}^2}$$

Average Partial Correlation



Basic Idea: Break Into Pieces

- smoothness assumption: cov. $\mathbf{C}[f] \approx$ constant over L samples
- split samples $\mathbf{x}[1], \ldots, \mathbf{x}[N]$ evenly into size-L blocks
- do sparse neighbourhood regression block-wise
- new test statistic

$$Z(\mathcal{I}) := \underset{|\operatorname{supp}(\mathbf{a}^{(b)})| \subseteq \mathcal{I}}{\arg\min} (1/N) \sum_{b=1}^{N/L} \sum_{t \in \mathrm{Block}\, b} (x_i[t] - \sum_j a_j^{(b)} x_j[t])^2$$

Outline

Just Relax

- consider (vectorized) sparse neighbourhood regression

$$\mathcal{N}(i) \approx \underset{|\operatorname{supp}(\mathbf{a})| \leq s_{\max}}{\arg\min} \|\mathbf{x}_i - \sum_j a_j \mathbf{x}_j\|_2^2$$
$$= \arg\min_{\mathbf{a}} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}\|_2^2 + \lambda |\operatorname{supp}(\mathbf{a})|$$

  with $\mathbf{X} := \{\mathbf{x}_j\}_{j \neq i}$ and some multiplier λ

- involves intractable search over $\binom{p}{s_{\max}}$ supports $\operatorname{supp}(\mathbf{a})$
- RELAX penalty $\lambda |\operatorname{supp}(\mathbf{a})|$ with convex function $\lambda \|\mathbf{a}\|_1$

The LASSO

- convex relaxation of sparse neighbourhood regression

$$\mathbf{a}_{\text{Lasso}} \in \arg\min_{\mathbf{a}} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_j\|_2^2 + \lambda\|\mathbf{a}\|_1$$

  for some suitable multiplier λ

- convex optimization problem
- (Lagrangian form) least absolute shrinkage and selection operator (LASSO)
- LASSO found widespread use in statistics/signal processing

Sample Complexity of LASSO

- consider LASSO estimator

$$\mathbf{a}_{\text{Lasso}} \in \arg\min_{\mathbf{a}} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}_j\|_2^2 + \lambda\|\mathbf{a}\|_1$$

- assume that z_i are only weakly correlated

$$|\mathrm{E}\{z_i z_j\}| \leq \sqrt{\mathrm{E}\{z_i^2\}\mathrm{E}\{z_j^2\}}/(2s_{\max})$$

- we have $\text{supp}(\mathbf{a}_{\text{Lasso}}) = \mathcal{N}(i)$ with high prob. if

$$N \geq (c_1 s_{\max} + c_2/(\kappa\rho_{\min}^2))\log(p - s_{\max})$$

- thus, for $\rho_{\min}^2 \ll 1/s_{\max}$ the LASSO is sample-size optimal!

Structure of LASSO

- consider LASSO

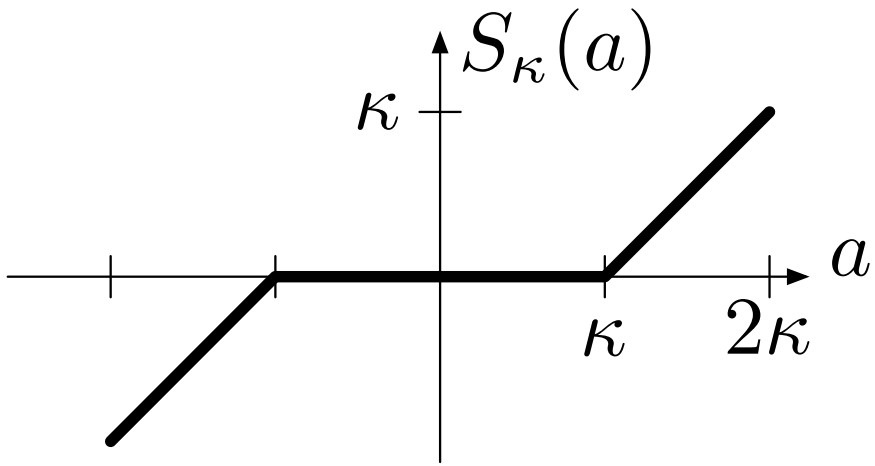
$$\arg\min_{\mathbf{a}} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1$$

- sum of smooth term $\|\mathbf{x}_i - \mathbf{X}\mathbf{a}\|_2^2$ and non-smooth $\|\mathbf{a}\|_1$
- highly developed methods around for such problems
- e.g., proximal gradient method, alternating direction method of multipliers (ADMM), Pock-Chambolle primal-dual, etc....
- basic idea: splitting of minimization of the two terms

ADMM for LASSO

- LASSO: $\mathbf{a}_{\text{Lasso}} \in \arg\min_{\mathbf{a}=\mathbf{b}} \|\mathbf{x}_i - \mathbf{X}\mathbf{a}\|_2^2 + \lambda\|\mathbf{b}\|_1$
- ADMM amounts to iterating (for some $\rho > 0$)

$$\mathbf{a}^{(k+1)} = \left(\mathbf{X}^T\mathbf{X} + \rho\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{x}_i + \rho(\mathbf{b}^{(k)} - \mathbf{c}^{(k)})\right)$$
$$\mathbf{b}^{(k+1)} = S_{\lambda/\rho}(\mathbf{a}^{(k+1)} + \mathbf{c}^{(k)})$$
$$\mathbf{c}^{(k+1)} = \mathbf{c}^{(k)} + \mathbf{a}^{(k+1)} - \mathbf{b}^{(k+1)}$$



- convergence under very general conditions

Further Reading at Arxiv

- A. Jung, "Sparse Label Propagation", Dec. 2016.
- A. Jung, "Learning the Conditional Independence Structure of Stationary Time Series: A Multitask Learning Approach", Jan. 2015.
- N. Tran Quang and A. Jung, "Learning conditional independence structure for high-dimensional uncorrelated vector processes", Sep. 2016.
- G. Hannak and A. Jung and N. Goertz, "On the Information-theoretic Limits of Graphical Model Selection for Gaussian Time Series", Mar. 2014.

Thank You for Your Attention

Frohe Weihnachten

und einen

Guten Rutsch!