ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Candidacy exam

Rodrigo C. G. Pena

Advisor: Prof. Pierre Vandergheynst

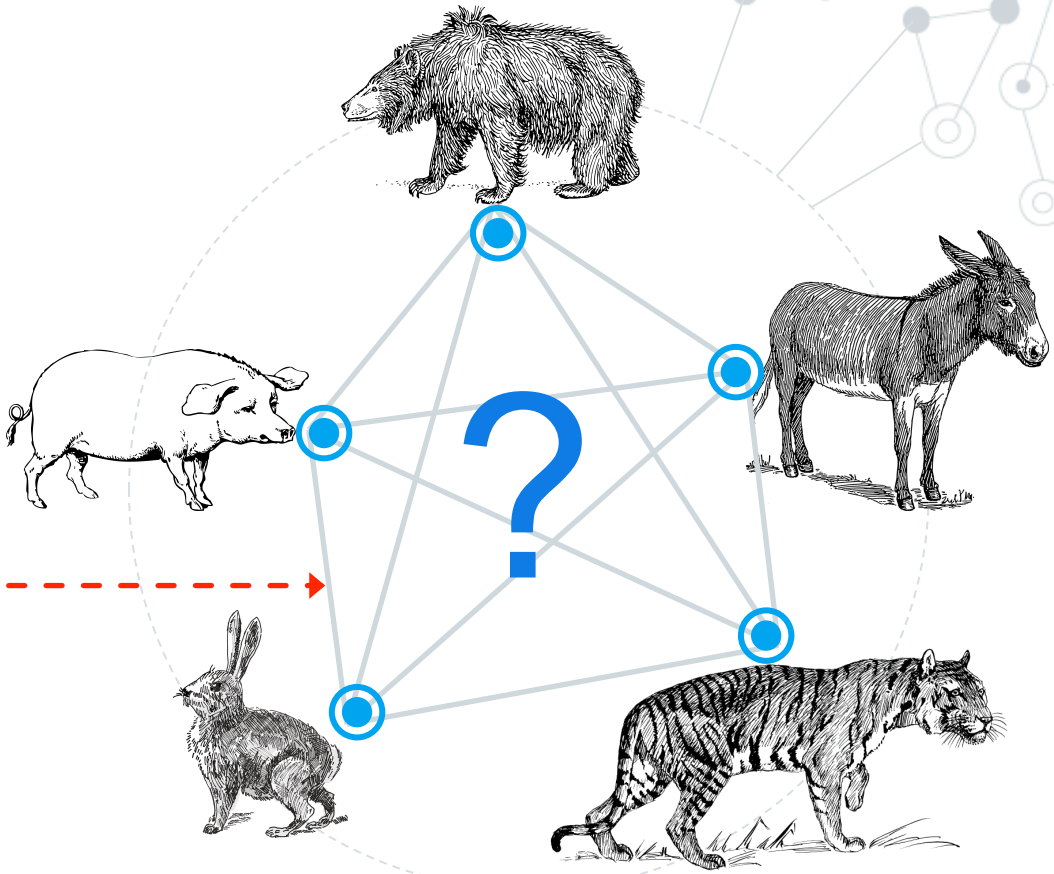
05 Aug 2016

Graph learning perspectives in statistics and signal processing

Motivation

- Graph structure is useful
  - Not always available
- Simple and popular approaches
  - Gaussian kernel
  - *k*-NN

$$W_{ij} = \exp\left(\frac{-\|x_i - x_j\|_2^2}{2\sigma^2}\right)$$



Notation

◎ Data:

$$X = (x_1|...|x_N)^T \in \mathbb{R}^{N \times d}$$

◎ Graph:

$$G = (\mathcal{V}, \mathcal{E}, \mathcal{W})$$

◎ Combinatorial graph Laplacian:

$$L = D - W = U \Lambda U^T$$

◎ Normalized graph Laplacian:

$$L_n = D^{-1/2} L D^{-1/2}$$

◎ Graph filtering

$$z = g(L)z_0 = U g(\Lambda) U^T z_0 = U g(\lambda) \hat{z_0}$$

1.

Model selection for Gaussian and binary data with sparsity prior

[Banerjee *et al.*, 2008]

Model selection via MAP estimation

Select model that best explains the data, while enforcing a prior.

$$\begin{aligned}\tilde{M} &= \arg\max_{M} \quad p(M|X) \\ &= \arg\max_{M} \quad p(X|M)p(M) \\ &= \arg\max_{M} \quad \log p(X|M) + \log p(M)\end{aligned}$$

For [Banerjee *et al.*, 2008]: Gaussian variables, sparsity prior.

$$\tilde{\Sigma}^{-1} = \arg\max_{M \succ 0} \quad \log\det M - tr(SM) - \lambda \|M\|_{1,1}$$

◎ Justified choice of lambda.

Dual problem and BCD

Use definition of dual norm, and Sion's min-max theorem

$$\tilde{\Sigma} = \arg\max_{W} \quad \log\det W$$
$$\text{subject to} \quad \|W - S\|_\infty \leq \lambda.$$

Algorithm 1 Block Coordinate Descent (BCD) for solving 4

Input: S, λ, ϵ
Output: $\tilde{\Sigma}$

1: $W \leftarrow S + \lambda I$
2: **while** $tr(W^{-1}S) - N + \lambda\|W^{-1}\|_{1,1} > \epsilon$ **do**
3: **for** $j = 1$ to N **do**
4: Solve the problem

$$\tilde{w} \leftarrow \arg\min_{y} \{y^T (W_{\setminus j \setminus j})^{-1} y : \|y - s_j\|_\infty \leq \lambda\} \quad (5)$$

5: Replace j-th column (resp. row) of W with $\tilde{w}$ (resp. $\tilde{w}^T$)
6: **end for**
7: **end while**
8: $\tilde{\Sigma} \leftarrow W$

Restriction to the free row/column

Complexity: $O(N^4)$, per sweep.
Faster alternatives exist, e.g., glasso, QUIC.

Binary data

Ising model:

$$p(x|\theta) = \exp\left[\sum_{i=1}^{N}\theta_i x_i + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\theta_{ij}x_i x_j - A(\theta)\right],$$

Model matrix is now

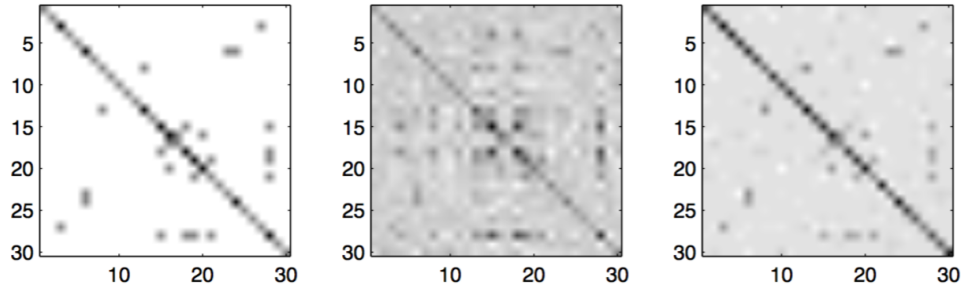
$$\Theta = \begin{bmatrix} 0 & \theta_1 & \theta_2 & \dots & \theta_N \\ \theta_1 & 0 & \theta_{12} & \dots & \theta_{1N} \\ \theta_2 & \theta_{12} & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \theta_N & \theta_{1N} & \dots & \dots & 0 \end{bmatrix}$$

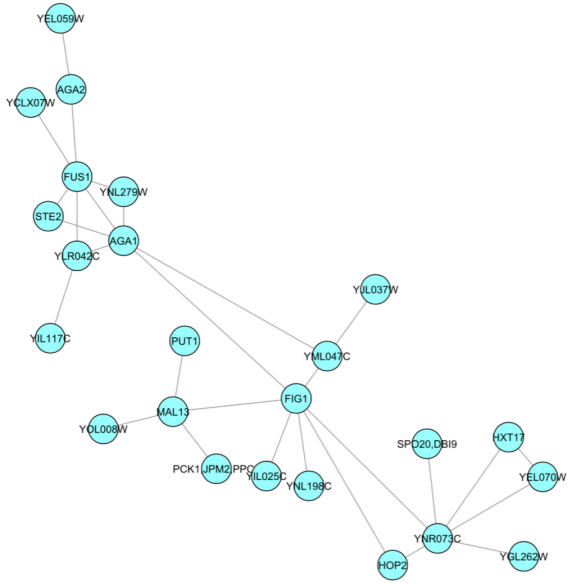
Same algorithm as before, with different initialization

Results

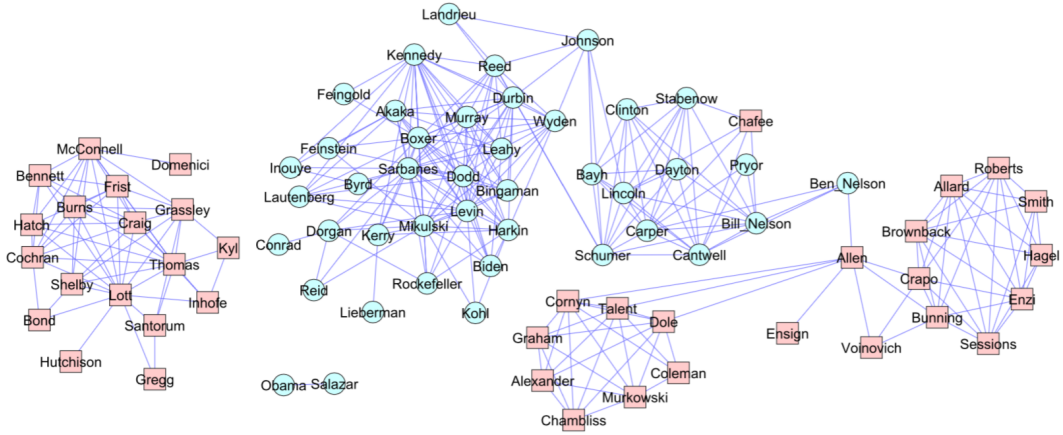
Simulations

Σ^{-1} S^{-1} $\tilde{\Sigma}^{-1}$





Gene expression



US senate voting data

2.

How to learn graphs from smooth signals

[Kalofolias, 2016]

Smoothness as a prior

Smoothness as measure of local differences, with L = D - W.

$$(1/2)\sum_{i,j} W_{ij}\|x_i - x_j\|_2^2 = tr(X^T L X)$$

Used in graph-regularized signal denoising/recovery.

Complementary problem:

$$\min_{L\in\mathcal{L}} \quad tr(X^T L X) + f(L)$$

Smoothness of the signals = graph sparsity prior

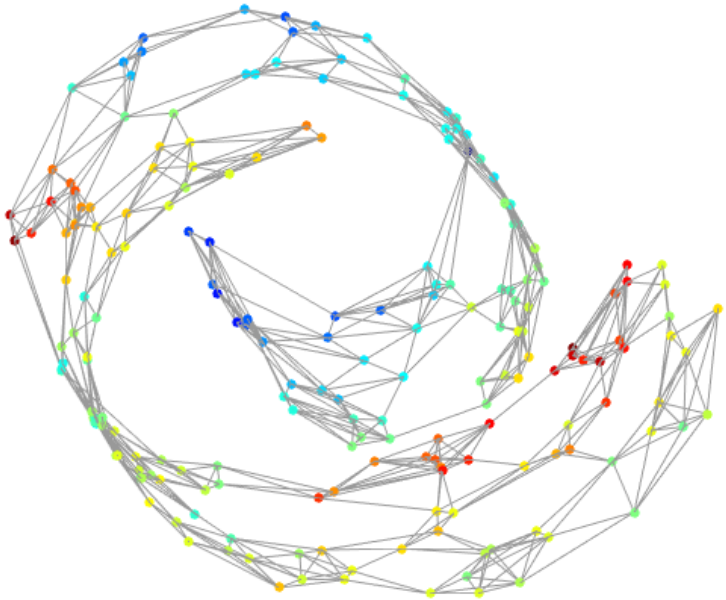
$$tr(X^T L X) = \frac{1}{2}\|W \odot Z\|_{1,1}$$

$$Z_{ij} = \|x_i - x_j\|_2^2$$

Smooth signals & graph filtering

Examples in the literature:

$$\begin{aligned}
\text{Tikhonov regularization:} \quad & z = (\alpha L + I)^{-1} z_0 \\
\text{[Dong } et \text{ } al.\text{, 2015]:} \quad & z = \sqrt{L^\dagger} z_0,\ z_0 \sim \mathcal{N}(0, I) \\
\text{Heat diffusion:} \quad & z = \exp(-tL) z_0
\end{aligned}$$



Degree prior & algorithm

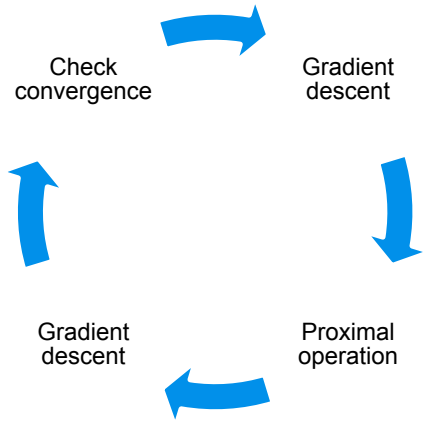
Problem to be solved: $\min_{W \in \mathcal{W}_m} \quad \|W \odot Z\|_{1,1} - \alpha \mathbf{1}^T \log(W\mathbf{1}) + \beta \|W\|_F^2,$

Free variables: lower triangle of W. Vectorize it!

$$\min_{w \in \mathbb{R}^{N(N-1)/2}} \quad \mathbb{I}_{\{w \succeq 0\}}(w) + 2w^T z - \alpha \mathbf{1}^T \log(d) + \beta \|w\|_2^2,$$

Forward-backward-forward primal-dual algorithm

◎ Close the primal-dual gap.



"state-of-the-art"

◎ [Dong *et al.*, 2015] put on the same framework

$$\min_{W \in \mathcal{W}_m} \quad \|W \odot Z\|_{1,1} - \alpha \|W\mathbf{1}\|_2^2 + \alpha \|W\|_F^2$$
$$\text{s. t.} \quad \|W\|_{1,1} = s$$

◎ Complexity: $O(N^2)$ per iteration

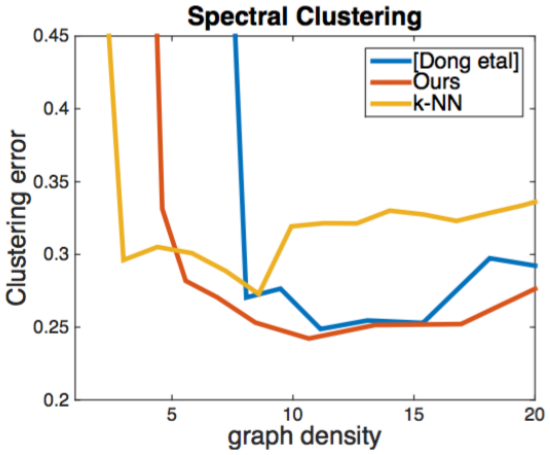
Results

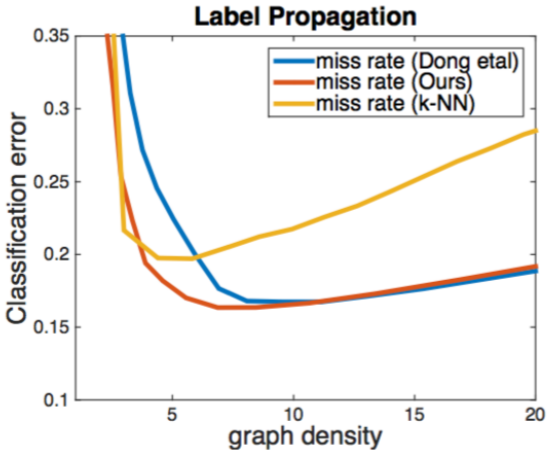
Simulations

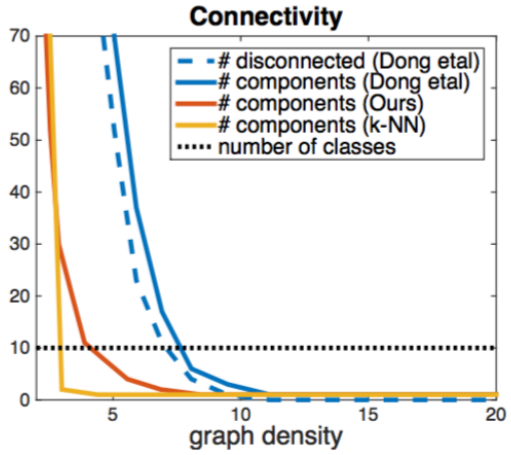
- 4 graph types: Uniform and Non-uniform Random Geometric, Erdös-Rényi, and Barabási-Albert.
- 3 smooth signal models seen before.
- Smaller errors than baseline and state-of-the-art.

Real data

- Spectral clustering and label propagation to measure performance.







3.

Compressive spectral clustering

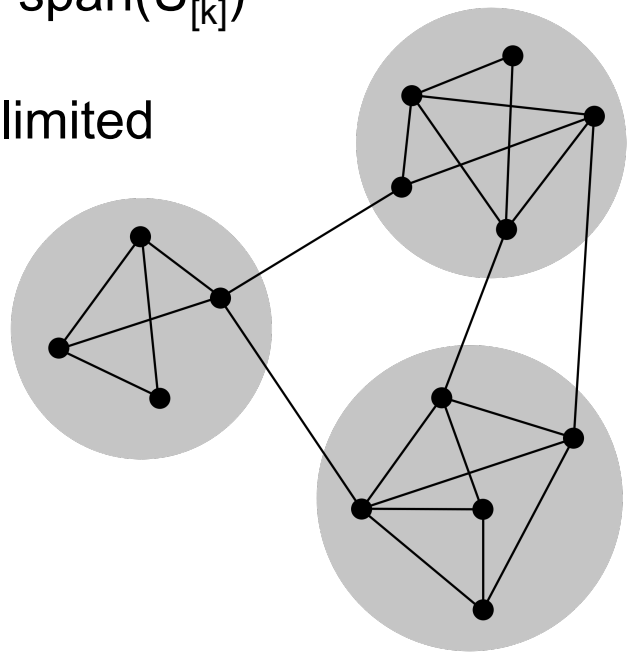
[Tremblay *et al.*, 2016]

Indicator vectors

Objective is to estimate each of the *k* cluster indicator vectors

$$(c_j)_i := \begin{cases} 1 & \text{if} \quad i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases}$$

- No inter-cluster connection: c_j's live on span($U_{[k]}$)
- Assumption: c_j's live close span($U_{[k]}$)
- GSP parlance: c_j's are close to *k*-bandlimited



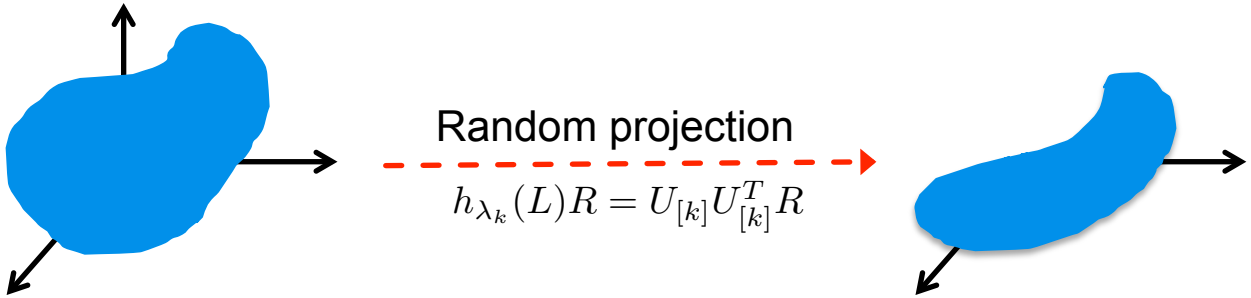
Spectral clustering

In a nutshell…

Eigendecomposition bottleneck

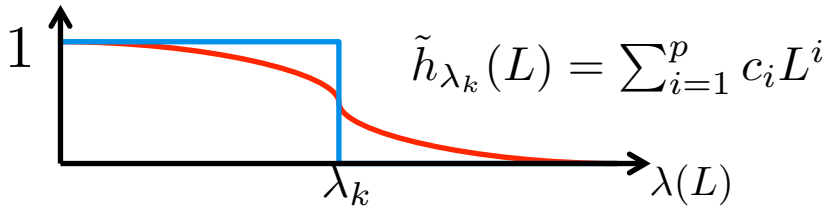
Fast filter O(*log N*) random Gaussian signals

◎ Random projections



◎ Johnson-Lindenstrauss: distances are preserved w.h.p.

◎ Fast filtering: Jackson-Chebyshev polynomials

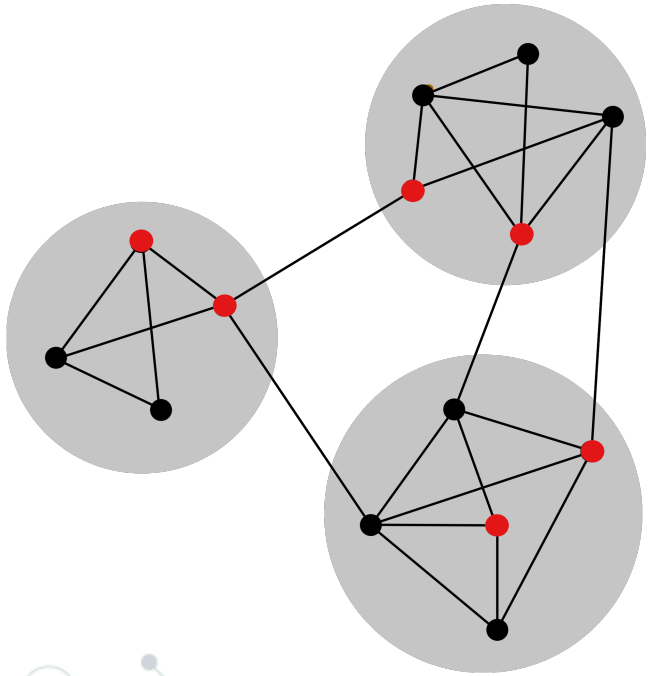


k-means bottleneck

Sample randomly O(*k log k*) features

◎ Sampling matrix M satisfies RIP

$$\tilde{c}_j \leftarrow \arg\min_y \|My - \tilde{c}_j^r\|_2^2 + \gamma y^T (1 - \tilde{H}_{\lambda_k}) y$$



Results

Complexity:

- CSC: $O(k^2 \log^2 k + pN(\log N + k))$
- SC: $O(Nk^2 + k^3)$

Simulations: Stochastic Block Model (SBM)

- CSC slightly less accurate than SC
- Huge gain in speed for large k

Real data: Amazon co-purchasing network

- CPU time for fixed modularity
- CSC much faster for large k

4. Research directions

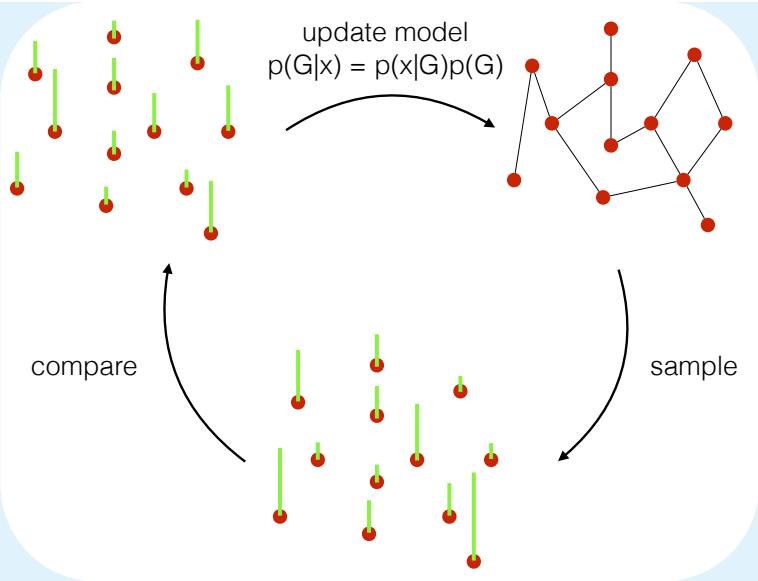
Model and samples

Kalofolias's problem as MAP estimation.

◎ Smoothness closely related to Gaussianity.

Idea: enforce instead sparse dictionary representation.

$$\min_{W,Y} \quad \|X - g(L)Y\|_F^2 + \gamma\|Y\|_{1,1} - \alpha \mathbf{1}^T \log(W\mathbf{1}) + \beta\|W\|_F^2.$$



◎ Sample signals from graphical model.

◎ "Pattern theory": analysis by synthesis.

Thank you

Q&A time