Lecture 10

Learning Graphs from Data
Introduction

• What is the problem of graph learning?
Introduction

• What is the problem of graph learning?
  - Given observations on a number of variables and some prior knowledge (distribution, model, etc)
Introduction

• What is the problem of graph learning?
  - Given observations on a number of variables and some prior knowledge (distribution, model, etc)
  - Build/learn a measure of relations between variables (correlation/covariance, graph topology/operator or equivalent)
Introduction

- Why is it important?
  - Learning relations between entities benefits numerous application domains
Introduction

• Why is it important?
  - Learning **relations between entities** benefits numerous application domains

**Objective:** functional connectivity between brain regions

**Input:** fMRI recordings in these regions

**Objective:** behavioural similarity/influence between people

**Input:** history of individual activities

image credit:
http://blog.myesr.org/mri-reveals-the-human-connectome/
https://www.iconexperience.com
Introduction

• Why is it important?
  - Learning relations between entities benefits numerous application domains
  - The learned relations can help us predict future observations

**Objective:** functional connectivity between brain regions

**Input:** fMRI recordings in these regions

**Objective:** behavioural similarity/influence between people

**Input:** history of individual activities

image credit:
http://blog.myesr.org/mri-reveals-the-human-connectome/
https://www.iconexperience.com
Introduction

Why is it important?
- Learning relations between entities benefits numerous application domains
- The learned relations can help us predict future observations

Objective: functional connectivity between brain regions
Input: fMRI recordings in these regions

Objective: behavioural similarity/influence between people
Input: history of individual activities

How do we build/learn the graph?

image credit:
http://blog.myesr.org/mri-reveals-the-human-connectome/
https://www.iconexperience.com
Outline

• A (very partial) literature review
• A signal processing perspective
• Concluding remarks
Outline

• A (very partial) literature review

• A signal processing perspective

• Concluding remarks
A (very partial) literature review

- Simple and intuitive methods
  - Sample correlation
  - Similarity function (e.g., Gaussian RBF)
A (very partial) literature review

• Simple and intuitive methods
  - Sample correlation
  - Similarity function (e.g., Gaussian RBF)

• Need for a meaningful data model: $\mathbf{X} \sim \mathcal{F}(\mathcal{G})$
A (very partial) literature review

• Simple and intuitive methods
  - Sample correlation
  - Similarity function (e.g., Gaussian RBF)

• Need for a meaningful data model: $\mathbf{X} \sim \mathcal{F}(\mathcal{G})$

• Two general approaches in the literature
  - Statistical models: $\mathbf{F}$ represents a distribution determined by $\mathbf{G}$ (e.g., probabilistic graphical models)
A (very partial) literature review

• Simple and intuitive methods
  - Sample correlation
  - Similarity function (e.g., Gaussian RBF)

• Need for a meaningful data model: \( \mathbf{X} \sim \mathcal{F}(\mathcal{G}) \)

• Two general approaches in the literature
  - Statistical models: \( \mathbf{F} \) represents a distribution determined by \( \mathbf{G} \) (e.g., probabilistic graphical models)
  - Physically-motivated models: \( \mathbf{F} \) represents a physical generative model on \( \mathbf{G} \) (e.g., diffusion process on graphs)
A (very partial) literature review

- Learning graphical models

Undirected graphical models: Markov random fields (MRF)

Directed graphical models: Bayesian networks (BN)

Factor graphs
A (very partial) literature review

- Learning graphical models

- Directed graphical models: Bayesian networks (BN)
- Undirected graphical models: Markov random fields (MRF)
- Factor graphs
A (very partial) literature review

- Learning pairwise MRF
A (very partial) literature review

- Learning pairwise MRF

conditional independence:

$$(v_i, v_j) \notin \mathcal{E} \iff x_i \independent x_j \mid \mathbf{x} \setminus \{x_i, x_j\}$$
A (very partial) literature review

• Learning pairwise MRF

conditional independence:

\[(v_i, v_j) \notin \mathcal{E} \iff x_i \perp x_j \mid \mathbf{x} \setminus \{x_i, x_j\}\]

probability parameterised by \(\Theta\):

\[
P(\mathbf{x}|\Theta) = \frac{1}{Z(\Theta)} \exp\left( \sum_{v_i \in \mathcal{V}} \theta_{i,i} x_i^2 + \sum_{(v_i,v_j) \in \mathcal{E}} \theta_{i,j} x_i x_j \right)
\]
A (very partial) literature review

- Learning pairwise MRF

conditional independence:

\((v_i, v_j) \notin \mathcal{E} \Leftrightarrow x_i \perp x_j \mid x \setminus \{x_i, x_j\}\)

probability parameterised by \(\Theta\):

\[
P(x|\Theta) = \frac{1}{Z(\Theta)} \exp\left(\sum_{v_i \in \mathcal{V}} \theta_{i,i} x_i^2 + \sum_{(v_i,v_j) \in \mathcal{E}} \theta_{i,j} x_i x_j\right)
\]

Gaussian MRF with precision \(\Theta\):

\[
P(x|\Theta) = \frac{|\Theta|^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2} x^T \Theta x\right)
\]
A (very partial) literature review

- Learning pairwise MRF

conditional independence:

\[(v_i, v_j) \notin \mathcal{E} \iff x_i \perp x_j \mid x \setminus \{x_i, x_j\}\]

probability parameterised by \(\Theta\):

\[
P(x \mid \Theta) = \frac{1}{Z(\Theta)} \exp\left( \sum_{v_i \in V} \theta_{i,i} x_i^2 + \sum_{(v_i, v_j) \in \mathcal{E}} \theta_{i,j} x_i x_j \right)
\]

Gaussian MRF with precision \(\Theta\):

\[
P(x \mid \Theta) = \frac{|\Theta|^{1/2}}{(2\pi)^{N/2}} \exp\left( -\frac{1}{2} x^T \Theta x \right)
\]

Learning a sparse \(\Theta\):

- interactions are mostly local
- computationally more tractable
A (very partial) literature review

Prune the smallest elements in sample precision (inverse covariance) matrix
Prune the smallest elements in sample precision (inverse covariance) matrix

\[ \Theta \]  
ground-truth precision

\[ X \sim \mathcal{N}(0, \Theta) \]  
data matrix

\[ S^{-1} \]  
inverse of sample covariance
A (very partial) literature review

Learning a graph = learning neighbourhood of each node
A (very partial) literature review

Learning a graph = learning neighbourhood of each node
A (very partial) literature review

covariance selection

$\ell_1$-regularised neighbourhood regression

Dempster Meinshausen & Buhlmann

Learning a graph = learning neighbourhood of each node

LASSO regression: $\min_{\beta_1} ||X_1 - X_{\setminus 1}\beta_1||^2 + \lambda||\beta_1||_1$
A (very partial) literature review

Learning a graph = learning neighbourhood of each node

\[
\text{LASSO regression: } \min_{\beta_1} \left\| X_1 - X_1 \beta_1 \right\|^2 + \lambda \left\| \beta_1 \right\|_1
\]

Logistic regression for discrete variables
A (very partial) literature review

- Covariance selection
- $\ell_1$-regularised neighbourhood regression
- $\ell_1$-regularised log-determinant
- $\ell_1$-regularised logistic regression

- Dempster
- Meinshausen & Buhlmann
- Banerjee
- Friedman
- Ravikumar

Estimation of sparse precision matrix

$X$, $S$
A (very partial) literature review

- **covariance selection**
  - Dempster

- **\( \ell_1 \)-regularised neighbourhood regression**
  - Meinshausen & Buhlmann
  - Banerjee Friedman

- **\( \ell_1 \)-regularised log-determinant**

- **\( \ell_1 \)-regularised logistic regression**
  - Ravikumar

1972 2006 2008 2010

**Estimation of sparse precision matrix**

graphical LASSO maximises likelihood of precision matrix \( \Theta \):

\[
|\Theta|^{M/2} \exp(-\sum_{m=1}^{M} \frac{1}{2} x(m)^T \Theta x(m))
\]
A (very partial) literature review

covariance selection

\( \ell_1 \)-regularised
neighbourhood regression

\( \ell_1 \)-regularised
log-determinant

\( \ell_1 \)-regularised
logistic regression

1972

2006

2008

2010

Dempster

Meinshausen & Buhlmann

Banerjee Friedman

Ravikumar

Estimation of sparse precision matrix

graphical LASSO maximises likelihood of precision matrix \( \Theta \):

\[
\max_{\Theta} \log \det \Theta - \text{tr}(S\Theta) - \rho \| \Theta \|_1
\]

log-likelihood function
A (very partial) literature review

- Covariance selection
- \( \ell_1 \)-regularised neighbourhood regression
- \( \ell_1 \)-regularised log-determinant
- \( \ell_1 \)-regularised logistic regression
- Quadratic approx. of Gauss. neg. log-likelihood

- Estimation of sparse precision matrix

- Graphical LASSO maximises likelihood of precision matrix \( \Theta \):

\[
\max_{\Theta} \log \det \Theta - \text{tr}(S\Theta) - \rho \| \Theta \|_1
\]

- Log-likelihood function
A (very partial) literature review

- Learning graphical models
  - Classical approaches lead to both positive/negative relations
  - Learning graphs with non-negative weights?
A (very partial) literature review

• Learning graphical models
  - Classical approaches lead to both positive/negative relations
  - Learning graphs with non-negative weights?

• Learning graphs with non-negative weights
  - M-matrices (sym., p.d., non-pos. off-diag.) have been used as precision, leading to attractive GMRF (Slawski and Hein 2015)
A (very partial) literature review

• Learning graphical models
  - Classical approaches lead to both positive/negative relations
  - Learning graphs with non-negative weights?

• Learning graphs with non-negative weights
  - M-matrices (sym., p.d., non-pos. off-diag.) have been used as precision, leading to attractive GMRF (Slawski and Hein 2015)
  - The combinatorial graph Laplacian $L$ belongs to M-matrices and is equivalent to graph topology
A (very partial) literature review

From arbitrary precision matrix to graph Laplacian
A (very partial) literature review

From arbitrary precision matrix to graph Laplacian
Common setting in graph signal processing (GSP)
Outline

• A (very partial) literature review

• A signal processing perspective

• Concluding remarks
GSP viewpoint for graph learning
GSP viewpoint for graph learning

$G_1$  $G_2$  $G_3$

Which graph to choose?
GSP viewpoint for graph learning

Which graph to choose?
- depends on the signal/graph model
- idea: choose one that enforces certain signal characteristics
Why GSP for graph learning?

- Existing approaches have limitations
  - Simple correlation/similarity functions are not enough
  - Statistical models often do not lead to non-negative edge weights
  - Many impose a “global” distribution or behaviour
Why GSP for graph learning?

• Existing approaches have limitations
  - Simple correlation/similarity functions are not enough
  - Statistical models often do not lead to non-negative edge weights
  - Many impose a “global” distribution or behaviour

• Opportunity and challenge for GSP
  - GSP tools such as frequency analysis and filtering offer another “regulariser” for learning graphs
  - Filtering-based approaches can provide generative models for signals with complex (non-Gaussian) behaviour
GSP for graph learning

- Signal processing is about $F \mathbf{c} = \mathbf{x}$
GSP for graph learning

- Graph signal processing is about $\mathbf{F}(\mathbf{G}) \mathbf{c} = \mathbf{x}$
GSP for graph learning

- Forward: Given $\mathbf{G}$ and $\mathbf{x}$, design $\mathbf{F}$ to study $\mathbf{c}$

$\mathbf{F}(\mathbf{G}) \times \mathbf{c} = \mathbf{x}$

- Fourier/wavelet atoms
- Graph Fourier/wavelet coefficient
- Trained dictionary atoms
- Graph dictionary coefficient

References:

- [Coifman06, Narang09, Hammond11, Shuman13, Sandryhaila13]
- [Zhang12, Thanou14]
GSP for graph learning

- Backward (graph learning): Given $\mathbf{x}$, design $\mathbf{F}$ and $\mathbf{c}$ to infer $\mathbf{G}$

\[
\mathcal{F}(\mathcal{G}) \times \mathbf{c} = \mathbf{x}
\]

$\mathbf{G}$
GSP for graph learning

- Backward (graph learning): Given $x$, design $F$ and $c$ to infer $G$

The key is a signal/graph model behind $F$
- Designed via graph operators (adjacency/Laplacian matrices, graph shift operators)
GSP for graph learning

- Backward (graph learning): Given $\mathbf{x}$, design $\mathbf{F}$ and $\mathbf{c}$ to infer $\mathbf{G}$

- The key is a signal/graph model behind $\mathbf{F}$
- Designed via graph operators (adjacency/Laplacian matrices, graph shift operators)
- Choice of/assumption on $\mathbf{c}$ often determines signal characteristics
Model 1: Global smoothness

- Signal values vary smoothly between all pairs of nodes that are connected
- Example: Temperature of different locations in a flat geographical region
- Usually quantified by the Laplacian quadratic form:

\[ x^T L x = \frac{1}{2} \sum_{i,j} W_{ij} (x(i) - x(j))^2 \]
Model 1: Global smoothness

- Signal values vary smoothly between all pairs of nodes that are connected
- Example: Temperature of different locations in a flat geographical region
- Usually quantified by the Laplacian quadratic form:

\[ x^T L x = \frac{1}{2} \sum_{i,j} W_{ij} (x(i) - x(j))^2 \]
Model 1: Global smoothness

- Signal values vary smoothly between all pairs of nodes that are connected
- Example: Temperature of different locations in a flat geographical region
- Usually quantified by the Laplacian quadratic form:

\[
x^T L x = \frac{1}{2} \sum_{i,j} W_{ij} (x(i) - x(j))^2
\]

Similar to previous approaches:

Lake (2010):

\[
\max_{\Theta = L + \frac{1}{\sigma^2} I} \log \det \Theta - \frac{1}{M} \text{tr}(XX^T \Theta) - \rho \|\Theta\|_1
\]

Daitch (2009):

\[
\min_L X^T L^2 X
\]

Hu (2013):

\[
\min_L \text{tr}(X^T L^s X) - \beta \|W\|_F
\]
Model 1: Global smoothness

- Dong et al. (2016) & Kalofolias (2016)
  - $\mathcal{F}(\mathcal{G}) = \chi$ (eigenvector matrix of $\mathbf{L}$)
  - Gaussian assumption on $\mathbf{c}$: $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, \Lambda)$
  - $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{L}^\dagger + \sigma^2 \mathbf{I})$ \textbf{Gaussian Markov Random Field}
Model 1: Global smoothness

- Dong et al. (2016) & Kalofolias (2016)

  - \( \mathcal{F}(\mathcal{G}) = \chi \) (eigenvector matrix of \( L \))
  - Gaussian assumption on \( c \): \( c \sim \mathcal{N}(0, \Lambda) \)
  - \( x \sim \mathcal{N}(0, L^\dagger + \sigma_c^2 I) \) \[ \text{Gaussian Markov Random Field} \]
  - Maximum a posteriori (MAP) estimation on \( c \) leads to minimisation of Laplacian quadratic form:

    \[
    \min_c | | x - \chi c | |_2^2 - \log P_c(c)
    \]
Model 1: Global smoothness

- Dong et al. (2016) & Kalofolias (2016)

  \[ \mathcal{F}(G) = \chi \quad \text{(eigenvector matrix of } L) \]

- Gaussian assumption on \( c \): \( c \sim \mathcal{N}(0, \Lambda) \)

- \( x \sim \mathcal{N}(0, L^\dagger + \sigma_\epsilon^2 I) \) \quad \text{Gaussian Markov Random Field}

- Maximum a posteriori (MAP) estimation on \( c \) leads to minimisation of Laplacian quadratic form:

  \[
  \min_c ||x - \chi c||_2^2 + \alpha \ c^T \Lambda c
  \]
Model 1: Global smoothness

- **Dong et al. (2016) & Kalofolias (2016)**

  - $\mathcal{F}(\mathcal{G}) = \chi$ (eigenvector matrix of $L$)
  
  - Gaussian assumption on $c$: $c \sim \mathcal{N}(0, \Lambda)$
  
  - $x \sim \mathcal{N}(0, L^\dagger + \sigma_\epsilon^2 I)$ \textbf{Gaussian Markov Random Field}

- Maximum a posteriori (MAP) estimation on $c$ leads to minimisation of Laplacian quadratic form:

  $$\min_c \|x - \chi c\|^2_2 + \alpha c^T \Lambda c$$

  $$\min_{L, Y} \|X - Y\|^2_F + \alpha \text{tr}(Y^TLY) + \beta \|L\|^2_F$$

  \text{data fidelity} \quad \text{smoothness on } Y \quad \text{regularisation}
Model 1: Global smoothness

- Dong et al. (2016) & Kalofolias (2016)
  
  - $\mathcal{F}(\mathcal{G}) = \chi$ (eigenvector matrix of $L$)
  - Gaussian assumption on $c$: $c \sim \mathcal{N}(0, \Lambda)$
  - $x \sim \mathcal{N}(0, L^\dagger + \sigma^2 \varepsilon I)$ Gaussian Markov Random Field
  - Maximum a posteriori (MAP) estimation on $c$ leads to minimisation of Laplacian quadratic form:

$$\min_c ||x - \chi c||^2_2 + \alpha c^T \Lambda c$$

$$\min_{L, Y} ||X - Y||^2_F + \alpha \text{tr}(Y^TLY) + \beta ||L||^2_F$$

- data fidelity 
- smoothness on $Y$ 
- regularisation
Model 1: Global smoothness

- Dong et al. (2016) & Kalofolias (2016)
  - $F(G) = \chi$ (eigenvector matrix of $L$)
  - Gaussian assumption on $c$: $c \sim \mathcal{N}(0, \Lambda)$
  - $x \sim \mathcal{N}(0, L^\dagger + \sigma^2_\epsilon I)$ Gaussian Markov Random Field
  - Maximum a posteriori (MAP) estimation on $c$ leads to minimisation of Laplacian quadratic form:

\[
\min_c ||x - \chi c||_2^2 + \alpha c^T \Lambda c
\]

\[
\min_{L,Y} ||X - Y||_F^2 + \alpha \text{tr}(Y^TLY) + \beta ||L||_F^2.
\]

data fidelity  smoothness on $Y$  regularisation
Model 1: Global smoothness

- Dong et al. (2016) & Kalofolias (2016)
  - $\mathcal{F}(\mathcal{G}) = \chi$ (eigenvector matrix of $L$)
  - Gaussian assumption on $c$: $c \sim \mathcal{N}(0, \Lambda)$
  - $x \sim \mathcal{N}(0, L^\dagger + \sigma^2 I)$  
    Gaussian Markov Random Field
  - Maximum a posteriori (MAP) estimation on $c$ leads to minimisation of Laplacian quadratic form:

$$\min_c \|x - \chi c\|^2_2 + \alpha c^T \Lambda c$$

$$\min_{L,Y} \|X - Y\|^2_F + \alpha \text{tr}(Y^T LY) + \beta \|L\|^2_F$$

data fidelity  smoothness on $Y$  regularisation

Learning enforces signal property (global smoothness)!
Model 2: Spectral filtering

• Signals are outcome of applying filtering to latent (input) signals
• The filtering often corresponds to a diffusion process on graphs (different spectral characteristics or localisation properties)
• Example: Movement of people/vehicles in transportation network
Model 2: Spectral filtering

- Signals are outcome of applying filtering to latent (input) signals
- The filtering often corresponds to a diffusion process on graphs (different spectral characteristics or localisation properties)
- Example: Movement of people/vehicles in transportation network

![Diagram showing a transportation network with nodes and edges, labeled v1, v2, ..., v9, with initial stage highlighted]
Model 2: Spectral filtering

- Signals are outcome of applying filtering to latent (input) signals
- The filtering often corresponds to a diffusion process on graphs (different spectral characteristics or localisation properties)
- Example: Movement of people/vehicles in transportation network
Model 2: Spectral filtering

- Signals are outcome of applying filtering to latent (input) signals
- The filtering often corresponds to a diffusion process on graphs (different spectral characteristics or localisation properties)
- Example: Movement of people/vehicles in transportation network

![Diagram of heat diffusion and graph shift operator](image-url)
Model 2: Spectral filtering

- Thanou et al. (2017)
  - $\mathcal{F}(\mathcal{G}) = e^{-\tau L}$ (localisation in vertex domain)
  - Sparsity assumption on $c$
Model 2: Spectral filtering

- Thanou et al. (2017)
  - $\mathcal{F}(\mathcal{G}) = e^{-\tau L}$ (localisation in vertex domain)
  - Sparsity assumption on $\mathbf{c}$
  - Each signal is a combination of several heat diffusion processes at time $\tau$

\[ \mathbf{e}^{-\tau L} \times \mathbf{c} = \mathbf{x} \]

[Thanou17]
Model 2: Spectral filtering

- Thanou et al. (2017)

\[ \mathcal{F}(\hat{G}) = e^{-\tau L} \] (localisation in vertex domain)

- Sparsity assumption on \( c \)

- Each signal is a combination of several heat diffusion processes at time \( \tau \)

\[
\min_{L,C,\tau} \left\| X - \mathcal{F}C \right\|_F^2 + \alpha \sum_{m=1}^{M} \| c_m \|_1 + \beta \| L \|_F^2
\]

s.t. \[ \mathcal{F} = [e^{-\tau_1 L}, e^{-\tau_2 L}, \ldots, e^{-\tau_S L}] \]
Model 2: Spectral filtering

- Thanou et al. (2017)
  - $\mathcal{F}(\mathcal{G}) = e^{-\tau L}$ (localisation in vertex domain)
  - Sparsity assumption on $c$
  - Each signal is a combination of several heat diffusion processes at time $\tau$

$$\min_{L,C,\tau} \|X - \mathcal{F}C\|_F^2 + \alpha \sum_{m=1}^{M} \|c_m\|_1 + \beta \|L\|_F^2$$

s.t. $\mathcal{F} = [e^{-\tau_1 L}, e^{-\tau_2 L}, \ldots, e^{-\tau_S L}]$

data fidelity  sparsity on $c$  regularisation
Model 2: Spectral filtering

- Thanou et al. (2017)
  - $\mathcal{F}(\mathcal{G}) = e^{-\tau L}$ (localisation in vertex domain)
  - Sparsity assumption on $c$
  - Each signal is a combination of several heat diffusion processes at time $\tau$

$$\min_{L,C,\tau} \left\| X - \mathcal{F} \mathcal{C} \right\|_F^2 + \alpha \sum_{m=1}^M \|c_m\|_1 + \beta \|L\|_F^2$$

Data fidelity \hspace{2cm} Sparsity on $c$ \hspace{2cm} Regularisation

[Thanou17]

Can be extended to general polynomial case (Maretic et al. 2017)
Model 2: Spectral filtering

- Pasdeloup et al. (2017)
  - $\mathcal{F}(\mathcal{G}) = \mathbf{T}^k = \mathbf{W}^k_{\text{norm}}$
  - $\{c_m\}$ are i.i.d. samples with independent entries
Model 2: Spectral filtering

- Pasdeloup et al. (2017)
  
  - $F(G) = T^k = W_{\text{norm}}^k$
  
  - $\{c_m\}$ are i.i.d. samples with independent entries
  
  - Two-step approach:
    
    - Estimate eigenvector matrix of graph operator from sample covariance:
      
      $$
      \Sigma = \mathbb{E} \left[ \sum_{m=1}^{M} X(m)X(m)^T \right] = \sum_{m=1}^{M} W_{\text{norm}}^{2k(m)} \quad \text{(polynomial of } W_{\text{norm}}) 
      $$
Model 2: Spectral filtering

- Pasdeloup et al. (2017)

- $\mathcal{F}(G) = T^k = W^k_{\text{norm}}$

- $\{c_m\}$ are i.i.d. samples with independent entries

- Two-step approach:
  - Estimate eigenvector matrix of graph operator from sample covariance:
    \[
    \Sigma = \mathbb{E} \left[ \sum_{m=1}^{M} X(m)X(m)'^T \right] = \sum_{m=1}^{M} W_{\text{norm}}^{2k(m)} \text{ (polynomial of } W_{\text{norm}} \text{ )}
    \]
  - Optimise for eigenvalues given constraints of $W_{\text{norm}}$ (mainly non-negativity of off-diagonal of $W_{\text{norm}}$ and eigenvalue range) and some priors (e.g., sparsity)
Model 2: Spectral filtering

- Pasdeloup et al. (2017)

\[ F(G) = T^k = W^{k}_{\text{norm}} \]

- \{c_m\} are i.i.d. samples with independent entries

- Two-step approach:
  
  - Estimate eigenvector matrix of graph operator from sample covariance:
    
    \[ \Sigma = \mathbb{E} \left[ \sum_{m=1}^{M} X(m)X(m)^T \right] = \sum_{m=1}^{M} W^{2k(m)}_{\text{norm}} \quad \text{(polynomial of } W_{\text{norm}}) \]

  - Optimise for eigenvalues given constraints of \( W_{\text{norm}} \) (mainly non-negativity of off-diagonal of \( W_{\text{norm}} \) and eigenvalue range) and some priors (e.g., sparsity)

"Graph-centric": Cost on graph components (also Segarra et al. 2017)

No assumption on signal structure and sparsity, but on eigenvectors/stationarity
Model 3: Causal dependency on graphs

• Signals are causal outcome of current or past observations (spectral characteristics depending on dependence structure)

• Example: Evolution of individual behaviour due to influence of different friends at different timestamps

• Characterised by vector autoregressive models (VARMs) or structural equation models (SEMs)
Model 3: Causal dependency on graphs

- Mei and Moura (2017)
  - $\mathcal{F}_s(\mathcal{G}) = P_s(\mathbf{W})$: polynomial of $\mathbf{W}$ of degree $s$
  - Define $c_s$ as $x[t-s]$
Model 3: Causal dependency on graphs

- Mei and Moura (2017)

\[ \mathcal{F}_s(G) = \mathbf{P}_s(W) : \text{polynomial of } W \text{ of degree } s \]

- Define \( c_s \) as \( x[t-s] \)

\[
\sum_{s=1}^{S} \left( \mathbf{P}_s(W) \times x[t-s] \right) = x
\]

\[
\begin{align*}
&x[t] \\
&x[t-1] \\
&x[t-S]
\end{align*}
\]
Model 3: Causal dependency on graphs

- Mei and Moura (2017)

\[ \mathcal{F}_s(\mathcal{G}) = \mathbf{P}_s(\mathbf{W}) : \text{polynomial of } \mathbf{W} \text{ of degree } s \]

- Define \( \mathbf{c}_s \) as \( \mathbf{x}[t - s] \)

\[
\sum_{s=1}^{S} \left( \begin{array}{c} P_s(\mathbf{W}) \\ x[t-s] \\ x \end{array} \right) = \mathbf{G}
\]

\[
\min_{\mathbf{W}, \mathbf{a}} \frac{1}{2} \sum_{k=S+1}^{K} \| \mathbf{x}[k] - \sum_{s=1}^{S} P_s(\mathbf{W}) \mathbf{x}[k - s]\|^2 + \lambda_1 \| \text{vec}(\mathbf{W}) \|_1 + \lambda_2 \| \mathbf{a} \|_1
\]
Model 3: Causal dependency on graphs

- Mei and Moura (2017)

\[ \mathcal{F}_s(\mathcal{G}) = \mathbf{P}_s(\mathbf{W}) : \text{polynomial of } \mathbf{W} \text{ of degree } s \]

- Define \( c_s \) as \( \mathbf{x}[t - s] \)

\[
\begin{align*}
\sum_{s=1}^{S} \left( \mathbf{P}_s(\mathbf{W}) \mathbf{x}[t - s] \right) = & \quad \mathbf{x} \\
\end{align*}
\]

\[
\begin{align*}
\mathbf{x}[t] &= \mathbf{x}[t - 1] + \ldots + \mathbf{x}[t - S] \\
\end{align*}
\]

\[
\begin{align*}
\min_{\mathbf{W}, \mathbf{a}} \frac{1}{2} \sum_{k=S+1}^{K} \|\mathbf{x}[k] - \sum_{s=1}^{S} \mathbf{P}_s(\mathbf{W}) \mathbf{x}[k - s]\|_2^2 + \lambda_1 \|\text{vec}(\mathbf{W})\|_1 + \lambda_2 \|\mathbf{a}\|_1 \\
\text{data fidelity} & \quad \text{sparsity on } \mathbf{W} \quad \text{sparsity on } \mathbf{a}
\end{align*}
\]
Model 3: Causal dependency on graphs

- Mei and Moura (2017)
  - $\mathcal{F}_s(\mathcal{G}) = P_s(W)$: polynomial of $W$ of degree $s$
  - Define $c_s$ as $x[t - s]$

$$\min_{W,a} \frac{1}{2} \sum_{k=S+1}^{K} \left( \|x[k] - \sum_{s=1}^{S} P_s(W)x[k - s]|_2^2 + \lambda_1 \|\text{vec}(W)|_1 + \lambda_2 \|a\|_1 \right)$$

- Good for inferring causal relations between signals
  - Can be combined with SEMs and kernelised: Shen et al. (to appear)
Comparison of different GSP methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Signal model</th>
<th>Assumption on $F$</th>
<th>Assumption on $c$</th>
<th>Learning outcome</th>
<th>Edge directionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Dong16]</td>
<td>global smoothness</td>
<td>eigenvector matrix</td>
<td>Gaussian</td>
<td>Laplacian</td>
<td>undirected</td>
</tr>
<tr>
<td>[Kalofolias16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Thanou17]</td>
<td>spectral filtering</td>
<td>heat kernel</td>
<td>sparsity</td>
<td>Laplacian</td>
<td>undirected</td>
</tr>
<tr>
<td>[Pasdeloup17]</td>
<td>spectral filtering</td>
<td>normalised adj./Laplacian</td>
<td>IID Gaussian</td>
<td>normalised adj./Laplacian</td>
<td>undirected</td>
</tr>
<tr>
<td>[Mei17]</td>
<td>causal dependency</td>
<td>polynomials of adj. matrix</td>
<td>past signals</td>
<td>adj. matrix</td>
<td>directed</td>
</tr>
</tbody>
</table>
Connection with broad literature

- Global smoothness of graph signals is also promoted in Graphical LASSO.
- Models based on spectral filtering or causal dependency lead to generative process of signals, similarly to traditional physically-motivated models.
- GSP approaches offer design flexibility and extend beyond a Gaussian statistical model or a simple diffusion model.
Some examples

- Image coding and compression (review of [Chung18])
  - images are natural graph signals on regular grid
  - learning adaptive edge weights for more efficient coding

[Fracastoro17]
Some examples

- Brain signal analysis (review of [Huang18])
  - voxels in 3D-PET images as vertices
  - learning functional connectivity of brain regions
Some examples

- Brain signal analysis (review of [Huang18])
  - voxels in 3D-PET images as vertices
  - learning functional connectivity of brain regions

- Other application domains
  - learning meteorology graph using temperatures
  - learning commuting graph using traffic volume
  - learning political relations using voting data
Outline

• A (very partial) literature review
• A signal processing perspective
• Concluding remarks
Future directions

GSP for graph learning
Future directions

Input signals
- partial observations
- sequential observations

GSP for graph learning
Future directions

Input signals
- partial observations
- sequential observations

Learning outcome
- directed graphs
- time-varying (dynamic) graphs
- graphs with certain properties
- intermediate graph representation
- probabilistic structure

GSP for graph learning
Future directions

Input signals
- partial observations
- sequential observations

Signal/graph model
- beyond smoothness: localisation in vertex-frequency domain
- adapt to specific input/output

Learning outcome
- directed graphs
- time-varying (dynamic) graphs
- graphs with certain properties
- intermediate graph representation
- probabilistic structure

GSP for graph learning
Future directions

Theoretical consideration
- performance guarantee
- computational efficiency

Input signals
- partial observations
- sequential observations

Signal/graph model
- beyond smoothness: localisation in vertex-frequency domain
- adapt to specific input/output

GSP for graph learning

Learning outcome
- directed graphs
- time-varying (dynamic) graphs
- graphs with certain properties
- intermediate graph representation
- probabilistic structure
Future directions

Theoretical consideration
- performance guarantee
- computational efficiency

Objective of graph learning
- for traditional graph-based learning, e.g., clustering, dim. reduction, ranking
- integrate inference with subsequent data analysis (targeted applications)

Input signals
- partial observations
- sequential observations

Signal/graph model
- beyond smoothness: localisation in vertex-frequency domain
- adapt to specific input/output

Learning outcome
- directed graphs
- time-varying (dynamic) graphs
- graphs with certain properties
- intermediate graph representation
- probabilistic structure

GSP for graph learning
References

Learning Graphs from Data: A Signal Representation Perspective
Xiaowen Dong*, Donna Tamayo*, Michael Rabbat, and Pascal Frossard

The construction of a meaningful graph topology plays a crucial role in the effective representation, processing, analysis and visualization of structured data. When a natural choice of the graph is not readily available from the dataset, it is then desirable to infer a graph topology from the data. In this tutorial overview, we survey solutions to the problem of graph learning, including classical viewpoints from statistics and physics, and more recent approaches that adopt a graph signal processing (GSP) framework. We further emphasize the conceptual similarities and differences between classical and GSP graph inference methods and highlight the potential advantage of the latter to a number of theoretical and practical scenarios. We conclude with several open issues and challenges that are keys to the design of future signal processing and machine learning algorithms for learning graphs from data.

1. INTRODUCTION

Modern data analysis and data processing tasks typically involve large sets of structured data, where the structure carries critical information about the nature of the data. One can find numerous examples of such datasets in a wide diversity of application domains, including transportation networks, social networks, computer networks, and brain networks. Typically, graphs are used as mathematical tools to describe the structure of such datasets. They provide a flexible way for representing relationships between data entities. Numerous signal processing and machine learning algorithms have been introduced in the past decade for analyzing structured data on a priori known graphs [3, 4]. However, there are often settings where the graph is not readily available, and the structure of the data has to be estimated in order to permit effective representation, processing, analysis or visualization of graph data. In this case, a crucial task is to infer a graph topology that describes the characteristics of the data observations, hence capturing the underlying relationship between these entities.

The problem of graph learning is the following: given M observations of N variables or data entities, represented in a data matrix X ∈ ℝM×N, and given some prior knowledge e.g. distribution, data model, and tools to fit and computationally feasible online learning algorithms, which include some form of filtering over graphs. Real data examples highlight the impact of the network and domain models on complex and temporal network dynamics, and leveraging appropriate algorithms and methodologies to handle learning over such networks is a crucial step in the development of innovative tools for various application domains.

1. INTRODUCTION

The purpose of minimizing the network connections is not only to understand the behavior of complex systems [35, 49, 60] but also to model and analyze complex and temporal network dynamics. The challenges in modeling complex and temporal network dynamics include efficient modeling of complex networks and their interactions with other complex networks, as well as the modeling of complex networks and their interactions with other complex networks and their interactions with other complex networks.

Keywords: network modeling, network dynamics, network analysis.