GRAPH SIGNAL PROCESSING: A STATISTICAL VIEWPOINT

Cha Zhang

Joint work with Dinei Florêncio and Philip A. Chou

Microsoft Research

Outline

- Gaussian Markov Random Field
- Graph construction
- Graph transform
- Graph down-sampling
- Graph prediction
- Graph-based regularization
- Conclusions

Gaussian Markov Random Field (GMRF)

A random vector $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ is called a GMRF with respect to the graph $G = (V = \{1, \cdots, n\}, E)$ with mean $\boldsymbol{\mu}$ and a precision matrix $\boldsymbol{Q} > 0$ (positive definite), if and only if its density has the form:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{Q}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

$$\text{and } Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in E \text{ for all } i \neq j.$$

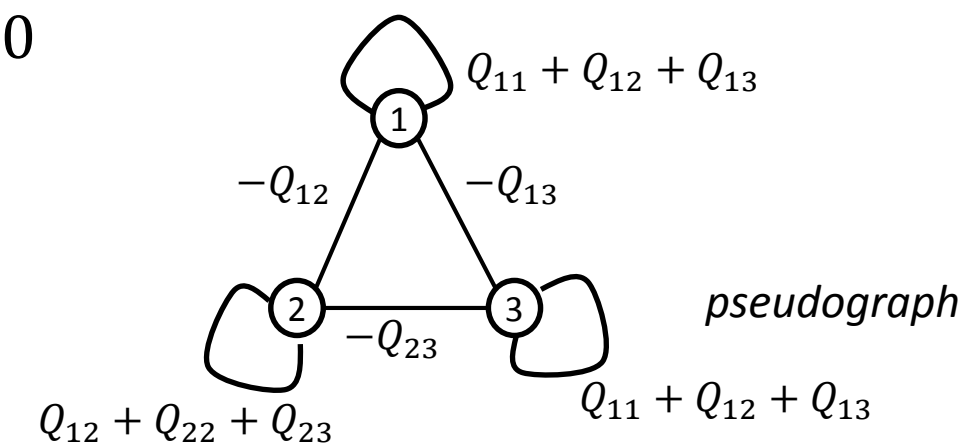
Gaussian Markov Random Field (GMRF)

- It is defined on a graph
- It is a multivariate Gaussian distribution
- Regarding the precision matrix $\boldsymbol{Q}$
  - $\mathbf{Q} = \mathbf{\Sigma}^{-1}$, $\mathbf{\Sigma}$ is the covariance matrix
  - Conditional interpretation
    - $E(x_i|\boldsymbol{x}_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}\left(x_j - \mu_j\right)$
    - $Prec(x_i|\boldsymbol{x}_{-i}) = Q_{ii}$
    - $Corr\left(x_i, x_j \middle| \boldsymbol{x}_{-ij}\right) = \frac{-Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, i \neq j$

A Bijective Mapping

- Zero-mean GMRF ⇔ a subset of bi-directed graphs
  - Random variable ⇔ node
  - $Q_{ij} \neq 0$ ⇔ edge
  - Assign graph weights:
    - $W_{ij} = -Q_{ij}$, for $i \neq j$
    - $W_{ii} = \sum_{j=1}^{n} Q_{ij}$
  - Mapping between $\boldsymbol{W}$ and $\boldsymbol{Q}$ is bijective
  - Subset because $\boldsymbol{Q} > 0$

$$\boldsymbol{Q} = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{bmatrix}$$



pseudograph

Intrinsic GMRF (IGMRF)

A random vector $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ is called an intrinsic GMRF of kth order if it has density:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n-k}{2}} (|\boldsymbol{Q}|^*)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Where $\boldsymbol{Q}$ is of rank $n-k$, and $|\cdot|^*$ denotes the generalized determinant (product of the non-zero eigenvalues).

IGMRF can be interpreted as limits of proper GMRFs, thus many of the properties of GMRF holds for IGMRF.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005

A Special GMRF

- Zero mean, and all rows of $\boldsymbol{Q}$ sum to zero
  - For all i, $\sum_j Q_{ij} = 0$
- When $\boldsymbol{Q}$ is of rank $n-1$: IGMRF of first order
- No self-loops in the graph → Simple Graph
- Local behavior

$$E(x_i|\boldsymbol{x}_{-i}) = -\frac{1}{Q_{ii}}\sum_{j:j\sim i} Q_{ij}x_j, \text{ with } -\frac{\sum_{j:j\sim i} Q_{ij}}{Q_{ii}} = 1$$

- IGMRF of first order closely related to first-order random walk

So…

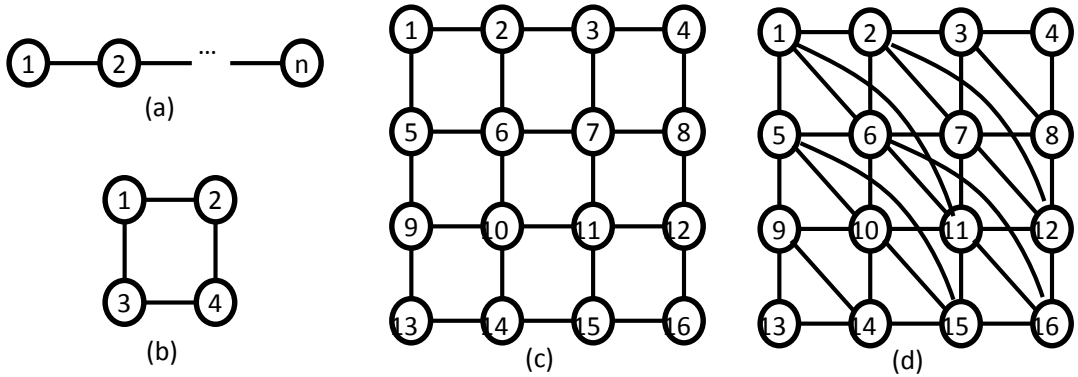
- For many bi-directed graphs, we can give them a GMRF (IGMRF) interpretation
- It's a Gaussian, isn't it too simple?
  - Real-world signals are hardly Gaussian…
  - What can we do with such a simple model?

So…

- For many bi-directed graphs, we can give them a GMRF (IGMRF) interpretation
- It's a Gaussian, isn't it too simple?
  - Real-world signals are hardly Gaussian…
  - What can we do with such a simple model?
    - A lot!
    - Many are already using it without realizing it…

Graph Construction

- Data-driven
  - Collect enough examples to estimate $\boldsymbol{\mu}$ and $\boldsymbol{Q}$
  - $\boldsymbol{Q}$ is unlikely sparse → densely connected graph
- Intuitive model-based
  - Widely used in 1D/2D regular grid signals
  - IGMRF, no self-loops: nice local behavior
  - $\boldsymbol{Q}$ identical to the Laplacian matrix



Graph Construction

- Model-constrained data-driven approach
- Maximum a-posterior estimation

$$\widehat{\boldsymbol{Q}} = \arg\max_{\boldsymbol{Q}} p(\boldsymbol{Q} | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$$
$$= \arg\max_{\boldsymbol{Q}} p(\boldsymbol{Q}) \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{Q})$$

- $p(\boldsymbol{x}_n | \boldsymbol{Q})$: likelihood, GMRF model
- $p(\boldsymbol{Q})$: prior knowledge
  - Application dependent
  - For example, $\boldsymbol{Q}$ is sparse, know graph structure, etc.

X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst. Learning graphs from signal observations under smoothness prior. In *submitted to: IEEE Transacctions on Signal Processing*, 2014.

Graph Transform

- Karhunen-Loève transform (KLT)

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{\mathrm{T}}$$

  Where $\boldsymbol{\Lambda} = diag(\lambda_0, \cdots, \lambda_n)$ is the diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}$.

- Since:

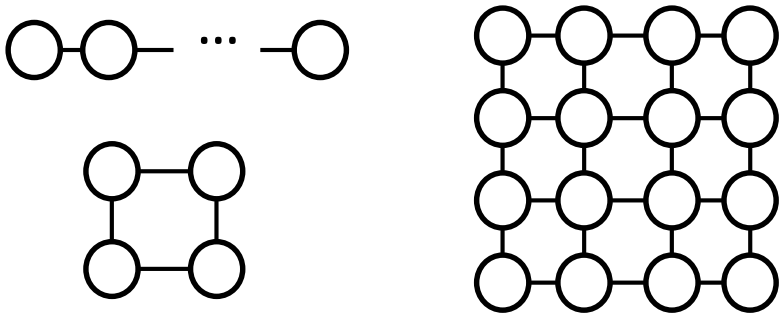
$$\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$$

  The eigenvector matrix of $\boldsymbol{Q}$ is the same as the KLT, ***optimal*** for decorrelation.

- Since GMRF is defined on a graph, we call the eigenvector matrix of the precision matrix $\boldsymbol{Q}$ the *Graph Transform*.

Graph Transform and DCT

- When $\boldsymbol{Q}$ is the Laplacian of a regular grid graph
  - 1D signal ⇒ Graph transform = 1D DCT
  - 2D signal on regular grid ⇒ 2D DCT is ***one*** of the eigenvector matrices of $\boldsymbol{Q}$
  - Thus under such an image model, 1D DCT and 2D DCT are optimal for decorrelation
  - **First known proof of 2D DCT optimality!**



C. Zhang and D. Florêncio. Analyzing the optimality of predictive transform coding using graph-based models. *IEEE Signal Processing Letters*, 20(1), Jan. 2013.

Graph Down-sampling

Given random vector $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ on graph $G = (V = \{1, \cdots, n\}, E)$, keep $\boldsymbol{x}_1 = (x_1, \cdots, x_{n-k})^T$, remove $\boldsymbol{x}_2 = (x_{n-k+1}, \cdots, x_n)^T$, what is the graph for $\boldsymbol{x}_1$?

Graph Down-sampling

Given GMRF random vector $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ on graph $G = (V = \{1, \cdots, n\}, E)$, keep $\boldsymbol{x}_1 = (x_1, \cdots, x_{n-k})^T$, remove $\boldsymbol{x}_2 = (x_{n-k+1}, \cdots, x_n)^T$, what is the graph for $\boldsymbol{x}_1$?

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}^{-1}$$

$$\rightarrow \mathbf{Q}_1 = \boldsymbol{\Sigma}_{11}^{-1} = \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}$$

Identical to *Kron Reduction*, a well-know technique for graph down-sampling originally derived for electronic networks.

F. Dörfler and F. Bullo. Kron reduction of graphs with applications to electrical networks. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 60(1):150–163, 2013.

Graph Prediction

- Random vector $\boldsymbol{x} = (x_1, \cdots, x_n, x_{n+1}, \cdots, x_m)^T$ follows GMRF with precision matrix $\boldsymbol{Q}$.
  - $\boldsymbol{x}_1 = (x_1, \cdots, x_n)^T$ unknown
  - $\boldsymbol{x}_2 = (x_{n+1}, \cdots, x_m)^T$ known

Graph Prediction

- Random vector $\boldsymbol{x} = (x_1, \cdots, x_n, x_{n+1}, \cdots, x_m)^T$ follows GMRF with precision matrix $\boldsymbol{Q}$.
  - $\boldsymbol{x}_1 = (x_1, \cdots, x_n)^T$ unknown
  - $\boldsymbol{x}_2 = (x_{n+1}, \cdots, x_m)^T$ known
  - Partition $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}$$

- $\boldsymbol{x}_1 | \boldsymbol{x}_2$ is also a GMRF with respect to its own subgraph with mean $\boldsymbol{\mu}_{\boldsymbol{x}_1|\boldsymbol{x}_2}$ and precision matrix $\boldsymbol{Q}_{\boldsymbol{x}_1|\boldsymbol{x}_2} > 0$:

$$\boldsymbol{\mu}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \boldsymbol{\mu}_{\boldsymbol{x}_1} - \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_{\boldsymbol{x}_2}\right)$$

$$\boldsymbol{Q}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \boldsymbol{Q}_{11}$$

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005

Predictive Transform Coding

- In the above formulation:
  - $\boldsymbol{x}_1$: pixels to be encoded
  - $\boldsymbol{x}_2$: know pixels nearby
- New conditional mean: best prediction

$$\boldsymbol{\mu}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \boldsymbol{\mu}_{\boldsymbol{x}_1} - \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_{\boldsymbol{x}_2}\right)$$

- New best decorrelation transform after applying best prediction:
  - Eigenvector matrix of $\boldsymbol{Q}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \boldsymbol{Q}_{11}$

Motion Prediction

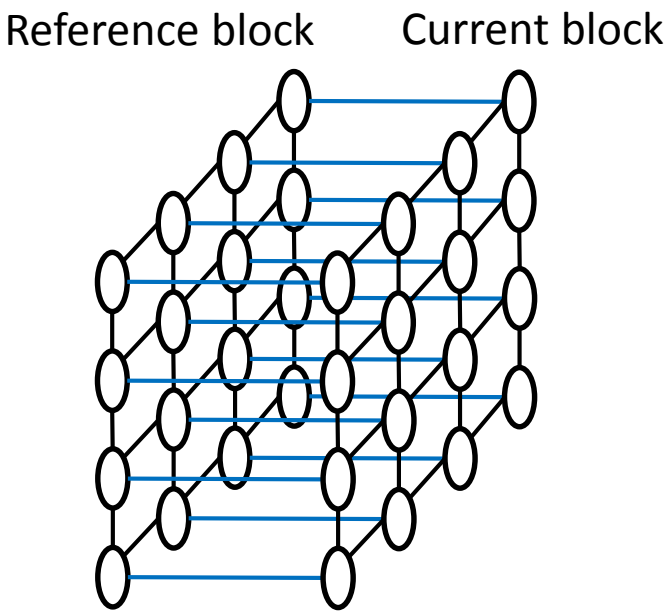
- Reference block and current block, both zero mean GMRF, with $\boldsymbol{Q}_{ref}$ and $\boldsymbol{Q}_c$
- Form a 3D GMRF model
- Precision matrix:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_c + \boldsymbol{I} & -\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{Q}_{ref} + \boldsymbol{I} \end{bmatrix}$$

- Regarding $\boldsymbol{x}_1 | \boldsymbol{x}_2$:

$$\boldsymbol{\mu}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = -(\boldsymbol{Q}_c + \boldsymbol{I})^{-1}\, \boldsymbol{x}_2$$

$$\boldsymbol{Q}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \boldsymbol{Q}_c + \boldsymbol{I}$$

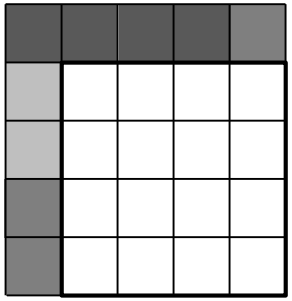


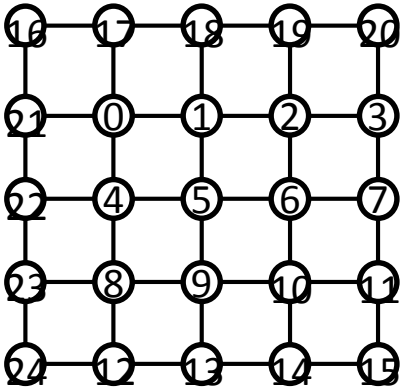
Motion Prediction

- When $\boldsymbol{Q}_c$ is the Laplacian matrix $\boldsymbol{Q}_c = \delta \boldsymbol{L}$
  - New mean: $\boldsymbol{\mu}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = -(\delta \boldsymbol{L} + \boldsymbol{I})^{-1}\, \boldsymbol{x}_2$
    - **Low pass filter on $\boldsymbol{x}_2$**
    - (Partial) reason that half-pixel prediction is better!
  - New precision matrix: $\boldsymbol{Q}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \delta \boldsymbol{L} + \boldsymbol{I}$
    - Its eigenvector matrix is same as the eigenvector matrix of $\boldsymbol{L}$
    - **First known proof that DCT is still optimal for residue compression in motion compensation!**

Intra-Frame Predictive Coding

- Zero mean GMRF
- New mean: $\boldsymbol{\mu}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{x}_2$
  - **Optimal prediction given the GMRF model**
- New precision matrix: $\boldsymbol{Q}_{\boldsymbol{x}_1|\boldsymbol{x}_2} = \boldsymbol{Q}_{11}$
  - DCT no longer optimal!
  - **First general analysis of the optimal transform for intra-frame predictive coding**





C. Yeo, Y. H. Tan, Z. Li, and S. Rahardja. Mode-dependent transforms for coding directional intra prediction residuals. *IEEE Trans. on Circuits and Systems for Video Technology*, 22(4):545–554, 2012.

Graph-Based Regularization

- Estimate a signal $\boldsymbol{x}$ on graph, such that a cost function is minimized, subject to certain regularization

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda S_p(\boldsymbol{x})$$

A popular smoothness term (p-Dirichlet form):

$$S_p(\boldsymbol{x}) \coloneqq \frac{1}{\mathrm{p}}\left[\sum_i \sum_{j:j\sim i} W_{ij}\left(x_i - x_j\right)^2\right]^{\frac{p}{2}}$$

A Prob. Viewpoint on Regularization

- Target function:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda S_p(\boldsymbol{x})$$

$$= \arg\max_{\boldsymbol{x}} e^{-\frac{f(\boldsymbol{x})}{\lambda} - S_p(\boldsymbol{x})}$$

- Mapping to MAP:
  - Likelihood: $e^{-\frac{f(\boldsymbol{x})}{\lambda}}$
  - Prior: $e^{-S_p(\boldsymbol{x})}$
    - When $p = 2$ → Laplacian GMRF prior
    - When $p \neq 2$ → Generalized GMRF prior

Conclusions

- GMRF provides a statistical viewpoint for GSP
- Provides valuable insights for
  - Graph construction
  - Graph transform
  - Graph down-sampling
  - Graph prediction
  - Graph-based regularization
- Always be aware of the statistical implications!