Learning Gaussian Graphical Models With Domain Specific Priors

or

Learning the Structure of Complex Data

Jean Honorio and Dimitris Samaras
Stony Brook University
Probabilistic Graphical Models

Graphical models are a marriage between probability theory and graph theory

- **Representation**: how can a graphical model compactly represent a joint probability distribution?
- **Inference**: how can we efficiently infer the hidden states of a system, given partial and possibly noisy observations?
- **Learning**: how do we estimate the parameters and structure of the model?
One Application: fMRI

**Problem:** find functional connectivity in fMRI data

- Fully exploratory: entire brain without ROI selection
- Learning BNs or MRFs on large number of variables is difficult

- Find coarse and fine grain connectivity
  - fMRI data seems to be Gaussian
  - Sparse, voxel to clusters, clusters to clusters
Example Application: Brain fMRI

- Activated brain regions show a change in blood supply
- Blood carries oxygen
- fMRI provides 3D brain images:
  - intensities are blood oxygenation levels

Examine the connectivity and interactivity between brain regions
Brain fMRI: monetary reward and cocaine addiction

- Reveals brain circuits of inhibitory control
  - 16 cocaine addicted subjects and 12 control subjects
  - each subject: 6 sessions
  - each session: 87 scans taken every 3.5 secs
  - 157 Talairach brain regions (nodes)

- **edges represent functional connectivity among brain regions**

- **12,500 parameters to be estimated for each task**
GGMs

- Pairwise *Markov random field* of jointly Gaussian variables
  - Tractable partition function (determinant)
GGMs

- Structure learning: learn graphical model topology (and parameters) from data

- Penalized maximum likelihood estimation:
  - sparseness (Meinshausen & Bühlmann, 2006; Banerjee et al., 2006; Friedman et al., 2007; Yuan & Lin, 2007)
  - diagonal structure (Levina et al., 2008)
  - block structure for known block-variable assignments (Duchi et al., 2008)
  - unknown block-variable assignments (Marlin & Murphy, 2009; Marlin et al., 2009)
  - spatial coherence (Honorio et al., 2009)
Learning GGMs

- We add regularizers that lead to convex problems.
- We propose "block coordinate descent" algorithms.
- The precision matrices that we generate are positive definite.
Most methods assume
- only sparseness (or simplicity)
- or too strong spatial priors, e.g.: if two pixels are far, they are assumed independent

Local constancy:

2D dataset (spatial neighborhood in black dashed lines)
*Local constancy does not discourage long range interactions*
Functional Brain MRI

Structures learnt from functional brain MRI of (a) drug addicted subjects versus (b) control subjects in a monetary reward task. Drug addicted subjects have more functional connections in the cerebellum (in yellow) when compared to control subjects, while control subjects have more connections in the prefrontal cortex (in green).

- **edges represent functional connectivity among brain regions**
- **undersampled to 869 voxels = 378,000 parameters to be estimated**
Our technique captures in (b): groupwise similar displacements from each independent leg (positive interaction in blue), and opposite displacements between legs and hands/feet (negative interaction in red) in the original sequence (a).
Cardiac MRI

(a) Original sequence
(b) Learnt structure

Our technique captures in (b): similar displacements between neighbor pixels (positive interaction in blue) caused by rotation, and opposite displacements across walls (negative interaction in red) caused by shrinking in the original sequence (a)
Multi-task learning

- Related tasks share common sparseness pattern of features:
  - regression (Liu et al., 2009)
  - classification (Jebara, 2004)
  - compressive sensing (Qi et al., 2008)
  - reinforcement learning (Wilson et al., 2007)
  - Bayesian networks (Niculescu-Mizil & Caruana, 2007)
Multi-task learning of GGMs

- Related tasks share common sparseness pattern of edges
- Maximum likelihood estimation with a $\ell_{1,\infty}$-norm penalty for corresponding edges across tasks
- Concave and leads to continuous quadratic knapsack problem with closed form solution
Results

- Subgraph of 10 brain regions for 3 cocaine subjects (both randomly selected), for our multi-task method (top) and GLasso (bottom)
- *Sparseness pattern of the structures produced by our technique is consistent across subjects*
GGMs as feature points

- Is the log-likelihood “well behaved”?
  - i.e. small changes in the parameters produce small changes in the objective function.

- Does the $l_p$ distance between the learnt parameters and the ground truth provide some guarantee on their generalization ability,
  - i.e. the expected log-likelihood.

- How does the $l_p$-norm of the difference of parameters between two graphical models relate to the similarity between probability distributions,
  - i.e. the Kullback-Leibler divergence.
If the log-likelihood is \textit{Lipschitz continuous with respect to the parameters} of a graphical model

- It is reasonable to use those parameters (e.g. weights of GGMs) as features for classification, dimensionality reduction and clustering

We prove that almost all graphical models have \textit{Lipschitz parameterization}

- Bayesian networks, Markov random fields and factor graphs for discrete and continuous random variables. Dynamic models such as dynamic Bayesian networks and conditional random fields
Proposed Projects

- General library for GGM learning (6 months)
  - Develop techniques for Variable Selection (3 months)
  - Work on data from CDDA members! (12 months)
  - Use VS and MT to detect variations in Inhibitory Control
  - SPM toolbox for detecting changes in longitudinal FMRI studies (12 months)

- Modeling and Activity Recognition using Kinect data for human interaction
  - Kinect dataset for humans interacting (3 months)
  - Find the GGM features for action recognition (9 months)
  - Develop techniques for subtle facial expression modeling using GGMs to model interaction between facial parts
Variable Selection for Interpretability

- Datasets with thousands of variables: gene expression, stock prices, world weather
- Enforce that only a small number of nodes interact with each other

We want to select these *important* nodes, and find their interaction pattern
One Motivation

- Activity recognition from time sequences: video, motion capture, accelerometers
- A proposed model:

Feature space (parameters of learnt models)
Temporal Segmentation

- Is it possible to segment a complex sequence?
- CMU MoCap: subject 86
- Small windows of 0.75 secs of activity
- **Features:** weights of a GGM
- Dimensionality reduction by PCA
- k-means clustering of 3 first eigenvectors
- In the figure, each point is a GGM