FODAVA-Lead: Dimension Reduction and Data Reduction: Foundations for Visualization Low Rank Matrix Learning Problems and Entropy Penalization

Vladimir Koltchinskii and Pedro Rangel

School of Mathematics
Georgia Institute of Technology
Fodava Annual Meeting

December 2010

Research Directions

- Sparse recovery in infinite dictionaries (Koltchinskii and Minsker, *COLT 2010*, 420–433)
- Active learning (Koltchinskii, *J. Machine Learning Research*, 2010, 11, 2387–2415)
- Sparsity in multiple kernel learning (Koltchinskii and Yuan, *Annals of Statistics*, 2010, 38, 3660–3695)
  - applications of multiple kernel learning to heterogeneous data fusion and multi language document analysis (with Haesun Park, Pedro Rangel)
- Low rank matrix learning:
  - nuclear norm approach (Koltchinskii, Lounici and Tsybakov, 2010)
  - von Neumann entropy penalization (Koltchinskii, 2010)
  - learning of low rank kernels on graphs (Koltchinskii and Rangel)

Low Rank Matrix Recovery

- Suppose that A is a large matrix and
  - either it has low rank,
  - or it can be well approximated by a low rank matrix.
- The goal is to estimate A based on noisy measurements of linear functionals of A

- Matrix Completion: Candes and Recht (2009), Candes and Tao (2009), ...
- Matrix Regression: Candes and Plan (2009), Rohde and Tsybakov (2009), Koltchinskii, Lounici and Tsybakov (2010), ...
- Quantum State Tomography: Gross et all (2009), Gross (2009), Koltchinskii (2010)
- Learning Kernels based on Empirical Data
- Covariance Matrix Estimation

Quantum State Tomography

- $\mathbb{M}_m(\mathbb{C})$ the set of all $m \times m$ matrices with complex entries
- $\mathcal{S} := \left\{ S \in \mathbb{M}_m(\mathbb{C}) : S = S^*, S \geq 0, \mathrm{tr}(S) = 1 \right\}$
- $\rho \in \mathcal{S}$ a **density matrix**
- $X_1, \ldots, X_n$ i.i.d. Hermitian matrices (observables) with distribution Π independent of $\xi_1, \ldots, \xi_n$ (for instance, a sample from the Pauli basis)
- **Regression Model:**

$$Y_j := \mathrm{tr}(\rho X_j) + \xi_j, \; j = 1, \ldots, n$$

- **Random Noise:** $\{\xi_j\}$ i.i.d. random variables with $\mathbb{E}\xi_j = 0$, $\sigma_\xi^2 := \mathbb{E}\xi_j^2 < +\infty$
- **Goal:** estimate ρ based on $(X_1, Y_1), \ldots, (X_n, Y_n)$

von Neumann Entropy Penalization

- **von Neumann entropy:**

$$\mathcal{E}(S) := -\mathrm{tr}(S \log S), S \in \mathcal{S}.$$

- **Entropy penalized least squares method: Trade-off between minimizing the empirical risk and maximizing the entropy**

$$\hat{\rho}^{\varepsilon} := \mathrm{argmin}_{S \in \mathcal{S}} \left\{ n^{-1} \sum_{j=1}^{n} (Y_j - \langle S, X_j \rangle)^2 + \varepsilon \, \mathrm{tr}(S \log S) \right\}.$$

- **Koltchinskii (2010) von Neumann Entropy Penalization and Low Rank Matrix Estimation**, arXiv:1009.2439v1: bounds on the error of $\hat{\rho}^{\varepsilon}$ in terms of the rank or “approximate rank” of ρ.

From Quantum State Tomography to Learning Kernels on Graphs

- **Learning Kernels**: Cristianini et al (2002), Lancriet et al (2004)): the goal is to find a symmetric nonnegatively drfinite kernel “well aligned” with the data.
- **Applications:**
  - prediction of similarities between points outside of the observed sample;
  - embeddings of the data into a Euclidean feature space
  - design kernel machines for classification and other learning problems
- **Tsuda and Noble, Learning kernels from biological networks by maximizing entropy,** *Bioinformatics*, 2004: applications of von Neumann entropy maximization to design of locally constrained diffusion kernels for prediction problems on protein and gene networks.

Learning Kernels on Graphs

- (V, E) a graph
- $X_1, \dots, X_n$ an i.i.d. sample from a probability distribution in V;
- S an $n \times n$ symmetric **empirical similarity matrix**;
- each entry s_{ij} describes how “similar” the vertices X_i and X_j are.
- **Goal:** to design a kernel K (=symmetric nonnegatively definite matrix $(K(u,v))_{u,v\in V}$) that approximates the similarity matrix and, at the same time, reflects the geometry of the graph.
- $K(u,v) = \langle \phi(u), \phi(v) \rangle,\ u, v \in V$, ϕ an embedding of V into a Euclidean feature space;
- **energy** of the embedding ϕ :

$$\mathrm{tr}(KL) = \sum_{u,v\in V, u\sim v} \|\phi(u) - \phi(v)\|^2,$$

where L is the **Laplacian** of the graph.

Entropy Penalization Approach to Learning

- Similarity matrix S is properly normalized;
- $\mathrm{tr}(K) = 1$, i.e., K is a density matrix;
- von Neumann entropy can be used as a complexity penalty to find a low rank estimate of K :

$$\hat{K} = \mathrm{argmin}_{K \in \mathcal{S}} \left[\frac{\lambda_1}{n(n-1)} \sum_{i \neq j} (s_{ij} - K(X_i, X_j))^2 + \right.$$

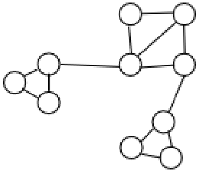
$$\left. \lambda_2 \, \mathrm{tr}(KL) + \lambda_3 \, \mathrm{tr}(K \log K) \right],$$

  $\lambda_1, \lambda_2, \lambda_3 > 0$ are regularization parameters.

- **The estimator $\hat{K}$ provides a trade-off between fitting the kernel for the data, minimizing the energy of the embedding defined by the kernel and maximizing the von Neumann entropy of the kernel**

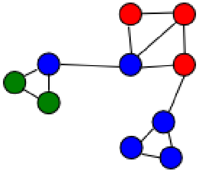
Learning kernels for prediction on graphs

- A social network $G = (V, E)$.



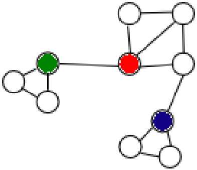
Learning kernels for prediction on graphs

- A social network $G = (V, E)$.
- Assume that a function on V represents political preferences of the individuals.

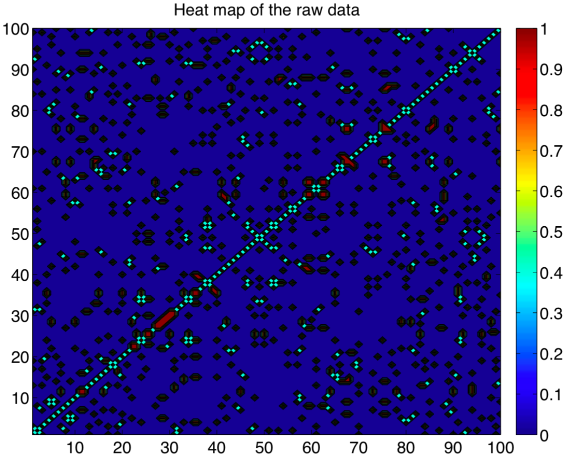


Learning kernels for prediction on graphs

- A social network $G = (V, E)$.
- Assume that a function on V represents political preferences of the individuals.
- The goal is to design a kernel for predicting how the individuals are going to vote in the forthcoming elections.
- This approach takes into account both the political preferences and the interactions between the individuals reflected in the geometry of the graph

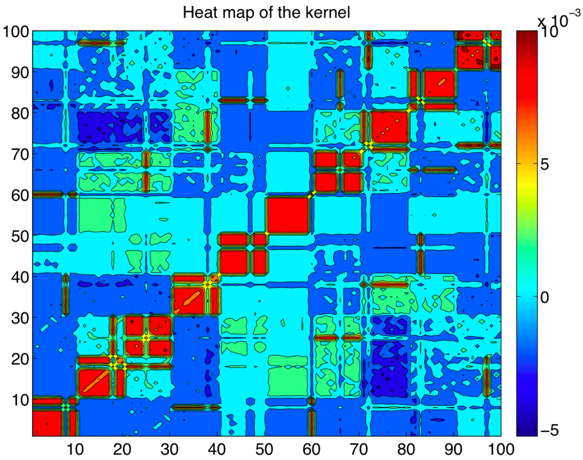


Experiments



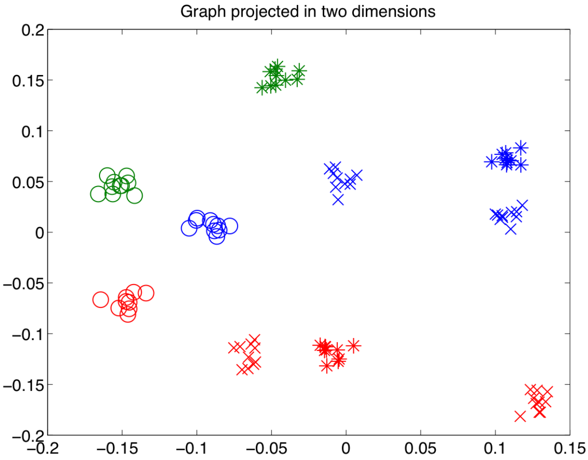
Heat map for the adjacency matrix of the graph. Note that this heat map does not show any particular structure

Experiments



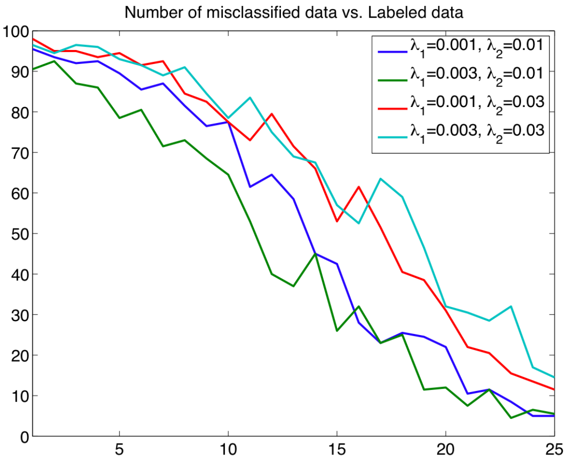
Heat map for the estimated kernel (the vertices are reordered to reveal the clusters of the graph)

Experiments



The estimated kernel naturally induces an embedding of the graph in a high dimensional feature space projected further into two dimensions. The clusters of the network and voting preferences become visible

Experiments



Using the kernel to solve a learning problem. Classification error vs number of labeled data.

Future Directions

- Estimation error bounds, in particular, oracle inequalities for $\hat{K}$ (such as in quantum state tomography and other low rank matrix estimation problems)
- Tuning methodology for regularization parameters
- Simultaneous learning of the classifier and of the kernel in kernel machine design for binary and multiclass classification
- Learning graph laplacians based on samples of vertices and edges
- Simultaneous learning of similarity kernels and laplacians
- Methods of embedding and visualization of graphs based on kernel learning