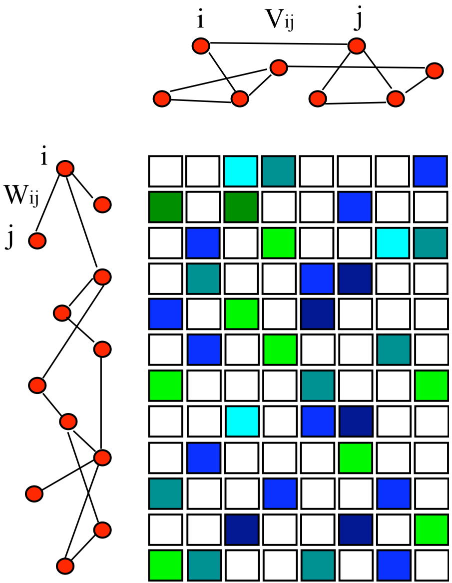
Low Rank Estimation of Smooth Kernels on Weighted Graphs

Vladimir Koltchinskii
Based on a joint work with Pedro Rangel

School of Mathematics
Georgia Institute of Technology

July 2013

Matrix Completion on Weighted Graphs



Matrix Completion

- V a set, $\mathrm{card}(V) = m$
- (X, X', Y) a random variable in $V \times V \times [-a, a]$, $a > 0$
- X, X' independent points sampled from the uniform distribution Π in V
- $Y \in [-a, a]$ a response variable

Matrix Completion

- $S_*(u,v) := \mathbb{E}(Y|X = u, X' = v)$
- Assume that $S_* \in \mathcal{S}_V$, where

$$\mathcal{S}_V := \Big\{ S : S(u,v) = S(v,u), u, v \in V \Big\}$$

- $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$ i.i.d. copies of (X, X', Y) (data)
- **Goal**: estimate S_* based on the data
- **Assumption**: S_* is low rank

Noiseless Problem: Nuclear Norm Minimization

- **Noiseless problem:**

$$Y_j := S_*(X_j, X_j'),\ j = 1, \dots, n$$

- $$\hat{S} := \operatorname{argmin}\Big\{\|S\|_1 : Y_j = S(X_j, X_j'), j = 1, \dots, n, S \in \mathcal{S}_V\Big\},$$

  where $\|\cdot\|_1$ is **the nuclear norm:**

$$\|S\|_1 := \mathrm{tr}(|S|) = \mathrm{tr}\Big(\sqrt{S^2}\Big)$$

  - Matrix Completion and Other Problems: Recht, Fazel and Parrilo (2010), Candes and Recht (2009), Candes and Tao (2010), Gross et all (2010), Gross (2011)

Noiseless Matrix Completion: Low Coherence Conditions

- $$S_* = \sum_{j=1}^{r} \mu_j (\psi_j \otimes \psi_j), \quad \mathrm{sign}(S_*) := \sum_{j=1}^{r} \mathrm{sign}(\mu_j)(\psi_j \otimes \psi_j)$$
- $L := \mathrm{l.s.}(\{\psi_j : 1 \leq j \leq r\})$
- $\{e_1, \dots, e_m\}$ the canonical basis of $\mathbb{R}^V$ with standard Euclidean inner product $\langle \cdot, \cdot \rangle$
- **Low Coherence.** For some $\nu > 0$,

$$\|P_L e_j\|^2 \leq \frac{\nu r}{m}, \ j = 1, \dots, m$$

and

$$\left| \left\langle \mathrm{sign}(S_*) e_i, e_j \right\rangle \right|^2 \leq \frac{\nu r}{m^2}, i, j = 1, \dots, m.$$

Noiseless Low Rank Recovery: Matrix Completion

- The following result is due to Candes and Tao (2010), Gross (2011):

Theorem

There exists a constant $C > 0$ such that

$$\mathbb{P}\{\hat{S} \neq S_*\} \leq m^{-2}$$

for all

$$n \geq C\nu rm \log^2 m.$$

Nuclear Norm Penalization in Noisy Matrix Completion (Matrix LASSO)

- Candes and Plan (2011), Rohde and Tsybakov (2011), Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2011), Koltchinskii (2011, 2012), ...
- Let $\mathcal{S}_V^a := \Big\{S : S \in \mathcal{S}_V, \max_{u,v\in V} |S(u,v)| \leq a\Big\}$
- $$\hat{S}^{\varepsilon} := \mathrm{argmin}_{S\in\mathcal{S}_V^a}\Big\{n^{-1}\sum_{j=1}^{n}(Y_j - S(X_j, X_j'))^2 + \varepsilon\|S\|_1\Big\}$$
- $\varepsilon > 0$ regularization parameter
- The error of this estimator will be expressed in terms of the squared $L_2(\Pi^2)$-norm:

$$\|S\|^2_{L_2(\Pi^2)} = m^{-2}\sum_{u,v\in V}|S(u,v)|^2.$$

A Low Rank Oracle Inequality

- Koltchinskii (2012)
- Given $t > 0$, $\bar{t} := t + 3\log\log\Big(m \vee n \vee a \vee a^{-1} \vee 2\Big)$.

Theorem

There exist constants $C, D > 0$ such that, for all $t > 0$ and for all

$$\varepsilon \geq Da\left(\sqrt{\frac{t+\log(2m)}{mn}} \bigvee \frac{t+\log(2m)}{n}\right),$$

with probability at least $1 - e^{-t}$,

$$\|\hat{S}^{\varepsilon} - S_*\|^2_{L_2(\Pi^2)} \leq \inf_{S \in \mathcal{S}_V^a}\left[\|S - S_*\|^2_{L_2(\Pi^2)} + C\left(m^2\varepsilon^2 \mathrm{rank}(S) + \frac{a^2\bar{t}}{n}\right)\right].$$

Remarks

- Take

$$\varepsilon = Da\sqrt{\frac{t+\log(2m)}{nm}}$$

and suppose that

$$\frac{m(t+\log(2m))}{n} \le 1, \quad \bar{t} \le m(t+\log(2m))$$

- Then, with probability at least $1-e^{-t}$

$$\|\hat{S}^{\varepsilon} - S_*\|^2_{L_2(\Pi^2)} \le C\frac{a^2 m\, \mathrm{rank}(S_*)(t+\log(2m))}{n}$$

Lower Bounds: Low Rank Case

- Koltchinskii, Lounici and Tsybakov (2011)
- Let $1 \leq r \leq m$ and let $\mathcal{P}_{r,a}$ be the set of all distributions P of (X, X', Y) such that X, X' are independent, $X, X' \sim \Pi$, $|Y| \leq a$ a.s. and $\mathbb{E}(Y|X, X') = S_P(X, X')$, where $S_P \in \mathcal{S}_V$, $\mathrm{rank}(S_P) \leq r$.

Theorem

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}} \sup_{P \in \mathcal{P}_{r,a}} \mathbb{P}_P \left\{ \|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \geq c_1 \frac{a^2 mr}{n} \right\} \geq c_2,$$

where the infimum is taken over all estimators $\hat{S}$ based on i.i.d. data $(X_j, X'_j, Y_j), j = 1, \dots, n$ sampled from P.

Weighted Graphs and Laplacians

- (V, A) **weighted graph** with a weight matrix $A := (a(u, v))_{u,v \in V}$, $a(u, v) \geq 0, a(u, v) = a(v, u), u, v \in V$.
- $\mathrm{deg}(u) = \sum_{v \in V} a(u, v)$ the degree of vertex $u \in V$.
- $\Delta := D - A$ (**the Laplacian**), D the diagonal matrix with $\mathrm{deg}(u), u \in V$ on the diagonal
- Δ can be viewed as an operator from $\mathbb{R}^V$ into itself
- $$\langle \Delta f, f \rangle = \frac{1}{2} \sum_{u,v \in V} a(u, v)(f(u) - f(v))^2, f : V \mapsto \mathbb{R}$$

Discrete Sobolev Norms and Smooth Kernels

- $S \in \mathcal{S}_V$ a kernel with spectral representation

$$S = \sum_{j=1}^{m} \mu_j (\psi_j \otimes \psi_j),$$

- $\{\mu_j\}$ are its eigenvalues
- $\{\psi_j\}$ are its orthonormal (in $\mathbb{R}^V$) eigenfunctions
- The "smoothness" of S can be characterized by **discrete Sobolev norms** $\|\Delta^{p/2} S\|_{L_2(\Pi^2)}, p > 0$:

$$\|\Delta^{p/2} S\|^2_{L_2(\Pi^2)} = \sum_{j=1}^{m} \mu_j^2 \|\Delta^{p/2} \psi_j\|^2_{L_2(\Pi)}.$$

Spectral Characteristics

- $W = \Delta^p$, $p > 0$ is a fixed constant.
- **Spectral Representation**

$$W = \sum_{j=1}^{m} \lambda_j (\phi_j \otimes \phi_j)$$

  $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ the eigenvalues of W
- $\phi_1, \phi_2, \ldots, \phi_m$ the corresponding orthonormal (in $\mathbb{R}^V$) eigenfunctions of W (and of Δ)
- $\lambda_{k+1} \leq c\lambda_k$ (if $\lambda_k > 0$)

Estimation of Smooth and Low Rank Kernels

- **Goal**: estimate the kernel S_* based on the training data $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$, assuming that S_* is low rank and "smooth" on the graph
- Let $r := \mathrm{rank}(S_*)$
- The "smoothness" of S_* will be characterized by

$$\rho := \|W^{1/2} S_*\|_{L_2(\Pi^2)}.$$

Heuristic Considerations

- $S_{*,l} := \sum_{i,j=1}^{l} \langle S_* \phi_i, \phi_j \rangle (\phi_i \otimes \phi_j)$
- $\|S_* - S_{*,l}\|^2_{L_2(\Pi)} \leq \frac{2\rho^2}{\lambda_{l+1}}, l = 1, \ldots, m$
- If, for each l, one can estimate $S_{*,l}$ with the squared $L_2(\Pi^2)$-error

$$\sim \frac{a^2 (r \wedge l) l}{n},$$

then one can expect that the squared $L_2(\Pi^2)$-error of estimation of S_* would be

$$\min_{1 \leq l \leq m} \left[\frac{a^2 (r \wedge l) l}{n} \bigvee \frac{\rho^2}{\lambda_{l+1}} \right]$$

- It is easy to see that

$$\min_{1 \leq l \leq m} \left[\frac{a^2 (r \wedge l) l}{n} \bigvee \frac{\rho^2}{\lambda_{l+1}} \right] \asymp \max_{1 \leq l \leq m} \left[\frac{a^2 (r \wedge l) l}{n} \bigwedge \frac{\rho^2}{\lambda_l} \right]$$

Lower Bounds for Low Rank and Smooth Kernels

- $\mathcal{S}_{r,\rho} := \{S \in \mathcal{S}_V : \mathrm{rank}(S) \leq r, \|W^{1/2}S\|_{L_2(\Pi^2)} \leq \rho\}$

$$\mathcal{P}_{r,\rho,a} := \Big\{P : (X, X', Y) \sim P, X, X' \text{ independent} \sim \Pi, |Y| \leq a,$$
$$\mathbb{E}(Y|X, X') = S_*(X, X') = S_P(X, X'), S_P \in \mathcal{S}_{r,\rho}\Big\}$$

- $\bar{\phi}_j := \sqrt{m}\phi_j,\ j = 1, \dots, m$ ($\{\bar{\phi}_j\}$ are orthonormal in $L_2(\Pi)$).

$$Q_p := \max_{1 \leq j \leq m} \|\bar{\phi}_j\|^2_{L_p(\Pi)}.$$

Lower Bounds for Smooth and Low Rank Kernels

- $$\delta_n(r,\rho,a) := \max_{1\le l\le m}\left[\frac{a^2(r\wedge l)l}{n}\bigwedge\frac{\rho^2}{\lambda_l}\bigwedge\frac{1}{(p-1)Q_p^2 m^{4/p}}\frac{a^2(r\wedge l)}{l}\right].$$

Theorem

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat S}\sup_{P\in\mathcal{P}_{r,\rho,a}}\mathbb{P}_P\left\{\|\hat S - S_P\|^2_{L_2(\Pi^2)} \ge c_1\delta_n(r,\rho,a)\right\} \ge c_2,$$

where the infimum is taken over all estimators $\hat S$ based on i.i.d. data $(X_j, X_j', Y_j), j=1,\dots,n$ sampled from P.

Lower Bounds for Smooth and Low Rank Kernels: Example

- $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$.
- $n \geq C' Q_\infty^{(\beta+1)/\beta} (\log m)^{(\beta+1)/2\beta} \left(\frac{\rho}{a}\right)^{1/\beta}$
- $$\delta_n(r,\rho,a) \asymp$$ $$\left(\left(\frac{a^2\rho^{1/\beta} r}{n}\right)^{2\beta/(2\beta+1)} \bigwedge \left(\frac{a^2\rho^{2/\beta}}{n}\right)^{\beta/(\beta+1)} \bigwedge \frac{a^2 rm}{n}\right) \bigvee \frac{a^2}{n}.$$

Least Squares Estimators with Nonconvex Penalties

- $$\mathcal{S}_r(l;a) := \left\{ S \in \mathcal{S}_V : \mathrm{rank}(S) \le r, \|S\|_{L_2(\Pi)} \le a, S = \sum_{i,j=1}^{l} s_{ij}(\phi_i \otimes \phi_j) \right\}$$

- S^a truncation of $S \in \mathcal{S}_V$:

$$S^a(u,v) = S(u,v)I(|S(u,v)| \le a) + aI(S(u,v) > a)$$
$$-aI(S(u,v) < -a)$$

- $\bar{\mathcal{S}}_r(l;a) := \{S^a : S \in \mathcal{S}_r(l;a)\}$

Least Squares Estimators with Nonconvex Penalties

- $$\hat{S}_{r,l,a} := \mathrm{argmin}_{S \in \bar{\mathcal{S}}_r(l;a)} \frac{1}{n} \sum_{j=1}^{n} (Y_j - S(X_j, X_j'))^2$$

- $$(\hat{r}, \hat{l}) :=$$
  $$\mathrm{argmin}_{r,l} \left[\frac{1}{n} \sum_{j=1}^{n} (Y_j - \hat{S}_{r,l,a}(X_j, X_j'))^2 + B \frac{a^2 (r \wedge l) l}{n} \log \left(\frac{Bnm}{(r \wedge l) l} \right) \right],$$

  $B > 0$ is a constant

- $$\hat{S} := \hat{S}_{\hat{r}, \hat{l}, a}$$

Least Squares Estimators with Nonconvex Penalties: An Upper Bound

- Recall that

$$\mathcal{S}_{r,\rho} := \{S \in \mathcal{S}_V : \mathrm{rank}(S) \le r, \|W^{1/2}S\|_{L_2(\Pi^2)} \le \rho\}$$

$$\mathcal{P}_{r,\rho,a} := \Big\{P : (X, X', Y) \sim P, X, X' \text{ independent} \sim \Pi, |Y| \le a,$$
$$\mathbb{E}(Y|X, X') = S_*(X, X') = S_P(X, X'), S_P \in \mathcal{S}_{r,\rho}\Big\}$$

Theorem

For all $P \in \mathcal{P}_{r,\rho,a}$ and for all $t > 0$ with probability at least $1 - e^{-t}$

$$\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \le C\left(\min_{1\le l\le m}\left[\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Bnm}{(r\wedge l)l}\right)\bigvee\frac{\rho^2}{\lambda_{l+1}}\right]\bigvee\frac{a^2(t+\log m)}{n}\right).$$

Upper Bound: Example

- $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$.
- $$\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \leq$$ $$C\Big[\Big(\Big(\frac{a^2\rho^{1/\beta} r}{n}\log\frac{Bnm}{r}\Big)^{2\beta/(2\beta+1)} \bigwedge \Big(\frac{a^2\rho^{2/\beta}\log(Bnm)}{n}\Big)^{\beta/(\beta+1)}$$ $$\bigwedge \frac{a^2 rm\log(Bnm)}{n}\Big) \bigvee \frac{a^2(t+\log m)}{n}\Big].$$
- Compare with the lower bound:

$$\Big(\Big(\frac{a^2\rho^{1/\beta} r}{n}\Big)^{2\beta/(2\beta+1)} \bigwedge \Big(\frac{a^2\rho^{2/\beta}}{n}\Big)^{\beta/(\beta+1)} \bigwedge \frac{a^2 rm}{n}\Big) \bigvee \frac{a^2}{n}.$$

Least Squares Estimators with Double Penalization: Nuclear Norm and Discrete Sobolev Norm

- $$\hat{S}_{\varepsilon,\bar{\varepsilon}} := \mathrm{argmin}_{S \in \mathcal{S}_V^a} \left\{ \sum_{j=1}^{n} (Y_j - S(X_j, X_j'))^2 + \varepsilon \|S\|_1 + \bar{\varepsilon} \|W^{1/2} S\|_{L_2(\Pi^2)}^2 \right\}$$
- $\|S\|_1 := \mathrm{tr}(|S|), S := \sqrt{S^2}$ **nuclear norm**
- $\varepsilon, \bar{\varepsilon}$ regularization parameters
- If $\bar{\varepsilon} = 0$, then $\hat{S}_{\varepsilon,0}$ is a **matrix LASSO estimator**

Choice of $\bar{\varepsilon}$: aggregation

- Divide the sample $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$ into two parts,

  $$(X_j, X_j', Y_j), j = 1, \dots, n' \ \text{ and } \ (X_{n'+j}, X_{n'+j}', Y_{n'+j}), j = 1, \dots, n - n',$$

  where $n' := [n/2] + 1$.
- $\hat{S}_l := \hat{S}_{\varepsilon, \bar{\varepsilon}_l}$, $\bar{\varepsilon}_l := \lambda_l^{-1}$, $l = 1, \dots, m$ is based only on the first n' observations.
- $$\hat{l} := \mathrm{argmin}_{l=1,\dots,m} \frac{1}{n - n'} \sum_{j=1}^{n-n'} \Big(Y_{n'+j} - \hat{S}_l(X_{n'+j}, X_{n'+j}') \Big)^2.$$
- $\hat{S} := \hat{S}_{\hat{l}}$.

Coherence Function

- **Spectral Function:** $F(\lambda) := \sum_{j=1}^{m} I(\lambda_j \le \lambda)$
- $S_* = \sum_{k=1}^{r} \mu_k (\psi_k \otimes \psi_k)$
- $L := \mathrm{l.s.}(\psi_1, \dots, \psi_r)$, $\dim(L) = \mathrm{rank}(S_*) = r$
- P_L the orthogonal projector onto L
- $E(\lambda) := \sum_{\lambda_j \le \lambda} (\phi_j \otimes \phi_j)$
- **Coherence Function**

$$\varphi(S_*; \lambda) := \langle P_L, E(\lambda) \rangle = \sum_{\lambda_j \le \lambda} \|P_L \phi_j\|^2, \lambda \ge 0$$

- $\varphi(S_*; \lambda) \le F(\lambda)$
- **A Low Coherence Assumption:** for some $\nu(S_*) \ge 1$,

$$\varphi(S_*; \lambda) \le \frac{\nu(S_*) r F(\lambda)}{m}, \lambda \ge 0.$$

Bounds under Low Coherence Assumption

- Given $t > 0$, let $t_{n,m} := t + 3\log\left(2\log_2 n + \frac{1}{2}\log_2 \frac{\lambda_m}{\lambda_1} + 2\right)$.
- Suppose $\frac{m(t+\log(2m))}{n} \le 1$ and $\varepsilon = Da\sqrt{\frac{t+\log(2m)}{nm}}$

Theorem

With probability at least $1 - e^{-t}$,

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le$$

$$C \min_{1\le l\le m}\left(\frac{\nu(S_*)\mathrm{rank}(S_*)F(\lambda_l)(t+\log(2m))}{n} + \frac{\|W^{1/2}S_*\|^2_{L_2(\Pi^2)}}{\lambda_l}\right)$$

$$+C\frac{a^2(\log(m+1)+t_{n,m})}{n}.$$

Bounds under Low Coherence Assumption: Example

- $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$
- Then

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \leq$$
$$C\Biggl(\biggl(\Bigl(\frac{\nu a^2 \rho^{1/\beta} r \log(2m)}{n}\Bigr)^{2\beta/(2\beta+1)} \bigwedge \Bigl(\frac{\nu a^2 \rho^{2/\beta} \log(2m)}{n}\Bigr)^{\beta/(\beta+1)}$$
$$\bigwedge \frac{a^2 rm}{n}\biggr) \bigvee \frac{a^2(\log(m+1) + t_{n,m})}{n}\Biggr)$$

- Compare with the lower bound:

$$\biggl(\Bigl(\frac{a^2 \rho^{1/\beta} r}{n}\Bigr)^{2\beta/(2\beta+1)} \bigwedge \Bigl(\frac{a^2 \rho^{2/\beta}}{n}\Bigr)^{\beta/(\beta+1)} \bigwedge \frac{a^2 rm}{n}\biggr) \bigvee \frac{a^2}{n}.$$