Research Update for MURI Project

Richard A. Davis
Columbia University

What am I working on?

1. Random matrices with heavy-tails (big n big p): (joint work with Thomas Mikosch and Oliver Pfaffel)

- Large dimensional data sets appear in many quantitative fields like finance, environmental sciences, wireless communications, fMRI, and genetics.
- Structure in this data can often be analyzed via sample covariances. PCA is used to transform data to a new set of variables, the principal components, ordered such that the first few retain most of the variation of the data.

Goal: In this work, we study the asymptotic properties of the largest eigenvalues from the sample covariance matrix.

What am I working on?

Setup (simplified version for this talk):

Let $\{\mathbf{X}_t\}$ be an iid sequence of random vectors of length p with Pareto tails and index $\alpha \in (2,4)$. That is,

$$P(X_{t,i} > x) \sim \frac{C}{x^{\alpha}}, \qquad as\ x \to \infty.$$

Assume the rv's have mean 0, so that the covariance matrix is given by

$$\Gamma = E(\mathbf{X}_{\mathrm{t}}\mathbf{X}_{\mathrm{t}}^{T}).$$

Estimate Γ by the sample covariance matrix

$$\hat{\Gamma} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_{\mathrm{t}}\mathbf{X}_{\mathrm{t}}^{T}$$

What am I working on?

Classical results: Assume the vector consists of IID N(0,1) rvs.

- For $n \to \infty$ and *p* fixed, Anderson (1963) proved that

$$\left(\frac{n}{2}\right)^{-1/2} (\hat{\lambda}_{(1)}/n - 1) \xrightarrow{d} N(0,1),$$

where $\hat{\lambda}_{(1)}$ is the largest eigenvalue of the empirical covariance matrix $n\widehat{\Gamma}$.

- Johnstone [2001] showed that for $p, n \to \infty, \frac{p}{n} \to \gamma \in (0,2)$, then

$$\frac{\sqrt{n}+\sqrt{p}}{\sqrt[3]{\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{p}}}} \left(\frac{(\widehat{\lambda}_{(1)}}{(\sqrt{n}+\sqrt{p})^2} \quad - 1 \right) \xrightarrow{d} \text{Tracy-Widom distribution.}$$

Theorem: Let $\{\mathbf{X}_t\}$ be an iid sequence of random vectors of length p w/

$$X_{t,j} = \sum_{i=0}^{\infty} \theta_{i Z_{j-i}},$$

where $\{Z_i\}$ is an iid sequence of rvs satisfying $P(|Z_i| > x) \sim \frac{1}{x^{\alpha}}, as\ x \to \infty$.

Let $\hat{\lambda}_{(1)} > \cdots > \hat{\lambda}_{(p)}$ be the decreasing eigenvalues of the empirical covariance matrix $n\hat{\Gamma} = \sum_{t=1}^{n} \mathbf{X}_\mathrm{t}\mathbf{X}_\mathrm{t}^T$. Then for almost ***any*** sequence $p_n \to \infty$ $(\limsup_n \frac{p_n}{\exp\{cn\}} < \infty)$, then with $c = \theta_0^2 + \theta_1^2 + \cdots,$

$$(np)^{-\frac{2}{\alpha}} \hat{\lambda}_{(i)} \xrightarrow{d} \mathrm{c}\Gamma_i^{-\frac{2}{\alpha}}, \qquad if\ \alpha \in (0,2)$$

$$(np)^{-\frac{2}{\alpha}} (\widehat{\lambda}_{(i)} - n\lambda_i) \xrightarrow{d} \mathrm{c}\Gamma_i^{-\frac{2}{\alpha}}, \qquad if\ \alpha \in (2,4),$$

where $\Gamma_i = E_1 + \cdots + E_i$, $E_1, E_2, \ldots,$ are iid unit exponentials, λ_i is the i^{th} largest eigenvalue of Γ.

What am I working on?

$$(np)^{-\frac{2}{\alpha}}\hat{\lambda}_{(i)} \xrightarrow{d} c\Gamma_i^{-\frac{2}{\alpha}}, \qquad if\ \alpha \in (0,2)$$

$$(np)^{-\frac{2}{\alpha}}(\hat{\lambda}_{(i)} - n\lambda_i) \xrightarrow{d} c\Gamma_i^{-\frac{2}{\alpha}}, \qquad if\ \alpha \in (2,4),$$

Remarks:

- Limits only depend on dependence between the rows via the constant *c*.
- In the first case, $\hat{\lambda}_{(1)}/\hat{\lambda}_{(2)} \xrightarrow{d} \Gamma_2^{\frac{2}{\alpha}}/\Gamma_1^{\frac{2}{\alpha}}$, which is the same regardless of the dependence between the rows!
- Results generalize to linear dependence between columns.
- Looking at problems where dependence is more than linear between rows.

What am I working on?

2. Indegree/outdegree analysis for growth networks

Collaborators(?): Sid, Bo Jiang, and Don

Krapivsky et al 2001 model:

The model is recursively defined. If V_n represents the nodes at *time* *n*, then

- with prob *p*, a new node attaches to existing node *v* WP

  $$\frac{in(v)+\lambda}{\sum_{v \in V_n}(in(v)+\lambda)}$$

  (indegree of *v* goes up by 1.)

- with prob 1-*p*, a new edge $u \rightarrow v$ is created between two existing nodes *u* and *v* with prob

  $$\frac{(in(v)+\lambda)(out(u)+\mu)}{\sum_{u \neq v \in V_n}(in(v)+\lambda)(out(u)+\mu)}$$

What am I working on?

Results from Krapivsky:

Let

$$N_{i,j} = number\ of\ nodes\ with\ indegree\ i\ and\ outdegree\ j$$

$$N_{i,\cdot} = number\ of\ nodes\ with\ ind\ i;$$

$$N_{\cdot,j} = number\ of\ nodes\ with\ out\ j;$$

Defining the relative frequencies (assuming they exist) via,

$$f_{i,j} = \lim_{n\to\infty} \frac{N_{i,j}}{n}$$

Marginal distributions:

- $f_{i,\cdot} \sim i^{-v_{in}}$, $v_{in} = 2 + p\lambda$ and $f_{\cdot,j} \sim j^{-v_{out}}$, $v_{out} = 2 + p(1+\mu)/(1-p)$

Joint distributions:

- If $v_{in} = v_{out}$, then $f_{i,j} \sim C \frac{i^{\lambda-1} j^{\mu}}{(i+j)^{2\lambda+1}}$

What am I working on?

Similar results for reciprocity models (Bo and Don): At each step, new node *v* is added such that

- with prob 1-*p*, it has an outgoing edge $v \to u$ WP

$$\frac{(in(u)+\lambda)}{\sum_{u \in V_n}(in(u)+\lambda)}$$

- with prob *p*, an edge $v \to u$ is created and a reciprocal edge $u \to v$ between two existing nodes *u* and *v* is created with prob

$$\frac{(in(u)+\lambda)}{\sum_{u \in V_n}(in(u)+\lambda)}$$

What am I working on?

Similar results for reciprocity models (Bo and Don):

Let

$$N_{i,j} = number\ of\ nodes\ with\ indegree\ i\ and\ outdegree\ j$$

$$N_{i,\cdot} = number\ of\ nodes\ with\ ind\ i;$$

$$N_{\cdot,j} = number\ of\ nodes\ with\ out\ j;$$

Defining the relative frequencies (assuming they exist) via,

$$f_{i,j} = \lim_{n \to \infty} \frac{N_{i,j}}{n}$$

Marginal distributions:

- $f_{i,\cdot} \sim (A+B) i^{-\nu_{in}}$, $\nu_{in} = 2 + p + \lambda$ and $f_{\cdot,j} \sim C i^{-\nu_{out}}$, $\nu_{out} = ?$

What am I working on?

Two views:

1. Double limit--empirical process:

$$F_n(\cdot) = \frac{1}{n}\sum_{t=1}^{n} \epsilon_{inv(v_t),out(v_t)}(\cdot)$$

So for $g\colon \mathbb{N}^+ \times \mathbb{N}^+ \to \mathbb{R}$,

$$F_n(g) = \frac{1}{n}\sum_{t=1}^{n} g(\,inv(v_t), out(v_t))$$

$$= \sum_{i,j=1}^{n} n^{-1} N_{i,j}\, g(\,i,j)$$

$$\to \sum_{i,j=1}^{\infty} f_{i,j}\, g(\,i,j)$$

$i.e.,$ $$F_n(\cdot) \to \sum_{i,j=1}^{\infty} f_{i,j}\, \epsilon_{i,j}(\cdot)$$

What am I working on?

2. Single sequence limit–point process:

$$T_n(\cdot) = \sum_{t=1}^{n} \epsilon_{\left\{\frac{inv(v_t)}{a_n}, \frac{out(v_t)}{b_n}\right\}}(\cdot)$$

so that for g: $\mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$, bounded with support on $[x, \infty) \times [y, \infty)$.

$$T_n(g) = \sum_{t=1}^{n} g\left(\frac{inv(v_t)}{a_n}, \frac{out(v_t)}{b_n}\right) = \sum_{i \geq a_n x, j \geq b_n y} n^{-1} N_{i,j} n g\left(\frac{i}{a_n}, \frac{j}{b_n}\right)$$

$$\sim \int_{a_n x}^{\infty}\int_{b_n y}^{\infty} f(u,v) n g\left(\frac{u}{a_n}, \frac{v}{b_n}\right) du dv$$

$$= \int_{x}^{\infty}\int_{y}^{\infty} n f(u a_n, v b_n) n a_n b_n g(u,v) du dv$$

What am I working on?

2. Single sequence limit–point process (cont):

$$= \int_x^\infty \int_y^\infty nf(ua_n, vb_n) na_n b_n g(u,v)\,du\,dv$$

$$\to \int_x^\infty \int_y^\infty c(u,v) g(u,v)\,du\,dv$$

provided we can choose a_n and b_n such that

$$na_n b_n f(ua_n, vb_n) \to c(u,v).$$

In particular,

$$T_n(du,dv) \overset{v}{\to} c(u,v)\,du\,dv \quad a.s.$$

What am I working on?

Krapivsky's example:

Assume $v_{in} = v_{out}$, in which case

$$f(i,j) \sim C \frac{i^{\lambda-1} j^{\mu}}{(i+j)^{2\lambda+1}}$$

Take $a_n = b_n = n^{1/(1+p\lambda)}$, and since $\mu - \lambda = -(1+p\lambda)$,

$$n a_n^2 f(a_n u, a_n v) \sim C \frac{n a_n^2 a_n^{\lambda-1} a_n^{\mu} u^{\lambda-1} v^{\mu}}{a_n^{2\lambda+1} (u+v)^{(2\lambda+1)}} = C \frac{n a_n^{\mu-\lambda} u^{\lambda-1} v^{\mu}}{(u+v)^{(2\lambda+1)}}$$

$$= C \frac{u^{\lambda-1} v^{\mu}}{(u+v)^{(2\lambda+1)}} =: c(u,v).$$

Observe that

$$c(au, av) = a^{-2} a^{\mu-\lambda} c(u,v)$$

so that the corresponding bivariate regularly measure is homogeneous with index $-(1+p\lambda)$,, i.e., regularly varying with index $\alpha = 1 + p\lambda$.

What am I working on?

Summary:

$$T_n(\cdot) = \sum_{t=1}^{n} \epsilon_{\left\{\frac{inv(v_t)}{a_n}, \frac{out(v_t)}{b_n}\right\}}(\cdot) \xrightarrow{v} c(u,v)dudv \quad a.s.$$

where $a_n = b_n = n^{1/(1+p\lambda)}$.

Note that the marginals have similar behavior, i.e.,

$$\sum_{t=1}^{n} \epsilon_{\left\{\frac{inv(v_t)}{a_n}\right\}}(\cdot) \xrightarrow{v} Cu^{-(1+p\lambda)-1}du \quad a.s.$$

i.e., regularly varying with the same index $1 + p\lambda$.

What do I intend to work on?

Attachment models:

- Develop the theory to establish the convergences (double limit or single sequence limit (see Sid)
- Attempt to understand the genesis of the power laws for the indegree/outdegrees using methods outside of solutions for recurrence equations (see Sid).
- Establish CLTs associated with the empirical and point process convergences.
- Estimation procedure (MLE) for estimating parameters of the model (p, λ, μ)
- Develop procedures for checking goodness of fit.

How can MURI team members provide support and expertise?

- Data!

Where do I need help?

- Data