Multivariate Heavy Tails: Modeling, Diagnostics and Applications

Sidney Resnick
School of Operations Research and Information Engineering
Rhodes Hall, Cornell University
Ithaca NY 14853 USA

MURI website: software, papers, talks:
http://www.orie.cornell.edu/research/groups/multheavytail
sir1@cornell.edu

MURI ARL October 27
October 25, 2017
1. **Outline**

1.1. **Particular 5-parameter directed edge preferential attachment model for social network growth**

Helps us organize our thinking about various issues.

- Computing the multivariate heavy tail of (in,out)-degree in the generative model as a function of the 5 parameters. Resnick and Samorodnitsky (2015); Samorodnitsky, Resnick, Towsley, Davis, Willis, and Wan (2016); Wan, Wang, Davis, and Resnick (2017); Wang and Resnick (2016)

- Calibrating the model to data; statistical analysis. Wan, Wang, Davis, and Resnick (2017)
  
  - MLE fitting of the model when data consists of directed edges & timestamps is excellent–efficient when model correct.
  
  - Approximate MLE fitting of the model when snapshot of directed edges is available is effective when model correct.
  
  - Ongoing (WWDRII): asymptotic EVT estimation methods are competitive when there is model error or model noise.

  * **But**<sub>1</sub>: Why do EVT methods work since devised for iid.
  * **But**<sub>2</sub>: How to deal with tradeoffs between model error and MLE efficiency.
• Probabilistic methods of computing model descriptors like degree distributions are fragile: Change assumptions and must start over.

  – Already happened to us (WWDR): 3 scenarios → 5-scenarios after looking at data.

  – Embedding methods (in pure birth, birth-immigration, Markov branching processes) to better understand how to mathematically compute preferential attachment model tail indices (ongoing–WR). Seek analysis methods more robust to changes in preferential attachment assumptions.

  – Need behavior of tail empirical measure for large degrees. Success for one model (WR) but ongoing.

    * Needed to understand asymptotics of degree distributions.
    * Needed to understand asymptotics of Hill estimator.
– Does the Hill estimator of tail indices always work?
  * Widely used (Eg KONECT Kunegis (2013)) with religious conviction.
  * Ongoing (WR): Hill provably consistent for one model but
    → methods depend on concentration inequalities holding between expected number of nodes of large degrees (finite network) and asymptotic number of nodes of large degree
    → We suspect these inequalities do not always hold.
  * Ongoing: Which diagnostics devised for repeated sampling (iid) can be used for network data.

• Can you expect a 5-parameter model to explain 1.5 million data?
  – In our experience, parameters change.
  – Ongoing (WWDR): need to identify change points.
\((1, 1) = \text{unclean}, \quad (1, 2), (2, 1), (2, 2) = \text{cleaned data.}\)

- (Stylized) Multivariate risk estimation based on discerning an asymptotic model that sits at asymptopia smilingly taunting the analyst.

- Based on multivariate regular variation: Imagine $\mathbf{Z} \geq \mathbf{0}$ is a risk vector (losses, pollutant levels, ...) and for risk regions $A$ bounded away from $\mathbf{0}$:

$$\lim_{t \to \infty} tP \left[ \frac{\mathbf{Z}}{b(t)} \in A \right] \to \nu(A).$$

The limit measure $\nu(\cdot)$ always concentrates on a cone $\mathbb{C}$. and

$$P[\widehat{\mathbf{Z}} \in A] \approx \frac{1}{t} \hat{\nu}(A/b(t)).$$

If the cone of concentration satifies $\mathbb{C} \subsetneq \mathbb{R}_+^2$ and $A \cap \mathbb{C} = \emptyset$; risk estimate of being in $A$ is 0.

- This is a function of the asymptotic method but is the risk really 0?
• Special cases:
  – Asymptotic independence of variables as in the Gaussian copula dependence model. $\mathbb{C} =$ axes.
  – Strong dependence or full asymptotic dependence. $\mathbb{C} =$ narrow wedge.
1.3. Other topics

1.3.1. No time: Threshold selection by the minimum distance method

- When doing tail estimation in one or more dimensions, what portion of the data should be used?
- Proposals by (among others)
  - Nguyen and Samorodnitsky (2012, 2013)
  - Wan and Davis (2017) and the minimum distance method of
  - Clauset (Clauset et al. (2009); Virkar and Clauset (2014))
  common in CS, ECE and networking.
- Clauset: for one dimensional data coming from $F(x)$ with power law tail $1 - F(x) \sim x^{-\alpha}$.
  - With data $X_1, \ldots, X_n$ and order-statistics $X_{(1)} \geq \cdots > X_{(n)}$, use $X_{(1)} \geq \cdots > X_{(k)}$.
  - What $k$?
  - Suggestion: Define KS distance between empirical tail CDF and Pareto tail using $k$ order statistics and best fitting Pareto tail using the Hill estimator $\hat{\alpha}(k)$:

$$D_k := \sup_{y \geq 1} \left| \frac{1}{k} \sum_{i=1}^{n} \epsilon_{X_i/X_{(k)}} (y, \infty) - y^{-\hat{\alpha}(k)} \right|, \quad 1 \leq k \leq n.$$
Choose the optimal $k^*$ as the one that minimizes the KS distance, that is,

$$k^* := \arg\min_{k \in I} D_k,$$

Seems to work well in practice; handy R-package. Not much is proven. Quite hard to prove anything.

* $k^*/n$ seems to converge to random limit in pure Pareto case.
* Is $\hat{\alpha}(k^*)$ consistent? (Probably). Asymptotically normal? (Maybe not or at least maybe not for network data.)
* Does it work if the tail is regularly varying and not pure Pareto or even 2nd order regularly varying. (Who knows?)
  * Stable distribution tails are 2nd order regularly varying.
• 500 simulations of undirected network with $10^5$ edges for various values of parameter.

• 500 estimates of tail index of degree distribution using Clauset method.

• For some values of parameters, estimates do not look normal

• Due to?
  
  – Clauset?
  
  – This is network and not iid data?
  
  – Both?
1.3.2. No time: Trimming a Lévy process by removing large jumps.

- Robustify estimation?
- Ongoing: Joint limit behavior for rth largest Lévy jump and trimmed Lévy process.
2. Preferential Attachment as Model for Network Growth

Resnick and Samorodnitsky (2015); Samorodnitsky, Resnick, Towsley, Davis, Willis, and Wan (2016); Wan, Wang, Davis, and Resnick (2017); Wang and Resnick (2016)

2.1. A model

Bollobás et al. (2003); Krapivsky and Redner (2001)

- Model parameters: $\alpha, \beta, \gamma, \delta_{in}, \delta_{out}$ with $\alpha + \beta + \gamma = 1$.
- $G(n) = (V_n, E_n)$ is a directed random graph with $n$ edges, $N(n)$ nodes, node set $V_n$ and edge set
  \[ E_n = \{(u, v) \in V_n \times V_n : (u, v) \in E_n\}. \]

- Node degree:
  - In-degree of $v$ in $G(n)$ is $D_{in}^{(n)}(v)$;
  - Out-degree of $v$ in $G(n)$ is $D_{out}^{(n)}(v)$.
- Obtain graph $G(n)$ from $G(n - 1)$ in a Markovian way as follows:
1. \( \alpha \) scenario: With probability \( \alpha \), append to \( G(n-1) \) a new node \( v \notin V_{n-1} \) and create directed edge \( v \mapsto w \in V_{n-1} \) with probability
\[
\frac{D^{(n-1)}_{in}(w) + \delta_{in}}{n - 1 + \delta_{in}N(n-1)}.
\]

2. \( \gamma \) scenario: With probability \( \gamma \), append to \( G(n-1) \) a new node \( v \notin V_{n-1} \) and create directed edge \( w \in V_{n-1} \mapsto v \notin V_{n-1} \) with probability
\[
\frac{D^{(n-1)}_{out}(w) + \delta_{out}}{n - 1 + \delta_{out}N(n-1)}.
\]

3. \( \beta \) scenario: With probability \( \beta \), create new directed edge between existing nodes
\[
v \in V_{n-1} \mapsto w \in V_{n-1}
\]
with probability
\[
\left( \frac{D^{(n-1)}_{out}(v) + \delta_{out}}{n - 1 + \delta_{out}N(n-1)} \right) \left( \frac{D^{(n-1)}_{in}(w) + \delta_{in}}{n - 1 + \delta_{in}N(n-1)} \right).
\]
2.2. Background.

Set

\[ N_{ij}(n) = \text{# nodes with in-degree}=i \text{ and out-degree}=j \text{ in } G(n). \]

Then (eg, Bollobás et al. (2003)) the limiting proportion of nodes with in-degree\(=i \) and out-degree\(=j \) is

\[
\lim_{n \to \infty} \frac{N_{ij}(n)}{N(n)} = p(i, j) = \text{a prob mass function.}
\]

2.2.1. Marginal behavior.

The limiting degree frequency \((p(i, j))\) has power-law tails: For some finite positive constants \(C_{\text{in}}\) and \(C_{\text{out}}\),

\[
p_{i}(\text{in}) := \sum_{j=0}^{\infty} p(i, j) \sim C_{\text{in}} i^{-t_{\text{in}}} \quad \text{as } i \to \infty, \text{ as long as } \alpha \delta_{\text{in}} + \gamma > 0,
\]

\[
p_{j}(\text{out}) := \sum_{i=0}^{\infty} p(i, j) \sim C_{\text{out}} j^{-t_{\text{out}}} \quad \text{as } j \to \infty, \text{ as long as } \gamma \delta_{\text{out}} + \alpha > 0,
\]

where

\[
t_{\text{in}} = 1 + \frac{1 + \delta_{\text{in}}(\alpha + \gamma)}{\alpha + \beta}, \quad t_{\text{out}} = 1 + \frac{1 + \delta_{\text{out}}(\alpha + \gamma)}{\gamma + \beta}.
\]
2.2.2. Joint behavior.

Resnick and Samorodnitsky (2015); Samorodnitsky, Resnick, Towsley, Davis, Willis, and Wan (2016); Wan, Wang, Davis, and Resnick (2017); Wang and Resnick (2016)

Set\[
    c_1 = \frac{1}{t_{in} - 1}, \quad c_2 = \frac{1}{t_{out} - 1}, \quad a = \frac{c_2}{c_1}.
\]

For \(x > 0, y > 0,\)

\[
    \lim_{m \to \infty} \frac{p([m^{c_1} x], [m^{c_2} y])}{m^{-(1+c_1+c_2)}} = \frac{\gamma}{\alpha + \gamma} \frac{x^{\delta_{out} - 1}}{c_1 \Gamma(\delta_{in} + 1) \Gamma(\delta_{out})} \int_0^\infty z^{-(2+1/c_1+\delta_{in}+a\delta_{out})} e^{-(\frac{x}{z} + \frac{y}{z^\alpha})} dz
\]

\[
    + \frac{\alpha}{\alpha + \gamma} \frac{x^{\delta_{in} - 1} y^{\delta}}{c_1 \Gamma(\delta_{in}) \Gamma(\delta_{out} + 1)} \int_0^\infty z^{-(1+a+1/c_1+\lambda+a\delta_{out})} e^{-(\frac{x}{z} + \frac{y}{z^\alpha})} dz
\]

\[
    = f(x, y; \alpha, \beta, \gamma, \delta_{in}, \delta_{out}) = f(x, y; \theta).
\]
2.3. Model Calibration/Fitting/Estimation

2.3.1. Issues, approaches, thoughts:

- What data is available?
  - Sometimes: Full history of edge creation with time stamps?
    * Sometimes available with real data. (Occasionally time stamps seem flakey.)
    * Full MLE methodology implemented since can write a likelihood. Works well when model is correct (eg. simulated).
- Simulate 5000 data sets with $10^5$ edges from model with $\theta = (0.3, 0.5, 0.2, 2, 1)$.

- For each data set, estimate with full MLE $\theta$.

- Make normal QQ-plot for 5000 normalized MLE estimates

- The fitted lines in black is R’s qq-line function; the red line is the 45-degree line through the origin.

- Conclude: Estimates are normal(0, 1).
Data available? (continued)

- Fixed time snapshot of the network; effectively observe at time $n$ and NOT at times $1, \ldots, n$.
  
  * Approximate [can’t write a full likelihood] MLE works well; estimators CAN but unsurprisingly there is noticeable loss of efficiency compared to MLE on full history.

  * Simulate 5000 data sets with $10^5$ edges from model with $\theta = (0.3, 0.5, 0.2, 2, 1)$.

  * For each data set, estimate $\theta$ with snapshot MLE.

  * Make normal QQ-plots for 5000 normalized MLE estimates

  * The fitted line in black is R’s qq-line function; the red line is the 45-degree line through the origin.

  * Conclude: Estimates are normal but variance increased due to loss of info.
• Should we use asymptotics to do estimation? Knowing

\[
\lim_{m \to \infty} \frac{p([mc_1x], [mc_2y])}{m^{-(1+c_1+c_2)}} = f(x, y; \theta)
\]

means asymptotic EVT estimation is an alternative to MLE. Note \(f(x, y; \theta)\) results from a double limit requiring 2 levels of pretend.

- Taking \(\lim_{n \to \infty} N_n(i, j)/N(n)\) to get \(p(i, j)\).
- Letting \(i \to \infty\) and \(j \to \infty\) in a controlled way in \(p(i, j)\).
- How close is \(EN_n(i, j)\) to \(np(i, j)\)?
- Asymptotics philosophy implemented: Use Hill estimator to estimate
  * \(\ell_{in}\);
  * \(\ell_{out}\);

and then the other parameters based on estimated angular measure corresponding to \(f(x, y; \theta)\).

- Asymptotic methods hold their own in the face of model error or data noise but inefficient compared to MLE when model is correct.
• Other issues?
  
  – Is the data from a stationary model? Some success fitting using parameters that are piecewise constant over subintervals of time. Ongoing: change point identification.
  
  – Our model of preferential attachment is linear in the in- and out-degree. Other forms of preferential attachment?
  
  – Wrestling with fitting real data to the model.
    
    * Fit struggling.
    
    * Some data have more than 3 scenarios and should have 5:
      
      \[(\alpha, \beta, \gamma, \delta, \psi)\]
      
      adding to 1.
3. Risk estimation Using Hidden Regular Variation: Strong Asyptotic Dependence

Das and Resnick (2015); Das and Resnick (2017); Das, Mitra, and Resnick (2013)

3.1. Regular variation on the first quadrant.

$Z \geq 0$ has a distribution which is regularly varying (has a multivariate heavy tail) if

- $\exists b(t) \in RV_{1/\alpha}$;
- $\exists$ limit measure $\nu(\cdot)$ on $\mathbb{R}_+^2 \setminus \{0\}$;
- As $t \to \infty$, for nice sets $A$ bounded away from 0:

$$tP\left[ \frac{Z}{b(t)} \in A \right] \to \nu(A).$$
The limit measure always concentrates on a cone $C$.

- What if $C \subseteq \mathbb{R}^2_+$?
- If $A \cap C = \emptyset$, risk estimation of being in $A$ is 0:
  \[
P[\hat{Z} \in A] \approx \frac{1}{t} \hat{\nu}(A/\hat{b}(t)) = 0.
\]
3.2. Cases

Main cases:

1. Asymptotic independence: Limit measure $\nu$ concentrates mass on $\mathbb{C} = \text{two axes}$. Results from using Gaussian copula.

\[ d = 2 \text{ and } \mathbb{C} = \text{axes} \]
\[ A = (x, \infty] = (x_1, \infty] \times (x_2, \infty] \]

and

\[ P[X \in A] = P[X_1 > x_1, X_2 > x_2] = 0. \]

Risk contagion: Can two or more components of the risk vector $X$ be simultaneously large?

- Not if the model has asymptotic independence.
- Achilles heel of the Gaussian copula.
2. **Asymptotic full dependence:** Limit measure concentrates on diagonal.
   - Hard to find data examples.

3. **Asymptotic strong dependence:** Limit measure concentrates on a narrow cone or wedge $C$.
   
   Example: Returns Exxon vs Chevron.
   
   Alternative to numerical summaries of strong dependence: coefficient of tail dependence, extremogram, EDM.
Summary and strategy.

- If the risk region $A$ is disjoint from $C$ where the limit measure $\nu(\cdot)$ concentrates, the risk estimate of

\[
P[\mathbf{Z} \in A] \approx \frac{1}{t} \hat{\nu}(A/\hat{b}(t)) = 0.
\]

- Concentration on a narrow cone is evident in mathematical and data examples; present when modeling via Gaussian copula.

- Strategy:
  - Decide that thresholded data is from model whose limit measure concentrates on a cone $C$ that is a proper subset of $\mathbb{R}_2^+$.  
  - Estimate and then remove $C$ from the state space and use remaining data to infer a 2nd (lighter) heavy tail property on $\mathbb{R}_2^+ \setminus C$.  

– Make non-zero risk estimates based on 2nd property.
– Create diagnostics to reveal:
  * Presence of 2nd heavy tail property (Hillish plot).
  * Estimated cone $C$ (Diamond plot).

• A second regular variation on $\mathbb{R}^2_+ \setminus C$ allows non-zero estimate of, for example,

$$P[Z_2 - 2a_u Z_1 > x],$$

ie, the probability of a loss when one buys

– 1 unit of security $I_2$ with risk $Z_2$ per unit; and
– sell $2a_u$ units of security $I_1$ with risk $Z_1$. 

3.3. (exxonr, chevronr)

- 1316 daily prices of Exxon and Chevron.
- October 10, 2001 to December 29, 2006 daily returns.
- Called (exxonr, chevronr).
- One expects strong dependence from two big companies engaged in similar activities.

Figure 1: Stock prices and scatterplot of Chevron and Exxon returns.
3.3.1. Diamond plots

- Map $(x, y) = (\text{exxonr, chevronr})$ onto $L_1$ unit sphere after discarding points below a threshold value of $x + y$.

- Use
  
  $$(x, y) \mapsto \left( \frac{x}{|x| + |y|}, \frac{y}{|x| + |y|} \right) = \theta = (\theta_1, \theta_2).$$

  from
  
  $$\mathbb{R}^2 \mapsto \mathbb{N}_0 = [\text{diamond}] \subset \mathbb{R}^2.$$

- where the $L_1$ unit sphere is

  $$[\text{diamond}] = \{ (\theta_1, \theta_2) : |\theta_1| + |\theta_2| = 1 \}.$$  

- Experiment with mapping at various thresholds determined by $k$, the number of order statistics of the norms $|x| + |y|$.

- Use thresholds $k = 400$ and $k = 70$.

- Model for the angular measure $S$ of limit measure $\nu$ is that $S$ concentrates in the first and third quadrants.

- Use range of $\theta_1$ in these quadrants as estimators. Get

  1. for the first quadrant

     $$(\hat{\theta}_1, \hat{\theta}_2) = (0.312, 0.701)$$
and

2. in the third quadrant

\[(\hat{\theta}_1, \hat{\theta}_2) = (-0.814, -0.284).\]

• These \(\hat{\theta}\)'s correspond to slopes of rays in Cartesian coordinates of \((\hat{a}_1, \hat{a}_2) = (0.429, 2.226)\) for the first quadrant.

Figure 2: Empirical angles (diamond plot) for 400 largest values under \(L_1\) norm for (exxonr,chevronr) with histogram (left two plots) and the same for 70 largest values (right two plots).
3.3.2. Ongoing:

- With post-doc: algorithms for identifying cone of concentration.
- How to search in higher dimensions.
- How to overcome data reduction resulting from thresholding.
- Quality of risk estimates?
References


B. Das, A. Mitra, and S.I. Resnick. Living on the multi-dimensional


