MURI Update Meeting
Multivariate Heavy Tail Phenomena: Modeling and Diagnostics
October 16, 2015,
Columbia University,

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

http://people.orie.cornell.edu/∼sid
607 255 1210   sir1@cornell.edu

October 15, 2015

1. Highly Dependent Models

1.1. Multivariate heavy tails

$$
\begin{aligned}
\boldsymbol{X} &= \text{multivariate vector in } \mathbb{R}_+^p \quad (\text{here } p = 2) \\
A &= \text{subset in } \mathbb{R}_+^p \\
b(t) &= \text{scaling function} \\
\nu &= \text{limit measure on subsets on } \mathbb{R}_+^p
\end{aligned}
$$

and $\boldsymbol{X}$ has a multivariate regularly varying distribution if as $t \to \infty$,

$$tP[\boldsymbol{X}/b(t) \in A] \to \nu(A). \tag{1}$$

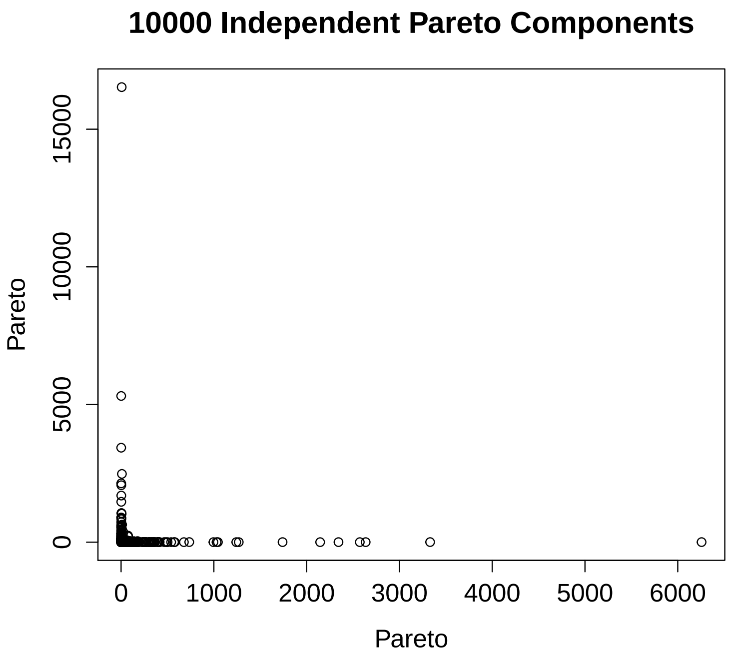
Resolved issues:

- What does "$\to$" mean? How do you define convergence?
- What sets A can we put in (1)?
- Such sets define *tail regions*.

Follow up Das and Resnick (2015).

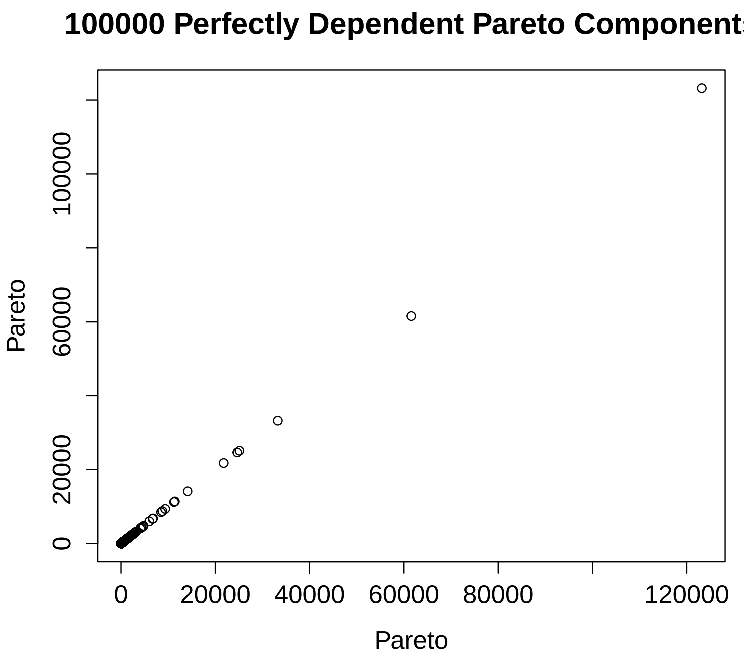
1.2. Extreme cases of $\nu(\cdot)$.

1.2.1. Asymptotic independence: ν concentrates on the axes. Mystery Data 1.



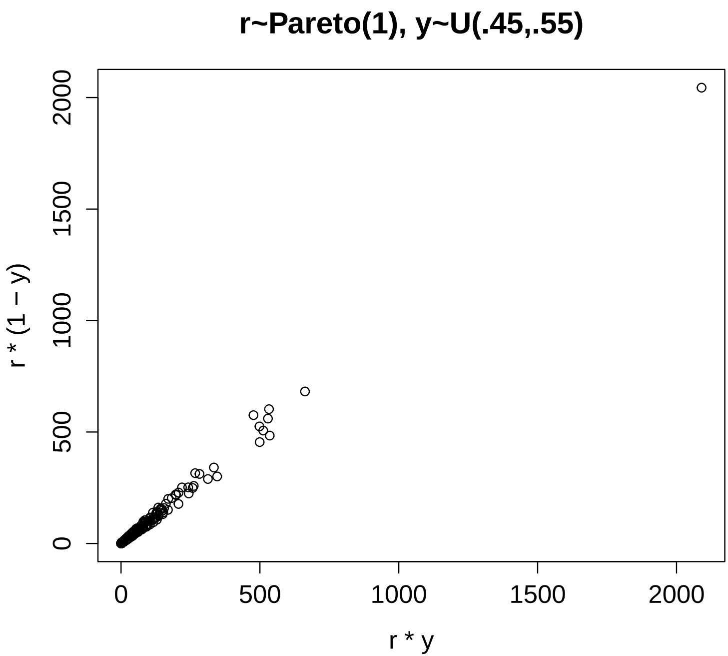
Typical of Gaussian dependence copulas.

1.2.2. Asymptotic full dependence: ν concentrates on the diagonal. Mystery Data 2.



Typical of exchange rate return data, eg. (Chinar,Australiar) vs USD.

1.2.3. Asymptotic high dependence: ν concentrates on a narrow cone about the diagonal. Mystery Data 3.



1.2.4. What's common?

- Limit measure concentrates on a small region of the state space, the support.
- If a risk region A is disjoint from the support, we estimate the risk probability as 0.
- Frequently there is a 2nd multivariate heavy tail regime on

$$\mathbb{R}_+^p \setminus [\text{support}].$$

Gives potential for improved risk estimates.

1.3. Strongly dependent data: Exxon vs Chevron returns.

Raw data highly dependent.

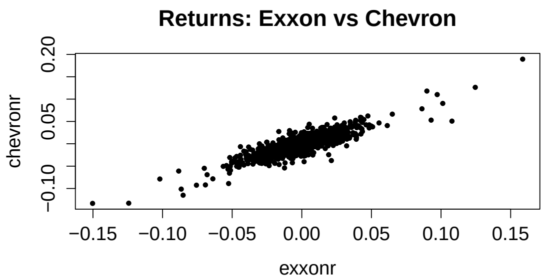
- 3124 (positive & negative) returns

  (exxonr,chevronr)

  from daily prices 2001–2014.

- Tail α's all approx 2.5.

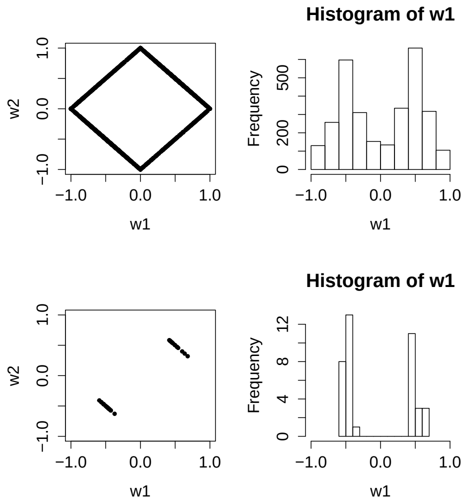
- Scatterplot shows high degree of dependence.



Dependence analysis and diamond graph (L_1-unit sphere):

$$(x, y) \mapsto \left(\frac{x}{|x| + |y|}, \frac{y}{|x| + |y|} \right) = \mathbf{w} = (w_1, w_2),$$

1.3.1. Empirical angles for 3000 largest values of the L_1 norm in $\mathbb{R}^2$ (left)and for the 40 largest values (right) .



Evidence from *Hillish plot* of a second heavy tail regime on

$$\mathbb{R}_+^2 \setminus [\text{narrow cone about diagonal}].$$

2. Asymptotic normality for node counts

2.1. Undirected model

Resnick and Samorodnitsky (2016)
Rules of attachment: Conditional on knowing the graph G_n, at stage $n+1$ a new node $n+1$ appears and with a parameter $\delta > -1$, either

1. The new node $n+1$ attaches to $v \in V_n$ with probability
$$\frac{D_n(v)+\delta}{n(2+\delta)+(1+\delta)}, \tag{2}$$
or
2. $n+1$ attaches to itself with probability
$$\frac{1+\delta}{n(2+\delta)+(1+\delta)}. \tag{3}$$

2.1.1. Known: SLLN for node frequencies.

For

$$N_n(k) = \ \#\text{ nodes with degree } k$$

we have as $n \to \infty$

$$\frac{N_n(k)}{n} \to p_k.$$

2.1.2. CLT:

$$\Big(\sqrt{n}\Big(\frac{N_n(k)}{n} - p_k\Big),\ k = 1, 2, \dots\Big) \Rightarrow \big(Z_k,\ k = 1, 2 \dots\big)$$

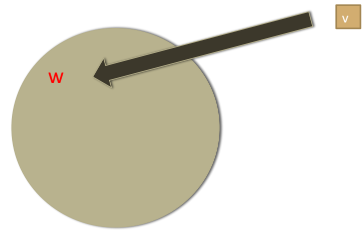
where $\big(Z_k,\ k = 1, 2 \dots\big)$ is a centered Gaussian process with covariance function R_Z that you won't love but can teach a computer to love.

2.2. Directed model

Wang and Resnick (2015).

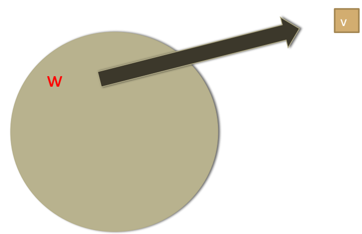
1. With probability α, append to $G(n-1)$ a new node $v \notin V_{n-1}$ and create directed edge $v \mapsto w \in V_{n-1}$ with probability

$$\frac{D_{in}(v) + \lambda}{(1+\lambda)n}.$$



2. With probability γ, 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

$$\frac{D_{out}(v) + \mu}{(1+\mu)n}.$$



2.2.1. Known: SLLN for node counts.

For

$$N_n(i,j) = \ \#\text{ nodes with in-degree } i, \text{ out-degree } j$$

we have as $n \to \infty$

$$\frac{N_n(i,j)}{n} \to p_{i,j}.$$

2.2.2. CLT:

Fix positive integers I, O. Provided that K_{IO} is invertible, we have

$$\left(\sqrt{n}\left(\frac{N_n(i,j)}{n} - p_{ij}\right) : 0 \le i \le I,\, 0 \le j \le O\right) \Rightarrow N(0, K_{IO}^{-1}\Sigma_{IO}K_{IO}^{-T}).$$

Matrices K_{IO} and Σ_{IO} are are specified in the paper and K_{IO} must be invertible. Expressions are unlovable but can be taught to a computer.

2.3. Follow-on

Apply to more formal inference methods.

3. pmf vs distribution

3.1. Known:

Samorodnitsky, Resnick, Towsley, Davis, Willis, and Wan (2016), Resnick and Samorodnitsky (2015)
For
$$N_n(i,j) = \text{ \# nodes with in-degree } i, \text{ out-degree } j$$
we have as $n \to \infty$
$$\frac{N_n(i,j)}{n} \to p_{i,j},$$
where $\{p_{i,j}\}$ is a pmf.

Let $(I,O) \sim \{p_{i,j}\}$. Then (I,O) has a regularly varying distribution:
$$tP\Big[\Big(\frac{I}{b_I(t)}, \frac{O}{b_O(t)}\Big) \in A\Big] \to \nu(A),$$
where $\nu(\cdot)$ is a measure with explicit continuous density on $\mathbb{R}_+^2$.

What is the asymptotic behavior of $p_{i,j}$?

Wang & Resnick (in progress).

- If $\{p_{i,j}\}$ is a regularly varying pmf satisfying something (eg. monotonicity) then $\{p_{i,j}\}$ is *embeddable* in a continuous pdf which is also regularly varying.
- A regularly varying density which is monotone generates a regularly varying measure.
- A regularly varying measure with a monotone density or monotone mass function means the density or mass function is regularly varying.

Note:

1. At least for some special cases, $\{p_{i,j}\}$ in the preferential attachment model is (eventually) monotone.
2. What is a minimal set of conditions or most useful set of conditions to make this all work.

4. Min K-S method and slashdot

Clauset, Shalizi, and Newman (2009)
Back to the Clauset method in one dimension: Can do for continuous version or (harder) discrete.

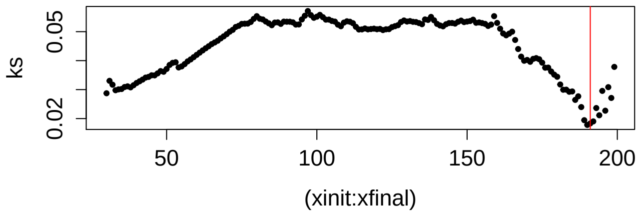
J. Sun & Resnick (preliminary).

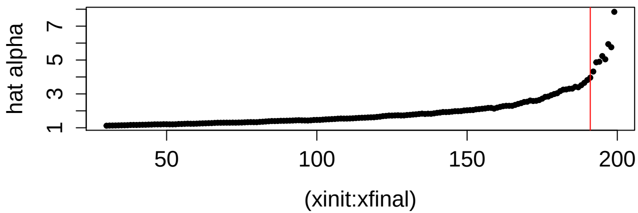
- Fix left endpoint x_l and fit Pareto $\hat{\alpha}$ on $[x_l, \infty)$ by, say, MLE. Note $\hat{\alpha}$ is a function of x_l.
- Compute the K-S distance between the fitted Pareto and the empirical cdf. This distance is function of x_l.
- Minimize the K-S distance over x_l.
- Report that value of $\hat{\alpha}$.
- Can be adapted to censoring on the right.
- Haven't thought yet about higher dimensions.

4.1. Slashdot data

- Recommender network.
- # followers ≤ 200 (unless you pay).
- Study what corresponds to (I, O), in- and out-degree where out-degree is censored.
- Ignore discrete nature of data.

```
minks(0,30,199)
min ks estimate of alpha is  3.958039
min ks at left endpoint= 191
```





References

A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, 2009. ISSN 0036-1445. doi: 10.1137/070710111. URL http://dx.doi.org/10.1137/070710111.

B. Das and S.I. Resnick. Models with hidden regular variation: Generation and detection. *Stochastic Systems*, 0:1–44, 2015. doi: 10.1214/14-SSY141. URL http://adsabs.harvard.edu/abs/2014arXiv1403.5774D.

S.I. Resnick and G. Samorodnitsky. Tauberian Theory for Multivariate Regularly Varying Distributions with Application to Preferential Attachment Networks. *Extremes*, 18(3):349–367, 2015. doi: 10.1007/s10687-015-0216-2. URL http://adsabs.harvard.edu/abs/2014arXiv1406.6395R.

S.I. Resnick and G. Samorodnitsky. Asymptotic normality of degree counts in a preferential attachment model. *Journal of Applied Probability*, 2016.

G. Samorodnitsky, S. Resnick, D. Towsley, R. Davis, A. Willis, and P. Wan. Nonstandard regular variation of in-degree and out-degree in the preferential attachment model. *Journal of Applied Probability*, 53(1), March 2016. http://arxiv.org/pdf/1405.4882.pdf.

T. Wang and S. I. Resnick. Asymptotic Normality of In- and Out-Degree Counts in a Preferential Attachment Model. *ArXiv e-prints*, October 2015.