Hidden Risk Estimation, Network Model Calibration

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

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

MURI Natick Nov 21, 2016

November 15, 2016

1. Outline

- Hidden regular variation (HRV): a semi-parametric asymptotic approximation method for improving risk estimates.
  - Case 1: Asymptotic independence of variables as in the Gaussian copula dependence model.
  - Case 2: Strong dependence or full asynptotic dependence.
- Preferential attachment as a model for social network growth.
  - Understanding the multivariate heavy tail of (in,out)-degree.
  - Simulation of preferential attachment growing networks.
  - Statistical analysis social network data and calibration of a linear preferential attachment model.
- No time: Threshold selection by the minimum distance method;
  - the limitations of the Clauset (Virkar and Clauset (2014)) method.
  - When doing tail estimation, what portion of the data should be used?

CORNELL

2. Hidden Regular Variation: Asymptotic Independence and Strong Asyptotic Dependence

Das and Resnick (2015); Das and Resnick (2016); Das, Mitra, and Resnick (2013)

2.1. Regular variation on the first quadrant.

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

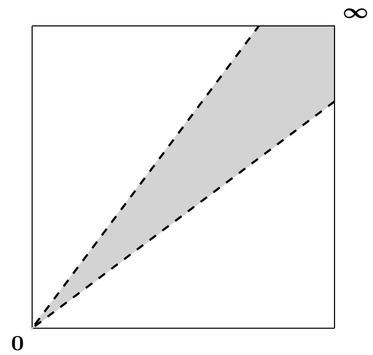
- $\exists\, b(t) \in RV_{1/\alpha}$;
- $\exists$ limit measure $\nu(\cdot)$ on $\mathbb{R}_+^2 \setminus \{\mathbf{0}\}$;
- As $t \to \infty$, for nice sets A bounded away from $\mathbf{0}$:

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

The limit measure always concentrates on a cone $\mathbb{C}$.

- What if $\mathbb{C} \subsetneq \mathbb{R}_+^2$?
- If $A \cap \mathbb{C} = \emptyset$, risk estimation of being in A is 0:

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



2.2. Cases

Consider cases:

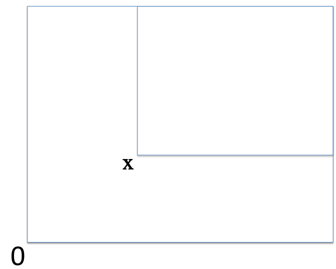
1. **Asymptotic independence:** Limit measure ν concentrates mass on $\mathbb{C}$ = two axes. Results from using Gaussian copula.

$d = 2$ and $\mathbb{C}$ = axes and

$$A = (\mathbf{x}, \boldsymbol{\infty}] = (x_1, \infty] \times (x_2, \infty]$$

and

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

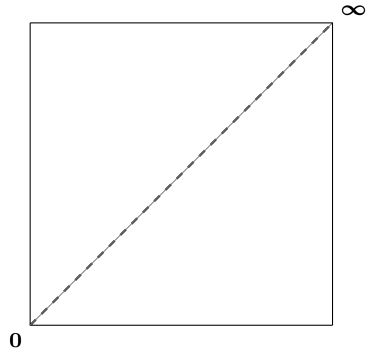


Risk contagion: Can two or more components of the risk vector $\boldsymbol{X}$ be simultaneously large?

- Not if the model has asymptotic independence.
- This is the Achilles heel of the Gaussian copula.

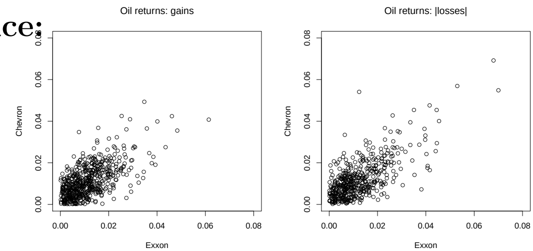
2. **Asymptotic full dependence:** Limit measure concentrates on diagonal.

    - Hard to find data examples.



3. **Asymptotic strong dependence:** Limit measure concentrates on a narrow cone or wedge $\mathbb{C}$.

    Example: Returns Exxon vs Chevron.



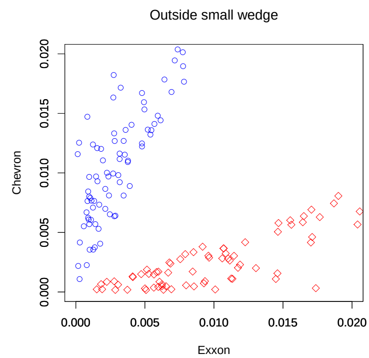
Summary and strategy.

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

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

- Concentration on a narrow cone is evident in many 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 $\mathbb{C}$ that is a proper subset of $\mathbb{R}_+^2$.
  - Estimate and then remove $\mathbb{C}$ from the state space and use remaining data to infer a 2nd (lighter) heavy tail property on $\mathbb{R}_+^2 \setminus \mathbb{C}$.



Outline
Hidden Regular . . .
Preferential . . .
Threshold
Title Page
Page 7 of 25
Go Back
Full Screen
Close
Quit

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

• A second regular variation on $\mathbb{R}_+^2 \setminus \mathbb{C}$ 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.

2.3. (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.

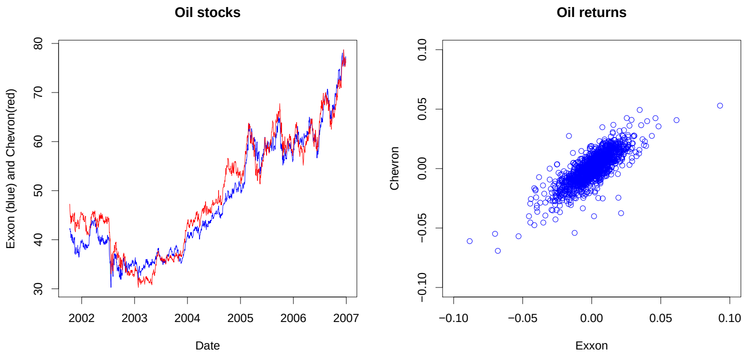


Figure 1: Stock prices and scatterplot of Chevron and Exxon returns.

Outline
Hidden Regular . . .
Preferential . . .
Threshold
Title Page
Page 9 of 25
Go Back
Full Screen
Close
Quit

2.3.1. Diamond plots

- Map (x, y) =(exxonr,chevronr) onto L_1 unit sphere after discarding points below a threshold value of $\mathtt{x}+\mathtt{y}$.
- Use
$$(x, y) \mapsto \left(\frac{x}{|x|+|y|}, \frac{y}{|x|+|y|}\right) = \boldsymbol{\theta} = (\theta_1, \theta_2).$$
from
$$\mathbb{R}^2 \mapsto \aleph_{\mathbf{0}} = [\text{diamond}] \subset \mathbb{R}^2.$$
- 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 ν is that S concentrates in the first and third quadrants.
- Use range of θ_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.

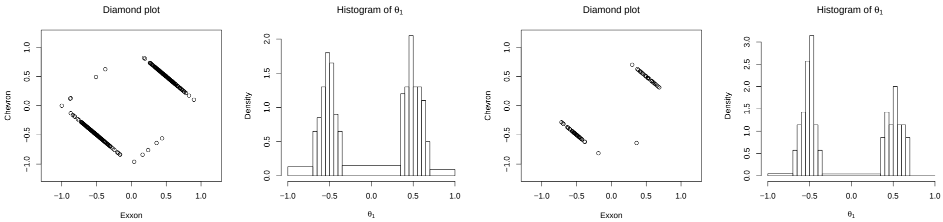


Figure 2: Empirical angles (diamond plot) for 400 largest values under L_1 norm for (exxonr,chevronr) with histogram (left two plots) and the same for 70 largest values (right two plots).

3. Preferential Attachment as Model for Network Growth

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

3.1. A model

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

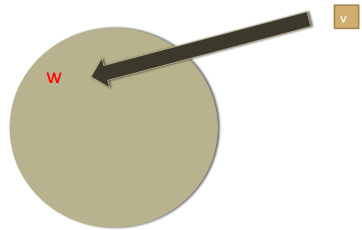
- Model parameters: $\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}}$ with $\alpha + \beta + \gamma = 1$.
- $G(n) = (V_n, E_n)$ is a directed random graph with n edges, $N(n)$ nodes, node set V_n and edge set

$$E_n = \{(u, v) \in V_n \times V_n : (u, v) \in E_n\}.$$

- Node degree:
  - In-degree of v in $G(n)$ is $D_{\text{in}}^{(n)}(v)$;
  - Out-degree of v in $G(n)$ is $D_{\text{out}}^{(n)}(v)$.
- Obtain graph $G(n)$ from $G(n-1)$ in a Markovian way as follows:

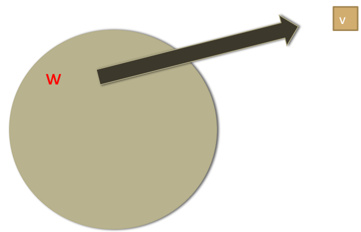
1. α scenario: With probability α, append to $G(n-1)$ a new node $v \notin V_{n-1}$ and create directed edge $v \mapsto w \in V_{n-1}$ with probability

$$\frac{D_{\text{in}}^{(n-1)}(w) + \delta_{\text{in}}}{n - 1 + \delta_{\text{in}} N(n-1)}.$$



2. γ scenario: 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

$$\frac{D_{\text{out}}^{(n-1)}(w) + \delta_{\text{out}}}{n - 1 + \delta_{\text{out}} N(n-1)}.$$

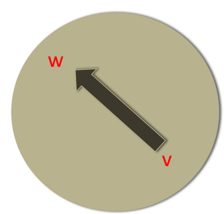


3. β scenario: With probability β, create new directed edge between existing nodes

$$v \in V_{n-1} \mapsto w \in V_{n-1}$$

with probability

$$\left(\frac{D_{\text{out}}^{(n-1)}(v) + \delta_{\text{out}}}{n - 1 + \delta_{\text{out}} N(n-1)}\right)\left(\frac{D_{\text{in}}^{(n-1)}(w) + \delta_{\text{in}}}{n - 1 + \delta_{\text{in}} N(n-1)}\right)$$



3.2. 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

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

3.2.1. Marginal behavior.

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

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

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

where

$$\alpha_{\text{in}} = 1 + \frac{1+\delta_{\text{in}}(\alpha+\gamma)}{\alpha+\beta}, \quad \alpha_{\text{out}} = 1 + \frac{1+\delta_{\text{out}}(\alpha+\gamma)}{\gamma+\beta}.$$

3.2.2. Joint behavior.

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

Set

$$c_1 = \frac{1}{\alpha_{\text{in}} - 1}, \quad c_2 = \frac{1}{\alpha_{\text{out}} - 1}, \quad a = c_2/c_1.$$

For $x > 0$, $y > 0$,

$$\lim_{m\to\infty} \frac{p([m^{c_1}x],[m^{c_2}y])}{m^{-(1+c_1+c_2)}} = \frac{\gamma}{\alpha+\gamma} \frac{x^{\delta_{in}} y^{\delta_{\text{out}}-1}}{c_1\Gamma(\delta_{\text{in}}+1)\Gamma(\delta_{\text{out}})} \int_0^\infty z^{-(2+1/c_1+\delta_{in}+a\delta_{\text{out}})} e^{-\left(\frac{x}{z}+\frac{y}{z^a}\right)} \mathrm{d}z$$
$$+ \frac{\alpha}{\alpha+\gamma} \frac{x^{\delta_{\text{in}}-1} y^{\delta}_{\text{out}}}{c_1\Gamma(\delta_{\text{in}})\Gamma(\delta_{\text{out}}+1)} \int_0^\infty z^{-(1+a+1/c_1+\lambda+a\delta_{\text{out}})} e^{-\left(\frac{x}{z}+\frac{y}{z^a}\right)} \mathrm{d}z$$
$$= f(x,y;\alpha,\beta,\gamma,\delta_{in},\delta_{\text{out}}) = f(x,y;\boldsymbol{\theta}).$$

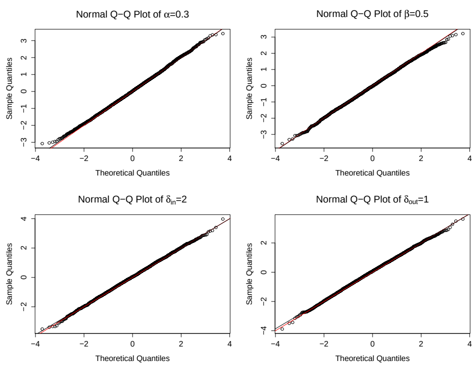
3.3. Model Calibration/Fitting/Estimation

Issues, approaches, thoughts:

- Should we use asymptotics to do estimation? Note $f(x, y; \boldsymbol{\theta})$ results from essentially a double limit:
  - 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 $p(i,j)$.
  - Asymptotics philosophy can be implemented and requires using $f(x, y; \boldsymbol{\theta})$. Could use tail methods to estimate
    * α_{in};
    * α_{out};

    and then the other parameters based on estimated angular measure corresponding to $f(x, y; \boldsymbol{\theta})$.
  - Asymptotic methods would be more robust against inevitable model error but suffer in accuracy compare to model based estimation when the model is correct (ie simulated).

- What data is available?
  - Full history of edge creation with time stamps?
    - Available when simulate network (Atwood, Ribeiro, and Towsley (2015), J. Roy, P. Wan)
    - Available with real data; time stamps can be unreliable.
    - Full MLE methodology implemented and works well when model is correct (simulated).



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

· For each data set, estimate with full MLE $\boldsymbol{\theta}$.

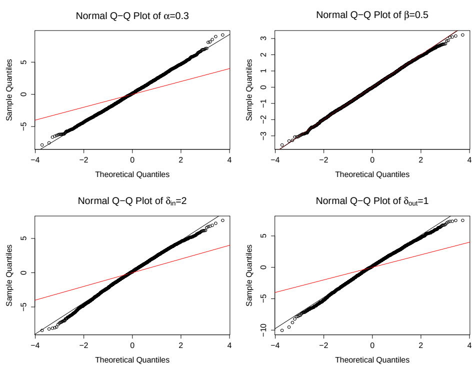
· Make normal QQ-plot for 5000 normalized MLE estimates

· 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 $1, \dots, n$.
  - ∗ MLE (approximate) still 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 $\boldsymbol{\theta} = (0.3, 0.5, 0.2, 2, 1)$.

∗ For each data set, estimate $\boldsymbol{\theta}$ 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.

- Other issues?
  - Is the data from a stationary model? Some success fitting using piecewise parameters that are piecewise constant over time.
  - 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.

4. Minimum distance threshold selection

- For heavy tailed data, what part of the data should be used?
- Rule: use k upper order statistics.
- Clausett method (Clauset et al. (2009); Virkar and Clauset (2014))
  - With data $X_1, \dots, X_n$ 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:

$$D_k := \sup_{y \geq 1} \left| \frac{1}{k} \sum_{i=1}^{n} \epsilon_{X_i/X_{(k)}}(y, \infty] - y^{-\hat{\alpha}(k)} \right|, \quad 1 \leq k \leq n.$$

Choose the optimal k^* as the one that minimizes the KS distance, that is,

$$k^* := \mathrm{argmin}_{k \in \mathcal{I}} D_k,$$

- But: If data is really Pareto $k^* \sim cn$ so what is the point?
- If data is Pareto but only from some point on, still have the challenge of finding the endpoint. The min distance method does a reasonable job.
- If data is heavy tailed but not Pareto? Not clear this works in the case of second order regular variation (eg. stable).

Contents

References

J. Atwood, B. Ribeiro, and D. Towsley. Efficient network generation under general preferential attachment. *Computational Social Networks*, 2(1):7, 2015. ISSN 2197-4314. doi: 10.1186/s40649-015-0012-9. URL http://dx.doi.org/10.1186/s40649-015-0012-9.

B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, 2003)*, pages 132–139, New York, 2003. ACM.

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*, 5(2):195–238, 2015. ISSN 1946-5238. doi: 10.1214/14-SSY141.

B. Das and S.I. Resnick. Hidden regular variation under full and strong asymptotic dependence. *ArXiv e-prints*, February 2016. To appear: Extremes.

B. Das, A. Mitra, and S.I. Resnick. Living on the multi-dimensional edge: Seeking hidden risks using regular variation. *Advances in Applied Probability*, 45(1):139–163, 2013.

P.L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123:1–14, 2001.

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.

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):146–161, March 2016. doi: 10.1017/jpr.2015.15.

Y. Virkar and A. Clauset. Power-law distributions in binned empirical data. *Ann. Appl. Stat.*, 8(1):89–119, 2014. ISSN 1932-6157. doi: 10.1214/13-AOAS710. URL `http://dx.doi.org/10.1214/13-AOAS710`.

P. Wan, T. Wang, R. Davis, and S. Resnick. Calibrating the linear preferential attachment model. Technical report, Cornell University, 2016. In preparation.

T. Wang and S.I. Resnick. Multivariate regular variation of discrete mass functions with applications to preferential attachment networks. *Methodology and Computing in Applied Probability*, 2016. ISSN 1573-7713. doi: 10.1007/s11009-016-9503-x. URL http://dx.doi.org/10.1007/s11009-016-9503-x.

CORNELL

Title Page

Page 25 of 25

Go Back

Full Screen

Close

Quit