the long tau-path for detecting monotone association in an unspecified subpopulation

Joe Verducci

Current Challenges in Statistical Learning Workshop
Banff International Research Station
Tuesday, December 13, 2011

Joint work with Stephen Bamattre

Outline

1. Basic Tau Path
    1. motivation and successes
    2. estimation of an ordering
    3. (screening) test for association
2. Simulations to Learn the Functional Form of the Boundary for Large Samples
3. Normal Extension of the Beta Distribution
4. Continuous formulation and role of the Beta

Problem and Examples

Develop a general method for detecting *monotone association* in an *unspecified subpopulation.*

Examples:

1. Cancer cells in which gene expression controls chemo-resistance
   A. Activity of Quassinoids correlated with expression of the IGFBP6 gene over Ovarian and CNS cell-lines
   B. Newly differentiated cell-lines form an unanticipated subgroup over which several associations are discovered.
2. Subpopulations of 210 US Designated Marketing Areas
   A. sales are associated with specific types of advertizing campaigns, but only in certain DMAs
   B. retention is associated with economic descriptors, in some DMAs
   C. Results are replicated in different years, which some changes in the subpopulations involved.

Basic Tau Path Formulation as an estimation problem

Let $\mathcal{F}$ be a family of bivariate copula models (joint distributions with uniform margins) indexed by the population Kendall τ, defined by

$$\tau = E[\text{sign}(X_i - X_j) * \text{sign}(Y_i - Y_j)]$$

when (X_i, Y_i) and (X_j, Y_j) are independently sampled from $F_\tau \in \mathcal{F}$

Underlying Model

$$(X_i, Y_i) \sim F_{\tau_i} \quad , \; i = 1,\ldots,n$$

where

$$\tau_{\pi(1)} \leq \cdots \leq \tau_{\pi(n)}$$

for some permutation π: {1,…,n} → {1,…,n} .

The problem is to estimate π based on the sample $\{(X_i, Y_i) \mid i = 1,\ldots,n\}$

Estimation is driven by the idea of concordance, which underlies Kendall's τ.

Concordance

$\{(X_i, Y_i) \mid i=1,\ldots,n\}$ are assumed to be continuous variables, with no two values exactly equal.

Two observations *i* and *j*

are concordant with respect to *(X,Y)* if

$$(Y_i - Y_j)(X_i - X_j) > 0$$

and discordant if

$$(Y_i - Y_j)(X_i - X_j) < 0$$

Kendall's τ Correlation Coefficient

- Definition: Let

  N = ${}_nC_2$ = n(n-1)/2 be the number of distinct pairs of observations in the sample of size *n*.

  C be the number of concordant pairs

  D be the number of discordant pairs

  Then $$t = \frac{C - D}{N}$$

  is Kendall's Coefficient for the sample.

Kendall's Test for Independence

- Data: A pair of variables (X, Y) measured on a sample from a population.
- Null Hypothesis (H_0): X and Y are independent in the population
- Alternative Hypothesis H_A: not all $\tau_i = 0$
- Test Statistic: Kendall's t Coefficient for the sample
- Sampling Distribution: of Kendall's t Coefficient under H_0 is well known.
- Rejection Rule: Reject if test statistic is unusually high or low

Extend Kendall's Test

- Idea: rearrange the order of observations so that the sets of observations over which X and Y are most concordant appear early on in the re-sequenced data.
- We need to look carefully at the details of how Kendall's statistic is calculated.

Concordance Matrix

For a fixed (X, Y) pair, let

$\mathbb{C} = [C_{ij}] = n \times n$ concordance matrix

C_{ij} = *1* if cell-lines *i* and *j* are concordant

= *-1* if cell-lines *i* and *j* are discordant

= *0* if $i = j$

Notes:

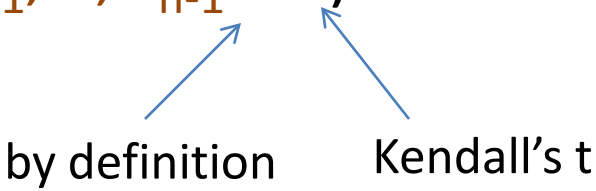
– $\mathbb{C}$ is symmetric.

– Kendall's t coefficient

= average of the off diagonal elements of $\mathbb{C}$.

tau–path

- Let $\mathbb{C}_k$ denote the *k x k* matrix formed from the first *k* rows and columns of $\mathbb{C}$.
- The average t_k of the off diagonal elements of $\mathbb{C}_{k+1}$ is Kendall's τ coefficient of correlation for the subpopulation represented by the first *k+1* cell lines.
- The sequence $(t_1, \ldots, t_{n-1} = t)$ is called a tau-path.



Ordering of Observations

We would like to re-order the cell-lines to make the tau-path ($t_1,\ldots,t_{n-1}$) decreasing.

$$\tau_1 \geq \cdots \geq \tau_{n-1}$$

Example

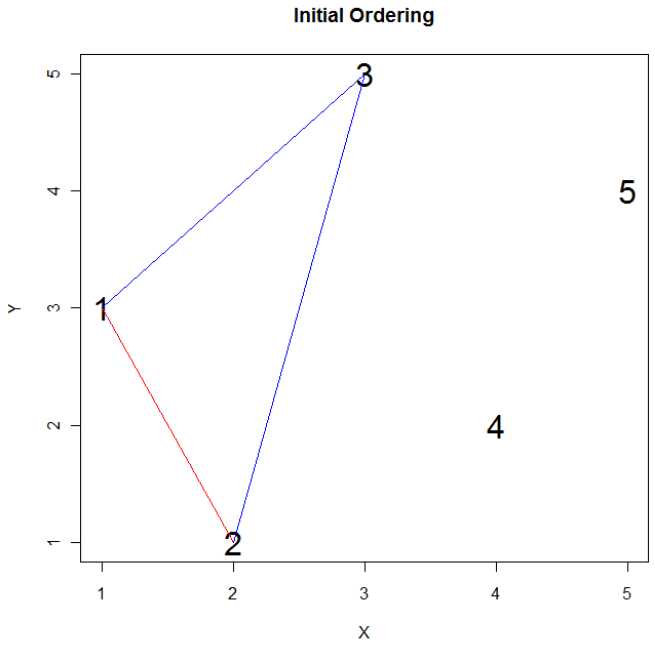
Initial Ordering:

X = [1 2 3 4 5]

Y = [3 1 5 2 4]

Tau-path starts

(-1,1/3,…)



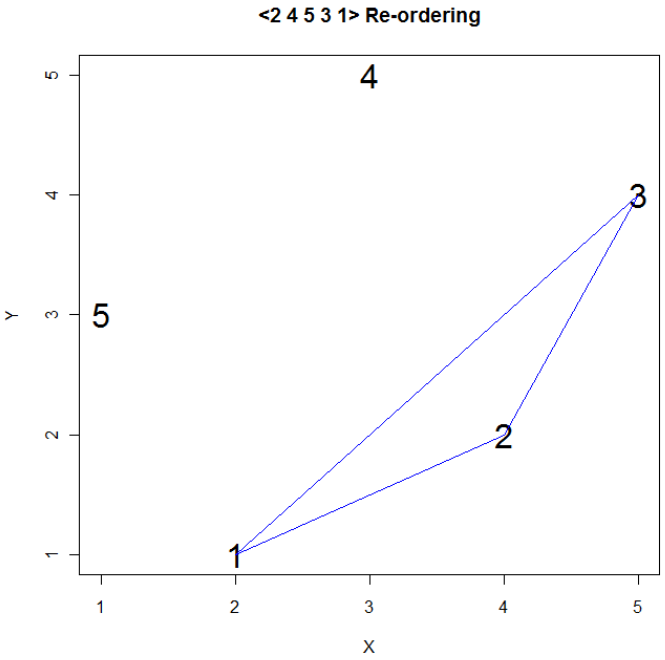
Re-Ordering of Observations

<2 4 5 3 1> Re-ordering

X* = [2 4 5 3 1]

Y* = [1 2 4 5 3]

Tau path starts (1,1,…)

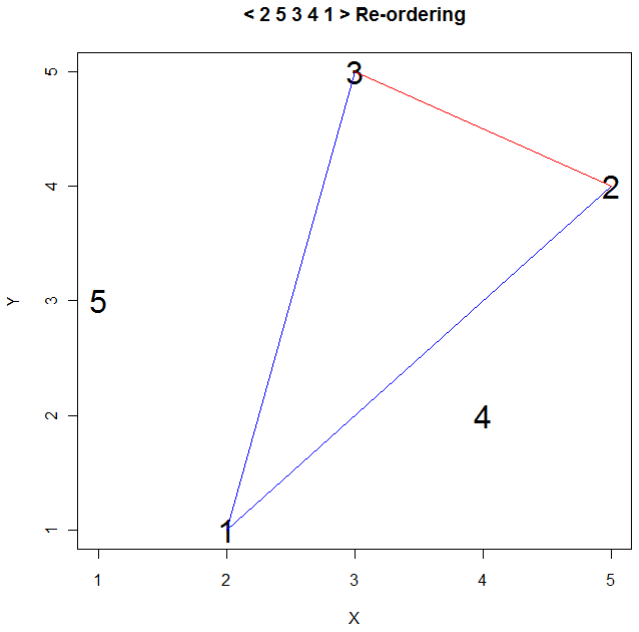


<2 5 3 4 1> Re-ordering

X** = [2 5 3 4 1]

Y** = [1 4 5 2 3]

Tau path starts (1,1/3,…)



Basic Backdrop Algorithm

1. Set $C_0 = C$, the original concordance matrix.
2. At step j ($j = 1, \ldots, n-1$)
   A. Let m_j be the observation corresponding to the row/column of C_{j-1} with the minimal sum
   B. Set $\pi(n+1-j) = m_j$
   C. Let C_j be the matrix formed by removing row and column m_j of C_{j-1} .
3. Set $\pi(n)$ = the last remaining observation number.

The average of the entries in C_j is non-decreasing in j.

Other good properties come from handling ties carefully.

Example

- X = [1 2 3 4 5] Y = [3 1 5 2 4]

Concordance Matrix

0	-1	1	-1	1
-1	0	1	1	1
1	1	0	-1	-1
-1	1	-1	0	1
1	1	-1	1	0

τ–path:
(-1,1/3,0,1/5)

<2 4 5 3 1> Reordered Concordance Matrix

0	1	1	1	-1
1	0	1	-1	-1
1	1	0	-1	1
-1	-1	1	0	1
1	-1	-1	1	0

τ–path:
(1,1,1/3,1/5)

More Refined Algorithms

- Fast Backward Conditional Search (BCS) Algorithm
  - Very quick estimation of π
  - "locally admissible" not dominated by nearby estimates
  - Good for screening to detect any significant association
- Full BCS Algorithm for estimating π
  - Slower, but still reasonably fast for a few pairs
  - Admissible
  - Good for estimating associated subset after detecting high association.
- For details, get a copy of the paper
  - Yu, L., Verducci, J. and Blower, P. "The Tau-Path Test for Monotone Association in an Unspecified Subpopulation: Applications to Chemogenomic Data Mining," *Statistical Methodology* Special Issue on Data Mining, 2011.

Nonparametric Development

- Non-parametric Mixture Model
  - Two "association-homogeneous" components
    1. (X,Y) have common population τ^*
    2. (X,Y) are independent
  - Proportion of observations in which (X,Y) are associated is not known
- Hypothesis Test
  - Null Hypothesis is that (X,Y) are independent
  - Test should be sensitive to this particular type of diversion from independence.

Fig 1a. Mixture of normals: 35 correlated and 65 uncorrelated

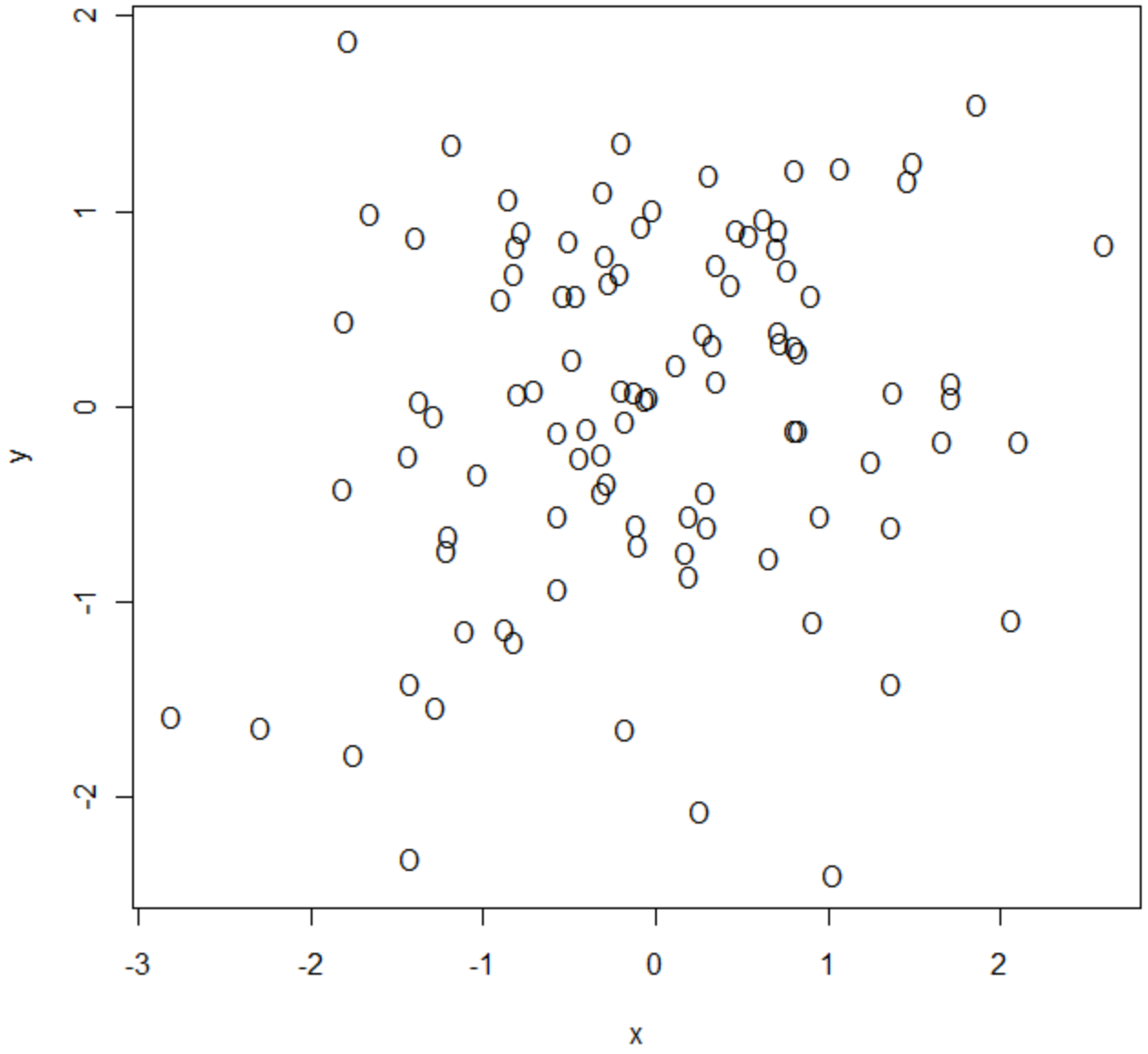
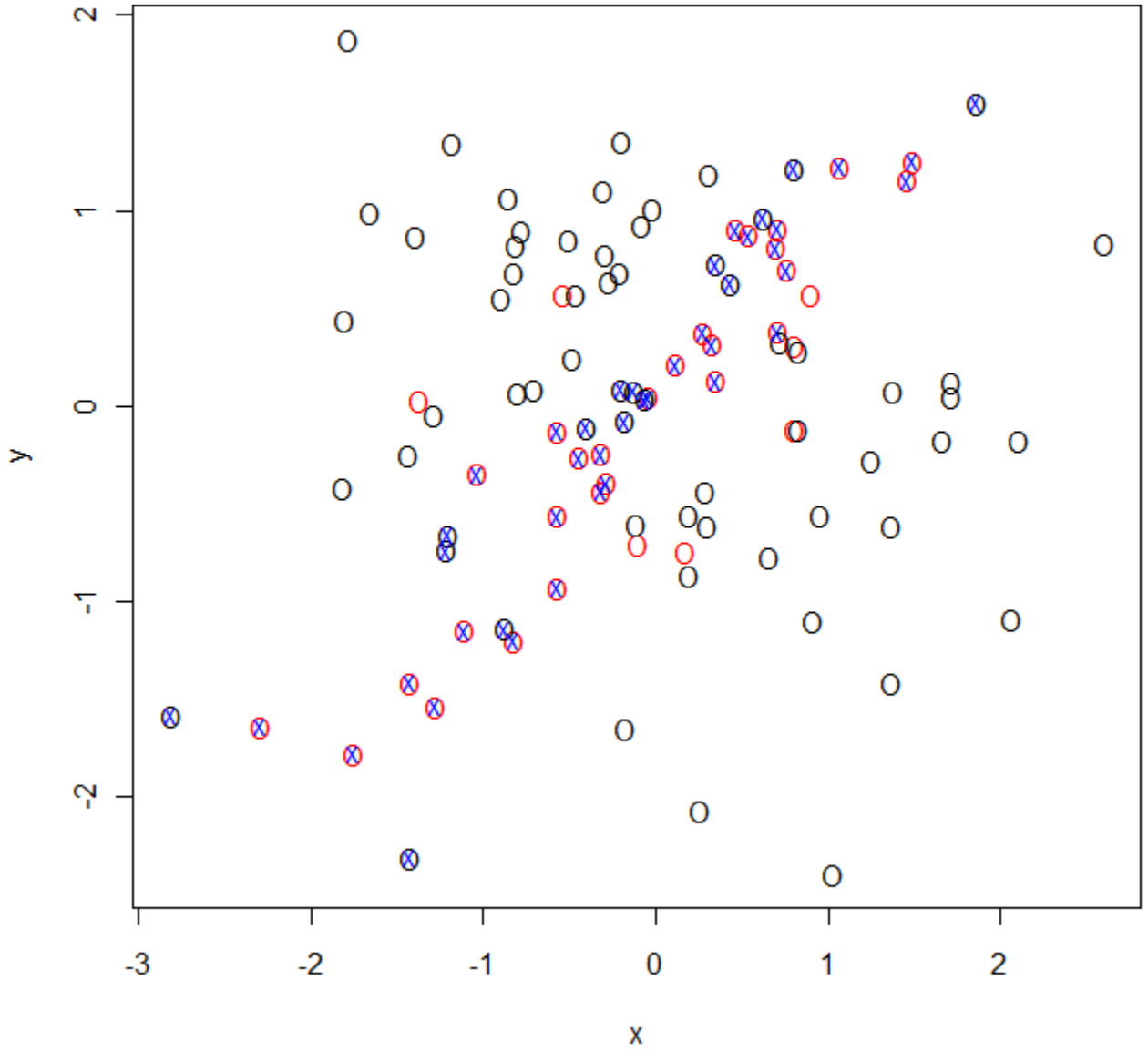


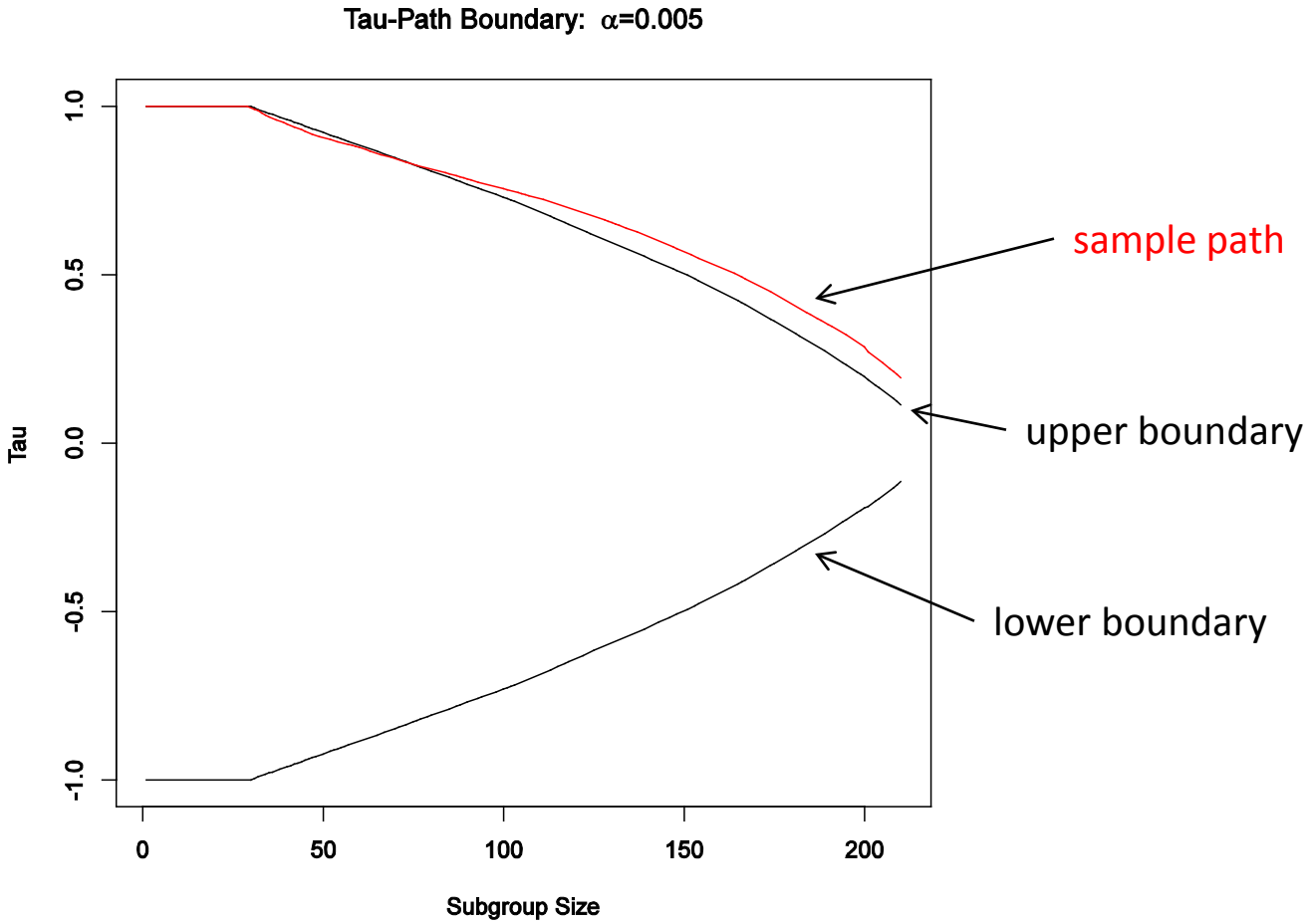
Fig 1b. 35 correlated obs. in red; predicted marked by blue x



Tau Path: Rejection Bounds for the test

- Generate a large number of independent *n*-dimensional samples from two independent populations.
- Estimate π, and the resulting ordered Tau-Path, for each sample.
- Construct $(1-\alpha)^{th}$ percentile of the sampling distribution of each element in ordered Tau-Path to obtain *upper boundary.*
- Reflect about axis to obtain *lower boundary,* which will be used to detect a decreasing relationship when the Tau-Path is reordered to be decreasing in *discordance*.
- Thus at any point *k* along the boundary, 100α percent of the decreasing in concordance paths are expected to exceed the upper boundary by chance; and 100α percent of the decreasing in concordance paths are expected to exceed the lower boundary by chance.

Tau Path: Graph of Boundary



Tau Path: Recalibration of the Boundary

- Pathwise Confidence:
  - Generate large number of independent *n*-dimensional samples from two independent populations
  - For the previously constructed boundary, determine the number of sample Tau-Paths that break the boundary at any point *k*, and calibrate the *pointwise* significance value α to obtain *pathwise* significance α^*.

- For observed data, obtain (positive and negative) Tau-Path, and reject H_0 if it breaks either boundary at any *k*.

Tau Path: Stopping Rule: *"Top K"*

- **Goal:** Determine an 'optimal' choice k^* for the number of observations in the subsample A under which X and Y are associated.
- **Issues:**
  - The tau path itself is not particularly informative
    - First break of the boundary
    - Largest gap from path to boundary (what scale?)
  - It's usually best to overestimate k* because false positives are unavoidable.
- **Idea** (new point of view): Let
  - ω be the ranking of $[X_1,\ldots,X_n]$ after reordering the observations (by π)
  - ν be the ranking of $[Y_1,\ldots,Y_n]$ after reordering the observations (by π)
  - Decompose the ranking ν given ω into independent *stages*, and find the stage at which ω is informative for ν. This gives an estimate k_ω for k^*.
  - Symmetrize by switching the roles of ω and ν; this gives an estimate k_ν.
  - Use $\max(k_\omega, k_\nu)$

Tau Path: Multistage Model for Stopping Rule

Assume ν is random with distribution centered at the true ranking ω, and n observations are ordered in sequence:

- **Stage 1:** Pick i[th] best of all objects (according to ω),
  incurring cost $V_1 = \nu = i - 1$

$$P(V_1 = v) = \left(\frac{1 - r_1}{1 - r_1^{n-1}} \right) r_1^{\nu} I_{\{0,\ldots,n-1\}}(v), \quad r_1 \in (0,1)$$

- **Stage j∈{2,…,n-1} :** Pick *i[th]* best of remaining objects,
  incurring cost: $V_j = \nu = i - 1$

$$P(V_j = v) = \left(\frac{1 - r_j}{1 - r_j^{n-1}} \right) r_j^{\nu} I_{\{0,\ldots,n-j\}}(v), \quad r_j \in (0,1)$$

Assume that $(V_1, \ldots, V_{n-1})$ are independent.

Tau Path: Multistage Model (continued)

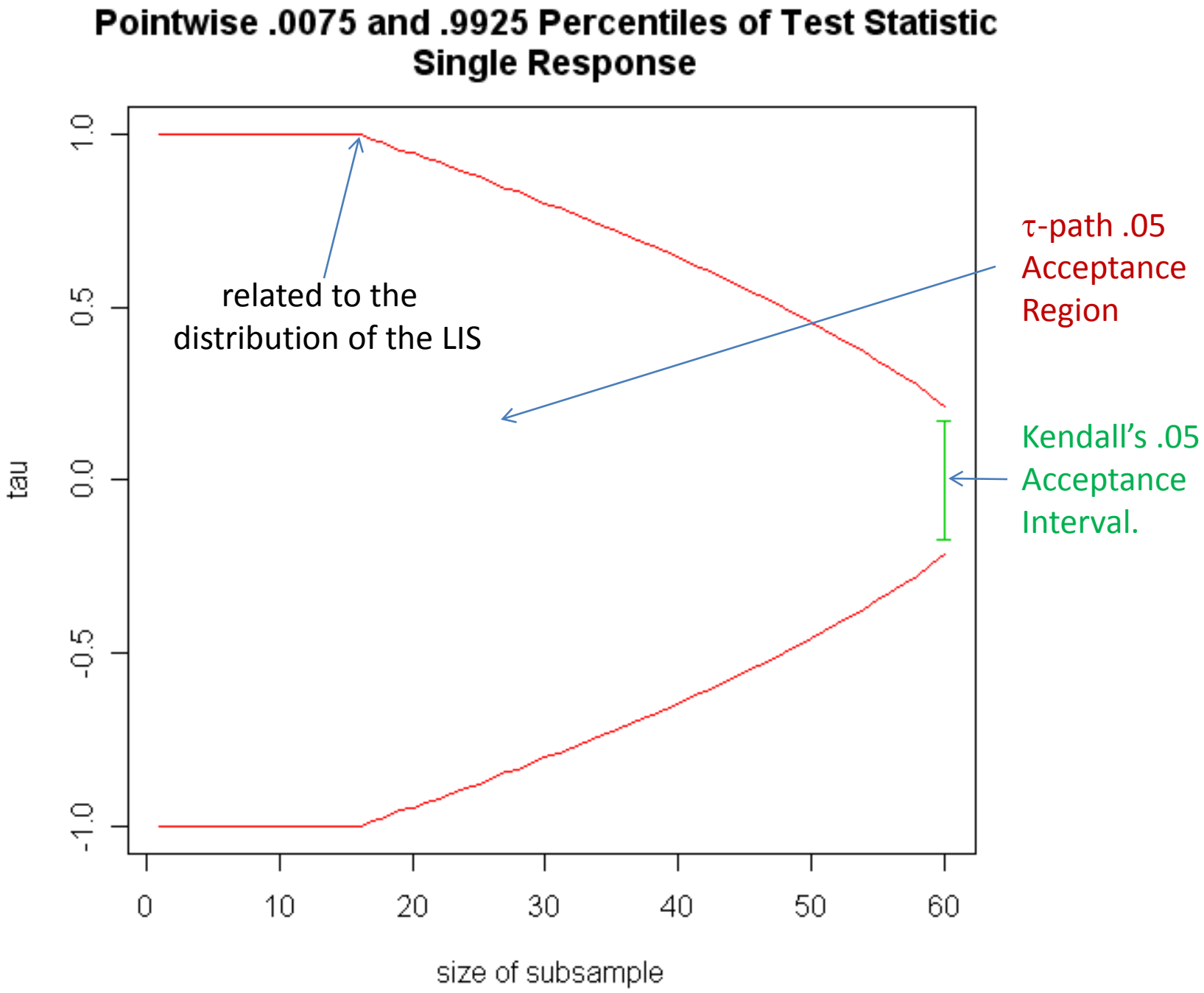
- With a few additional assumptions (notably, assuming a common value for $\theta_j = -\log r_j = \theta$ over a window of length w), we obtain local MLEs $\hat{\theta}_j$
- The limiting distribution for V_j as $\theta \to 0$ is uniform on {*0,…, n-k*}

  $\Rightarrow$ Determine k* so that: $\theta_{k^*} > 0,\ \theta_j = 0 \text{ for all } j > k^*$

Simple Estimator for k_ω:

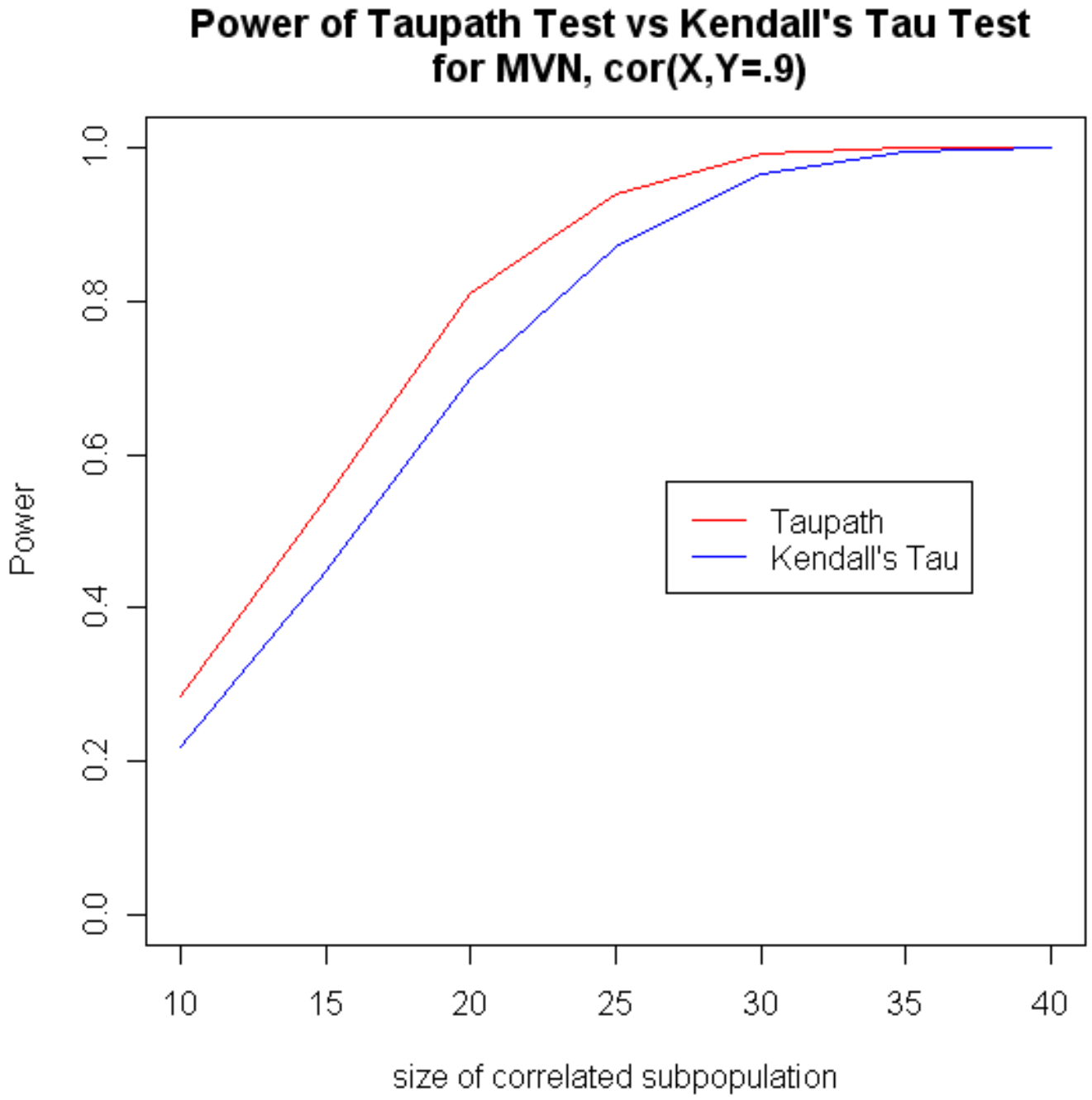
$$\min\left\{k \middle| \hat{\theta}_k > q(k), \theta_k < q(j) \text{ for all but } \alpha(n-k) \text{ of the } j > k\right\}$$

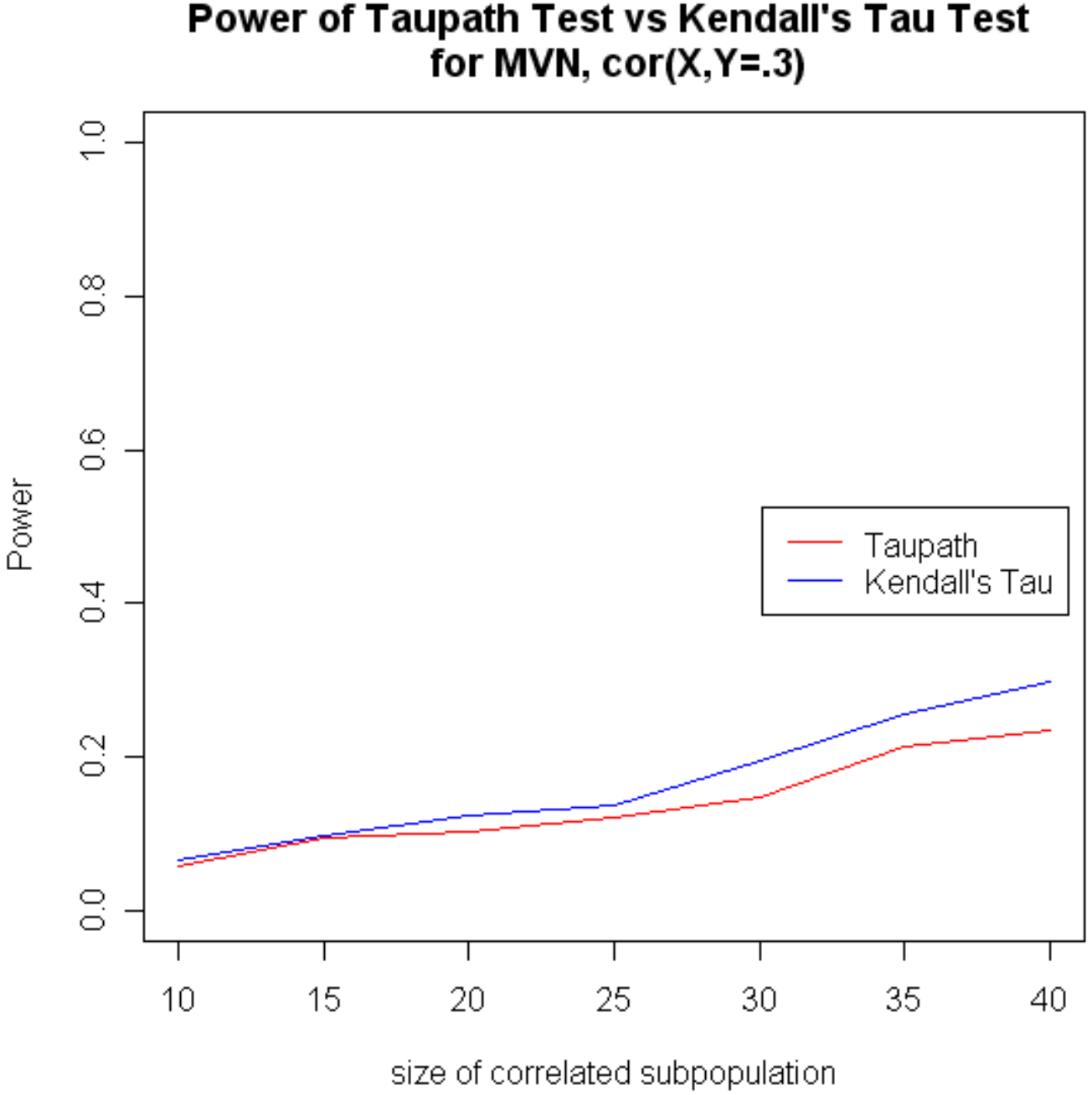
where *q(j)* denotes a α^{th} quantile of $\hat{\theta}_j$, if $\theta_j = \cdots = \theta_{j+w-1} = 0$

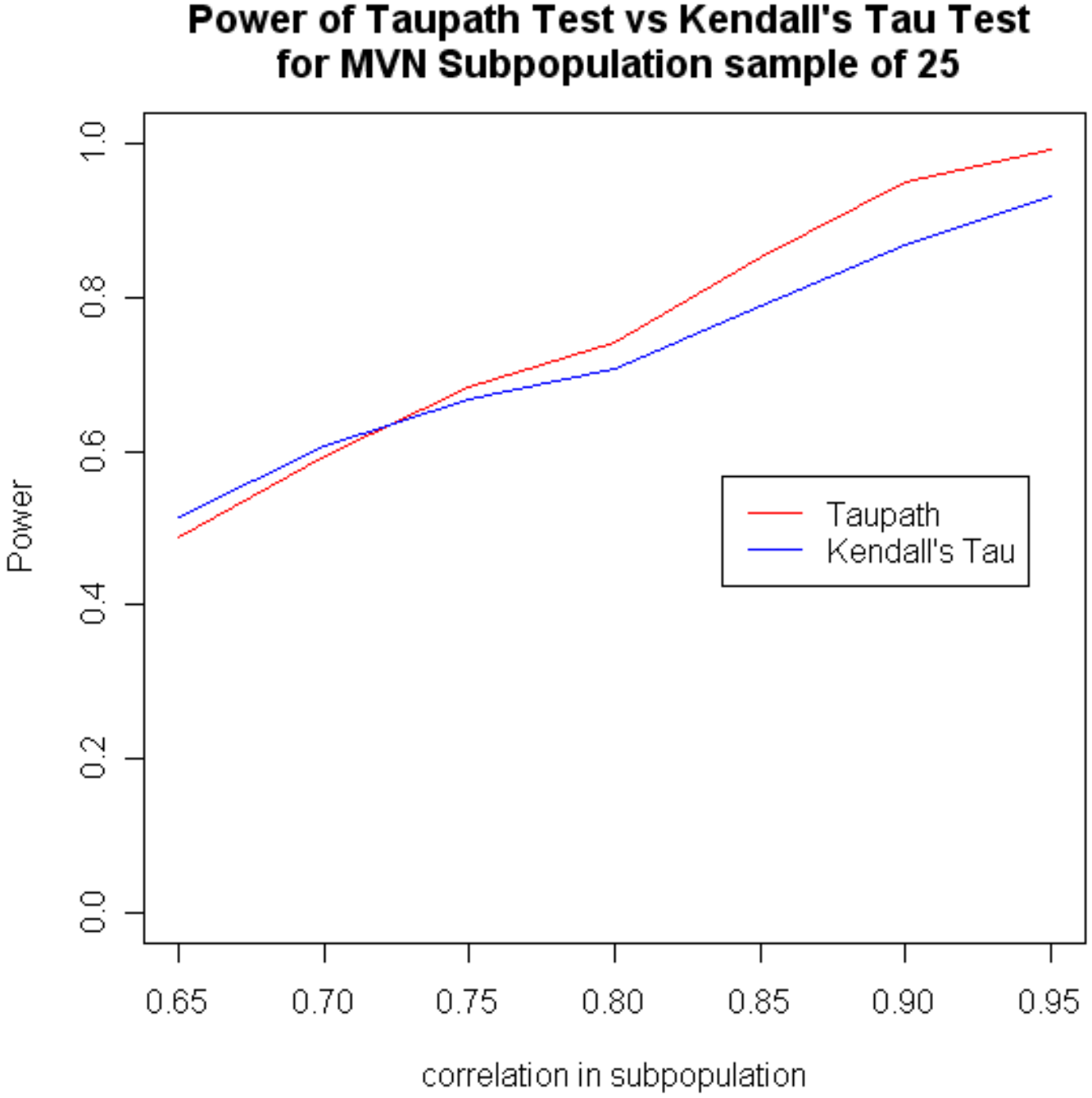

Pointwise .0075 and .9925 Percentiles of Test Statistic
Single Response
tau
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
60
size of subsample
related to the
distribution of the LIS
τ-path .05
Acceptance
Region
Kendall's .05
Acceptance
Interval.

Measuring Performance: Power

- Pick some configuration of interest.
  - *(X,Y)* normally distributed
  - In *k* of the pairs, *X* and *Y* have correlation $\rho = .9$
  - In the remaining *(60 – k)* of the pairs, *X* and *Y* are independent
- Generate 1000 samples from this distribution
  - In what percentage of these sample does the τ-path test reject?
  - In what percentage does the standard Kendall tau test reject?


Power of Taupath Test vs Kendall's Tau Test
for MVN, cor(X,Y=.9)
Power
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
30
35
40
size of correlated subpopulation
Taupath
Kendall's Tau


Power of Taupath Test vs Kendall's Tau Test
for MVN, cor(X,Y=.3)
Power
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
30
35
40
size of correlated subpopulation
Taupath
Kendall's Tau

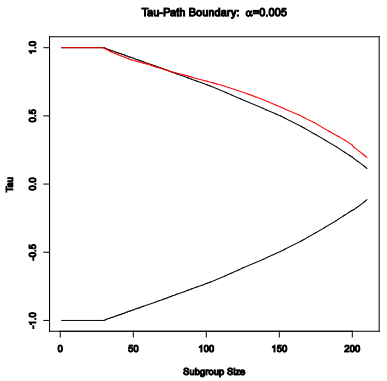

Power of Taupath Test vs Kendall's Tau Test
for MVN Subpopulation sample of 25
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.65
0.70
0.75
0.80
0.85
0.90
0.95
correlation in subpopulation
Taupath
Kendall's Tau

Extension to large samples (Asymptotics)

- Recall, the rejection bounds for the Tau-Path test are obtained by generation of a large number independent *n*-dimensional samples from two independent populations

- This is computationally reasonable for analysis on 210 DMAs, but for agency-level analysis *n* is at least an order of magnitude larger.

- We can avoid this simulation step by obtaining an approximate functional form for the Tau-Path Boundary as a function of only *n* and pointwise (or pathwise) significance α.



Asymptotics: Approximating the Boundary

- The boundary curve is only defined at discrete points k∈{2,…,n}, so let $(B_{2\alpha},\ldots,B_{n\alpha})$ denote the (Upper) Tau-Path Boundary with pointwise significance α.
- On the basis of simulation results, a reasonable approximation for the boundary is given by the following hierarchical relationships:

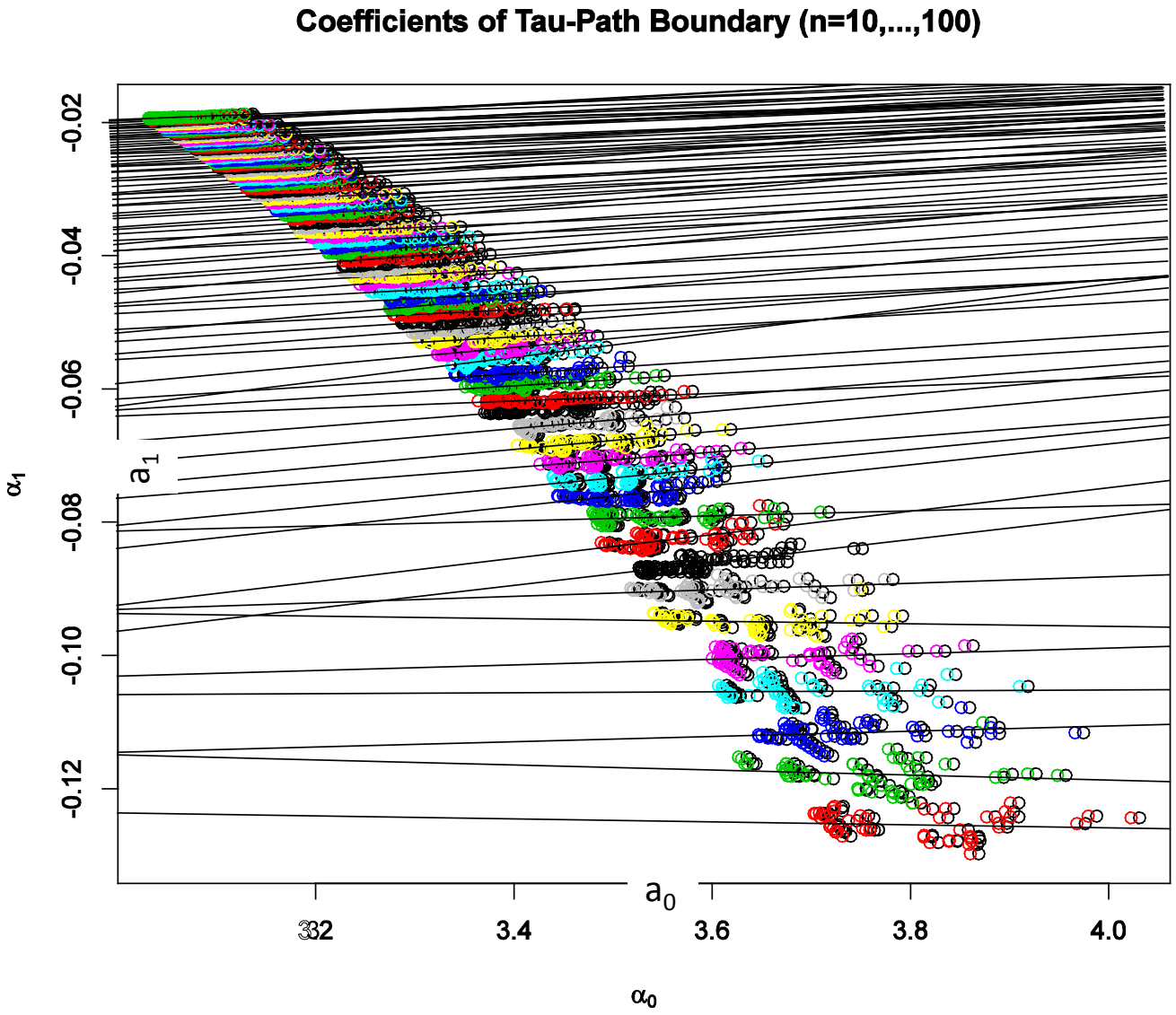
$$B_{k\alpha} \approx \begin{cases} \log(a_0 + a_1 k), & \log(\alpha_0 + \alpha_1 k) < 1 \\ 1, & \log(\alpha_0 + \alpha_1 k) \geq 1 \end{cases}$$

$$a_1 \approx b_0 + b_1 a_0$$

$$b_1 \propto n^{-3/4}, b_0 \propto n^{-1}$$

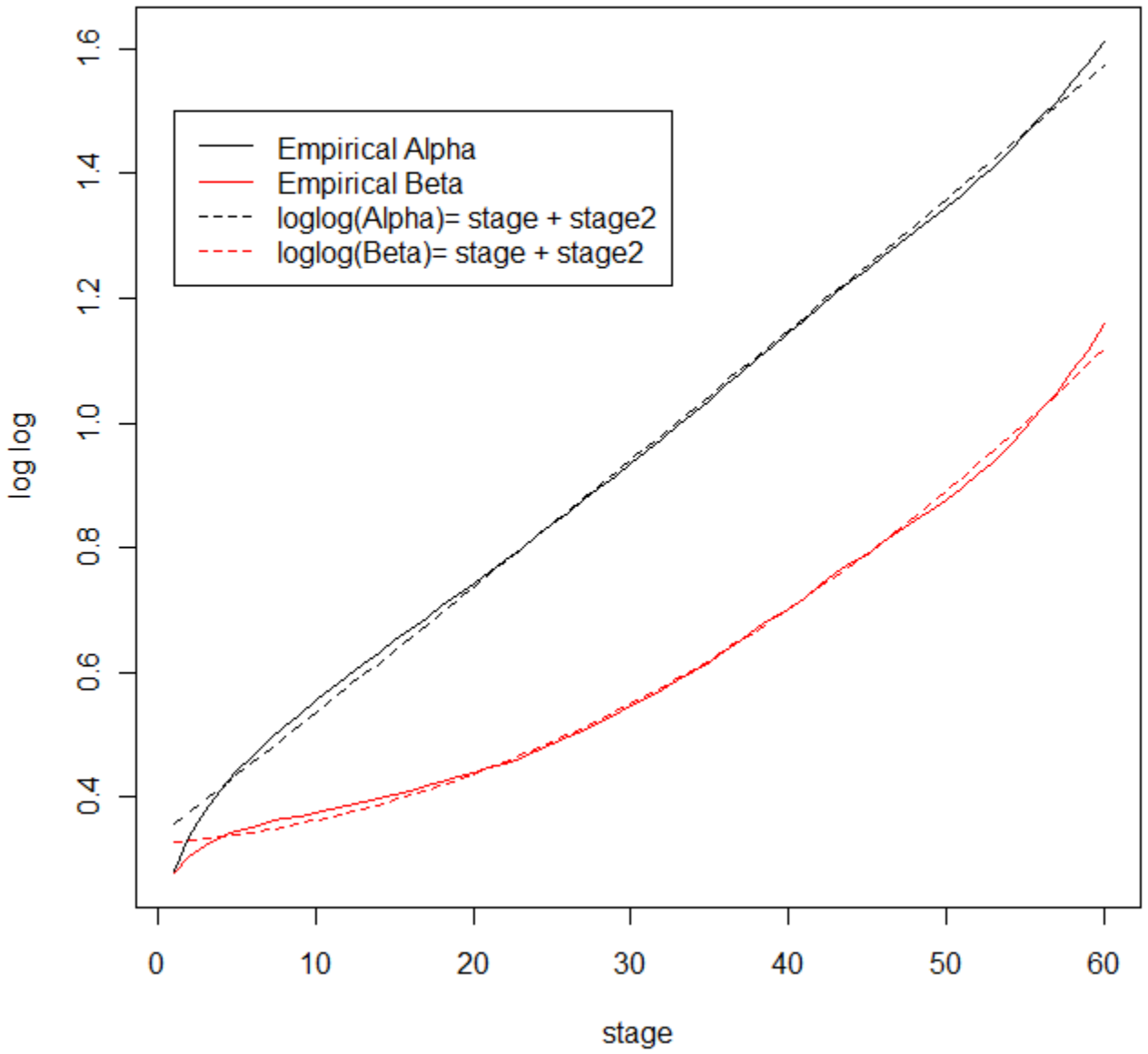
- In this formulation, the rejection bounds depend on α only through a_0
  - Current Work: obtain form for this term that is appropriately sensitive to the significance level $a_0 = f(n,\alpha)$

Asymptotics: Estimating the Boundary



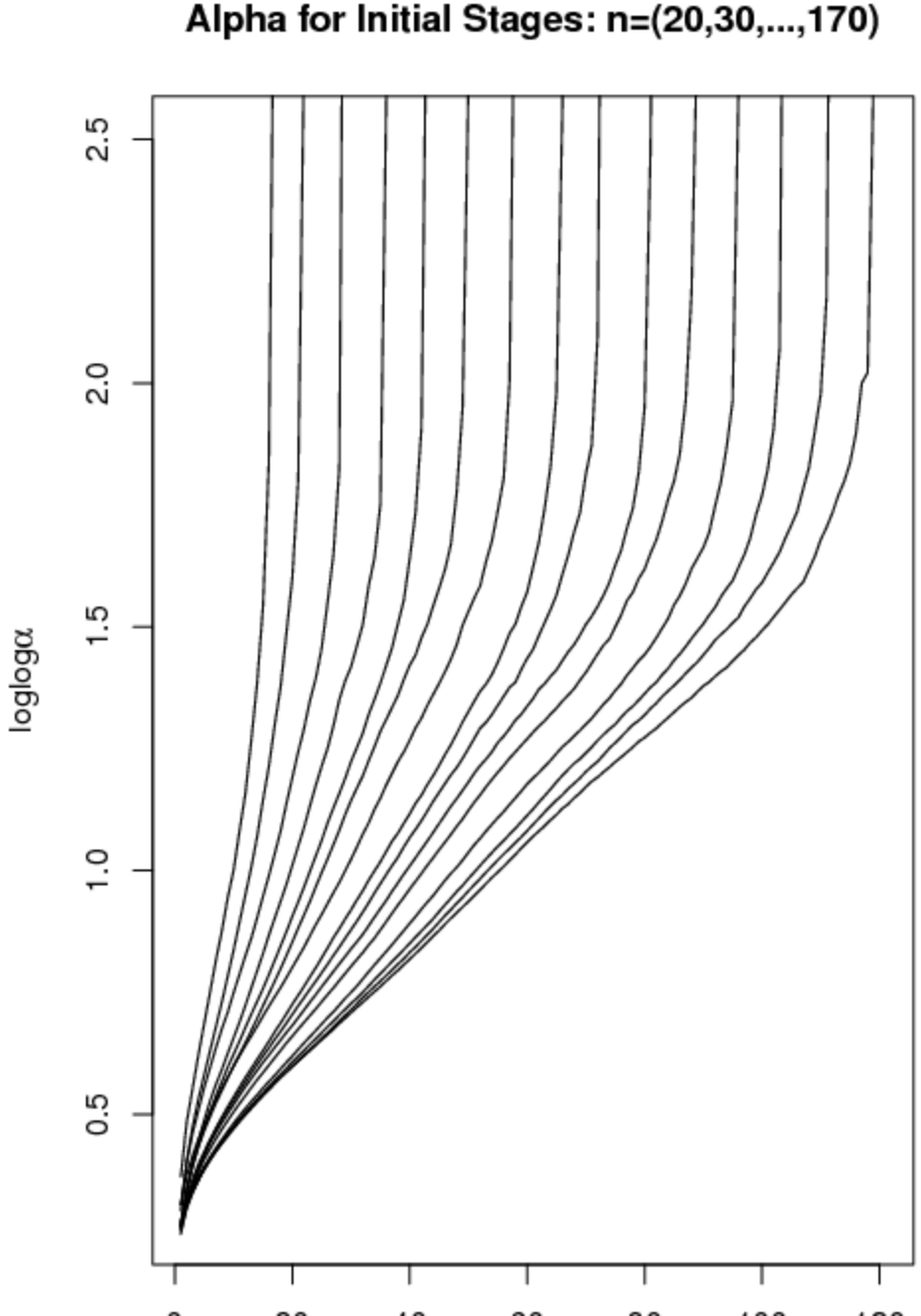
Marginal Distributions of the t_k

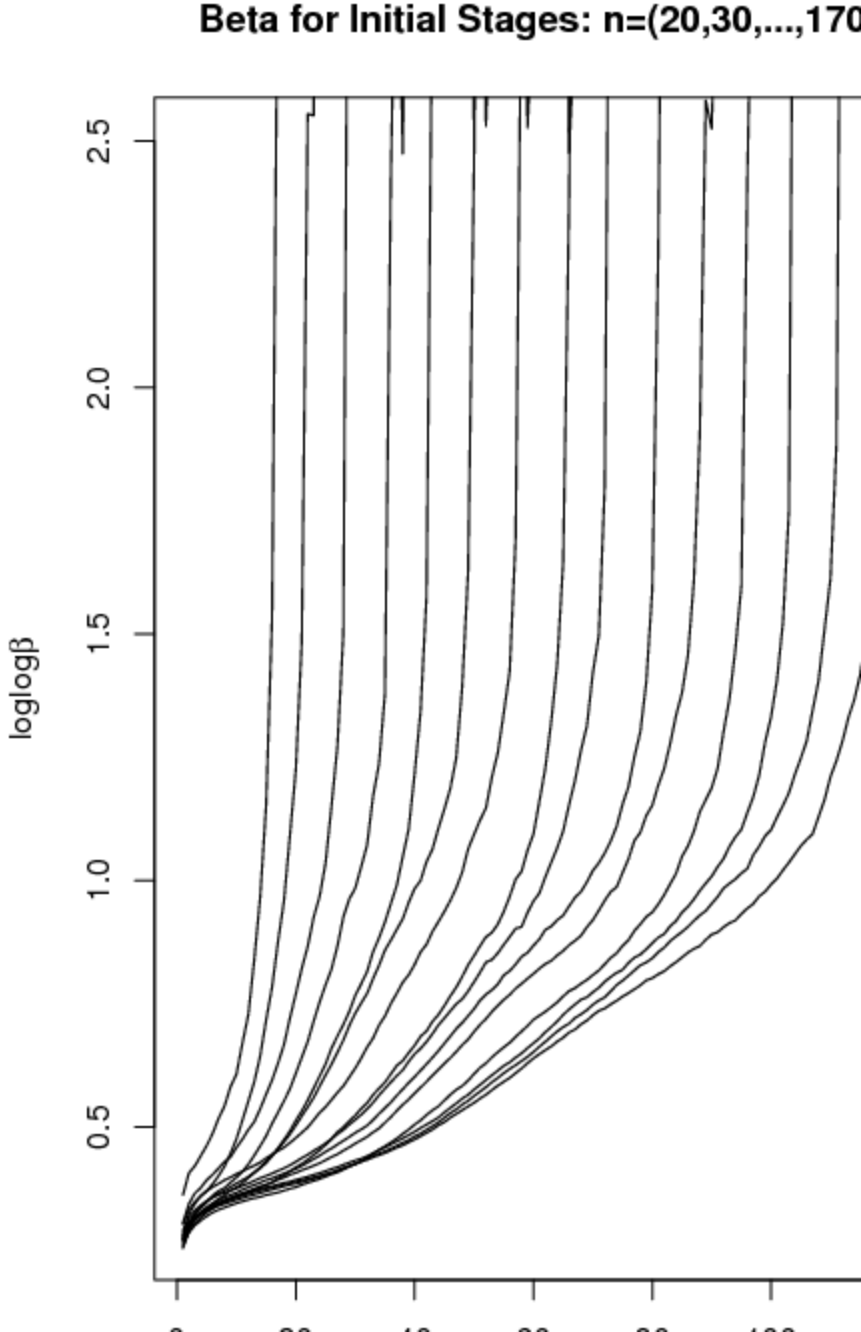
- To understand better the nature of the boundaries, we try to find the marginal distributions of the t_k ($k = 2,\ldots,n$)
- Work on the scale $p_k = (t_k + 1)/2$ with range $(0,1)$
- It is difficult to get a recurrence relationship because of the increasing dependencies among the column sums of the concordance matrices as minimal

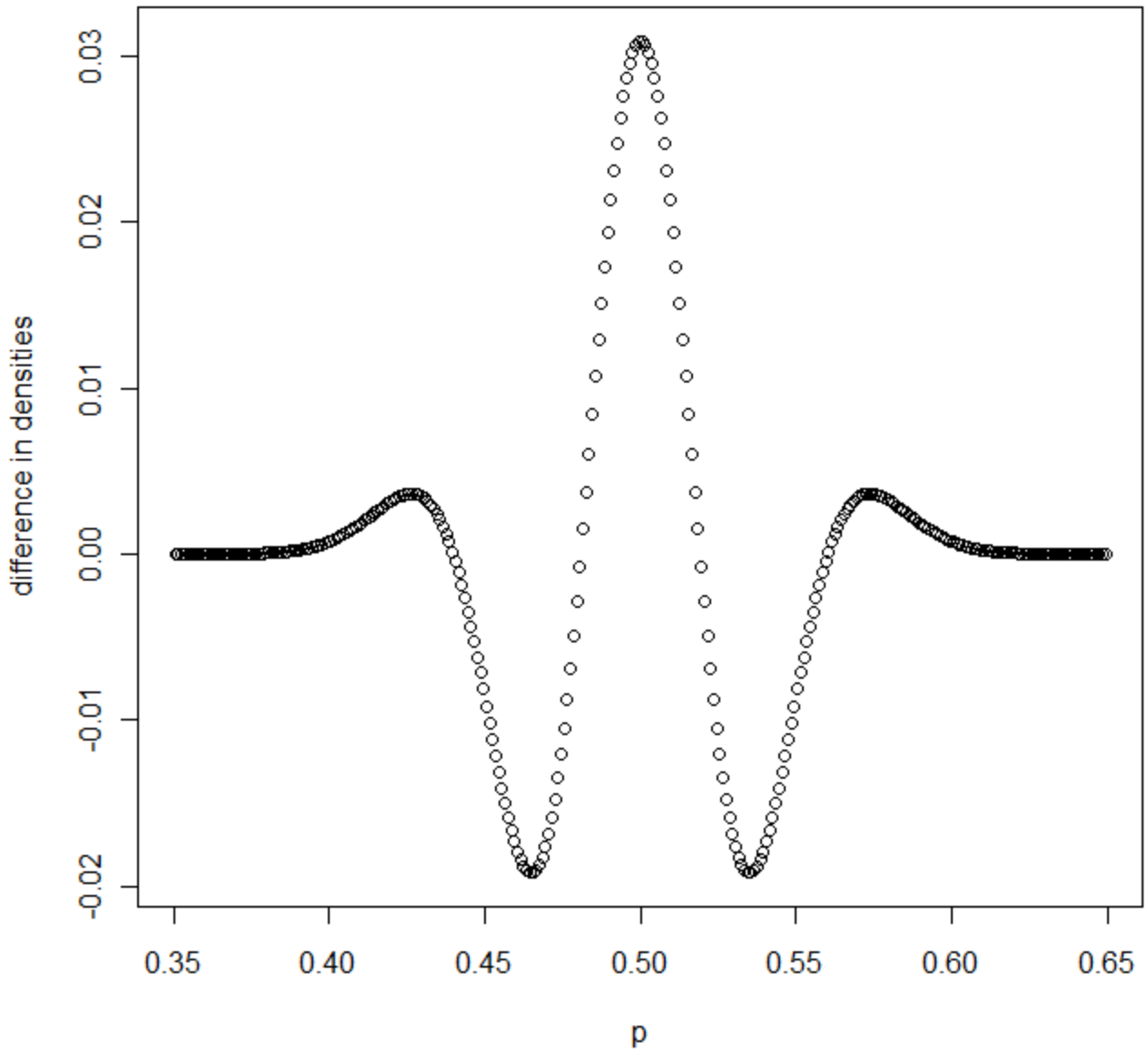

Alpha and Beta at initial stages for n=100
Empirical Alpha
Empirical Beta
loglog(Alpha)= stage + stage2
loglog(Beta)= stage + stage2
log log
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
10
20
30
40
50
60
stage

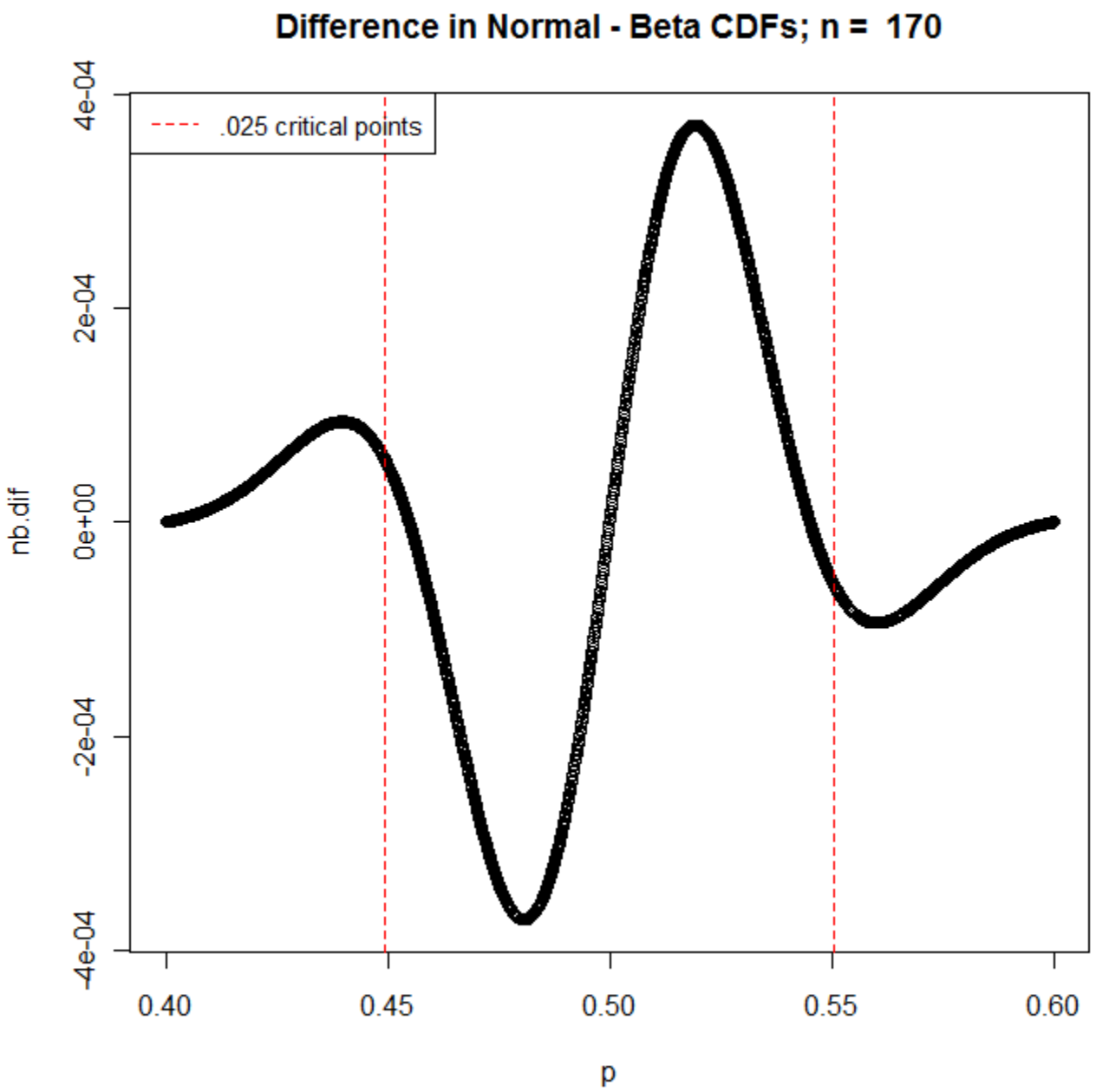
Beta Approximations to the p_k Distributions

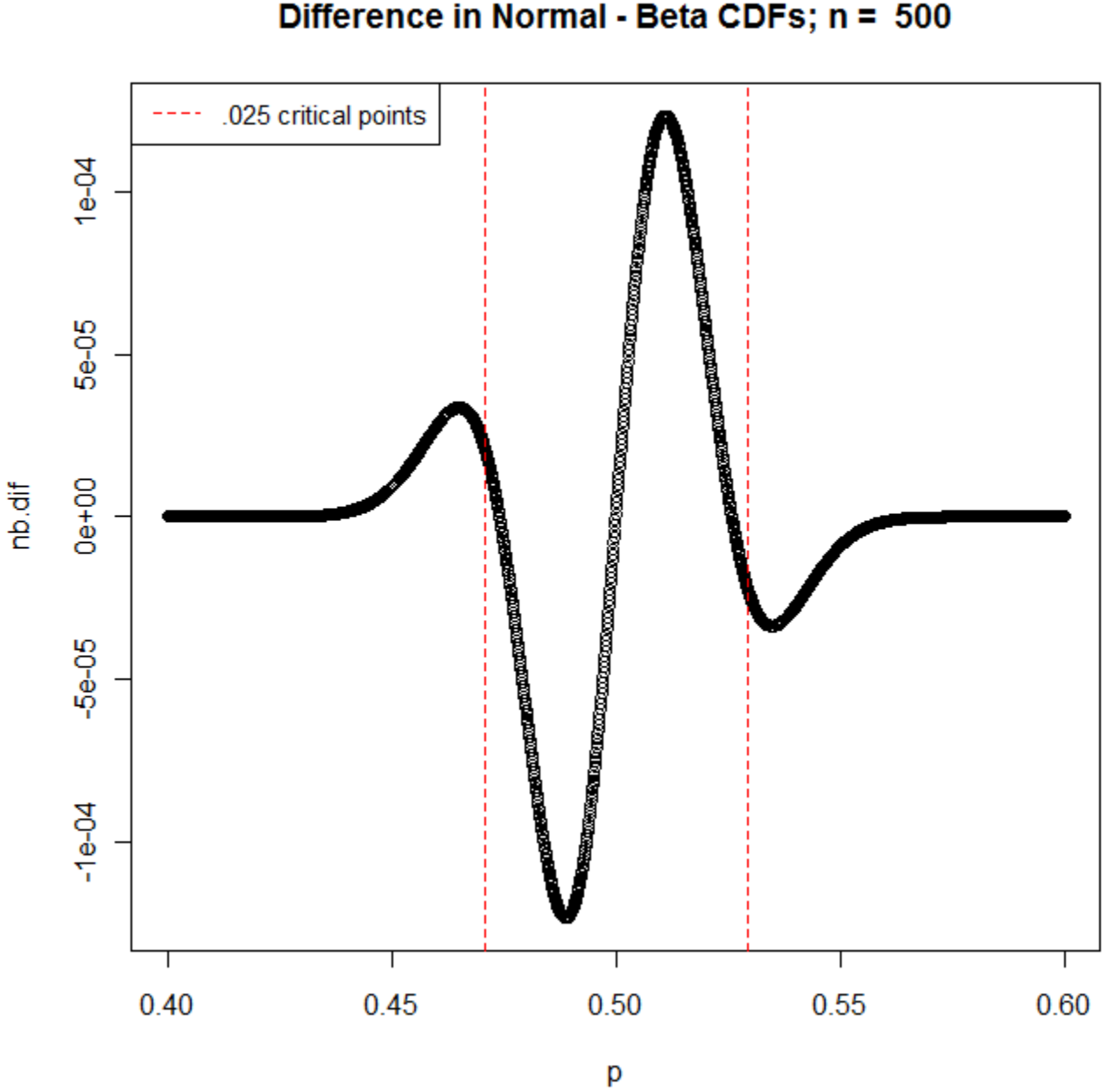
QQ_plot_n170.mp4

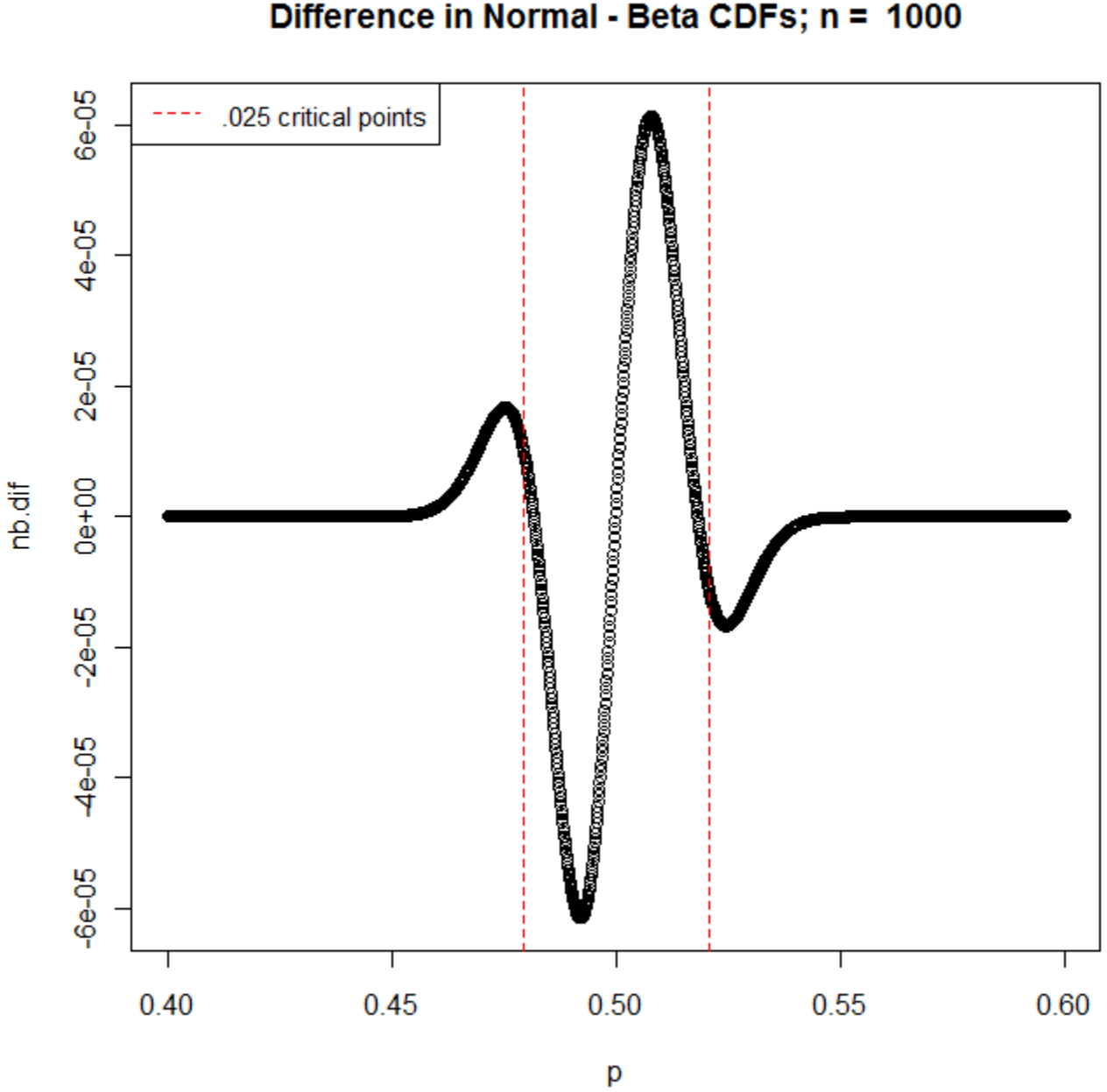

Alpha for Initial Stages: n=(20,30,...,170)
loglogα
2.5
2.0
1.5
1.0
0.5


Beta for Initial Stages: n=(20,30,...,170
loglogβ
2.5
2.0
1.5
1.0
0.5


Normal - Beta densities with null p variance
difference in densities
-0.02
-0.01
0.00
0.01
0.02
0.03
0.35
0.40
0.45
0.50
0.55
0.60
0.65
p


Difference in Normal - Beta CDFs; n = 170
.025 critical points
nb.dif
4e-04
2e-04
0e+00
-2e-04
-4e-04
0.40
0.45
0.50
0.55
0.60
p


Difference in Normal - Beta CDFs; n = 500
.025 critical points
nb.dif
1e-04
5e-05
0e+00
-5e-05
-1e-04
0.40
0.45
0.50
0.55
0.60
p

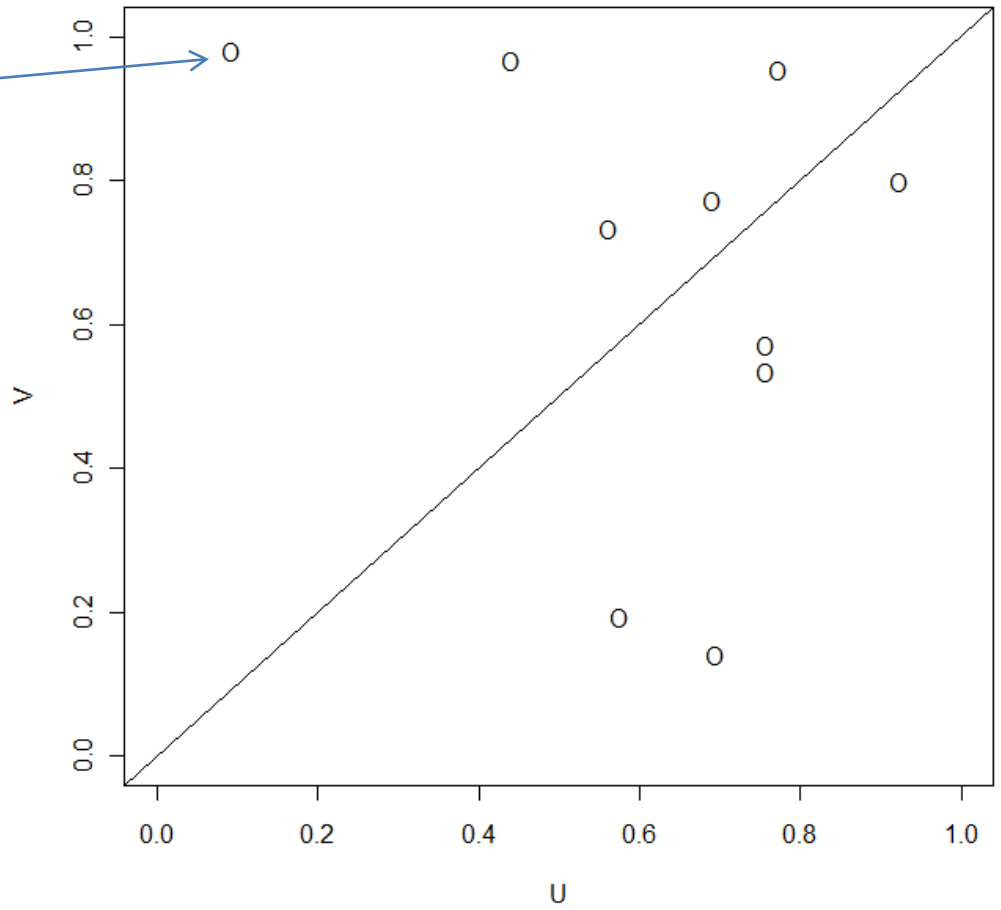

Difference in Normal - Beta CDFs; n = 1000
.025 critical points
nb.dif
6e-05
4e-05
2e-05
0e+00
-2e-05
-4e-05
-6e-05
0.40
0.45
0.50
0.55
0.60
p

A related approach using order statistics

Let $U = F_X(X) \sim \text{Uniform}(0,1)$
and $V = F_Y(Y) \sim \text{Uniform}(0,1)$ be independent

Note:
Using empirical estimates for F_X and F_Y reduces the order statistics to ranks.

Minimal point



Alternative Generation Defining Minimal Points

Generate a sample $\{(U_i, V_i) \mid i = 1,\ldots n\}$ as follows:

1. Generate $S_i \sim 2(1-s) \quad 0 < s < 1$

   This corresponds to $|U_i+V_i-1|$

2. Generate $D_i \sim (1-s_i)\text{Uniform}(0,1)$

   This corresponds to $=|V_i-U_i|$

3. Randomly reflect in each of the lines $V=U$, V=-U.

The j^{th} minimal point is the one with the j^{th} smallest value among

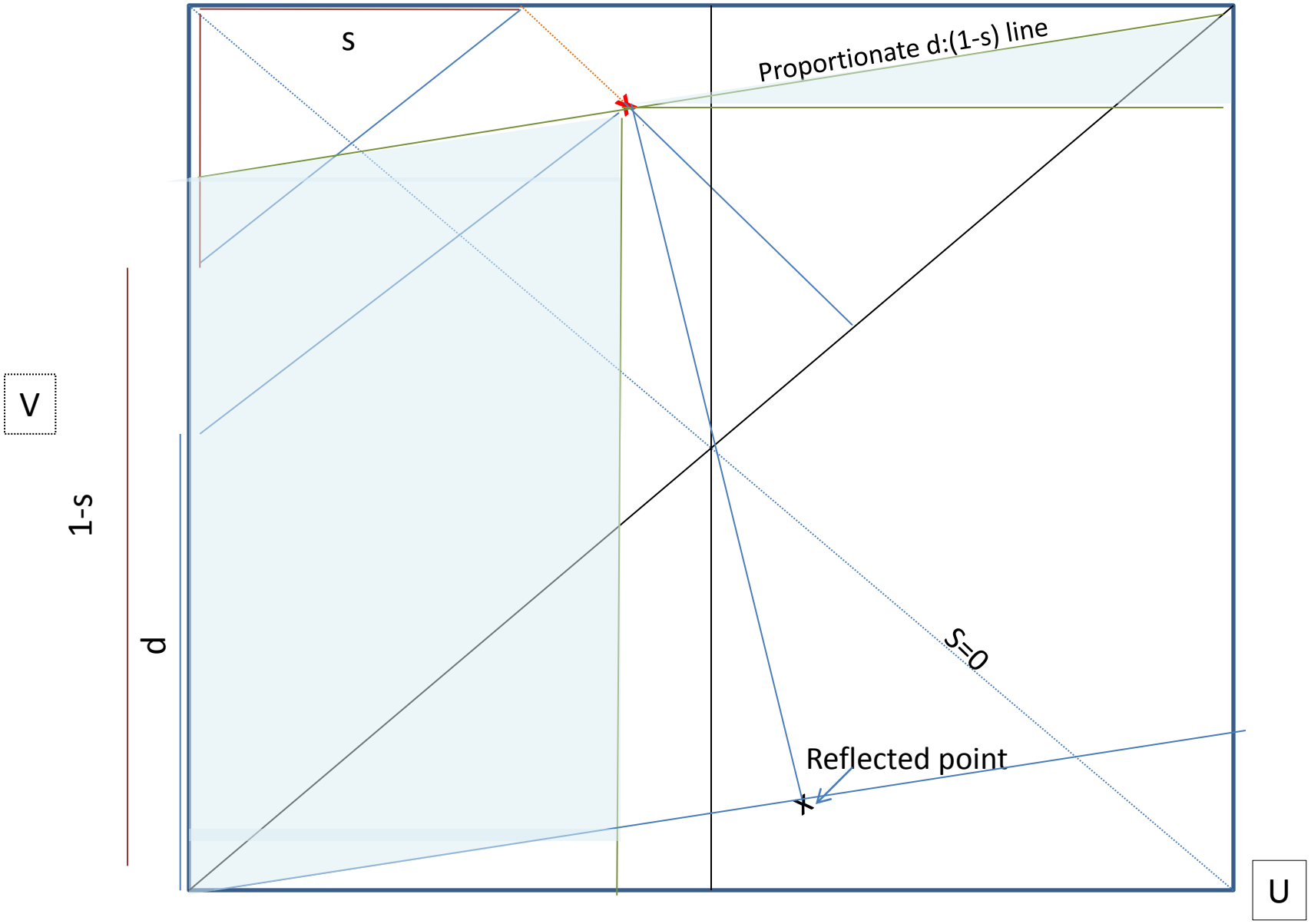
$$\{ d_i / (1-s_i) \mid i = 1,\ldots, n \}$$

To generate a j^{th} minimal point:

$$S \sim 2(1-s)$$

$$D_j \mid s \sim (1-s)\text{Beta}(j, n+1-j)$$

Truncated uniform probability of concordance with *j*-th min point *x* from a sample of n



$U, V \sim \text{Unif}(0,1)$ $S = |U+V-1| \sim 2(1-s)$ $D = |V-U|$ $D_j|s \sim (1-s)\text{Beta}(j,n+1-j)$

The Distribution of P_j for Stage j

$$P_j = \frac{3}{4}(1 - D_j) + \frac{1}{4}S$$

$$(1 - D_j) \sim Beta(j, n + 1 - j)$$

$$S \sim Beta(1,2)$$

$j=1,\ldots n\text{-}2$

D_j and S are independent.