

# **Detecting Change in Multivariate Data Streams Using Minimum Subgraphs**

**Robert Koyak**

**Operations Research Dept.**

**Naval Postgraduate School**

**Collaborative work with Dave Ruth,  
Emily Craparo, and Kevin Wood**

# Basic Setup

Have  $N$  observations assumed to be sampled independently from unknown, multivariate distributions,  $F_j =$  distribution of observation  $j$

**Homogeneity Hypothesis ( $H_0$ ):**

$$F_1 = F_2 = \dots = F_N$$

**Heterogeneity Hypothesis ( $H_1$ ):**

There exists some  $k \in \{2, \dots, N\}$  such that

$$F_1 = F_2 = \dots = F_{k-1}, F_k \neq F_{k-1}, \text{ and}$$

$$\delta(F_1, F_j) - \max_{k \leq r \leq j-1} \delta(F_r, F_j) \text{ is}$$

strictly positive and nondecreasing for

$$j \in \{k + 1, \dots, N\}$$

# Heterogeneity includes:

- A single change in distribution at a known change point (“two-sample problem”)
- A single change in distribution at an unknown change point
- Directional drift (in mean or other features) that begins at an unknown point in the observation sequence

# Distance Matrix

$D = [d_{ij}] = N \times N$  distance matrix (Euclidean, Manhattan, etc.)

$$d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\| = d(\mathbf{y}_i, \mathbf{y}_j)$$

Maa, Pearl, and Bartoszynski (1996):

$Y_1, Y_2, Y_3$  independent,  $\sim F$

$Z_1, Z_2, Z_3$  independent,  $\sim G$

$F = G$  if and only if

$$d(Y_1, Y_2) \stackrel{\mathcal{L}}{=} d(Z_1, Z_2) \stackrel{\mathcal{L}}{=} d(Y_3, Z_3)$$

The distance matrix has the information needed to express departure from the homogeneity hypothesis. For the types of departure we want to detect, this information should be expressed in particular ways.

How can we unlock it?

The strategy we will explore is to fit a minimum subgraph (of some type) to the data treated as vertices in a complete, undirected graph. From the subgraph a statistic is derived that is sensitive to the departures from homogeneity that we wish to detect.

# A Graph-Theoretic Approach

Complete undirected graph  $G_N = (V, E)$ ,  $V = \square_N$ ,

$$|E| = N(N-1)/2$$

Subgraph family  $\mathcal{G}_N$  (e.g. spanning trees,  $k$ -factors, Hamiltonian paths or circuits)

Minimum subgraph  $\hat{G} = (\hat{V}, \hat{E}) \in \mathcal{G}_N$  is defined by

$$\hat{G} = \operatorname{argmin}_{G=(V,E) \in \mathcal{G}_N} \sum_{(i,j) \in E} d_{ij}$$

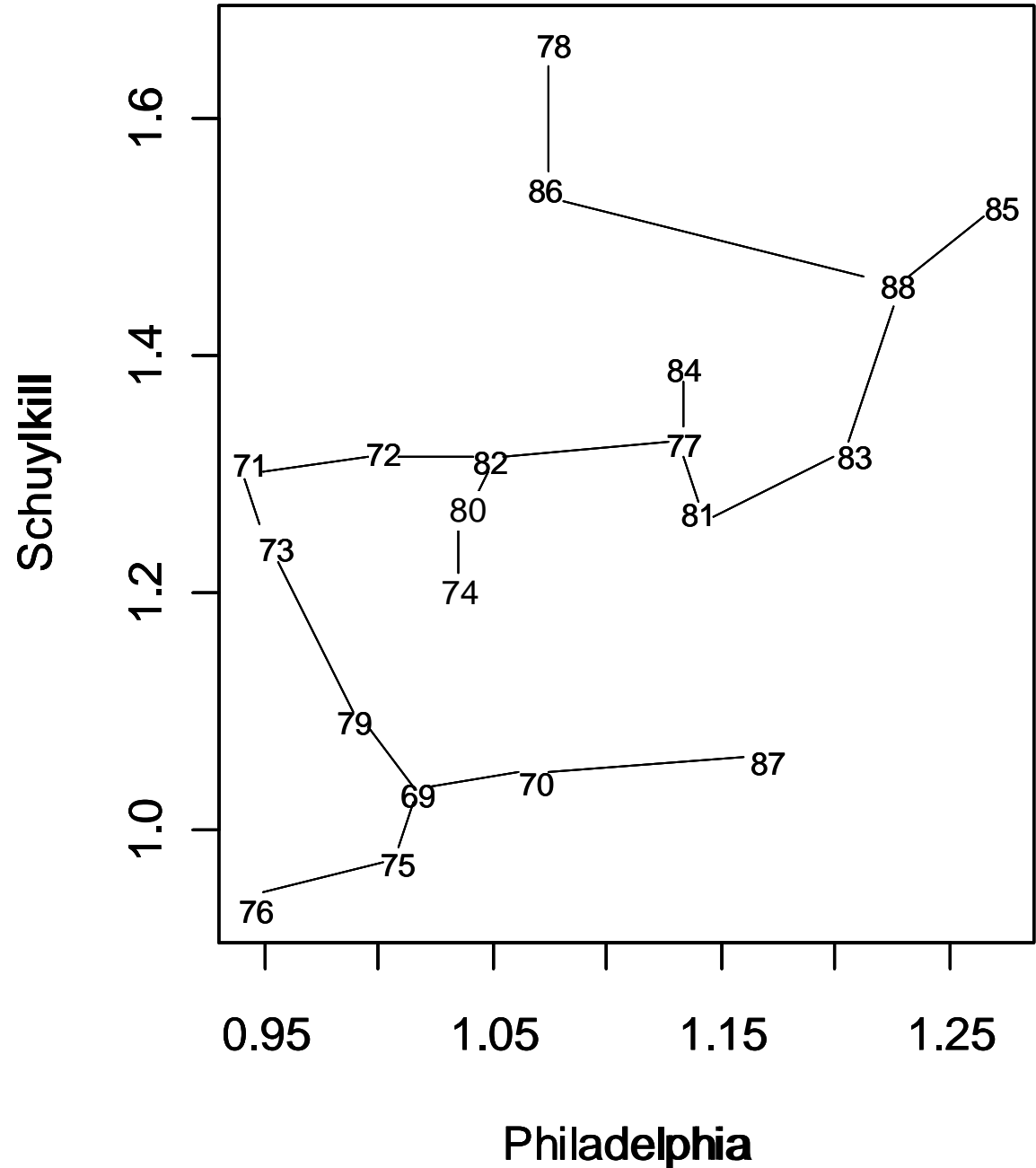
The test statistic is  $\hat{\psi} = \psi(\hat{G})$

# Minimum Spanning Trees (MSTs)

- Friedman and Rafsky (1979) used MSTs to define a multivariate extension of the runs test in the context of the two-sample problem
- The test statistic is the number of edges in the MST that join vertices belonging to different samples
- Small values of the statistic are evidence against homogeneity

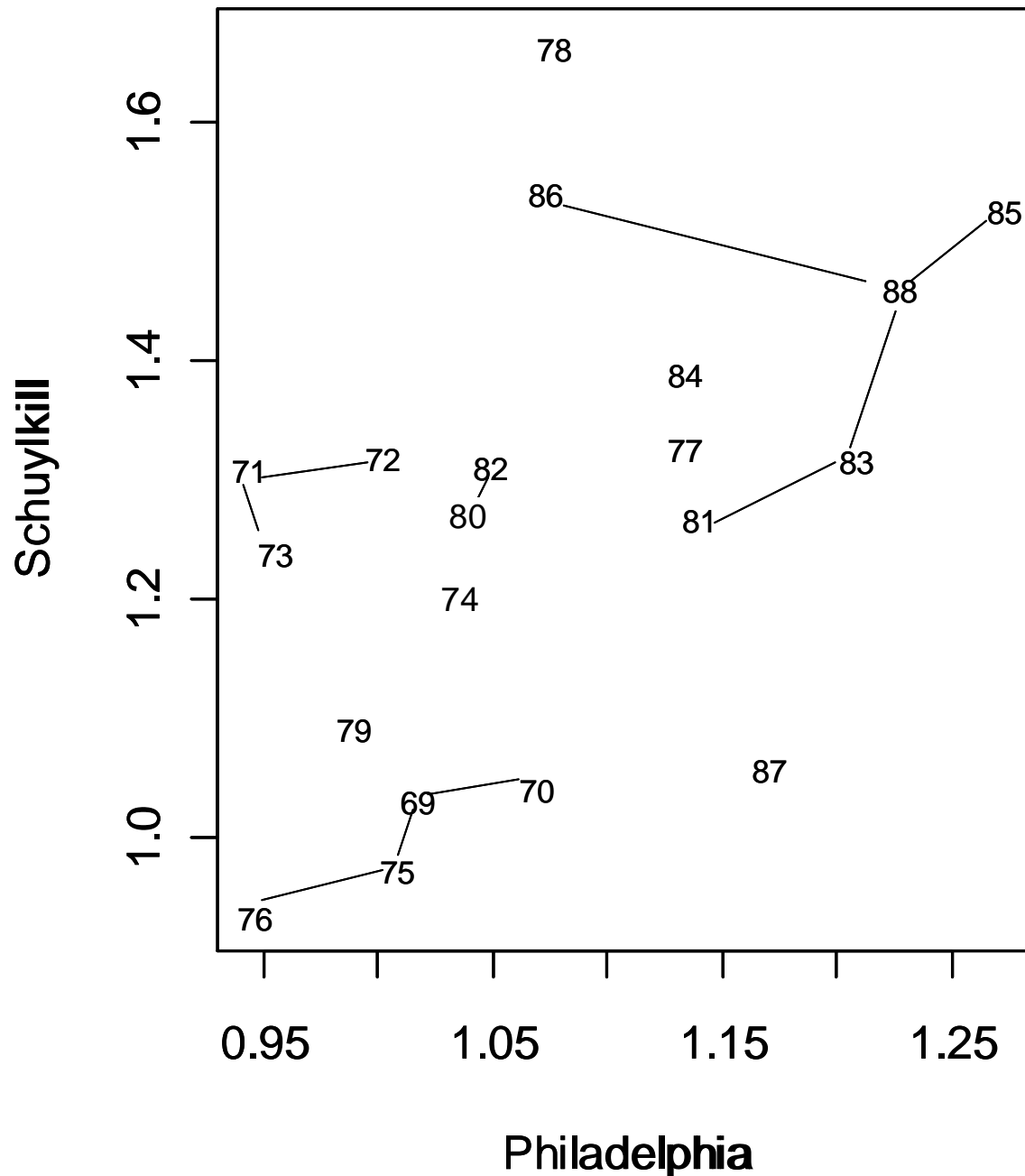
MST for breast cancer mortality rates, 1969 to 1988 ( $N = 20$ ), relative to 1968 base.

Next, treat Sample 1 as the years 1969-1978 and Sample 2 as the years 1979-1988



There are  $\hat{\psi}_{\text{MST}} = 11$  edges that join vertices in different samples.

The p-value, obtained by a permutation test, is about 0.41



# Is anything really happening?

Spearman rank correlations vs. time,  
p-values:

Philadelphia  $\approx$  .0004

Schuylkill  $\approx$  .01

# Minimum Non-bipartite Matching (MNBM)

- Also known as unipartite matching, 1-factor
- Rosenbaum (2005) defined a “cross-match” test using MNBM analogous to that of Friedman and Rafsky
- The test statistic is the number of edges in the MNBM that join vertices belonging to different samples
- Small values of the statistic are evidence against homogeneity

# Cross-match test (Rosenbaum)

$n = \lfloor N / 2 \rfloor$  (number of matching edges)

Group 1 has  $k$  observations

Group 2 has  $N - k$  observations

$M^C$  = number of cross - matches

$M$  = number of matches within Group 1

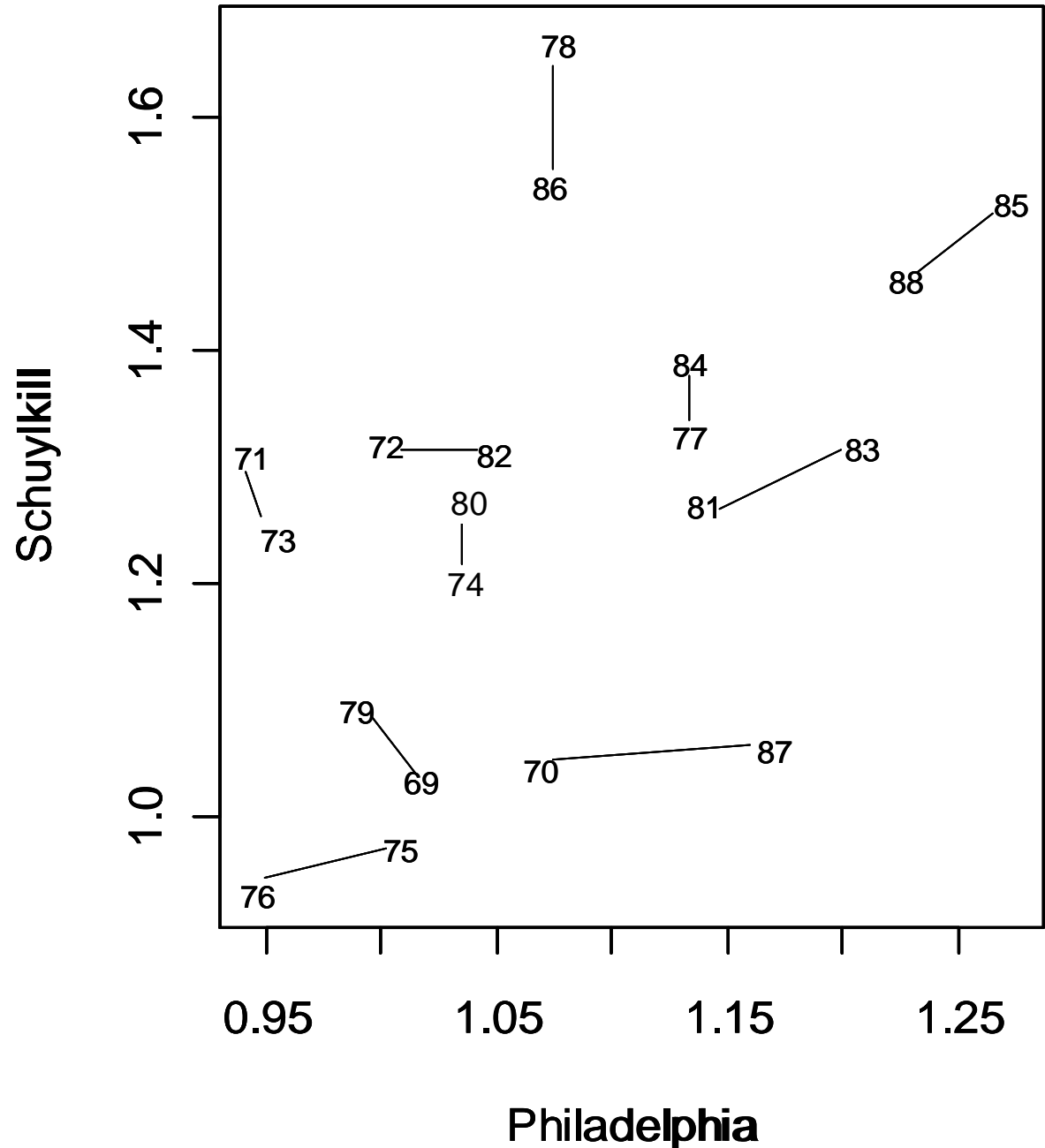
$$2M + M^C = k$$

$$P(M = r) = 2^{k-2r} \binom{n}{k-r} \binom{k-r}{r} \binom{N}{k}^{-1},$$

$$r = 0 \vee (k - n), \dots, \lfloor k / 2 \rfloor$$

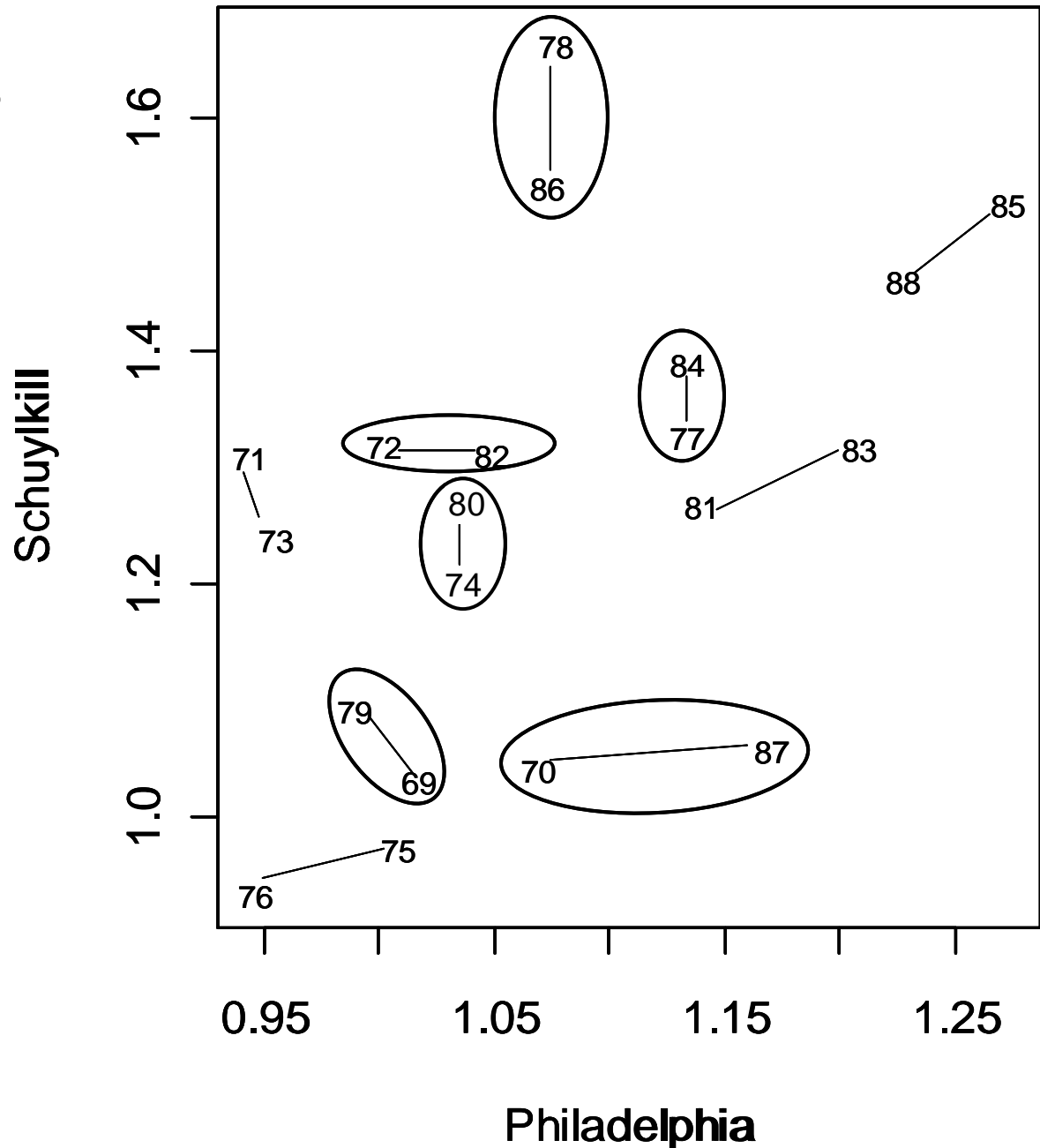
MNBM fit to the breast cancer mortality data.

Count the number of edges that join vertices in different groups



There are  $\hat{\psi}_{CM} = 6$  edges that join vertices in different samples.

The p-value, obtained from the exact null distribution, is about 0.87



# Extensions of the Cross-Match Test

Ruth (2009) and Ruth & Koyak (2011) introduce two extensions of the cross-match test to detect departures from homogeneity in the direction of  $H_1$  :

- (1) An exact, simultaneous cross-match test for an unspecified change-point

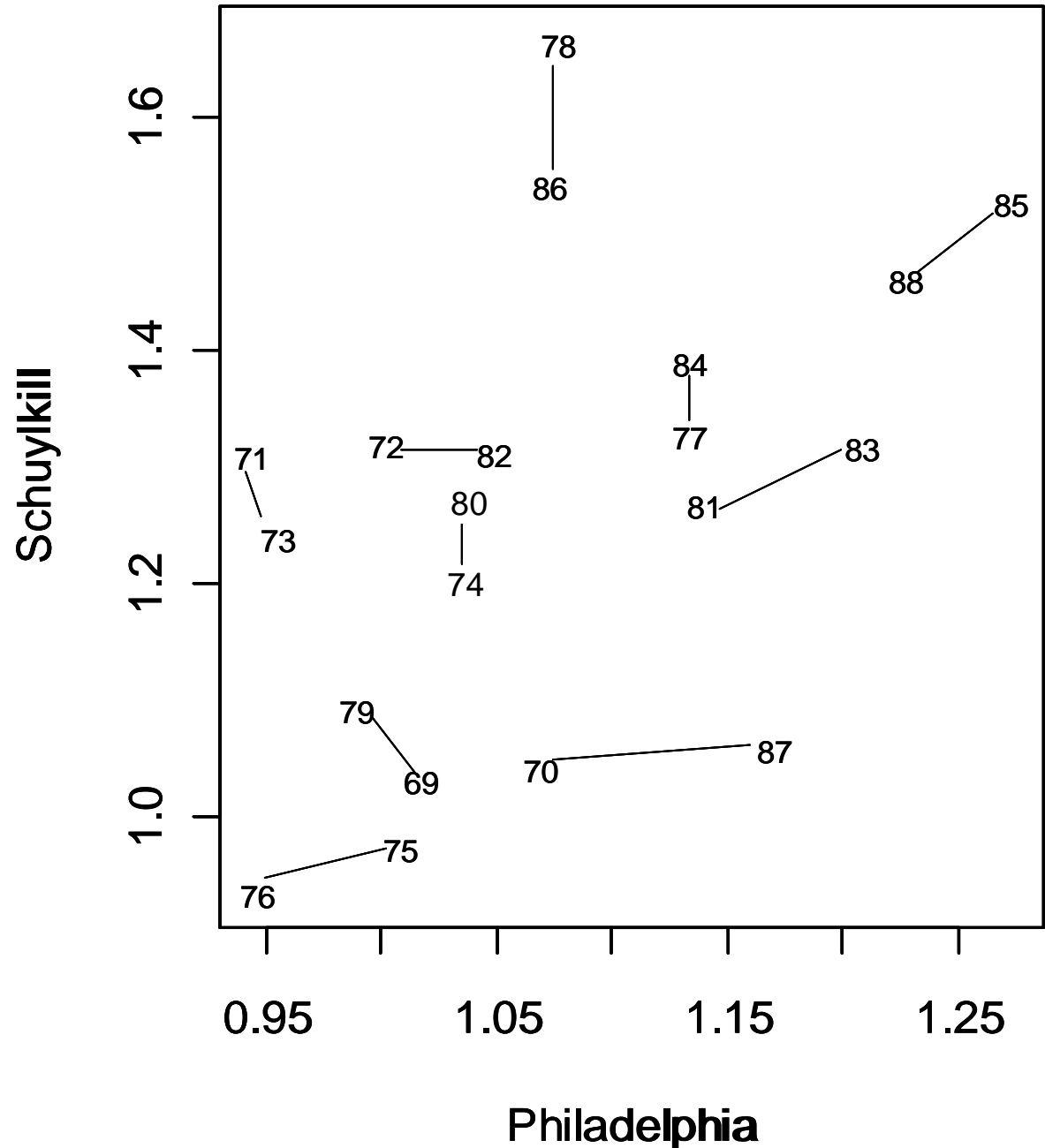
$$\hat{\psi}_{\text{SCM}}(\alpha) = \min_{k_0 \leq k \leq k_1} \left\{ \Psi_{\text{CM}}(k) - q_k(\alpha, k_0, k_1) \right\}$$

- (2) A sum of (vertex) pair maxima test

$$\begin{aligned} \hat{\psi}_{\text{SPM}} &= \sum_{(i,j) \in \hat{E}} i \vee j \\ &= \frac{1}{2} \sum_{(i,j) \in \hat{E}} |i - j| + \frac{1}{4} N(N + 1) \end{aligned}$$

SCM test has exact p-value of 0.59 for testing against an unspecified change-point

SPM test has approximate p-value of 0.41



# Some Theory

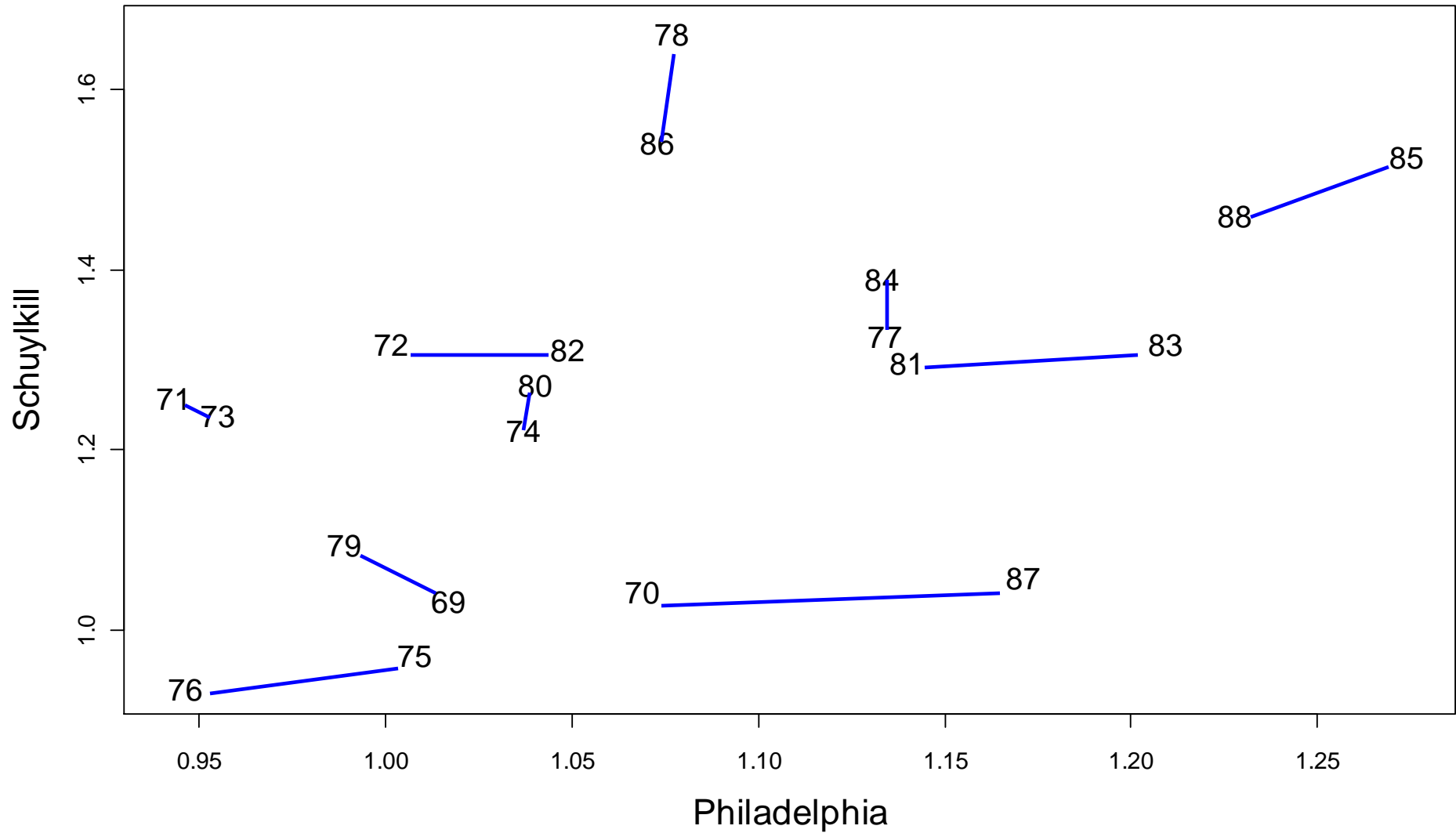
- Friedman & Rafsky's  $\hat{\psi}_{\text{MST}}$ 
  - Asymptotic normality under  $H_0$
  - Universal consistency under  $H_1$  for the two-sample problem (Henze & Penrose, 1999)
- Rosenbaum's  $\hat{\psi}_{\text{CM}}$ 
  - Asymptotic normality under  $H_0$
  - Consistency under restrictive assumptions
- Ruth's SPM test  $\hat{\psi}_{\text{SPM}}$ 
  - Asymptotic normality under  $H_0$
  - Consistency remains to be proven

# Ensemble Tests

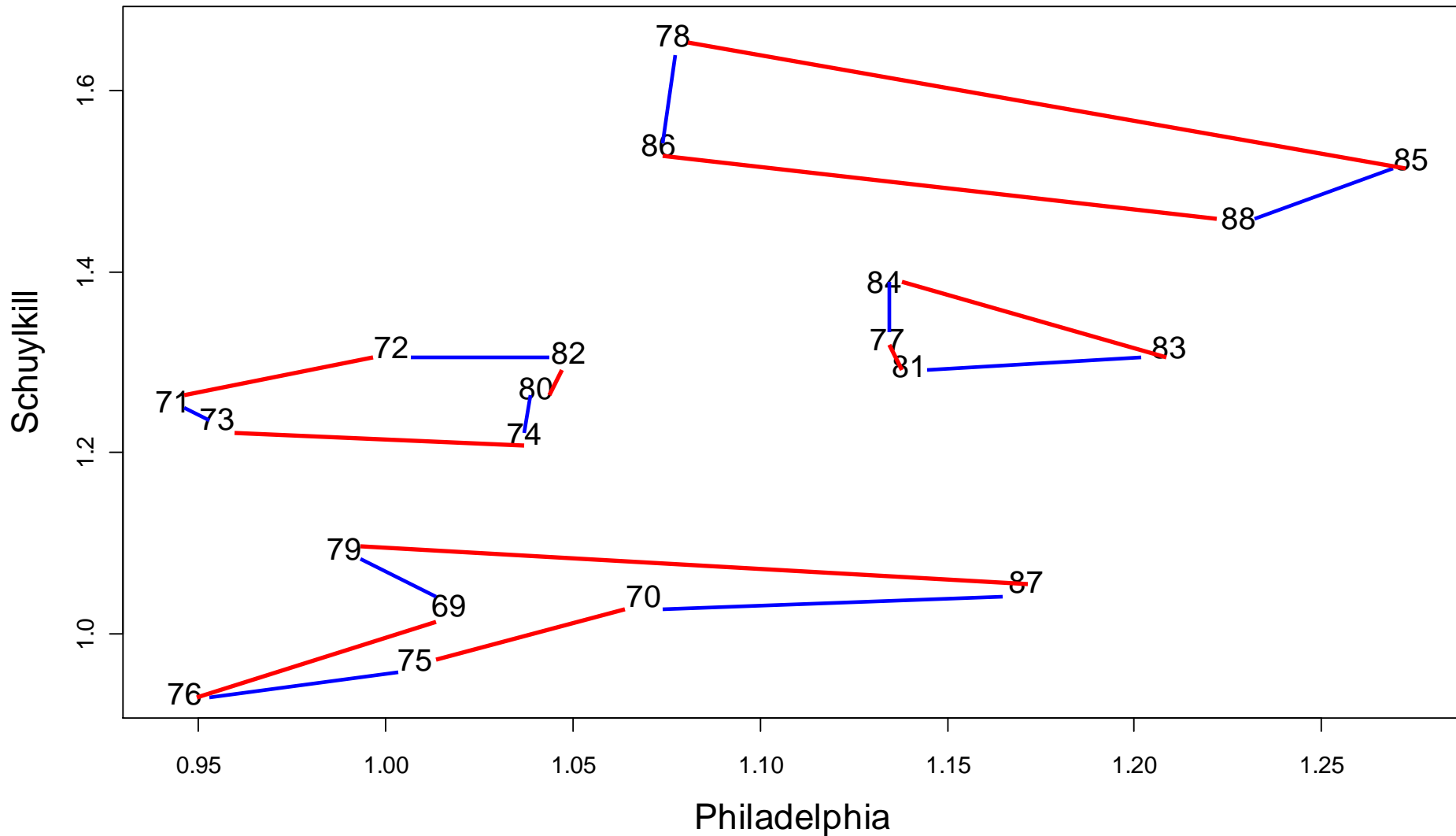
Problem with graph-theoretic tests: a single minimum subgraph contains very limited information about  $D$  and as such these tests are not very powerful

- Tukey suggested fitting multiple "orthogonal" MSTs in Friedman & Rafsky's test and combining them (in a manner that was not specified)
- Two subgraphs are orthogonal if they share no common edges
- For MSTs this is problematic: existence of a fixed number of orthogonal MSTs (even two) is not assured!
- For MNBMs we are assured at least  $\lfloor N/2 \rfloor$  orthogonal subgraphs (Anderson, 1971) constructed sequentially

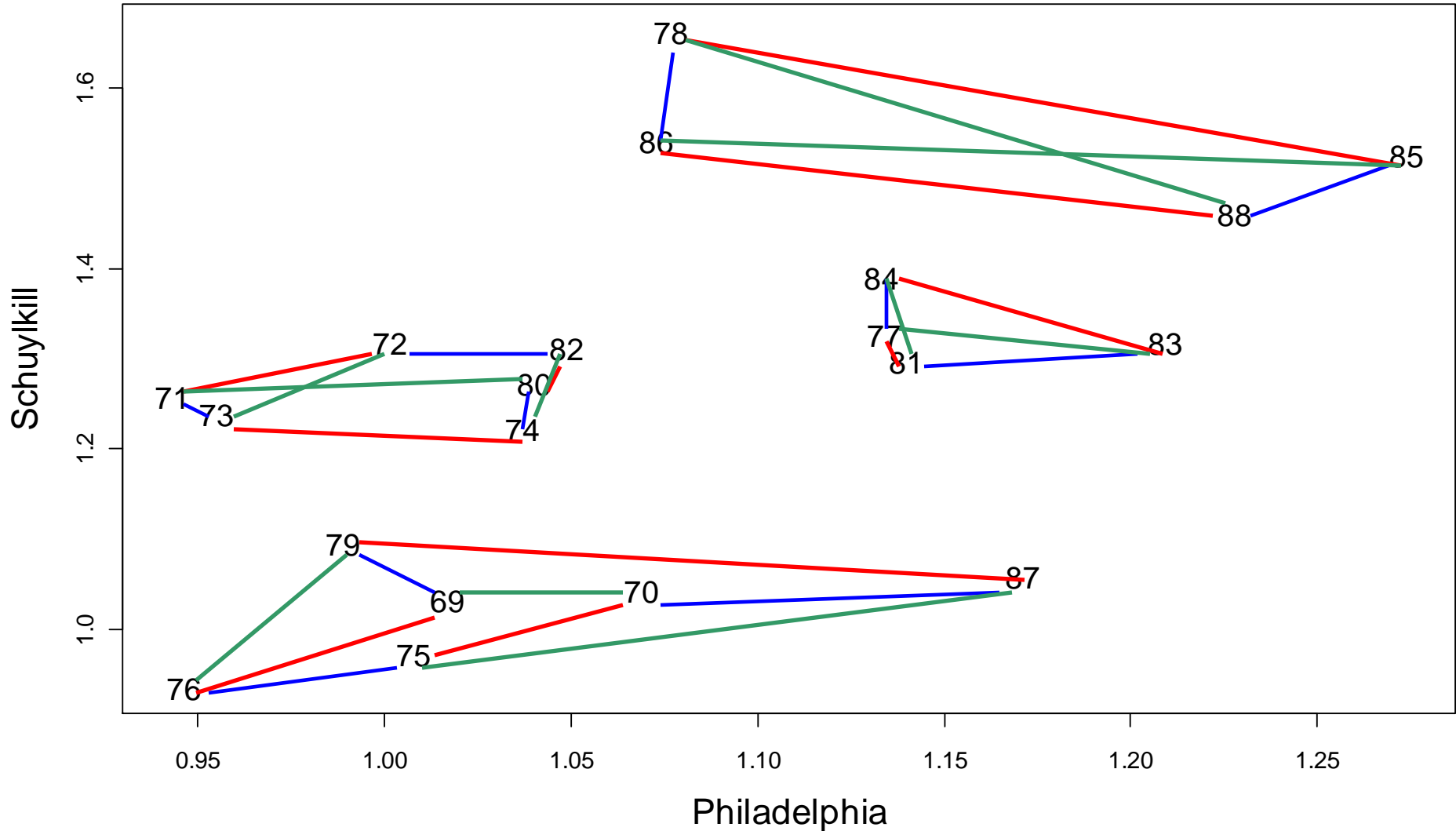
# First MNBM Fit to the Breast Cancer Mortality Data



# First Two MNBMs Fit to the Breast Cancer Mortality Data



# First Three NMBMs Fit to the Breast Cancer Mortality Data



# Structure of Ensembles

- Ensemble pairs decompose into Hamiltonian cycles each having an even number of vertices
  - Under  $H_0$  all 1-factors are equally likely but it is not true that all ensemble 2-factors are equally likely!
  - However, conditional on the cyclic structure uniformity is true
  - Second-order properties do not depend on the cyclic structure
- Ensemble 3-factors have more complex cyclic behavior and also exhibit triangles
  - Prevalence of triangles depends on the dimensionality of the data:  
lower dimension = more triangles

# Ensemble Tests

Ruth (2009) proposed an Ensemble Sum of Pair Maxima (ESPM) test based on fitting a sequence of  $n = \lfloor N / 2 \rfloor$  orthogonal MNBMs and taking the cumulative sums of the SPM statistics. The test takes the following form:

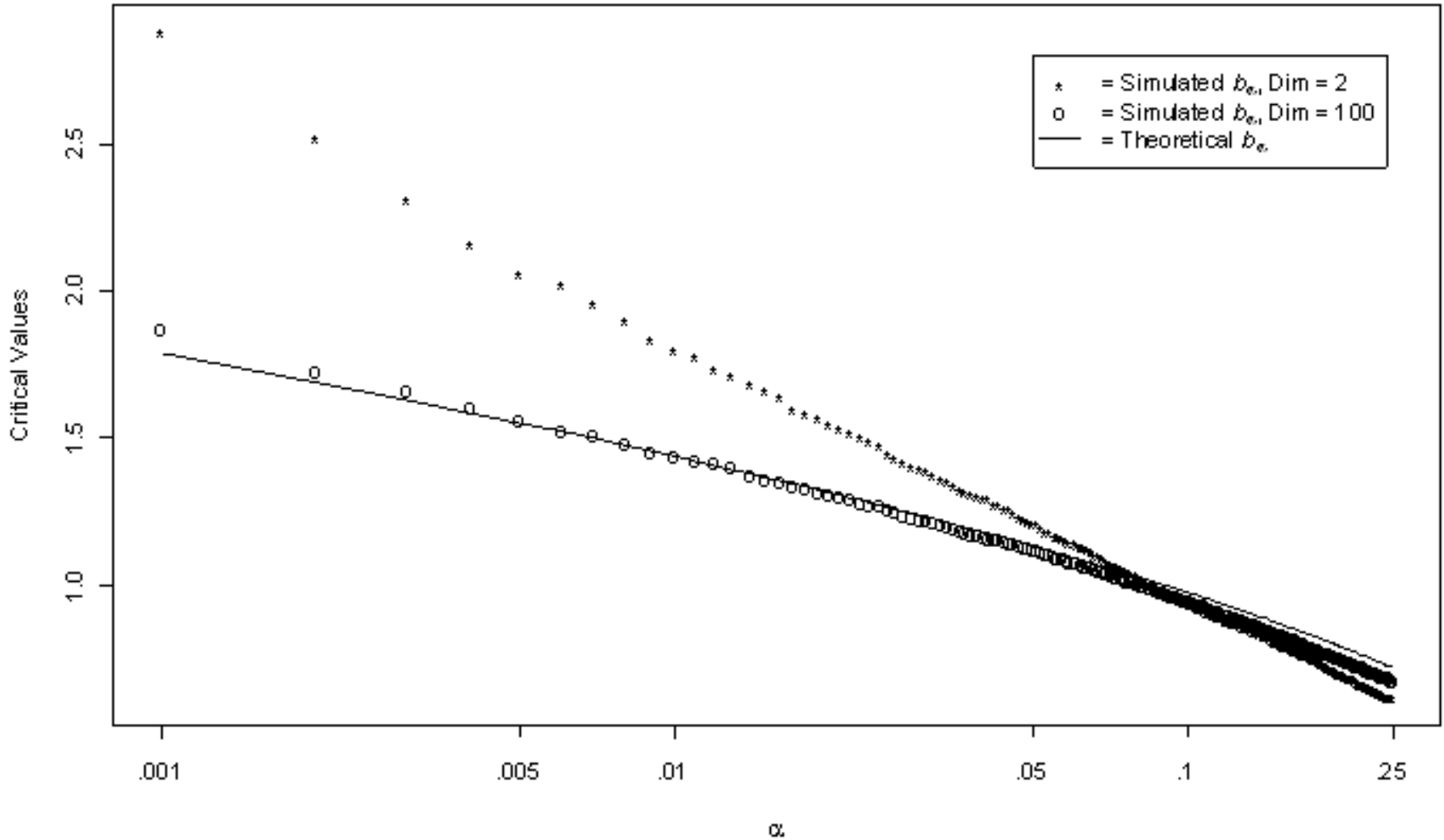
$$\hat{\psi}_{\text{ESPM}} = c_N^{-1} \max_{k \in \{1, \dots, n\}} \left( \xi_{k,N} - \sum_{j=1}^k \hat{\psi}_{\text{SPM}}(j) \right)$$

$$c_N^2 = N(N+1)(N-1)^2 / 180, \quad \xi_{k,N} = kN(N+1) / 3$$

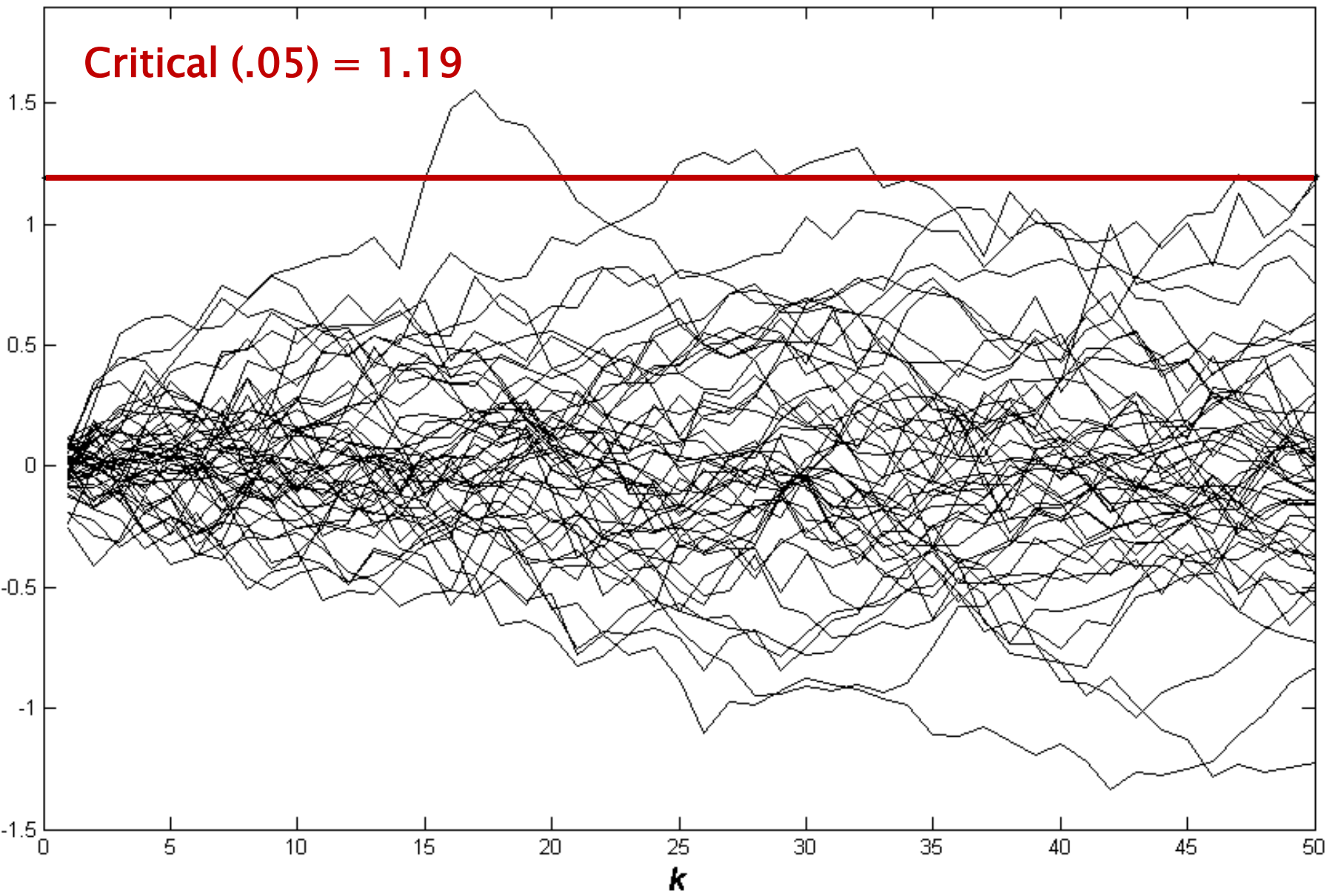
# Ensemble Tests

- (1) Under  $H_0$  the process  $B_N(t_k) = c_N^{-1} \left( \xi_{k,N} - \sum_{j=1}^k \hat{\psi}_{\text{SPM}}(j) \right)$  has the same first two moments as a Brownian bridge,  $t_k = k / (N - 1)$
- (2) Although the summands individually are asymptotically normal, the same is not true of the process itself!
- (3) Unless the dimensionality of the observations is very large, classical Brownian bridge theory (Shorack & Wellner, 1987) produces critical values that violate the nominal level
- (4) Ruth (2009) produced critical values for different values of  $N$  and  $d =$  dimensionality using extensive simulations

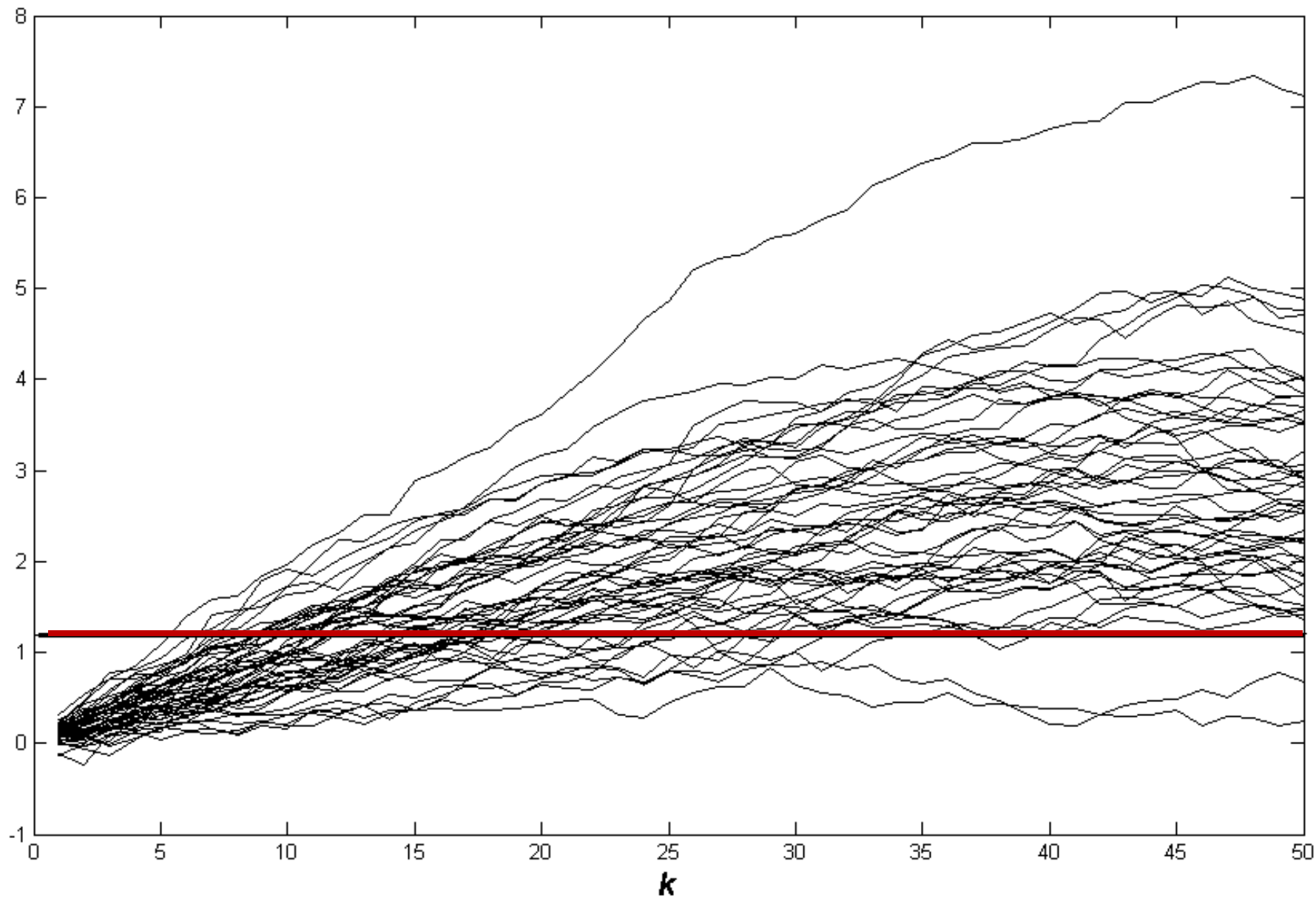
# Simulated critical values for $N = 200$



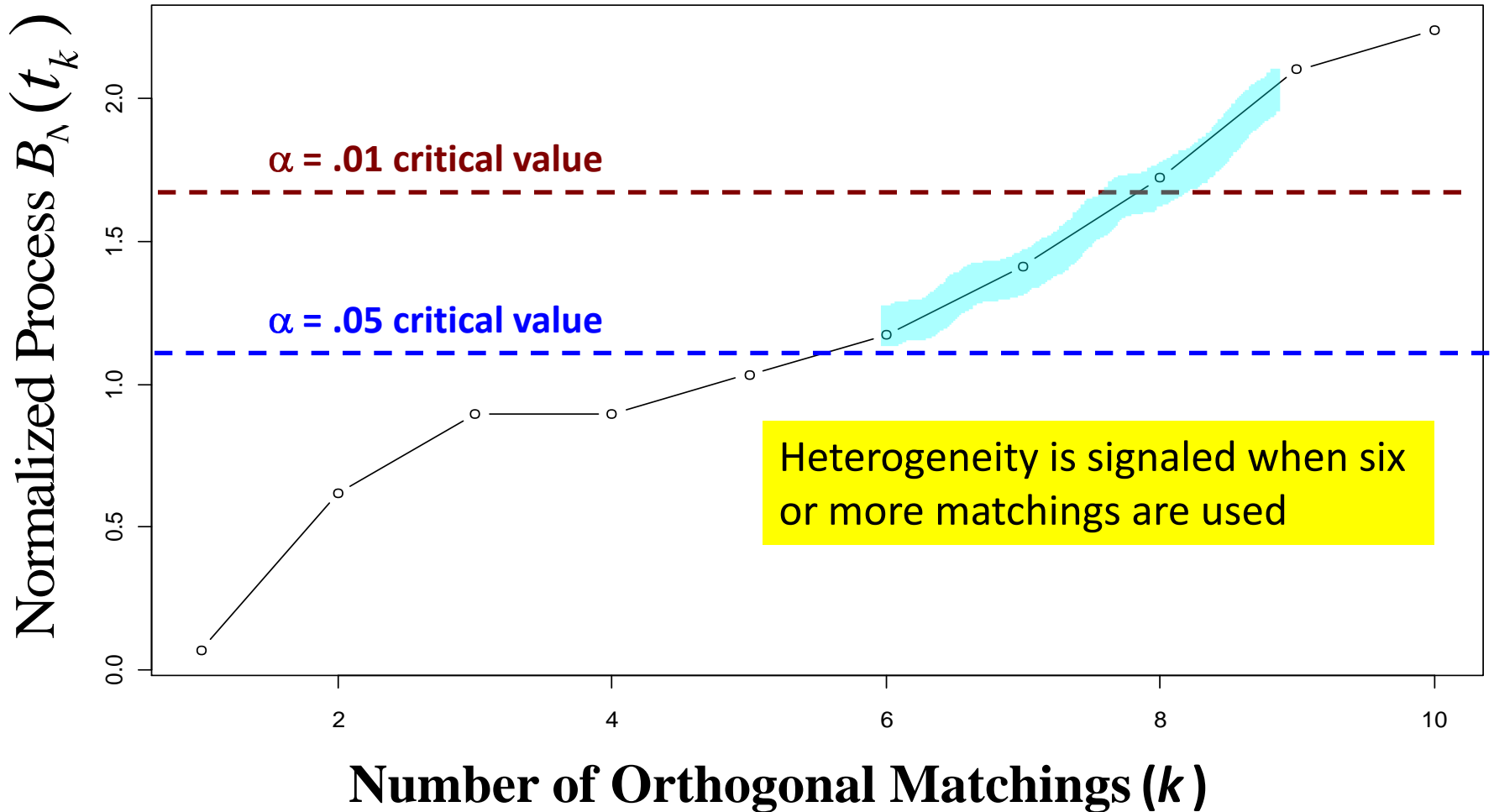
# 100 Simulated $B_N(t_k)$ , Bivar. Normal, Homogeneous



# 100 Simulated $B_N(t_k)$ , Bivar. Normal, Mean Jump



$\hat{\psi}_{\text{ESPM}} = 2.24$  has p-value less than .01



**Power simulations,  $N = 200$ , jump at observation 101,  $\Delta = \text{norm}$  of mean vector after the jump, nominal .05-level tests**

Multivariate normal,  $\theta = \text{mean}$ ,  $p = 5$

$\Delta$	SCM	SPM	ESPM	JJS
<b>0</b>	.05	.06	.04	.05
<b>.5</b>	.09	.10	.60	.52
<b>1.0</b>	.33	.41	1.00	1.00

Multivariate normal,  $\theta = \text{mean}$ ,  $p = 20$

	SCM	SPM	ESPM	JJS
<b>0</b>	.05	.05	.05	.03
<b>.5</b>	.07	.09	.33	.20
<b>1.0</b>	.16	.22	.95	.95

## Power simulations, $N = 200$ , jump at observation 101, nominal .05-level tests

Multivariate normal,  $\theta =$  covariance matrix,  $p = 5$

$1+\Delta$  mult.

	SCM	SPM	ESPM	JJS
<b>0</b>	.05	.06	.05	.04
<b>.5</b>	.42	.51	.97	.15
<b>1.0</b>	.99	.99	1.00	.24

Multivariate normal mixture,  $\theta =$  mean,  $p = 5$

$\Delta$  norm

	SCM	SPM	ESPM	JJS
<b>0</b>	.05	.05	.04	.27
<b>.5</b>	.08	.09	.56	.38
<b>1.0</b>	.25	.36	.99	.85

# Graph-theoretic Tests: Some Challenges and Possible Directions

1. Computational
2. Theoretical
3. Alternate graph-theoretic approaches
4. Adaptation to real-world problems

# Computational Challenges

Finding a MNBM requires  $O(Nm \log(N))$  computation time using the Blossom V algorithm (Kolmogorov, 2009). For the complete graph,  $m \propto N^2$ . For ensemble tests the order of computation is about  $O(N^4 \log(N))$  which is prohibitive with large sample sizes (e.g.  $N > 1000$ ).

## Possible strategies:

- (1) Use a greedy algorithm
- (2) Restrict the edge set ( $m \propto N$ )
- (3) Try something else

# Faster Matchings?

- Simple greedy heuristics are difficult to extend to multiple matchings
- Edge restriction heuristics. Sufficient conditions for a perfect matching to exist ( $N$  even) include
  - A regular graph of degree  $\geq \lfloor N/2 \rfloor$
  - A connected, claw-free graph
  - A Delaunay triangulation
- Necessary and sufficient conditions: Tutte's Theorem
$$\text{odd}(V - S) \leq |S| \text{ for all } S \subset V$$

# Are MNBM tests universally consistent?

- Asymptotic theory for MNBM is not straightforward even for a single matching, let alone ensembles.
- Aldous & Steele (1992) theory for MSTs exploits perturbation localizability of MSTs (not applicable to matchings).
- Interesting recent work: "Poisson Matching" (Holroyd *et al.* 2008)

MNBM is a solution to the integer linear program

Minimize: 
$$f(\mathbf{x}) = \sum_{i < j}^n \sum x_{ij} d_{ij}$$

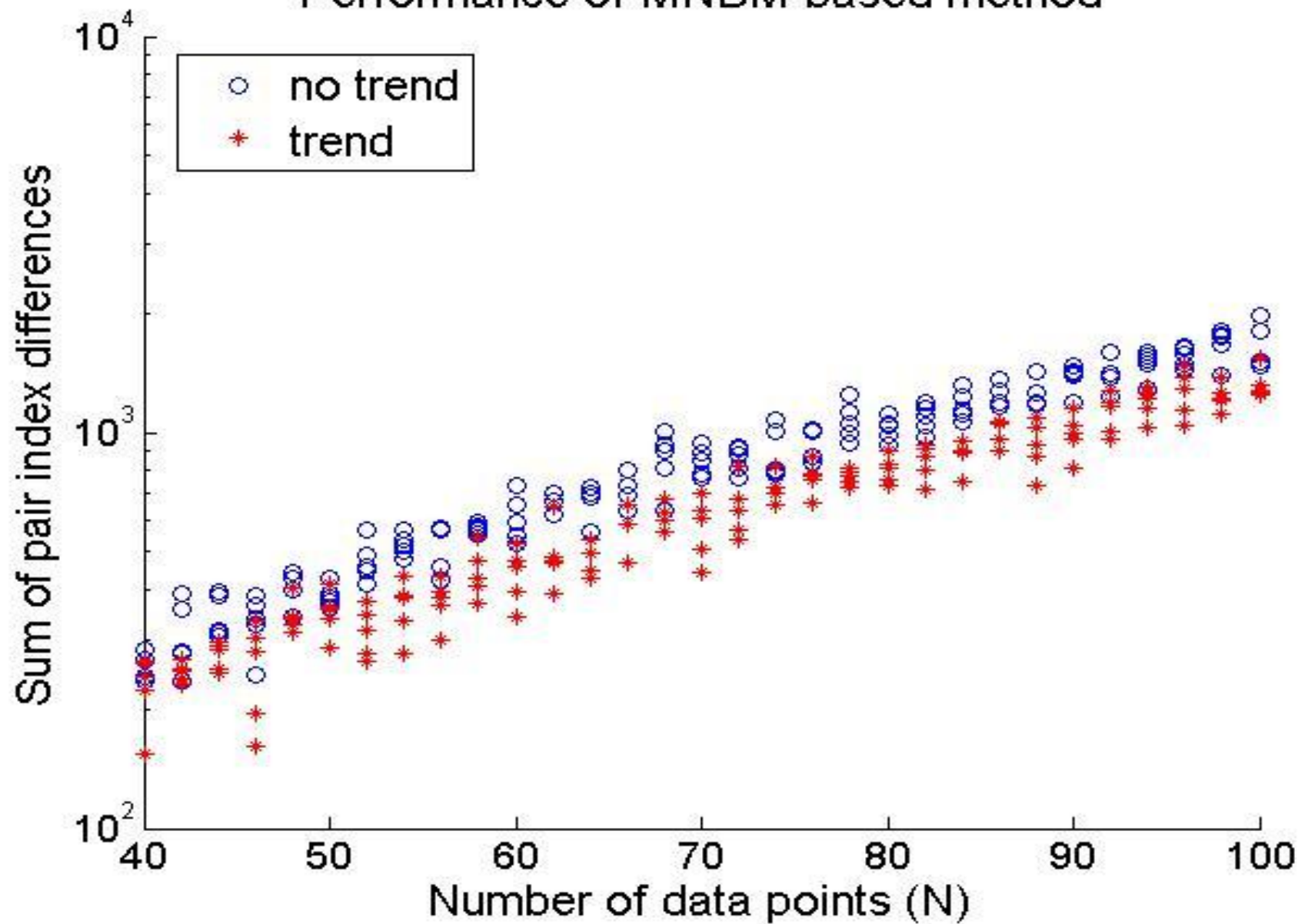
Subject to: 
$$x_{ij} \in \{0, 1\}, \sum_{\substack{i=1 \\ i \neq j}}^n x_{i \wedge j, i \vee j} = 1, j = 1, \dots, n$$

By replacing the integrality constraints with the interval constraints  $0 \leq x_{ij} \leq 1$  a solution can be obtained using LP. A "relaxed" SPM statistic can be defined by

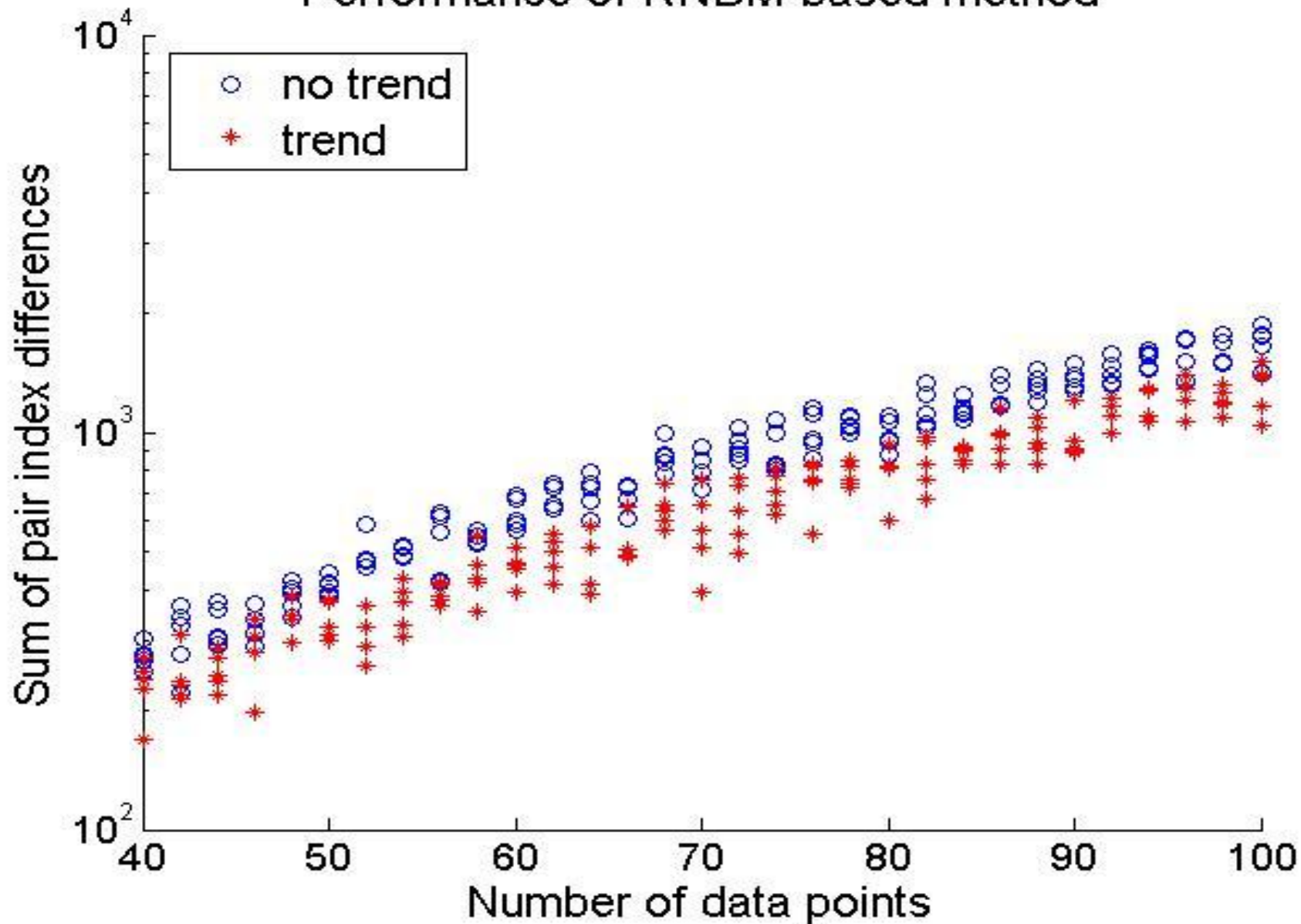
$$\hat{\psi}_{\text{RSPM}} = \sum_{i < j}^n j \cdot \hat{x}_{ij} = \frac{1}{2} \sum_{i < j}^n |i - j| \hat{x}_{ij} + \frac{1}{4} N(N + 1)$$

- Solutions to RNBM satisfy  $\hat{x}_{ij} \in \{0, \frac{1}{2}, 1\}$
- To fit ensembles enforce the constraints  $0 \leq \hat{x}_{ij} \leq k, k = 1, \dots, n$  over a sequence of problems. There is no assurance that solutions will be "nested", however, which complicates theory
- Performance of relaxed MNBM statistics compares favorably with that of regular MNBM
- What about nearest neighbors?

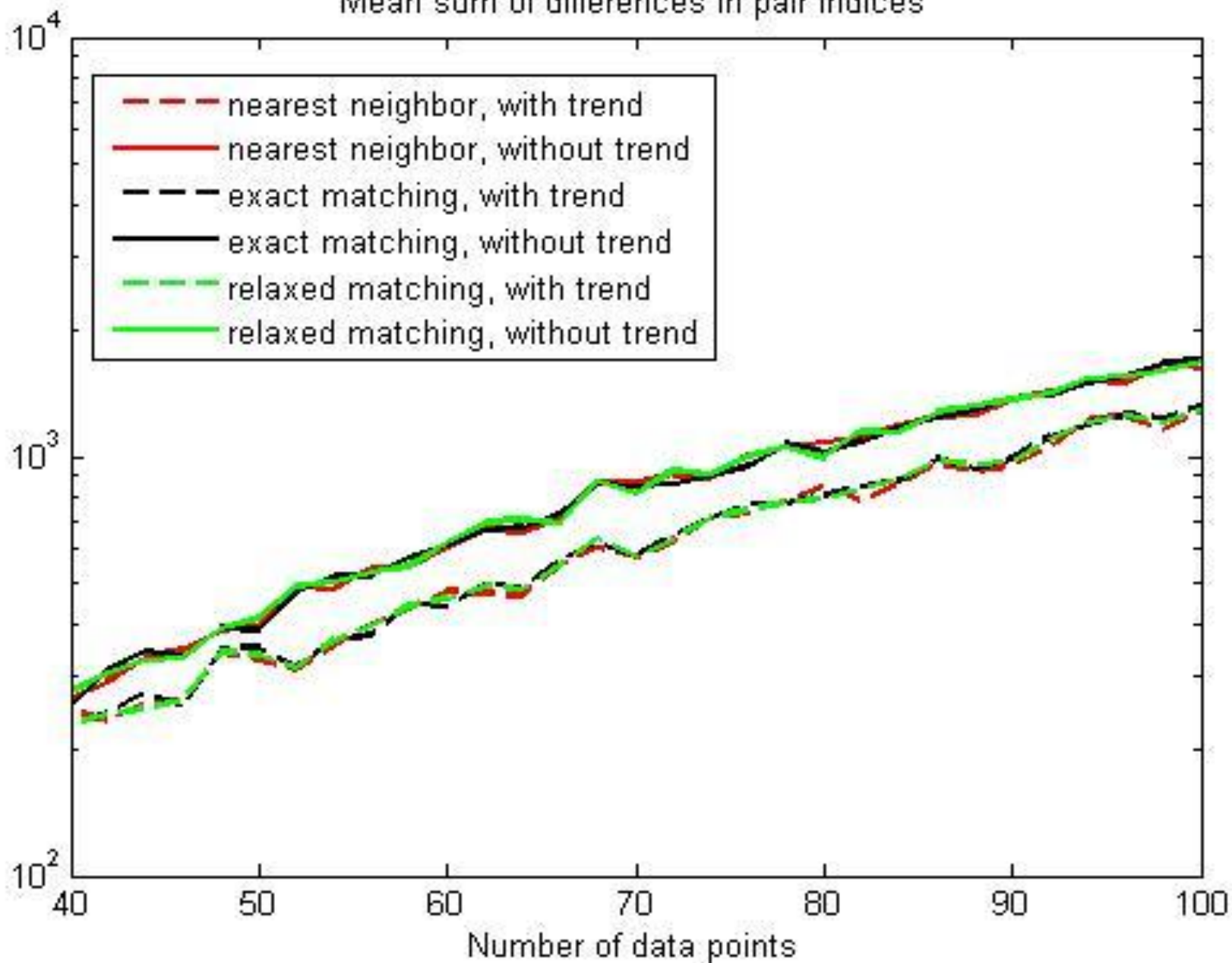
# Performance of MNBM-based method



# Performance of RNBM-based method



Mean sum of differences in pair indices



# Possible Applications

- Process control (off-line, on-line)
- Mechanical prognostics
- Threat detection
- Syndromic surveillance

In high-dimensional problems, it may be useful to couple graph-theoretic methods with methods to reduce dimensionality

# Dimension reduction

Consider the optimization problem

$$\min_{\mathbf{w}} \sum_{(i,j) \in E} |i - j| x_{ij}(\mathbf{w})$$

$$\text{s.t. } \mathbf{x}(\mathbf{w}) = \operatorname{argmin}_{\mathbf{x} \in X} \sum_{(i,j) \in E} \left\| \mathbf{w}^T (\mathbf{y}_i - \mathbf{y}_j) \right\| x_{ij}$$

$$\mathbf{w} \in \{0, 1\}^p$$

$$\sum_r w_r = p'$$

Vector  $\mathbf{w}$  projects into a low – dimensional space to minimize the sum of pair index differences in the resulting minimum – weight matching

- Simplification 1: use Manhattan distance:

$$d_{ij} = \sum_r d_{ijr} w_r, \quad d_{ijr} = |y_{ir} - y_{jr}|$$

- Simplification 2: use relaxed matching instead of exact matching; enforce minimum-weight matching using strong duality.

$$\min_{\mathbf{w} \in \{0,1\}^p, \mathbf{x} \geq \mathbf{0}, \boldsymbol{\pi}} \sum_{(i,j) \in E} |i - j| x_{ij}$$

$$\text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{1}$$

$$\sum_{v \in V} a_{v,(i,j)} \pi_v \leq \sum_r d_{ijr} w_r \quad \forall (i,j) \in E$$

$$\sum_{v \in V} \pi_v = \sum_{(i,j) \in E} \sum_r d_{ijr} w_r x_{ij}$$

$$\sum_r w_r = p'$$