LP Approach to Statistical Modeling

Deep Mukhopadhyay

Department of Statistics
Temple University

deep@temple.edu

Joint work with Emanuel Parzen

May 2, 2014

The Mixed Data Problem or The 'Variety' Problem

Simple Scenario. We have (X, Y). Goal: Understand the Relationship.

Y	X	**Statistical Measures**
Continuous	Continuous	Spearman's correlation
Discrete	Discrete	Chi-squared Statistic
Binary	Binary	Pearson's ϕ-coefficient
Binary	Continuous	Wilcoxon Statistic
Discrete	Continuous	Kruskal–Wallis H-Statistic

GOAL: To Design Mixed Data Algorithm.

Algorithms that *do not require* the users to specify the 'types' of variables (e.g., binary/discrete/continuous...)

Brute-force Naive Attempt & Challenges

Statistical Measure	R-function
Spearman's correlation	`cor`$(X, Y,$ method$=$"spearman"$)$
Chi-squared Statistic	`chisq.test`(X, Y)\$statistic
Pearson's ϕ-coefficient	`phi`(X, Y)
Wilcoxon Statistic	`wilcox.test`(X, Y)\$statistic
Kruskal–Wallis H-Statistic	`kruskal.test`(X, Y)\$statistic

R-code

```
if(case==1) return(cor(X, Y, method="spearman"));
if(case==2) return(chisq.test(X, Y)$statistic);
    ⋮     ⋮
if(case==5) return(kruskal.test(X, Y)$statistic);
```

The Challenge of 'Variety' Problem

Few Facts already been Recognized by Big Data Experts:

- "Research problem! Killing most CIO's that I know. If there is any achilles heel in big data, this is it!", White House Office of Science & Technology Policy and MIT, March 3, 2014.

- "Variety Is the Unsolved Problem in Big Data, especially difficult to solve Programmatically"

- "Today's Big Data Challenge Stems From Variety, Not Volume or Velocity"

- "Of the famous big data Vs, it's the variety in data that holds the most potential for exploitation."

Specific Goals of this Talk

- Application of the *General Theory* (LP Statistical Data Science) to construct Dependence Measures, Models and Graphical Exploratory Tool that is valid for Mixed Data (discrete or continuous).

- LP Algorithm *Simultaneously* tackles: Copula density estimation, Correlation measures, Contingency table modeling, Nonlinear regression, Quantile Regression, Correspondence analysis, ...Thus permits *easy way to establish relationship* among various methods

Design "Single" General Algorithm that tackles different varieties of *data types, data patterns, and data structures.*

Specific Goals of this Talk

- Application of the *General Theory* (LP Statistical Data Science) to construct Dependence Measures, Models and Graphical Exploratory Tool that is valid for Mixed Data (discrete or continuous).

- LP Algorithm *Simultaneously* tackles: Copula density estimation, Correlation measures, Contingency table modeling, Nonlinear regression, Quantile Regression, Correspondence analysis, ...Thus permits *easy way to establish relationship* among various methods

Design "Single" General Algorithm that tackles different varieties of *data types, data patterns, and data structures.*

LP StatScience: Simplify, Generalize and Unify

LP Unification: Traditional Statistical Measures

- Spearman Correlation (discrete/continuous): $\mathrm{LP}(1,1;X,Y)$.
- Wilcoxon Statistics: $\mathrm{LP}(1,1;X,Y)$.
- Pearson ϕ for 2×2 table: $\mathrm{LP}(1,1;X,Y)$.
- χ^2 Divergence Statistics: $\sum_{j,k}\left|\mathrm{LP}(j,k;X,Y)\right|^2$.
- Gini Correlation: $\mathrm{LP}(1,0;X,Y)$, $\mathrm{LP}(0,1;X,Y)$.
- Pearson Correlation: $\sum_{j,k}\mathrm{LP}[j,0;X,X]\,\mathrm{LP}[j,k;X,Y]\,\mathrm{LP}[k,0;Y,Y]$.
- Maximal Correlation: $\max_{j,k>0}\left|\mathrm{LP}[j,k;X,Y]\right|$.

LP STAT provides easy way to establish relationship among various 'isolated' statistical ideas; more systematic and automatic algorithm, simplifies teaching and practice of statistics by unification.

Deserve to be: FUNDAMENTAL methods of Statistical Learning !

LP[1,1] as Generalized Spearman Correlation

- How to define a "properly normalized" Spearman correlation for discrete data with ties is an ongoing question.
- Denuit and Lambert (2005), Nešlehová(2007), Mesfioui and Quessy (2010), Blumentritt and Schmid (2012), and Genest et al. (2013)...
- Range of Spearman for discrete data with ties depends on marginals and typically does not reach the bounds ± 1

The Genest Example

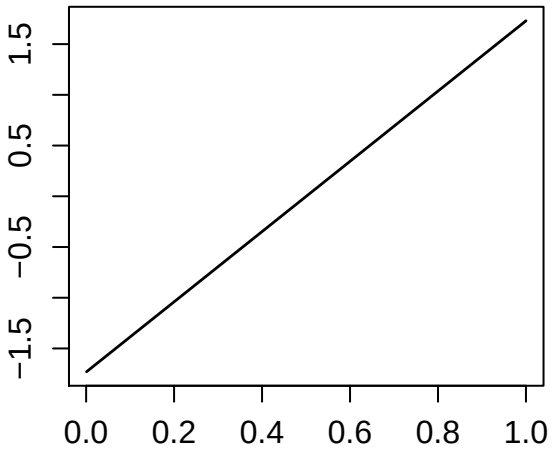
(X, Y) Bernoulli $\Pr(X = 0) = \Pr(Y = 0) = \Pr(X = 0, Y = 0) = p$, which implies $Y = X$ almost surely, still the traditional Spearman $p(1 - p) < 1$. Similarly, if $\Pr(X = 0) = p = 1 - q = \Pr(Y = 1)$ and $\Pr(X = 0, Y = 1) = p$, then $Y = 1 - X$ almost surely, yet Spearman $-p(1 - p) > -1$.

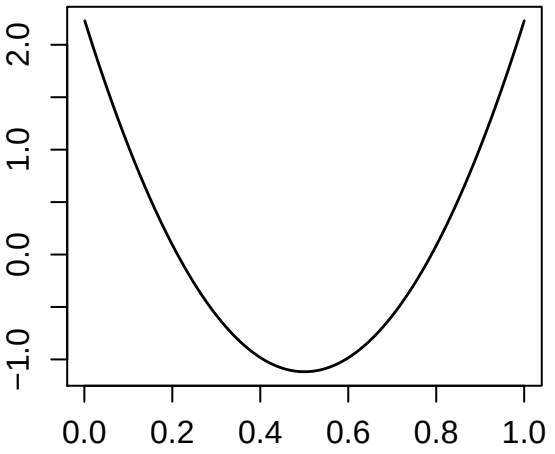
The LP Answer ± 1.

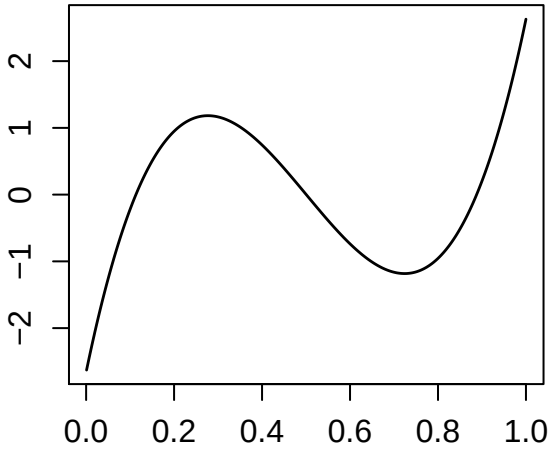
Construction of LP Orthonormal Score Functions

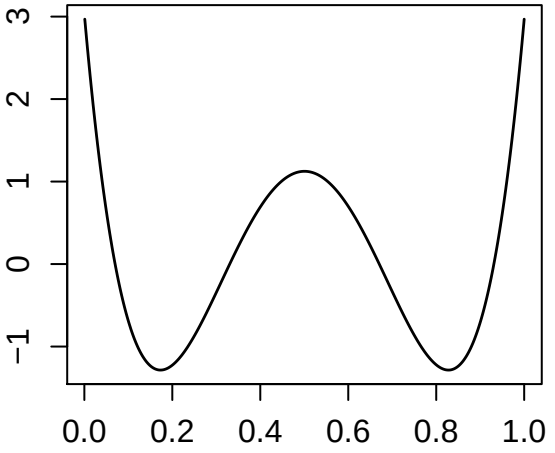
- Random variable X, distribution $F(x) = F(x; X)$.
- Quantile $Q(u) = Q(u; X)$, $0 < u < 1$.
- Mid-distribution $F^{\mathrm{mid}}(x; X) = F(x; X) - .5p(x; X)$.
- $\mathbb{E}[F^{\mathrm{mid}}(X; X)] = .5, \mathrm{Var}[F^{\mathrm{mid}}(X; X)] = (1/12)(1 - \sum_x p^3(x; X))$.
- LP SCORE Function $T_j(x; X)$.
- LP UNIT Score Function $S_j(u; X) = T_j(Q(u; X); X)$, $0 < u < 1$.
- $T_j(X; X)$ Gram Schmidt orthonormalization powers of $T_1(X; X)$.
- $T_1(X; X) = \mathcal{Z}(F^{\mathrm{mid}}(X; X))$ where $\mathcal{Z}(X) = (X - \mathbb{E}(X))/\sigma(X)$.
- $S_j(u; X)$ are orthonormal in $L^2(F)$, the Hilbert space of square integrable functions w.r.t the measure F (discrete or continuous).

$X \leftarrow \texttt{rnorm}(n = 200,\ \mu = 0,\ \sigma = 1);\quad \texttt{LP.poly}(X, m = 4)$

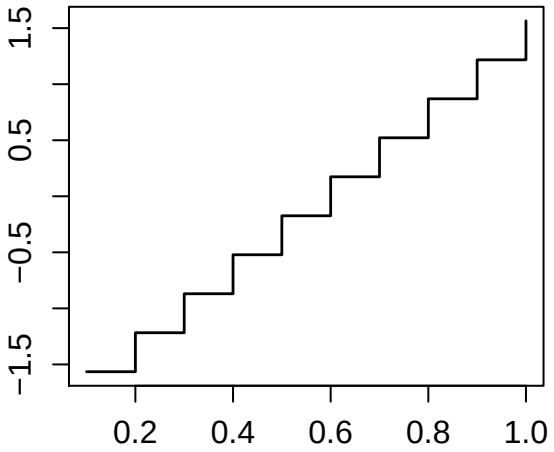


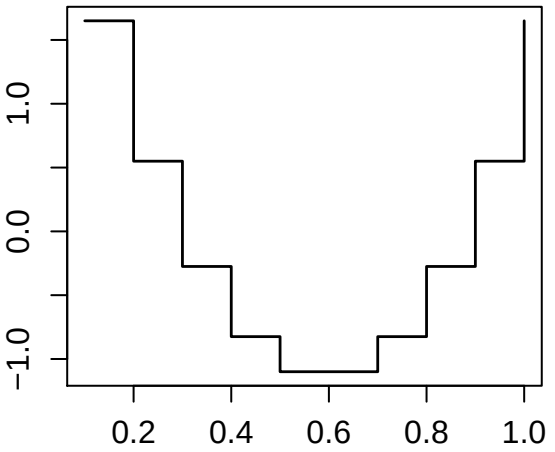


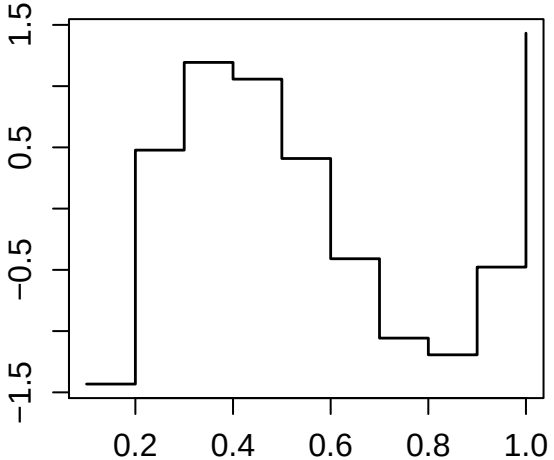


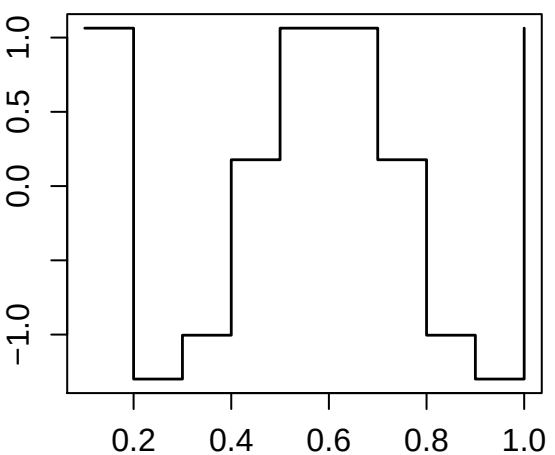


$X \leftarrow$ `Discrete.Uniform`$\{1, 2, \ldots, 10\}$; `LP.poly`$(X, m = 4)$

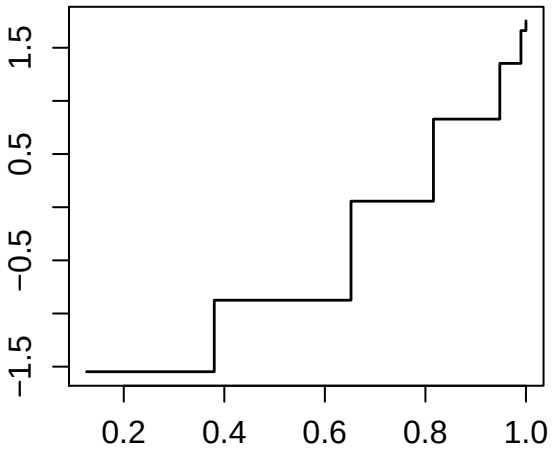


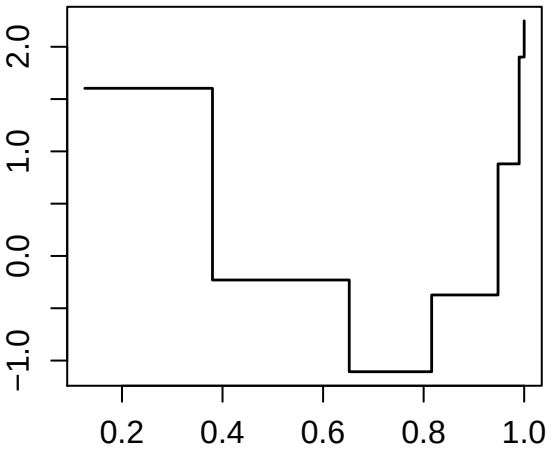


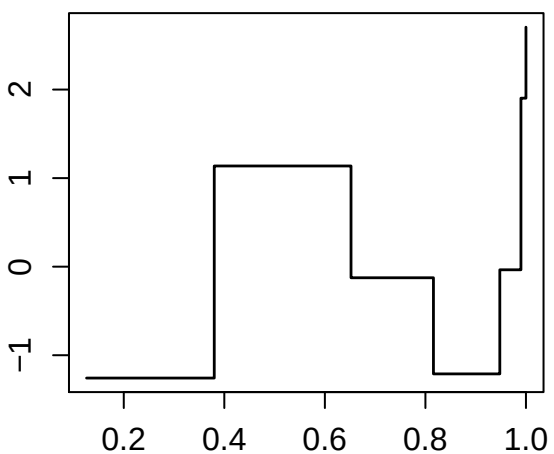


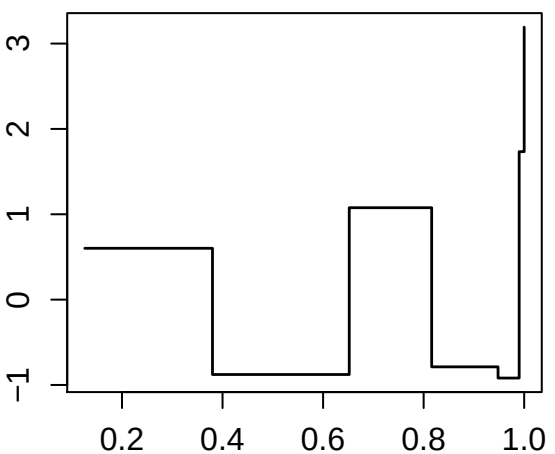


$X \leftarrow$ `rpois`$(n = 200,\ \lambda = 2)$; `LP.poly`$(X, m = 4)$









LP Moments, LP Comoments, Copula Density, LPINFOR

- $\mathrm{LP}(j; X) = \mathrm{LP}(j, 0; X, X) = \mathbb{E}[X T_j(X; X)]$.
- $\mathrm{LP}(j, k; X, Y) = \mathbb{E}[T_j(X; X)\, T_k(Y; Y)]$.
- Variance Decomposition: $\mathrm{Var}[X] = \sum_j |\mathrm{LP}(j; X)|^2$.
- $\mathrm{Cov}(X, Y) = \sum_{j,k>0} \mathrm{LP}(j; X)\, \mathrm{LP}(k; Y)\, \mathrm{LP}(j, k; X, Y)$.
- Dependence (X, Y): $\mathrm{LPINFOR}(X, Y) = \sum_{j,k>0} \left|\mathrm{LP}(j, k; X, Y)\right|^2$.
- FUNDAMENTAL LP Representation Theorem (For *'Mixed'* data)

$$\mathrm{cop}(u, v; X, Y) - 1 = \sum_j \mathrm{LP}(j, k; X, Y)\, S_j(u; X)\, S_k(v; Y),$$

$$\log \mathrm{cop}(u, v; X, Y) = \sum_{j,k>0} \theta_{j,k}\, S_j(u; X)\, S_k(v; Y) - K(\theta).$$

- Piecewise constant Checkerboard plot of $\mathrm{cop}(u, v; X, Y)$.

LP Moments, LP Comoments, Copula Density, LPINFOR

- $\mathrm{LP}(j;X) = \mathrm{LP}(j,0;X,X) = \mathbb{E}[XT_j(X;X)]$.
- $\mathrm{LP}(j,k;X,Y) = \mathbb{E}[T_j(X;X)\,T_k(Y;Y)]$.
- Variance Decomposition: $\mathrm{Var}[X] = \sum_j |\,\mathrm{LP}(j;X)|^2$.
- $\mathrm{Cov}(X,Y) = \sum_{j,k>0} \mathrm{LP}(j;X)\,\mathrm{LP}(k;Y)\,\mathrm{LP}(j,k;X,Y)$.
- Dependence (X,Y): $\mathrm{LPINFOR}(X,Y) = \sum_{j,k>0} \left|\,\mathrm{LP}(j,k;X,Y)\right|^2$.
- FUNDAMENTAL LP Representation Theorem (For *'Mixed'* data)

$$\mathrm{cop}(u,v;X,Y) - 1 = \sum_j \mathrm{LP}(j,k;X,Y)\,S_j(u;X)\,S_k(v;Y),$$

$$\log \mathrm{cop}(u,v;X,Y) = \sum_{j,k>0} \theta_{j,k}\,S_j(u;X)\,S_k(v;Y) - K(\theta).$$

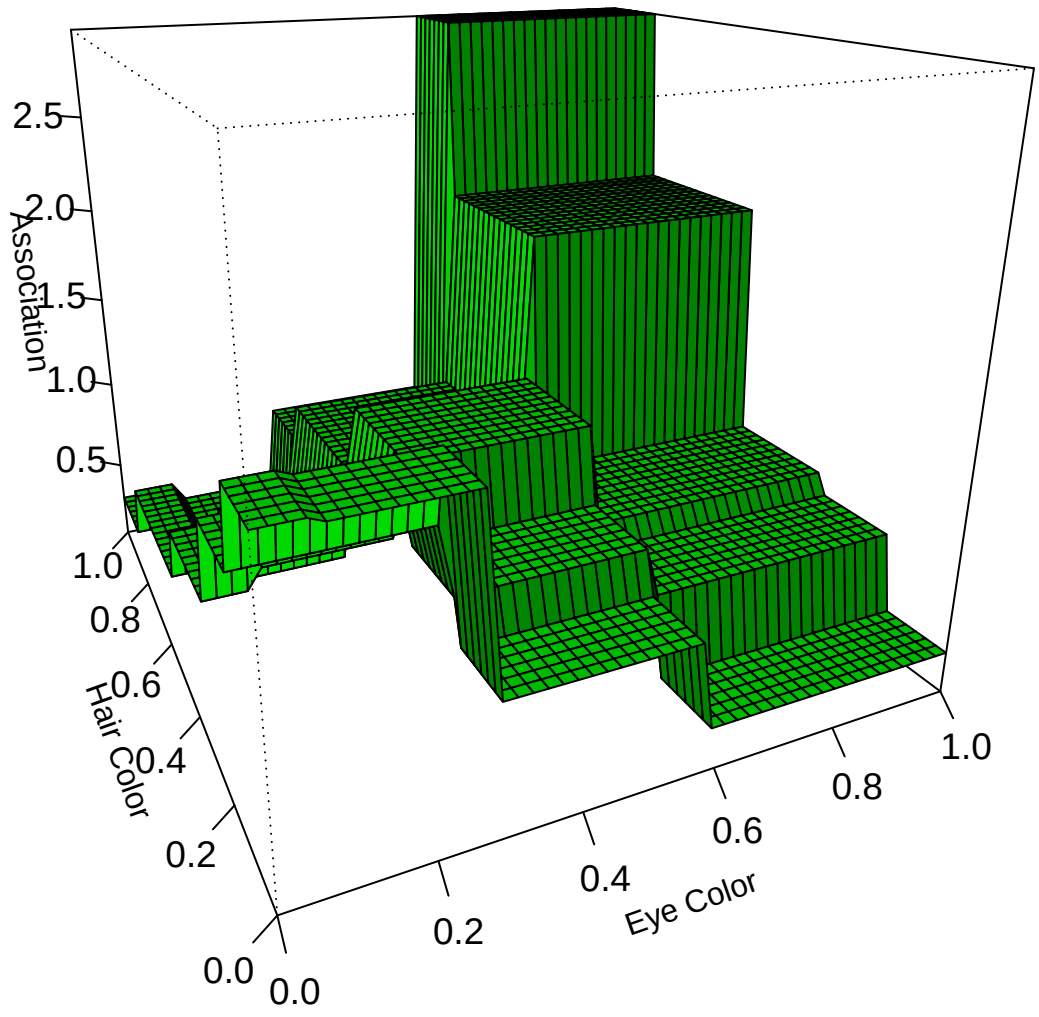
- Piecewise constant Checkerboard plot of $\mathrm{cop}(u,v;X,Y)$.

Fisher's Hair and Eye Color Data (1940)

Eye Color	**Hair Color**				
	Fair	Red	Medium	Dark	Black
Blue	326	38	241	110	3
Light	688	116	584	188	4
Medium	343	84	909	412	26
Dark	98	48	403	681	85

Table: Two way contingency table classifying 5387 children of Scotland.

LP Nonparametric Checkerboard Copula Estimate



$$\widehat{\mathrm{cop}}(u, v; X, Y) = 1 + .42 S_1(u)\, S_1(v) + .12 S_2(u)\, S_1(v) + .16 S_2(u)\, S_2(v)$$

LP Canonical Copula Expansion, Correspondence Analysis

- LP Representation: Singular values of LP-comoment kernel

$$\text{cop}(u,v) - 1 = \sum_{k=1}^{\infty} \lambda_k \, \phi_k(u;X) \, \psi_k(v;Y),$$

$$\log \text{cop}(u,v) = \sum_{k=1}^{\infty} \gamma_k \, \phi_k(u;X) \, \psi_k(v;Y) - K(\gamma).$$

- (X, Y) Discrete (contingency table): Unifies Goodman (1991,1996).

Correspondence Analysis: Given a $I \times J$ two-way contingency table, GOAL graphically display the association among the row and column categories.

Step A. Estimate LP canonical copula density, perform SVD of LP(X, Y).

Step B. LP profile coordinates $\mu_{ik} = \lambda_k \phi_k(u_i; X)$ and $\nu_{jk} = \lambda_k \psi_k(v_j; Y)$.

Step C. Jointly display the row and column profiles in the same plot.

LP Canonical Copula Expansion, Correspondence Analysis

- LP Representation: Singular values of LP-comoment kernel

$$\text{cop}(u,v) - 1 = \sum_{k=1}^{\infty} \lambda_k \, \phi_k(u;X) \, \psi_k(v;Y),$$

$$\log \text{cop}(u,v) = \sum_{k=1}^{\infty} \gamma_k \, \phi_k(u;X) \, \psi_k(v;Y) - K(\gamma).$$

- (X, Y) Discrete (contingency table): Unifies Goodman (1991,1996).

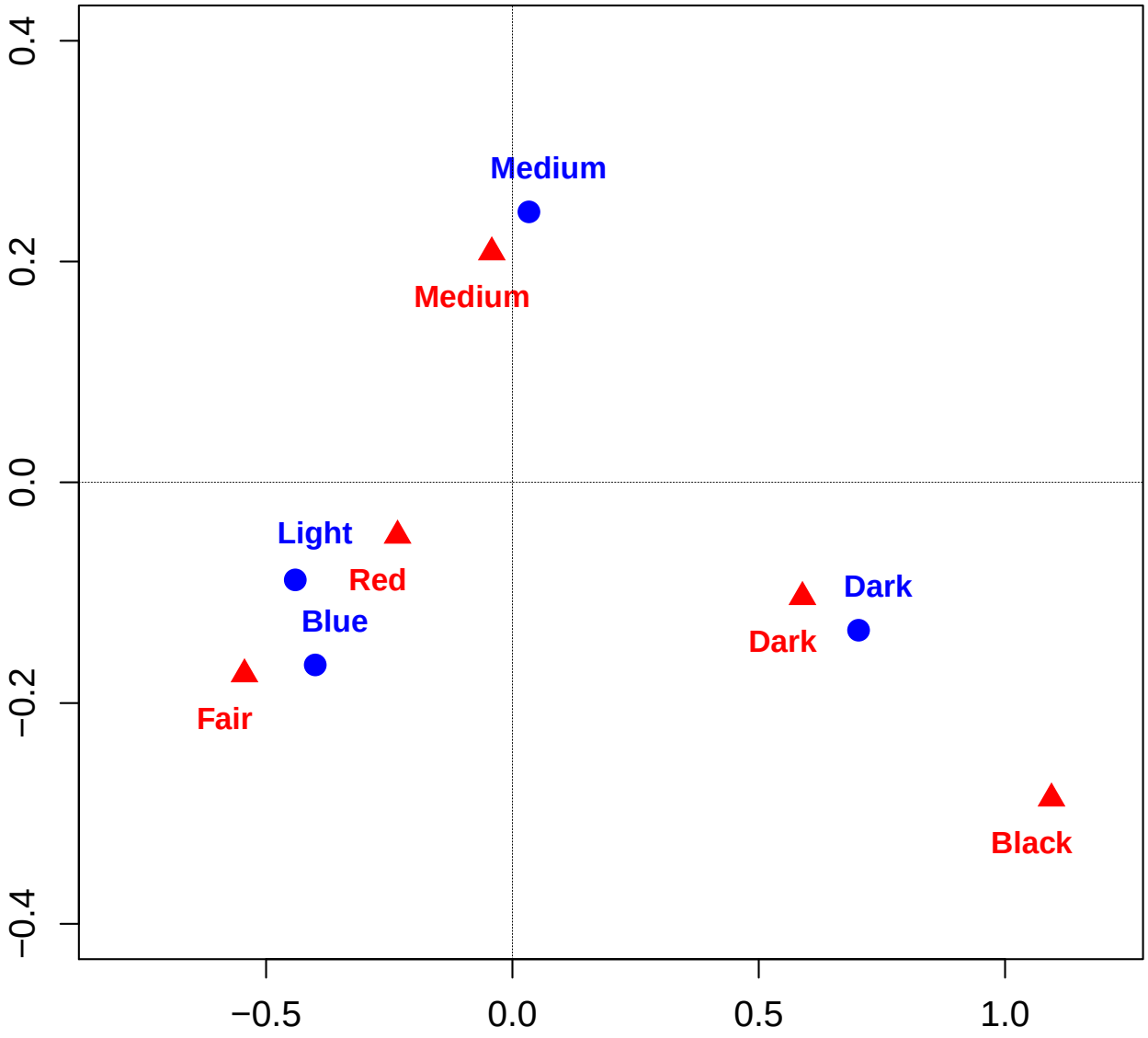
Correspondence Analysis: Given a $I \times J$ two-way contingency table, GOAL graphically display the association among the row and column categories.

Step A. Estimate LP canonical copula density, perform SVD of LP(X, Y).

Step B. LP profile coordinates $\mu_{ik} = \lambda_k \phi_k(u_i; X)$ and $\nu_{jk} = \lambda_k \psi_k(v_j; Y)$.

Step C. Jointly display the row and column profiles in the same plot.

LP Correspondence Map



LP Unification of Two Cultures of Correspondence Analysis

- LP L^2 canonical copula-based row and column scoring reproduces the classical correspondence analysis pioneered by Hirschfeld (1935) and Benzecri (1969).

- LP canonical exponential copula reproduces Goodman's profile coordinates, Goodman (1981) and Goodman (1985), which leads to the graphical display sometimes known as a weighted logratio map (Greenacre, 2010) or the spectral map (SM) (Lewi, 1998).

- 'Two Cultures': Anglo multivariate analysis (usual normal theory) and French multivariate analysis (matrix algebra data analysis), using LP representation theory of discrete checkerboard copula density.

Low-Rank Smoothing Model, Smart Computational Algorithm

- Traditional Approach: SVD *raw* distance matrix (also known as contingency ratios) order I and J

$$\begin{aligned} \mathrm{cop}(F(x), F(y)) &= \frac{p(x, y; X, Y)}{p(x; X)p(y; Y)} \\ \log \mathrm{cop}(F(x), F(y)) &= \log \frac{p(x, y; X, Y)}{p(x; X)p(y; Y)} \end{aligned}$$

- High computational complexity and storage requirements large data
- Numerically unstable for sparse tables.
- LP Approach: SVD on LP comoment matrix, which is often a *far smaller order than the dimensional of the observed frequency matrix*.
- LP algorithm compresses the structure of large tables, enormous computational gain for large structured data.

Low-Rank Smoothing Model, Smart Computational Algorithm

- Traditional Approach: SVD *raw* distance matrix (also known as contingency ratios) order I and J

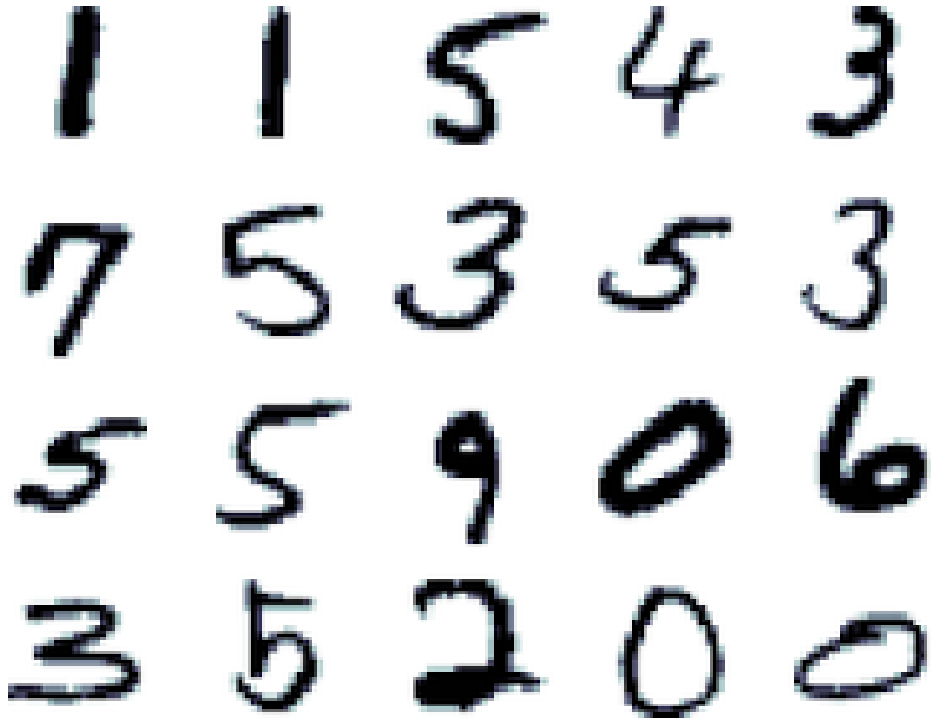
$$\operatorname{cop}(F(x), F(y)) = \frac{p(x, y; X, Y)}{p(x; X)p(y; Y)}$$

$$\log \operatorname{cop}(F(x), F(y)) = \log \frac{p(x, y; X, Y)}{p(x; X)p(y; Y)}$$

- High computational complexity and storage requirements large data
- Numerically unstable for sparse tables.

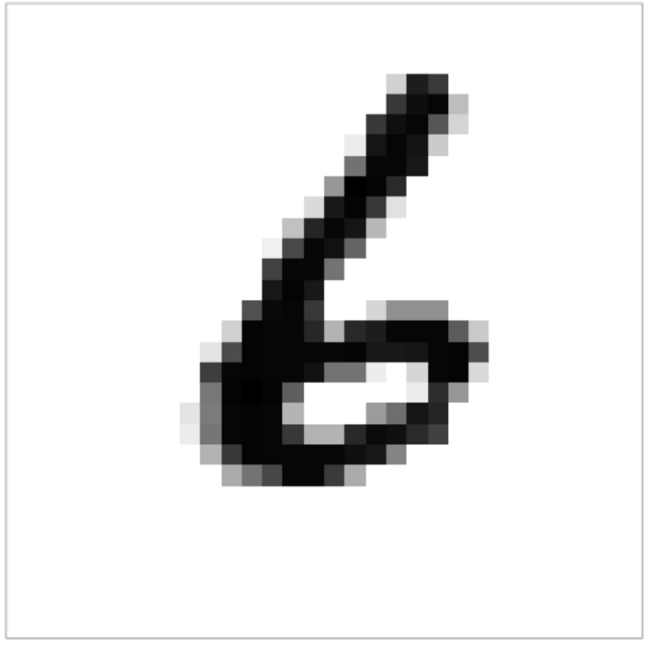
- LP Approach: SVD on LP comoment matrix, which is often a *far smaller order than the dimensional of the observed frequency matrix*.
- LP algorithm compresses the structure of large tables, enormous computational gain for large structured data.

MNIST: Image of Handwritten digits



- Famous database of $70,000$ images of handwritten digits.
- Goal is to build classifier: take an image of a handwritten single digit, and determine what that digit is.
- How to create sparse features to capture the structure ?

Traditional Approach to Build Sparse Model

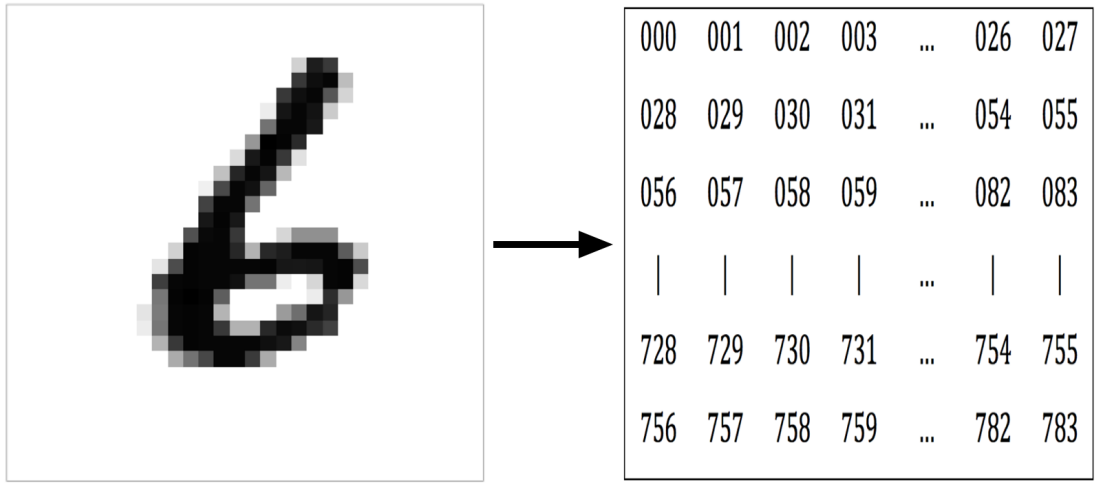


000	001	002	003	...	026	027
028	029	030	031	...	054	055
056	057	058	059	...	082	083
|	|	|	|	...	|	|
728	729	730	731	...	754	755
756	757	758	759	...	782	783

- Traditional Approach: Pixel matrix $\rightarrow$ Long vector of size $p \times p$
- Destroy the structure !
- Fit overparameterized model $\rightarrow$ Lasso $\rightarrow$ Simple (sparse) model.

Can we build a Sparse Nonparametric Model directly instead?

LPImage Compression: Sparse Coding, DIRECT Approach



$$\widehat{\mathrm{LP}}[X, Y] = \begin{bmatrix} 0.15 & 0.09 & -0.14 & -0.04 & -0.04 \\ 0.0 & -0.60 & -0.13 & 0.19 & 0.11 \\ -0.26 & -0.05 & 0.30 & 0.00 & -0.06 \\ 0.19 & 0.16 & -0.11 & 0.14 & 0.09 \\ 0.04 & -0.09 & 0.0 & 0.12 & -0.09 \end{bmatrix}$$

- Apply Logistic regression on the LP-features: $94 - 97\%$ accuracy.
- Traditional Lasso-logistic regression yields $88 - 92\%$ accuracy.

Table: IQ by males of various ages. Is there any pattern ? Data described in Mack and Wolfe (1981) and Rayner and Best (1996).

	Intelligence Scale														
Age group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$16-19$	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0
$20-34$	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0
$35-54$	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
$55-69$	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
$69+$	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0

χ^2 yields 60, with df $= 56$, p-value $= 0.3329$, fails to capture the pattern.

$$\widehat{\mathrm{LP}}[X = \mathrm{Age}, Y = \mathrm{IQ}] = \begin{bmatrix} -0.316 & 0.173 & 0.168 & -0.114 \\ -0.618^{*} & -0.031 & -0.101 & 0.068 \\ 0.087 & 0.136 & 0.077 & 0.037 \\ 0.165 & 0.215 & 0.042 & 0.289 \end{bmatrix}$$

Table: IQ by males of various ages. Is there any pattern ? Data described in Mack and Wolfe (1981) and Rayner and Best (1996).

	Intelligence Scale														
Age group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$16-19$	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0
$20-34$	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0
$35-54$	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
$55-69$	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
$69+$	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0

χ^2 yields 60, with df $= 56$, p-value $= 0.3329$, fails to capture the pattern.

$$\widehat{\mathrm{LP}}\left[X = \mathrm{Age}, Y = \mathrm{IQ}\right] = \begin{bmatrix} -0.316 & 0.173 & 0.168 & -0.114 \\ -0.618^{*} & -0.031 & -0.101 & 0.068 \\ 0.087 & 0.136 & 0.077 & 0.037 \\ 0.165 & 0.215 & 0.042 & 0.289 \end{bmatrix}$$

Nonlinear Correlation, LPINFOR, Power comparison

$$y = f(x) + \textbf{NOISE}, \quad X \sim U[0,1].$$

Setting 1. Under Gaussian error (σ), $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma \uparrow 3$.

Setting 2. Proposition of outliers(η), $\eta \uparrow .4$

$$\epsilon \sim (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(\mu, 3), \text{ where } \mu = \pm 5 \text{ w.p } 1/2$$

Setting 3. Level of 'bad' leverage points (η), $\eta \uparrow .4$

$$\epsilon \sim (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(\mu, 3), \text{ where } \mu \in \{\pm 20, \pm 30\} \text{ w.p } 1/4.$$

Setting 4. Heavy-Tailed error (σ), $\epsilon \sim \text{Cauchy}(0, \text{Scale} = \sigma)$, $\sigma \uparrow 2$.

Nonlinear Correlation, LPINFOR, Power comparison

$$y = f(x) + \textbf{NOISE}, \quad X \sim U[0,1].$$

Setting 1. Under Gaussian error (σ), $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma \uparrow 3$.

Setting 2. Proposition of outliers(η), $\eta \uparrow .4$

$$\epsilon \sim (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(\mu, 3), \text{ where } \mu = \pm 5 \text{ w.p } 1/2$$

Setting 3. Level of 'bad' leverage points (η), $\eta \uparrow .4$

$$\epsilon \sim (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(\mu, 3), \text{ where } \mu \in \{\pm 20, \pm 30\} \text{ w.p } 1/4.$$

Setting 4. Heavy-Tailed error (σ), $\epsilon \sim \text{Cauchy}(0, \text{Scale} = \sigma)$, $\sigma \uparrow 2$.

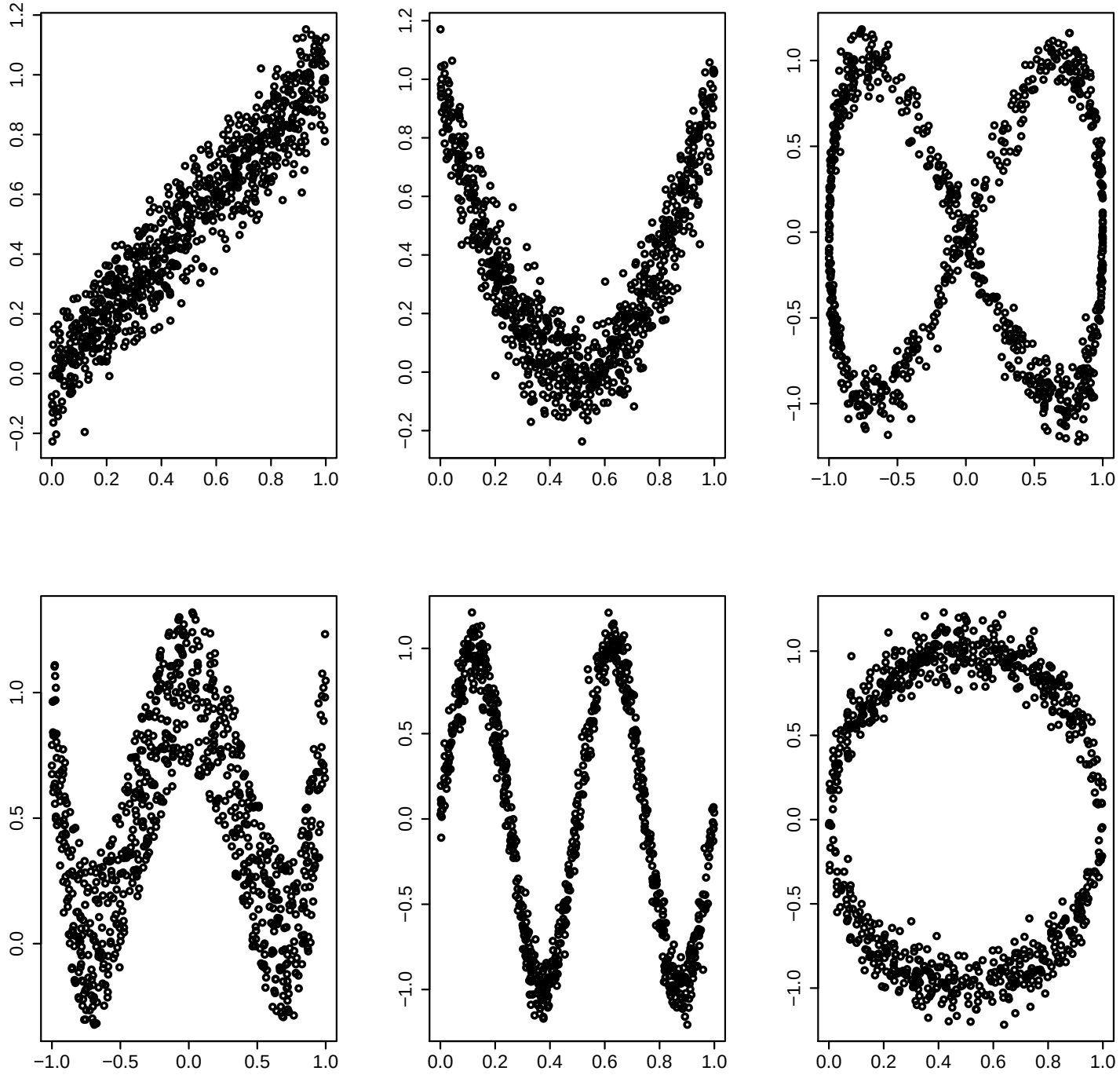


Table: Performance summary table: six competing methods, compared over six bivariate relationships and four different noise settings.

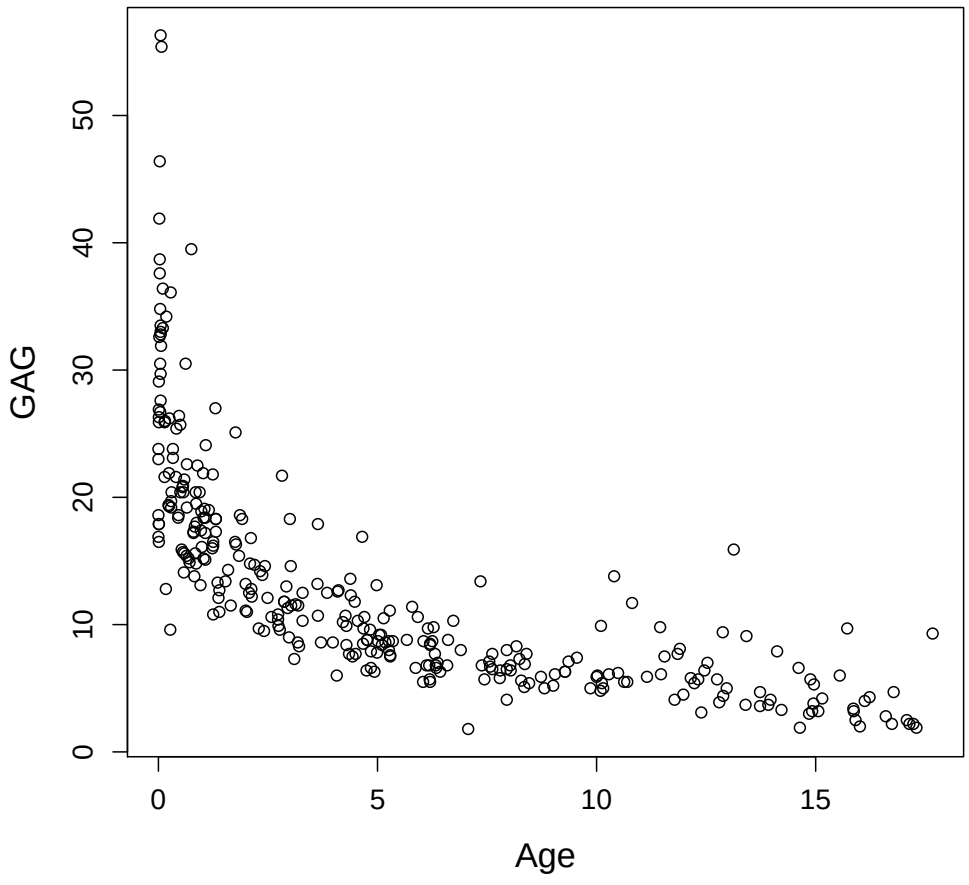
		Functional Relation					
Noise	Performance	**Linear**	**Quadratic**	**Lissajous**	**W-shaped**	**Sine**	**Circle**
E_1	Winner	Pearson	LPINFOR	LPINFOR	Dcor	Dcor	LPINFOR
	Runner-up	Spearman	Dcor	MIC	LPINFOR	MIC	MIC
E_2	Winner	Spearman	LPINFOR	LPINFOR	LPINFOR	MIC	LPINFOR
	Runner-up	Dcor	MIC	MIC	Dcor	Dcor	MIC
E_3	Winner	Spearman	LPINFOR	LPINFOR	LPINFOR	MIC	LPINFOR
	Runner-up	LPINFOR	MIC	MIC	MIC	LPINFOR	MIC
E_4	Winner	Spearman	LPINFOR	LPINFOR	LPINFOR	MIC	LPINFOR
	Runner-up	LPINFOR	MIC	Dcor	MIC	LPINFOR	MIC

Is it a Scalable Solution ? BIG 'n'

Table: Computational Complexity: uniformly distributed, independent samples of size n, averaged over 50 runs based on Intel(R) Core(TM) i7-3540M CPU @ 3.00GHz 2 Core(s) processor. **Timings are reported in seconds**.

	Size of the data sets				
Methods	$n = 500$	$n = 1000$	$n = 2500$	$n = 5000$	$n = 10,000$
LP	0.001 (.004)	0.002 (.005)	0.005 (.007)	0.011 (.007)	0.018 (.007)
Dcor	0.094 (.013)	0.371 (.013)	1.773 (.450)	7.756 (.810)	44.960 (12.64)
MIC	0.628 (.008)	1.357 (.035)	5.584 (.110)	19.052 (.645)	65.170 (2.54)

LPSmoothing (X,Y): The Ripley data



- X = Age and Y = GAG concentration in urine measured for $n = 314$ Children between $0 - 18$ years.
- Question: *What are "normal levels" of GAG in children of each age* ?

LP Copula-Based Nonparametric Regression

- Parametric Copula based Approach: Noh et al. 2013, Nikoloulopoulos and Karlis, 2010.
- OPEN PROBLEM: Remarkable finding by Dette et al. (2013) "commonly used parametric copula families are not flexible enough to produce non-monotone regression function ! ", which makes it unsuitable for any practical use.
- Compact easy to compute representation of LP Copula based Regression, coefficients defined as $\mathbb{E}[YT_j(X;X)] = \mathrm{LP}(j,0;X,Y)$

$$\mathbb{E}\big[Y \mid X = x\big] - \mathbb{E}[Y] = \sum_{j>0} T_j(x;X)\ \mathrm{LP}(j,0;X,Y).$$

- LP nonparametric regression algorithm *bypasses the problem of misspecification bias and can adapt to any nonlinear shape*.

LP Copula-Based Nonparametric Regression

- Parametric Copula based Approach: Noh et al. 2013, Nikoloulopoulos and Karlis, 2010.

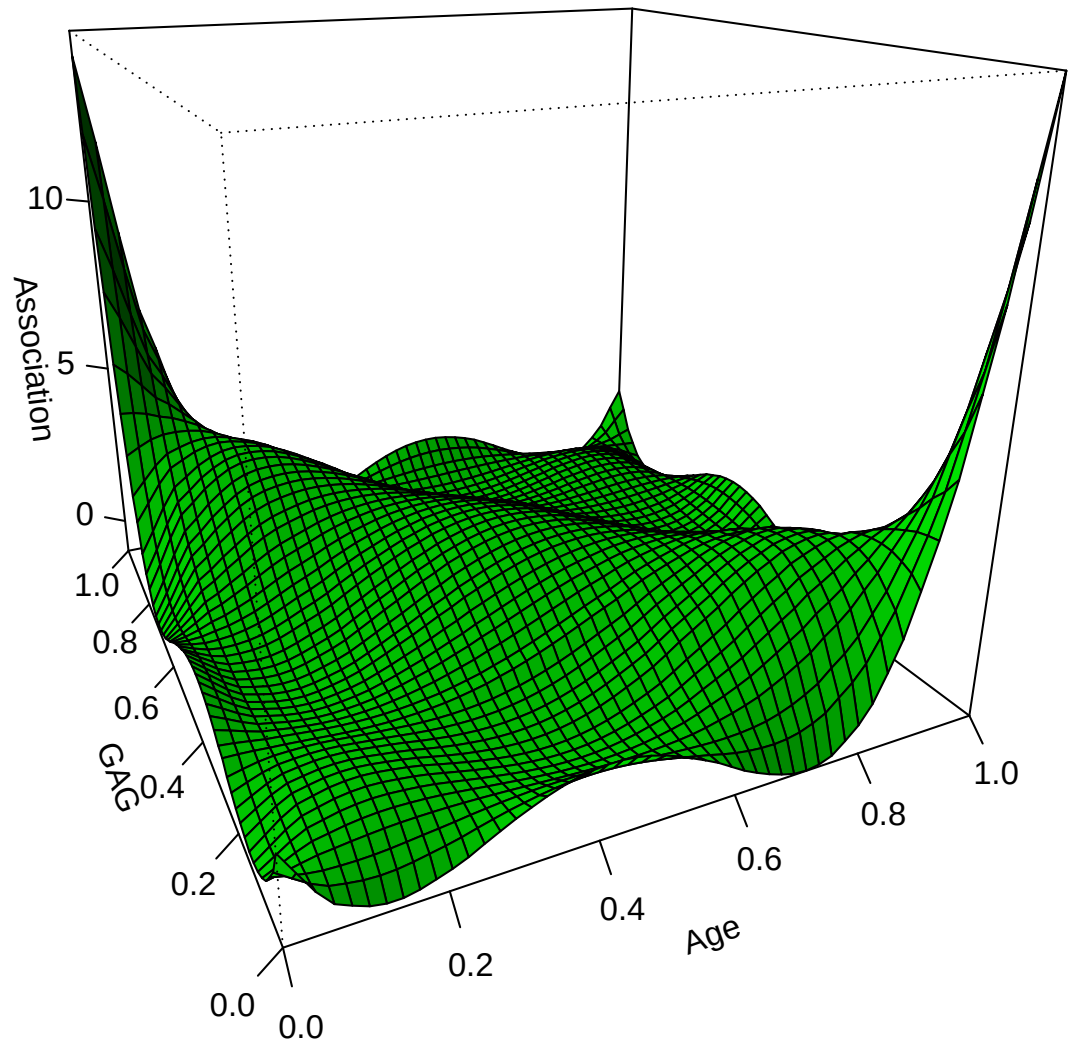
- OPEN PROBLEM: Remarkable finding by Dette et al. (2013) "commonly used parametric copula families are not flexible enough to produce non-monotone regression function ! ", which makes it unsuitable for any practical use.

- Compact easy to compute representation of LP Copula based Regression, coefficients defined as $\mathbb{E}[YT_j(X;X)] = \mathrm{LP}(j,0;X,Y)$

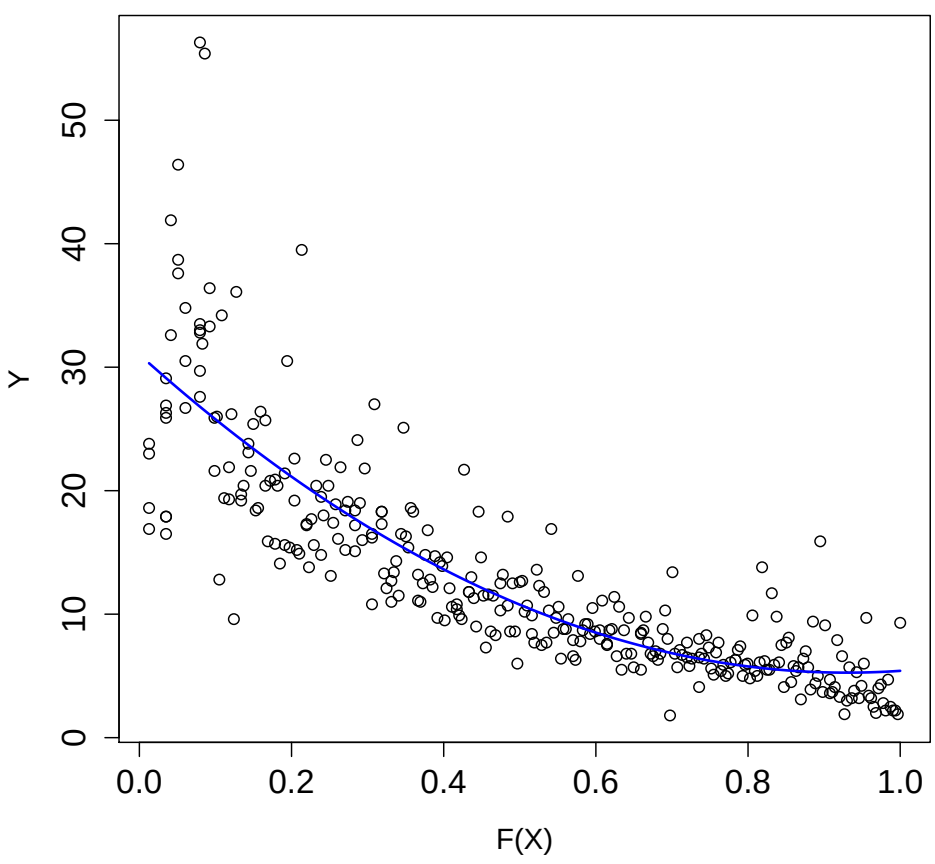
$$\mathbb{E}\big[Y \mid X = x\big] - \mathbb{E}[Y] = \sum_{j>0} T_j(x;X)\ \mathrm{LP}(j,0;X,Y).$$

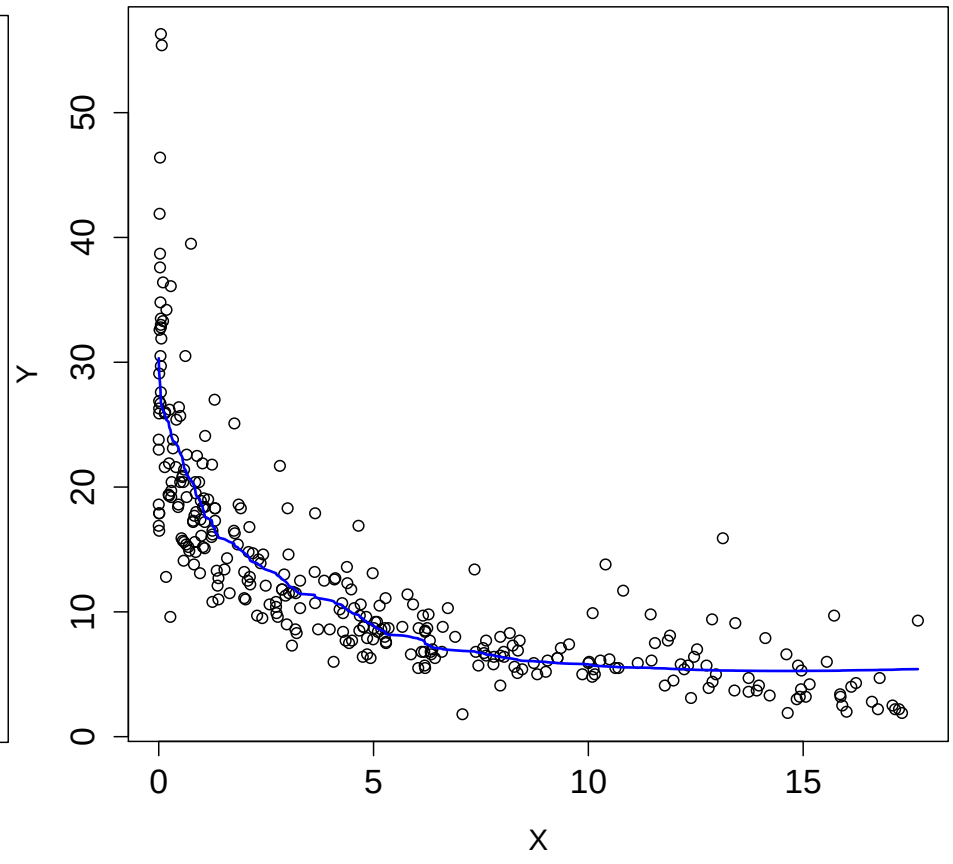
- LP nonparametric regression algorithm *bypasses the problem of misspecification bias and can adapt to any nonlinear shape*.

LP Smooth Nonparametric Copula Density



LP Copula Based Nonparametric Regression





$$\widehat{\mathbb{E}}[Y|X=x] \;=\; 13.1 - 7.32\, T_1(x) + 2.20\, T_2(x)$$

Alternative to popular Regression Splines, pioneered by GRACE WAHBA.

Conditional Distribution, Conditional Quantile

- OPEN PROBLEM: Our approach generalizes the seminal work by Koenker and Bassett (1978) on parametric (linear) quantile regression to nonparametric nonlinear setup, which guaranteed to produce non-crossing quantile curves, a longstanding practical problem.

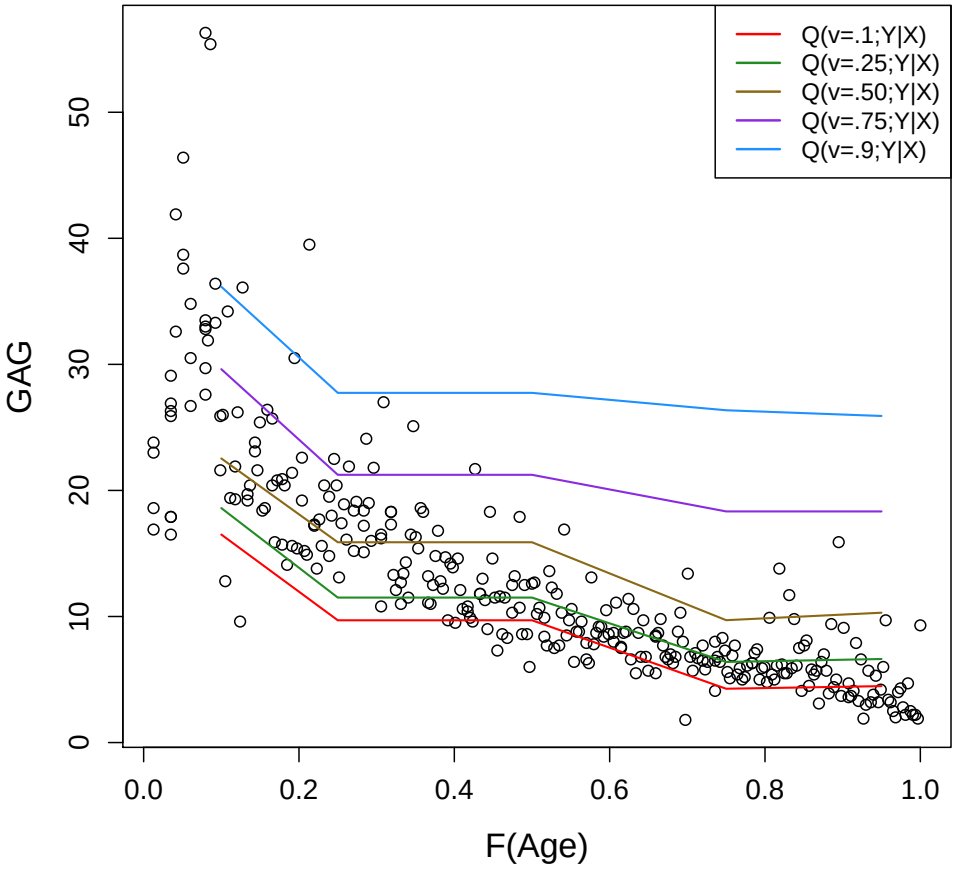
- Nonparametric estimation of conditional density $f(y; Y|X = x)$

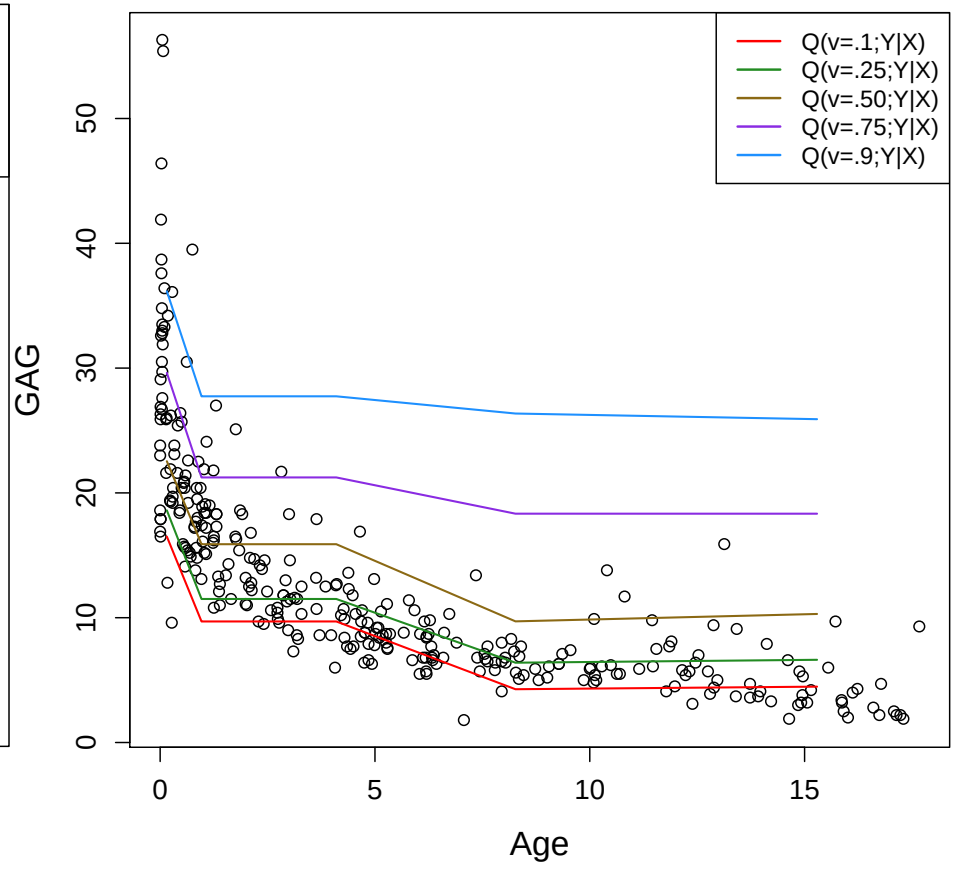
$$f(y; Y|X = x) = f(y; Y)\,\mathrm{cop}(F(x), F(y)) = f(y; Y)\,d\big(F(y); Y, Y|X = x\big).$$

- LP copula estimation simultaneously estimate all the copula slices or $d(v; Y|X = Q(u; X))$ for various u.

- Conditional quantiles $Q(v; Y, Y|X = Q(u; X))$ can be simulated by accept-reject sampling from conditional comparison density.

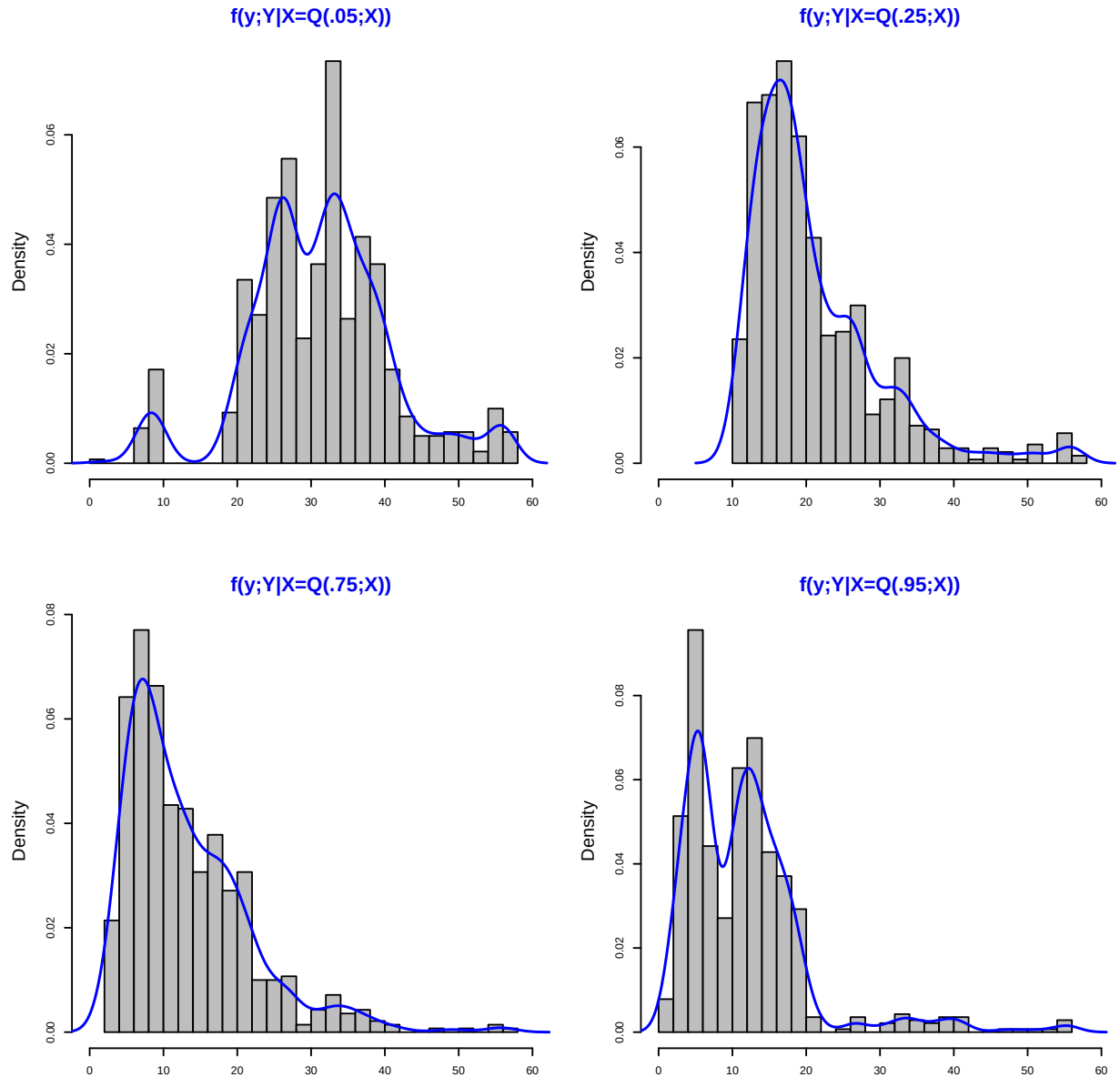
LP Non-crossing Quantile Regression





Answer to The Scientific Question: Conditional Quantile Bands.

For More Insight: Conditional Densities



LP-Coherence Matrix: Multivariate Nonlinear Association

1. SET UP, Random Vectors: $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$.

2. Define: $\text{COH}(X, Y) \;=\; \Sigma_{XX}^{-1}\, \Sigma_{XY}\, \Sigma_{YY}^{-1}\, \Sigma_{YX}$.

3. LP-Coherence Dependence Measure: $\text{Trace}\big[\, \text{COH}(TX,\, TY) \,\big]$.

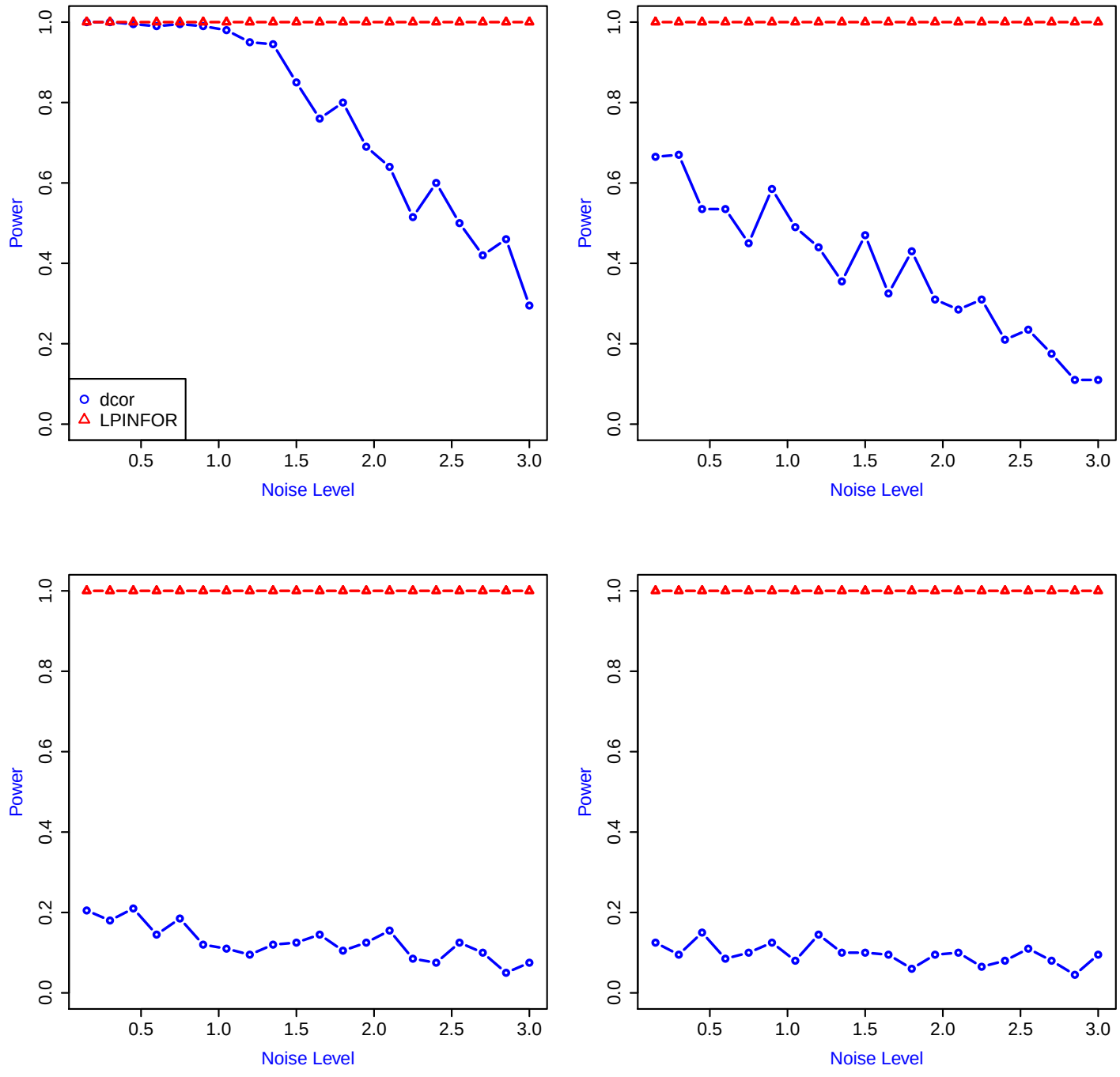
4. Unifying Univariate and Multivariate. For X, Y Univariate Verify that:

$$\text{LPINFOR}(X, Y) \;=\; \text{Trace}\big[\, \text{COH}(TX,\, TY) \,\big].$$

5. Works for Mixed Multivariate Random Variables. Nonlinear, Robust.

6. Székely (2007) Example: Fixed $n = 200,\; p = q = \big\{5, 10, 25, 50\big\}$.

$$Y_k \;=\; \log(X_k^2) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2), \;\; \sigma \uparrow 3.$$


Power
Noise Level
dcor
LPINFOR
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0

The BIG Idea: LP STATISTICAL DATA SCIENCE

- Mukhopadhyay, S. and Parzen, E. (2014) LP Approach to Statistical Modeling, *Preprint*.
- Mukhopadhyay, S. (2013), Nonparametric Inference for High Dimensional Data, Ph.D. Thesis, Texas A&M University, College Station, TX.
- Parzen, E. and Mukhopadhyay, S. (2013) LP Mixed Data Science : Outline of Theory, *arXiv:1311.0562*.
- Parzen, E. and Mukhopadhyay, S. (2013) United Statistical Algorithm, Small and Big Data: Future of Statistician, *arXiv:1308.0641*.
- Parzen, E. and Mukhopadhyay, S. (2012) Modeling, Dependence, Classification, United Statistical Science, Many Cultures, *arXiv:1204.4699*.

Long-Term Research Goals

Currently we are investigating how LP Theory can be adapted to

- (Nonlinear Non-Gaussian) Time Series Modeling; [Winner 2014 IEEE International Biometric Identification Data Challenge]
- Nonparametric Functional Statistical Approach to Network Modeling; [Preliminary results will be presented in ISNPS 2014, Spain]
- Conditional Independence and Pearl's Causal Inference;

$\vdots \quad \vdots \quad \vdots$

We plan to approach these important statistical problems one at a time

GOAL: *Comprehensively Unify them using LP Theory and Tools*.

——— Thanks.