LP Nonparametric Modeling Exploratory Big Data Analysis

Emanuel Parzen[1] **and Deep Mukhopadhyay**[2]

[1]Texas A&M University, College Station, TX, USA

[2]Temple University, Philadelphia, PA, USA

eparzen@stat.tamu.edu, deep@temple.edu

CONFERENCE TO HONOR GRACE WAHBA

I am proud to have been the PhD thesis advisor of world famous Professor GRACE WAHBA I have been her friend for 50 years, and an admirer of her originality and world class accomplishments in statistical theory and applications. She is acclaimed by many honors: member of the National Academy of Sciences, mentor of many leading statisticians (also members Natonal Academy of Sciences), great as scholar, researcher, human relationships, collaborator in many scientific fields described in her COPSS autobiographical essay about RKHS and Ah Hah experiences. Her research is the foundation of the interface between statistics and machine learning.

Abstract

We present an approach to statistical data modeling and exploratory data analysis called 'LP Statistical Data Science.' It aims to generalize and unify traditional and novel statistical measures, methods, and exploratory tools. This article outlines fundamental concepts along with real-data examples to illustrate how the 'LP Statistical Algorithm' can systematically tackle different varieties of data *types*, data *patterns*, and data *structures* under a coherent theoretical framework. A fundamental role is played by specially designed orthonormal basis of a random variable X for linear (Hilbert space theory) representation of a general function of X, such as $\mathbb{E}[Y \mid X]$.

PAST & FUTURE Statistical Science

What is Statistics ? Answers & Questions

"What Do You Know ? How Do You Know It ?"

Statisticians in the 21st century (if they take advantage of their opportunities) can look forward to very bright futures for the discipline and profession of statistics by solving the important problem of teaching thousands statistical data scientist aspirants, what are the fundamental methods of statistical learning. We are developing a framework that unifies parametric and nonparametric models, frequentist and Bayesian inference, small data and big data, beginning courses that differ from traditional courses by teaching methods for simple data in ways that extend to complex high dimensional data (similar to the goal of teaching finite dimensional math in notation that extend to Hilbert space).

Wholesale and Retail Problem Solving

Retail: Solving Real Data Scientific problems one at a time for clients.

Wholesale: Theory and Solutions Applicable for Many clients.

To accomplish above goals a theory, called LP STATISTICAL SCIENCE, is being developed by me and Subhadeep(Deep) Mukhopadhyay, faculty member at Temple University Fox School of Business (2013 Ph.D). The theory is unifying because it applies to mixed data (variables which are discrete or continuous) and to population and sample distributions. It practices "plug in estimation" computed for sample distributions by the same definitions used for population distributions. It shows fundamental role of $\mathrm{Leg}_j(u)$, $0 < u < 1$, Orthonormal Legendre Polynomials of u, basis of $L^2(0,1)$.

Quantiles and Mid-Distributions

- Random variable X, distribution $F(x) = F(x; X)$.
- Quantile $Q(u) = Q(u; X)$, $0 < u < 1$.
- Mid-distribution $F^{\mathrm{mid}}(x; X) = F(x; X) - .5p(x; X)$
- $\mathbb{E}[F^{\mathrm{mid}}(X; X)] = .5, \mathrm{Var}[F^{\mathrm{mid}}(X; X)] = (1/12)(1 - \sum_x p^3(x; X))$
- $F(Q(u)) = u$, at $u = F(x)$ for some x.
- $Q(F(X)) = X$, probability 1.
- $\mathbb{E}[Y|X] = \mathbb{E}[Y|F(X)]$, probability 1.
- $\mathbb{E}[Y|X] = \sum_j T_j(X : X)\,\mathbb{E}[YT_j(X; X)]$, $T_j(x; X)$ orthonormal $L^2(F)$
- $X \;=\; \sum_j T_j(X; X)\;\mathbb{E}[XT_j(X; X)]$.

LP INSIGHT: X Can be represented Linear Function of $T_j(X; X)$, that are orthonormal polynomials of $F(X; X)$.

Construction of Score Functions

- SCORE Function $T_j(x; X)$
- UNIT Score Function $S_j(u; X) = T_j(Q(u; X); X)$
- X Continuous construct $T_j(x : X) = \mathrm{Leg}_j(F(x; X))$
- X Discrete $T_j(X; X)$ Gram Schmidt orthonormalization powers of $T_1(X; X)$
- $T_1(X; X) = \mathcal{Z}(F^{\mathrm{mid}}(X; X))$ where $\mathcal{Z}(X) = (X - \mathbb{E}(X))/\sigma(X)$
- $\mathbb{E}[Y|X = Q(u; X)] = \sum_j S_j(u; X)\ \mathrm{LP}(j, 0; X, Y)$
- $\mathrm{LP}(j, 0; X, Y) \ = \ \mathbb{E}[Y\ T_j(X; X)]$

LP Moments, LP Comoments, Linear Nonparametric Models

- $\mathrm{LP}(j;X) = \mathrm{LP}(j,0;X,X) = \mathbb{E}[XT_j(X;X)]$
- $\mathrm{LP}(j,k;X,Y) = \mathbb{E}[T_j(X;X)T_k(Y;Y)]$
- Spearman Correlation: $\mathrm{LP}(1,1;X,Y) = R(F^{\mathrm{mid}}(X;X), F^{\mathrm{mid}}(Y;Y)]$
- Gini Correlation: $\mathrm{LP}(1,0;X,\mathcal{Z}(Y)), \mathrm{LP}(0,1;\mathcal{Z}(X),Y)$
- UNITY INTRO STAT Methods: Gini, Spearman, LPINFOR (Chi-square, AOV)
- LPEDA: Diagnose Non-normality $\mathrm{LP}(j;X)$, and Non-stationarity $\mathrm{LP}(0,j;X,\mathrm{index}\,X)$.
- Deserve to be: FUNDAMENTAL methods of Statistical Learning !

To Question "What is Statistics ?"

Answer "What is Statistical Learning ?"

- Variance Decomposition: $\text{Var}[X] = \sum_j |\,\text{LP}(j;X)|^2$
- $\text{Cov}(X,Y) \;=\; \sum_{j,k>0} \text{LP}(j;X)\,\text{LP}(k;Y)\,\text{LP}(j,k;X,Y)$
- $X \sim \mathcal{N}(\mu,1)$: $\text{LP}(1;X) = 2\sqrt{3}\;\mathbb{E}[\phi(X);X] \;=\; \sqrt{3/\pi}$
- Tail Index: X smallest m, $\sum_1^m \Big|\,\text{LP}[j;\mathcal{Z}(X)]\Big|^2 > .95.$
- $\text{Var}[\mathbb{E}[Y|X]] = \sum_j |\,\text{LP}(j,0;X,Y)|^2$
- Dependence (X,Y) $\text{LPINFOR}(X,Y) = \sum \Big|\,\text{LP}(j,k;X,Y)\Big|^2.$
- Nonparametric Model of Many Variables X_i by Multivariate Analysis of Score Vectors $TX_i = \big\{\,T_1(X_i;X_i),\ldots,T_m(X_i;X_i)\big\}$

Comparison Density, LP Skew Model, Significant Scores

- LP SKEW Model X Continuous $f(x;X) = g(x)\,d(G(x))$
- LP SKEW Model X Discrete $p(x;X) = g(x)\,d(G(x))$
- Comparison Density X continuous G continuous

$$d(u) = f(Q(u;G);X)/g(Q(u;G)) = d(u;G,F(.;X))$$

- Comparison Distribution $D(u) = F(Q(u;G);X)$
- Formula ("E=mc^2 quality, Ah Hah Wahba"): GOF components

$$\mathrm{LP}(j;G,F) = \int T_j(x;G)\,\mathrm{d}F(x) = \int S_j(u;G)\,\mathrm{d}D(u;G,F)$$

- Observe sample distribution $\widetilde{F}(x;X)$
- EMPIRICAL PROCESS Theorem: When model $G(x) = F(x;X)$

$$\sqrt{n}\,\mathrm{LP}(j;G,\widetilde{F}) \xrightarrow{d} \mathcal{N}(0,1)$$

L2, Max Entropy, Bayesian Estimation of Comparison Density

- Select Significant Components $\mathrm{LP}(j; G, \widetilde{F})$ *Data-Driven* AIC Ledwina.
- Significant Component Scores $S_j(u; G)$ are Sufficient Statistics for $d(u)$ in LP SKEW model for true $f(x; X)$ or $p(x; X)$
- L^2 Comparison Density $d(u) - 1 = \sum_j \mathrm{LP}(j; G, \widetilde{F})\, S_j(u; G)$
- MaxEnt Comparison Density $\log d(u) = \sum_j \theta_j S_j(u) - K(\theta)$
- θ_j Canonical Parameters, TRUE $\mathrm{LP}(j; G, F)$ Mean Parameters.
- Bayesian Estimator of Exponential $d(u)$: Conjugate Prior on $\mathrm{LP}[j, G, F]$, update the mean parameter $\mathrm{LP}(j; G, \widetilde{F})$.
- Smooth χ^2 Statistic is sum of squares of SIGNIFICANT components

$$\mathrm{LP}(j; G, \widetilde{F}) = \sum_x T_j(x; G)\, \widetilde{p}(x; X)$$

- Achieve Advice to Look at the Data, DON'T just compute RAW Chi-square, Smooth Chi Square *has reduced degrees of freedom*.

Modeling (X,Y), Dependence, Copula Density

- Define Copula Distribution at $u = F(x; X), v = F(y; Y)$ for some x, y

$$\text{Cop}(u, v; X, Y) = F(Q(u; X), Q(v; Y); X, Y)$$

- Define Copula Density valid for Mixed (X, Y):

$$\text{cop}(u, v; X, Y) = d(v; Y, Y|X = Q(u; X)) = d(u; X, X|Y = Q(v; Y))$$

- FUNDAMENTAL LP Representation Theorem

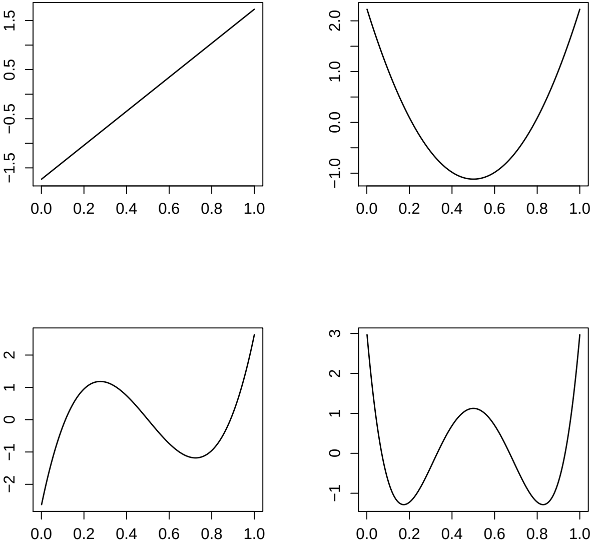
$$\text{cop}(u, v; X, Y) - 1 = \textstyle\sum_j \text{LP}(j, k; X, Y)\, S_j(u; X)\, S_k(v; Y)$$

- Piecewise constant Checkerboard plot of $\text{cop}(u, v; X, Y)$
- Select Significant Comoments $\text{LP}(j, k; X, Y)$, Data-Driven
- LP SKEW Conditional Density at $x = Q(u; X), y = Q(y; Y)$

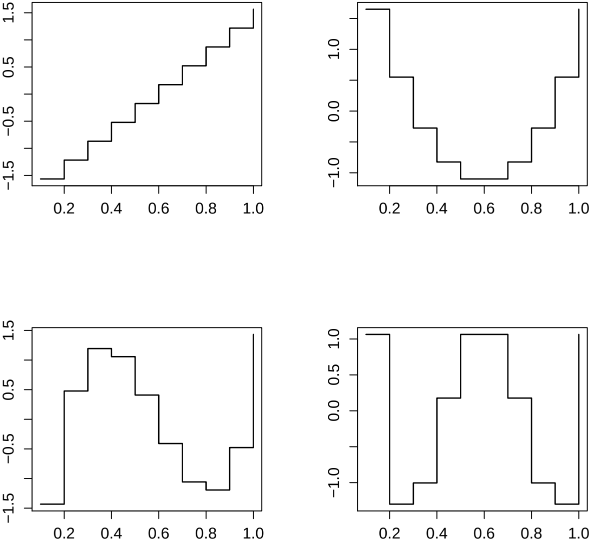
$$f(y; Y|X = x) \ = \ f(y; Y)\, \text{cop}(u, v; X, Y)$$

- Insight: Plotting “Slices” of $\text{cop}(u, v; X, Y)$, for Selected u.

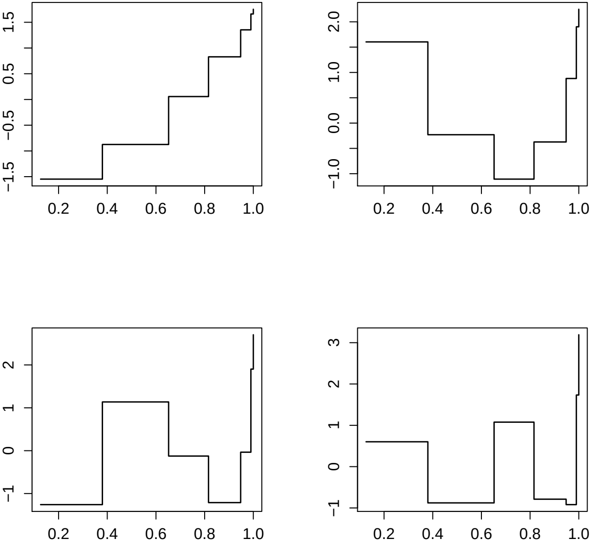
$X \;\leftarrow\; \texttt{rnorm}(n = 200,\ \mu = 0,\ \sigma = 1);\quad \texttt{LP.poly}(X, m = 4)$

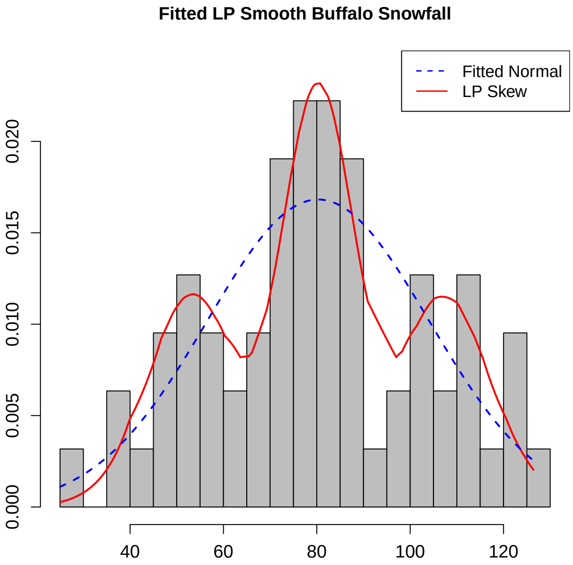


$X \ \leftarrow \ \texttt{Discrete.Uniform}\{1,2,\ldots,10\}; \ \ \texttt{LP.poly}(X, m=4)$



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





$$\widehat{f}(x) = g(x)\left[1 - .34\, S_6(u)\right], \quad g(x) = \mathcal{N}(80.29, 23.72^2)$$

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

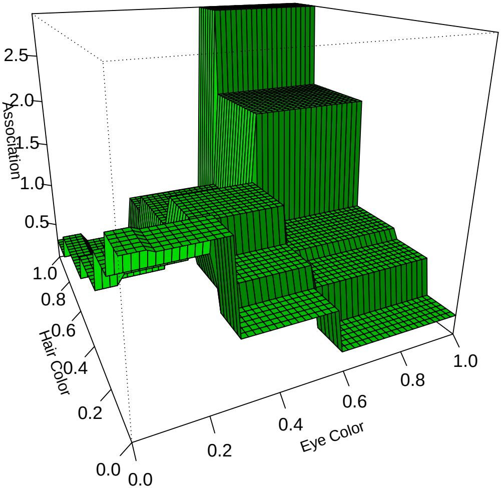
Table: Two way contingency table classifying 5387 children of Scotland.

LP-Comoments, LPINFOR, Degrees of freedom

$$\widehat{\mathrm{LP}}\big[X, Y\big] \;=\; \begin{bmatrix} 0.423^* & 0.024 & 0.039 & -0.009 \\ 0.115^* & 0.157^* & 0.001 & -0.021 \\ -0.050 & 0.085 & 0.017 & -0.032 \end{bmatrix}$$

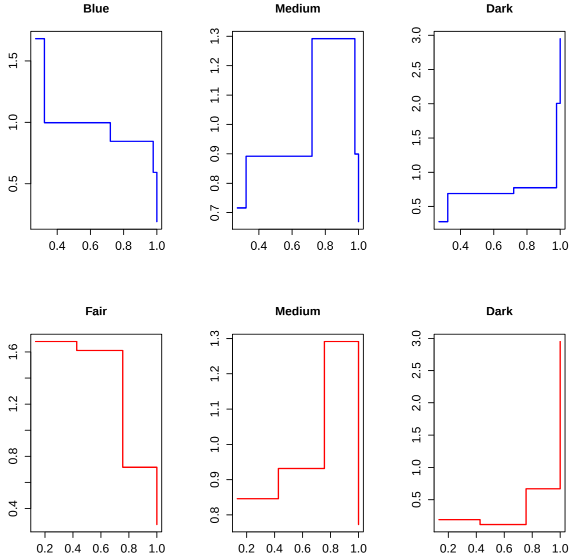
- Raw/Saturated ChisqDiv $= \chi^2/n =$ **.230**,
- df $(4-1)\times(5-1) = 12$
- "Smooth"-LPINFOR = **.220** with effective df 3.
- df drops dramatically from 12 to 3 ! Boost power.

LP Nonparametric Checkerboard Copula Estimate



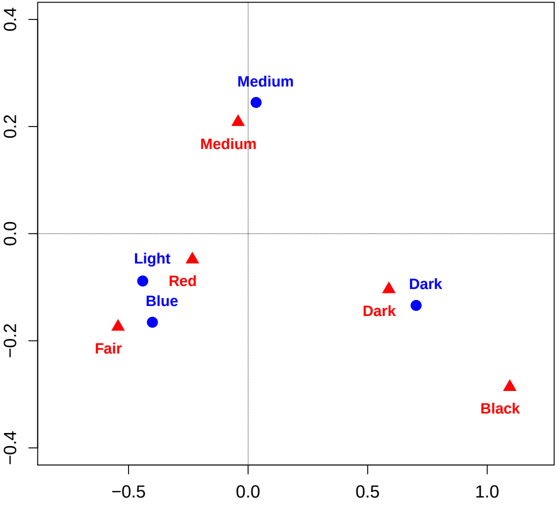
$$\widehat{\mathrm{cop}}(u,v;X,Y) \,=\, 1 + .42 S_1(u)\, S_1(v) + .12 S_2(u)\, S_1(v) + .16 S_2(u)\, S_2(v)$$

Slices of Copula Densities



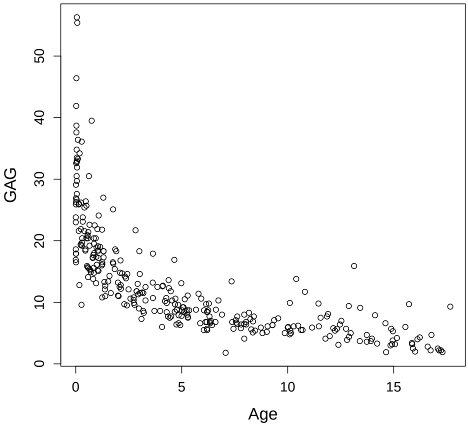
“Blue”: Row categories, “Red”: Column categories.

LP Correspondence Map



Compare "shapes" of copula density "slices" for Row and Col categories.

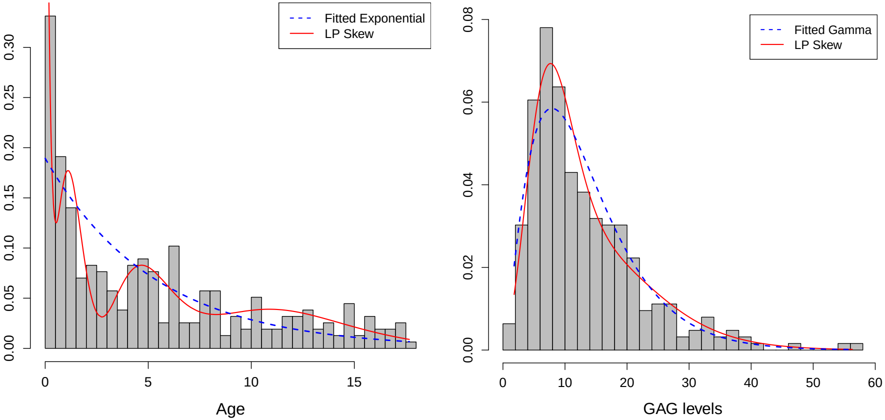
The Gag Urine Problem



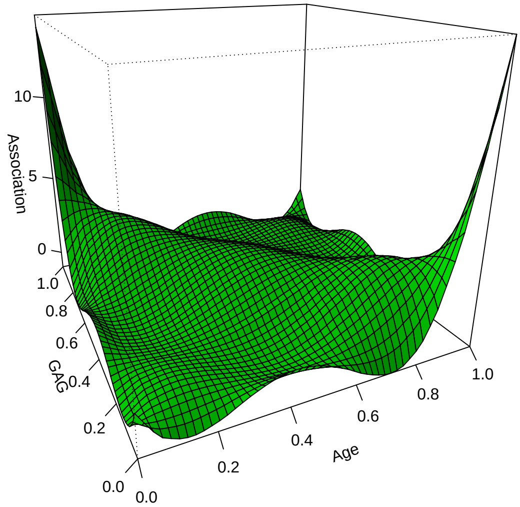
- X = Age and Y = GAG concentration in urine measured for $n = 314$ Children between $0 - 18$ years.

- Question: *What are "normal levels" of GAG in children of each age* ?

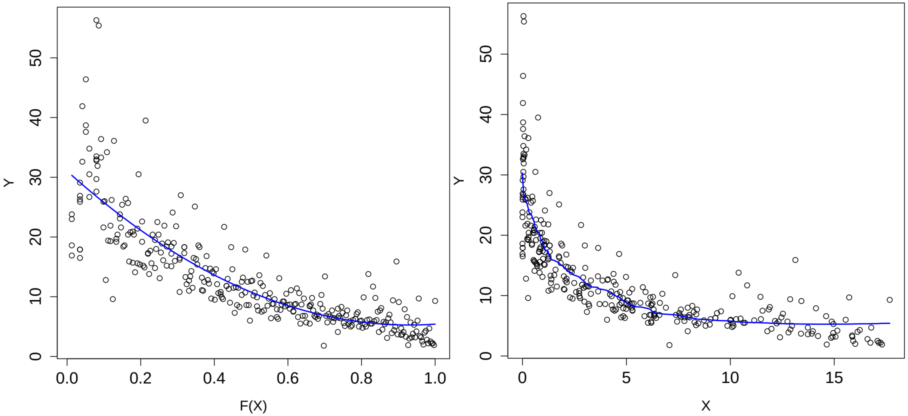
LP Smooth Marginal Densities



LP Smooth Nonparametric Copula Density



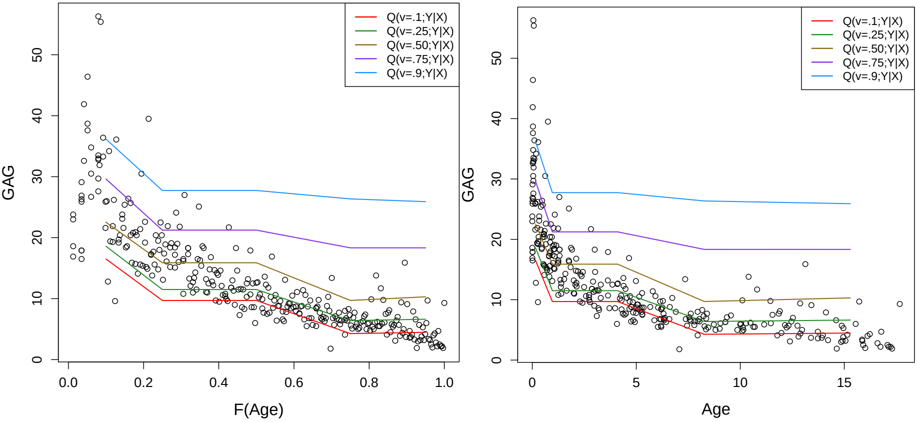
LP Copula Based Nonparametric Regression



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

Alternative to popular Regression Splines, pioneered by GRACE WAHBA.

LP Non-crossing Quantile Regression



Answer to The Scientific Question: Conditional Quantile Bands.

Other Ongoing Successful Applications

1. LPTime: Non-stationary, Non-Gaussian Time series Modeling. [Winner IEEE EMVIC 2014 Competition]

2. LPImage: Image Compression, Sparse Coding. [Remarkable performance in MNIST data]

3. LPNetwork: Network Modeling, Community detection. [Scalable for massive network, will be presented in ISNPS 2014]

$\vdots \qquad\qquad \vdots$