Correspondence Analysis Using Neural Networks

Hsiang Hsu[1], Salman Salamatian[2] and Flavio P. Calmon[1]

[1]School of Engineering and Applied Science, Harvard University

[2]Research Laboratory of Electronics , Massachusetts Institute of Technology

Background

- Extracting correlation/ correspondence in data is at the core of machine learning and data science
- Traditional machine learning:
  - PCA/ Canonical Correlation Analysis (CCA): restricted to single variable, hard to find non-linear relation
  - Kernel methods (e.g. kernel PCA, kernel CCA): restricted to the pre-selected kernel family
- Modern machine learning: Variational Auto-Encoder, non-linear embedding, etc.
  - Suffering from entanglement between representations
  - Hard to visualize, interpret
- Let's revisit an exploratory multivariate statistical tool: Correspondence Analysis

Which functions of a hidden variable can be estimated with small mean-squared error?

$$\min_f \quad \mathrm{mmse}(f(X)|Y) = 1-\lambda_i$$
$$\text{s.t.} \quad \mathbb{E}[f(X)] = 0,\quad \|f(X)\|_2 = 1,\quad \mathbb{E}[f(X)f_1(X)] = 0,\ \vdots,\ \mathbb{E}[f(X)f_{i-1}(X)] = 0$$

Maximizing function: $f_i(X)$

Correspondence Analysis (CA)

- Similar to PCA, Canonical Correlation Analysis (CCA), but
  - Produces low-dimensional representation of data that captures non-linear relationships
  - Enables visualization and interpretability
- Widely used in Genealogy, Epidemiology, Social and Environmental Sciences (see Selected Reference)
- Consider two random variables X and Y with their joint probability $p_{X,Y}$ and supports $\mathcal{X} = [n]$, $\mathcal{Y} = [m]$

Contingency Table

$$Q = \mathsf{D}_X^{-1/2}(\mathsf{P}_{X,Y} - \mathsf{P}_X\mathsf{P}_Y^T)\mathsf{D}_Y^{-1/2} = \begin{bmatrix} \frac{p_{X,Y}(1,1)-p_X(1)p_Y(1)}{\sqrt{p_X(1)p_Y(1)}} & \cdots & \frac{p_{X,Y}(1,m)-p_X(1)p_Y(m)}{\sqrt{p_X(1)p_Y(m)}} \\ \vdots & \ddots & \vdots \\ \frac{p_{X,Y}(1n,1)-p_X(n)p_Y(1)}{\sqrt{p_X(n)p_Y(1)}} & \cdots & \frac{p_{X,Y}(n,m)-p_X(n)p_Y(m)}{\sqrt{p_X(n)p_Y(m)}} \end{bmatrix}$$

Diagonal matrices with marginals as entries

- Singular Value Decomposition: σ_i: singular values

$$Q = \mathsf{D}_X^{-1/2}(\mathsf{P}_{X,Y} - \mathsf{P}_X\mathsf{P}_Y^T)\mathsf{D}_Y^{-1/2} = \mathsf{U}\mathsf{Z}\mathsf{V}^T$$

Orthogonal factors of X: $\mathsf{L} \triangleq \mathsf{D}_X^{-1/2}\mathsf{U}$ Orthogonal factors of Y: $\mathsf{R} \triangleq \mathsf{D}_Y^{-1/2}\mathsf{V}$

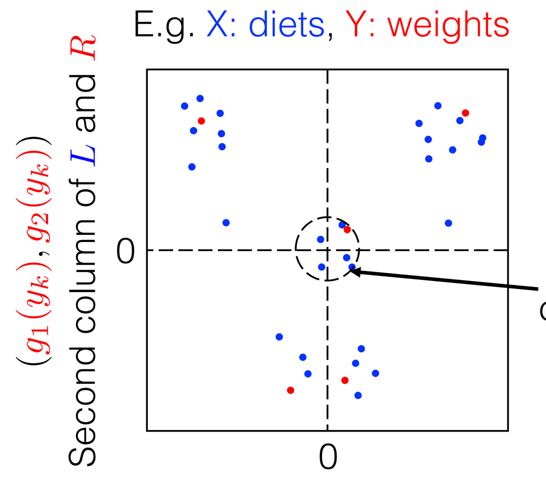
$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & & \sqrt{\lambda_d} \end{bmatrix}$$

Factor scores: $\lambda_i = \sigma_i^2$

Factor score ratios: $\frac{\lambda_i}{\sum_i \lambda_i}$

First Factoring Plane

E.g. X: diets, Y: weights



- Limitations of SVD-based Correspondence Analysis
  1. Limited to categorical (discrete) data with finite support
  2. Reliably estimation of the contingency table (approximation of $p_{X,Y}$) may be infeasible due to limited number of samples
  3. Not scalable for high-dimensional data

Principal Inertia Components (PICs)

- Starting from Maximal Correlation

$$\max_{f,g} \quad \mathbb{E}[f(X)g(Y)]$$
$$\text{subject to} \quad \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0,\quad \|f(X)\|_2 = \|g(Y)\|_2 = 1$$

If f and g are linear functions: CCA

Maximal Correlation $= \rho_m(X;Y) = \sqrt{\lambda_1} \quad = \sqrt{\lambda_2} \quad = \sqrt{\lambda_3}$

Maximizing functions: $f_1(X)$ and $g_1(Y)$

$\mathbb{E}[f(X)f_1(X)] = \mathbb{E}[g(Y)g_1(Y)] = 0$ Maximizing functions: $f_2(X)$ and $g_2(Y)$

$\mathbb{E}[f(X)f_2(X)] = \mathbb{E}[g(Y)g_2(Y)] = 0$ Maximizing functions: $f_3(X)$ and $g_3(Y)$

$\cdots$ And so forth.

$f_1, f_2, \cdots$ and $g_1, g_2, \cdots$ are Principal Functions $\quad \lambda_1, \lambda_2, \cdots$ are Principal Inertia Components

- The principal functions are in the Hilbert space of finite-variance functions
- The principal functions are low-dimensional orthogonal (disentangled) representations of data
- Reconstitution Formula: Decomposition of a joint distribution. Let $d = \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$

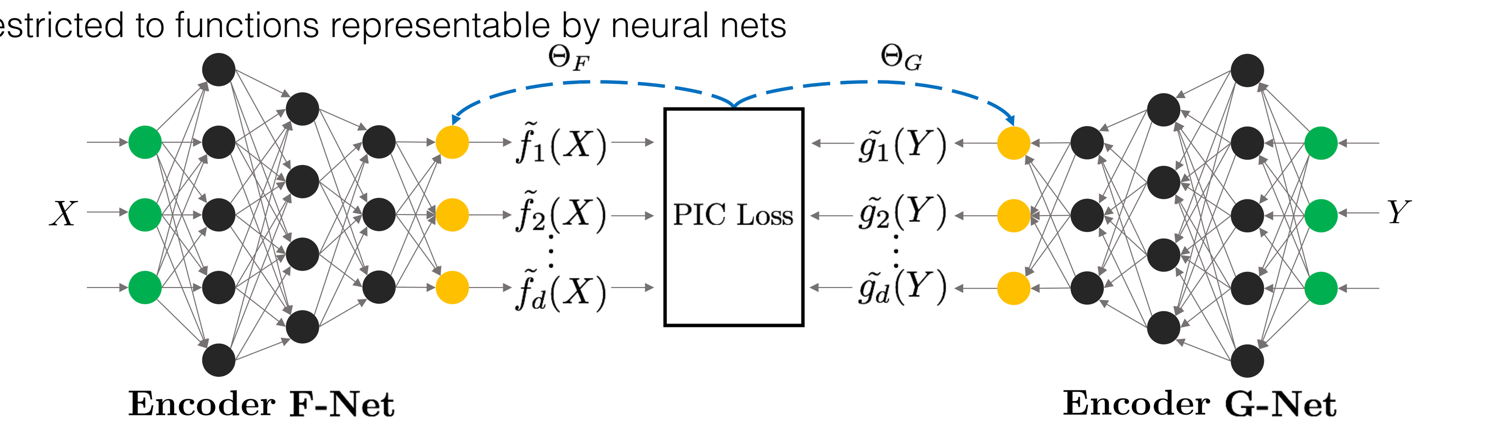
$$p_{X,Y}(x,y) = p_X(x)p_Y(y)\left(1 + \sum_{l=1}^{d} \sqrt{\lambda_i} f_i(x) g_i(y)\right)$$

- PICs for Discrete Distributions: Proposition 2

Principal Functions $f_1, f_2, \cdots$	Principal Functions $g_1, g_2, \cdots$	PICs $\lambda_1, \lambda_2, \cdots$
||	||	||
Orthogonal factors of X $\mathsf{L} \triangleq \mathsf{D}_X^{-1/2}\mathsf{U}$	Orthogonal factors of Y $\mathsf{R} \triangleq \mathsf{D}_Y^{-1/2}\mathsf{V}$	Factor scores: $\lambda_i = \sigma_i^2$

Correspondence Analysis Neural Net (CA-NN)

- Searching the whole Hilbert space for f and g is infeasible
- Restricted to functions representable by neural nets



Proposition 3: The PIC Loss is given by

$$\min_{\tilde{\mathsf{f}},\tilde{\mathsf{g}}} -2\|\mathsf{C}_f^{-\frac{1}{2}}\mathsf{C}_{fg}\|_d + \mathbb{E}[\|\tilde{\mathsf{g}}(Y)\|_2^2],$$

where $\mathsf{C}_f = \mathbb{E}[\tilde{\mathsf{f}}(X)\tilde{\mathsf{f}}(X)^\intercal]$, $\mathsf{C}_{fg} = \mathbb{E}[\tilde{\mathsf{f}}(X)\tilde{\mathsf{g}}(Y)^\intercal]$, and $\|\mathsf{Z}\|_d$ is the d-th Ky-Fan norm, defined as the sum of the singular values of Z. Denoting by A and B the whitening matrices for $\tilde{\mathsf{f}}(X)$ and $\tilde{\mathsf{g}}(X)$, the principal functions are given by $\mathsf{f}(X) = [f_0(X), \cdots, f_d(X)]^\intercal = \mathsf{A}\tilde{\mathsf{f}}(X)$ and $\mathsf{g}(Y) = [g_0(Y), \cdots, g_d(Y)]^\intercal = \mathsf{B}\tilde{\mathsf{g}}(Y)$.

Synthetic Data

- Analytical solutions of principal functions are in general hard
- Discrete case (Binary Symmetric Channel):

$X \sim Bernoulli(p)$, $Z \sim Bernoulli(\delta)$, $Y = X \oplus Z$

- Continuous case (Gaussian):

X, Y jointly Gaussian

Principal Functions: Hermite polynomials

$$H_l(x) \triangleq (-1)^l e^{\frac{x^2}{2}} \frac{d^l}{dx^l} e^{-\frac{x^2}{2}}.$$

Discrete PICs				
CA-NN	0.8011	0.7942	0.7918	0.7883
Analytic value	0.8000	0.8000	0.8000	0.8000
Gaussian PICs				
CA-NN	0.7007	0.4938	0.3376	0.2037
Analytic value	0.6977	0.4675	0.2979	0.2113

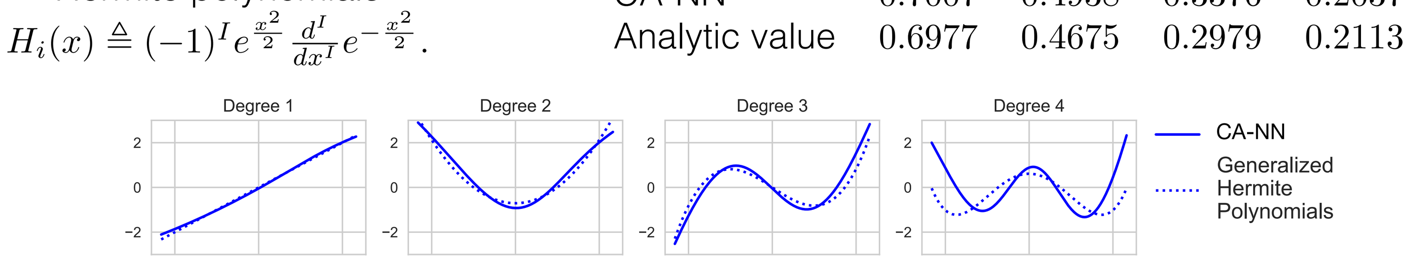
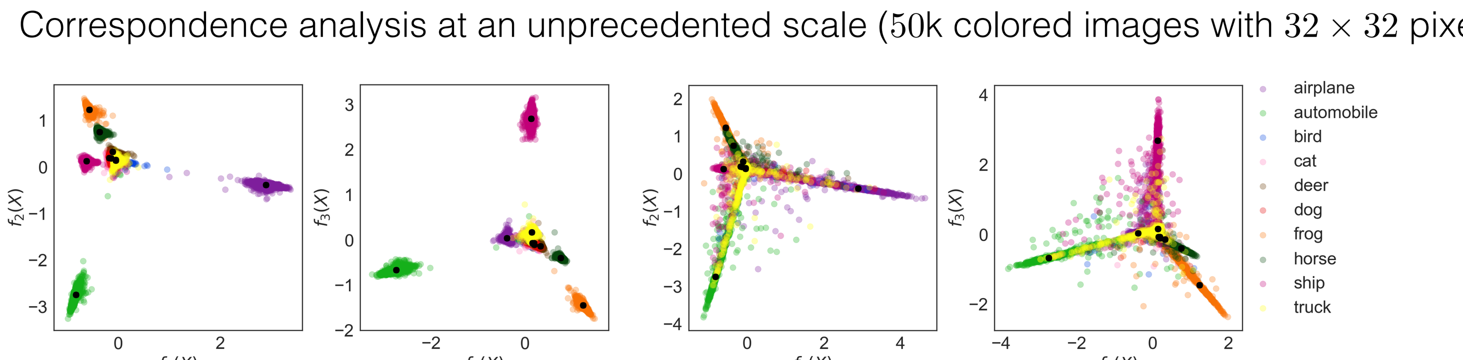


Image Dataset – CIFAR-10

- Correspondence analysis at an unprecedented scale (50k colored images with 32×32 pixels)



(a) Training set (b) Test set

Selected Reference

- Greenacre, M. (2017). Correspondence analysis in practice. Chapman and Hall/CRC.
- ter Braak, C. J. et al. (2004). Co-correspondence analysis: a new ordination method to relate two community compositions. Ecology, 85(3):834–846.
- Ferrari, A. et al. (2016). A whole-genome sequence and transcriptome perspective on her2-positive breast cancers. Nature communications, 7:12222.
- Sourial, N. et al. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. Journal of clinical epidemiology, 63(6):638–646.
- Rényi, A. (1959). On measures of dependence. Acta mathematica hungarica, 10(3-4):441–451.
- Witsenhausen, H. S. (1975). On sequences of pairs of dependent random variables. SIAM Journal on Applied Mathematics, 28(1):100–113.
- Andrew, G., et al. (2013). Deep canonical correlation analysis. In International Conference on Machine Learning, pages 1247–1255.
- Abbe, E. et al. (2012). A coordinate system for gaussian networks. IEEE Transactions on Information Theory, 58(2):721–733.
- Calmon, F. P. et al. (2017). Principal inertia components and applications. IEEE Transactions on Information Theory, 63(8):5011–5038.
- Hsu, H. et al. (2018). Deep orthogonal representations: Fundamental properties and applications. arXiv preprint arXiv:1806.08449.

Contact

Hsiang Hsu Paper GitHub and Data

Kaggle Food Recipe Dataset

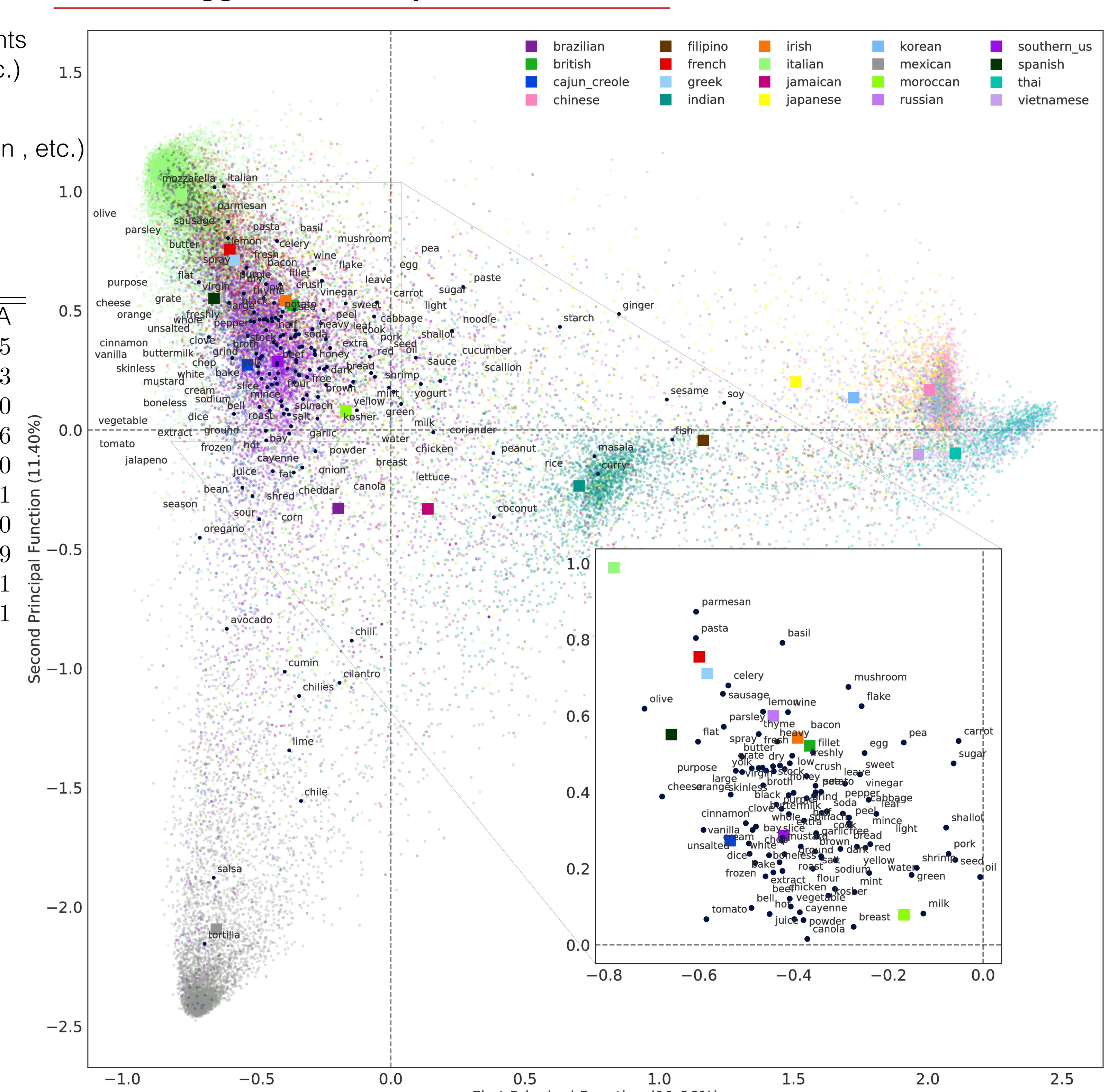
Data Description

- X: 39774 recipes, 6714 ingredients (e.g. peanuts, sesame, beef, etc.)
- Y: 20 types of cuisines (e.g. Japanese, Greek, Jamaican , etc.)

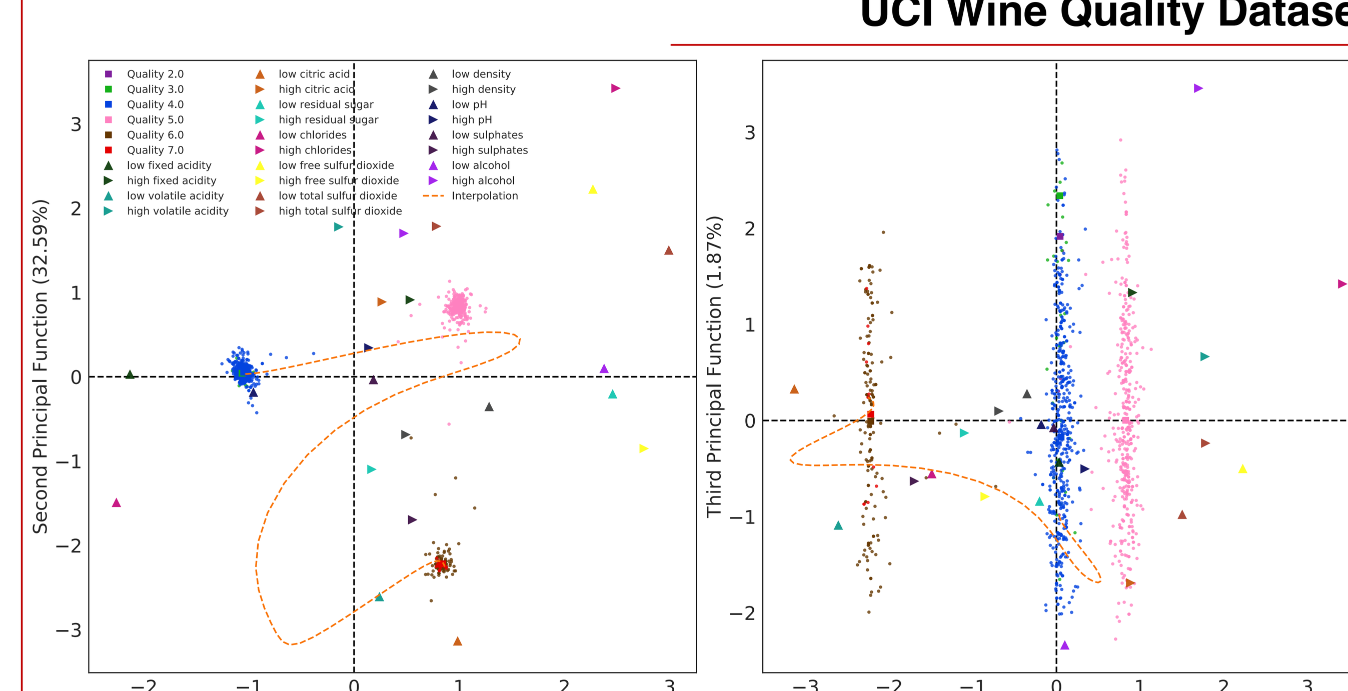
Top ten PICs		Correlations	
CA-NN	SVD	CCA	KCCA
0.9092	0.4504	0.1915	0.6585
0.8667	0.3894	0.1751	0.1223
0.8412	0.3149	0.1342	0.0860
0.7932	0.2943	0.1083	0.0636
0.7391	0.2413	0.1050	0.0320
0.6413	0.1958	0.0823	0.0131
0.6018	0.1547	0.0623	0.0090
0.4792	0.1191	0.0488	0.0089
0.4508	0.1146	0.0485	0.0051
0.2821	0.1035	0.0431	0.0011

Key Observations/ Interpretations

- Clusters: Asian v.s. Western
- First principal function: Asian v.s. Rest
- Second principal function: Mexican v.s. Rest
- Signature ingredients for different cuisines



UCI Wine Quality Dataset



Data Description

- X: 4898 red wines with 11 physico-chemical attributes (e.g. pH value, acid, alcohol, etc.)
- Y: 6 levels of qualities (2-7)
- Continuous-valued attributes: SVD fails

Key Observations/ Interpretations

- 3 sub-clusters of qualities, not 6
- 2 hidden orthogonal factors, not 11 attributes
- Interpolation: From bad wine to good wine