Kernel Representations and Kernel Estimation

Peter Bickel

U.C. Berkeley

December, 2008

with assistance of C. Bae

Probable collaborators: Bin Yu, Liza Levina and others
Outline

1. Kernel Representations and Laplacians Kernel Matrices
2. Applications to Statistics
3. Clustering: Sample and Population Versions
4. Asymptotics as $n \to \infty$. (Koltchinskii, Giné; Belkin-Niyogi, Belkin, Von Luxburg; Nadler et al.)
5. Heuristics for Scale Going to 0: Kernels
6. Heuristics for Scale Going to 0: Laplacians (Nadler et al.)
7. Connections with Large $d$ (El Karoui)
8. Some Future Directions
Kernel Representations

Given: $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$, i.i.d. $\mathbb{P}$, $\frac{d\mathbb{P}}{d\lambda} \equiv p$.

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

- Symmetric

- $\|K(\mathbf{x}_i, \mathbf{x}_j)\|$ positive definite for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in general position.

- Represent $\mathbf{x}$ by $K(\mathbf{x}, \cdot)$: Feature vector
  - Prototypical $K$: $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$
  - Usual $K$: $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{|\mathbf{x} - \mathbf{y}|^2}{w^2}}$
Application in Statistics

1. Estimation of functions using splines, Bayes priors on Sobolev spaces

Closure in $\| \cdot \|_K$ of linear span of $\{ K(x, \cdot) \}$: RKHS

$$\| f \|_K^2 = \sum_j \alpha_j^2 K(x_j, x_j), \ f \equiv \sum_j \alpha_j K(x_j, \cdot)$$

G Wahba (e.g. CBMS-NSF Series, SIAM (1990))

2. Support vector machines: Computation

Vapnik (1998), Statistical Learning Theory

Polynomial regression of order $d$ on $\mathbb{R}$

$$K : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \ K(x, y) = \sum_{r=0}^{d} x^{r} y^{r} = \frac{[xy]^{d+1} - 1}{xy - 1}$$
3. Kernel ICA

Independence $\equiv \text{Corr}(K(X, x), K(Y, y)) = 0$ for all $x, y$. (for “dense” RKHS)

Empirically:

$$\frac{1}{n} \sum_{i=1}^{n} \left( K(X_i, X_j) - \bar{K}_j^X \right) \left( K(Y_i, Y_k) - \bar{K}_k^Y \right) \leq 0$$

for “all” $j, k$, $\bar{K}_j^X \equiv \frac{1}{n} \sum_{i=1}^{n} K(X_i, X_j)$ and $\bar{K}_k^Y \equiv \frac{1}{n} \sum_{i=1}^{n} K(Y_i, Y_k)$.

Jordan, Bach (2002), JMLR
4. Clustering and Spectral Clustering

- $K_n = [K(X_i, X_j)]_{n \times n}$ (Kernel)
- $L_n = D_n - K_n$ (Laplacian)
- $D_n = \text{diag}(\sum_{j=1}^{n} K(X_i, X_j))$
- $\mathcal{L}_n \equiv I - D_n^{-1/2} K_n D_n^{-1/2}$ (Normalized Laplacian)
- $\tilde{\mathcal{L}}_n = K_n D_n^{-1}$ (Normalized Laplacian)
- $K_n$: symmetric and positive definite if $K$ is
- $L_n, \mathcal{L}_n$: symmetric and positive definite if $K \geq 0$
Clustering

Eigenstructures of $K_n, L_n, \mathcal{L}_n, \tilde{\mathcal{L}}_n$ used for

(1) Manifold estimation (Belkin, Niyogi (2003), Neural Computation)

(2) Graph spectral clustering (Shi, Malik (2000), IEEE Transactions on Pattern Analysis and Machine Intelligence)

(3) Clustering (Nadler, Lafon, Coifman, Kevrekidis, NIPS (2005) (NLCK))

Many others

Recent contributions:

- Tutorial: Von Luxburg (2007), Statistics and Computation
Clustering

Observations

1) $I - \mathcal{L}_n$ and $\tilde{\mathcal{L}}_n$ have same eigenvalues

2) Right eigenvector of $\tilde{\mathcal{L}}_n = D_n^{-1/2}$ (eigenvector of $\mathcal{L}_n$) and similarly for left eigenvectors.
Theory and Heuristics for $K_n, \tilde{\mathcal{L}}_n$

Sample version

- $K^w(x, y) \equiv K^w(x - y) = \frac{1}{w^d} e^{-\frac{|x-y|^2}{2w^2}}$
- $\hat{K}^w_n \equiv \frac{1}{n} K^w(X_i, X_j)_{n \times n}$
- $\hat{K}^w_n : L_2(\mathbb{P}_n) \to L_2(\mathbb{P}_n)$
- $\mathbb{P}_n \equiv$ Empirical distribution function

\[
\hat{K}^w_n(f)(x) = \int \hat{K}_w(x, y) f(y) d\mathbb{P}_n(y) = \frac{1}{n} \sum_{i=1}^{n} K_w(x - X_i) f(X_i)
\]

- $\hat{K}^w_n(1)(x) = \hat{p}_w(x)$: kernel density estimate
- $\left( \hat{K}^w_n(f)(X_1), \ldots, \hat{K}^w_n(f)(X_n) \right)^T = \|\hat{K}^w_n\|(f(X_1), \ldots, f(X_n))^T$
- $\tilde{\mathcal{L}}^w_n(f)(x) = \frac{1}{\hat{p}_w(x)} K^w_n(f)(x)$
- $\tilde{\mathcal{L}}^w_n(f)(1) = 1$
Implementation of Clustering

- Compute $k$ eigenvectors
- $[e_{ij}] \equiv (e_1, e_2, \ldots, e_k)_{n \times k}$
- Represent $X_i$ by $i^{th}$ row
- **Motivation** If $X_i = K(X_i, \cdot)$ in RKHS, eigenvector $e_t \in$ RKHS,
  $<e_t, K(X_i, \cdot)> = e_{it}$.
- Use $k$ means clustering
  **Ideal**: $e_{is}e_{it} = 0, s \neq t$. 
Theory and Heuristics for $K_n$, $\tilde{L}_n$

Population version

- $K^w : L_2(\mathbb{P}) \to L_2(\mathbb{P})$

$$K^w(f)(x) \equiv \int K^w(x - y)f(y)d\mathbb{P}(y)$$

$$= \int K^w(x - y)f(y)p(y)d\mu(y),$$

$\mu = \text{Lebesque measure}$.

Notes:

1. $K^w$ is Hilbert Schmidt

2. $K^w$ is self adjoint, positive definite

$\therefore K^w$ has a point spectrum $\{0 < \lambda_1^w < \lambda_2^w < \cdots \}$ with associated eigenfunctions $\{e_{ij}(\cdot), i \geq 1, j = 1, \ldots, n_i\}$, $n_i \equiv \text{multiplicity of } \lambda_i$. 
Theory and Heuristics for $K_n, \tilde{L}_n$


If $\Lambda_n = \sum_{i=1}^{n} \delta_{\hat{\lambda}_{in}}, \hat{\lambda}_{in} \geq \cdots \geq \hat{\lambda}_{nn}$, eigenvalues of $\hat{K}_n^w$.

$\Lambda = \sum_{i=1}^{\infty} \delta_{\lambda_i}, \lambda_1 > \lambda_2 > \cdots$, eigenvalues of $K^w$

$$\{\Lambda_n | i \leq K\} \Rightarrow \{\Lambda | i \leq K\}$$

$$\sum_{i=1}^{n} \lambda_{in}^2 \rightarrow \sum_{i=1}^{\infty} \lambda_i^2$$
Theory and Heuristics for $K_n$, $\tilde{L}_n$

- $\tilde{L}^w(f)(x) = \frac{1}{p^w(x)} K^w(f)(x)$
- $p^w(x) = \int K^w(x - y)p(y)d\mu(y)$
  - $L^w$ has same structure as $K^w$
  - Spectrum of $I - L^w =$ spectrum of $\tilde{L}^w$
  - (Right) eigenfunctions of $\tilde{L}^w = [p^w]^{-1/2}$(eigenfunctions of $L^w$)

Koltchinskii-Giné: $L_n$ spectrum converges to $L$ spectrum.
Convergence of Eigenvectors

- $\mathbf{v}_{jn} \leftrightarrow \lambda_{jn} \rightarrow \lambda_j$, $\lambda_{1n} > \lambda_{2n} > \cdots$
- $\tilde{\mathbf{v}}_{jn} \leftrightarrow \tilde{\lambda}_{jn} \rightarrow \tilde{\lambda}_j$, $0 < \tilde{\lambda}_{1n} < \tilde{\lambda}_{2n} < \cdots$
- $\lambda_j, \tilde{\lambda}_j$ isolated
- (LBB): $|\mathbf{v}_{jn} - (e_j(x_1), \ldots, e_j(x_n))^T|_\infty \rightarrow 0$
  $e_j(\cdot)$—eigenfunction $\leftrightarrow \lambda_j$
Clustering Controversy

Which works better?

- (LBB) argue that $\mathcal{L}, \tilde{\mathcal{L}}_n$ better than $L_n$. Don’t explain why clustering occurs - but see (L Tutorial)
- (NLCK) argue convincingly for $\tilde{\mathcal{L}}_n$

1. Clustering according to distance defined by Markov Chain on $\mathbb{R}^d$, $p(y|x) = \frac{K^w(x,y)}{p^w(x)}$.

2. Consider $w \to 0$, relate to diffusion.
Population Heuristics for $w \to 0$

- $K^w(f)(x) \to f(x)p(x)$ as $w \to 0$ (Diagonal operator)
- $\int K_w(x, y)f(y)p(y)d\mu(y) - f(x)p(x) = \int K(0, z)(p(x + wz)f(x + wz) - p(x)f(x))dz$

By $L_2$ convergence theorem, if $\mu = \text{Lebesque measure},$

$$||K_w(f) - fp||_2 \to 0, \text{ } fp \in L_2(\mu)$$
Spectral Theory for Normal Operator \( T : L_2 \to L_2 \)

Definition:

\[
\sigma(T) = \{ \lambda \in \mathbb{C} : (T - \lambda I) \text{ does not have bounded inverse} \}
\]

Facts

1. \( \sigma(T) \) is compact
2. \( \exists E_T : \mathcal{B} \to \mathcal{P}(L_2), \mathcal{B} \equiv \text{Borel } \sigma \text{ field on } \sigma(T) \),
   
   \[ \mathcal{P}(L_2) = \{ \text{Projection operators on } L_2 \}, \quad E_T(\emptyset) = 0, \quad E_T(\sigma(T)) = \text{Identity}, \quad E_T(A \cap B) = E_T(A)E_T(B), \]
   
   \[ E_T(\sum_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} E_T(A_j) \text{ if } A_j \cap A_k = \emptyset, \ j \neq k, \]
   
   such that \( E_T \) has uniquely property

\[
T = \int_{\sigma(T)} \lambda dE_T(\lambda)
\]
Spectral Theory for Diagonal $T : L_2 \rightarrow L_2$

3. If $T(f) = gf$, $g$ fixed, $\sigma(T) = \text{Range of } g$

4. If $\sigma(T) = \sum_{i=1}^{L} [a_i, b_i]$, $g^{-1}[a_i, b_i] \equiv X_i$. $E_T[a_i, b_i](f) = 1 \chi_i f$.

5. If $||T_n(f) - T(f)|| \rightarrow 0 \ \forall f$, then $E_{T_n}[a, b](f) \rightarrow E_T[a, b](f)$
   $\forall f, a, b$ with $\mathbb{E}\{a\} = \mathbb{E}\{b\} = 0$. 
Perturbation Theory

If $T_n \leftrightarrow K^{w_n}, w_n \to 0$

If $E_{jn} \leftrightarrow$ Projection on eigenspace corresponding to $\lambda_{jn}$, $\lambda_{1n} > \cdots > \lambda_{kn} > \cdots$, eigenvalues of $T_n$,

$E_{T_n}[c, d] = \sum \{E_{jn} : \lambda_{jn} \in [c, d]\}$

By (1)

$E_{T_n}[c, d](f) \to 1_{\mathcal{X}} f, \mathcal{X} = p^{-1}[c, d]$
One Dimensional Clustering

Permute $X_1, \ldots, X_n$. Let

- $T_n f(x) \equiv \frac{1}{nw} \sum_{i=1}^{n} K \left( \frac{x-X_i}{w} \right) f(X_i)$
- $Tf \equiv pf$
- $IMSE \equiv \mathbb{E} \int ((T_n f)(x) - (Tf)(x))^2 \, dx$

Conditions

(i) $pf \in L_1$, $[pf]' \in L_2$, $p, f$ bounded

(ii) $\int K^2(z)dz < \infty$, $\int z^2 K(z)dz < \infty$

(iii) $w_n \to 0$, $nw_n \to \infty$
**Claim**

Under $(i) - (iii)$, $IMSE \to 0$.

Further, if $E_n \leftrightarrow T_n$, $E \leftrightarrow T$, then

$$
\mathbb{E} \left[ \int ((E_n - E)[a, b]f(x))^2 \, dx \right] \to 0,
$$

for $a, b$ with $\mathbb{E}\{a\} = \mathbb{E}\{b\} = 0$.

**Proof**

By $L_2$, $L_1$ continuity theorems and Taylor expansion.
One Dimensional Clustering

In limit:

Eigenvalues $\leftrightarrow p_1$ have last set of coordinates $\approx 0$

Eigenvalues $\leftrightarrow p_2$ have first set of coordinates $\approx 0$
One Dimensional Clustering

In general

(BSY) suggest look for eigenvectors with constant signs not necessarily among top eigenvalues.
Our Analysis for One Dimensional Clustering

Suppose $X_1 < \cdots < X_n$.

Let first $N$ eigenvectors of $K_n \leftrightarrow \{E_{nN}\}$

1) Threshold coordinates of eigenvectors

2) Select $k$ which
   
   a) Have constant sign
   
   b) Are sparse
   
   c) Have connected support

3) (Better) Find $k$ elements of linear span of first $N$ eigenvectors which satisfy a), b), c) maximally
Clustering Examples - 1 Dimension

A. Sample of 400 from $\mathcal{N}(0, 1)$

B. First eigenvector of $K_n$ when data is ordered

C. First eigenvector for random sample
Clustering Examples

Kernel Eigenvectors (Data ordered)

Eigenvectors carry information

0.5N(0,1)+0.5N(2,1)

0.5N(0,1)+0.5N(2,0.25)
Clustering Examples

Kernel method: $0.5\mathcal{N}(0,1)+0.5\mathcal{N}(2,0.25)$

Results of 2 means clustering using kernel clustering for sample of 400 from $\frac{1}{2}\mathcal{N}(0,1)+\frac{1}{2}\mathcal{N}(2,0.25)$

- **Blue** Classified as $\mathcal{N}(0,1)$ correctly
- **Yellow** Misclassified, in fact was drawn from $\mathcal{N}(0,1)$
- **Red** Classified as $\mathcal{N}(2,0.25)$ correctly
- **Green** Misclassified, in fact was drawn from $\mathcal{N}(2,0.25)$
Clustering Examples

Normalized Laplacian Eigenveotrs(Data ordered)(A,B)
Results of 2 means clustering using Normalized Laplacian for sample of 400 from $\frac{1}{2} \mathcal{N}(0, 1) + \frac{1}{2} \mathcal{N}(2, 0.25)$

- **Blue**: Classified as $\mathcal{N}(0, 1)$ correctly
- **Yellow**: Misclassified, in fact was drawn from $\mathcal{N}(0, 1)$
- **Red**: Classified as $\mathcal{N}(2, 0.25)$ correctly
- **Green**: Misclassified, in fact was drawn from $\mathcal{N}(2, 0.25)$
More than 1 dimension

- Difficulties with $K_n$ (and unnormalized Laplacian) because there is no clear guide to which high eigenvalue, eigenvectors to keep (LBB)
- Constant sign seems insufficient and as yet no analogue to ordering as in 1 dimension
Theory for $\tilde{\mathcal{L}}_n$(NLCK)

- $\tilde{\mathcal{L}}^w \rightarrow \text{Id}$
- $\frac{\text{Id} - \tilde{\mathcal{L}}^w}{w} \rightarrow$ Differential operator on dense subset of $L_2$
- $f \rightarrow \triangle f - C(\nabla \log p, \nabla f)$, $\triangle f \equiv \sum_{j=1}^d \frac{\partial^2 f}{\partial^2 x_j}$,
  $\nabla g = \left( \frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_d} \right)^T$
- With suitable boundary conditions, discrete eigenvalues, $0 < \lambda_2 < \cdots$, eigenfunctions are nearly constant and drop off quickly around maxima ($\nabla \log p = 0$). Hessian nd.
- Clustering is possible but $k$ means has to be applied since eigenvectors uninterpretable unless data are structured.
Clustering Examples

Truth:

- Class 1: $X_1 \sim \mathcal{N}(0, 1), X_2 = 0$, Class 2: $X_1 = 0, X_2 \sim \mathcal{N}(3, 1)$,
  Class 3: $X_1 \sim \mathcal{N}(6, 1), X_2 \sim \mathcal{N}(6, 1)$, independent
- Each sample has probability $1/3$ from these 3 classes

Kernel clustering with default $w$
5 eigenvectors obey (BSY) rule.
### Clustering Examples

**Truth:**

- Class 1: $X_1 \sim \mathcal{N}(0, 1)$, $X_2 = 0$,
- Class 2: $X_1 = 0$, $X_2 \sim \mathcal{N}(3, 1)$,
- Class 3: $X_1 \sim \mathcal{N}(6, 1)$, $X_2 \sim \mathcal{N}(6, 1)$, independent

- Each sample has probability 1/3 from these 3 classes

---

**Laplacian clustering**
Clustering using $\hat{\mathcal{L}}_0$ (Diffusion Operator) (Nadler et al.)

Figure 1: Diffusion map results on different datasets. Top - the datasets. Middle - the eigenvalues. Bottom - the first eigenvector vs. $x_1$ or the first and second eigenvectors for the case of three Gaussians.
Questions

1. Show that sample theory converges to limit as \( n \to \infty \) and \( w \to 0 \).

2. Rate results for \( \tilde{L} \)

3. Is there a way of rescuing \( K \) for \( d > 1 \)?

4. What happens as \( d \to \infty \)?
The case $d \to \infty$

N. El Karoui (2008), Annals of Statistics

(Sample) If $\mathbf{X}_{d \times 1} \sim \mathcal{N}(0, \text{Id})$,

$$\sigma \left( \frac{1}{n} \| K(\mathbf{X}_i, \mathbf{X}_j) \| \right) \approx \sigma \left( \frac{1}{n} \mathbf{X}_i^T \mathbf{X}_j \right) = \sigma \left( \frac{1}{n} \sum \mathbf{X}_i \mathbf{X}_i^T \right)$$

No gain

$\sigma(\mathcal{L}_n)$ behaves the same way
5. What happens if, for example,

- $X_{d \times 1} \sim \sum_{j=1}^{K} \pi_j F_j$
- $Z_j \sim F_j$, $A_j : d \times p_j$, $Z_j \sim \mathcal{N}_{p_j}(\mu_j, \Sigma_j)$
- $\sum_{j=1}^{K} p_j \leq p < \infty$ all $d$
Some Support for This View

The Zip Code Digits

• 16 × 16 images of handwritten Zip code images

• 25 randomly drawn images for each digit

• For each digit histogram of local dimension estimates for
  4 × 4 patches (144 patches per images)

• 600 – 1200 images per digit

• Dimension estimate (Levina, Bickel (2005), NIPS)
Some Support for This View (A. Rothman, E. Levina)

Average histograms for all digits with patches treated as units
Some Support for This View (A. Rothman, E. Levina)

Histograms for all digits images treated as units
**Questions**

1. If this view holds do the eigenvectors of \( \tilde{L}_n \) provide as \( w_n \to 0 \) successful clustering in \( p \) dimensional space? (See Goldberg, Ritov Zakkai (2008), JMLR)

2. Can this be established asymptotically in a reasonable way by considering \( \tilde{L} = \lim_{w \to 0} \tilde{L}^w \) and letting \( n \to \infty, w_n \to 0 \) in \( \tilde{L}_n \)?

3. And/or can \( K_n \) be adapted to give similar results?

4. As for \( d = 1 \), can preclustering yield low dimensional representation for high dimensional data?
Clustering Examples

Truth:

- Class 1: $X_1 \sim \mathcal{N}(0, 1), X_2 = 0$, Class 2: $X_1 = 0, X_2 \sim \mathcal{N}(3, 1)$,
- Class 3: $X_1 \sim \mathcal{N}(6, 1), X_2 \sim \mathcal{N}(6, 1)$, independent
- Each sample has probability $1/3$ from these 3 classes

Kernel clustering with small $w$