Graph Wavelets

F. Göbel
joint work with G. Blanchard, U. von Luxburg and C. Gehrke

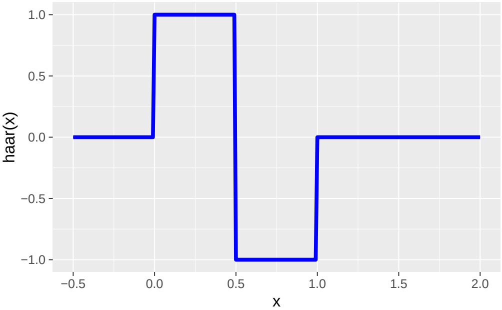
Universität Potsdam

Research Unit 1735
Structural Inference in Statistics:
Adaptation and Efficiency

Stochastic processes and statistical machine learning I
Workshop in Potsdam 14/02/2018

Outline

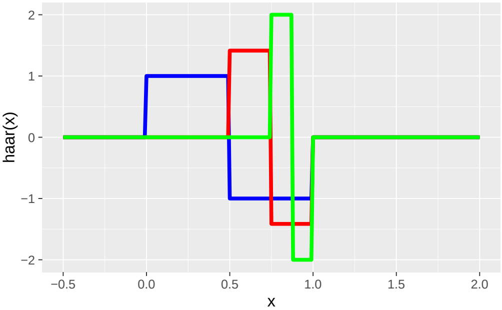
WAVELETS ON $\mathbb{R}$



- Classical example: Haar wavelet, $\psi(t) = \mathbf{1}\{t \in [0, \frac{1}{2})\} - \mathbf{1}\{t \in [\frac{1}{2}, 1)\}$
- Let $\psi_{m,n}(t) := 2^{m/2}\psi(2^m t - n)$
- The $(\psi_{m,n})_{m,n\in\mathbb{Z}^2}$ form an orthonormal basis of $L^2(\mathbb{R})$

$$f \in L^2(\mathbb{R}) \;\Rightarrow\; f(\cdot) = \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} \langle \psi_{m,n}, f\rangle \, \psi_{m,n}(\cdot)\,.$$

Classical wavelets: properties



- $\psi_{m,n}(t) := 2^{m/2}\psi(2^m t - n)$
- $(\psi_{m,n})$ orthogonal basis
- $(\psi_{\cdot,n})$ are rescaled versions of ψ
- $(\psi_{m,\cdot})$ are translations of each other

What are wavelets good for?

- Visualization of signal in space(time)/frequency plot
- Data compression
  Local adaptivity
  Smoothing (e.g. nonlinear by thresholding)
- Denoising:
  - Observe: $Y_i = f(x_i) + \varepsilon_i\,, \qquad \text{with } x_i = \frac{i}{N}\,, \qquad i = 1, \ldots, N$
  - Compute empirical/noisy wavelets coefficients:

$$\widehat{\alpha}_{m,k} = \langle \mathbf{Y}, \psi_{m,k} \rangle_{P_N} = \frac{1}{N} \sum_{i=1}^{N} Y_i \psi_{m,k}(x_i)$$

  - Threshold empirical coefficients:

$$\widetilde{\alpha}_{m,k} = \widehat{\alpha}_{m,k} \mathbf{1}\{|\widehat{\alpha}_{m,k}| \geq \tau\}$$

  - Reconstruct with thresholded coefficients:

$$\widehat{f}(\cdot) = \sum_{m,k} \widetilde{\alpha}_{m,k} \psi_{m,k}(\cdot)$$

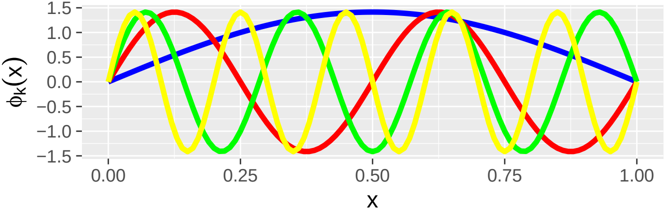
Why wavelets?

- There exist other classical bases, for instance Fourier basis. Why use wavelets?
- Fourier functions are not localized. A truncated Fourier expansion approximates well a signal that is “uniformly smooth over the domain”.
- Wavelets give better approximations for signals of inhomogeneous smoothness, e.g. piecewise smooth functions with some discontinuities.
- How to extend the wavelet approach if points x_i are in high dimension, not uniformly distributed?

Littlewood-Paley decomposition

- Laplace-operator $-\Delta f = -\frac{\partial^2}{\partial x^2} f(x)$
- Eigenvalues $\lambda_k = k^2\pi^2$
- ONB/Eigenfunctions $\Phi_k(x) = \sqrt{2}\sin(k\pi x)$
- Fourier decomposition

$$f(x) = \sum_{k\geq 1} \langle f, \Phi_k \rangle \, \Phi_k(x)$$

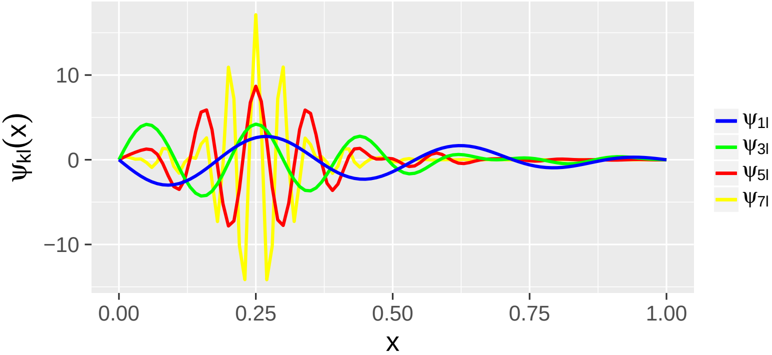


- Eigenfunctions are not localised

Littlewood-Paley decomposition

- Take a point x_ℓ fixed, construct localized functions via

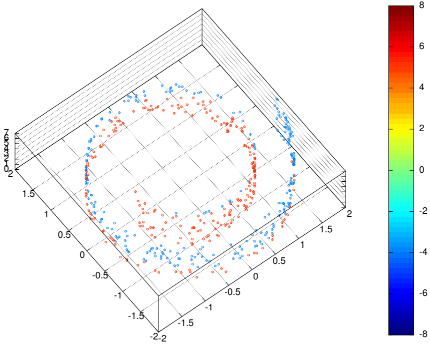
$$\Psi_{i\ell}(x) = \sum_k \sqrt{\zeta(2^{-i}\lambda_k)}\Phi_k(x_\ell)\Phi_k(x)$$

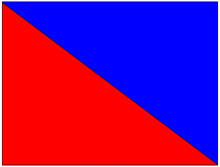


- Note: the Meyer wavelet is constructed following a similar principle
- Extend to more general case?

Outline

TOY EXAMPLE





Setting

Model n observations,

$$y_i = f(x_i) + \epsilon_i$$

- $x_i \in \mathbb{R}^d$, d can be big, x_i realization of $X \sim P$
- $D = \{x_1, \ldots, x_n\}$
- $y_i \in \mathbb{R}$ noisy observation
- ϵ_i noise, iid, $\mathbb{E}[\epsilon] = 0, \textbf{Var}(\epsilon_i) = \sigma^2$

Assumption $x_i \in M \subset \mathbb{R}^d$

e.g. M low-dimensional compact submanifold

Task (Denoising): Estimate $(f(x_i))_i$ given $(y_i)_i$ using the unknown geometric structure of M

(f may have inhomogeneous regularity)

Literature

- On constructing localised, wavelet-like frames for manifolds

  (Regular manifolds) Narcowich, Petrushev, and Ward (2006): on spheres; Petrushev and Xu (2008), Baldi, Kerkyacharian, Marinucci and Picard (2009): on balls
  (on compact homogeneous manifolds) Geller and Mayeli (2009), Geller and Pesenson (2011): based on Laplacian operator; Kerkyacharian, Le Pennec and Picard (2011): on more general operators
  Coulhon, Kerkyacharian and Petrushev (2012): Develop band limited well-localised frames
  "Heat Kernel Generated Frames in the Setting of Dirichlet Spaces"

- On data-adapted wavelet-like frames

  Hammond, Vandergheynst and Gribonval (2011) (graph-based)
  Gavish, Nadler and Coifman (2010) (tree-based)

WISHLIST

- Consider $f \in L^2(D)$
- We want a decomposition of f with respect to a set of functions

$$f(x) = \sum_j \langle f, g_j \rangle \, g_j$$

  Properties of the dictionary
  - (Over)complete,
  - Adapted to the structure of the domain of f
  - Ideally: the dictionary exhibits the features of a wavelet basis (multiscale, localization, ...)
- Application in statistics:

$$y = \sum_j \langle y, g_j \rangle \, g_j = \sum_j \left(\langle f, g_j \rangle + \langle \epsilon, g_j \rangle \right) g_i$$

  then estimating f corresponds to estimate the coefficients $\langle f, g_j \rangle$ given $\langle y, g_j \rangle$

PLAN OF ATTACK

- We want to use a “Fourier analysis” adapted to the domain of f (and possibly to the covariate distribution)
- Fourier analysis exists on manifolds, but the manifold containing the data is unknown a priori
- **Solution:** use approximation by a neighborhood graph constructed on the data + graph Laplacian (principle underlying Laplacian Eigenmaps).
- Then apply the “frequency decomposition” device to the obtained spectral decomposition

Approach mainly based on work of Coulhon et al. (2012)
Similar approach: Hammond et al. (2011) (general frame)
Different approach: Gavish et al. (2010) (hierarchical tree, Haar-like basis)

Steps of the construction

We need:

- Neighborhood graph A:
  - Finite undirected (weighted) graph
  - represented by symmetric adjacency matrix $A = (a_{ij})$
  - Degree of a vertex v_i:$d_i = \sum_j a_{ij}$, $G := diag(d_1, \ldots, d_n)$
  - Graph types: (weighted) k-NN, (weighted) ϵ, complete weighted

STEPS OF THE CONSTRUCTION

We need:

- Neighborhood graph A
- Graph Laplacian:
  Unnormalized
  $$L_u = G - A$$
  Normalized
  $$L_{norm} = I - G^{-1/2}AG^{-1/2}$$

  - Properties of L: symmetric, positive semi-definite
  - Spectral theorem for matrices:
    The eigenvectors Φ_i of L are an orthonormal basis of $L^2(D) = \mathbb{R}^n$ and the eigenvalues λ_i are ≥ 0

Laplacian Eigenmaps

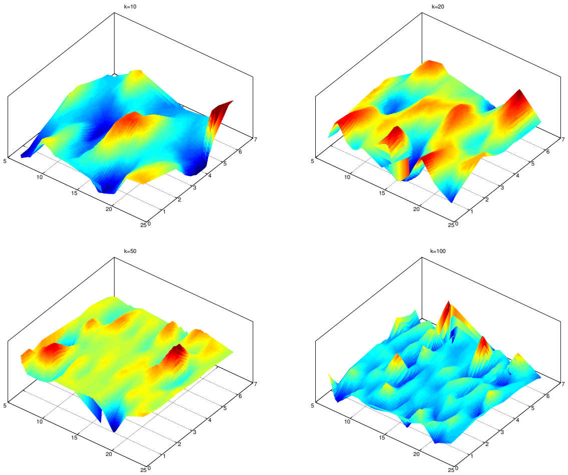


Figure: Swiss roll data: eigenvectors Φ_j for $j = 10, 30, 50, 100$.

STEPS OF THE CONSTRUCTION

We need:

- Neighborhood graph A
- Graph Laplacian: L, $\{\Phi_i\}_i$, $\{\lambda_i\}_i$
- Function system (decomposition of unity):
  $\{\zeta_k\}_{k\in\mathbb{N}}$ is a sequence of functions $\zeta_k : \mathbb{R}_+ \to [0,1]$ satisfying
  (DoU) $\sum_{j\geq 0} \zeta_j(x) = 1$ for all $x \geq 0$;
  (FD) $\#\{\zeta_k : \zeta_k(\lambda_i) \neq 0\} < \infty$ for $i = 1, \dots, n$.

Definition of the dictionary

Definition

The dictionary $\{\Psi_{k\ell} \in \mathbb{R}^n, 0 \leq k \leq Q, 1 \leq \ell \leq n\}$ is defined by

$$\Psi_{k\ell} = \sum_{i=1}^{n} \sqrt{\zeta_k(\lambda_i)}\Phi_i(x_\ell)\Phi_i. \tag{1}$$

with $Q := \max\{k : \exists i \in \{1, \dots, n\} \text{ with } \zeta_k(\lambda_i) > 0\}$.

RESULT: TIGHT FRAME

Theorem

The dictionary $\{\Psi_{k\ell}\}_{k,\ell}$ is a Parseval frame for $\mathbb{R}^n$, that is:

(a) For all $x \in \mathbb{R}^n$:

$$\|x\|^2 = \sum_{k,\ell} |\langle x, \Psi_{k\ell}\rangle|^2 .$$

(b) For all $y \in \mathbb{R}^n$ ($y : D \to \mathbb{R}$) the recovery formula holds:

$$y = \sum_{k,\ell} \langle y, \Psi_{k\ell}\rangle \, \Psi_{k\ell}. \tag{2}$$

(c) $\forall k, \ell : \; \|\Psi_{k\ell}\| \leq 1$

Choice of $\{\zeta_k\}_k$ - Multiscale bandpass filter

Choice corresponds to Coulhon et al. (2012)
(smooth Littlewood-Paley decomposition)

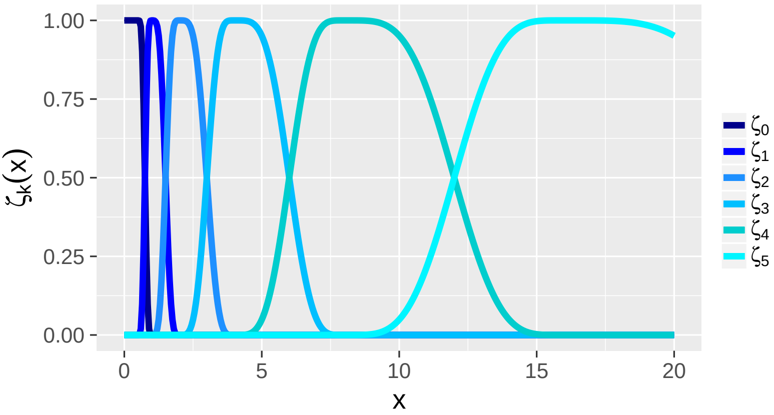
Definition (Multiscale bandpass filter)

Let $g \in C^\infty(\mathbb{R}_+)$, $\mathrm{Supp} g \subset [0,1]$, $0 \le g \le 1$, $g(u) = 1$ for $u \in [0, 1/b]$ (for some constant $b > 1$). For $k \in \mathbb{N} = \{0, 1, \ldots\}$ the functions $\zeta_k : \mathbb{R}_+ \to [0,1]$ are defined by

$$\zeta_k(x) := \begin{cases} g(x) & \text{if } k = 0 \\ g(b^{-k}x) - g(b^{-k+1}x) & \text{if } k > 0 \end{cases} \tag{3}$$

The sequence $\{\zeta_k\}_{k\ge0}$ is called multiscale bandpass filter.

Choice of $\{\zeta_k\}_k$ - Multiscale bandpass filter



Properties: $\zeta_k \in C^\infty(\mathbb{R}_+)$, $0 \leq \zeta_k \leq 1$,
$\zeta_k(x) = \zeta_1(b^{-(k-1)}x)$ for $k \geq 1$
$\mathrm{Supp}\zeta_0 \subset [0, 1]$, $\mathrm{Supp}\zeta_k \subset [b^{k-2}, b^k]$ for $k \geq 1$

Outline

Localization

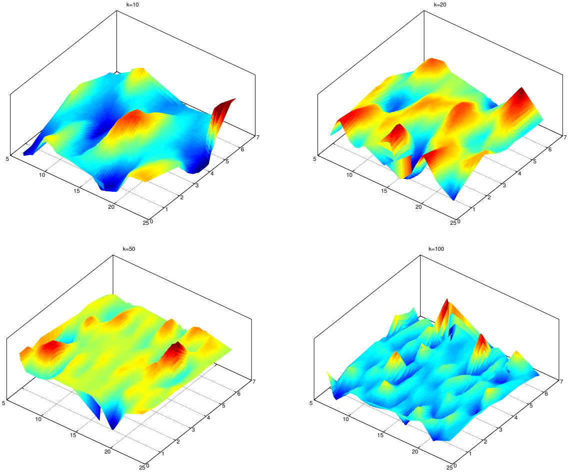


Figure: Swiss roll data: eigenvectors Φ_j for $j = 10, 30, 50, 100$; frame elements Ψ_{kl} for l fixed and $k = 0, 2, 5, 7$.

Localization

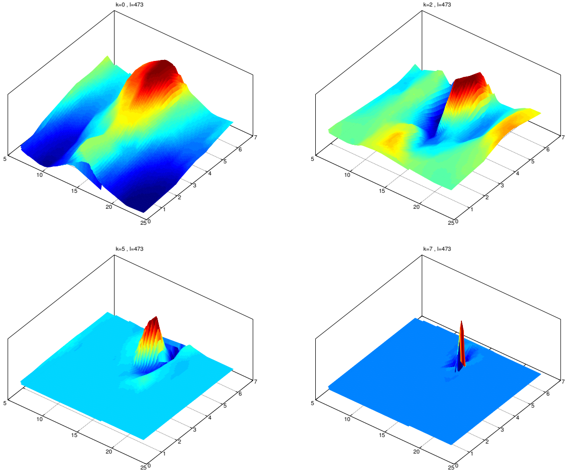


Figure: Swiss roll data: eigenvectors Φ_j for $j = 10, 30, 50, 100$; frame elements Ψ_{kl} for l fixed and $k = 0, 2, 5, 7$.

Why Doubling Condition and Poincaré Inequality

- in [CKP12]: heat kernel bounds important ingredient for localization in their setting: DC+ Poincare $\leftrightarrow$ Harnack inequality $\leftrightarrow$ Gaussian estimate for heat kernel
- for graph setting: Delmotte (1997), Barlow and Chen (2016) - similar results
- Question: If manifold satisfies DC, does the graph satisfy a DC as well?
  When does the graph satisfy a local Poincaré inequality?

SPATIAL LOCALIZATION?

Coulhon et al. (2012): the almost exponential localization of $\Psi_{k\ell}$ follows from 2 sufficient geometrical conditions:
If M compact and μ finite and

- Doubling measure:

$$\mu(B(x,2r)) \leq 2^d \mu(B(x,r)) \text{ for all } x \text{ and } r > 0.$$

- Local Poincaré inequality:

$$\int_{B(x,r)} (f(y) - f_B)^2 d\mu(y) \leq Cr \int_{B(x,r)} \|\nabla f\|^2 d\mu \text{ for all } x \text{ and } r > 0,$$

  with f_B mean of f over $B(x,r)$

Do we have an appropriate discrete analogue on a geometrical graph based on $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mu$?

ASSUMPTIONS

Assumption

(B1) choose $\lambda_1, \lambda_2 \in (0, 1)$, $q \in (0, 1)$

(B2) $\mathcal{M} \subset \mathbb{R}^k$ compact submanifold (with parameters r_0, s_0), $(\mathcal{M}, d_{\mathcal{M}}, \mu)$

(B3) $\mathcal{M}$ is geodesically convex

(B4) (finite sample of $\mathcal{M}$ as vertex set of an ϵ-graph)

(B5) ϵ-graph with parameter $\epsilon > 0$ such that $\epsilon < s_0$ and $\epsilon \leq (2/\pi) r_0 \sqrt{24\lambda_1}$

(B6) choose sample size $n = n(q, \lambda_2, \epsilon, \mu)$ such that

$$n \geq \frac{\ln(q \inf_{y \in \mathcal{M}} \mu(B(y, \epsilon\lambda_2/16)))}{\ln((1 - \inf_{z \in \mathcal{M}} \mu(B(z, \epsilon\lambda_2/8))}$$

minimum radius of curvature $r_0 = r_0(\mathcal{M}) := \left(\max_{\gamma,t} \|\ddot{\gamma}(t)\|\right)^{-1}$ (γ unit-speed geodesics)
minimum branch separation
$s_0 := \max\{s : s > 0 \text{ and } \|x - y\| < s \Rightarrow d_{\mathcal{M}}(x, y) \leq \pi r_0 \text{ for } x, y \in \mathcal{M}\}$.

Doubling Condition

Theorem

Under (B1-B6), with $\lambda_1 = \lambda_2 = 0.5$*, with prob at least* $1 - 2q$*, for all* $x \in V$ *and for all* $r \geq 2$ *with* $\hat{\mu}_n(B_{G,SP}(x,r)) \geq 2(\sqrt{\frac{-\ln q + \ln 3n^2}{n}} + \frac{2}{n})^2$ *it holds for* n *large enough*

$$\hat{\mu}_n(B_{G,SP}(x,2r)) \leq 2^{3.2+6v}\hat{\mu}_n(B_{G,SP}(x,r)).$$

The proof is based on

- approximation of distances $d_M, d_{G,E}, d_{SP}$ (using $d_M \approx d_E$ by [BSLT00]),
- Lemma: uniform bound

$$\mathbf{P}\left(\sup_{i=1..n}\sup_{r>0}\left|\sqrt{\hat{\mu}_n(B_{\mathcal{M}}(X_i,r))} - \sqrt{\mu(B_{\mathcal{M}}(X_i,r))}\right| > \delta\right) \leq \alpha.$$

- and volume doubling on the manifold.

Sketch of proof

Distance approximation (see [BSLT00, Main Theorem B])

$$(1-\lambda_1)d_{\mathcal{M}}(x,y) \le d_{G,E}(x,y) \le (1+\lambda_2)d_{\mathcal{M}}(x,y) \text{ whp}$$

and

$$\frac{1}{4}\,\epsilon\,\left(d_{G,SP}(x,y)-1\right) \le d_{G,E}(x,y) \le \epsilon\, d_{G,SP}(x,y).$$

Then, for fixed $s \ge 0$, we can derive the inequalities

$$\begin{aligned}
\hat{\mu}_n(B_{G,SP}(x,2r)) &\le \hat{\mu}_n\left(B_{\mathcal{M}}\left(x,(1-\lambda_1)^{-1}\epsilon 2r\right)\right)\\
&\le \frac{3}{2}\mu_n\left(B_{\mathcal{M}}\left(x,(1-\lambda_1)^{-1}\epsilon 2r\right)\right)+3\delta^2\\
&\le \frac{3}{2}2^{\lceil s\rceil v}\mu_n\left(B_{\mathcal{M}}\left(x,\frac{(1-\lambda_1)^{-1}}{2^s}\epsilon 2r\right)\right)+3\delta^2\\
&\le \frac{3}{2}2^{\lceil s\rceil v}\left(\frac{3}{2}\hat{\mu}_n\left(B_{\mathcal{M}}\left(x,\frac{(1-\lambda_1)^{-1}}{2^s}\epsilon 2r\right)\right)+3\delta^2\right)+3\delta^2\\
&\le \frac{3}{2}2^{\lceil s\rceil v}\left(\frac{3}{2}\hat{\mu}_n\left(B_{G,SP}\left(x,\frac{4(1+\lambda_2)(1-\lambda_1)^{-1}}{2^s}2r+1\right)\right)+3\delta^2\right)+3\delta^2
\end{aligned}$$

which holds with high probability.

SKTECH OF PROOF - LEMMA

key words: conditioning on center point and radius, reduction to random radii: ordered/non-ordered, Okamoto's inequality

$$\mathbf{P}\left(\sup_{i=1..n}\sup_{r>0}\left|\sqrt{\hat{\mu}_n(B_{\mathcal{M}}(X_i,r))}-\sqrt{\mu(B_{\mathcal{M}}(X_i,r))}\right|>\delta\right)=\mathbf{P}\left(\sup_{i=1..n}\sup_{r>0}|T_{ir}|>\delta\right)$$
$$=\mathbf{P}\left(\bigcup_{i=1}^{n}\{\sup_{r>0}|T_{ir}|>\delta\}\right)\leq\sum_{i=1}^{n}\mathbf{P}\left(\sup_{r>0}|T_{ir}|>\delta\right)=\sum_{i=1}^{n}\mathbf{E}_{X_i}\left(\mathbf{P}\left(\sup_{r>0}|T_{ir}|>\delta\,\middle|\,X_i\right)\right)$$

decompose for fixed i the set

$$\{r>0\}=\{r_i^{(j)}:r_{ij}\neq 0, j\leq n\}\cup\bigcup_{j=1}^{n}\left(r_i^{(j)},r_i^{(j+1)}\right)$$

upper bound $\sup_{r>0}|T_{ir}|$ by $\max\{E_{1,i},E_{2,i},E_{3,i}\}$ for fixed i where

$$E_{1,i}:=\max_{j=1..n:r_{ij}>0}\left|T_{ir_{ij}}\right|,\ E_{2,i}:=\max_{j=1..n}\overline{T}_{ir_{ij}}\text{and } E_{3,i}:=\max_{j=1..n+1,j\neq i}-T_{ir_{ij}}.$$

Sketch of proof II - Lemma

Introduce non-biased random variable $\tilde{\mu}_n$

$$\begin{aligned}\mathbf{P}\left(\left|T_{ir_{ij}}\right| > \delta \;\middle|\; X_i, X_j\right) &= \mathbf{P}\left(\left|\sqrt{\hat{\mu}_n(B_{\mathcal{M}}(X_i, r_{ij}))} - \sqrt{\mu_n(B_{\mathcal{M}}(X_i, r_{ij}))}\right| > \delta \;\middle|\; X_i, X_j\right) \\ &\leq \mathbf{P}\left(\left|\sqrt{\tilde{\mu}_n(B_{\mathcal{M}}(X_i, r_{ij}))} - \sqrt{\mu_n(B_{\mathcal{M}}(X_i, r_{ij}))}\right| > \delta - \frac{1}{\sqrt{n}} \;\middle|\; X_i, X_j\right) \\ &\leq 2\exp\left(-(n-2)\left(\delta - \frac{1}{\sqrt{n}}\right)^2\right)\end{aligned}$$

using

Lemma (Okamoto Inequality)

Let $Y_i \sim B(p)$ iid with $\mathbf{E}(Y_i) = p$ and set $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} Y_i$. Then, for $\delta > 0$,

$$\mathbf{P}\left(\left|\sqrt{p} - \sqrt{\hat{p}}\right| > \delta\right) \leq 2\exp(-m\delta^2).$$

Local Poincaré Inequality

additional assumption: Ahlfors-regularity of μ

Definition (k-Ahlfors)

A measure μ on $(\mathcal{M}, d_{\mathcal{M}})$ is said to be $k-$ *Ahlfors* if

$$\exists c_l > 0, c_u > 0 \; \forall B_{\mathcal{M}}(x,r) : \quad c_l r^k \leq \mu(B_{\mathcal{M}}(x,r)) \leq c_u r^k$$

holds.

LPI

Theorem (main theorem - lpi in d_{SP} -version 1/n)

Assumptions: Ahlfors regular μ, existence of Bi-lipschitz homeomorphism, connected unweighted ϵ-graph, measures $\pi_x = 1/n$ on $V(G)$ and $\tilde{\pi}_x = 1/n_A$ for some $A = B_{\mathcal{M}}(r) \subset V(G)$, lower bound on edge weights (a).
Under B1-B6, then, whp, $\forall f, \forall x_i \in V(G)$,

$$\sum_{x \in \overline{B}_{G,SP}(x_0, r_{SP})} (f_x - \bar{f}_B)^2 1/n \leq C r_{SP}^2 \sum_{x,y \in \overline{B}_{G,SP}(x_0, \lambda r_{SP})} (f_x - f_y)^2 1/n.$$

Assumptions for results:

- existence of bi-Lipschitz homeomorphism: A compact, $\exists h : A \to [0,1]^k$ bi-Lipschitz-homeomorphism
  the existence of the bi-lip homeo is ensured by the condition $r < r_{max}$ in $d_{\mathcal{M}}$ distance
- $\exists\, 0 < L_{min} < L_{max} < \infty$ Lipschitz-constants, they should be global (independent of A), including factor $1/r_{\mathcal{M}}$

Sketch of proof

- general structure of the inequality [DS91]

$$\sum_{x\in A}(f_x - \overline{f}_A)^2\tilde{\pi}_x) \leq \kappa_A \frac{1}{2}\sum_{x\in A}\sum_{y\in A, y\sim x} a_{xy}(f_x - f_y)^2\tilde{\pi}_x$$

with $\tilde{\pi}_x := \frac{\pi_x}{\mu(A)} = \frac{\pi_x}{\sum_{y\in A}\pi_y}$ and $\overline{f}_A = \sum_{x\in A} f_x\tilde{\pi}_x$

$$\kappa_A := \max_{e=(a,b),a,b\in A} \sum_{\gamma_{xy}\ni e, \gamma_{xy}\in A} Q^A_{xy}\tilde{\pi}_x\tilde{\pi}_y \text{ with } Q^A_{xy} := \sum_{e\in\gamma^A_{xy}} \frac{1}{a_{kl}\tilde{\pi}_k}.$$

- lpi for d_M given bound on kappa: assume κ_A can be bounded by $C \cdot r^2$ whp for $A = \overline{B}_{\mathcal{M}}(r)$,
- lpi for d_{SP}

SKETCH OF PROOF II

- bound on kappa given existence of bi-lip:
  for $\pi_x = 1/n$ and unweighted graph: $c_A = C_A = 1/n_A$, $a_A = 1$

$$Q_{xy}^A \leq \frac{1}{c_A \cdot a_A} l_{max}(A) \qquad \text{and} \qquad \kappa_A \leq \frac{C_A^2}{c_A \cdot a_A} l_{max}(A) b_{max}(A)$$

  with

$$l_{max}(A) := \max_{x,y \in A} \mathrm{NE}\left(\gamma_{xy}^A\right) \qquad \text{and} \qquad b_{max}(A) := \max_{e \in G_A} \sum_{\gamma_{xy}^A \ni e} 1.$$

  bound $l_{max}(A)$ and $b_{max}(A)$ using random Hamming pathes [vLRH14]
- existence of bi-lip: upper bound for radius r

Outline

DENOISING - APPROACH

- Observe $y_i = f(x_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$
- Use recovery formula for y and f

$$f = \sum_{k,l} a_{kl} \Psi_{kl} \text{ with } a_{kl} = \langle \Psi_{kl}, f \rangle$$

$$y = \sum_{k,l} b_{kl} \Psi_{kl} \text{ with } b_{kl} = \langle \Psi_{kl}, y \rangle = a_{kl} + \langle \Psi_{kl}, \epsilon \rangle$$

- Apply thresholding method to the coefficients b_{kl}:

$$\hat{a}_{kl} = Thr(b_{kl})$$

- Define estimate:

$$\hat{f} := \sum_{k,l} \hat{a}_{kl} \Psi_{kl}$$

Theorem (Oracle-type inequality)

With soft-thresholding S_S and threshold $t_{kl} = \sigma^2 \|\Psi_{kl}\| \sqrt{2 \log n}$

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|^2\right] \leq (1 + 2\log n)\left(\sigma^2 + \sum_{k,l} \min(\sigma^2 \|\Psi_{kl}\|^2, \langle f, \Psi_{kl}\rangle^2)\right)$$

(See also Candes (2006))

Class of reference estimators: linear projection estimators (keep-or-kill)

$$\tilde{f}_J = \sum_{(k,l)\in J} b_{kl}\Psi_{kl}$$

Then

$$\inf_J \mathbb{E}\left[\left\|\tilde{f}_J - f\right\|^2\right] \leq \sum_{k,l} \min(\sigma^2 \|\Psi_{kl}\|^2, \langle f, \Psi_{kl}\rangle^2)$$

Proof Ingredients

- Property of Parseval frame

$$\left\|\sum_i a_i z_i\right\|^2 \leq \|a\|^2 = \sum_i a_i^2, \tag{3}$$

- Property of denoising model

$$\frac{\langle y, \Psi_{kl}\rangle}{\sigma\|\Psi_{kl}\|} \sim \mathcal{N}\left(\frac{a_{kl}}{\sigma\|\Psi_{kl}\|}, 1\right). \tag{4}$$

- Result from Donoho and Johnstone (1994):
  For $0 \leq \delta \leq 1/2$,$t = \sqrt{2\log(\delta^{-1})}$ and $X \sim \mathcal{N}(\mu, 1)$

$$\mathbb{E}\left[(S_s(X,t)-\mu)^2\right] \leq (t^2+1)\left(\exp\left(-\frac{t^2}{2}\right) + \min(1,\mu^2)\right). \tag{5}$$

- and $\sum_{k,l}\|\Psi_{kl}\|^2 = n$

OUTLINE

SIMULATIONS - BASICS

- simulations based on empirical mean squared error (*MSE*)
  $MSE(\hat{f}, f) = \frac{1}{n} \left\| \hat{f} - f \right\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(x_i) - f(x_i))^2$
- every method depends on one tuning parameter
- so far no prediction for our method
- optimize MSE wrt tunig parameter ($t_o = \operatorname{Arg}\underset{t}{\operatorname{Min}} MSE(\hat{f}_t, f)$) and compare the "optimal" MSEs

Comparison Frame Thr vs ONB Thr and ONB embedding I

Question: Does the frame lead to better results than ONB-based methods?

Example: sphere, jump function, $\sigma^2 = 1, n = 500, m = 50$

Graph	L	**FrTh**		**LETh**		**LETr**	
kNN	U	0.510	(0.050)	0.693	(0.061)	0.905	(0.108)
kNN	N	0.538	(0.046)	0.712	(0.055)	0.931	(0.094)
WkNN	U	0.521	(0.049)	0.652	(0.050)	0.800	(0.097)
WkNN	N	0.530	(0.049)	0.674	(0.057)	0.749	(0.091)
CGK	U	0.520	(0.055)	0.638	(0.065)	0.821	(0.107)
CGK	N	0.530	(0.052)	0.670	(0.050)	0.725	(0.081)
ϵG	U	0.505	(0.058)	0.650	(0.068)	0.865	(0.115)
ϵG	N	0.557	(0.052)	0.710	(0.059)	0.902	(0.106)
WϵG	U	0.482	(0.055)	0.622	(0.064)	0.787	(0.111)
WϵG	N	0.530	(0.049)	0.674	(0.057)	0.749	(0.091)

Smoothing Kernel Regression: min. MSE = 0.612 (0.066)
Kernel Ridge Regression: min. MSE = 0.594 (0.051)

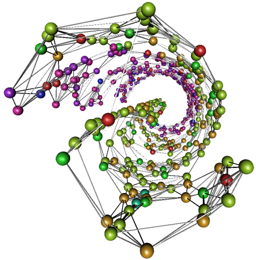
Comparison Frame Thr vs ONB Thr and ONB embedding II

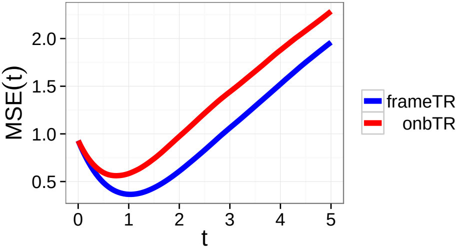
Example: swiss roll, jump function, $\sigma^2 = 1, n = 500, m = 50$

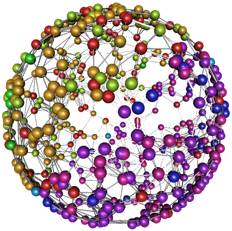
Graph	L	FrTh		LETh		LETr	
kNN	U	0.462	(0.043)	0.647	(0.039)	0.876	(0.079)
kNN	N	0.494	(0.043)	0.676	(0.043)	0.902	(0.071)
WkNN	U	0.443	(0.045)	0.600	(0.050)	0.790	(0.102)
WkNN	N	0.500	(0.043)	0.659	(0.045)	0.775	(0.079)
CGK	U	0.491	(0.053)	0.625	(0.057)	0.844	(0.096)
CGK	N	0.520	(0.047)	0.648	(0.049)	0.713	(0.079)
ϵG	U	0.459	(0.049)	0.610	(0.053)	0.872	(0.095)
ϵG	N	0.532	(0.045)	0.681	(0.050)	0.884	(0.089)
WϵG	U	0.441	(0.049)	0.574	(0.049)	0.793	(0.113)
WϵG	N	0.503	(0.045)	0.643	(0.051)	0.744	(0.089)

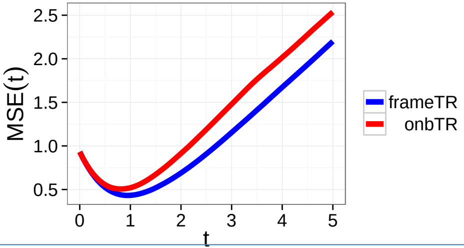
Smoothing Kernel Regression: min. MSE = 0.589 (0.082)
Kernel Ridge Regression: min. MSE = 0.779 (0.052)

Comparison Frame and ONB Thresholding







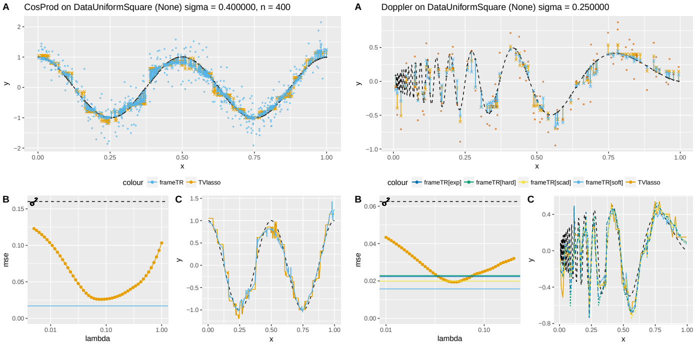


Comparison to Total variation denoising

Setup: test functions (not normalized) with specific sigmas, 1d, Total variation denoising:

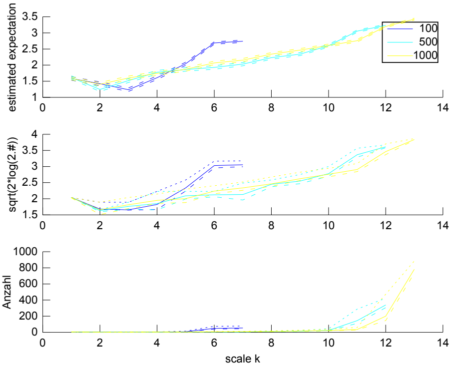
$$\hat{f}_{TV} \in \operatorname*{Arg\,Min}_{f \in \mathbb{R}^n} \frac{1}{n} \|f - y\|_2^2 + \lambda \|Wf\|_1$$

with W incidence matric: $L_{un} = W^t W = D - A$ (undir. unw. graph)



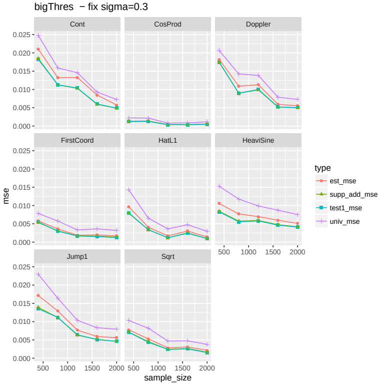
THRESHOLD - UNIVERSAL OR SCALE DEPENDENT?

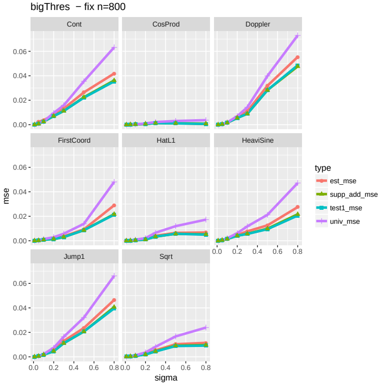
Q: How does $\sup_l |\langle \Psi_{kl}, \epsilon \rangle|$ behave for various k? Consider expectation

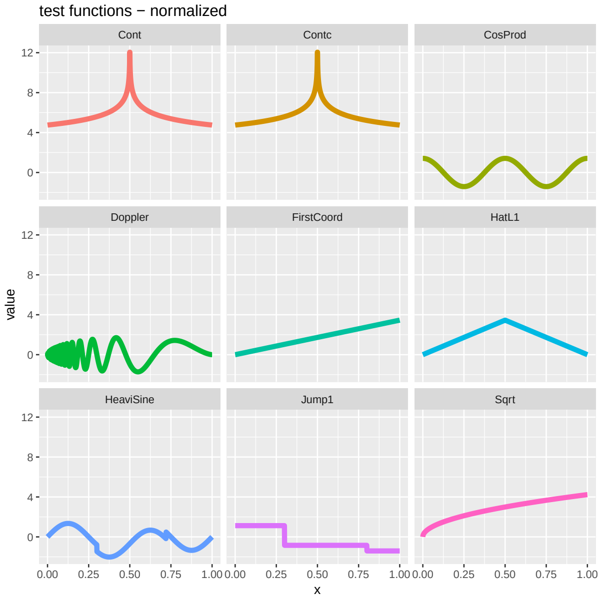


- try scale-dependent threshold

Comparision of different thresholding strategies and thresholds

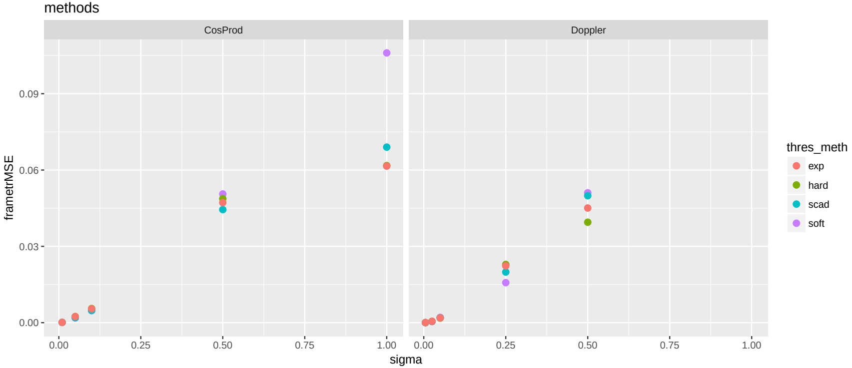





test functions − normalized
Cont
Contc
CosProd
Doppler
FirstCoord
HatL1
HeaviSine
Jump1
Sqrt
value
x
0
4
8
12
0.00
0.25
0.50
0.75
1.00

Soft, hard, SCAD, ...

n=100



Summary and Outlook

- Method to construct a Parseval frame exhibiting wavelet-like properties (multiscale, localised) while adapting to the instrinsic geometry of the data.
- This frame can be used in the denoising setting:
  simple coefficient thresholding method which satisfies an oracle-type inequality
  (with superior performance in simulations for denoising as compared to usual (spectral and non spectral) approaches)
- Doubling Condition and LPI hold whp for random ϵ-graph (under some assumptions)

- Extension of this methodology to semi-supervised learning setting?
- Proof of spatial localization?

Thank you.

Mira Bernstein, Vin De Silva, John C. Langford, and Joshua B. Tenenbaum.
Graph approximations to geodesics on embedded manifolds, 2000.

Thierry Coulhon, Gerard Kerkyacharian, and Pencho Petrushev.
Heat kernel generated frames in the setting of dirichlet spaces.
June 2012.

Persi Diaconis and Daniel Stroock.
Geometric bounds for eigenvalues of Markov chains.
The Annals of Applied Probability, 1(1):36–61, Feb 1991.

Ulrike von Luxburg, Agnes Radl, and Matthias Hein.
Hitting and commute times in large random neighborhood graphs.
Journal of Machine Learning Research, 15(1):1751–1798, 2014.