Additional Simulation, SVM for Functional Data Classification

Fabrice Rossi, Nathalie Villa

August 27, 2019

Contents

1 Penalized Spline Parameter(λ) Optimization

2 Paper Summary

1. Penalized Spline Parameter(λ) Optimization

- To reduce the MSE of smoothed curve, we need to reduce the sampling variance at the cost of some increase in bias
- One way is to find the curve that minimize the penalized residual sum of squares using roughness penalty parameter λ
- A natural measure of a function's roughness(PEN) is the integrated squared second derivative

$$PEN(x) = \int [D^2 x(s)]^2 ds$$

- Then, the penalized residual sum of squares is

$$PENSSE_\lambda(x \mid \boldsymbol{y}) = [\boldsymbol{y} - x(\boldsymbol{t})]' \boldsymbol{W} [\boldsymbol{y} - x(\boldsymbol{t})]^2 + \lambda \times PEN(x)$$

($\boldsymbol{y}$: observation, $x(\boldsymbol{t})$: basis function, $\boldsymbol{W}$: weight matrix)

1. Penalized Spline Parameter(λ) Optimization

- Proper λ can be found by the generalized cross-validation(GCV)

- The criterion is usually expressed as,

$$GCV(\lambda) = \frac{n^{-1}SSE}{[n^{-1}trace(\boldsymbol{I} - \boldsymbol{S_\lambda})]^2}$$

where $\boldsymbol{S_\lambda}$ is the smoothing operator

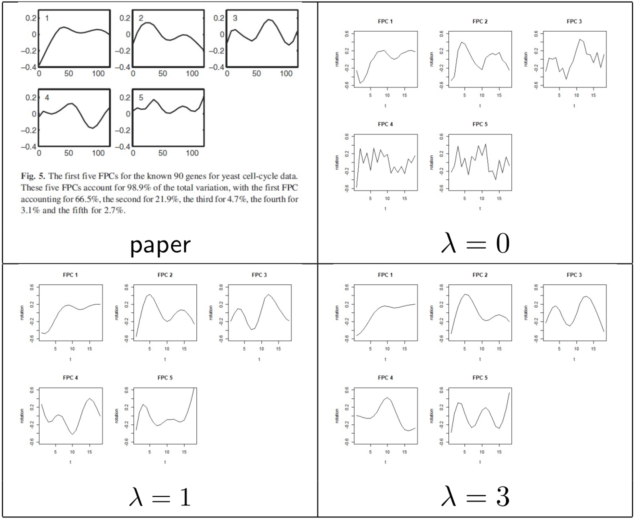
- With smooth.basisPar R function in fda package, appropriate λ was found to be 3

	lambda	df	gcv
1e-05	1e-05	3.995391	4100.560
1e-04	1e-04	3.955736	3801.695
0.001	1e-03	3.682808	2601.715
0.01	1e-02	3.164761	1984.170
0.1	1e-01	2.902563	1776.631
1	1e+00	2.428404	1441.917
10	1e+01	2.069369	1425.225
100	1e+02	2.007396	1433.415
1000	1e+03	2.000745	1434.435
10000	1e+04	2.000075	1434.539
1e+05	1e+05	2.000007	1434.549
1e+06	1e+06	2.000001	1434.550
1e+07	1e+07	2.000000	1434.550
1e+08	1e+08	2.000000	1434.550
1e+09	1e+09	2.000000	1434.550

	lambda	df	gcv
1	1	2.428404	1441.917
2	2	2.271926	1418.573
3	3	2.199204	1416.916
4	4	2.157175	1418.151
5	5	2.129792	1419.736
6	6	2.110536	1421.198
7	7	2.096255	1422.459
8	8	2.085242	1423.531
9	9	2.076491	1424.444
10	10	2.069369	1425.225

1. Penalized Spline Parameter(λ) Optimization

- The comparison of the five FPC curves(λ)



1. Penalized Spline Parameter(λ) Optimization

Table: Classification error rates($\lambda = 1$)

No. of FPCs or base functions	Group 1 FPCA	B-S	Group 2 FPCA	B-S	overall FPCA	B-S
1	32.72 (8.41)	27.32 (7.86)	32.70 (8.31)	26.90 (7.86)	32.71 (5.26)	27.11 (4.58)
2	22.16 (6.65)	24.08 (6.37)	22.06 (6.15)	24.80 (6.55)	22.11 (4.33)	24.44 (3.97)
3	7.58 (4.58)	7.92 (3.96)	8.26 (5.34)	8.76 (4.71)	7.92 (3.35)	8.34 (2.70)
4	7.14 (4.14)	8.18 (4.18)	7.62 (5.10)	8.98 (5.00)	7.38 (3.11)	8.58 (2.96)
5	7.40 (4.07)	7.68 (4.29)	7.86 (5.26)	8.58 (5.01)	7.63 (3.06)	8.13 (3.15)

Table: Classification error rates($\lambda = 3$)

No. of FPCs or base functions	Group 1 FPCA	B-S	Group 2 FPCA	B-S	overall FPCA	B-S
1	32.94 (8.39)	26.46 (7.84)	32.82 (8.57)	26.74 (6.98)	32.88 (5.33)	26.60 (4.57)
2	21.40 (8.39)	25.40 (6.98)	21.68(5.97)	25.54 (6.41)	21.54 (4.17)	25.47 (4.27)
3	8.68 (5.06)	25.14 (7.08)	9.04 (5.20)	25.18(7.06)	8.86 (3.54)	24.66 (4.36)
4	8.06 (4.54)	15.48 (6.70)	8.70 (5.02)	15.48 (5.70)	8.38 (3.20)	15.37 (4.18)
5	7.04 (4.06)	14.56 (6.13)	7.62 (5.15)	14.56 (6.42)	7.33 (3.01)	14.01 (4.20)

2. SVM for Functional Data Classification
2.1 Introduction

- This paper adapts support vector machines(SVM) to functional data classification
- Functional SVM takes advantages of using kernels considering the functional nature of the data
- Those kernels allow us to apply the expert knowledge on the data and construct a consistent training procedure
- An observation is an element of $L^2(\mu)$, the Hilbert space of μ-square-integrable real valued functions defined on $\mathbb{R}$ (μ: a known finite positive Borel measure on $\mathbb{R}$)
- That is, $L^2(\mu)$ is an arbitrary Hilbert space H with $\langle u, v\rangle = \int u(\mu)v(\mu)d\mu$

2.2 Support Vector Machines for FDA

The goal is to classify data, (x_1, y_1),..., (x_N, y_N), into two predefined classes, where X has values in Hilbert space $\mathbb{H}$ and Y in $\{-1, 1\}$

- Hard margin SVM
  - Find an element $w \in \mathbb{H}$ with the following conditions:

$$\min_{w,b} \langle w, w \rangle, \;\; y_i(\langle w, x_i \rangle + b) \geq 1, \; 1 \leq i \leq N, \;\; b : \text{a real value}$$

- Soft margin SVM
  - Modify the hard margin SVM allowing some classification errors(ξ_i):

$$\min_{w,b,\xi} \langle w, w \rangle + C \sum_{i=1}^{N} \xi_i, \;\; C : \text{regulation parameter}$$

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \; 1 \leq i \leq N, \;\; \xi_i \geq 0$$

2.2 Support Vector Machines for FDA

- Nonlinear SVM
  - Some classification problems don't have a satisfactory linear solution but have a nonlinear one
  - Nonlinear SVMs are obtained by transforming the original data(x_i) using a function ϕ

- Dual formulation and Kernels
  - According to C.-J. Lin, soft margin SVM can be optimized by following conditions:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$$

  - Now the optimization problem is reduced to N-dimension rather than infinite dimension
  - Above conditions can be transformed by Kernel function, $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$

2.3 Kernels for FDA

1. Classical kernels
   - Many standard kernels can be applied such as the Gaussian kernel: $K(u, v) = e^{\sigma \|u-v\|^2}$

2. Using the functional nature of the data
   - Apply the transformation operator P to the standard kernel K: $Q(u, v) = K(P(u), P(v))$

   - To reduce the dimensionality of the input space, one can use projection on d-dimensional subspace V_d with basis$\{\Psi_j\}_{1,...,d}$

   - The transformation P_{V_d} is defined as $P_{V_d}(x) = \sum_{j=1}^{d} \langle x, \Psi_j \rangle \Psi_j$
   - Then a standard $\mathbb{R}^d$ SVM can be applied to the transformed vector data $\langle x, \Psi_1 \rangle, ..., \langle x, \Psi_d \rangle$

   - Derivative transformation is also possible: $Q(u, v) = K(D(u), D(v))$

2.4 Consistency of Functional SVM

- The generalization error of a classifier f is defined as $L(f) = P(f(X) \neq Y)$

- The minimal generalization error is achieved by the optimal classifier, $f^*(x) = \begin{cases} 1 & \text{when } P(Y = 1 \mid X = x) > \frac{1}{2}, \\ -1 & \text{otherwise} \end{cases}$

- With a learning sample of size N, one can construct an algorithm which finds a classification rule f_N chosen from a set of admissible classifiers

- This algorithm is said to be consistent if $L(f_N) \xrightarrow{N \to +\infty} L^* = L(f^*)$

2.4 Consistency of Functional SVM

- A learning algorithm for functional SVM

(1) Choose several parameters consisting a set A:
the weights(α_i, b), the projection size d, the regularization parameter C, and the fully specified kernel K(Gaussian exponential, etc.)

(2) The data($(x_i, y_i), i = 1, ..., N$) are split into l_N training set and $N - l_N$ validation set

(3) For each $a \in A$, the training set is used to calculate the SVM classification rule $f_a(x) = \text{sign}(\sum_{i=1}^{l_N} \alpha_i^* y_i K(P_{V_d}(x), P_{V_d}(x_i)) + b^*)$, where α_i^*, b^* are the solution of the soft margin SVM

2.4 Consistency of Functional SVM

- A learning algorithm for functional SVM

  (4) The validation set is used to select the optimal value of a, a^* satisfying:

  $$\arg\min_{a \in A} \hat{L}(f_a) + \frac{\lambda_a}{\sqrt{N - l_N}},$$

  $$\text{where } \hat{L}(f_a) = \frac{1}{N - l_N} \sum_{n=l_N+1}^{N} I_{\{f_a(x_n) \neq y_n\}}$$

  and λ_a is a penalty term used to avoid selecting the most complex models(the highest d in general)

- This algorithm is proved to be consistent with several conditions

2.5 Applications

- Speech recognition
  - G.Biau *et al.* studied classifying speech samples("yes" vs "no", "boat" vs "goat", and "sh" vs "ao")

  - Each digitized speech function(data) is a vector in $\mathbb{R}^{8192}$

  - As the data have temporal patterns, the Fourier basis was used

  - The penalty term λ_d is 0 for $d \leq 100$, and a high value for $d > 100$

  - The error rates for functional SVM based methods are compared with those for k-nn and QDA(Quadratic Discriminant Analysis) based methods

2.5 Applications

- Speech recognition

Table 2
Error rate for reference methods for the speech recognition problem (leave-one out)

Problem	k-nn (%)	QDA (%)
Yes/no	10	7
Boat/goat	21	35
sh/ao	16	19

Table 3
Error rate for SVM based methods for the speech recognition problem (leave-one out)

Problem/ Kernel	Linear (direct) (%)	Linear (projection) (%)	Gaussian (projection) (%)
Yes/no	58	19	10
Boat/goat	46	29	8
sh/ao	47	25	12

- The functional SVM methods with projection show better classification

2.5 Applications

- Using wavelet basis
  - Another study is performed on the data of 32ms phonemes($\mathbb{R}^{256}$) from TIMIT database

  - The problem is classifying "aa" vs "ao": the training set is 519 samples for "aa", and 759 samples for "ao"

  - As the data are very noisy, a wavelet basis was used

Table 4
Error rate for all methods on the test set

Functional Gaussian SVM (%)	Functional linear SVM (%)	Linear SVM(%)
22	19.4	20

- Functional kernels are not as useful as in the previous application
- One possible reason is smaller dimension of the input space than the number of training set

2.5 Applications

- Spectrometric data set
  - The data are 100 channel spectrum of absorbances in the wavelength range 850-1050nm

  - The problem is to separate high fat(more than 20%) samples from low fat(less than 20%) samples

  - Functional SVMs with standard kernels(linear and Gaussian) are compared to functional SVMs with derivative based kernels

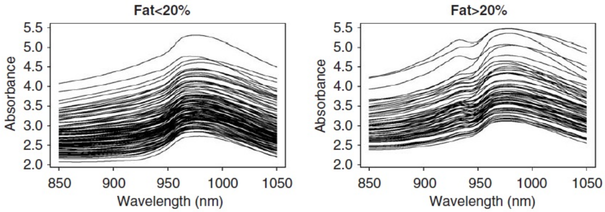


Fig. 1. Spectra for both classes.

2.5 Applications

- Spectrometric data set

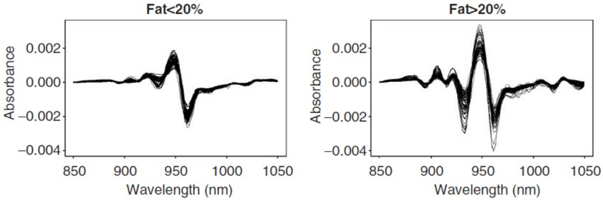


Fig. 2. Second derivatives of the spectra for both classes.

Table 5
Mean test error rate for all methods on the spectrometric dataset

Kernel	Mean test error (%)
Linear	3.38
Linear on second derivatives	3.28
Gaussian	7.5
Gaussian on second derivatives	2.6

- It appears that functional transformation(derivative) improves classification error rate

References I

B. Silverman J. Ramsay.
Functional data analysis.
Springer Series in Statistics, 2005.

Nathalie Villa Fabrice Rossi.
Support vector machine for functional data classification.
Neurocomputing, 69:730–742, 2006.

C.-J. Lin.
Formulations of support vector machines: a note from an optimization point of view.
Neural Comput.j, 2(13):307–317, 2001.

M. Wegkamp G. Biau, F. Bunea.
Functional classification in hilbert spaces.
IEEE Trans. Inf. Theory, 51:2163–2172, 2005.

Thank You!