CONSTRUCTIONS OF LOCAL ORTHONORMAL BASES FOR CLASSIFICATION AND REGRESSION

RONALD R. COIFMAN AND NAOKI SAITO

Abstract. We describe extensions to the "best-basis" method to construct orthonormal bases which either maximize a class separability for signal classification problems or minimize an estimation error for regression problems. These algorithms reduce the dimensionality of these problems by using basis functions which are well localized in time-frequency plane as feature extractors.

Constructions de Bases Orthonormées Locales de Classification et Régression

Résumé. Nous décrivons une extension de la méthode de la "meilleure base" qui permet de sélectionner une base orthonormée optimale qui maximise la séparation entre classes dans un problème de classification et minimise l'erreur d'estimation dans les problèmes de régression. Ces algorithmes réduisent la dimension de ces problèmes en utilisant des fonctions de base bien localisées en "temps-fréquence" pour extraire les traits charactéristiques des signaux.

Version française abrégée

Nous décrivons un algorithme rapide de sélection de base orthogonale adaptée à la classification. La méthode de Coifman et Wickerhauser [3] s'adapte facilement. En effet pour la classification il suffit d'avoir une mesure de discrimination qui permet de comparer l'efficacité relative de deux systèmes de coordonnées (restreints à un sous-espace). Ceci nous permet de remonter dans l'arbre des sous-espaces de la librairie de paquets d'ondelettes ou de bases trigonométriques locales, en choisissant à chaque branchement la discrimination la plus efficace.

Nous disposons de plusieurs choix de mesures de discrimination (voir ex., [1]). Considérons le cas de deux classes. Soit $\boldsymbol{p} = \{p_i\}_{i=1}^n$, $\boldsymbol{q} = \{q_i\}_{i=1}^n$ deux suites positives avec $\sum p_i = \sum q_i = 1$ (qui peuvent être interprétées comme des distributions d'énergies normalisées des signaux de la classe 1 ou 2 respectivement, dans un système de coordonnées). Une fonctionnelle de discrimination $\mathcal{D}(\boldsymbol{p}, \boldsymbol{q})$ entre ces deux suites devrait mesurer l'écart entre $\boldsymbol{p}$ et $\boldsymbol{q}$. Un choix naturel $\mathcal{D}$ est donné par *l'entropie relative* (*entropie croisée*, *distance de Kullback-Leibler*, ou *I-divergence*) [8]:

$$I(\boldsymbol{p}, \boldsymbol{q}) \stackrel{\Delta}{=} \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i},$$

nous convenons que $\log 0 = -\infty$, $\log(x/0) = +\infty$ pour $x \gneqq 0$, $0 \cdot (\pm\infty) = 0$. Il est clair que $I(\boldsymbol{p}, \boldsymbol{q}) \geq 0$ avec égalité si et seulement si $\boldsymbol{p} \equiv \boldsymbol{q}$. Cette quantité n'est pas une métrique. Pour obtenir un algorithme efficace de calcul numérique il est suffisant que la

mesure de discrimination soit *additive*, c-à-d devrait satisfaire pour tout j, $1 \leq j \leq n$,

$$\mathcal{D}(\{p_i\}_{i=1}^n, \{q_i\}_{i=1}^n) = \mathcal{D}(\{p_i\}_{i=1}^j, \{q_i\}_{i=1}^j) + \mathcal{D}(\{p_i\}_{i=j+1}^n, \{q_i\}_{i=j+1}^n).$$

Nous utilisons cette quantité pour construire la base locale de discrimination, en choisissant d'abord une librairie de bases et un ensemble de données d'apprentissage.
Étape 0: *Choisir un mode de décomposition temps-fréquence (paquets d'ondelettes ou bases trigonométriques locales).*
Étape 1: *Former une table d'énergie temps-fréquence de chaque classe. En associant à chaque sous-espace de l'arbre l'énergie normalisée de la classe.*
Étape 2: *Calculer la discrimination entre classes à chaque branchement de l'arbre.*
Étape 3: *A chaque branchement choisir le système de coordonnées le plus discriminant.*
Étape 4: *Ordonner les fonctions de bases par leur puissance de discrimination.*
Étape 5: *Garder les k ($\ll n$) fonctions de bases de plus grande discrimination.*

Un algorithme semblable pour la regression s'obtient en minimisant à chaque branchement de l'arbre l'erreur de regression.

Les deux tables du textes comparent l'efficacité des méthodes LDA [5], [6] et CART [2] utilisées sur l'échantillonage originel à l'efficacité de ces méthodes dans les coordonnées de la base locale de discrimination.

1. Introduction

Extracting relevant features from signals or images is an important process for data analysis, such as classifying signals into known categories (*classification*) or predicting a response of interest based on these signals (*regression*). In this paper, we focus our attention on methods of selection of coordinate systems to enhance the performance of a few classification schemes. The "best-basis" paradigm permits a rapid (e.g., $O(n \log n)$) search among a large collection of orthogonal bases to find that basis which extracts the most discriminating features for a classification problem or which permits the best approximation for a regression problem. Our method extracts the most significant coordinates which are then fed into a traditional classifier such as Linear Discriminant Analysis (LDA) of R. A. Fisher [5] (see also [6]), or a Classification and Regression Tree (CART) [2].

2. Construction of Local Discriminant Basis

In this section, we describe a fast algorithm to construct an adaptive orthonormal basis which is localized in the time-frequency plane and which discriminates given signal classes.

The *best-basis* algorithm of Coifman and Wickerhauser [3] was developed mainly for signal compression. This method first expands a given signal or a given collection of signals into a redundant set of wavelet packet bases or local trigonometric bases having a binary tree structure where the nodes of the tree represent subspaces with different time-frequency localization characteristics. Then a complete basis called a *best basis* which minimizes a certain information cost functional (e.g., entropy) is searched in this binary tree using the divide-and-conquer algorithm. This cost functional measures the flatness of the energy distribution of the signal so that minimizing this leads to an efficient representation (or coordinate system) for the signal. Because of this cost functional,

the best-basis algorithm is good for signal compression but is not necessarily good for classification problems: for classification, we need a measure to evaluate the power of discrimination of the nodes (or subspaces) in the tree-structured bases.

There are many choices for the discriminant measure (see e.g., [1]); all of them essentially measure "statistical distances" among classes. For simplicity, let us first consider the two-class case. Let $\boldsymbol{p} = \{p_i\}_{i=1}^n$, $\boldsymbol{q} = \{q_i\}_{i=1}^n$ be two nonnegative sequences with $\sum p_i = \sum q_i = 1$ (which can be viewed as normalized energy distributions of signals belonging to class 1 and class 2 respectively in a coordinate system). The discriminant information functional $\mathcal{D}(\boldsymbol{p}, \boldsymbol{q})$ between these two sequences should measure how differently $\boldsymbol{p}$ and $\boldsymbol{q}$ are distributed. One natural choice for $\mathcal{D}$ is the so-called *relative entropy* (also known as *cross entropy*, *Kullback-Leibler distance*, or *I-divergence*) [8]:

$$I(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}, \tag{1}$$

with the convention, $\log 0 = -\infty$, $\log(x/0) = +\infty$ for $x \gneqq 0$, $0 \cdot (\pm\infty) = 0$. It is clear that $I(\boldsymbol{p}, \boldsymbol{q}) \geq 0$ and equality holds iff $\boldsymbol{p} \equiv \boldsymbol{q}$. This quantity is not a metric since it is not symmetric and does not satisfy the triangle inequality. But it measures the discrepancy of $\boldsymbol{p}$ from $\boldsymbol{q}$. Note that if $q_i = 1/n$ for all i, i.e., q_i are distributed uniformly, then $I(\boldsymbol{p}, \boldsymbol{q}) = -H(\boldsymbol{p})$, the negative of the entropy of the sequence $\boldsymbol{p}$ itself.

Remark. The relative entropy (1) has an important interpretation for denoising applications. Let class 1 consist of a signal plus noise or a signal plus "background" and let class 2 consist of a pure noise or "background" (not necessarily a random signal). Then, by selecting a basis maximizing $\mathcal{D}$ between class 1 and class2, we can construct the best basis for denoising arbitrary noise or pulling a signal out of a textured background. In this context, an asymmetric measure such as (1) makes sense. If, however, a symmetric quantity is preferred, one should use the *J-divergence* between $\boldsymbol{p}$ and $\boldsymbol{q}$ [8]:

$$J(\boldsymbol{p}, \boldsymbol{q}) \triangleq I(\boldsymbol{p}, \boldsymbol{q}) + I(\boldsymbol{q}, \boldsymbol{p}). \tag{2}$$

To obtain a fast computational algorithm, it is convenient if the measure $\mathcal{D}$ is *additive*: for any j, $1 \leq j \leq n$,

$$\mathcal{D}(\{p_i\}_{i=1}^n, \{q_i\}_{i=1}^n) = \mathcal{D}(\{p_i\}_{i=1}^j, \{q_i\}_{i=1}^j) + \mathcal{D}(\{p_i\}_{i=j+1}^n, \{q_i\}_{i=j+1}^n). \tag{3}$$

The measures (1) and (2) are both additive. In Section 3, we will consider a non-additive measure.

For measuring discrepancies among L distributions, $\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(L)}$, one may take $\binom{L}{2}$ pairwise combinations of $\mathcal{D}$:

$$\mathcal{D}(\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(L)}) \triangleq \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \mathcal{D}(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)}). \tag{4}$$

Now we use this quantity to construct a good local orthonormal basis for classifying signals (vectors) of length n into L classes:

Algorithm 1. *Given a training dataset,*
Step 0: *Choose a time-frequency decomposition method (i.e., choose which wavelet packet or local cosine/sine basis is to be used).*

Step 1: *Construct a time-frequency energy map for each class by: normalizing each signal by the total energy of all signals of that class, expanding that signal into the tree-structured subspaces, and accumulating the energy at each coefficient position.*
Step 2: *At each node, compute the discriminant measure among L time-frequency energy maps.*
Step 3: *Prune the binary tree: eliminate children nodes if the sum of their discriminant measures is smaller than or equal to the discriminant measure of their parent node.*
Step 4: *Order the basis functions by their power of discrimination (see below).*
Step 5: *Use $k\,(\ll n)$ most discriminant basis functions for constructing classifiers.*

Step 3 produces a complete basis which permits us to select good coordinates for classification in terms of time-frequency energy distributions; we call this basis the *local discriminant basis* (LDB).

In Step 4, the power of discrimination of an individual basis function must be measured. There are several choices for this measure: 1) the discriminant measure of a single basis function: $\mathcal{D}(\alpha_i^{(1)}, \ldots, \alpha_i^{(L)})$, where $\alpha_i^{(l)}$ is a coefficient of the time-frequency energy map of class l, 2) the Fisher's class separability: the ratio of the between-class variance to the within-class variance using the inner products between the signals and a single basis function, or 3) the robust version of 2) obtained by replacing mean by median and variance by median absolute deviation. See [1], [7] for details and more examples. We note that this step can also be viewed as a restricted version of the projection pursuit algorithm [7].

Step 5 reduces the dimensionality of the problem from n to k. How to select the best k is a tough interesting question. One possibility is to use model selection methods such as the minimum description length (MDL) criterion [9].

3. Construction of Local Regression Basis

In this section, we describe an algorithm to construct an adaptive orthonormal basis which is localized in the time-frequency plane and which minimizes a measure of prediction error (such as ℓ^2 error) for the regression problem. In contrast with the LDB algorithm of the previous section where the statistical (classification) method is used after the basis selection, the algorithm described in this section integrates the statistical (regression) method into the basis selection mechanism.

Algorithm 2. *Given a training dataset,*
Step 0: *Choose a time-frequency decomposition method (i.e., choose which wavelet packet, or local cosine/sine basis is to be used).*
Step 1: *Expand each signal into the tree-structured subspaces.*
Step 2: *At each node, invoke a regression method $\mathcal{R}$, fit a model, and then compute the residual error between the given response vector and the prediction using the expansion coefficients in this node.*
Step 3: *Prune the binary tree: eliminate children nodes if the prediction error computed from the union of the coefficients at these nodes (using the same method $\mathcal{R}$) is larger than that of their parent node.*
Step 4: *Use $k\,(\ll n)$ most important basis functions for the problem at hand.*

Step 3 produces a complete basis which gives the smallest prediction error (using $\mathcal{R}$) in the set of all possible bases obtainable by the divide-and-conquer algorithm; we call this basis the *local regression basis* (LRB) with respect to $\mathcal{R}$.

Also in Step 3, note that the prediction error computed from the union of the two nodes is not equal to the sum of the individual errors at these nodes in general since the prediction error is not additive in the sense of (3).

Step 4 is the so-called "selection-of-variables" problem. Again the MDL criterion [9] might be a good candidate for obtaining the optimal k.

4. Examples

To demonstrate the capability of the local discriminant basis, we conducted two classification experiments using synthetic signals. In both cases, we specified three classes of signals by analytic formulas. For each class, we generated 100 training signals and 1000 test signals. We first applied the LDA and Classification Tree (CT) to the training signals of the original coordinate (i.e., standard Euclidean) system, and obtained the classification rules. Then the test signals were fed into these classifiers and the misclassification rates were computed. Next we computed the LDB (using the relative entropy as $\mathcal{D}$) from the training signals, selected a small number of most discriminant basis functions (in terms of the component-wise relative entropy), and applied the LDA and CT to the resulting coefficients. Finally the test signals were projected onto these basis functions and fed into these classifiers; then the misclassification rates were computed. For each method, we also computed the misclassification rate on the training dataset. The details of the experiments using real datasets as well as the examples of LRB can be found in [10].

Example 1. *Triangular waveform classification.* This is an example for classification originally examined in [2]. The dimensionality of the signal was extended from 21 in [2] to 32 for the dyadic dimensionality requirement of the bases under consideration. Three classes of signals were generated by the following formulas:

$$\begin{aligned} x^{(1)}(i) &= uh_1(i) + (1-u)h_2(i) + \epsilon(i) \quad \text{for Class 1,} \\ x^{(2)}(i) &= uh_1(i) + (1-u)h_3(i) + \epsilon(i) \quad \text{for Class 2,} \\ x^{(3)}(i) &= uh_2(i) + (1-u)h_3(i) + \epsilon(i) \quad \text{for Class 3,} \end{aligned}$$

where $i = 1, \ldots, 32$, $h_1(i) = \max(6 - |i-7|, 0)$, $h_2(i) = h_1(i-8)$, $h_3(i) = h_1(i-4)$, u is a uniform random variable on the interval $(0, 1)$, and $\epsilon(i)$ are standard normal variates. The LDB was computed from the wavelet packet coefficients with the 6-tap coiflet filter [4]. Then the five most discriminant coordinates were selected. These five basis functions look similar to the functions h_j or their derivatives. The misclassification rates are given in the table:

Method	Training Data	Test Data
LDA on the standard coordinate system	13.33 %	20.90 %
CT on the standard coordinate system	6.33 %	29.87 %
LDA to Top 5 LDB coordinates	14.33 %	15.90 %
CT to Top 5 LDB coordinates	7.00 %	21.37 %

The best result so far was obtained using the LDA on the LDB coordinates. We would like to note that according to Breiman et al. [2], the Bayes error of this example is about 14 %.

Example 2. *Signal shape classification.* The second example is a signal shape classification problem. In this example, we try to classify synthetic noisy signals with various amplitudes, lengths, and positions into three possible classes. More precisely, signals of three classes were generated by:

$$\begin{aligned} c(i) &= (6+\eta)\cdot\chi_{[a,b]}(i) + \epsilon(i) && \text{for "cylinder" class,} \\ b(i) &= (6+\eta)\cdot\chi_{[a,b]}(i)\cdot(i-a)/(b-a) + \epsilon(i) && \text{for "bell" class,} \\ f(i) &= (6+\eta)\cdot\chi_{[a,b]}(i)\cdot(b-i)/(b-a) + \epsilon(i) && \text{for "funnel" class,} \end{aligned}$$

where $i = 1, \ldots, 128$, a is an integer-valued uniform random variable on the interval $[16, 32]$, $b - a$ also obeys an integer-valued uniform distribution on $[32, 96]$, η and $\epsilon(i)$ are standard normal variates, and $\chi_{[a,b]}(i)$ is the characteristic function on the interval $[a, b]$. The 12-tap coiflet filter [4] was used for the LDB selection. Then the 10 most important coordinates were selected. They captured local features of these signals such as edge positions and type of edges. The misclassification rates in this case are:

Method	Training Data	Test Data
LDA on the standard coordinate system	0.33 %	13.17 %
CT on the standard coordinate system	3.00 %	13.37 %
LDA to Top 10 LDB coordinates	3.67 %	6.20 %
CT to Top 10 LDB coordinates	3.00 %	3.83 %

As expected, the LDA applied to the original coordinate system was almost perfect with respect to the training data, but it adapted too much to the training data, so it lost flexibility; when applied to the new test dataset, it did not work well. The best result was obtained using the CT on the LDB coordinates in this case.

5. Conclusion

We have described two algorithms to construct adaptive local orthonormal bases for classification and regression problems. The basis functions generated by these algorithms can capture relevant local features (in both time and frequency) in data. These bases provide us with better insight and understanding of relationships between the essential features of the input signals and the corresponding outputs (class names or response values), and permit us to build rudimentary data-driven models. Therefore, they can enhance both traditional and modern statistical methods.

References

1. M. Basseville, *Distance measures for signal processing and pattern recognition*, Signal Processing **18** (1989), no. 4, 349–369.
2. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman & Hall, Inc., New York, 1993, previously published by Wadsworth & Brooks/Cole in 1984.
3. R. R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 713–719.

4. I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, SIAM, Philadelphia, 1992.
5. R. A. Fisher, *The use of multiple meaurements in taxonomic problems*, Ann. Eugenics **7** (1936), 179–188.
6. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, San Diego, CA, 1990.
7. P. J. Huber, *Projection pursuit*, Ann. Statist. **13** (1985), no. 2, 435–525, with discussions.
8. S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Statist. **22** (1951), 79–86.
9. J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.
10. N. Saito, *Local Feature Extraction and Its Applications Using a Library of Bases*, Ph.D. thesis, Dept. of Mathematics, Yale University, New Haven, CT 06520 USA, 1994, in preparation.

(R.R.C.) Department of Mathematics, Yale University, 10 Hillhouse Avenue, New Haven, CT 06520

(N.S.) Schlumberger-Doll Research, Old Quarry Road, Ridgefield, CT 06877
and Department of Mathematics, Yale University, 10 Hillhouse Avenue, New Haven, CT 06520