# ON LOCAL ORTHONORMAL BASES FOR CLASSIFICATION AND REGRESSION

*Naoki Saito*

Schlumberger-Doll Research
Old Quarry Road
Ridgefield, CT 06877
saito@ridgefield.sdr.slb.com

*Ronald R. Coifman*

Department of Mathematics
Yale University
New Haven, CT 06520
coifman@math.yale.edu

## ABSTRACT

We describe extensions to the "best-basis" method to select orthonormal bases suitable for signal classification and regression problems from a large collection of orthonormal bases. For classification problems, we select the basis which maximizes relative entropy of time-frequency energy distributions among classes. For regression problems, we select the basis which tries to minimize the regression error. Once these bases are selected, the most significant coordinates are fed into a traditional classifier or regression method such as Linear Discriminant Analysis (LDA) or Classification and Regression Tree (CART$^{TM}$). The performance of these statistical methods is enhanced since the proposed methods reduce the dimensionality of the problems without losing important information for the problem at hand by using the basis functions which are well-localized in the time-frequency plane as feature extractors. Finally, we compare their performance with the traditional methods using a synthetic example.

## 1. INTRODUCTION

Extracting relevant features from signals is important for signal analysis such as classification or regression (prediction). Often, important features for these problems, such as edges, spikes, or transients, are characterized by local information in the time (space) domain and the frequency (wave number) domain. The *best-basis* algorithm of Coifman and Wickerhauser [1] was developed to extract such local information mainly for signal compression. This method first expands a given signal into a *dictionary* of orthonormal bases, i.e., a redundant set of wavelet packet bases or local sine/cosine bases having a binary tree structure. 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 function (e.g., entropy) is searched in this binary tree using the divide-and-conquer algorithm. This cost function 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 function, the best-basis algorithm is good for signal compression but is not necessarily good for classification or regression problems.

## 2. LOCAL DISCRIMINANT BASES

### 2.1. Measures of Discrimination Power

For classification, we need a measure to evaluate the discrimination power of the nodes (i.e., subspaces) in the tree-structured bases. There are many choices for the discriminant measure $\mathcal{D}$ (see e.g., [2]). One natural choice is the so-called *relative entropy* (also known as *cross entropy* or *Kullback-Leibler distance*). 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). Then, relative entropy is defined as:

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

If a symmetric quantity is preferred, one can use the *J-divergence* between $\boldsymbol{p}$ and $\boldsymbol{q}$:

$$J(\boldsymbol{p}, \boldsymbol{q}) \stackrel{\Delta}{=} D(\boldsymbol{p}, \boldsymbol{q}) + D(\boldsymbol{q}, \boldsymbol{p}).$$

The measures $D$ and $J$ are both *additive*: for any $j$, $1 \leq j \leq n$,

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

For measuring discrepancies among $L$ distributions, a simple way is to take $\binom{L}{2}$ pairwise combinations of $\mathcal{D}$.

### 2.2. The Local Discriminant Basis Algorithm

The following algorithm selects an orthonormal basis (from the dictionary) which maximizes the discriminant measure on the time-frequency energy distributions of classes. We call this a *local discriminant basis* (LDB).

**Algorithm 1** *Given $L$ classes of training signals,*
Step 0: *Choose a dictionary of orthonormal bases (i.e., specify QMFs for a wavelet packet dictionary or decide to use either the local cosine dictionary or the local sine dictionary).*
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 signal energy in each coordinate.*
Step 2: *At each node, compute the discriminant measure $\mathcal{D}$ 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 discrimination power and use $k\,(\ll n)$ most discriminant basis vectors for constructing classifiers.* Step 5: *Use $k\,(\ll n)$ most discriminant basis vectors for constructing classifiers.*

The selection (or pruning) process in Step 3 is fast, i.e., $O(n)$ since the measure $\mathcal{D}$ is additive. After this step, we have a complete orthonormal basis LDB.

**Proposition 1** *The basis obtained by Step 3 of Algorithm 1 maximizes the additive discriminant measure $\mathcal{D}$ on the time-frequency energy distributions among all the bases in the dictionary obtainable by the divide-and-conquer algorithm.*

See [3] for the proof.

## 3. LOCAL REGRESSION BASES

For regression problems, we need a different measure to access the goodness of the subspaces. Here, we use regression (or prediction) error as a criterion: the smaller the error using a chosen regression method on the data belonging to a subspace, the better that subspace is. In particular, we use *residual sum of squares* (i.e., $\ell^2$ error); however, one may use other type of error measure as well, e.g., $\ell^1$ error. As a regression method used at each subspace, we use CART [4] in this paper. Again, one may use any other type of regression method such as linear regression, artificial neural networks, etc. We note that the prediction error computed from the union of the two subspaces is not equal to the sum of the individual errors at these subspaces in general since the prediction error is not an additive measure (this argument is similar to Cover [5]). 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 dictionary of orthonormal bases (i.e., specify QMFs for a wavelet packet dictionary or decide to use either the local cosine dictionary or the local sine dictionary).*
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.*

We refer to the basis thus obtained as the *local regression basis* (LRB) relative to $\mathcal{R}$. Unlike LDB, LRB may not give the smallest prediction error (using $\mathcal{R}$) in the set of all possible bases obtainable by the divide-and-conquer algorithm from the dictionary. This is not because the prediction error is non-additive but because the best prediction error of the union of the two individually-best subspaces may not be necessarily smaller than the best prediction error of the union of the two subspaces each of which is not individually-best by itself. In this sense, the LRB is still a first step toward the general regression problem using the best-basis paradigm. Step 4 is the so-called "selection-of-variables" problem. The MDL criterion [6] may be a good candidate for obtaining the optimal $k$.

## 4. AN EXAMPLE

We tested our methods using the triangular waveform classification often referred to as "waveform" described in [4]. The dimensionality of the signal was extended from 21 in [4] to 32 for the dyadic dimensionality requirement of the bases under consideration. We generated 100 training signals and 1000 test signals for each

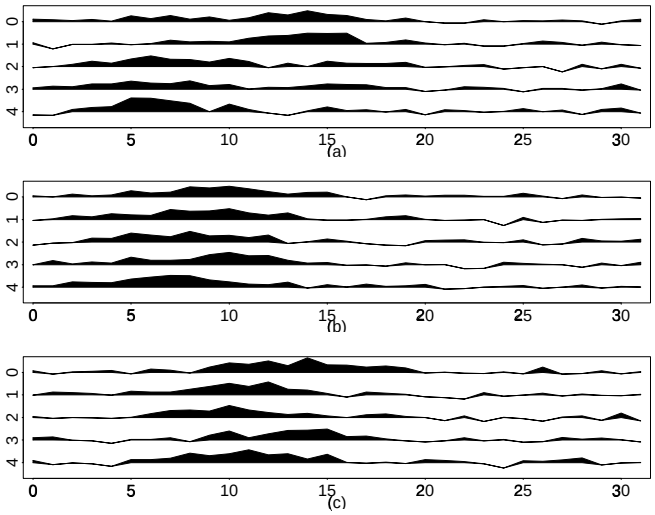

Figure 1: Five sample waveforms from (a) Class 1, (b) Class 2, and (c) Class3.

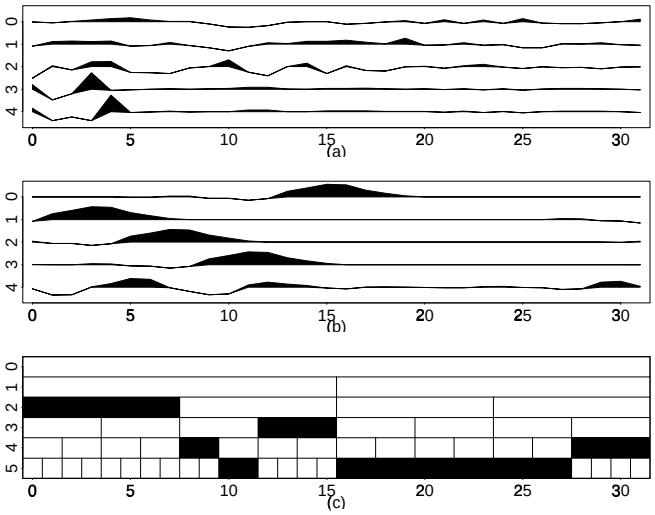

Figure 2: Plots from the analysis of the example "waveform": (a) Top five LDA vectors. (b) Top 5 LDB vectors. (c) The subspaces selected as the LDB.

class by the following formula:

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

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 the standard normal variates. Figure 1 shows five sample waveforms from each class. We first constructed LDA-based classifier and Classification Tree (CT) (with and without pruning) using the training signals represented in the original coordinate (i.e., standard Euclidean) system, and computed resubstitution (or apparent) error rates. We used the pruning algorithm based on the MDL principle described in [3]. Then we fed the test signals into these classifiers and computed the error rates. Next we computed the LDB (using the 6-tap coiflet filter [7] and asymmetric relative entropy) as a discriminant measure using the training signals. Then we selected five individually-most-discriminant basis functions, and used these coordinates to construct LDA-based classifier and CTs. Finally the test signals were projected onto these selected LDB functions and then fed into these classifiers. In Figure 2, we compare the top five vectors from LDA and LDB. Only top two vectors were useful in LDA in this case. The top five LDB vectors look similar to the functions $h_j$ or their derivatives whereas it is difficult to interpret the LDA vectors. For LRB, we used CT and misclassification rates as a regression method $\mathcal{R}$ and regression errors, respectively. At each subspace, we examined two possibilities: one is to use the fully-grown CT, the other is to use the pruned CT. Also, for the final over all classification given the LRB coordinates, we applied both the fully-grown and pruned CTs. The best LRB result was obtained by the fully-grown CT on all the LRBP coordinates. Here LRBP means the LRB selected by Algorithm 2 with pruned CT as $\mathcal{R}$ for subspace evaluation. Figure 3 shows these selected coordinates as well as the subspace pattern.

The misclassification rates are summarized in Table 1. The best result so far was obtained by applying LDA to the top 5 LDB coordinates. We note that according to Breiman et al. [4], the Bayes error of this example is about 14 %. Comparing with the LDB and LRB methods from these results, we observe the following: (1) The misclassification rates except the one by the LDA-based classification in Table 1 are comparable, and (2) seven functions out of 11 selected LRB functions have larger scale features than the top 5 LDB functions. In fact the LRB functions try to combine the elementary triangular waves $h_1, h_2, h_3$, e.g., the LRB function #6 has two major positive peaks around the functions $h_1$ and $h_2$ and a major negative peak around $h_3$.

The details as well as other examples and applications of LDB/LRB can be found in [3], [8], and [9].

## 5. CONCLUSION

We have described two algorithms to construct adaptive local orthonormal bases for classification and re-

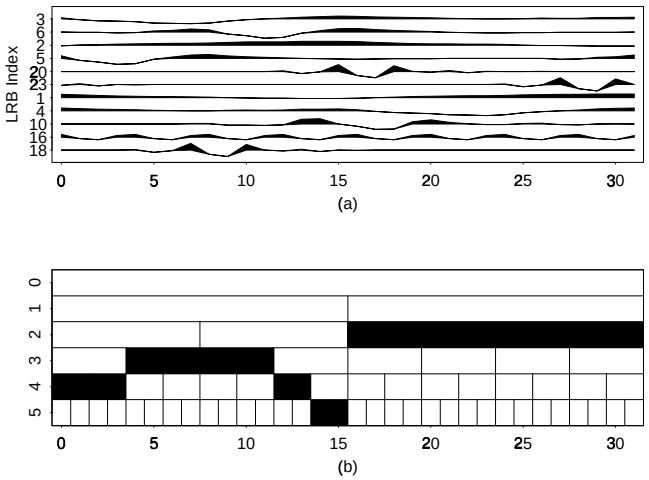


Figure 3: (a) The best LRB functions for the example "waveform." (b) The selected subspaces as the LRB.

| Method | Error rate (%) Training | Test |
|---|---|---|
| LDA on STD | 13.33 | 20.90 |
| FCT on STD | 6.33 | 29.87 |
| PCT on STD | 29.33 | 32.97 |
| LDA on LDB5 | 14.33 | **15.90** |
| FCT on LDB5 | 7.00 | 21.37 |
| PCT on LDB5 | 17.00 | 25.10 |
| FCT on LDB | 7.33 | 23.60 |
| PCT on LDB | 17.00 | 25.10 |
| FCT on LRBF | 4.33 | 24.33 |
| PCT on LRBF | 17.00 | 25.10 |
| FCT on LRBP | 4.33 | 22.13 |
| PCT on LRBP | 16.67 | 25.00 |

Table 1: Misclassification rates of the example "waveform". In Method column, FCT and PCT denote the full and pruned classification trees, respectively. STD, LDB5, and LDB represent the standard Euclidean coordinates, the top 5 LDB coordinates, and all the LDB coordinates, respectively. LRBF and LRBP represent all the LRB coordinates obtained by the subspace evaluation using FCTs and PCTs, respectively. We do not show the error rates of LDA on all the LDB coordinates since this is the same as the ones of LDA on STD theoretically. The smallest error on the test dataset is shown in bold font.

gression 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. The LDB method is computationally faster and generated a better result using the specific; however, the LRB method is much more flexible and general than the LDB method although it is more computationally intensive than the LDB method.

## 6. REFERENCES

[1] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", IEEE Trans. Inform. Theory, vol. 38, no. 2, pp. 713–719, 1992.

[2] M. Basseville, "Distance measures for signal processing and pattern recognition", Signal Processing, vol. 18, no. 4, pp. 349–369, 1989.

[3] 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.

[4] 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.

[5] T. M. Cover, "The best two independent measurements are not the two best", IEEE Trans. Syst. Man Cybern., vol. SMC-4, no. 1, pp. 116–117, 1974.

[6] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[7] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, SIAM, Philadelphia, 1992.

[8] R. R. Coifman and N. Saito, "Constructions of local orthonormal bases for classification and regression", Comptes Rendus Acad. Sci. Paris, Série I, vol. 319, no. 2, pp. 191–196, 1994.

[9] N. Saito and R. R. Coifman, "Local discriminant bases", in *Mathematical Imaging: Wavelet Applications in Signal and Image Processing* (A. F. Laine and M. A. Unser, eds.), Proc. SPIE 2303, pp. 2–14, 1994.