Local wavelet-vaguelette-based functional logistic regression for classification of gene expression data

M.M Rincón[1], M.D. Ruiz-Medina[2,*]

[1] *mmrh@ugr.correo.es*
[2] *mruiz@ugr.es*
[*] *Corresponding author*

Abstract. *This paper focuses on the problem of functional statistical classification of gene expression curves. A local wavelet-vaguelette-based functional logistic regression approach is presented. This approach offers an alternative to the Functional-Principal-Component-Analysis-based logistic regression (see [4]). The performance of the methodology proposed is illustrated by implementing it for classification of yeast cell-cycle temporal gene expression data from [5] data set, where leave-one-out cross-validation error shows high accuracy of the model.*

Keywords. *Functional data; Functional logistic regression; Gene expression profile; Local wavelet-vaguelette decomposition; Yeast cell cycle gene expression data.*

1. INTRODUCTION

Functional wavelet bases have been widely used in the analysis of fractal biological signals, since their provide a localized multiscale decomposition of such signals. The wavelet transform of a random biological signal $\{X(t),\ t \in \mathbb{R}\}$ leads to a sequence of correlated random wavelet coefficients. To avoid redundancy in such coefficients a local version of the wavelet-vaguelette decomposition of a random signal is considered (see [1]), to obtain suitable response variables for a functional logistic regression, providing low-error rate classification for the yeast cell-cycle gene expression profiles analyzed.

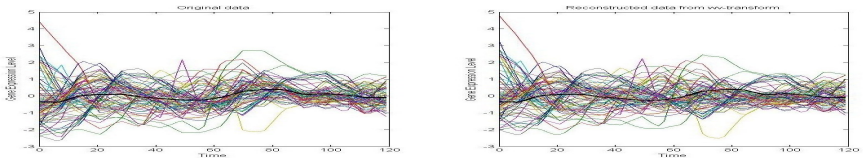

Figure 1: Left panel: Temporal gene expression profiles of yeast cell cycle. Right panel: Reconstruction of the temporal gene expression profiles in left panel from wavelet-vaguelette transform.

2. MODELS AND METHODS

The sample curves are assumed to be independent realizations of a mean-square integrable stochastic process $X(t)$ on $[0,S]$. Let $X_i(t_h)$ be the observation of the ith sample function at time t_h, for $h=1,\ldots,n$, and $i=1,\ldots,M$. Non-parametric kernel-based estimators, $\widehat{\mu}(t)$ and $\widehat{C}_X(s,t)$, are computed from a grid with $N=2^p$, $p\in\mathbb{N}$, equally spaced points in $[0,S]$. Diagonal elements $\sigma^2(s)=\widehat{C}_X(s,s)$, $s\in[0,S]$, are approximated by interpolated values $\widehat{\sigma}^2(s)$, $s\in[0,S]$.

Multiresolution-like Analysis

The empirical eigenvalues $\widehat{\lambda}_l$, $l=1,\ldots,N$, and the corresponding empirical eigenvectors $(\widehat{\rho}_l(t_1),\ldots,\widehat{\rho}_l(t_N))$, $l=1,\ldots,N$, of the covariance estimate $\widehat{C}=\widehat{C}(t_l,t_m)$, $l,m=1,\ldots,N$, allow us to define the empirical kernel $\widehat{t}_X$, factorizing the covariance function $\widehat{C}_X(s,t)$, and the empirical kernel $\widehat{l}_X$, approximating the inverse $L_X=\mathcal{T}_X^{-1}$ of operator $\mathcal{T}_X$, respectively as follows:

$$\widehat{t}_X(t_h,t_m)=\sum_{l=1}^{N}\widehat{\lambda}_l^{1/2}\widehat{\rho}_l(t_h)\widehat{\rho}_l(t_m),\quad \widehat{l}_X(t_h,t_m)=\sum_{l=1}^{N}\widehat{\lambda}_l^{-1/2}\widehat{\rho}_l(t_h)\widehat{\rho}_l(t_m). \tag{1}$$

for $h,m=1,\ldots,N$. The construction of the empirical wavelet-vaguelette functions is given in terms of kernels $\widehat{t}_X$ and $\widehat{l}_X$, and a given orthonormal wavelet basis. We have chosen Haar system, with *the father wavelet*, $\phi(x)=I_{[0,1)}(x)$, and *the mother wavelet*, $\psi(x)=I_{[0,1/2)}(x)-I_{[1/2,1)}(x)$. Thus,(see [1],[6]), for $h=1,...,N$,

$$\widehat{\varphi}_0(t_h)=\sum_{m=1}^{N}\widehat{t}_X(t_h,t_m)\phi(t_m),\quad \widehat{\gamma}_{j,k}(t_h)=\sum_{m=1}^{N}\widehat{t}_X(t_h,t_m)\psi_{j,k}(t_m),\quad k=0,\ldots,2^j-1,\ j=0,\ldots,p-1.$$

In matrix form, we denote by $\varphi_0=\{a_h\}$ the vector with entries $a_h=\widehat{\varphi_0}(t_h)$, for $h=1,\ldots,N$, given by the product of the matrix $\widehat{T}=\{b_{h,m}\}$, with $b_{h,m}=\widehat{t}_X(t_h,t_m)$, for $h,m=1,\ldots,N$, and the vector $\Phi=\{c_m\}$, with $c_m=\phi(t_m)$, for $m=1,\ldots,N$. Similarly, for $j=0,\ldots,p-1$, the matrix $\Gamma_j=\{d_{h,k+1}\}$, with entries $d_{h,k+1}=\widehat{\gamma}_{j,k}(t_h)$, for $h=1,\ldots,N$, and $k=0,\ldots,2^j-1$, is the product of matrices $\widehat{T}$ and Ψ_j, where $\Psi_j=\{l_{m,k+1}\}$ has entries $l_{m,k+1}=\psi_{j,k}(t_m)$, for $m=1,\ldots,N$, and $k=0,\ldots,2^j-1$. Additionally, we have $\varphi^0=[\widehat{T}^{-1}]^T\times\Phi$ and $\Gamma^j=[\widehat{T}^{-1}]^T\times\Psi_j$.

For each sample curve X_i, evaluated at time $t\in[0,S]$, the following local empirical coefficients are computed:

$$\widehat{X}_i^{\widehat{\sigma}(t);\widehat{\varphi}^0}=\sum_{m=1}^{N}(X_i(t_m)-\widehat{\mu}(t_m))\widehat{\varphi}^{\widehat{\sigma}(t);0}(t_m)\quad \widehat{X}_i^{\widehat{\sigma}(t);j,k,\widehat{\gamma}}=\sum_{m=1}^{N}(X_i(t_m)-\widehat{\mu}(t_m))\widehat{\gamma}^{\widehat{\sigma}(t);j,k}(t_m)$$

where $\{\widehat{\varphi}^{\widehat{\sigma}(t);0},\widehat{\gamma}^{\widehat{\sigma}(t);j,k},\ k=0,\ldots,2^j-1,\ j=0,\ldots,p-1\}=\{\widehat{\sigma}(t)\widehat{\varphi}^0,\widehat{\sigma}(t)\widehat{\gamma}^{j,k},\ k=0,\ldots,2^j-1,\ j=0,\ldots,p-1\}$ denotes the locally re-scaled empirical dual Riesz basis of $\{\widehat{\varphi}_{\widehat{\sigma}(t);0},\widehat{\gamma}_{\widehat{\sigma}(t);j,k},\ k=0,\ldots,2^j-1,\ j=0,\ldots,p-1\}=\{(1/\widehat{\sigma}(t))\widehat{\varphi}_0,(1/\widehat{\sigma}(t))\widehat{\gamma}_{j,k},\ k=0,\ldots,2^j-1,\ j=0,\ldots,p-1\}$. Note that

$$\langle\widehat{\varphi}_{\widehat{\sigma}(t);0},\widehat{\varphi}^{\widehat{\sigma}(t);0}\rangle=1,\quad \langle\widehat{\gamma}_{\widehat{\sigma}(t);j_1,k_1},\widehat{\gamma}^{\widehat{\sigma}(t);j_2,k_2}\rangle=\delta_{j_1,j_2}\delta_{k_1,k_2},\quad \langle\widehat{\varphi}_{\widehat{\sigma}(t);0},\widehat{\gamma}^{\widehat{\sigma}(t);j,k}\rangle=0,\quad \langle\widehat{\varphi}^{\widehat{\sigma}(t);0},\widehat{\gamma}_{\widehat{\sigma}(t);j,k}\rangle=0. \tag{2}$$

The M sample curves can then be approximated in terms of the following empirical local wavelet-vaguelette decomposition: For $i = 1,\ldots,M$, and for each $t \in [0,S]$,

$$X_i(t) \simeq \widehat{\mu} + \widehat{X}^{\widehat{\varphi}^0}_{\widehat{\sigma}(t);i}\widehat{\varphi}_{\widehat{\sigma}(t);0}(t) + \sum_{j=0}^{p-1}\sum_{k=0}^{2^j-1} \widehat{X}^{j,k,\widehat{\gamma}}_{\widehat{\sigma}(t);i}\widehat{\gamma}_{\widehat{\sigma}(t);j,k}(t), \quad t \in [0,S]. \tag{3}$$

This decomposition will be considered in the implementation of functional logistic regression to classify the data into two groups, G_0 and G_1.

Functional Logistic Regression

Consider a response variable Y with Bernoulli distribution, having mean μ and variance $\sigma^2 = \mu(1-\mu)$. The response variable Y takes the value $Y = 1$ if the sample curve is in group G_1, or $Y = 0$ if it isn't. We define $\eta_i = g(\mu_i) = \alpha + \int \beta(t)Z_i(t)dt$, for α a constant, g the *logit* function, $g^{-1}(x) = e^x/(1+e^x)$, and $Z_i(t) = X_i(t) - \hat{\mu}(t)$. Thus, $Y_i = g^{-1}(\eta_i) + e_i$, with errors $e_i, i = 1,\ldots,M$, considered as independent random variables with zero-mean and finite variance.

Due to the square integrability of β, the functional parameter β admits the local decomposition:

$$\beta(t) = \beta_{\sigma(t);\varphi_0}\varphi^{\sigma(t);0}(t) + \sum_{j=0}^{p-1}\sum_{k=0}^{2^j-1} \beta_{\sigma(t);j,k,\gamma}\gamma^{\sigma(t),j,k}(t), \quad t \in [0,S], \tag{4}$$

in terms of the dual local Riesz bases $\{\varphi_{\sigma(t);0}, \gamma_{\sigma(t);j,k},\ k = 0,\ldots,2^j-1,\ j = 0,\ldots,p-1\}$ and $\{\varphi^{\sigma(t);0,k}, \gamma^{\sigma(t);j,k},\ k = 0,\ldots,2^j-1,\ j = 0,\ldots,p-1\}$. In the development below, the local Fourier coefficients of parameter function β, with respect to the empirical scaled basis $\{(\varphi_{\widehat{\sigma}(t);0},\ \gamma_{\widehat{\sigma}(t);j,k},\ k = 0,\ldots,2^j-1,\ j = 0,\ldots,p-1\}$, will be denoted as $\widehat{\beta}^t_{\varphi_0} = \beta_{\widehat{\sigma}(t);\varphi_0}$, $\widehat{\beta}^t_{j,k,\gamma} = \beta_{\widehat{\sigma}(t);j,k,\gamma}$, $k = 0,\ldots,2^j-1,\ j = 0,\ldots,p-1$, for each $t \in [0,S]$. The above approximations of $Z_i(t)$ from (3), and of $\beta(t)$ from (4), considering (2), lead to the following estimation of $\eta_i(t)$:

$$\eta_i(t) \simeq \widehat{\alpha}^t + \sum_k \widehat{Z}_i^{\widehat{\sigma}(t);\widehat{\varphi}^0}\widehat{\beta}^t_{\varphi_0} + \sum_{j=0}^{p-1}\sum_{k=0}^{2^j-1} \widehat{Z}_i^{\widehat{\sigma}(t);j,k,\widehat{\gamma}}\widehat{\beta}^t_{j,k,\gamma}.$$

The functional model is then reduced to a generalized linear model (see [2]), for each $t \in [0,S]$, where iterated weighted least square estimation is usually applied to compute $\widehat{\beta}^t$, from the following equations: For $i = 1,\ldots,M$,

$$\sum_i (Y_i - \mu_i(t)) = 0 \quad \sum_i (Y_i - \mu_i(t))(\widehat{Z}^t_i)^T = 0,$$

where $(\widehat{Z}^t_i)^T$ is the vector of Fourier coefficients of $Z_i(t)$ on the empirical locally scaled wavelete-vaguelette basis $\{\widehat{\varphi}^{\widehat{\sigma}(t);0}, \widehat{\gamma}^{\widehat{\sigma}(t);j,k},\ k = 0,\ldots,2^j-1,\ j = 0,\ldots,p-1\}$.

The mean $\widehat{\beta}$ over t of the obtained $\widehat{\beta}^t$, for each $t = t_1,\ldots,t_M$, is computed. A prior probability p_0 is considered for G_0 memberships, and similarly, a prior probability p_1 is considered for G_1 memberships. Thus, if $\hat{pr}(Y_i = 1|X_i(t)) = \widehat{\eta}_i = g^{-1}((\widehat{Z}^T_i, 1) * \widehat{\beta}) \geq p_1$, the ith curve is a member of G_1. Otherwise, it belongs to G_0.

3. RESULTS

Application to the analysis of yeast cell cycle gene expression profiles

We use the temporal gene expression data (α factor synchronized) for $M = 90$ genes involved in the yeast cell cycle obtained by [5] as sample curves. The gene expression is measured every 7 minutes between 0 and $S = 119$ minutes (both time instants included), thus, $n = 18$ observations for each gene. It is known that 44 of these genes are related to G_1 phase regulation and 46 to the $S, S/G_2, G_2/M$ and M/G_1 phases.

Figure (1) displays the original data with their approximation in terms of the local wavelet-vaguelette decomposition, considering a grid with $N = 64 = 2^6$ equally spaced time points. Convergence of the iterated weighted least squares algorithm is achieved for every point t in the grid after 100 iterations or less controlled by de deviance. $\widehat{\beta}^t$ for a grid with $N = 64$ equally spaced time points, are displayed with the mean vector $\widehat{\beta}$ in Figure 2. In order to measure the accuracy of the model, the cross-validation

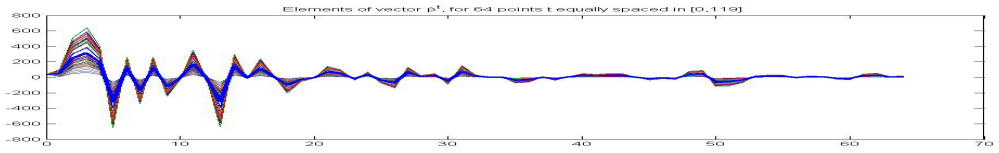

Figure 2: Components of $\widehat{\beta}^t$ for 64 equally spaced points $t \in [0,119]$, the coarsest line is the mean vector $\widehat{\beta}$, note bigger influence of the first coefficients in the response variable.

classification error rate (*CVE*) is obtained. Suppose the i-th gene is missing, the mean and the covariance function estimates, based on the other 89 genes, and parameters $\widetilde{\beta}$, from the reduced functional sample, are then computed. These parameters are tested to obtain the approximation $\widehat{\eta}^{-i}$ of $\widehat{\eta}$, based on the sample information provided by the 89 gene expression curves, removing the i-th gene.This procedure is repeated with every gene, if $g^{-1}(\eta^{(-i)}) \geq p_1$ the i-th gen is member of G_1, otherwise is from G_0. The *CVE* is defined as the quotient between the total number of genes misclassified under cross-validation, and the total number of genes. High accuracy of the model is assured by a $CVE = 0.13$.

4. CONCLUSIONS

In this paper, a local wavelet-vaguelette decomposition is considered for the non-redundant representation of gene expression profiles, since it holds for a large class of stochastic process, including processes with fractal and heavy-tailed covariance functions (see [3]).

Acknowledgments. This work has been supported in part by COLFUTURO, Colombia; projects MTM2009-13393 of the DGI, MEC, and P09-FQM-5052 of the Andalousian CICE, Spain.

References

[1] Angulo, J.M. & Ruiz-Medina, M.D. (1999). Multiresolution approximation to the stochastic inverse problem. *Advanced Applied Probability* **31**, 1039–1057.

[2] McCullagh, P. & Nelder J.A. (1989). *Generalized Linear Models*. Chapman & Hall

[3] Kelbert, M., Leonenko, N.N. & Ruiz-Medina, M.D. (2005). Fractional random fields associated with stochastic fractional heat equations. *Advanced Applied Probability* **37**, 108–133.

[4] Leng, X. & Müller, H.G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** (1), 68–76. *Annals of Statistics* **33**, 774–805.

[5] Spellman, P.T. , Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders,K., Eisen, M.B., Brown,P.O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell-cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hibridization. *Molecular Biology of the Cell* **9**, 3273–3297.

[6] Vidakovic, B. (2006) *Statistical Modelling by Wavelets.* John Wiley and sons. New York.