Two-stage nonparametric regression for longitudinal data

Eva Ferreira*, Vicente Núñez-Antón

Departamento de Econometría y Estadistíca
Universidad del País Vasco - Spain

Juan Rodríguez-Póo

Departamento de Economía
Universidad de Cantabria - Spain

Summary

In the analysis of longitudinal data it is of main interest to investigate the existence of group and individual effects under correlated observations across time. In this paper, we develop a nonparametric two-step procedure that enables us to estimate group effects under a very general form of correlation across time. Moreover, we propose several methods to estimate the bandwidth and show their asymptotyc optimality. Since the asymptotic distribution is untractable, we develop a randomization test that is suitable for testing the group effects. Finally, we apply the estimation procedure, the bandwidth selection criteria and the randomization test to the data from the Iowa Cochlear Implant Project.

Keywords: Kernel estimation; Bandwidth selection; Nonstationary errors; Group effects; Randomization test.

*All correspondence should be directed to Eva Ferreira. Dpto. de Econometría y Estadística. Universidad del País Vasco. Avda. Lehendakari Aguirre, 83. 48015 Bilbao, Spain. E-mail: etpfegae@bs.ehu.es

1 INTRODUCTION

The analysis of growth and dose response curves, where experimental units (normally allocated to different treatments) are observed over a period of time, has been studied under the hypothesis that the observed sequence of measurements $y(t)$ is sampled from a realization of a continuous time stochastic process $Y(t)$.

In this type of analysis, it is of main interest to estimate the mean response of $Y(t)$ and whether there exist significant differences between the treatment groups. One way to test for differences between groups is to assume that the mean response $Y(t)$ is decomposed into two additive components. The first one captures the so called common effect and the second one the group effect. To estimate both common and specific effects, in a fully parametric context, Laird and Ware (1982) propose linear additive components and Lindstrom and Bates (1990) nonlinear ones. More recently, Boularan et al. (1994) have introduced a method to estimate the components nonparametrically, but they do not incorporate dependence along time in the response variable. On the other hand, Hart and Wehrly (1986) estimate the mean response of a growth curve nonparametrically accounting for correlated errors, but they do not study the possible difference between treatment groups. Moreover, no discussion about bandwidth selection criteria is provided. Rice and Silverman (1991) and Hart and Wehrly (1993) propose cross validation based methods, and Ferreira et al. (1997) introduce a selection method that is based on a modification of the Rice criterion. Moreover, this last paper considers nonstationary correlated errors as suggested by Núñez-Antón and Woodwoth (1994). All the previous error covariance structures fall within the general class considered by Fraiman and Meloche (1994) and, therefore, the result in all cases is that smoothing dependent observations is in general not consistent unless we use repeated measurements. As pointed out in Ferreira et al. (1997), this represents an important problem when analyzing asymptotic optimality of a data driven bandwidth selection method since the bandwidth only has a second order effect.

In order to overcome the previous problem and, at the same time, be able to estimate and test group effects, we propose an additive nonparametric model where the dependency of observations along time is captured by using discrete nostationary covariance structures. This type of covariance structures are motivated in Hart (1991), and analyzed for the stationary case. We extend the results to the nonstationary case, as motivated by Núñez-Antón and Woodwoth (1994). The discrete covariance structure is particularly useful when analyzing data driven bandwidth selection criteria. The additive components will be estimated nonparametrically following the procedure proposed by Boularan et al. (1994) and the bandwidth will be chosen by two different data-driven criteria. We develop a method to test for differences between groups. A standard procedure to carry these tests out will be to obtain the asymptotic distribution of the estimators. Unfortunately, under the hypotheses of our model, this is not a trivial issue. In addition to this, practical reasons have led us to propose a different and well know approach for these “more complicated” types of situations. The proposal we make for these tests is very simple and, also, very easy to implement. We are in fact proposing to carry out a randomization test to detect if the group difference is statistically significant. This test is based on permutation of the rows in the data matrix. This technique appears as a very natural way to test for the group effect. Finally, with this procedure we can also test for the density structure of the time density for each group, since we will propose a model under a random time index.

Thus, the paper structure is as follows. In Section 2 we propose the two-stage model and present the methodology for the estimation of the unknown functions. In Section 3 we present the main results that provide the consistency of the estimations for the two stages. The proposals for the bandwidth selection are described in Section 4. In Section 5 the randomization test for the group effect will be developed. In Section 6 we present the results when applying this methodology to our data set. Finally, in the Appendix, we give the details for the proofs of the main results.

2 STATISTICAL MODEL AND HYPOTHESES

All the considerations in Section 1 lead us to propose the following model

$$Y_{ij} = r(t_{ij}) + r_G(t_{ij}) + \epsilon_i(t_{ij}), \tag{1}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, where n is the maximum number of measurements for each individual, and n_i is the number of available data for the i-th subject. Y_{ij} is the response for subject i at time t_{ij}. If the observation at t_{ij} is missing, we put $Y_{ij} = 0$. We will also let $N = \sum_{i=1}^{m} n_i$ and $N_G = \sum_{i \in G} n_i$. That is, N is the total number of available observations and N_G is the number of available observations in group G. Note that $r(\cdot)$ denotes the common effect and $r_G(\cdot)$ the group effect. We assume that:

- **(M.1)** $\mathrm{E}(Y_{ij}|t_{ij}) = r(t_{ij})$ and $\mathrm{E}(Y_{ij}|i \in G, t_{ij}) = r(t_{ij}) + r_G(t_{ij})$.

The $\epsilon_i(t_{ij})$'s are assumed to model any yet unexplained effect due to the possible existence of a correlation between any two measurements on the same individual.

In addition, we need the following identification condition for r_G:

- **(M.2)** $\mathrm{E}(r_G(t)) = \sum_G r_G(t)P(G) = 0$, for any $t \in [0, b]$,

where $P(G)$ denotes the probability of being in group G. The t_{ij}'s are the times at which the i-th subject is observed for the j-th time. All time points are assumed to belong to a compact, let us say $[0, b]$. As for the time points, we assume that:

- **(T.1)** $t_{ij} \in [s_j, s_{j+1}], \forall i$, with $s_1 = 0$ and $s_{n+1} = b$.
- **(T.2)** $t_{ij} = (s_{j+1} + s_j)/2 + U_{ij}$, where the U_{ij}'s are independent random variables with unconditional density g_j, and the support of the g_j is $[(s_j - s_{j+1})/2; (s_{j+1} - s_j)/2]$. Thus, the density of the random variable representing the time can be written as:

$$f(t) = \frac{1}{n}\sum_{j=1}^{n} g_j\left(t - \frac{s_{j+1} + s_j}{2}\right).$$

- **(T.3)** The conditional density of U_{ij}, when we know that i belongs to group G, will be denoted by g_{jG}. Therefore,

$$f_G(t) = \frac{1}{n}\sum_{j=1}^{n} g_{jG}\left(t - \frac{s_{j+1} + s_j}{2}\right)$$

  will be the density of the time index in group G, since they might be different. It should also be noticed that $f(t) = \sum_G f_G(t)P(G)$.

Here, it is also assumed that:

- **(R.1)** r, r_G, and f are all bounded, p times continuously differentiable functions and $\inf_t f(t) > 0$; $\inf_t f_G(t) > 0$, for any $t \in [0, b]$ and for any group G.
- **(R.2)** For $k \geq 1$ there is a constant C_k so that for $t \in [0, b]$

$$E\left[Y^k | T = t\right] \leq C_k < \infty$$

- **(I.1)** All the $n_i's$ are of the same order, i.e. $n_i = O(n)$.

We will study the consistency of the estimators under a very general structure, allowing for non-stationary correlation between observations on the same individual. In order to obtain the main results, we will need the following assumptions for $i = 1, \ldots, m$ and $j = 1, \ldots, n$:

- **(E.1)** $\mathrm{E}\{\epsilon_i(t_{ij})|t_{ij}\} = 0$.

- **(E.2)** There is independence across individuals and for the same individual we have that

$$\text{cov}\{\epsilon_i(t_{ij}), \epsilon_i(t_{is}) | t_{ij}, t_{is}\} = \gamma(j, s). \tag{2}$$

- **(E.3)** There exists a finite positive constant S_γ such that $\lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} \sum_{s=1}^{n} |\gamma(j,s)| = S_\gamma$

- **(E.4)** $\sum_{s=1}^{n} |j-s||\gamma(j,s)| < C < \infty$, uniformly in j.

- **(E.5)**

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{j,l} \sum_{j',l'} |\mu(j,l,j',l')| < C < \infty,$$

where $\mu(j,l,j',l') = E(\epsilon(t_{ij})\epsilon(t_{il}) - \gamma(j,l))(\epsilon(t_{ij'})\epsilon(t_{il'}) - \gamma(j',l'))$

Remark 1. One can see from these assumptions that we are working with independent individuals, allowing for a discrete nonstationary covariance between observations on the same individual. We do not assume any parametric structure for the covariance, but we restrict by **(E.3)** and **(E.4)** the behavior when the number of observations per subject increases.

Remark 2. We might notice that if $\gamma(j,s) = \gamma(|j-s|)$ we are in the stationary case (Hart, 1991). In this sense, our results can be interpreted as a generalization of those obtained in his paper.

Remark 3. A discrete structure like (2) is in between the continuous structure and the independent case, letting the data values at two design points separated by a fixed distance, be asymptotically uncorrelated.

We now propose the estimates for the components of our model, providing their theoretical properties.

3 KERNEL ESTIMATORS AND FIRST PROPERTIES

Each of the two components, the common and the group effect, in model (1) will be estimated using a kernel smoothing technique.

The function $r(\cdot)$ will be estimated by using all the data as if they had been collected from one single individual. We use as estimates of $r(\cdot)$ the familiar Naradaya-Watson kernel regression smoother (Härdle, 1990). These estimates are defined for a kernel function $K(\cdot)$ and for a bandwidth h by

$$\hat{r}(t) = \frac{\frac{1}{Nh}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij} K\left(\frac{t_{ij}-t}{h}\right) I(Y_{ij} \neq 0)}{\hat{f}(t)} \tag{3}$$

with the usual Rosenblatt-Parzen kernel density estimator method (Rosenblatt, 1956; Parzen, 1962):

$$\hat{f}(t) = \frac{1}{Nh}\sum_{i=1}^{m}\sum_{j=1}^{n} K\left(\frac{t_{ij}-t}{h}\right) I(Y_{ij} \neq 0) \tag{4}$$

Here $I(\cdot)$ is an indicator function for the condition used as argument. Once we have estimated the common effect $r(\cdot)$, we estimate the group component $r_G(\cdot)$ by doing kernel regression on the residuals for each group, i.e. for any group G we define the estimate:

$$\hat{r}_G(t) = \frac{\frac{1}{N_G h_G}\sum_{i\in G}\sum_{j=1}^{n} (Y_{ij} - \hat{r}(t_{ij})) K\left(\frac{t_{ij}-t}{h_G}\right) I(Y_{ij} \neq 0)}{\hat{f}_G(t)} \tag{5}$$

Here the bandwidth h_G may vary from one group to the other. The corresponding density estimator is

$$\hat{f}_G(t) = \frac{1}{N_G h_G}\sum_{i\in G}\sum_{j=1}^{n} K\left(\frac{t_{ij}-t}{h}\right) I(Y_{ij} \neq 0) \tag{6}$$

The usual assumptions in kernel smoothing techniques for the kernel K and the different bandwidths are as follows:

- **(K.1)** K is a kernel having compact support on $[-c, c]$, with $c > 0$.
- **(K.2)** K is a bounded kernel of order p.
- **(K.3)** $\int_{-c}^{c} K(z)dz = 1$.
- **(H.1)** $h \longrightarrow 0$, as N tends to infinity.
- **(H.2)** $h_G \longrightarrow 0$, as N_G tends to infinity.

Now we need to study the asymptotic properties of our nonparametric estimators given by equations (3) and (5). That is, we give asymptotic bounds for the consistency properties for these estimates. When giving these properties, and as typical in kernel smoothing techniques, a weight function $w(t)$ with the following assumption will be used:

- **(W.1)** w is positive, bounded and integrable over $[0, b]$.

In order to study the asymptotic properties, let us consider the mean square error and the mean integrated square error of any estimator $s(t)$ that is being estimated by $\hat{s}(t)$.

$$\begin{aligned}
\mathrm{MSE}\{\hat{s}(t)\} &= \mathrm{E}\{\hat{s}(t) - s(t)\}^2 \\
&= [\mathrm{E}\{\hat{s}(t)\} - s(t)]^2 + \mathrm{var}\{\hat{s}(t)\} \\
&= \mathrm{bias}^2\{\hat{s}(t)\} + \mathrm{var}\{\hat{s}(t)\}, \\
\mathrm{MISE}(h) &= \int_0^b \mathrm{E}\{\hat{s}(t) - s(t)\}^2 w(t) dt, \\
\text{and} \quad \mathrm{ASE}(h) &= \frac{1}{N} \sum_{i,j} \{\hat{s}(t_{ij}) - s(t_{ij})\}^2 w(t_{ij}).
\end{aligned}$$

Theorem 1 *Under assumptions* **(E.1)-(E.4)**, **(H.1)**, **(I.1)**, **(K.1)-(K.3)**, **(M.1)-(M.2)**, **(R.1)**, *and* **(T.1)-(T.3)**, *we have that, for any $t \in (0, b)$, as Nh goes to infinity,*

$$MSE\{\hat{r}(t)\} = c_K(t)^2 h^{2p} + \frac{1}{Nhf(t)} S_\gamma d_K + o(h^{2p}) + o((Nh)^{-1}),$$

where $d_K = \int K^2(u)du$ and $c_K(t) = (f(t)p!)^{-1} \int_{-c}^{c} \eta^p K(\eta) d\eta \{(rf)^{(p)}(t)\}$.

Remark 4. The order of the MSE is the standard in a nonparametric estimator. The relevance of our result is that this order can be obtained under a very flexible covariance structure.

Remark 5. The results show the consistency whenever h goes to zero and Nh goes to infinity. Note that for this result to hold it is not necessary to let both, n and m, go to infinity.

Remark 6. The greater the variance term is, the greater the smoothing parameter will be, and the variance term increases with S_γ. So, if we have correlation between units, the smoothing parameter should be greater than in the independent case. This is consistent with the analysis made by Hart and Wehrly (1986), since, as they point out "the less smooth the correlation function is at zero, the larger the optimum bandwidth is." Clearly, the discrete structure is not even continuous and, therefore, not smooth at all.

It is also possible to give a global convergence result that we present as a corollary.

Corollary 1 *Under the assumptions of Theorem 1, as Nh goes to infinity,*

$$\mathrm{MISE}(h) = C_K h^{2p} + D_K \frac{1}{Nh} + o(h^{2p}) + o((Nh)^{-1}),$$

as N tends to infinity. Here, the constants are $C_K = \int_0^b c_K(t)dt$ and $D_K = S_\gamma d_K \int_0^b f(t)^{-1}dt$.

The next theorem gives the consistency for the kernel regression estimate $\hat{r}_G(\cdot)$ for the second stage (i.e. the estimate for the group effect).

Theorem 2 *Assuming that* **(H.2)** *and the conditions in Theorem 1 hold, we have that, for any group G, as $N_G h_G$ goes to infinity,*

$$\mathrm{MSE}\{\hat{r}_G(t)\} \quad = \quad c_G(t)h_G^{2p} + \frac{1}{N_G h_G f_G(t)} S_\gamma d_K + o(h_G^{2p}) + o((N_G h_G)^{-1}),$$

where $c_G(t) = (f_G(t)p!)^{-1} \int_{-c}^{c} \eta^p K(\eta)d\eta\{(r_G f_G)^{(p)}(t)\}$.

Remark 7. Again, we see that the convergence of the nonparametric estimator to the underlying function goes as rapidly as possible under the standard conditions.

Finally, we give a global convergence result for the second stage estimator.

Corollary 2 *Under the assumptions of Theorem 2, as $N_G h_G$ goes to infinity,*

$$\mathrm{MISE}_G(h) = E_K h_G^{2p} + F_K \frac{1}{N_G h_G} + o(h_G^{2p}) + o((N_G h_G)^{-1}) + O((h_G^{2p} + (N_G h_G)^{-1})(h^{2p} + (Nh)^{-1})).$$

Here, the constants are $E_K = \int_0^b c_G(t)dt$ and $F_K = S_\gamma d_K \int_0^b f_G(t)^{-1}dt$.

4 SELECTION OF THE SMOOTHING PARAMETER

In the previous sections, the estimator and the consistency results have been obtained assuming fixed bandwidths for the estimator of the common effect, h, and the group effect, h_G. In this section, we will introduce different procedures to compute these bandwidths from data. This is of particular interest when considering, in the next section, the randomization approach to test for equality of the different groups. We define one optimal global bandwidth for the estimator of the mean effect as the solution to the following optimization problem $h^* = \mathrm{argmin}_h \mathrm{MISE}(h)$, and for the group G, $h_G^* = \mathrm{argmin}_h \mathrm{MISE}_G(h)$. Unfortunately, these optimal bandwidths can not be determined since they are functions of unknown quantities. To overcome these problems, several methods have been proposed in the literature. Among them, we have decided to use the cross validation method, and the method proposed by Rice (1984). The reason is that both bandwidth selection methods can be easily transformed to be used in a longitudinal data framework. For cross validation see Hart and Wehrly (1993) and for the Rice criterion Ferreira et al. (1997).

In the cross validation setting, the bandwidth for the mean effect will be chosen according to the following criterion $\hat{h}^{CV} = \operatorname{argmin}_h \mathrm{CV}(h)$. The cross validation function is defined as $\mathrm{CV}(h) = N^{-1}\sum_{i,j}\left(Y_{ij} - \hat{r}^{-i}(t_{ij})\right)^2 w(t_{ij})$, where

$$\hat{r}^{-i}(t) = \frac{\displaystyle\sum_{k\neq i}\sum_{j=1}^{n} Y_{kj,G} K\left(\frac{t_{kj}-t}{h}\right) I(Y_{kj,G}\neq 0)}{Nh\hat{f}^{-i}(t)} \tag{7}$$

and

$$\hat{f}^{-i}(t) = \frac{\displaystyle\sum_{k\neq i}\sum_{j=1}^{n} K\left(\frac{t_{kj}-t}{h}\right) I(Y_{kj}\neq 0)}{Nh}. \tag{8}$$

As for the group effect, we have selected the bandwidth $\hat{h}_G^{CV} = \operatorname{argmin}_h \mathrm{CV}_G(h)$. The cross validation function is defined as $\mathrm{CV}_G(h) = N_G^{-1}\sum_{i\in G,j}\left(Y_{ij} - \hat{r}(t_{ij}) - \hat{r}_G^{-i}(t_{ij})\right)^2 w(t_{ij})$, where

$$\hat{r}_G^{-i}(t) = \frac{\displaystyle\sum_{k\neq i;k\in G}\sum_{j=1}^{n} Y_{kj} K\left(\frac{t_{kj}-t}{h}\right) I(Y_{kj}\neq 0)}{N_G h\hat{f}_G^{-i}(t)} \tag{9}$$

and

$$\hat{f}_G^{-i}(t) = \frac{\displaystyle\sum_{k\neq i;k\in G}\sum_{j=1}^{n} K\left(\frac{t_{kj}-t}{h}\right) I(Y_{kj}\neq 0).}{N_G h} \tag{10}$$

If we want to use the estimated bandwidths propertly, it is necessary to establish some relationship between respectively h^* and h_G^* and $\hat{h}^{CV}$ and $\hat{h}_G^{CV}$. Following Shibata (1981), the next theorems state the asymptotic optimality, with respect to the MISE, of $\hat{h}^{CV}$, $\hat{h}_G^{CV}$.

Theorem 3 *Assume that conditions in Theorem 1 hold. In addition assume that* **(E.5)**, **(R.2)** *and* **(W.1)** *also hold. Then, as m tends to infinity,*

$$\frac{\mathrm{MISE}\left(\hat{h}^{CV}\right)}{\inf_{h\in H_N}\mathrm{MISE}(h)} \longrightarrow 1 \qquad \text{a.s.} \tag{11}$$

for some $0<\epsilon<\frac{1}{2p+1}$ *and* $H_N = [aN^{-\frac{1}{2p+1}-\epsilon}, bN^{-\frac{1}{2p+1}+\epsilon}]$.

Theorem 4 *Assume that conditions in Theorem 2 hold. In addition, assume that* **(R.2)** *and* **(W.1)** *also hold. Then, as* $m_G = \#\{i\in G\}$ *tends to infinity*

$$\frac{\mathrm{MISE}_G\left(\hat{h}_G^{CV}\right)}{\inf_{h_G\in H_{N_G}}\mathrm{MISE}_G(h_G)} \longrightarrow 1 \qquad \text{a.s.} \tag{12}$$

for some $0 < \epsilon < \frac{1}{2p+1}$ *and* $H_{N_G} = [a{N_G}^{-\frac{1}{2p+1}-\epsilon}, b{N_G}^{-\frac{1}{2p+1}+\epsilon}]$.

These results were originally proved for the i.i.d. case by Härdle and Marron (1985), and they were extended to the case of dependence by Härdle and Vieu (1992). In the framework of longitudinal data, Hart and Wehrly (1993) show the consistency of cross-validation under a continuous and stationary structure for the covariance function. Boularan et al. (1994) show the same consistency property when estimating the different components of an additive model, but assuming independence among the different curves. The main interest of the results shown in Theorems 3 and 4 is that the asymptotic optimality of the cross validated bandwidths is shown under the assumptions that the covariance structure is discrete and nonstationary. This generalizes the previous results and highlights a very important feature in the asymptotic properties in the cross-validation method for longitudinal data. That is, cross-validation when deleting one curve is not consistent unless we let the number of individuals go to infinity.

As pointed out in Section 1, we are very interested in the applicability of the methodology presented in this work. The dataset we are interested in has been analyzed before and presents a correlation structure that can be well modelled in a parametric way. Therefore, as Hart and Wehrly (1993) argue, if we use a modified Rice criterion (see Rice, 1984), the resulting smoothing parameter should be better. Roughly speaking, selecting cross validation is like using the sample covariance as an estimator for the covariance structure, whereas the Rice criterion uses a consistent estimator of the covariance.

For this modified Rice criterion, the bandwidth for the common effect will be selected according to the following criterion $\hat{h}^R = \text{argmin}_h \text{RICE}\,(h)$, where

$$\text{RICE}(h) \quad = \quad \frac{1}{N}\sum_{i,j}(Y_{ij} - \hat{r}(t_{ij}))^2 w(t_{ij}) + \frac{2}{N^2 h}\sum_{i,j}\frac{1}{\hat{f}(t_{ij})}\sum_{k=1}^{n_i} K\left(\frac{t_{ij}-t_{ik}}{h}\right)\hat{\gamma}(k,j)w(t_{ij}).$$

Theorem 5 *Assume that conditions in Theorem 1 hold. In addition that assume* **(E.5)**, **(R.2)** *and* **(W.1)** *also hold, and that* $\hat{\gamma}$ *is a consistent estimator of* γ. *Then, as* m *tends to infinity,*

$$\frac{\text{RICE}\left(\hat{h}^R\right)}{\inf_{h \in H_N} \text{MISE}\,(h)} \longrightarrow 1 \qquad \text{a.s.} \tag{13}$$

for some $0 < \epsilon < \frac{1}{2p+1}$ *and* $H_N = [aN^{-\frac{1}{2p+1}-\epsilon}, bN^{-\frac{1}{2p+1}+\epsilon}]$.

Now, we can consider the same criterion for the second stage. In this case, the bandwidth for the group effect will be $\hat{h}_G^R = \text{argmin}_h \text{RICE}_G\,(h)$.

$$\text{RICE}_G(h) = \frac{1}{N_G}\sum_{i\in G,j}(Y_{ij} - \hat{r}(t_{ij}) - \hat{r}_G(t_{ij}))^2 w(t_{ij}) + \frac{2}{N_G^2 h}\sum_{i\in G,j}\frac{1}{\hat{f}_G(t_{ij})}\sum_{k=1}^{n_i} K\left(\frac{t_{ij}-t_{ik}}{h}\right)\hat{\gamma}(k,j)w(t_{ij}).$$

For the second stage, we state the next theorem.

Theorem 6 *Assume that conditions in Theorem 2 hold. In addition, assume that* **(R.2)** *and* **(W.1)** *also hold, and that $\hat{\gamma}$ is a consistent estimator of γ. Then, as m_G tends to infinity*

$$\frac{\mathrm{MISE}_G\left(\hat{h}_G^R\right)}{\inf_{h_G \in H_{N_G}} \mathrm{MISE}_G\left(h_G\right)} \longrightarrow 1 \qquad \text{a.s.} \tag{14}$$

for some $0 < \epsilon < \frac{1}{2p+1}$ and $H_{N_G} = [aN_G{}^{-\frac{1}{2p+1}-\epsilon}, bN_G{}^{-\frac{1}{2p+1}+\epsilon}]$.

5 TESTING FOR THE GROUP EFFECT

Once we have estimated the mean and the group effect, it can be of interest to analyze the difference between groups. That is, we would like to test if the difference between groups is statistically significant. This has been an issue receiving some consideration from other authors that have analyzed this dataset in the past (see e.g., Zimmerman et al., 1998 and Núñez-Antón and Woodworth, 1994). Thus, there are two issues regarding possible tests to be carried out, for the two different groups:

- A contrast to determine if the difference between groups is statistically significant.
- A contrast to determine if the difference between the densities for the times at which individuals are observed is statistically different for the two groups. Regarding this issue, we have assumed, for estimation purposes, that the densities are in fact different and, thus, we have used different densities for the estimation of the group effects.

That is, we are interested in testing both the statistical significance of the difference between groups and between densities for each of the groups. Even though the methodology is similar, we will only concentrate on a randomization test for contrasting the difference between groups, assuming that each of the two groups has a different density.

A randomization test is a permutation test based on randomization (random assignment) to test a null hypothesis about treatment effects in a randomized experiment. In our case, individuals were randomly assigned to the different groups (i.e. different types of cochlear implant) and, thus, we are on this situation. The test is carried out as follows. A test statistic is computed for the experimental data, then the data are permuted (divided or rearranged) repeatedly and the test statistic is computed for each of the resulting data permutations. Those data permutations, including the one representing the obtained results, constitute the reference set for determining significance. The proportion of data permutations in the reference set that have test statistic values greater than or equal to (or, for certain test statistics, less than or equal to) the value for the experimentally obtained results is the P-value (significance or probability value). If, for example, the proportion is 0.02, the p-value is 0.02, and the results are significant at the 0.05 but not the 0.01 level of significance. Determining significance on the basis of a distribution of test statistics generated by permuting the data is characteristic of all permutations tests; it is when the basis for permuting the data is random assignment that a permutation test is called a randomization test. This is the type of test we will be using here. See Good (1994), Edgington (1995), Garthwaite et al. (1995, chapter 9) and Lehmann (1986, Chapter 5, Sec. 12, for references to complete some general details on this methodology.

The null hypothesis for a *traditional randomization test* is that the measurements for each person or other unit that is randomly assigned will be the same under one assignment to treatment as under any alternative assignment that could have resulted from the random assignment procedure. Thus, when the null hypothesis for the traditional randomization test, which is the hypothesis of no differencial treatment effect, is true, random assignment of subjects to treatments randomly divides the measurements among the treatments. Each data permutation in the reference set, which functions as a randomization test "significance table", represents the results that would have been obtained for a particular assignment if the null hypothesis is true.

The procedure we will use for our specific test to investigate the statistical significance of the treatment difference is as follows:

1. Define a "grid" of points where all effects (i.e. common and group effects) are estimated. Note that this is necessary because the two groups do not have common time points. The number of time points in the grid will be denoted by N_g.
2. Estimate the densities $f_A(\cdot)$, $f_B(\cdot)$, and the group effects, $\hat{r}_A(t)$ and $\hat{r}_B(t)$ for the original dataset, using the methods specified in Sections 3 and 4.
3. Define the test statistic as:
$$T_{obs} = \sum_{j=1}^{N_g} |\hat{r}_A(t_j) - \hat{r}_B(t_j)| = 2 \sum_{j=1}^{N_g} |\hat{r}_A(t_j)| \, . \tag{15}$$
The last equality holds because of the condition $\hat{r}_B(t) = -\hat{r}_A(t)$.
4. Keep the number of observations per group as 21, the initial value in the original dataset.

5. Permute randomly the rows of the data matrix (m individuals with the i-th individual having n_i observations available) Q times.

6. For each of the $q = 1, \ldots, Q$ permutations of the data, assign half of the individuals (i.e. 21) to treatment A, and the other half to treatment B. Let us denote these halves by A_q and B_q. This will constitute our new dataset originated from permuting the data.

7. Maintain the estimate of the common effect since the sampling is without replacement. Estimate $f_{A_q}(\cdot)$, $f_{B_q}(\cdot)$, $r_{A_q}(t)$ and $r_{B_q}(t)$, as indicated in Sections 3 and 4, keeping the values for the bandwidths we obtained for the original dataset, since the optimal bandwidth is not affected by the randomization process.

8. Compute the test statistic in equation (15) for each of the $q = 1, \ldots, Q$ permutations of the data.

9. Calculate the P-value or the critical value for our randomization test. The proportion of data permutations in the reference set that have test statistic values greater than or equal to the value for the experimentally obtained results is the P-value (significance or probability value). The critical value will be the $(1 - \alpha)$-th percentile of the test statistic values calculated from the respective permutations.

10. Conclude if the difference in group effect is statistically significant or not, at the α level of significance.

The same procedure can be used for testing the significance of the difference for the densities in the two groups.

6 Application to the Speech Recognition Data

In this section we present the application of our methodology to the speech recognition dataset from the Iowa Cochlear Implant Project (Gantz et al., 1988). In this data set, the data consist of scores (percentage of correct responses) on a sentence test administered under audition-only conditions to groups of human subjects wearing one of two types of cochlear implants, referred to here as A and B. Implants were surgically implanted five to six weeks prior to being electrically connected to an external speech processor. Subjects were profoundly, bilaterally deaf, thus pre-connection baseline values for the sentence test were all zero. Twenty-one subjects received implant A and 21 received implant B.

The data set consists of $N = 155$ observations available on $m = 42$ subjects, 21 subjects in each of the two groups for each type of cochlear implant. In the notation introduced in Section 2, we also have that the number of observations available for the subjects wearing cochlear implant type A was $N_{GA} = 76$ and the one for subjects wearing cochlear implant type B was $N_{GB} = 79$. The maximum number of available observations a subject can have is $n = 6$. As illustrative examples, we can point out that the values the n_i's (i.e. number of observations available on subject i) take on are different and small. For example, $n_1 = 2$, $n_2 = 3$, $\cdots$, $n_{41} = 2$, and $n_{42} = 2$. Thus, as can clearly be seen, most subjects do not have more than 2 or 3 observations available, and very few have all 6 observations.

Measurements were scheduled at 1, 9, 18, 30, 40, and 54 months after connection; there was some variation in actual follow-up times, however, so these times were not exact. Moreover, some subjects did not show up for one or more of their scheduled follow-ups. Thus, actual follow-up times could be very well considered to be randomly distributed around the scheduled follow-up times. It was also assumed that these observations were missing at random. We are interested in estimating the common and group (i.e. type of cochlear implant) effects believed to be present in this data set. As proposed in earlier sections, we would compare two well known methods for bandwidth selection: the cross validation criterion and the Rice criterion. The theoretical properties for these two methods have shown that they work well under the conditions stated in Section 4, which can be easily verified for this dataset (see, e.g. Núñez-Antón et al., 1999).

In addition, we would like to test for a group effect (i.e. if there is as significant difference between the two types of cochlear implants). We also want to investigate the features in the within subjects covariance structure, known to have a specific parametric nonstationary structure (see, e.g. Núñez-Antón and Woodworth, 1994, and Zimmerman et al., 1998). These ideas and the steps to carry them out can, of course, provide a clear picture of the main properties of the dependency between audiologic performance and time for this data set.

Figures 1 and 2 show the scatter plot and the profiles of the audiologic performance over time for all 42 subjects. From these figures, it is quite hard to reckon a precise parametric form to use for the model in this data set.

Núñez-Antón and Woodworth (1994) analyzed a previous version of this dataset. Their final model for the mean response was a quadratic in time model that did not find a significant group

Figure 1: Scatter plot of the speech recognition data set.

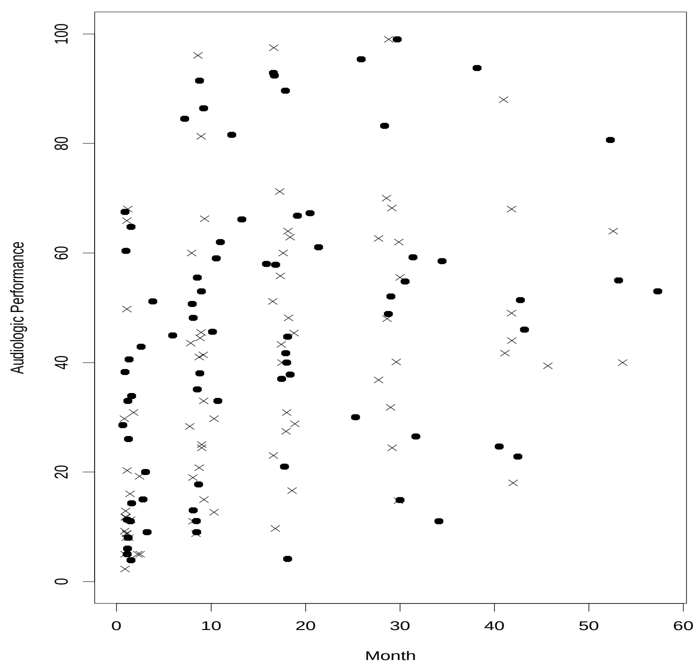
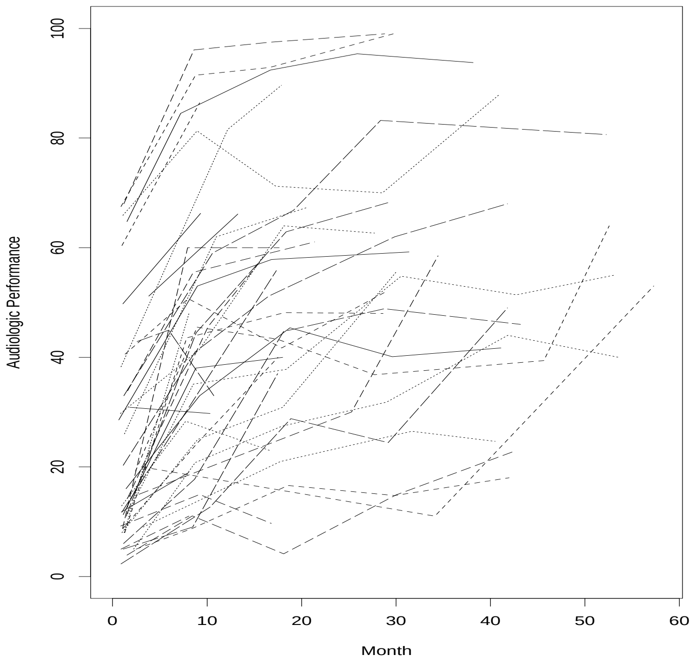


Figure 2: Profiles for the subjects in the speech recognition data set.



effect. The other known parametric approach for this data set has been taken by Zimmerman et al. (1998). They considered the updated and corrected version of the dataset (i.e. the one we will use here). They decided to consider only the observations available at 1, 9, 18 and 30 months and obtained a quadratic in time model for the mean response and a significant group effect. Núñez-Antón et al. (1999) proposed a three-stage estimation approach, assuming a continuous covariance structure and without justifying the selection of the smoothing parameter.

For our estimation process we have selected the Gaussian kernel and for the bandwidth selection, we present both cross validation and Rice criteria. For the latter, we have used estimates obtained from nonlinear least squares, following the ideas in Ferreira et al. (1997), for the theoretical model where:

$$\mathrm{cov}\{\epsilon_i(t_{ij}), \epsilon_r(t_{rs})|t_{ij}, t_{rs}\} = \begin{cases} \sigma^2 \rho^{\frac{|j^\lambda - s^\lambda|}{\lambda}} & \text{if } i = r \\ 0 & \text{if } i \neq r \end{cases}$$

Estimation of the Common Effect:

We have used the methodology develop in Section 3. The value obtained for the bandwidth using the leave-out-one-subject cross-validation criterion was $\hat{h}^{CV} = 3.05$, and the corresponding bandwidth using the Rice criterion was $\hat{h}^{R} = 2.62$. Figure 3 shows the scatter plot of the data together with the estimated curves $\hat{r}$ for the common effect for both methods of bandwidth selection.

As can be seen from this figure, there is no real difference for the estimates obtained for the common effect, when using either one of these two methods.

Figure 3 shows that up to about 30 months there is an improvement in the hearing of the subject, then it seems that the performance becomes worse as time goes by and it then increases again at about 40 months. This decrease could be due mainly to the fact that the subjects with significant improvement in their hearing did not come back to any more appointments, and the subjects that continued being observed were the ones showing improvement on the average lower than the mean of the whole group. We will be able to see if this hypothesis is reasonable when estimating the individual effects. The final increase in performance could be just due to the effect of the implant in the people that came back for the last appointments; i.e. people's hearing gets better as times goes by. Since the individuals left are not exactly, in general, the ones with higher performances, if the implants are somewhat helpful, they should improve their hearing in the long run.

Estimation of the Group Effect: According to the methodology in Sections 3 and 4, we estimate $\hat{r}_A(t)$ and $\hat{r}_B(t)$. Now, motivated by assumption **(M.2)** we have re-calculated the effects

Figure 3: Kernel regression estimate for the common effect $\hat{r}(t)$ for cross validation and Rice's criterion together with the scatter plot of the speech recognition data set.

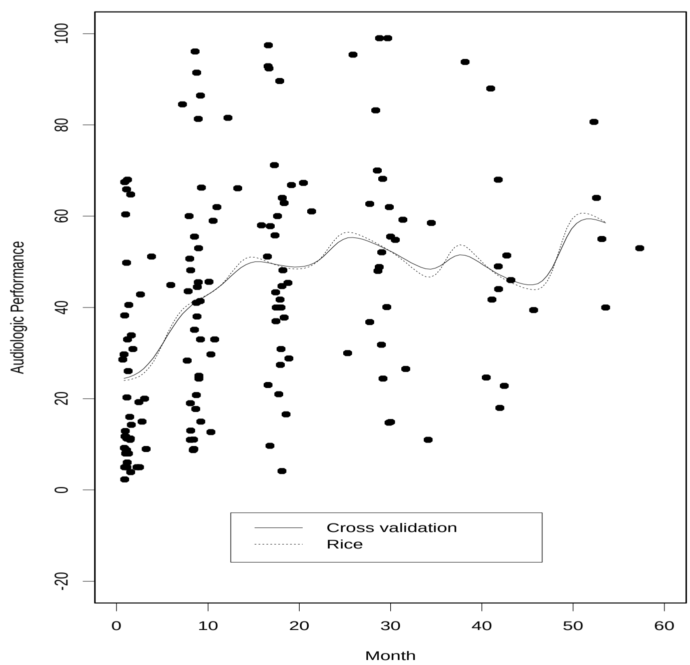
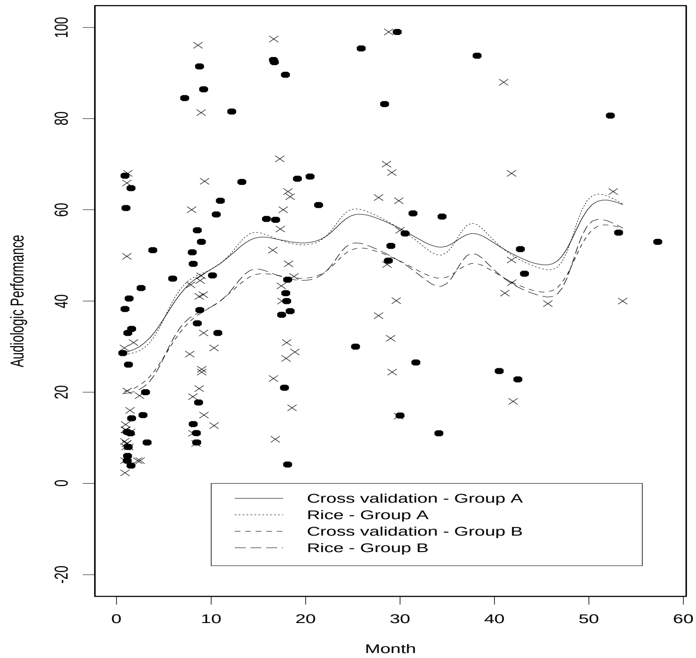


Figure 4: Kernel regression estimate for the common and group effect $\hat{r}_G(t) + \hat{r}(t)$ for cochlear implant groups A and B.



using:

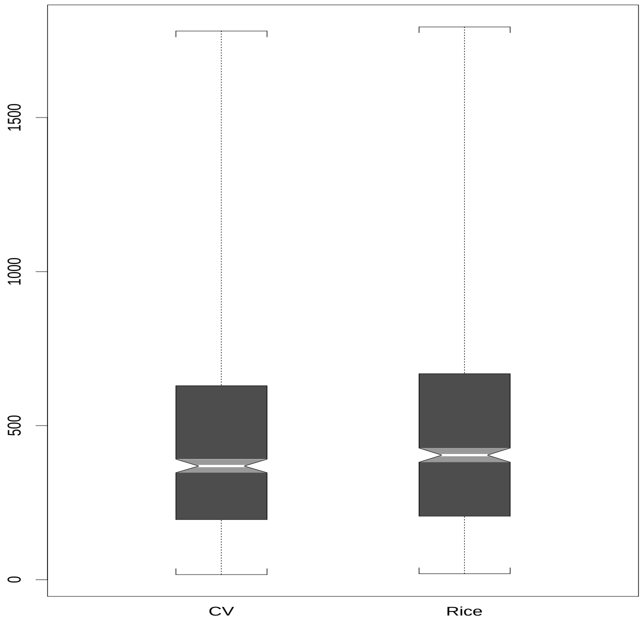
$$\hat{r}_A^*(t) = \frac{\hat{r}_A(t) - \hat{r}_B(t)}{2} \text{ and } \hat{r}_B^*(t) = -\hat{r}_A^*(t).$$

The bandwidths selected using the two methods led to the same large bandwiths values. The values for these bandwiths were $\hat{h}_A = \hat{h}_B = 25$.

Figure 4 shows the scatter plot of the data for both groups together with the estimated curves $\hat{r}(t) + \hat{r}_G(t)$ for the common and group effects, for the cross validation criterion and Rice's criterion. As mentioned before, there is no real difference in the results obtained by the two methods proposed here for this dataset.

From this figure we can see that it looks as if A is a better group, in terms of audiologic performance, than B. If we were analyzing the two groups in more detail, it could be easily seen that the "better" group (i.e. A) in terms of improvement of hearing, would be the one losing more people after 30 months. Basically, the individuals hearing better, most of them from group A, did not come back after 30 months. This last issue can be observed by just looking at the data from group A, also featured in this graphic.

Figure 5: Test statistic values for testing the group difference between cochlear implants using CV and Rice's criterion.



As mentioned in Section 5, we are also interested in testing the significance of the group effect. We carried out the procedure described in Section 5 using $N_g = 100$ and $Q = 1000$. We have obtained, when using cross validation, $T_{obs} = 717.53$ and the 95-th percentile of the test statistic values from the 1000 permutations was $T^*_{0.05} = 1042.18$. On the other hand, Rice's criterion was used, we obtained $T_{obs} = 720.11$ and the 95-th percentile of the test statistic values from the 1000 permutations was $T^*_{0.05} = 1164.72$. As can be seen, in both cases the null hypothesis is not rejected, concluding that the difference between the two groups (i.e. two types of cochlear implants) is not statistically significant at the $\alpha = 0.05$ level. Figure 5 shows the boxplots for the test statistic values from the permutations obtained using cross validation and Rice's criterion.

ACKNOWLEDGEMENTS

This work was supported by Dirección General de Enseñanza Superior del Ministerio Español de Educación y Cultura and Universidad del País Vasco (UPV/EHU) under research grant PB95-0346.

APPENDIX

Proof of Theorem 1: For the proof of this theorem we need the following three lemmas.

Lemma 1 *Under the assumptions established in Theorem 1, we have that for any compact set D in $(0, b)$, as N tend to infinity,*

$$\sup_{t \in D} |\hat{f}(t) - f(t)| \longrightarrow 0 \qquad \text{a.s.}$$

And, as N_G tends to infinity,

$$\sup_{t \in D} |\hat{f}_G(t) - f_G(t)| \longrightarrow 0 \qquad \text{a.s.}$$

Proof of Lemma 1: Assumptions **(R.1)** and **(E.1)** are sufficient for Theorem 3.5.1 from Györfi et al. (1989) and, therefore, the proof is done. ∎

Lemma 2 *Let $t > 0$. Under the assumptions of Theorem 1, we have that*

$$\text{var}\{\hat{r}(t)\} = \frac{1}{Nhf(t)} d_K S_\gamma + o\left(\frac{1}{Nh}\right).$$

Proof of Lemma 2: Letting $K_{ij} = K\left(\frac{t_{ij}-t}{h}\right)$, we can write

$$\begin{aligned} \text{var}\{\hat{r}(t)\} &= \frac{1}{N^2h^2f^2(t)} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n} E\left[E\{\epsilon_{ij}\epsilon_{il}|t_{ij}t_{il}\}K_{ij}K_{il}\right] I(Y_{ij} \neq 0) I(Y_{il} \neq 0) \\ & \quad + o\left(\frac{h\sum_{i=1}^{m} n_i^2}{N^2}\right) = A + o\left(\frac{h\sum_{i=1}^{m} n_i^2}{N^2}\right) \end{aligned}$$

From Assumptions **(E.1)** and **(E.2)**, the first term on the right side is

$$\begin{aligned} A &= \frac{1}{N^2h^2f^2(t)} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n} \gamma(j,l) E(K_{ij}K_{il}) I(Y_{ij} \neq 0) I(Y_{il} \neq 0) \\ &= \frac{1}{N^2h^2f^2(t)} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n} \gamma(j,l) E(K_{ij}^2) I(Y_{ij} \neq 0) I(Y_{il} \neq 0) \\ & \quad + \frac{1}{N^2h^2f^2(t)} \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n} \gamma(j,l) E\left[K_{ij}K'_{ij}\left|\frac{t_{ij}-t_{il}}{h}\right|\right] I(Y_{ij} \neq 0) I(Y_{il} \neq 0) \\ & \quad + O(\text{higher order terms}). \end{aligned}$$

Now, using Assumptions **(T.2)** and **(E.4)** we obtain that the last expression is

$$
\begin{aligned}
& \frac{1}{N^2h^2f^2(t)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\gamma(j,l)E(K_{ij}^2)I(Y_{ij}\neq 0)I(Y_{il}\neq 0) \\
& +\frac{1}{N^2h^2f^2(t)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\gamma(j,l)E\left[K_{ij}K'_{ij}O\left(\frac{|j-l|}{nh}\right)\right]I(Y_{ij}\neq 0)I(Y_{il}\neq 0) \\
& +O(\text{higher order terms}) \\
= \; & \frac{1}{N^2h^2f^2(t)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\gamma(j,l)E(K_{ij}^2)I(Y_{ij}\neq 0)I(Y_{il}\neq 0) \\
& +O\left(\frac{1}{N^2h^2f^2(t)}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{nh}\right)+O(\text{higher order terms})
\end{aligned}
$$

Now, using the change of variable $t_{ij}-t=hu$ and the first order Taylor expansion of $f(\cdot)$ and after some other standard manipulations, we have that the first and leading term can be written as:

$$
\begin{aligned}
& \frac{1}{N^2h^2f^2(t)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\gamma(j,l)\int K^2\left(\frac{\eta_j-t}{h}\right)f(\eta_j)d\eta_j \\
= \; & \frac{1}{N^2hf^2(t)}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\gamma(j,l)\int K^2(u)f(t)du+o((Nh)^{-1}) \\
= \; & \frac{1}{Nhf(t)}d_kS_\gamma+o\left(\frac{1}{Nh}\right).
\end{aligned}
$$

The last step uses the fact that when h goes to zero, the variable u will integrate over the whole domain. This concludes the proof of the lemma. ∎

Lemma 3 *Let $t>0$. Under the assumptions of Theorem 1, we have that*

$$
\mathrm{E}\{\hat{r}(t)-r(t)\}=c_K(t)h^p+o(h^p),
$$

and the $o(\cdot)$ is uniform over t in $]0,b[$.

Proof of Lemma 3:

The proof follows directly from Núñez-Antón et al. (1999), Lemma 2. ∎

Given the previous lemmas, the proof of Theorem 1 is direct using the decomposition of the MSE in terms of the square of the bias plus the variance, and the fact that

$$
\hat{r}(t)-r(t)=(\hat{r}(t)-r(t))\frac{\hat{f}(t)}{f(t)}+(\hat{r}(t)-r(t))\frac{f(t)-\hat{f}(t)}{f(t)}.
$$

∎

Proof of Theorem 2: Theorem 2 is a direct consequence of the following two lemmas.

Lemma 4 *Under the assumptions of Theorem 2 and for $t > 0$, we have that* $\mathrm{E}\{\hat{r}_G(t) - \tilde{r}_G(t)\}^2 = o(h^{2p} + (Nh)^{-1})$, *where the* $o(\cdot)$ *are uniform over* t *and*

$$\tilde{r}_G(t) = \frac{1}{N_G h_G f_G(t)} \sum_{i \in G} \sum_{j=1}^{n} \{Y_{ij} - r(t_{ij})\} K\left(\frac{t_{ij} - t}{h_G}\right) I(Y_{ij} \neq 0).$$

Proof of Lemma 4:

First note that $\hat{r}_G(t) - \tilde{r}_G(t)$ can be written as $(N_G h_G)^{-1} \sum_{i \in G} \sum_{j=1}^{n} K((t - t_{ij}) h_G^{-1})[r(t_{ij}) - \hat{r}(t_{ij})]$. Now, it is not hard to show that for any $s \neq t$, we have that $\mathrm{E}[(r(t) - \hat{r}(t))(r(s) - \hat{r}(s))] = o(\mathrm{E}(r(t) - \hat{r}(t))^2)$. That is, we are dealing with a smoother of $r(t_{ij}) - \hat{r}(t_{ij})$. From here, and using similar arguments to those in Lemmas 2 and 3, we have the final result stated in the lemma. ∎

Lemma 5 *Let $t > 0$. Under the assumptions of Theorem 2, we have that* $\mathrm{E}\{r_G(t) - \tilde{r}_G(t)\}^2 = c_G(t) h_G^{2p} + (N_G h_G f_G(t))^{-1} S_\gamma d_K + o(h_G^{2p}) + o((N_G h_G)^{-1})$, *where* $o(\cdot)$ *is uniform over* t *in* $]0, b[$.

Proof of Lemma 5:

This proof is straightforward by just using the result in Lemma 1 for $f = f_G$, Lemma 4, and the proofs of Lemmas 2 and 3, where we change Y_{ij} by $Z_{ij} = Y_{ij} - r(t_{ij})$, N by N_G, h by h_G, and $f(t)$ by $f_G(t)$. ∎

Proof of Theorems 3 and 5:

In order to simplify the proofs, for any estimator $\hat{s}(t)$ of $s(t)$, we define:

$$\begin{aligned}
\mathrm{ASE}^*(h) &= \frac{1}{N} \sum_i (\hat{s}(t_i) - s(t_i))^2 \left(\frac{\hat{f}(t_i)}{f(t_i)}\right)^2 w(t_i), \\
\overline{\mathrm{ASE}^*}(h) &= \frac{1}{N} \sum_i \left(\hat{s}^{-i}(t_i) - s(t_i)\right)^2 \left(\frac{\hat{f}^{-i}(t_i)}{f(t_i)}\right)^2 w(t_i), \\
\overline{\mathrm{ASE}}(h) &= \frac{1}{N} \sum_i \left(\hat{s}^{-i}(t_i) - s(t_i)\right)^2 w(t_i), \\
\text{and } \mathrm{MISE}^*(h) &= \int_0^b \mathrm{E}(\hat{s}(t) - s(t))^2 \left(\frac{\hat{f}(t)}{f(t)}\right)^2 w(t) dt
\end{aligned}$$

The proof of these theorems relies on the following lemmas.

Lemma 6 *Under the assumptions in Theorem 3, we have that*

$$\sup_{h \in H_N} \frac{\mathrm{MISE}^*(h) - \mathrm{MISE}(h)}{\mathrm{MISE}(h)} \longrightarrow 0 \qquad \text{a.s.},$$

and

$$\sup_{h\in H_N} \frac{\mathrm{ASE}^*(h) - \mathrm{ASE}(h)}{\mathrm{ASE}(h)} \longrightarrow 0 \qquad \text{a.s.}$$

as N tends to infinity.

Proof of Lemma 6: The proof is straightforward from Lemma 1. ∎

Lemma 7 *Under the assumptions in Theorem 3, we have that, as N tends to infinity,*

$$\sup_{h\in H_N} \frac{\mathrm{ASE}^*(h) - \mathrm{MISE}^*(h)}{\mathrm{MISE}^*(h)} \longrightarrow 0 \qquad \text{a.s.}$$

Proof of Lemma 7: The proof of this lemma is inmediate by considering assumptions **(R.1)**, **(R.2)**, **(K.1)**, **(K.2)**, **(K.3)** and **(W.1)**. Then, Theorem 2 from Marron and Härdle (1986), pag. 98, applies and the proof is done. ∎

Lemma 8 *Under the assumptions in Theorem 3, then we have that, as N tends to infinity,*

$$\sup_{h\in H_N} \frac{\mathrm{ASE}^*(h) - \overline{\mathrm{ASE}^*}(h)}{\mathrm{MISE}^*(h)} \longrightarrow 0 \qquad \text{a.s.}$$

Proof of Lemma 8: The proof of this lemma follows closely the proof of Lemma 3 from Härdle and Marron (1985). Let

$$\mathrm{ASE}^*(h) - \overline{\mathrm{ASE}^*}(h) = \frac{1}{N}\sum_{i,j}\left[A_{ij}^2 + 2A_{i,j}\left(r(t_{ij})\hat{f}(t_{ij}) - \hat{r}(t_{ij})\hat{f}(t_{ij})\right)\right] f\left(t_{ij}\right)^{-2} w(t_{ij}),$$

where

$$A_{ij} = r(t_{ij})\hat{f}^{-i}(t_{ij}) - \hat{r}^{-i}(t_{ij})\hat{f}^{-i}(t_{ij}) - \left(r(t_{ij})\hat{f}(t_{ij}) - \hat{r}(t_{ij})\hat{f}(t_{ij})\right).$$

Then,

$$\begin{aligned} \mathrm{ASE}^*(h) - \overline{\mathrm{ASE}^*}(h) &= \left[(m-1)^{-2} + 2(m-1)^{-1}\right] \mathrm{ASE}^*(h) \\ +(m-1)^{-2}N^{-1}\sum_{i,j}&\left\{\frac{1}{nh}\sum_{l=1}^{n} K\left(\frac{t_{ij}-t_{il}}{h}\right)(Y_{il} - r(t_{il}))\right\}^2 f(t_{ij})^{-2} + 2\left[(m-1)^{-2} + (m-1)^{-1}\right] \\ \times N^{-1}\sum_{i,j}&\left(r(t_{ij})\hat{f}(t_{ij}) - \hat{r}(t_{ij})\hat{f}(t_{ij})\right)\frac{1}{nh}\left[\sum_{l=1}^{n} K\left(\frac{t_{ij}-t_{il}}{h}\right)(Y_{il,G} - r(t_{il}))\right] f(t_{ij})^{-2}. \end{aligned}$$

For the second term, by strong law of large numbers, and following the same arguments than in the proof of Lemma 2, we have that, as m tends to infinity,

$$\frac{1}{m-1}\sum_{i=1}^{m}\left[\frac{1}{nh}\sum_{l=1}^{n}K\left(\frac{t_{ij}-t_{il}}{h}\right)\epsilon(t_{il})\right]^2 = O_p\left(\frac{1}{nh}\right)$$

Finally, apply the Cauchy-Schwartz inequality to the cross term, use the results from Corollary 1 and the lemma follows. ∎

The proofs of Theorems 3 and 5 follow closely Theorem 1 from Härdle and Marron (1985), pag. 1468 (H-M from now on). In order to keep the proofs as brief as possible, we will write the expressions such that we can compare them directly with H-M and, therefore, use the results that can be applied to our structure. The main idea will be to see where we use the fact that the errors are not independent, and use there the hypotheses of this more general structure.

For the proof of Theorem 3, it is enough to check that

$$\sup_{h,h'\in H_N}\left|\frac{(\mathrm{MISE}(h)-\mathrm{MISE}(h'))-(\mathrm{CV}(h)-\mathrm{CV}(h'))}{\mathrm{MISE}(h^*)}\right| \longrightarrow 0 \qquad \text{a.s.,}$$

and for the proof of Theorem 5, it is enough to check that:

$$\sup_{h,h'\in H_N}\left|\frac{\mathrm{MISE}(h)-\mathrm{MISE}(h')-(\mathrm{RICE}(h)-\mathrm{RICE}(h'))}{\mathrm{MISE}(h^*)}\right| \to 0\, a.s.$$

Now, given Lemmas 6, 7 and 8, this can be done by showing, respectively, that, as N tends to infinity,

$$\sup_{h,h'}\left|\frac{\left(\overline{\mathrm{ASE}}(h)-\overline{\mathrm{ASE}}(h')\right)-(\mathrm{CV}(h)-\mathrm{CV}(h'))}{\mathrm{MISE}(h^*)}\right| \longrightarrow 0 \qquad \text{a.s., and} \tag{16}$$

$$\sup_{h,h'}\left|\frac{\left(\overline{\mathrm{ASE}}(h)-\overline{\mathrm{ASE}}(h')\right)-(\mathrm{RICE}(h)-\mathrm{RICE}(h'))}{\mathrm{MISE}(h^*)}\right| \longrightarrow 0 \qquad \text{a.s.} \tag{17}$$

We show first (16). Since

$$\overline{\mathrm{ASE}}(h)-\mathrm{CV}(h) = \frac{1}{N}\sum_{i,j}(Y_{ij}-r(t_{ij}))^2\,w(t_{ij}) + \frac{2}{N}\sum_{i,j}(Y_{ij}-r(t_{ij}))\left(\hat{r}^{-i}(t_{ij})-r(t_{ij})\right)w(t_{ij}).$$

Thus, the theorem will be proved just by showing that

$$\sup_{h\in H_N}\frac{\left|N^{-1}\sum_{i,j}(Y_{ij}-r(t_{ij}))\,(\hat{r}^{-i}(t_{ij})-r(t_{ij}))\,w(t_{ij})\right|}{\mathrm{MISE}(h)} \longrightarrow 0 \qquad \text{a.s.}$$

Note that

$$\left| N^{-1} \sum_{i,j} (Y_{ij} - r(t_{ij})) \left(\hat{r}^{-i}(t_{ij}) - r(t_{ij}) \right) w(t_{ij}) \right| = A_1(h) + A_2(h),$$

where

$$A_1(h) = \left| \frac{1}{N^2 h} \left(\frac{m}{m-1} \right) \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k \neq i} \sum_{l=1}^{n} K \left(\frac{t_{ij} - t_{kl}}{h} \right) \frac{(r(t_{kl}) - r(t_{ij}))\epsilon(t_{ij})}{f(t_{ij})} \right|,$$

and

$$A_2(h) = \left| \frac{1}{N^2 h} \left(\frac{m}{m-1} \right) \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k \neq i} \sum_{l=1}^{n} K \left(\frac{t_{ij} - t_{kl}}{h} \right) \frac{1}{f(t_{ij})} \epsilon(t_{ij}) \epsilon(t_{kl}) \right|.$$

Hence, the proof of Theorem 3 will follow by showing that:

$$\sup_{h \in H_N} \frac{A_i(h)}{\mathrm{MISE}(h^*)} \to 0 \text{ a.s.}$$

for $i = 1, 2$. For $A_1(h)$, we write it as $N^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{l=1}^{n} V_{ijkl}$, where

$$V_{ijkl} = \frac{1}{h} \left(\frac{m}{m-1} \right) K \left(\frac{t_{ij} - t_{kl}}{h} \right) \frac{(r(t_{kl}) - r(t_{ij}))}{f(t_{ij})} \epsilon(t_{ij})$$

and $N^{-1} \sum_{ijkl} V_{ijkl} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \epsilon(t_{ij})$, with $a_{ij} = (Nhf(t_{ij}))^{-1} \sum_{k,i,l} K((t_{ij} - t_{kl})h^{-1})(r(t_{kl} - r(t_{ij}))I(k \neq i)$.

Hence, in our context, the order of a_{ij} is the same as the order of b_j in Lemma 4 on H-M. This means that if we prove that $E[(\sum_{ij} a_{ij} \epsilon(t_{ij}))^2] \leq C(\sum_{ij} a_{ij}^2)$ holds, the result follows straightforwardly, since the rest is immediate from their proof.

In order to do this, let us note that

$$\begin{aligned} E[\sum_{ij} a_{ij} \epsilon(t_{ij})]^2 &= E[\sum_{ij} \sum_{kl} a_{ij} c_{kl} \epsilon(t_{ij}) \epsilon(t_{kl})] = \sum_{ij} \sum_{l} a_{ij} c_{il} \gamma(j, l) \\ &= \sum_{ij} [a_{ij}^2 \gamma(j, j) + \sum_{l \neq j} a_{ij} a_{il} \gamma(j, l)] \leq C \sum_{ij} a_{ij}^2, \end{aligned}$$

by using Assumption **(E.4)**, and the same arguments as in the proof of Theorem 1.

For $A_2(h)$, we write it as $N^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{l=1}^{n} U_{ijkl}$, where

$$U_{ijkl} = \frac{1}{h} \left(\frac{m}{m-1} \right) K \left(\frac{t_{ij} - t_{kl}}{h} \right) \frac{\epsilon(t_{kl}) \epsilon(t_{ij})}{f(t_{ij})} I(k \neq i),$$

and $N^{-1} \sum_{ijkl} U_{ijkl} = \sum_{ijkl} b_{ijkl} \epsilon(t_{ij}) \epsilon(t_{kl})$, with $b_{ijkl} = (Nhf(t_{ij}))^{-1} K \left(\frac{t_{ij} - t_{kl}}{h} \right) I(k \neq i)$.

Hence, in our context, the order of b_{ijkl} is the same as the order of a_{ij} in Lemma 4 on H-M. This means that if we prove that $E[(\sum_{ijkl} b_{ijkl}\epsilon(t_{ij})\epsilon(t_{kl}))^2] \leq C(\sum_{ijkl} b_{ijkl}^2)$ holds, the result follows straightforwardly, since the rest is immediate from their proof.

For this note that $\mathrm{E}[\sum_{ijkl} b_{ijkl}\epsilon(t_{ij})\epsilon(t_{kl})]^2 = \mathrm{E}[\sum_{ijkl}\sum_{i'j'k'l'} b_{ijkl}b_{i'j'k'l'}\epsilon(t_{ij})\epsilon(t_{kl})\epsilon(t_{i'j'})\epsilon(t_{k'l'})]$. Using assumption **(E.2)** then the last term is equal to $\sum_{ijj'kll'} b_{ijkl}b_{ij'kl'}\gamma(j,j')\gamma(l,l')$, and using similar arguments than above this expression can be bounded by $C(\sum_{ijkl} b_{ijkl}^2)$. This closes the proof of Theorem 3. ∎

To finish the proof of Theorem 5, we point out that the numerator in equation (17) can be written as $C(h)+C(h')$, where

$$\begin{aligned} C(h) &= \frac{2}{N}\sum_{i,j}(Y(t_{ij})-r(t_{ij}))(\hat{r}(t_{ij})-r(t_{ij})) \\ &\quad -\frac{2}{N^2h}\sum_{i,j}\frac{1}{\hat{f}(t_{ij})}\sum_{l=1}^{n}K\left(\frac{t_{ij}-t_{il}}{h}\right)\hat{\gamma}(j,l). \end{aligned}$$

It is easy to see that the main terms of $C(h)$ are $2C_1(h)+2C_2(h)+2C_3(h)$, defined as

$$\begin{aligned} C_1(h) &= \frac{1}{N^2h}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{n}K\left(\frac{t_{ij}-t_{kl}}{h}\right)\frac{(r(t_{kl})-r(t_{ij}))\epsilon(t_{ij})}{f(t_{ij})} \\ C_2(h) &= \frac{1}{N^2h}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k\neq i}\sum_{l=1}^{n}K\left(\frac{t_{ij}-t_{kl}}{h}\right)\frac{1}{f(t_{ij})}\epsilon(t_{ij})\epsilon(t_{kl}) \\ C_3(h) &= \frac{1}{N^2h}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}K\left(\frac{t_{ij}-t_{il}}{h}\right)\frac{1}{f(t_{ij})}[\epsilon(t_{ij})\epsilon(t_{il})-\hat{\gamma}(j,l)]. \end{aligned}$$

Hence, the proof will be done when we show that

$$\sup_{h\in H_N}\frac{C_i(h)}{\mathrm{MISE}(h^*)}\to 0\; a.s.$$

for $i=1,2,3$. This result has already been proved for $i=1,2$. As for $C_3(h)$, we have that:

$$\begin{aligned} C_3(h) &= \frac{1}{N^2h\hat{f}(t_{ij})}\sum_{ij}\sum_{l}K\left(\frac{t_{ij}-t_{il}}{h}\right)(\epsilon(t_{ij})\epsilon(t_{il})-\gamma(j,l)+\gamma(j,l)-\hat{\gamma}(j,l)) \\ &= \frac{1}{N}\sum_{ijl}c_{ijl}Z_{ijl}+\frac{1}{N}\sum_{ijl}c_{ijl}\left(\gamma(j,l)-\hat{\gamma}(j,l)\right), \end{aligned}$$

where $c_{ijl}=b_{ijil}$ from the proof of Theorem 3, and $Z_{ijl}=\epsilon(t_{ij})\epsilon(t_{il})-\gamma(j,l)$.

Now, we have that $E(\sum_{ijl}c_{ijl}Z_{ijl})^2 = E(\sum_{i,j,l}\sum_{i'j'l'}c_{ijl}c_{i'j'l'}Z_{ijl}Z_{i'j'l'})$ $= \sum_{ijl}[c_{ijl}^2\mu(j,l,j,l)+o(c_{ijl}^2)]$, by using Assumption **(E.5)**. Now, the proof follows taking into account the order of c_{ijl}. The remainder term is straightforward since $E(\gamma(j,l)-\hat{\gamma}(j,l))^2=o(1)$, when m goes to infinity.

Proof of Theorems 4 and 6: First we write $\mathrm{CV}_G(h_G)$ as:

$$\mathrm{CV}_G(h_G) = \frac{1}{N_G}\sum_{i\in G}\sum_j \left[(Y_{ij} - r(t_{ij}) - \tilde{r}_G^{-i}(t_{ij})) + (r(t_{ij}) - \hat{r}(t_{ij})) + (\tilde{r}_G^{-i}(t_{ij}) - \hat{r}_G^{-i}(t_{ij}))\right]^2 w(t_{ij}),$$

where $\tilde{r}_G^{-i}$ has the same expression as $\hat{r}_G^{-i}$, by replacing $\hat{r}$ by r. Since the second term between brackets does not depend on h_G, and using Lemma 4 for the third term, the result is direct using the same arguments than in the proof in Theorems 3 and 5. In analogous way we get the same for Theorem 6. ∎

REFERENCES

Boularan, J., Ferré, L. and Vieu, P. (1994). Growth curves: a two-stage nonparametric approach. *Journal of Statistical Planning and Inference* **38**, 327-350.

Edgington, E.S. (1995). *Randomization Tests. Third Edition, Revised and Expanded.* New York: Marcel Dekker.

Ferreira, E., Núñez-Antón, V. and Rodríguez-Póo, J. (1997). Kernel regression estimates of growth curves using nonstationary correlated errors. *Statist. Proba. Let.* **34**, 413-423.

Fraiman, R. and Meloche, J. (1994). Smoothing dependent observations. *Statist. Proba. Let.* **21**, 203-214.

Gantz, B.J., Tyler, R.S., Knutson, J.F., Woodworth, G.G., Abbas, P., McCabe, B.F., Hinrichs, J., Tye-Murray, N., Lansing, C., Kuk, F. and Brown, C. (1988). Evaluation of five different cochlear implant designs: Audiologic assessment and predictors of performance. *Laryngoscope* **98**, 1100-1106.

Garthwaite, P.H., Jolliffe, I.T. and Jones, B. (1995). *Statistical Inference.* New York: Prentice Hall.

Good, P. (1994). *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses.* New York: Springer-Verlag.

Gýorfi, L., Härdle, W., Sardá, P. and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series.* Lecture Notes in Statistics, Vol. 70. New York: Springer-Verlag.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Härdle, W. and Marron, S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics* **13**, 1465-1481.

Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series. *J. Time Series Anal.* **13**, 209-232.

Hart, J.D. (1991). Kernel regression estimation with time series errors. *J. R. Statist. Soc. B* **1**, 173-187.

Hart, J.D. and Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080-1088.

Hart, J.D. and Wehrly, T.E. (1993). Consistency of cross-validation when the data are curves. *Stoch. Proc. Appl.* **45**, 351-361.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* Second Edition. New York: John Wiley and Sons.

Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects model for repeated measures data. *Biometrics* **46**, 673-687.

Marron, J.S. and Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *Journal of Multivariate Analysis* **20**, 91-113.

Núñez-Antón, V. and Woodworth, G.G. (1994). Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors. *Biometrics* **50**, 445-456.

Núñez-Antón, V., Rodriguez-Póo, J. and Vieu, P. (1999). Longitudinal data with nonstationary errors: a nonparametric three-stage approach. *Test.* Under revision.

Parzen, E. (1962). On estimation of a probability density and mode. *Annals of Mathematical Statistics* **35**, 1065-1076.

Rice, J.A. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* **84**, 1215-1230.

Rice, J.A. and Silverman, B.W. (1991). Estimating mean and covariance structure when the data are curves. *J.R. Statist. Soc. B*, **53**, 233-243.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density. *Annals of Mathematical Statistics* **27**, 642-669.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Zimmerman, D.L., Núñez-Antón, V. and El Barmi, H. (1997). Computational aspects of likelihood-based estimation of first-order antedependence models. *J. Statist. Comp. Simul.* **60**, 67-84.