Arribas-Gil, A., De la Cruz, R., Lebarbier, E., and Meza, C. (2015). Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. Biometrics, 71(2), 333–343

DOI: https://doi.org/10.1111/biom.12280

Web-based Supplementary Materials for "Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators"

by Ana Arribas–Gil[1,*], Rolando De la Cruz[2,], Emilie Lebarbier[3,***] , and Cristian Meza[4,****]**

[1]Departamento de Estadística, Universidad Carlos III de Madrid, Getafe, Spain

[2]Departamento de Salud Pública, Escuela de Medicina and Departamento de Estadística, Facultad de Matemáticas
Pontificia Universidad Católica de Chile, Santiago, Chile

[3]AgroParisTech UMR518, Paris 5e and INRA UMR518, Paris 5e, France

[4]CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile

**email:* ana.arribas@uc3m.es

***email:* rolando@med.puc.cl

****email:* emilie.lebarbier@agroparistech.fr

*****email:* cristian.meza@uv.cl

This Web-Appendix contains:

1. The choice of γ for prediction of pregnancy outcome from hormone levels (Section 5.1).
2. A simulation study demonstrating our method on a highly unbalanced design mimicking the pregnancy outcome dataset (Section 5.1).
3. A simulation study comparing our estimation method with the `R` function `slm` for balanced design (Section 5.2).

1. The choice of γ for prediction of pregnancy outcome from hormone levels (Section 5.1)

In order to choose the constant γ of the LASSO-type estimator, for the pregnant women dataset studied in this work we consider a grid of values for γ from 0.2 to 2.1 and we compute the misclassification error rate for each value of γ. It is important to note that our estimation method is based on a unified EM algorithm exploiting the fact that the model is linear, so the impact of the choice of the γ of the LASSO estimation part is hard to define in the complete iterative estimation procedure.

For our classification problem, we choose γ focusing on the misclassification error rate. In practice, for this dataset, γ is chosen so that the misclassification error rate is minimized. In Figure 1, we can see the misclassification error rate against different values of γ for this specific dataset. It is clear here that the best performance of our algorithm, for the classification problem, is obtained with $\gamma = 0.35$.

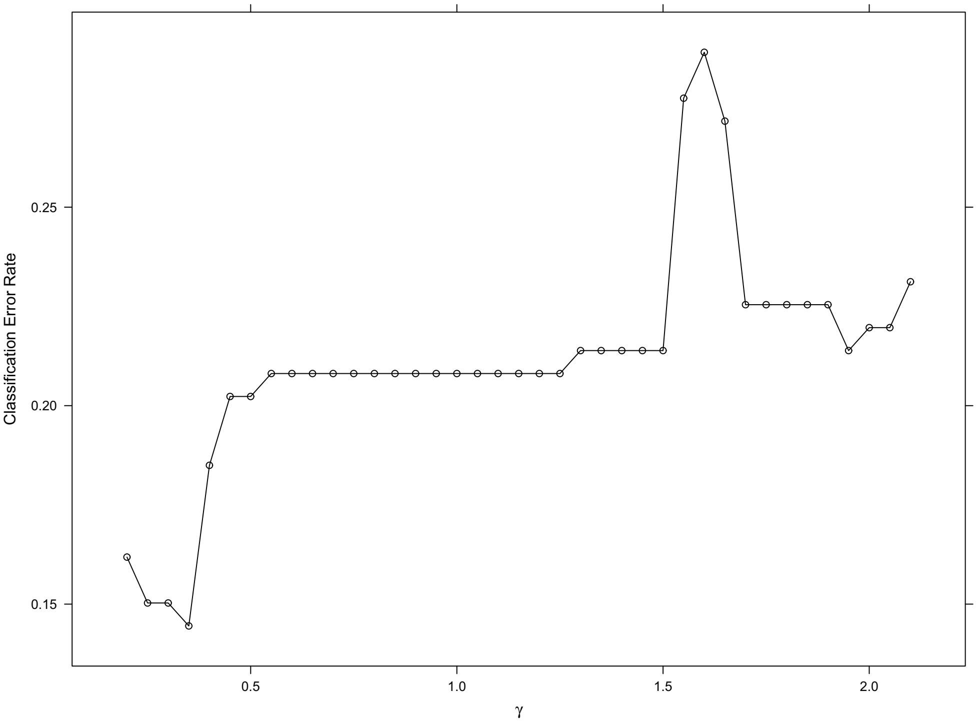

Figure 1. Pregnancy outcome from hormone levels: Error rate for different values of γ.

2. Simulation study on a highly unbalanced design (Section 5.1)

In order to assess the performance of our method in a scenario similar to the one provided by the real dataset we conducted the simulation study described below. The objective was to overcome the misspecification problems detected in the real dataset while still dealing with a highly unbalanced design and two uneven groups.

To this end, we used model (9) to sample observations from two groups that replicate the structure of the normal and abnormal groups in the real dataset. That is, we kept the number of individuals in each group and for the simulated normal group we also preserved the observed time points for each individual. For the simulated abnormal dataset, however, we did not retain the observed design. Indeed, we believe that its highly sparse structure makes the fitting of this group very difficult. Then, we replaced the observed time points of the abnormal group by those of 49 randomly chosen individuals from the normal group. In this way, the level of "unbalance" or "sparsity" in both simulated groups was similar. The data was simulated as follow:

- $Y_{ij}^{n} = \boldsymbol{\phi}_i + \bar{f}_n(t_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, 124, \;\; j = 1, \ldots, n_i$ with $\boldsymbol{\phi}_i \sim \mathcal{N}(0, \bar{\sigma}_{\phi,n}^2)$ i.i.d. and $\varepsilon_{ij} \sim \mathcal{N}(0, \bar{\sigma}_n^2)$ i.i.d. independent. The design points t_{ij}, $i = 1, \ldots, 124$, $j = 1, \ldots, n_i$ are those observed in the real dataset for the normal group, and $\bar{f}_n$, $\bar{\sigma}_{\phi,n}^2$, $\bar{\sigma}_n^2$ are the parameter estimates obtained for the normal group in the real dataset analysis.
- $Y_{ij}^{ab} = \boldsymbol{\phi}_i + \bar{f}_{ab}(t_{k_i j}) + \varepsilon_{ij}, \quad i = 1, \ldots, 49, \;\; j = 1, \ldots, n_{k_i}$ with $\boldsymbol{\phi}_i \sim \mathcal{N}(0, \bar{\sigma}_{\phi,ab}^2)$ i.i.d. and $\varepsilon_{ij} \sim \mathcal{N}(0, \bar{\sigma}_{ab}^2)$ i.i.d. independent. The design points $t_{k_i j}$, $i = 1, \ldots, 49$, $j = 1, \ldots, n_{k_i}$ correspond to 49 randomly chosen individuals from the normal group, that is $k_i \in \{1, \ldots, 124\}$, $i = 1, \ldots, 49,$. Also, $\bar{f}_{ab}$, $\bar{\sigma}_{\phi,ab}^2$, $\bar{\sigma}_{ab}^2$ are the parameter estimates obtained for the abnormal group in the real dataset.

We have adjusted $\bar{f}_n$ and $\bar{f}_{ab}$ up to an additive constant so that sample means of observations in each group of the simulated dataset equal those of the real dataset. Then, we have simulated 150 datasets in this way, with the 49 randomly chosen individuals requested to sample the abnormal group possibly varying from one dataset to another. We have ran the EM algorithm with the specifications used for the real dataset and described above. In Table 1 we present the average classification result over 150 samples. There is a clear improvement in classification with respect to the real dataset, with a decrease in the number of misclassified individuals in the two groups and a reduction in the global error rate of almost 3 points (11.8% versus 14.4% of the within sample error rate). However, there is still a large difference between the group specific misclassification error rates, which are 38.6% in the abnormal group versus 1.2% in the normal group. This is due to the complexity of the datasets (the real and the simulated one), for which many of the trajectories of individuals in the abnormal group present the shape characteristic of the normal group. In this framework any classification method will fail to achieve perfect classification.

Table 1
Classification in the simulation study: Within sample and cross–validation average error rates over 150 simulated datasets.

	Within sample error rate			*Leave-one out C.V.*		
Group	*Normal*	*Abnormal*	*Total*	*Normal*	*Abnormal*	*Total*
Normal	122.56	1.44	124	119.63	4.37	124
Abnormal	18.91	30.09	49	17.37	31.63	49
Total	141.47	31.53	173	137.00	36.00	173

3. Simulation study and comparison with `slm` function for balanced data (Section 5.2)

As described in Section 5.2, we have generated 100 datasets each one consisting on two groups of observations with values described in Table 2. The objective was to compare our method with that of Bayes classification in SLMM's based on smoothing splines estimation of f. The difference between this simulation study and the one presented in Section 5.2 is

Table 2
Parameter values used in the simulation study

	Group 1	Group 2
N	10	10
m	20	20
A	1	1.5
ω	50	80
a_1	10	20
a_2	11	21
a_3	50	40
a_4	60	70
a_5	62	72
b_1	1	2
b_2	1	2
b_3	-2	-1
b_4	4	3
b_5	5	4
σ^2	0.25	0.25
σ^2_ϕ	1	1
μ_X	10	10
σ^2_X	1	1
p_X	0.3	0.3
θ_1	0.5	0.3
θ_2	0.1	-0.1

that here we explore the case of a perfectly balanced dataset (same number of individuals in the two groups and same mean number of observations per individual).

Indeed, we considered 10 individuals for each group and the resulting median number of observations per individual was 17 in both groups. This design is kept fixed for all the datasets. We have then estimated the model parameters separately in each group. For our estimation procedure we have ran the EM algorithm with $\gamma = 0.70$ and using the same dictionary used for the real dataset adding 128 Haar functions defined by ($t \mapsto 2^{7/2}\mathbb{1}_{[0,1]}\left(\frac{2^7 t}{100} - k\right), k = 0, \ldots, 2^7 - 1$). For smoothing splines estimation we have considered the `slm` function of `assist` package with f is periodic with period equal to 1 and $\int_0^1 f = 0$, i.e. $f \in W_2^0(per) = W_2(per) \ominus span\{1\}$ where $W_2(per)$ is the periodic Sobolev space of order 2 in L^2 and $span\{1\}$ represents the set of constant functions.

In Table 3 we present the average classification results over 100 samples. There is a clear

Table 3
Classification in the simulation study: Average error rates over 100 simulated datasets by using semiparametric LME (SLMM) and smoothing splines.

	SLMM			*Smoothing Splines*		
Group	*Group 1*	*Group 2*	*Total*	*Group 1*	*Group 2*	*Total*
Group 1	9.97	0.03	10	7.83	2.17	10
Group 2	0.56	9.44	10	1.77	8.23	10
Total	10.53	9.47	20	9.6	10.4	20

improvement in classification with respect to Ke and Wang's approach since we obtained a small average number of misclassified individuals (smaller than one) using our method. Using smoothing splines in this simulation framework, we obtained a larger average number of misclassified individuals (close to four). It looks like our method's performance improves in a balanced scenario, compared to the results obtained under an unbalanced design (see Section 5.2). However, in both cases it outperforms the classification results obtained with the smoothing splines estimation based method.