A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications

Article *in* Computer Methods and Programs in Biomedicine · April 2017

DOI: 10.1016/j.cmpb.2017.02.019

CITATIONS
66

READS
396

4 authors:

Lizbeth Naranjo
Universidad Nacional Autónoma de México
25 PUBLICATIONS **224** CITATIONS
SEE PROFILE

C. J. Pérez
Universidad de Extremadura
104 PUBLICATIONS **1,253** CITATIONS
SEE PROFILE

Jacinto Martín
Universidad de Extremadura
87 PUBLICATIONS **1,593** CITATIONS
SEE PROFILE

Yolanda Campos-Roca
Universidad de Extremadura
51 PUBLICATIONS **651** CITATIONS
SEE PROFILE

A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications

Lizbeth Naranjo[a,1], Carlos J. Pérez[b], Jacinto Martín[b], Yolanda Campos-Roca[c]

[a]*Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, México D.F. (Mexico)*
[b]*Departamento de Matemáticas, Universidad de Extremadura, Cáceres (Spain)*
[c]*Departamento de Tecnologías de los Computadores y las Comunicaciones, Universidad de Extremadura, Cáceres (Spain)*

Abstract

Background and objectives: In the scientific literature, there is a lack of variable selection and classification methods considering replicated data. The problem motivating this work consists in the discrimination of people suffering Parkinson's disease from healthy subjects based on acoustic features automatically extracted from replicated voice recordings. **Methods:** A two-stage variable selection and classification approach has been developed to properly match the replication-based experimental design. The way the statistical approach has been specified allows that the computational problems are solved by using an easy-to-implement Gibbs sampling algorithm. **Results:** The proposed approach produces an acceptable predictive capacity for PD discrimination with the considered database, despite the fact that the sample size is relatively small. Specifically, the accuracy rate, sensitivity and specificity are 86.2%, 82.5%, and 90.0%, respectively. However, the most important fact is that there is an improvement in the interpretability of the results at the same time that it is shown a better chain mixing and a lower computation time with respect to the only-classification approaches presented in the scientific literature. **Conclusions:** To the best of the authors' knowledge, this is the first approach developed to properly consider intra-subject variability for variable selection and classification. Although the proposed approach has been applied for PD discrimination, it can be applied in other contexts with similar replication-based experimental designs.

Keywords: Bayesian binary regression; Gibbs sampling; Parkinson's disease; Replicated measurements; Variable selection; Voice features.

1. Introduction

People with Parkinson's Disease (PD) exhibit a chronic neurological disorder caused by the progressive degeneration and death of dopaminergic neurons. These neurons play a key role in coordinating movement at level of muscular tone. An estimated 7 to 10 million people worldwide are living with this medical condition.

Voice and speech are also affected in people with PD, since they depend on laryngeal, respiratory and articulatory functions. Vocal impairment is hypothesized to be one of the earliest signs of the disease ([1]). Since the early stages of PD, there are subtle abnormalities in speech that might not be perceptible to listeners, but they could be evaluated in an objective way by performing acoustic analyses on recorded speech signals ([2]).

In recent years numerous techniques have been developed to assess speech-related diseases. The monograph presented by [3] provides a view of the state of the art concerning automatic classification of speech signals for clinical purposes. Some authors have considered measures extracted from speech recordings to discriminate healthy people from those with PD (see, e.g., [4], [5], [6], [7], [8], [9] and [10]). In these investigations, the voices of the subjects are recorded to extract some specific features of the signals and use them to classify individuals by using different methods. The Parkinson's Voice Initiative has played an important role in the spread of this topic[2].

Successfully addressing early diagnosis of people with PD is a key issue to improve patients' quality of life. Note that making an accurate diagnosis of PD, particularly in its early stages, is difficult and may take years. Unfortunately, no single definitive diagnostic test is currently available for PD, and accurate diagnosis has been a significant challenge, especially among clinicians without particular expertise in movement disorders. Diagnosis relies on clinical information provided by PD patients and findings of neurological exams performed by expert

Email addresses: lizbethna@ciencias.unam.mx (Lizbeth Naranjo), carper@unex.es (Carlos J. Pérez), jrmartin@unex.es (Jacinto Martín), ycampos@unex.es (Yolanda Campos-Roca)

[1]Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Av. Universidad 3000, Circuito Exterior S/N, Delegación Coyoacán, D. F. México, 04510, Tel: +52 55 5622 3899, Ext.: 45735.

[2]http://www.parkinsonsvoice.org/

clinical staff. Classification techniques based on acoustic features have a great potential to provide an automatic early diagnosis of the disease ([11]). Undiagnosed subjects can benefit from these techniques, since it is estimated that 20% of people with PD remain undiagnosed ([12]).

In this context, it has been usual to make replicated voice recordings for each individual and treat the extracted features with independence-based classification methods (see, e.g., [4], [5] and [7]). Since features are extracted from multiple voice recordings from the same subject, in principle, the features from each individual should be identical. Technology imperfections and biological variability result in nonidentical replicated features that are more similar to one another than features from different subjects. This within-subject variability must be statistically addressed, which is not usually considered in the scientific literature. [9] and [10] proposed classification approaches for PD detection that take into account the underlying within-subject dependence. These methods are based on the introduction of latent variables that represent the true unknown features from which the replicated features are considered as measurements with error. [10] also performed a voice recording replication-based experiment to obtain a new feature database composed of 3 voice recording replications of the sustained /a/ phonation for 80 subjects (40 of them with PD). 44 features were extracted from each voice recording. All features were used in the classification approach, in spite of the fact that many of them were highly correlated. Even more, there are 5 groups of features that have related formulation. This suggests that variable selection considering replicated data must be performed in order to avoid redundant information.

The multicollinearity problem occurs when there are strong linear dependencies among the explanatory variables, leading to high correlations among these variables in such a way that the precision of some estimates of the correlated predictors can be degraded. Multicollinearity may bias the parameter estimates and makes them more unstable. The existence of multicollinearity may result in large standard errors for the parameter estimates or in parameter estimates with opposite signs of those expected. This can affect the parameter interpretability. Multicollinearity is a problem that can be solved with variable selection [13]. The linear regression model assumes each predictor has an independent effect on the response that can be included in the regression parameter. When predictors are highly correlated, the data do not contain much information on the independent effects of each variable. The data are deficient for determining the independent effects of a covariate on the response because the covariates themselves are not independent. One solution is using an algorithm as the one proposed by [13] that

accounts for correlation among the predictors by simultaneously performing predictor selection and clustering. One of the solutions summarized by [14] is to remove all but one of the highly correlated variables from the analysis. In order to address this problem and provide more interpretable results, a variable selection approach is necessary.

There are many variable selection methods proposed in the scientific literature from both frequentist and Bayesian perspectives, see e.g., [15] and [16]. [17] proposed a Bayesian approach to gene selection and classification using the logistic regression model. They used Gibbs sampling and Markov chain Monte Carlo (MCMC) methods to discover important genes. [18] proposed a Bayesian variable selection approach to multinomial probit models, where the number of predictors substantially exceeds the sample size. A two-level hierarchical Bayesian model for variable selection which assumes a prior that favors sparseness was presented by [19]. [20] proposed a simple Bayesian logistic regression approach that uses a Laplace prior to avoid overfitting and produces sparse predictive models for text data. A Bayesian stochastic variable selection approach for gene selection based on a probit regression model with a generalized singular g-prior distribution for the regression coefficients was proposed by [21]. [22] provided an overview of several Bayesian variable selection methods in the unified framework of Bayesian hierarchical models, and highlighted discrepancies and connections among them. Finally, [23] and [24] summarized several variable selection methods from a Bayesian perspective. In spite of the great number of variable selection approaches, up to the authors' best knowledge, replicated data have not been properly considered in any variable selection approach.

In this paper, a two-stage variable selection Bayesian approach that handles replicated measurements is proposed. The first stage is a pre-selection step based on a filter method to reduce the number of variables to one per feature group. The method chooses the variable that has minimal accumulated dissimilarity with respect to the other ones in its group and it is considered to represent it. Then, in the second stage, a regularization-based classification approach based on LASSO (Least Absolute Shrinkage and Selection Operator) is proposed. This approach integrates ideas from [20] and [10]. Computational difficulties are avoided by developing an easy-to-implement Gibbs sampling-based algorithm. In some contexts, the second stage can be used without the need of a pre-selection step.

The outline of the paper is as follows. In Section 2, the motivating problem is presented and discussed. Section 3 presents the proposed two-stage variable selection approach that handles replicated measurements. Section 4 shows the experimental results both with real and simulated data. A discussion is presented

in Section 5. Finally, Section 6 presents the main conclusion. Some technical details are presented in Appendix A.

2. Motivating problem

[10] conducted an experiment to discriminate PD subjects from healthy individuals by considering replicated voice recordings. A total of 40 people affected by PD and 40 healthy individuals were considered. The research protocol consisted of a brief questionnaire and three recording replications of the sustained /a/ phonation, leading to a total of 240 voice recordings. All subjects signed an informed consent and the protocol was approved by the Bioethical Committee from the University of Extremadura.

Each voice recording was processed to provide 44 acoustic features, i.e. a 44-dimensional vector per voice recording. The extracted features are divided into several groups based on whether they have related formulation or not. This leads to the following 9 groups, 4 of them composed by only one feature:

C_1: Pitch local perturbation measures. Relative jitter, absolute jitter, relative average perturbation (RAP), and pitch perturbation quotient (PPQ).

C_2: Amplitude perturbation measures. Local shimmer, shimmer in dB, 3-point amplitude perturbation quotient (APQ3), 5-point amplitude perturbation quotient (APQ5), and 11-point amplitude perturbation quotient (APQ11).

C_3: Harmonic-to-noise ratio measures. Harmonic-to-noise ratio in the frequency band 0-500 Hz (HNR05), in 0-1500 Hz (HNR15), in 0-2500 Hz (HNR25), in 0-3500 Hz (HNR35), and in 0-3800 Hz (HNR38).

C_4: Mel frequency cepstral coefficient-based spectral measures of order 0 to 12 (MFCC0, MFCC1, . . . , MFCC12).

C_5: Derivatives of Mel frequency cepstral coefficients, where the numbers 0-12 represent the order of the MFCC coefficient to which the derivative is applied (Delta0, Delta1, . . . , Delta12).

C_6: Recurrence period density entropy (RPDE).

C_7: Detrended fluctuation analysis (DFA).

C_8: Pitch period entropy (PPE).

C_9: Glottal-to-noise excitation ratio (GNE).

Within each one of the five first groups, the variables are highly pairwise correlated, providing similar information. For example, in C_1, the relative jitter is defined as the average absolute difference between consecutive periods, divided by the average period, i.e.:

$$\text{Relative jitter} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i},$$

where T_i is the pitch period, whereas RAP is the average absolute difference between a period and the average of it and its two neighbors, divided by the average period, i.e.:

$$\text{RAP} = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|T_i - \frac{T_{i-1}+T_i+T_{i+1}}{3}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i}.$$

In this case, the Pearson correlation coefficient of these two features is 0.99 (as defined in Subsection 3.1). All the pairwise correlations in this group C_1 are above 0.89. Then, in order to avoid a multicollinearity problem, it is reasonable that only one feature per group is selected in a first stage. This leads to one feature in each of the nine groups. In a second stage, these nine features are considered in a regularization-based variable selection and classification method.

If the feature vectors were not replicated, the problem of variable selection and classification could be performed by using traditional machine learning approaches based on independent instances. However, the data have a dependent nature that must be treated with a model that properly matches the experimental design. This excludes methods based on votes, i.e. to consider all feature vectors of each subject as if they were independent and use a traditional machine learning method that considers the 240 feature vectors as independent instances. After applying the method, each feature vector coming from each recording is classified as PD or healthy, but the final decision on whether the subject is classified as having PD or not is based on counting the number of PD or healthy classifications for each individual. This leads to some individuals with some voice recordings considered as healthy and some others considered as PD-affected. One option to avoid this conceptual problem would be to aggregate the feature vectors for each subject (e.g., by using means), leading to a unique vector per subject and to apply traditional machine learning methods. However, this approach removes the underlying within-subject variability.

In this context, and in many other ones, technology imperfections and the biological variability result in nonidentical replicated feature vectors for each individual. This suggests to use the replicated feature vectors as if they were measured with error and to consider them as surrogates of a true unknown feature vector representing each subject. Although the within-group variability is higher than the within-subject variability, the latter is important enough to be considered in the statistical approach.

This problem motivates the development and application of the proposed approach. To the best of the authors' knowledge, it is the first one that integrates variable selection and classification for replicated data by taking into account the within-subject variability. More information on participants, speech recordings, and feature extraction processes can be found in [10].

3. A two-stage variable selection approach

Suppose that n independent binary random variables $Y_1, \ldots, Y_n$ are observed, where Y_i is distributed as a Bernoulli with success probability $\mathrm{P}(Y_i = 1) = p_i$, $i = 1, \ldots, n$. The probabilities p_i are related to two sets of covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, where $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iG})'$ is a $G \times J_i$ matrix of a set of G covariates which have been measured with J_i replicates, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})'$ is a vector of a set of H covariates which are exactly known. The notation J_i means that the number of replications can be different for each individual. Let $\boldsymbol{x}_{ij_i} = (x_{i1j_i}, \ldots, x_{iGj_i})$ be the j_i-th replication of the unknown covariate vector $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iG})$, $j_i = 1, \ldots, J_i$, and assume that they have a linear relationship represented by an additive measurement error model (see e.g., [25] and [26]). This implies that instead of the covariates $\boldsymbol{w}_i$, their replicates $\boldsymbol{x}_{ij_i}$ are observed, i.e., the $\boldsymbol{x}_{ij_i}$ are the surrogates of $\boldsymbol{w}_i$. The parameters p_i are related to $\boldsymbol{x}_{ij_i}$ and $\boldsymbol{z}_i$ through the following hierarchical model, that extends the one proposed in [10] to allow for a different number of replications for each individual:

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(p_i), \\
\Psi^{-1}(p_i) &= \boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z, \\
\boldsymbol{x}_{ij_i} &= \boldsymbol{w}_i + \boldsymbol{\varepsilon}_{ij_i}, \\
\boldsymbol{\varepsilon}_{ij_i} &\sim \text{Normal}_G(\mathbf{0}, \boldsymbol{\Gamma}),
\end{aligned}
$$

for $i = 1, \ldots, n$ and $j_i = 1, \ldots, J_i$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_x', \boldsymbol{\beta}_z')'$ is a $(G + H)$ vector of unknown parameters, $\Psi^{-1}(\cdot)$ is a known nonnegative and nondecreasing function

ranging between 0 and 1, usually the inverse of the cumulative distribution function (cdf) of normal or logistic distributions, and $\mathbf{\Gamma}$ is a $G \times G$ matrix of variances and covariances. The error vector $\boldsymbol{\varepsilon}_{ij_i}$ is independent of $\boldsymbol{w}_i$, implying that $\boldsymbol{x}_{ij_i}$ is a surrogate of $\boldsymbol{w}_i$.

In the next subsection, a filter method is proposed to reduce the number of variables to one per group. This constitutes the first stage of the approach.

3.1. Correlation-based variable reduction

Let $C_1, C_2, \ldots, C_G$ be the groups of involved variables. It is assumed that the variables in each group share related formulation and they are highly pairwise correlated. This is the case for the motivating problem and, since the physical origin of the features within each group is the same, we propose to use only one variable per group in the subsequent classification approach. The objective is to reduce the number of redundant variables, while keeping the different information provided in all the variable groups. We define a simple criterion to perform this.

Let $\rho(x_k, x_l)$ be the correlation between variables x_k and x_l. For each group of variables, $g = 1, \ldots, G$, we propose to define the following measure of dissimilarity between x_k and x_l:

$$d(x_k, x_l) \quad = \quad 1 - |\rho(x_k, x_l)|\,, \qquad k, l \in C_g,$$

and define a measure of discrepancy for each variable with respect to the rest of the variables in its group as:

$$\delta(x_k) \quad = \quad \sum\nolimits_{l \in C_g} d(x_k, x_l), \qquad k \in C_g.$$

Then, the variable selection criterion consists in finding the variable k_g having the minimum discrepancy in each group, i.e., $x_{k_g} = \arg\min_{k \in C_g} \delta(x_k)$. This provides one representing variable per group. Obviously, when the group has only one variable, the particular application of this criterion provides the only variable as the representing one.

The remaining task is to determine the correlation estimate that will be used, taking into account that it must handle replicated data. We propose to use the technique presented by [27] in a gene expression array context for a different task. This method is based on a covariance structure that explicitly models within-subject and between-subject correlations of the data, i.e., it adequately handles replicated data. It requires that the feature vectors are distributed as a multivariate normal

in order to assure that the maximum likelihood estimation can be obtained. The statistical inference procedures are implemented in the `R` package `CORREP` ([28]). However, multivariate normality is a strong condition that may not be assumed for one or more groups of features. As an alternative, the Pearson correlation coefficient for replicated data, defined as follows, can be used:

$$\rho(x_k, x_l) \quad = \quad \frac{\sum_{i=1}^{n}(\bar{x}_{ik} - \bar{x}_k)(\bar{x}_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{n}(\bar{x}_{ik} - \bar{x}_k)^2}\sqrt{\sum_{i=1}^{n}(\bar{x}_{il} - \bar{x}_l)^2}},$$

where the notations $\bar{x}_{ik}$ and $\bar{x}_k$ are the means and the grand means, specifically, $\bar{x}_{ik} = \frac{1}{J_i}\sum_{j=1}^{J_i} x_{ikj}$ and $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_{ik}$.

The Pearson coefficient is a theoretically less efficient measure, since it introduces strong bias and reduces the variance. However, in practice, it provides good results and it is very easy to calculate. Therefore, for each particular group of features C_g, the method proposed by [27] is intended for application. When the applicability condition is not satisfied, the Pearson correlation coefficient for replicated data is applied.

After one variable per group has been obtained in this first stage, a regularization variable selection and classification approach is considered for the representing variables in the next subsection. Note that, in a different context, the particular case where each group contains only one variable directly leads to the application of the second stage.

3.2. LASSO-based approach

One of the most commonly used penalized regression methods is LASSO (see [29]). LASSO is a regularization technique for simultaneous estimation and variable selection. It does not remove variables, but favors the best predictors and penalizes the worst ones through a parameter regularization. A wide variety of Bayesian LASSO methods has been developed and published in the last years (see e.g., [30], [31], [32], and [33]).

The regularization-based variable selection and classification approach that will be proposed here integrates ideas from [20] and [10]. Considering the probit link function for the replication-based binary regression model presented in this section, a data augmentation framework based on [34] is defined. Specifically, n independent latent variables $u_1, \ldots, u_n$ are introduced, where u_i is distributed as $u_i \sim \text{Normal}(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z, 1)$, and it is defined $Y_i = 1$ if $u_i > 0$, and $Y_i = 0$ if $u_i \leq 0$. This leads to an augmented likelihood given by:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Gamma} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}) \quad = \quad f(\boldsymbol{y} \mid \boldsymbol{u}) f(\boldsymbol{u} \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}) f(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{\Gamma}) f(\boldsymbol{w}).$$

Now, the prior distributions must be defined. The prior distributions for $\mathbf{\Gamma}$ and $\boldsymbol{w}_i$ are the same as in [10], i.e., a conjugate prior distribution is considered for the variance and covariance matrix, $\mathbf{\Gamma} \sim \text{InvWishart}_G(\boldsymbol{V}, \nu)$, and the unknown covariate vectors are independently normal distributed, $\boldsymbol{w}_i \sim \text{Normal}_G(\boldsymbol{\mu}, \mathbf{\Sigma})$, where ν, $\boldsymbol{V}$, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are fixed values. The prior distribution for the regression parameters β_k is based on the proposal of [20], i.e., a Laplace prior distribution is considered, $\beta_k \sim \text{Laplace}\left(0, \gamma_k^{-1/2}\right)$, $k = 1, \ldots, G + H$, where γ_k can be fixed values or hyperprior distributions (e.g., $\gamma_k \sim \text{Gamma}(r, d)$). The Laplace prior distribution is represented as a scale mixture of normal distributions with independent exponentially distributed variances, i.e.:

$$\pi(\beta_k) \quad = \quad \int_0^\infty \pi(\beta_k|\tau_k)\pi(\tau_k)\, \mathrm{d}\tau_k,$$

where

$$\begin{aligned} \beta_k \mid \tau_k &\sim \text{Normal}(0, \tau_k), \\ \tau_k &\sim \text{Exp}(\gamma_k/2). \end{aligned}$$

Note that τ_k coefficients are nonnegative regularization parameters to control the size of the regression coefficients. The method shrinks the k-th regression coefficient toward 0 as τ_k decreases, i.e., a small posterior value of τ_k is associated with a small shrinkage of β_k. Note that when there is no initial information on the regularization parameters, the hyperparameters γ_k must be fixed in such a way that the variances of τ_k are high. When initial information is obtained based on the feature importance for classification, the hyperparameters can be selected to favor or penalize concrete variables. Then, the potentiality of the Bayesian paradigm can be exploited, even more.

The joint posterior density is not analytically tractable. However, the way this model has been specified allows to derive an easy-to-implement Gibbs sampling algorithm [35]. The full conditional distributions necessary to generate from the posterior distribution can be found in Appendix A. A graphical representation of the proposed regularization-based variable selection and classification approach is presented in Figure 1. This graphical representation is based on doodle objects of WinBugs [36]. It represents a direct acyclic graph, where the nodes are the model variables and the arrows show dependencies between them. There are two rectangular frames representing sets of identical repeating operations. One panel is indexed by i and ranges from 1 to n (subjects), whereas the second panel is indexed by j_i and ranges from 1 to J_i (replications for each subject). The stochastic variables Y_i and $\boldsymbol{w}_i$ (response and latent variables) are represented by oval nodes with

the heads of the simple arrows pointing to them. They are dependent variables in the hierarchical model, depending on the variables and parameters from which their arrows start. The covariates $\boldsymbol{z}_i$ and $\boldsymbol{x}_{ij_i}$ are represented by rectangular boxes, being independent variables in the hierarchical model. The parameter p_i depends on $\boldsymbol{\beta}$, $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ in a deterministic way by the relationship $p_i = \Psi(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z)$, and this is represented by double-lined arrows pointing to it. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ and $\boldsymbol{\Gamma}$ are stochastic, having their distributions depending on other hyperparameters. Moreover, $\boldsymbol{\beta}$ depends on $\boldsymbol{\tau}$. Finally, the hyperparameter of $\boldsymbol{\tau}$ is $\boldsymbol{\gamma}$, the hyperparameters of $\boldsymbol{\Gamma}$ are $\boldsymbol{V}$ and ν, and the hyperparameters of $\boldsymbol{w}_i$ are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. They have been represented by rectangular boxes because they are fixed values.

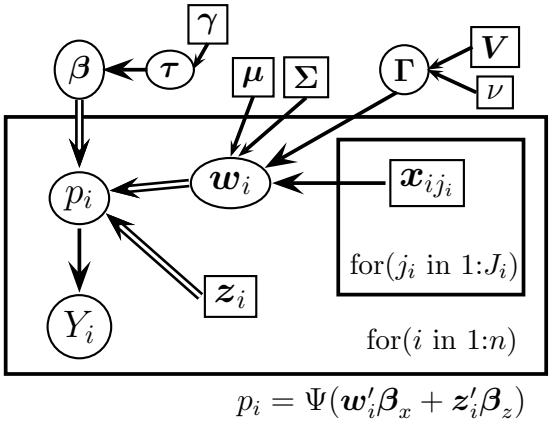

Figure 1: Flowchart for the proposed approach.

Probability predictions for future observations $\boldsymbol{y}^*$ are based on the predictive distribution. Specifically, latent variables $\boldsymbol{w}_i$ are generated from the multivariate normal distribution $\text{Normal}_G(\boldsymbol{m}_i, \boldsymbol{M}_i)$, where

$$\begin{aligned} \boldsymbol{m}_i &= \boldsymbol{M}_i\left(\sum\nolimits_{j_i=1}^{J_i} \boldsymbol{\Gamma}^{-1}\boldsymbol{x}_{ij_i} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \\ \boldsymbol{M}_i &= \left(J_i\boldsymbol{\Gamma}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}. \end{aligned}$$

Then, it is computed $p_i = \Psi(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z)$, and $y_i \sim \text{Bernoulli}(p_i)$. Note that $\boldsymbol{\Gamma}$, $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_z$ are the samples generated from the posterior distribution following the algorithm in Appendix A.

4. Experimental results

In this section the proposed methodology is applied to a PD real dataset and to simulation-based scenarios.

4.1. Parkinson's data application

The dataset considered here is the same as in [10], so the performances of both approaches can be compared. The acoustic variables have been individually normalized to have mean 0 and standard deviation 1, and the variable sex Z takes values $z = 0$ for men and $z = 1$ for women. The response variable Y takes values $y = 0$ for healthy subjects and $y = 1$ for people with PD.

Firstly, the filter procedure proposed in Subsection 3.1 is independently applied to groups C_g, $g = 1, \ldots, 5$, giving as representing features: relative jitter, local shimmer, HNR35, MFCC3, and Delta8. The other four groups have only one feature each, so all of them are considered, i.e.: RPDE, DFA, PPE, and GNE. Finally, variable sex is the only one that does not have replications and it is also considered since the approach is able to deal with covariates with and without replications. Then, a reduction from 45 variables to 10 is obtained ($G = 9$ and $H = 1$). These 10 variables are used in the second stage explained in Subsection 3.2.

The following hyperparameters were used for the prior distributions. $\boldsymbol{\mu}$ represents the sample mean vector of the normalized covariates, so all its elements are zeros. $\boldsymbol{\Sigma}$ is the sample correlation matrix of the normalized variables, so it is a matrix composed of a diagonal of ones. The hyperparameters of $\boldsymbol{\Gamma}$ are $\nu = G$ and $\boldsymbol{V} = \text{diag}_G(1)$. Finally, the hyperparameters of τ_k are $\gamma_k = 0.1$, so that the variances of τ_k are large for $k = 1, \ldots, G + H$. These specifications have been considered because the authors do not have any prior information on these parameters, so a noninformative setting is considered. This is usual in many situations, but the model also supports information on the prior parameters when it is available.

For comparative purpose, two approaches have been considered: A) the one proposed in this paper, and B) the one proposed by [10]. Gibbs sampling algorithms for both approaches have been applied. The convergence assessment has been performed by using BOA package (see [37]), with a total of 50,000 iterations for each approach. Firstly, the convergence diagnostic method in [38] was considered. Sample size requirements were sought to ensure that posterior estimates of the 0.025 tail probabilities (quantiles) would be within a ± 0.01 accuracy with probability equal to 0.95 with a precision of 0.01. The results for approach A (approach B) suggest that around 76,500 (91,000) samples should be generated,

the first 213 (238) of which are discarded as a burn-in sequence and every 56 (59) are saved as a thinning of the chain. For both models, the lower bound indicates that 937 independent samples are needed to estimate the posterior probability. On the other hand, the diagnostic in [39] with level of confidence of 0.05 and accuracy of 0.1 indicates that after a burn-in of 3,182 (4,565), all the iterations are retained for posterior inference. Note that a 1.44 times greater burn-in is necessary for approach B. Then, there is no significant evidence of non-stationarity based on Cramer-von Mises test statistics. Both diagnostics support that approach A needs a shorter chain to converge. The R code has been run on a computer with a 2.8 GHz Intel Core i7 processor and 4 GB 1333 MHz DDR3 RAM memory. The computation times for both approaches were 29.2 and 55.2 minutes, respectively. This means that Approach A takes, approximately, half the time as Approach B.

Once the posterior samples have been generated, predictive probabilities are obtained for each subject and a confusion matrix is built to obtain the classification measures. Note that all individuals are used in the training and testing sets. Later, a cross-validation procedure will be implemented. The accuracy rate obtained from the predictive distribution is 0.862. Sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)), and precision (TP/(TP+FP)) are 0.825, 0.900 and 0.891, respectively. The accuracy rate is 1.6% higher than the corresponding one in [10]. Slight differences with a maximum of 2.5%, favoring the proposed approach, are obtained for the remaining measures. Receiver Operating Characteristic (ROC) curves for both approaches are presented in Figure 2. The AUC (area under the curve) for approach A is 0.951, whereas for approach B is 0.956, which are practically the same.

AUC criterion is focused on classification. However, there are other usual goodness-of-fit criteria, such as, for example, the Bayesian information criterion (BIC, [40]) and the Akaike information criterion (AIC, [41]), which should be also taken into account. These criteria are evaluated as the following: $BIC = \overline{D(\eta)} + K\log(n)$ and $AIC = \overline{D(\eta)} + 2K$, where $D(\eta) = -2\log L(\eta)$ is the deviance of the model, $L(\eta)$ is the likelihood, $\overline{D(\eta)} = \mathrm{E}[D(\eta)|\text{data}]$ is the posterior mean of the deviance, K is the number of parameters in the model and n is the sample size. The BIC and AIC values are 107.735 and 55.331 for approach A, and 198.908 and 94.099 for approach B, respectively. Approach A provides the best performance in the BIC and AIC criteria (the lower they are, the better the approach is). Since, usually, the likelihood increases when adding parameters to the model, the BIC and AIC criteria are penalized through the number of parameters.

The prior distributions and their hyperparameters considered in both approaches are the same, except for the regression parameters (due to the way the approaches

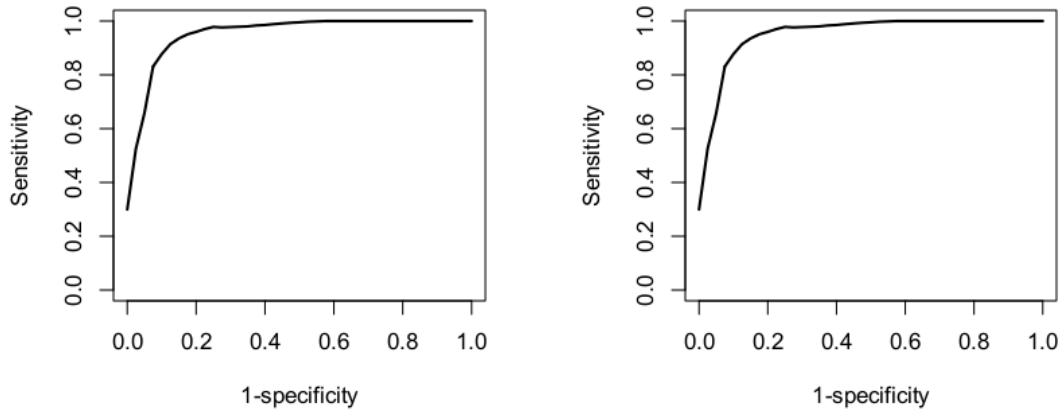

Figure 2: ROC curves for approaches A (left) and B (right).

have been built). They have been chosen in this way because no initial information on the parameters is available, and they promote that the likelihood and the data provide the most important contribution of the approach. In the new approach the regression parameters are involved in a regularization process by means of a scale mixture of normal distributions with independent exponentially distributed variances. A sensitivity analysis on the hyperparameters related to the regression parameter distributions is performed and presented here. Table 1 shows that, even with very important changes in the value of the hyperparameters, the results remain stable. In Figure 3 the regression parameter estimates and the AUC versus the hyperparameter γ_k are presented. The regression parameter estimates $\hat{\beta}_k$ tend to zero when the regularization hyperparameters γ_k are bigger. This happens because small values of γ_k allow higher variability in the estimates, and for big values of γ_k the variability is restricted. AUC values remain stable.

In order to validate the results, a stratified cross-validation framework is considered. Specifically, the dataset is randomly split into a training subset composed by 75% of the control subjects and 75% of the people with PD. The remaining individuals constitute the testing subset. The model parameters are determined using the training subset, and errors are computed using the testing subset. This procedure is repeated 100 times and the results are then averaged. Table 2 shows the results of the classification measures and the goodness-of-fit criteria.

γ_k	TN	FP	FN	TP	Accuracy rate	Sensitivity	Specificity	Precision	AUC
0.001	35	5	7	33	0.850	0.825	0.875	0.868	0.946
0.005	35	5	7	33	0.850	0.825	0.875	0.868	0.945
0.01	36	4	6	34	0.875	0.850	0.900	0.895	0.952
0.05	37	3	7	33	0.875	0.825	0.925	0.917	0.950
0.1	36	4	8	32	0.850	0.800	0.900	0.889	0.954
0.5	36	4	7	33	0.863	0.825	0.900	0.892	0.953
1	36	4	7	33	0.863	0.825	0.900	0.892	0.952

Table 1: Results for different values of the hyperparameter γ_k.

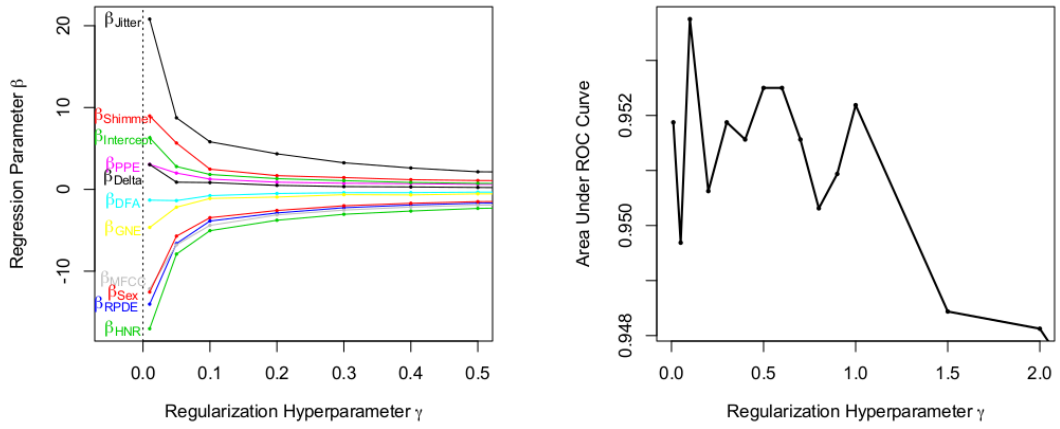

Figure 3: Regression parameter estimates (left) and AUC (right) for different values of the hyperparameter γ_k.

The obtained results show an advantage of the proposed approach. However, this is not the most important fact, but the better chain mixing, the lower computational time and, above all, the improvement in interpretability. Interpretation of results is crucial to understand which acoustic features are more important to discriminate PD subjects from healthy ones.

By using the approach in this paper, the number of redundant variables in the first step has been reduced, and in the second one, the remaining variables are penalized or favored to show their relative importance in the classification. This regularization approach provides information about the relative contribution of the variables to the classification prediction through the regression parameters. Table

	Approach A	Approach B
Accuracy rate	0.779 (0.080)	0.752 (0.086)
Sensitivity	0.765 (0.135)	0.718 (0.132)
Specificity	0.792 (0.150)	0.786 (0.135)
Precision	0.806 (0.115)	0.785 (0.118)
AUC	0.879 (0.067)	0.860 (0.070)
BIC	98.626 (2.336)	190.447 (0.178)
AIC	52.550 (2.336)	94.107 (0.178)

Table 2: Means (standard deviations) of the classification measures and the goodness-of-fit criteria for the two approaches with the considered stratified cross-validation framework.

3 presents the regression and regularization parameter estimation.

	β_k	τ_k
Relative jitter	5.886 (3.224)	4.940 (1.940)
Local shimmer	2.345 (2.150)	3.792 (1.869)
HNR35	-5.078 (2.356)	4.751 (1.867)
MFCC3	-4.369 (2.628)	4.456 (1.903)
Delta8	0.658 (1.685)	3.290 (1.966)
RPDE	-3.924 (1.963)	4.372 (1.870)
DFA	-0.797 (1.265)	3.112 (1.873)
PPE	1.202 (1.305)	3.268 (1.840)
GNE	-1.143 (1.391)	3.307 (1.910)
Sex	-3.609 (2.062)	4.260 (1.929)
Intercept	1.837 (1.284)	3.527 (1.859)

Table 3: Means (standard deviations) of the posterior estimates of the regression and regularization parameters.

The 95% Highest Probability Density (HPD) intervals for the the four regression parameters related to relative jitter, HNR35, RPDE, and MFCC3 do not contain the zero value (the remaining ones contain it). Therefore, these four regression parameters are the most important ones for Approach A. However, in Approach B, the HPD intervals for the regression parameters of all the 46 variables contain the zero value. Then, there is no way to obtain a clear interpretation about which variables are the most important to discriminate healthy controls from PD subjects.

Finding the most influential features through the proposed approach provides

information that can be related to physiological aspects of PD. The practical results will be discussed in Section 5.

4.2. Simulation-based scenarios

A simulation-based experiment is conducted to analyze the effects of the high correlations on the proposed approach and on the proposal in [10], as well as to perform a comparison between both approaches. The simulated scenarios contain some groups of highly correlated variables. These variables are highly correlated within their own groups, but they have been independently generated between groups. Although the generated datasets do not exactly mimic the motivating Parkinson's dataset, the idea of groups of highly correlated variables is based on it.

A set of $n = 100$ subjects having $J = 3$ replications each one are simulated by the following process. The covariate vectors $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{i,40})$, for $i = 1, \ldots, n$, are randomly generated by considering that the variables are clustered in $G = 8$ groups. For each group g, the latent vectors are generated from a 5-dimensional multivariate normal distribution having means $\mathrm{E}(w_{ik}) = 0$, variances $\mathrm{Var}(w_{ik}) = 0.25$, covariances $\mathrm{Cov}(w_{ik}, w_{ih}) = 0.2475$, and correlations 0.99. They are independently generated between groups. Two more scenarios are considered, but now it is assumed that correlations within groups are 0.95 and 0.90 and, therefore, covariances are 0.2375 and 0.225, respectively, within each group.

The variables x_{ikj_i}, for subject i, variable k and replication j_i, where $i = 1, \ldots, n$, $k = 1, \ldots, 40$ and $j_i = 1, 2, 3$, are randomly generated from a normal distribution with mean w_{ik} and variance 0.01, i.e., the x_{ikj_i} are generated from $\mathrm{Normal}(w_{ik}, 0.01)$ for $j_i = 1, 2, 3$.

The p_i are computed by $p_i = \Phi(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z)$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution (probit link), $\boldsymbol{z}_i = 1$ is referred to the intercept, and the regression parameters are $\boldsymbol{\beta} = (\boldsymbol{\beta}_x', \boldsymbol{\beta}_z)'$, with $\boldsymbol{\beta}_z = 1$ and

$$\boldsymbol{\beta}_x' = (\underbrace{0.5, \ldots}_{5\text{ times}}, \underbrace{-0.5, \ldots}_{5\text{ times}}, \underbrace{0.6, \ldots}_{5\text{ times}}, \underbrace{-0.6, \ldots}_{5\text{ times}}, \underbrace{0.7, \ldots}_{5\text{ times}}, \underbrace{-0.7, \ldots}_{5\text{ times}}, \underbrace{0.8, \ldots}_{5\text{ times}}, \underbrace{-0.8, \ldots}_{5\text{ times}}).$$

Finally, the responses Y_i are randomly generated by the following procedure: generate $u_i \sim \mathrm{U}(0, 1)$, and then define $Y_i = 1$ if $p_i > u_i$, else $Y_i = 0$. This procedure is performed 100 times for each scenario.

Two approaches have been considered: the one proposed in this paper, having 8 variables because of the variable reduction performed in the first stage (Approach A), and the one proposed by [10], having 40 variables highly correlated by

groups (Approach B). A noninformative framework as in the previous subsection is considered with the same hyperparameters.

A stratified cross-validation framework is considered. Each dataset is randomly split into a training subset, composed by 75% of the individuals, and a testing subset, composed by the remaining ones. The parameters are estimated by learning from the training subset, and the classification measures are computed by using the testing subset. The results for the 100 datasets in the three scenarios are summarized in Table 4, showing average values for the analyzed measures and their standard deviations.

	Correlation 0.99		Correlation 0.95		Correlation 0.90	
Approach	A	B	A	B	A	B
Accuracy rate	0.879	0.885	0.871	0.888	0.856	0.882
	(0.054)	(0.055)	(0.062)	(0.053)	(0.065)	(0.058)
Sensitivity	0.982	0.988	0.962	0.985	0.956	0.986
	(0.034)	(0.029)	(0.049)	(0.036)	(0.055)	(0.033)
Specificity	0.732	0.738	0.742	0.748	0.713	0.733
	(0.127)	(0.137)	(0.137)	(0.130)	(0.143)	(0.136)
Precision	0.842	0.846	0.844	0.851	0.830	0.844
	(0.070)	(0.073)	(0.077)	(0.070)	(0.075)	(0.070)
AUC	0.968	0.977	0.956	0.975	0.944	0.972
	(0.033)	(0.028)	(0.045)	(0.029)	(0.047)	(0.032)
BIC	43.601	80.246	47.983	80.000	52.847	79.848
	(5.212)	(0.952)	(8.090)	(0.786)	(10.976)	(0.653)
AIC	45.850	85.369	50.232	85.123	55.096	84.970
	(5.212)	(0.952)	(8.090)	(0.786)	(10.976)	(0.653)

Table 4: Simulated data: means (standard deviations) of the measures in the two approaches with the considered stratified cross-validation framework for the three scenarios.

Results show that the higher the correlation, the closer the results of both approaches. Approach A is better than B when AIC and BIC criteria are used and slightly worse when the other criteria are considered. For example, accuracy rates keep similar for Approach B when the correlation decreases, whereas a 2.3% of reduction is obtained for Approach A. Although slightly lower accuracy rates are obtained in Approach A, the multicollinearity problem is reduced and better estimates are obtained. Note that the opposite happened in the previous subsection where Approach A had a slightly better accuracy rate than Approach B. This simulation-based experiment has been conducted to show that the accuracy rates

are very close and that the advantage can be in the other direction, but always the differences are small and depend on the correlation. However, Approach A provides better estimates, more interpretable results and lower computation times as it is also shown next.

Here the convergence assessment has also been performed by using BOA package (see [37]), with a total of 50,000 iterations for each one of the approaches A and B and the same specifications as in the previous subsection. By using the convergence diagnostic method in [38], the results for approach A (approach B) suggest that around 45,000 (104,500) samples should be generated, the first 110 (300) of which are discarded as a burn-in sequence and every 18 (45) are saved as a thinning of the chain. On the other hand, the diagnostic in [39] indicates that after a burn-in of 0 (4,268) all the iterations are retained for posterior inference, with no significant evidence of non-stationarity based on Cramer-von Mises test statistics. Both diagnostics support that approach A needs a shorter burn-in and a shorter chain to converge. The computation times for both approaches were 11.8 and 25.9 minutes, respectively, i.e., being Approach A more than two times faster than Approach B. This is a consequence of a better chain mixing.

5. Discussion

A two-stage variable selection and classification approach has been developed to properly match the replication-based experimental design related to the database in hand. The first stage reduces the number of variables, whereas the second one applies a regularization-based variable selection and classification method. This regularization method is based on LASSO, one of the the most commonly used penalized regression methods [29, 30]. Recently, [42] proposed an alternative Bayesian analysis of LASSO by considering the scale mixture of uniform representation of the Laplace density and [43] proposed a Bayesian methodology for selecting the tuning parameters in penalized regression methods. Although none of both approaches consider a replication-based framework, the underlying ideas are useful. Even more, to the best of the authors' knowledge, no variable selection approaches properly considering replications have been defined up to now. In some contexts where there are no groups of highly correlated variables, the approach can also be used without applying the feature reduction proposed in the first stage, promoting sparseness when the number of variables is large.

The way the model with a probit link function has been specified allows to address the computational issues by deriving an easy-to-implement Gibbs sampling algorithm. Specifically, the development of this Gibbs sampling algorithm

is based on the use of latent variables in a data augmentation framework as in [10] and the choice of prior distribution for the regression parameters as in [20]. The proposed approach could be extended to other link functions as the logit one or even the skew probit one. However, in these cases, the Gibbs sampling algorithm would need a Metropolis-Hastings step in the implementation procedure.

The proposed approach gains in interpretability with respect to the one in [10], because many redundant variables have been removed and the remaining ones are regularized (penalized or favored). This also avoids multicollinearity problems, which is a major issue in this context. Recently, some authors considered this problem in different contexts [44, 45]. Besides, the number of latent variables introduced is also lower in the proposed approach. The introduction of auxiliary variables increases the autocorrelations, as was exemplified by [46]. [37] showed that high autocorrelations suggest slow mixing of chains and, usually, slow convergence to the posterior distribution. The proposed approach exhibits a better chain mixing and a lower computation time when compared with the only classification one. Besides, the experimental results show an acceptable predictive capacity with the considered database, despite the fact that the sample size is relatively small.

As mentioned before, the problem motivating this methodology consists in the discrimination of people suffering PD from healthy subjects based on acoustic features automatically extracted from replicated voice recordings. In addition to this concrete application, the proposed methodology can be applied to many other contexts with similar replication-based experimental designs.

The addressed application is based on the assumption that human voice is affected by PD due to disorders of laryngeal, respiratory and articulatory functions. This allows the acoustic features feeding the variable selection and classification approach to predict probabilities of belonging to PD or healthy classes for new subjects, and provide information on the most relevant acoustic features.

In people with PD, vocal fold stiffness and bowing cause changes in vocal fold mass and tension. As a consequence, the subject cannot properly keep a sustained phonation and therefore shows unstable fundamental frequency, i.e., high jitter [47]. This matches the obtained results, being the relative jitter one of the four most influential features. Besides, vocal fold bowing also produces incomplete glottal closure [47]. As a result, a noise excess appears due to unphonated air escaped through these leaks in the glottis, expecting lower HNR values [48]. This assumption also matches the experimental results since HNR35 is also found to be one of the four most relevant features. MFCCs are related to speech spectral envelope, which depends on placement of articulators (collectively referring to

the lips, teeth, tongue, alveolar ridges, velum). Standard deviations of this type of coefficients are used to model fluctuations in postural stability of articulators during sustained vowel phonation. In this work, MFCC3, which represents the MFCC-based feature group after the filter procedure, has been proven to be an important contributor to PD discrimination success. Another recent contribution that consider MFCC-based features extracted from sustained vowel recordings to discriminate between patients with PD and healthy people is [49]. Finally, due to the devastating impact of vocal fold dysfunction on the complex dynamical structure of speech signals, the use of complementary features that are able to reflect the nonlinear nature of speech has been proven useful. In [50], nonlinear features as well as jitter, HNR and MFCC-based features are used to discriminate PD. A very relevant nonlinear feature is RPDE, which extends the conventional concept of periodicity and substitutes it by the idea of recurrence [51]. This feature quantifies the uncertainty in the measurement of the pitch period and it has been identified by the proposed approach as a relevant feature for PD discrimination. Nonlinear features are promising not only for PD detection, but also to identify other voice-related diseases [52].

The research line about automatic speech signal analysis for PD diagnosis is of great interest. The Parkinson's Voice Initiative has played an important role in the spread of this topic. Many authors have provided different approaches to address this problem, e.g., [4, 5, 6, 7, 8]. More recently, [53] proposed a PD classification algorithm by combining a multi-edit-nearest-neighbor algorithm and an ensemble learning algorithm. Also, [54] presented a survey of algorithms for feature selection to identify PD. In this context, it has become usual to conduct experiments with replicated recordings and assess the performance of a certain approach by using independence-based classification methods. The misuse of features extracted from replicated voice recordings artificially increases the sample size and leads to over-optimistic results. Other authors have considered only one measure for each feature per subject [55] or they have aggregated the replicated features [56]. Anyway, the within-subject variability has not been or cannot be considered in all these studies, but the proposed approach is able to properly address it. The development of other approaches considering replications would be of interest for the near future, particularly, those based on nonlinear methodologies.

Although the voice recording procedure is noninvasive and not time-consuming, there is a difficulty of recruiting people suffering PD. The database, involving 40 healthy controls and 40 PD subjects, is considered here as relatively small from a cross-validation viewpoint, but it has a reasonable size when it is compared to other databases for similar purposes. For example, [4] considered 31 subjects (23

of which were PD subjects), [6] analyzed 40 subjects (20 with PD), [8] studied 46 subjects (24 with PD), and [55] considered three databases (Spanish, German, and Czech) with 100 (50 with PD), 176 (88 with PD) and 56 (20 with PD) subjects, respectively. Besides, there is an underlying difficulty for people suffering PD to keep sustained phonations during some seconds. This could cause that some of the replications can not be used in the posterior feature extraction procedure. This is not the case for the database in hand, for which three valid voice recordings have been obtained and processed for each subject. Although a balanced design is preferred, the approach has been designed to allow also a different number of replications per subject.

The proposed methodology provides a tool to assist physicians in making decisions based on acoustic biomarkers. If this methodology is applied in a primary health care center, family physicians may have an objective noninvasive low-cost tool providing further evidence to refer the patients to a neurological unit. Even more, due to the difficulty of PD diagnostic, this approach can also be useful as a complementary tool for neurologists in performing diagnostics, particularly in the case of early detection.

6. Conclusion

Voice recording replications have not been usually addressed in a proper way for PD discrimination due to the fact that the dependence nature of the data has generally been ignored. The proposed Bayesian approach fills in a gap on variable selection and classification in the presence of replicated data by properly matching the experimental design. It allows for a better interpretability and chain mixing and provides a lower computation time with respect to the only-classification approach presented in the scientific literature. Nowadays, computer assisted diagnostic systems are playing an important role to help in the diagnosis of PD.

Acknowledgement

Thanks to the anonymous participants and to Carmen Bravo and Rosa María Muñoz for carrying out the voice recordings and providing information from the people with PD. We are grateful to the *Asociación Regional de Parkinson de Extremadura* and *Confederación Española de Personas con Discapacidad Física y Orgánica* for providing support in the experiment development. We also thank the three referees and the editor for comments and suggestions which have highly improved both the readability and the content of this paper.

This research has been supported by *Ministerio de Economía y Competitividad*, Spain (Project MTM2014-56949-C3-3-R), *Gobierno de Extremadura*, Spain (Projects GR15052 and GR15106), UNAM-DGAPA-PAPIIT (Project IA106416), and *European Union* (European Regional Development Funds).

Appendix A. Full conditional distributions

The full conditional distributions of the Gibbs sampling algorithm are given by the following:

$$u_i \mid \boldsymbol{y}, \boldsymbol{x}, z, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\tau} \sim \begin{cases} \text{Normal}\left(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z, 1\right) I[u_i > 0] & \text{if } y_i = 1 \\ \text{Normal}\left(\boldsymbol{w}_i'\boldsymbol{\beta}_x + \boldsymbol{z}_i'\boldsymbol{\beta}_z, 1\right) I[u_i \leq 0] & \text{if } y_i = 0 \end{cases} \tag{A.1}$$

$$\boldsymbol{w}_i \mid \boldsymbol{y}, \boldsymbol{x}, z, \boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\tau} \sim \text{Normal}_G\left(\boldsymbol{m}_i, \boldsymbol{M}_i\right), \tag{A.2}$$

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x}, z, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{\Gamma}, \boldsymbol{\tau} \sim \text{Normal}_{G+H}\left(\boldsymbol{b}^*, \boldsymbol{B}^*\right), \tag{A.3}$$

$$\boldsymbol{\Gamma} \mid \boldsymbol{y}, \boldsymbol{x}, z, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\tau} \sim \text{InvWishart}\left(\boldsymbol{V}^*, \sum_{i=1}^{n} J_i + \nu\right), \tag{A.4}$$

$$\tau_k^{-1} \mid \boldsymbol{y}, \boldsymbol{x}, z, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\Gamma} \sim \text{InvGaussian}\left(\sqrt{\gamma_k}/|\beta_k|, \gamma_k\right), \tag{A.5}$$

where

$$\begin{aligned} \boldsymbol{m}_i &= \boldsymbol{M}_i\left(\boldsymbol{\beta}_x(u_i - \boldsymbol{z}_i'\boldsymbol{\beta}_z) + \sum_{j_i=1}^{J_i} \boldsymbol{\Gamma}^{-1}\boldsymbol{x}_{ij_i} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \\ \boldsymbol{M}_i &= \left(\boldsymbol{\beta}_x\boldsymbol{\beta}_x' + J_i\boldsymbol{\Gamma}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}, \\ \boldsymbol{b}^* &= \boldsymbol{B}^*\left((\boldsymbol{w}, z)'\boldsymbol{u}\right), \\ \boldsymbol{B}^* &= \left((\boldsymbol{w}, z)'(\boldsymbol{w}, z) + \text{diag}\left(\tau_1, \ldots, \tau_{G+H}\right)^{-1}\right)^{-1}, \\ \boldsymbol{V}^* &= \sum_{i=1}^{n}\sum_{j_i=1}^{J_i}(\boldsymbol{x}_{ij_i} - \boldsymbol{w}_i)(\boldsymbol{x}_{ij_i} - \boldsymbol{w}_i)' + \boldsymbol{V}. \end{aligned}$$

The final Gibbs sampling-based algorithm consists of choosing initial values $\boldsymbol{w}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Gamma}^{(0)}$ and $\boldsymbol{\tau}^{(0)}$, and iteratively sampling $\boldsymbol{u}^{(l)}$, $\boldsymbol{w}^{(l)}$, $\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\Gamma}^{(l)}$ and $\boldsymbol{\tau}^{(l)}$ from the full conditional distributions (A.1), (A.2), (A.3), (A.4) and (A.5), respectively.

References

[1] J. R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, Elsevier, 2005.

[2] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly, P. J. Snyder, Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment, Journal of Neurolinguistics 17 (6) (2004) 439–453.

[3] L. Baghai-Ravary, S. W. Beet, Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders, Springer Briefs in Electrical and Computer Engineering - Speech Tecnology, Springer, New York, 2013.

[4] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, IEEE Transactions on Biomedical Engineering 56 (4) (2009) 1015–1022.

[5] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, L. O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1264–1271.

[6] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, O. Kursun, Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings, IEEE Journal of Biomedical and Health Informatics 17 (4) (2013) 828–834.

[7] M. Hariharan, K. Polat, R. Sindhu, A new hybrid intelligent system for accurate detection of Parkinson's disease, Computer Methods and Programs in Biomedicine 113 (3) (2014) 904–913.

[8] M. Novotny, J. Rusz, R. Cmejla, E. Ruzicka, Automatic evaluation of articulatory disorders in Parkinson's disease, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (9) (2014) 1366–1378.

[9] C. J. Pérez, L. Naranjo, J. Martín, Y. Campos-Roca, A latent variable-based Bayesian regression to address recording replication in Parkinson's disease, in: EURASIP (Ed.), Proceedings of the 22nd European Signal Processing Conference (EUSIPCO-2014), IEEE, Lisbon, Portugal, 2014, pp. 1447–1451.

[10] L. Naranjo, C. J. Pérez, Y. Campos-Roca, J. Martín, Addressing voice recording replications for Parkinson's disease detection, Expert Systems With Applications 46 (2016) 286–292.

[11] J. Rusz, R. Cmejla, H. Ruzickova, E. Ruzicka, Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease, Journal of Acoustical Society of America 129 (1) (2011) 350–367.

[12] A. Schrag, Y. Ben-Shlomo, N. Quinn, How valid is the clinical diagnosis of Parkinson's disease in the community?, Journal of Neurology, Neurosurgery & Psychiatry 73 (5) (2002) 529–534.

[13] S. M. Curtis, S. K. Ghosh, A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression, Journal of Statistical Theory and Practice 5 (4) (2011) 715–735.

[14] H. Midi, S. K. Sarkar, S. Rana, Collinearity diagnostics of binary logistic regression model, Journal of Interdisciplinary Mathematics 13 (3) (2010) 253–267.

[15] J. Kadane, N. Lazar, Methods and criteria for model selection, Journal of the American Statistical Association 99 (465) (2004) 279–290.

[16] R. B. O'Hara, M. J. Sillanpää, A review of Bayesian variable selection methods: What, how and which, Bayesian Analysis 4 (1) (2009) 85–118.

[17] X. Zhou, K.-Y. Liu, S. T. C. Wong, Cancer classification and prediction using logistic regression with Bayesian gene selection, Journal of Biomedical Informatics 37 (2004) 249–259.

[18] N. Sha, M. Vannucci, M. G. Tadesse, P. J. Brown, I. Dragoni, N. Davies, T. C. Roberts, A. Contestabile, M. Salmon, C. Buckley, F. Falciani, Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage, Biometrics 60 (2004) 812–819.

[19] K. Bae, B. K. Mallick, Gene selection using a two-level hierarchical Bayesian model, Bioinformatics 20 (18) (2004) 3423–3430.

[20] A. Genkin, D. D. Lewis, D. Madigan, Large-scale Bayesian logistic regression for text categorization, Technometrics 49 (3) (2007) 291–304.

[21] Y. Ai-Jun, S. Xin-Yuan, Bayesian variable selection for disease classification using gene expression data, Bioinformatics 26 (2) (2010) 215–222.

[22] V. Rockova, E. Lesaffre, J. Luime, B. Löwenberg, Hierarchical Bayesian formulations for selecting variables in regression models, Statistics in Medicine 31 (2012) 1221–1237.

[23] M. Kyung, J. Gill, M. Ghosh, G. Casella, Penalized regression, standard errors, and Bayesian LASSOS, Bayesian analysis 5 (2) (2010) 369–412.

[24] E. Lesaffre, A. B. Lawson, Bayesian Biostatistics, John Wiley & Sons, Chichester, UK, 2012.

[25] J. P. Buonaccorsi, Measurement Error: Models, Methods and Applications, Chapman and Hall/CRC, Boca Raton, Florida, 2010.

[26] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, Measurement Error in Nonlinear Models: A Modern Perspective, 2nd Edition, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[27] D. Zhu, Y. Li, H. Li, Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data, Bioinformatics Advanced Access (2007) 1–8.

[28] D. Zhu, Y. Li, Correp: Multivariate correlation estimator and statistical inference procedures, R package version 1.36.0 (2007).

[29] R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society. Series B 58 (1) (1996) 267–288.

[30] T. Park, G. Casella, The Bayesian LASSO, Journal of the American Statistical Association 103 (482) (2008) 681–686.

[31] S. Balakrishnan, D. Madigan, Priors on the variance in sparse Bayesian learning: the demi-Bayesian LASSO, in: M.-H. Chen, P. Muller, D. Sun, K. Ye (Eds.), Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger, Springer, Berlin, 2010, pp. 346–359.

[32] A. Lykou, I. Ntzoufras, On Bayesian LASSO variable selection and the specification of the shrinkage parameter, Statistics and Computing 23 (3) (2013) 361–390.

[33] C. Leng, M.-N. Tran, D. Nott, Bayesian adaptive LASSO, Annals of Institute Statistical Mathematics 66 (2) (2014) 221–244.

[34] J. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88 (422) (1993) 669–679.

[35] W. R. Gilks, S. Richardson, D. J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman and Hall, London, 1996.

[36] D. Lunn, A. Thomas, N. Best, D. Spiegelhalter, Winbugs – a bayesian modelling framework: concepts, structure, and extensibility, Statistics and Computing 10 (2000) 325–337.

[37] B. J. Smith, BOA: an R package for MCMC output convergence assessment and posterior inference, Journal of Statistical Software 21 (11) (2007) 1–37.

[38] A. E. Raftery, S. M. Lewis, How many iterations in the Gibbs sampler?, in: J. M. Bernardo, A. F. M. Smith, A. P. Dawid, J. O. Berger (Eds.), Bayesian Statistics 4, Oxford University Press, New York, 1992, pp. 763–773.

[39] P. Heidelberger, P. Welch, Simulation run length control in the presence of an initial transient, Operations Research 31 (1983) 1109–1144.

[40] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461–464.

[41] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov, F. Csaki (Eds.), Proceedings of 2nd International Symposium on Information Theory, Académiai Kiado, Budapest, Hungary, 1973, pp. 267–281.

[42] H. Mallick, N. Yi, A new Bayesian Lasso, Statistics and Its Interface 7 (4) (2014) 571–582.

[43] V. Roy, S. Chakraborty, Selection of tuning parameters, solution paths and standard errors for Bayesian Lassos, Bayesian Analysis Advance Publication (2016) 1–25.

[44] C. K. Chandrasekhar, H. Bagyalakshmi, M. R. Srinivasan, M. Gallo, Partial ridge regression under multicollinearity, Journal of Applied Statistics 43 (13) (2016) 2462–2473.

[45] C.-C. L. Huang, Y.-J. Jou, H. Cho, A new multicollinearity diagnostic for generalized linear models, Journal of Applied Statistics 43 (11) (2016) 2019–2043.

[46] P. Damien, J. Wakefield, S. Walker, Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables, Journal of the Royal Statistical Society, Series B 61 (2) (1999) 331–344.

[47] D. Theodoros, L. Ramig, Communication and swallowing in Parkinson disease, Plural Publishing, 2011.

[48] M. Asgari, I. Shafran, Extracting cues from speech for predicting severity of Parkinson's disease, in: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2010, pp. 462–467.

[49] A. Benba, A. Jilbab, A. Hammouch, Analysis of multiple types of voice recordings in cepstral domain using mfcc for discriminating between patients with Parkinson's disease and healthy people, International Journal of Speech Technology 19 (3) (2016) 449–456.

[50] A. Benba, A. Jilbab, A. Hammouch, Voice assessments for detecting patients with parkinsons diseases using pca and npca, International Journal of Speech Technology 19 (4) (2016) 743–754.

[51] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, I. M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, BioMedical Engineering OnLine 6 (23) (2007) 1–19.

[52] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, J. B. Alonso-Hernandez, M. Faundez-Zanuy, et al., Robust and complex approach of pathological speech signal analysis, Neurocomputing 167 (2015) 94–111.

[53] H.-H. Zhang, L. Yang, Y. Liu, P. Wang, J. Yin, Y. Li, M. Qiu, X. Zhu, F. Yan, Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples, BioMedical Engineering OnLine 15 (122) (2016) 1–22.

[54] P. Shrivastava, A. Shukla, P. Vepakomma, N. Bhansali, A survey of nature-inspired algorithms for feature selection to identify Parkinson's disease, Computer Methods and Programs in Biomedicine 139 (2017) 171–179.

[55] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, E. Nöth, Automatic detection of Parkinson's disease in running speech spoken in three different languages, Journal of the Acoustic Society of America 139 (1) (2016) 481–500.

[56] T. Silva, I. Dutra, T-SPPA trended statistical preprocessing algorithm, in: V. Snasel, J. Platos, E. El-Qawasmeh (Eds.), The International Conference on Digital Information Processing and Communications, Vol. I, Springer-Verlag, 2011, pp. 118–131.