Addressing voice recording replications for tracking Parkinson’s disease progression

Article *in* Medical & Biological Engineering & Computing · May 2016

DOI: 10.1007/s11517-016-1512-y

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

Addressing voice recording replications for tracking Parkinson's disease progression

Lizbeth Naranjo · Carlos J. Pérez · Jacinto Martín

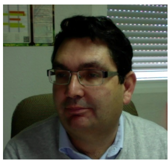
Received: date / Accepted: date

Abstract Tracking Parkinson's disease symptom severity by using characteristics automatically extracted from voice recordings is a very interesting and challenging problem. In this context, voice features are automatically extracted from multiple voice recordings from the same subjects. In principle, for each subject, the features should be identical at a concrete time, but the imperfections in technology and the own biological variability result in non-identical replicated features. The involved within-subject variability must be addressed since replicated measurements from voice recordings can not be directly used in independence-based pattern recognition methods as they have been routinely used through the scientific literature. Besides, the time plays a key role in the experimental design. In this paper, for the first time, a Bayesian linear regression approach suitable to handle replicated measurements and time is proposed. Moreover, a version favoring the best predictors and penalizing the worst ones is also presented. Computational difficulties have been avoided by developing Gibbs sampling-based approaches.

Keywords Bayesian regression models · Latent variables · Longitudinal data · Parkinson's disease · Replicated measurements · Variable selection · Voice features

Lizbeth Naranjo is an Assistant Professor in the Faculty of Science at the Universidad Nacional Autónoma de México (UNAM). She received her B.Sc. degree in Actuarial Science (2006) and her M.Sc. degree in Mathematical Science (2009) from the UNAM, and her Ph.D. degree in Mathematical Science (2014) from the Universidad de Extremadura, Spain.

Carlos J. Pérez is an Associate Professor in the Department of Mathematics at the Universidad de Extremadura. He received his Bachelor's degree in Mathematical Science in 1996 and his Ph. D. degree in Mathematics in 2003.



Jacinto Martín, Ph.D., is an Associate Professor in the Department of Mathematics at the Universidad de Extremadura since 1999. He received his B.Sc. degree in Mathematical Science in 1991 and his Ph.D. in Computer Science in 1995.

1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease. According to the Parkinson's Disease Foundation, an estimated 7 to 10 million people worldwide are living with this medical condition. PD produces a variety of motor and non-motor deficits. The most ostensible motor symptoms are bradykinesia (slowness of movement) or akinesia (lack of movement), resting tremor, postural instability or muscular rigidity. Other symptoms are loss of facial expression or a tendency to lean forward during walking. In addition to these motor symptoms, many PD patients develop non-motor deficits such as neuropsychiatric problems, autonomic dysfunctions, sleep disturbances or cognition disorders.

Voice and speech, as dependent on laryngeal, respiratory and articulatory functions, are also affected in people with PD. Non-dopaminergic changes can also affect language, cognition and mood, which can impact on communication. In fact, vocal impairment can be one of the earliest indicators of PD [10]. Since the very early stages of PD, there may be 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. In these investigations, the voices of the subjects are recorded to extract some specific characteristics of the speech signals and they are classified according to different methods.

The attempt to automatically find speech patterns in neurological patients traces back to more than 30 years [17]. In recent years numerous techniques have been developed to assess speech-related diseases. The monograph presented by [1] provides a current view of the state-of-the-art concerning automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. This research line has a special development for the diagnosis of PD. Some authors have considered measures extracted from speech recordings to discriminate healthy people from those with PD [13,18,21]. This has opened an interesting way to address early diagnoses of people with PD. Note that it is estimated that 20% of people with PD remain undiagnosed [22]. Besides, this also may help to combat misdiagnosis. The Parkinson's Voice Initiative has played an important role in the diffusion of this research topic[1].

As the disease progresses, the voice impairment increases. PD tracking is often performed by applying the Unified Parkinson's Disease Rating Scale (UPDRS [20]), which reflects the presence and severity of symptoms, but does not provide information about the underlying causes. It requires the patient's physical presence in a clinical center and the availability of expert clinical staff for a long time. For many people with PD, visits to hospital are an additional complication. Advances in broadband telecommunication systems offer the possibility of remote monitoring [9]. Telemonitoring is objective, simple, noninvasive and facilitates fast, frequent remote tracking of disease progression. On the other hand, it significantly alleviates the national health systems of excessive workload and the large associated costs of clinical human expertise. The development of accurate remote systems can be very useful to monitor the disease.

Some approaches have addressed the problem of finding a statistical mapping between speech parameters and UPDRS scores. However, up to date, all approaches have been performed based on only one multicenter study that tested the feasibility of a Computer based at-Home Testing Device (AHTD) in 52 early-stage unmedicated PD patients over a period of six months [9]. Note that the size of the voice recording database AHTD is limited in number of subjects. However, it is not as limited in number of voice recordings since each subject has multiple replicated measures in each recording time.

Features extracted from voice recordings of sustained /a/ from 42 out of 52 patients from the AHTD database were considered by Tsanas and collaborators in several publications for tracking purposes [25–29]. Voice characteristics considered in [25] were used by other authors for the same purpose [5, 7]. The collected voice recordings of the AHTD database are not publicly available, but data analyzed in [25] are available online at UCI Machine Learning Repository[2]. This dataset considers the information from at least 20 valid study sessions during the trial period. In each session, the subject's voice was recorded 6 times, providing replications for each individual. Since, in this context, features are extracted from multiple voice recordings from the same subject at a concrete time, in principle, the features should be identical. The imperfections in technology and the own biological variability result in non-identical replicated features that are more similar to one another than features from different subjects.

All the previous works based on the multicenter study [9] tried to find the minimum average difference between the UPDRS scores and the predicted estimations. They differ in several aspects: linear versus non-linear methods, treating all individuals jointly or splitting by sex, different feature selection algorithms or different types of cross-validation schemes. This provided different accuracies. However, all these approaches considered the features extracted from the recordings as if they were independent. Since each patient has provided many replications during the six months, there is no longer independence. Therefore, in this experimental design there is a within-subject variability that has not been considered up to now.

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

[2] https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

Besides, since there were, at least, 20 weekly sessions and only three UPDRS measures, the data in [25] provided interpolated UPDRS values for, at least, 17 sessions. Both considering the measures as independent and interpolating UPDRS values increase the sample size and provide unreliable accuracies.

Based on the problem addressed before, alternative modelling methods are required to address this kind of experimental design by considering replicated measurements and longitudinality. In this paper, we demonstrate for the first time an approach suitable to handle simultaneously replicated measurements and time. A longitudinal latent variable-based regression approach is proposed. A version favoring the best predictors and penalizing the worst ones is also proposed. Besides, Bayesian methodology has been considered in these approaches, what allows the incorporation of initial information through the prior distributions, when it is available. Otherwise noninformative prior distributions can be considered. When using Bayesian methodology, the posterior distribution of the parameters has to be computed. In general, the posterior distribution is not analytically tractable. Therefore, integrating over high-dimensional spaces is usually required. Markov Chain Monte Carlo (MCMC) methods ([6,8]) use Markov chains to draw samples from the required distributions that allow to compute numerical estimates. The difficulty of computing the posterior distribution has been avoided by using Gibbs sampling, that is a particular case of MCMC method. The way the model have been defined allows to derive the full conditional distributions necessary to generate from the posterior distribution in the Gibbs sampling-based algorithm.

The outline of the paper is as follows. Section 2 proposes the approach and highlights how it handles replicated measurements and time. Besides, the approach is extended to address a kind of variable selection by favoring the best predictors and penalizing the worst ones. Section 3 is devoted to the experimental study. In Section 4, the significance of the results is established and the study is placed in the context of the current knowledge in the field. Finally, the conclusion is presented in Section 5.

2 Methods

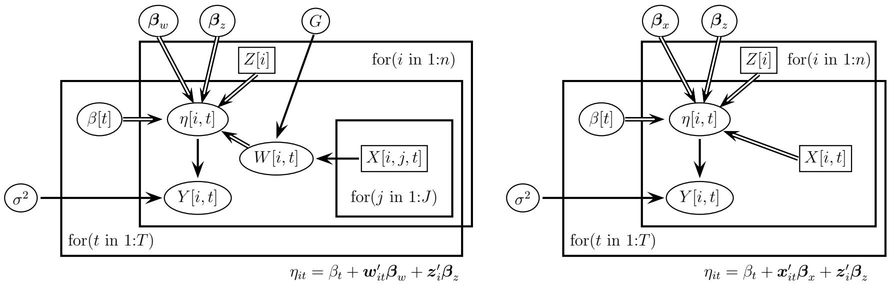
2.1 A latent variable-based regression approach

Suppose that n independent random variables $y_{1t}, \ldots, y_{nt}$ are observed in the time t, where y_{it} is normal distributed and $t = 1, \ldots, T$. The observations y_{it} are related to two sets of covariates $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_i$ through a linear regression model, where $\boldsymbol{x}_{it} = (\boldsymbol{x}_{i1t}, \ldots, \boldsymbol{x}_{iJt})$ is a $K \times J$ matrix of a set of K covariates which have been measured with J replicates, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})'$ is a vector of a set of H covariates which are exactly known. Suppose that $\boldsymbol{x}_{ijt} = (x_{i1jt}, \ldots, x_{iKjt})'$ is the j-th replication of the unknown covariates vector $\boldsymbol{w}_{it} = (w_{i1t}, \ldots, w_{iKt})'$ and assume that they have a linear relationship (additive measurement error model [3,4]), i.e., instead of the unknown covariates $\boldsymbol{w}$, their replicates $\boldsymbol{x}$ are observed. By this way, the $\boldsymbol{x}_{ijt}$ are the surrogates of $\boldsymbol{w}_{it}$. Therefore, the following hierarchical linear model is defined:

$$\begin{aligned} y_{it} &= \beta_t + \boldsymbol{w}'_{it}\boldsymbol{\beta}_w + \boldsymbol{z}'_i\boldsymbol{\beta}_z + \varepsilon_{it}, \\ \varepsilon_{it} &\sim \text{Normal}(0, \sigma^2), \\ \boldsymbol{x}_{ijt} &= \boldsymbol{w}_{it} + \boldsymbol{\delta}_{ijt}, \\ \boldsymbol{\delta}_{ijt} &\sim \text{Normal}_K(\boldsymbol{0}, \boldsymbol{G}), \end{aligned}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_w, \boldsymbol{\beta}'_z)'$ is a $(K{+}H)$-dimensional vector of unknown parameters, ε_{it} are independent and identically distributed random variables, and $\boldsymbol{G}$ is a $K \times K$ matrix of variances and covariances (the replicates among covariates are not independent). The error vector $\boldsymbol{\delta}_{ijt}$ is independent of $\boldsymbol{w}_{it}$, implying that $\boldsymbol{x}_{ijt}$ is a surrogate of $\boldsymbol{w}_{it}$. In this case, the vector $\boldsymbol{x}$ is said to be a surrogate since the conditional distribution $\boldsymbol{y}$ given $(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{x})$ only depends on $(\boldsymbol{w}, \boldsymbol{z})$. A classical error model is assumed, i.e., $\mathrm{E}(\boldsymbol{x}|\boldsymbol{w}) = \boldsymbol{w}$, and $\boldsymbol{x} = \boldsymbol{w} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is the measurement error.

In the standard Bayesian longitudinal linear regression model, the variables $\boldsymbol{x}_{it}$'s are exactly known, i.e., the voice measures should be exactly known without any replication. However, in the proposed model, the $\boldsymbol{x}_{ijt}$'s are replications of the voice measures, and they are the surrogates of $\boldsymbol{w}_{it}$. By tackling the problem in this way, replicated measurements can be considered, whereas with the standard method can not. The flowcharts displayed in Figure 1 represent both the proposed approach and the standard one.



(a) (b)

Fig. 1 Flowcharts for the proposed longitudinal latent variable-based approach (a) and for the standard longitudinal linear regression approach (b).

Without loss of generality and to simplify notation, the index t is omitted by introducing the parameters β_t into the term $\boldsymbol{z}'_i\boldsymbol{\beta}_z$, such that $\boldsymbol{y}$ is an $N-$dimensional vector, where $N = nT$. Therefore, the model is redefined by:

$$\begin{aligned} y_i &= \boldsymbol{w}'_i\boldsymbol{\beta}_w + \boldsymbol{z}'_i\boldsymbol{\beta}_z + \varepsilon_i, \\ \varepsilon_i &\sim \text{Normal}(0, \sigma^2), \\ \boldsymbol{x}_{ij} &= \boldsymbol{w}_i + \boldsymbol{\delta}_{ij}, \\ \boldsymbol{\delta}_{ij} &\sim \text{Normal}_K(\boldsymbol{0}, \boldsymbol{G}). \end{aligned}$$

The following step is to define the prior distributions. The usual approach for regression models assumes a multivariate normal distribution for the regression parameters, $\boldsymbol{\beta} \sim \text{Normal}_{K+H}(\boldsymbol{b}, \boldsymbol{B})$, with $\boldsymbol{b}$ and $\boldsymbol{B}$ fixed, and for the variance σ^2 an inverse gamma distribution is assumed, $\sigma^2 \sim \text{InvGamma}(a, d)$. Conjugate prior distributions for the variance and covariance parameters of the replications are considered, $\boldsymbol{G} \sim \text{InvWishart}_K(\boldsymbol{V}, \nu)$, where ν and $\boldsymbol{V}$ are fixed. Finally, the latent variables are distributed as $\boldsymbol{w}_i \sim \text{Normal}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are fixed.

The likelihood function considering the observed and latent variables is:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) = f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{G}) f(\boldsymbol{w}),$$

then, the joint posterior density is:

$$\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \propto \mathcal{L}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{G}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \pi(\boldsymbol{\beta}) \pi(\sigma^2) \pi(\boldsymbol{G}).$$

These specifications allow to derive an efficient Gibbs sampling algorithm to generate from the posterior distribution. The full conditional distributions are:

$$\begin{aligned}
\boldsymbol{w}_i |\ \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{G} &\sim \text{Normal}_K(\boldsymbol{m}_i\ ,\ \boldsymbol{M}), && (1)\\
\boldsymbol{\beta} |\ \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \sigma^2, \boldsymbol{G} &\sim \text{Normal}_{K+H}(\boldsymbol{b}^*\ ,\ \boldsymbol{B}^*), && (2)\\
\sigma^2 |\ \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{G} &\sim \text{InvGamma}\,(a^*\ ,\ d^*), && (3)\\
\boldsymbol{G} |\ \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}, \sigma^2 &\sim \text{InvWishart}(\boldsymbol{V}^*\ ,\ NJ + \nu), && (4)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{m}_i &= \boldsymbol{M}\left(\boldsymbol{\beta}_w (y_i - \boldsymbol{z}_i' \boldsymbol{\beta}_z)/\sigma^2 + \sum_{j=1}^{J} \boldsymbol{G}^{-1} \boldsymbol{x}_{ij} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right),\\
\boldsymbol{M} &= \left(\boldsymbol{\beta}_w \boldsymbol{\beta}_w' / \sigma^2 + J \boldsymbol{G}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1},\\
\boldsymbol{b}^* &= \boldsymbol{B}^* \big((\boldsymbol{w}, \boldsymbol{z})' \boldsymbol{y} / \sigma^2 + \boldsymbol{B}^{-1} \boldsymbol{b}\big),\\
\boldsymbol{B}^* &= \big((\boldsymbol{w}, \boldsymbol{z})' (\boldsymbol{w}, \boldsymbol{z}) / \sigma^2 + \boldsymbol{B}^{-1}\big)^{-1},\\
a^* &= a + N/2, \qquad d^* = d + \frac{1}{2} \sum_{i=1}^{N} (y_i - \boldsymbol{w}_i' \boldsymbol{\beta}_w - \boldsymbol{z}_i' \boldsymbol{\beta}_z)^2,\\
\boldsymbol{V}^* &= \sum_{i=1}^{N} \sum_{j=1}^{J} (\boldsymbol{x}_{ij} - \boldsymbol{w}_i)(\boldsymbol{x}_{ij} - \boldsymbol{w}_i)' + \boldsymbol{V}.
\end{aligned}$$

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

This approach is able to estimate the UPDRS values while the within-subject variability and time are taken into account. The proposed approach properly matches the experimental design. Next section modifies this approach to allow the inclusion of a variable selection procedure.

2.2 Penalization for variable selection

Identifying significant predictors will enhance the prediction performance of the fitted model. There have been many variable selection methods proposed in the scientific literature from both frequentist and Bayesian perspectives [11,16]. Among the variable selection methods of particular interest are the penalized regression approaches, which can unify many techniques with an easy-to-implement framework. One of the most commonly used methods is the Least Absolute Shrinkage and Selection Operator (LASSO) [24], that minimizes the squared error subject to the non-differentiable constraint expressed in terms of the L_1 norm of the coefficients.

LASSO estimator can be interpreted as the posterior mode in a Bayesian context. A wide variety of Bayesian LASSO methods has been developed and published in the last years [2,12,14]. 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 the regression coefficients. Bayesian adaptive LASSO approach proposed by [30] is adapted here to be integrated in the proposed regression approach by modifying the Gibbs sampling-based algorithm. The formulation is:

$$\begin{aligned}\beta_k|\sigma^2_{\beta_k} &\sim \text{Normal}(0,\sigma^2_{\beta_k}),\ \ \sigma^2_{\beta_k}=\sigma^2\tau^2_k,\\ \sigma^2 &\sim \text{InvGamma}(a,d),\\ \tau^2_k &\sim \frac{\lambda^2_k}{2}\exp\{-\lambda^2_k\tau^2_k/2\},\\ \lambda^2_k &\sim \text{Gamma}(r,s),\end{aligned}$$

where $k=1,\ldots,K+H$ and the penalty considered here is $\sum_{k=1}^{K+H}\lambda_k|\beta_k|/|\hat{\beta_k}|$, being $\hat{\beta_k}$ the least square estimate of β_k. The λ_k coefficients are nonnegative regularization parameters. The method shrinks the k-th regression coefficient toward 0 as λ_k increases, i.e., a large posterior value of λ_k is associated with a large shrinkage of β_k. Therefore, λ_k are penalty parameters to control the size of the regression coefficients. τ_k parameters are auxiliary variables chosen to allow this Bayesian hierarchical structure.

Now, the full conditional distributions for the new approach considering penalization are given below. The full conditional distributions for $\boldsymbol{w}_i$, $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{G}$ are the same as in (1), (2), (3) and (4), respectively, but now the hyperparameters for (2) and (3) changes to:

$$\begin{aligned}\boldsymbol{b}^* &= \boldsymbol{B}^*\big((\boldsymbol{w},\boldsymbol{z})'\boldsymbol{y}/\sigma^2\big),\\ \boldsymbol{B}^* &= \big((\boldsymbol{w},\boldsymbol{z})'(\boldsymbol{w},\boldsymbol{z})/\sigma^2+\text{diag}(\sigma^2\tau^2_1,\ldots,\sigma^2\tau^2_{K+H})^{-1}\big)^{-1},\\ a^* &= a+N/2+(K+H)/2,\\ d^* &= d+\frac{1}{2}\sum_{k=1}^{K+H}\frac{\beta^2_k}{\tau^2_k}+\frac{1}{2}\sum_{i=1}^{N}(y_i-\boldsymbol{w}_i'\boldsymbol{\beta}_w-\boldsymbol{z}_i'\boldsymbol{\beta}_z)^2,\end{aligned}$$

while the new full conditional distributions are:

$$\lambda^2_k|\boldsymbol{y},\boldsymbol{x},\boldsymbol{z},\boldsymbol{w},\boldsymbol{\beta},\sigma^2,\boldsymbol{G},\boldsymbol{\tau}\sim\text{Gamma}\left(r+1,\tau^2_k/2+s\right), \tag{5}$$

$$\frac{1}{\tau^2_k}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{z},\boldsymbol{w},\boldsymbol{\beta},\sigma^2,\boldsymbol{G},\boldsymbol{\lambda}\sim\text{InvGaussian}\left(\lambda_k\sigma/|\beta_k|,\lambda^2_k\right). \tag{6}$$

The final algorithm consists of choosing initial values $\boldsymbol{w}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\sigma^{2(0)}$, $\boldsymbol{G}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\boldsymbol{\tau}^{(0)}$, and iteratively sampling $\boldsymbol{w}^{(l)}$, $\boldsymbol{\beta}^{(l)}$, $\sigma^{2(l)}$, $\boldsymbol{G}^{(l)}$, $\boldsymbol{\lambda}^{(l)}$, $\boldsymbol{\tau}^{(l)}$ from the full conditional distributions: (1), (2), (3), (4), (5) and (6), respectively.

3 Application

3.1 Dataset

The AHTD database of voice recordings was presented in [9]. Originally, 52 subjects having idiopathic PD with diagnosis within the previous five years at trial onset were recruited from six USA medical centers that supervised the experiment. People with PD were physically assessed and given UPDRS scores at baseline, three months and six-months into the trial. During the six months the trial lasted, six phonations of the sustained /a/ were weekly recorded to each patient. This means that the patients produced a long vowel /a/ during some seconds, trying to keep the pitch and loudness as constant as possible. All patients signed an informed consent.

The collected voice recordings of the AHTD database are not publicly available, but data analyzed in [25] are available online at UCI Machine Learning Repository. This dataset considers the information from 42 out of 52 patients for which at least 20 valid study sessions were performed during the trial period. The data file contains name, age, gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS (in both cases, real and interpolated values), and 16 biomedical voice measures (5 measures of variation in fundamental frequency, 6 measures of variation in amplitude, 2 measures of ratio of noise to tonal components in the voice, one nonlinear dynamical complexity measure, one signal fractal scaling exponent, and one nonlinear measure of fundamental frequency variation). Each row corresponds to one of 5,875 signals of sustained /a/ for these individuals. The experimental design is summarized in Figure 2.

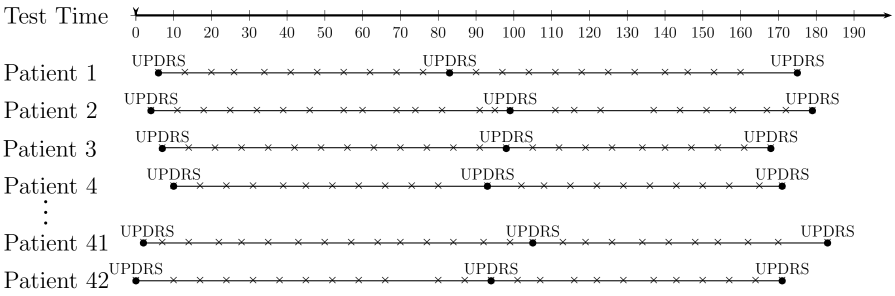


Fig. 2 Experimental design scheme in a time scale based on day. Black points represent replicated voice recordings and UPDRS measures. Crosses represent only replicated voice recordings.

3.2 Experimental setting

Some acoustic variables are highly correlated since they come from similar formulations (e.g. measures of variation in fundamental frequency), providing redundant information. Therefore, only one variable of each group is considered, i.e., jitter in percentage, local shimmer, HNR (Harmonic-to-Noise Ratio), RPDE (Recurrence Period Density Entropy), DFA (Detrended Fluctuation Analysis) and PPE (Pitch Period Entropy).

The following specifications will be used to estimate both the clinician's motor UPDRS scores (motor-UPDRS) and the clinician's total UPDRS scores (total UPDRS) from the acoustics variables and sex. Three scenarios are considered based on the physiological differences in the vocal apparatus of men and women: i) The acoustic variables and the sex are considered as predictor variables, ii) Only the acoustic variables for men are considered as predictor variables, and iii) Only the acoustic variables for women are considered as predictor variables. In each scenario, the acoustic variables are individually normalized to get mean equal to 0 and standard deviation equal to 1.

The prior distributions are: $\boldsymbol{w}_i \sim \text{Normal}(\mathbf{0}, \text{diag}(1))$, $\boldsymbol{\beta} \sim \text{Normal}(\mathbf{0}, \text{diag}(1000))$, $\sigma^2 \sim \text{InvGamma}(1,1)$, $\boldsymbol{G} \sim \text{InvWishart}(\text{diag}(1), K)$, $\lambda_k^2 \sim \text{Gamma}(1,1)$.

The algorithms have been implemented in `R` software. A total of 30000 iterations with a burn-in of 10000 and saving one out of 20 generated values have been considered. With these specifications, the chains generated by using the Gibbs sampling algorithms seem to have converged. The convergence analysis has been performed by using `BOA` package [23].

Cross-validation is used to assess the model generalization performance. Specifically, the dataset is randomly split into a training subset and a testing subset. The training subset is approximately composed by 75% of the men and 75% of the women, specifically, by 78.5% in both cases to provide numbers closer to integers in both groups. The individuals in the training subset are chosen without replacement. The remaining individuals constitute the testing subset. Note that the selection is based on individuals, including all their voice recording replications. The model parameters are determined using the training subset, and errors are computed using the testing subset. This process is independently performed 100 times and the results are then averaged. A analogous validation framework is considered also for men and women separately. Therefore, three models have been fitted: one for men, one for women, and one for both men and women. All of them have different sets of parameter estimates.

In order to measure the goodness-of-fit of the approaches, the following criteria are considered: Mean Absolute Error, $MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \widehat{y_i}|$, Mean Relative Error, $MRE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \widehat{y_i}| / y_i$, and Root Mean Squared Error, $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y_i})^2}$. The lower these measures are, the better the model fitting is.

3.3 Experimental results

Following the previous specifications, the approach in Subsection 2.1 is first applied. For the three considered goodness-of-fit criteria, Table 1 shows the means and the standard deviations (mean±sd) obtained with the specified cross-validation scheme.

	Motor-UPDRS		Total-UPDRS	
	Training	Testing	Training	Testing
	Both sexes			
MAE	5.825±0.362	7.624±1.158	7.668±0.471	9.808±1.679
MRE	0.362±0.031	0.470±0.130	0.337±0.029	0.426±0.123
RMSE	7.227±0.418	9.363±1.392	9.560±0.553	12.092±2.159
	Men			
MAE	6.156±0.469	8.440±1.582	8.278±0.650	11.139±2.123
MRE	0.376±0.042	0.522±0.195	0.349±0.039	0.478±0.181
RMSE	7.536±0.531	10.162±1.788	10.226±0.750	13.535±2.563
	Women			
MAE	4.340±1.276	10.272±4.700	5.228±1.607	12.250±6.535
MRE	0.287±0.089	0.608±0.261	0.255±0.081	0.518±0.244
RMSE	5.477±1.632	12.486±6.299	6.614±2.056	15.070±8.701

Table 1 Means and standard deviations of several goodness-of-fit criteria for the approach in Subsection 2.1.

As it is expected, the results are better when the approach is applied to the training set than when it is applied to the testing set. This happens because training subset is approximately composed by 75% of patients (exactly 78.5%), i.e. 22 men and 11 women, whereas the testing subset is constituted by only 6 men and 3 women. When applying the parameters obtained by using the training subset to the same training subset, more accurate results are obtained than when they are applied to a different subset (the testing set). The results in the testing set are considered to assess the model generalization performance.

We have obtained a MAE of 9.81 for total scale and 7.62 for motor scale, when considering both men and women. If models for men and women are considered separately, worse results are obtained since the sample size is reduced to 28 subjects for the men group and only to 14 subjects for women one. This increases the standard deviations and compromises the model generalization performance.

Table 2 shows the results for the approach considering penalization for predictors presented in Subsection 2.2. This approach favors the best predictors and penalizes the worst ones. In all cases, the results are better than the ones obtained with the non-penalized approach. The improvement ranges from 0.1 to 2 for MAE, from 0.007 to 0.087 for MRE, and from 0.158 to 2.593 for RMSE.

4 Discussion

In order to track PD symptom severity, voice recording replications are useful due to the existing within-subject variability. Although the inter-subject variability plays a fundamental role to track PD symptom severity (note that there are subjects with different stages of the disease), the within-subject variability is also important in this kind of experimental design. In principle, for each subject, the features from several voice recordings should be identical at a concrete time, but

	Motor-UPDRS		Total-UPDRS	
	Training	Testing	Training	Testing
		Both sexes		
MAE	5.901±0.343	7.522±1.104	7.729±0.461	9.638±1.636
MRE	0.366±0.030	0.463±0.128	0.339±0.028	0.419±0.122
RMSE	7.308±0.393	9.205±1.313	9.646±0.546	11.856±2.109
		Men		
MAE	6.281±0.438	8.226±1.487	8.415±0.637	10.815±2.069
MRE	0.382±0.042	0.509±0.195	0.355±0.040	0.466±0.183
RMSE	7.652±0.494	9.837±1.634	10.380±0.738	13.064±2.496
		Women		
MAE	5.109±0.560	8.788±2.883	6.043±0.630	10.238±3.887
MRE	0.335±0.046	0.521±0.189	0.293±0.037	0.440±0.169
RMSE	6.482±0.738	10.628±3.474	7.702±0.845	12.477±4.768

Table 2 Means and standard deviations of several goodness-of-fit criteria for the approach in Subsection 2.2.

the imperfections in technology and the own biological variability result in non-identical replicated features.

Besides the time, the dependence among the replicated measurements must be considered in the statistical model, since this dependence is naturally present in this experimental design. This has been ignored in the publications based on the multicenter experiment [5, 7, 15, 25–29]. It has also become usual when treating this dataset to use the weekly voice recordings with interpolated UPDRS scores, since only three UPDRS evaluations are available in the 6 month experiment. This has artificially increased the sample size to the equivalent of 5,875 individuals, instead of the real 42 individuals with his/her replications at baseline, third month and sixth month. Therefore, the results obtained by using an artificially increased sample size with methodologies that do not match the experimental design provide the wrong perception that the accuracy is higher than it actually is.

The impact of this investigation lies on the proposal of two approaches that allow to handle replicated measurements in a longitudinal study. For the first time this has been performed for this kind of experimental design. The key point to develop both approaches is the introduction of latent variables, what allows to derive efficient Gibbs sampling-based algorithms. This is based on the idea of considering that the replicated features are surrogates of the latent variables, i.e., they could be considered as if they were measured with error to approximate to the real unknown feature that is the latent variable.

The cross-validation scheme considered here is based on 22 men and 11 women as the training set, and only 6 men and 3 women as the testing set (each individual with his/her replicated measurements in times 0, 3, and 6 months) in each iteration. No interpolated UPDRS values have been considered here, but only real UPDRS scores. UPDRS assessment relies on the clinical rater's subjective evaluation and experience. Practice has shown that expert clinicians might differ as much as 4-5 UPDRS points in their evaluations [19]. The best obtained MAEs are 7.52 UPDRS points for motor scale and 9.64 UPDRS points for total scale. Therefore, MAEs are still far from the acceptable 4-5 points of the inter-rater variability. By using the artificially increased sample and a methodology that does not fit the experimental design, [25] obtained a MAE close to 7.5 for total scale, whereas [27] obtained a MAE close to 2. However, our results are the first reliable ones that are presented in the scientific literature on this database. This implies that the research considering features extracted from voice recordings to track PD progression is in an initial stage and much more must be done before these tools can be considered in real practice.

There is a need to perform new experiments different from the one in [9]. With only the already extracted features from the AHTD database and without access to the waveforms, it is not possible to test the performance of these and other future methods with possibly enhanced voice features. Although there is a difficulty to recruit patients for this kind of experiments, it is necessary more research on new patients before these techniques can be incorporated into protocols by neurological units, possibly in a remote way. This should happen when these methods provide an inter-rater variability for UPDRS lower than 4-5 points [19].

5 Conclusion

A novel Bayesian linear regression approach suitable to handle replicated measurements and time is proposed. A LASSO-based method for variable selection has also been proposed. To the authors' knowledge, these are the first longitudinal approaches considering the within-subject variability of replicated measures. Although this methodology has been applied for tracking PD progression, it can be applied to other datasets obtained from similar experimental designs. The results show that the approaches provide estimations higher than the inter-rater variability for UPDRS. Nevertheless, the obtained results are reliable and suggest the need of more research. It is necessary to conduct more experiments on new patients so that these proposed approaches and other that, eventually, may be developed can be applied. The final objective is to incorporate low-cost, noninvasive, self-administered technology into protocols by neurological units. This would lead to an improvement in the quality of life of people with PD, and a cost-reduction for the administration.

Acknowledgements Thanks to Athanasios Tsanas for providing information on the dataset. We thank four reviewers for comments and suggestions which have highly improved this paper. This research has been partially supported by *Ministerio de Economía y Competitividad*, Spain (Projects MTM2011-28983-C03-02 and MTM2014-56949-C3-3-R), *Gobierno de Extremadura*, Spain (Projects GR15052 and GR15106), and *European Union* (European Regional Development Funds).

References

1. L. Baghai-Ravary and 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.
2. S. Balakrishnan and D. Madigan. Priors on the variance in sparse Bayesian learning: the demi-Bayesian Lasso. In M.-H. Chen, P. Muller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, pages 346–359. Springer, Berlin, 2010.
3. J. P. Buonaccorsi. *Measurement Error: Models, Methods and Applications.* Chapman and Hall/CRC, Boca Raton, Florida, 2010.

4. R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, Florida, second edition, 2006.
5. M. Castelli, L. Vanneschi, and S. Silva. Prediction of the unified Parkinson's disease rating scale assessment using a genetic programming system with geometric semantic genetic operators. *Expert Systems with Applications*, 41(10):4608–4616, 2014.
6. M.-H. Chen, Q.-M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Series in Statistics. Springer, 2000.
7. Ö. Eskidere, F. Ertaç, and C. Hanilçi. A comparison of regression methods for remote tracking of Parkinson's disease progression. *Expert Systems with Applications*, 39(5):5523–5528, 2012.
8. D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, 2nd edition, 2006.
9. C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, A. D. Wu, P. H. Kraus, L. M. Blasucci, E. A. Shamim, K. D. Sethi, J. Spielman, K. Kubota, A. S. Grove, E. Dishman, and C. B. Taylor. Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device. *Movement Disorders*, 24(4):551–556, 2009.
10. B. Harel, M. Cannizzaro, and P. J. Snyder. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study. *Brain and Cognition*, 56(1):24–29, 2004.
11. J. Kadane and N. Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290, 2004.
12. C. Leng, M.-N. Tran, and D. Nott. Bayesian adaptive Lasso. *Annals of Institute Statistical Mathematics*, 66(2):221–244, 2014.
13. M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, 2009.
14. A. Lykou and I. Ntzoufras. On Bayesian LASSO variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, 23(3):361–390, 2013.
15. P Mohammadi, A. Hatamlou, and M. Masdari. A comparative study on remote tracking of Parkinson's disease progression using data mining methods. *International Journal in Foundations of Computer Science and Technology*, 3(6):71–83, 2013.
16. R. B. O'Hara and M. J. Sillanpää. A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–118, 2009.
17. M. Okada. Measurement of speech patterns in neurological disease. *Medical & Biological Engineering & Computing*, 21:145–148, 1983.
18. C. J. Pérez, L. Naranjo, J. Martín, and Y. Campos-Roca. A latent variable-based Bayesian regression to address recording replication in Parkinson's disease. In EURASIP, editor, *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO-2014)*, pages 1447–1451, Lisbon, Portugal, 2014. IEEE.
19. B. Post, M. P. Merkus, R. M. A. de Bie, R. J. de Haan, and J. D. Speelman. Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Movement Disorders*, 20(12):1577–1584, 2005.
20. C. Ramaker, J. Marinus, A. M. Stiggelbout, and B. J. van Hilten. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders*, 17(5):867–876, 2002.
21. B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and 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):828–834, 2013.
22. A. Schrag, Y. Ben-Shlomo, and N. Quinn. How valid is the clinical diagnosis of Parkinson's disease in the community? *Journal of Neurology, Neurosurgery & Psychiatry*, 73(5):529–534, 2002.
23. B. J. Smith. BOA: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11):1–37, 2007.
24. R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.
25. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *IEEE Transactions Biomedical Engineering*, 57(4):884–893, 2010.

26. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 594–597, Dallas, US, 2010. IEEE.
27. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *The Royal Society Interface*, 8(59):842–855, 2011.
28. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Robust parsimonious selection of dysphonia measures for telemonitoring of Parkinson's disease symptom severity. In Claudia Manfredi, editor, *Models and analysis of vocal emissions for biomedical applications: 7th international workshop*, pages 169–172, Firenze, Italy, 2011. Firenze University Press.
29. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Using the cellular mobile telephone network to remotely monitor Parkinson's disease symptom severity. *IEEE Transactions on Biomedical Engineering*, 2013. (Submitted).
30. H. Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.