Report 1: Regression on Parkinson's disease data

Francesco Conforte, s277683,
ICT for Health attended in A.Y. 2020/21

November 8th 2020

1 Introduction

Patients affected by Parkinson's disease cannot perfectly control their muscles. In particular they show tremor, they walk with difficulties and, in general, they have problems in starting a movement. Many of them cannot speak correctly, since they cannot control the vocal cords and the vocal tract.

Levodopa is prescribed to patients, but the amount of treatment should be increased as the illness progresses and it should be provided at the right time during the day, to prevent the freezing phenomenon. It would be beneficial to measure total UPDRS ((Unified Parkinson's Disease Rating Scale) many times during the day in order to adapt the treatment to the specific patient. This means that an automatic way to measure total UPDRS must be developed using simple techniques easily managed by the patient or his/her caregiver.

One possibility is to use patient voice recordings (that can be easily obtained several times during the day through a smartphone) to generate vocal features that can be then used to regress total UPDRS.

In the following, some linear regression techniques are analyzed, applied on the public dataset that can be downloaded at [1].

2 Data analysis

The 22 features available in the dataset are listed in table 1: of these, subject ID and test time are removed, total UPDRS is considered as regressand and the remaining 19 features are used as regressors. In particular, regressors include many voice parameters (jitter and shimmer measured in different ways, Noise to Harmonic Ratio NHR, etc) and motor UPDRS. The number of points in the dataset is 5875; data are shuffled and the first 50% of the points are used to train the linear model, 25% of the points are used for the validation and the remaining 25% are used to test the model performance. Data are normalized using mean and standard deviation measured on the training dataset.

Figure 1 shows the covariance matrix for the entire dataset (each feature normalized to have unit variance): correlation between total and motor UPDRS is evident, and strong

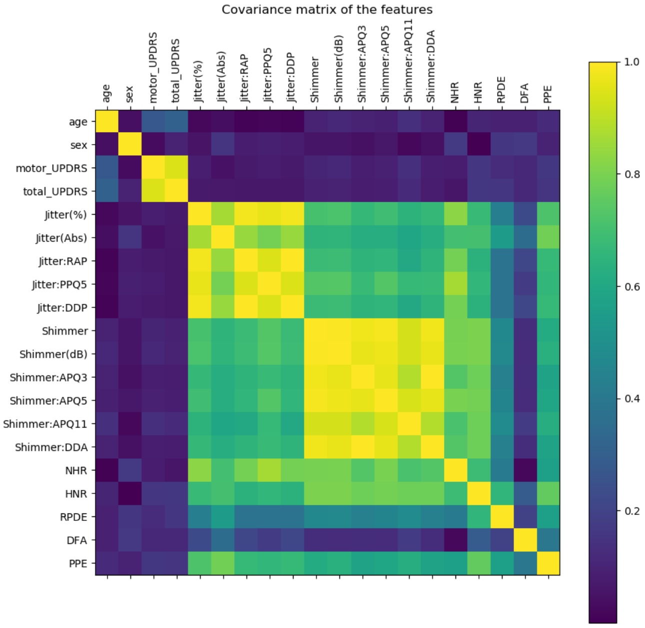


Figure 1: Covariance matrix of the features

correlation also exists among shimmer parameters and among jitter parameters (possible collinearity); on the other hand, only a weak correlation exists between total UPDRS and voice parameters.

3 Linear regression

The model assumed in linear regression is

$$y = w_1 x_1 + \ldots + w_F x_F = \mathbf{x}^T \mathbf{w} \tag{1}$$

1	subject	2	age	3	sex
4	test time	5	motor UPDRS	6	total UPDRS
7	Jitter(%)	8	Jitter(Abs)	9	Jitter:RAP
10	Jitter:PPQ5	11	Jitter:DDP	12	Shimmer
13	Shimmer(dB)	14	Shimmer:APQ3	15	Shimmer:APQ5
16	Shimmer:APQ11	17	Shimmer:DDA	18	NHR
19	HNR	20	RPDE	21	DFA
22	PPE				

Table 1: List of features

where y is the regressand (total UPDRS), $\mathbf{x}^T = [x_1, \ldots, x_F]$[1] stores the F regressors and $\mathbf{w}^T = [w_1, \ldots, w_F]$ is the weight vector to be optimized. In (1), $y, x_1, \ldots, x_F$ are all random variables.

3.1 Linear Least Squares (LLS)

In Linear Least Squares (LLS) $\mathbf{w}$ is found that minimizes the mean square error (MSE):

$$\hat{\mathbf{w}} = \arg\min \mathbb{E}\{(y - \mathbf{x}^T\mathbf{w})^2\} \tag{2}$$

and the mathematical solution is

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

where $\mathbf{X}$ is the matrix that stores the (normalized) training regressor points and $\mathbf{y}$ is the (normalized) training regressand vector. Given $\hat{\mathbf{w}}$, the normalized regressand is estimated as

$$\hat{y} = \mathbf{x}^T\mathbf{w} \tag{4}$$

Figure 2 shows the results obtained with LLS; in particular, Figure 2a shows the optimized weights: they are all very small, apart from two values related to 'Shimmer:APQ3' and 'Shimmer:DDA', which have values respectively around -25 and 25. This happens because of collinearity problem since these two features are strictly correlated as it was showed in Figure 1. This problem could be avoided by removing one of the two features, but they have not been removed for study purposes.

In Figure 2b, a scatter plot showing a comparison between the regressed and the true value of total UPDRS, after applying the model to the test subset, is reported. The estimated values are quite close to the true values, apart from some of them: when the true value of total UPDRS (y) becomes large, the estimated one ($\hat{y}$) tends to be smaller. This also explains the second small Gaussian pdf of Figure 2c.

Figure 2c, instead, offers a point of view about the error and its probability distribution, evaluated on training, validation and test datasets. Error pdfs are a mixture of Gaussian pdfs and there is not much difference in the three subsets, meaning that there is no overfitting.

3.2 Stochastic gradient algorithm with Adam optimization (SG)

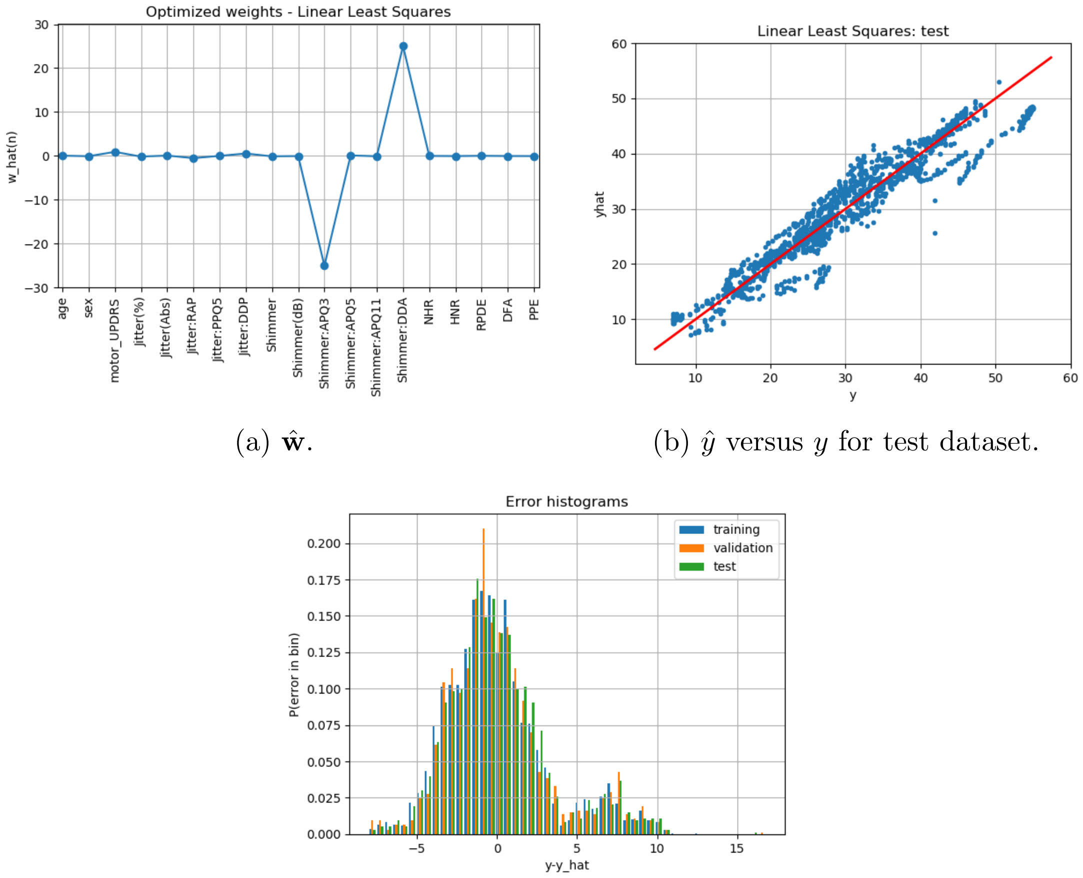
Minimization of eq. 2 can be obtained iteratively using the stochastic gradient algorithm. At the i-th step

$$\hat{\mathbf{w}}_{i+1} = \hat{\mathbf{w}}_i - \gamma\nabla f_i(\mathbf{x}(i))$$

where

$$f_i(\mathbf{x}(i)) = [\mathbf{x}^T(i)\hat{\mathbf{w}}_i - y(i)]^2, \quad \nabla f_i(\mathbf{x}(i)) = 2[\mathbf{x}^T(i)\hat{\mathbf{w}}_i - y(i)]\mathbf{x}(i)$$

[1] $\mathbf{x}$ is a column vector and $\mathbf{x}^T$ is its transpose



(a) $\hat{\mathbf{w}}$.

(b) $\hat{y}$ versus y for test dataset.

(c) Histogram of $y - \hat{y}$ for training, validation and test datasets.

Figure 2: Linear Least Squares results

being $\mathbf{x}^T(i)$ the i-th row of matrix $\mathbf{X}$ and $y(i)$ the i-th element of the regressand vector $\mathbf{y}$ for the normalized training dataset. Adam optimization was applied and therefore $\nabla f_i(\mathbf{x}(i))$ was substituted with its "mean" value, according to [2], using exponential decay rates $\beta_1 = 0.99$ (for the mean) and $\beta_2 = 0.999$ (for the mean square value). The used value of the learning coefficient γ was 10^{-4} and a number of iterations L equal to 300000 was used as a stopping condition to train the model over the training set.

Results are shown in Figure 3. In particular, Figure 3a shows the trend of the Mean Square Error over the number of iterations L, for training and validation datasets. This graph makes it possible to see the learning of the model as the iteration increases. Moreover, it can be seen that there is no overfitting because: as the precision on the training set increases, the results on the validation set improve as well. Therefore, the chosen stopping condition was a number of steps sufficient to build a good linear regression model.

The optimum weight vector obtained with the Stochastic gradient algorithm is shown in Figure 3b. It can be observed that this method reduces the effect of collinearity, by removing those very large values of Figure 2a. The most important feature to regress the total UPDRS is the motor UPDRS, as shown by the spike in correspondence of it.

Plots in Figure 3c and Figure 3d give an idea on the goodness of the obtained results. Like the LLS, the pdfs of the error for all the three subsets are a mixture of Gaussian pdfs and the model starts to give worst predictions for large values of total UPDRS.

3.3 Ridge Regression (RR)

In ridge regression, the estimated weight vector $\hat{\mathbf{w}}$ is obtained as

$$\hat{\mathbf{w}} = \arg\min \ \left[\mathbb{E}\{(y - \mathbf{x}^T\mathbf{w})^2\} + \lambda\|\mathbf{w}\|^2\right] \tag{5}$$

where λ is a hyper-parameter to be optimized through the minimization of the mean square error on the validation dataset. Thus the optimum weight vector becomes

$$\hat{\mathbf{w}}_\lambda = \arg\min \mathbb{E}\{(y - \mathbf{x}^T\mathbf{w})^2 + \lambda\mathbf{w}^T\mathbf{w}\} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{6}$$

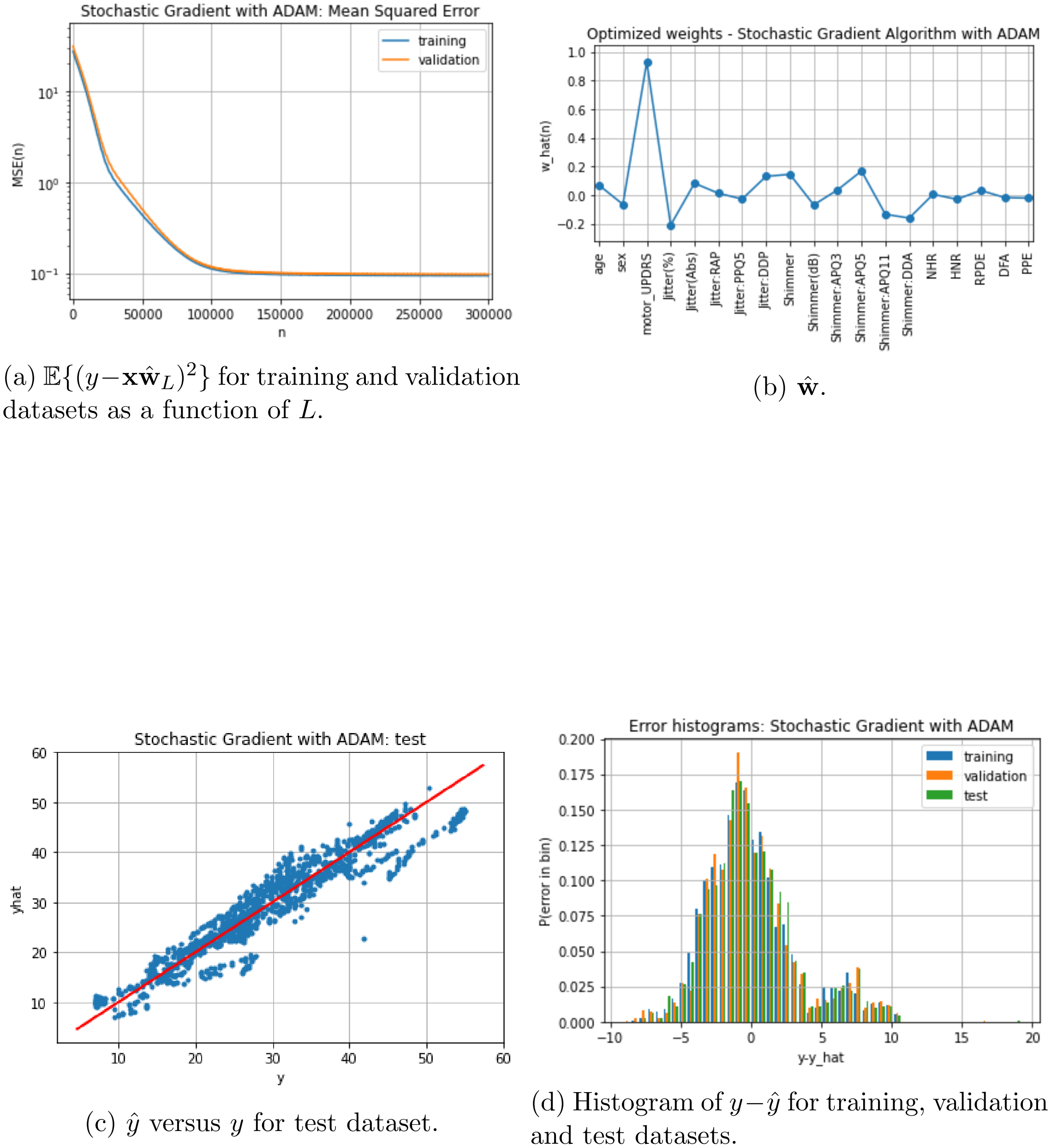
where $\mathbf{I}$ is the identity matrix and λ is chosen that minimizes

$$\mathbb{E}\{(y - \mathbf{x}^T\hat{\mathbf{w}}_\lambda)^2\} \tag{7}$$

on the validation dataset. Results are shown in Figure 4, and, in particular, Figure 4a shows that the optimum value of λ is 29.

With this value of the Lagrangian multiplier, the obtained optimum weight vector is shown in Figure 4b. As expected, with the ridge regression, the collinearity problem is overcome, indeed values of weights of correlated features 'Shimmer:APQ3' and 'Shimmer:DDA' are no more large values as it happened with LLS.

Furthermore, after applying this model on the test subset, results of the regression reported in Figure 4c were obtained and the pdfs of the error for training, validation and test datasets can be observed in Figure 4d. They are very similar to results obtained by applying LLS and Stochastic gradient algorithm.



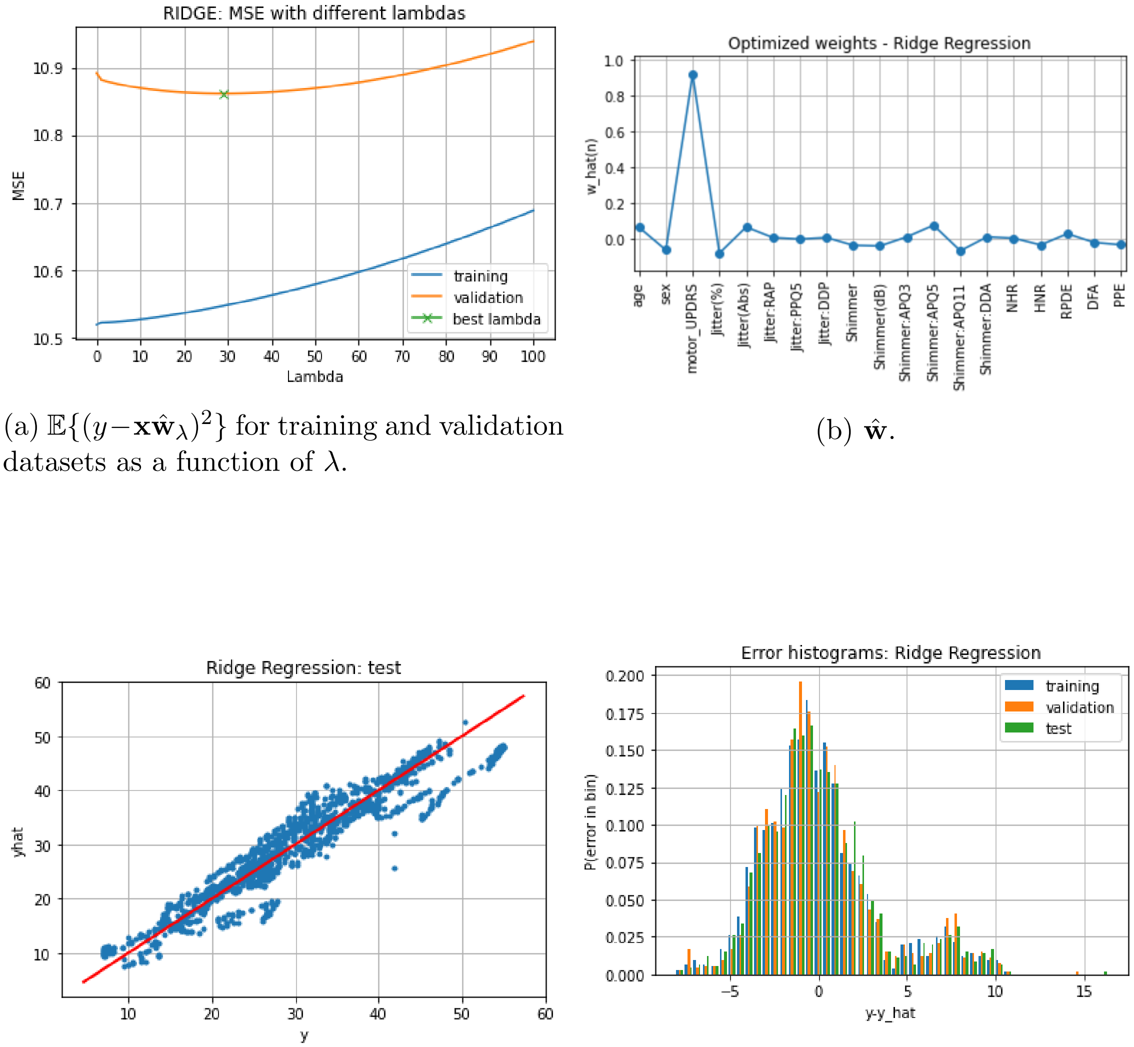
(a) $\mathbb{E}\{(y - \mathbf{x}\hat{\mathbf{w}}_L)^2\}$ for training and validation datasets as a function of L.

(b) $\hat{\mathbf{w}}$.

(c) $\hat{y}$ versus y for test dataset.

(d) Histogram of $y - \hat{y}$ for training, validation and test datasets.

Figure 3: Stochastic gradient results.



(a) $\mathbb{E}\{(y - \mathbf{x}\hat{\mathbf{w}}_\lambda)^2\}$ for training and validation datasets as a function of λ.

(b) $\hat{\mathbf{w}}$.

(c) $\hat{y}$ versus y for test dataset.

(d) Histogram of $y - \hat{y}$ for training, validation and test datasets.

Figure 4: Ridge regression results

3.4 Numerical comparison

The regression error $e = y - \hat{y}$ for the training, validation and test datasets can be seen as a random variable, that can be statistically described. Table 2 lists its main statistical parameters (mean, standard deviation, mean square value and coefficient of determination R^2) for the three analyzed methods.

	Dataset	Err. Mean	Err. St. dev.	MSE	R^2
LLS	Training	3.01×10^{-13}	3.244	10.519	0.989
	Validation	0.134	3.299	10.892	0.989
	Test	0.133	3.266	10.674	0.989
SG	Training	-8.15×10^{-14}	3.256	10.600	0.989
	Validation	0.122	3.323	11.049	0.988
	Test	0.130	3.278	10.753	0.989
RR	Training	-7.684×10^{-14}	3.248	10.547	0.989
	Validation	0.130	3.294	10.862	0.989
	Test	0.120	3.276	10.739	0.989

Table 2: Comparison among the regression methods.

4 Conclusions

By observing Table 2, the first consideration that can be made is that all methods provide similar performances on the regression of total UPDRS. For all the three methods, the error mean in the training subset is practically zero, whereas it is slightly positive in validation and test datasets because the mean of total UPDRS was evaluated using only the cases stored in the training subset.

A particular focus needs to be done on values of error standard deviation: all methods provide an error on the test set of around 3.2 points, which means that most of the times the regression error is around 3-6 points. From a medical point of view, the results of the prediction are acceptable, since the values of error standard deviations are lower than the standard deviation of true total UPDRS values stored in the test set, which is around 10 points.

However, even though the reliability of predicted total UPDRS is acceptable, it is important to recall that it is a prediction for medical purpose and there is always the danger of a not precise treatment (for instance a wrong dose of Levodopa) on the patient. Therefore, the obtained prediction cannot claim to replace an accurate visit of the neurologist, as it is also recalled in [3].

As it was noticed in Section 2, many features were correlated to each other. For example, "Shimmer" features are highly correlated with each other as well as "Jitter" ones but no one of them was correlated to "total UPDRS". After evaluating all the models, it is clear that the voice sampled features are not informative at all. Indeed, plots of optimum weight vectors obtained with Stochastic gradient algorithm and ridge regression show that the only feature that always presents an important weight is the "motor UPDRS". Therefore, in conclusion, it is not possible to claim that total UPDRS can be predicted by considering only dysphonia measures: they do not combine linearly to predict the UPDRS. Thus, nonlinear regression may be required.

References

[1] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. `https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring`.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[3] 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 on Biomedical Engineering*, 57(4):884–893, April 2010.