Feature Selection in Finite Mixture of Regression Models with Diverging Number of Parameters

Abbas Khalili

Department of Mathematics and Statistics,

McGill University, Montreal

(Joint work with Shili Lin, The Ohio State University.)

June 10, 2011.

Outline

- Introduction
- Finite mixture of regression (FMR) models with diverging number of parameters.
- Feature selection problem in FMR models with diverging number of parameters.
- A new feature selection method in FMR models and its statistical properties.
- Analysis of the Parkinson's disease data.

- Recent advancements in medical and other fields of scientific research have allowed scientists to collect data of unprecedented size and complexity.
- A common statistical problem in such applications is:

  To model a response variable Y as a function of a small subset of features $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$.
- This is often referred to as a *feature (or variable) selection problem*.

- Sometimes, in applications, many variables are introduced to reduce possible modeling biases, but the number of variables a model can accommodate is often limited by the amount of data available.

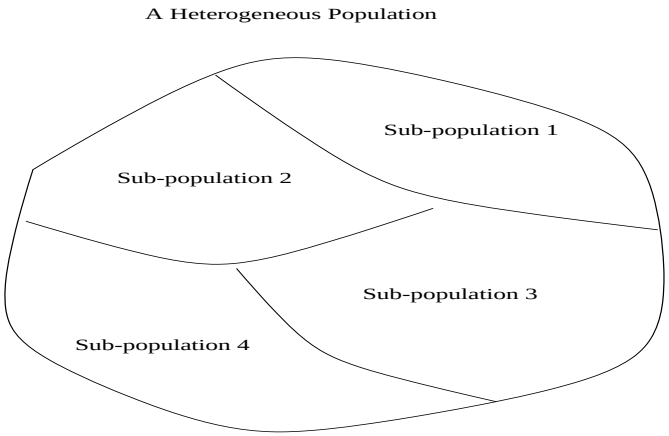
- In other words, the number of variables considered depends on the sample size, i.e. p_n, which reflects the estimability of the parametric model.

- The problem becomes even more complex when the population under study is made up of *subpopulations* and the relationship between Y and $\boldsymbol{x}$ varies across the subpopulations.

- *Finite mixture of regression (*FMR*) models* provide a flexible statistical tool in studying such relationships.

Finite Mixture of Regression (FMR) Models

- In some applications the population under study is heterogeneous, i.e. it is made of sub-populations:



- Each sub-popultation calls for its own regression modelling between Y and $\boldsymbol{x}$.
- FMR models provide a natural way to model unobserved heterogeneity in such populations.

Example: Parkinson disease (PD)

- Parkinson disease (PD) is a neurological disorder.
- Over one million people in North America have PD.
- No cure available for PD, although current medication are effective in controlling its symptoms.
- Early diagnostic and monitoring is crucial in controlling the disease and in improving the life quality for the patients.

Monitoring PD's symptom progression

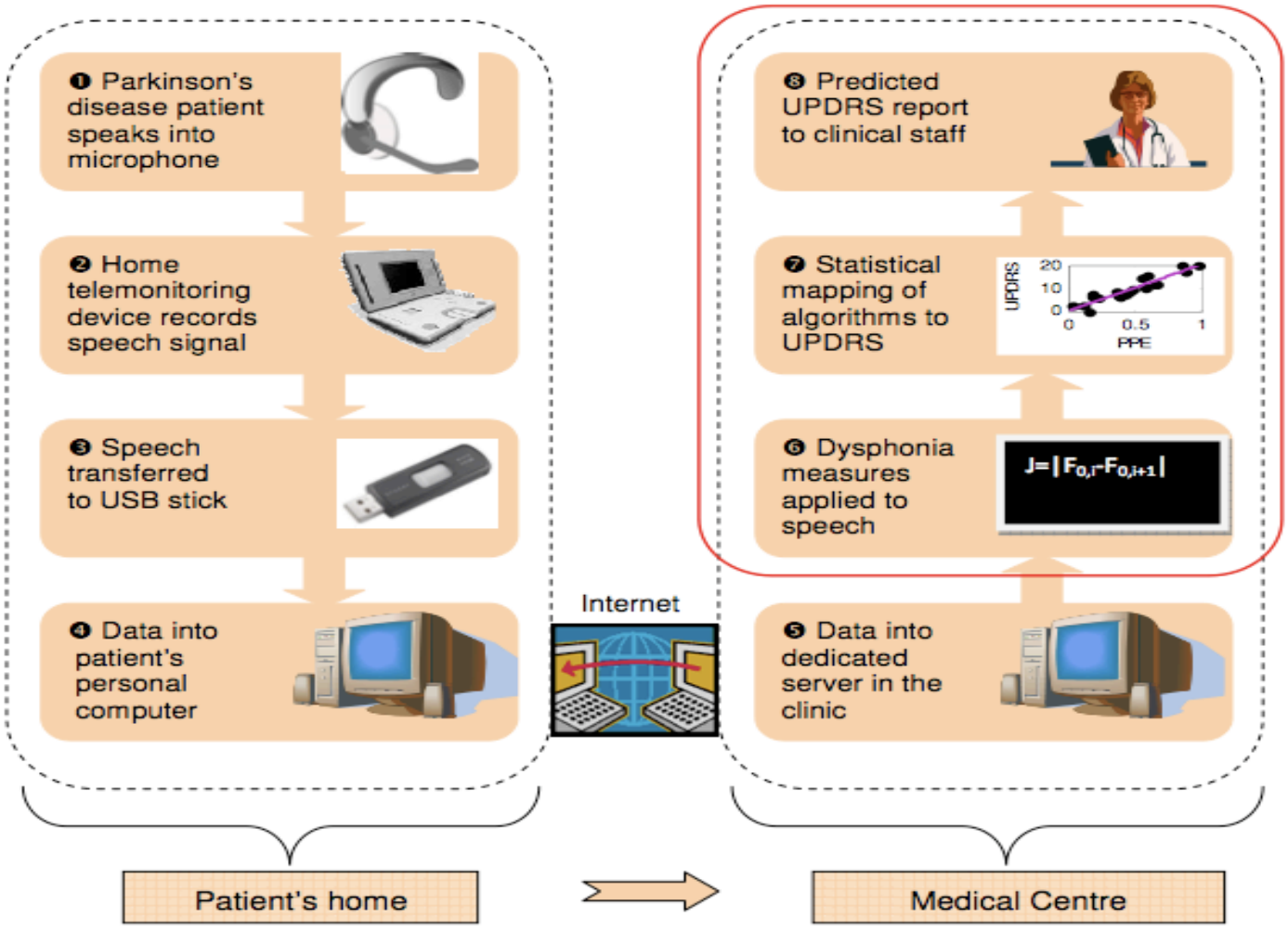
- The main symptoms of PD are tremor, rigidity and other general movement disorders, as well as vocal impairment.
- Tracking PD symptoms involves various physical examinations performed by trained clinical staff.
- A medical rater subjectively assesses the ability of a patient in performing certain tasks. The physical tests are mapped to a metric that is designed to follow PD progression.
- A typical metric of such is the Unified Parkinson's Disease Rating Scale (UPDRS), which reflects the presence and severity of symptoms of PD.

- However, UPDRS requires the patient's presence in the clinic and is also a time-consuming physical examination by trained clinical staff.
- Thus, symptom monitoring is costly and logistically inconvenient for both patients and clinical staff.
- An affordable, cost-effective and reliable alternative:

Telemonitoring

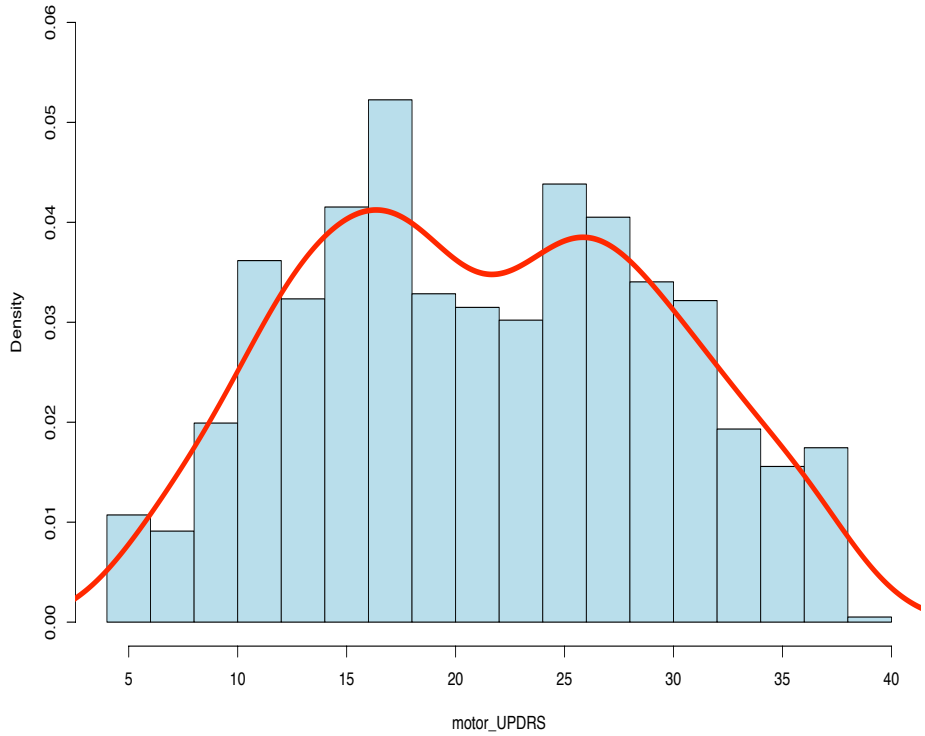
- Noninvasive telemonitoring is an emerging option in general medical care.
- Such options are also considered in PD research.
- The potential for telemonitoring of PD depends heavily on the design of simple tests that can be self-administrated quickly and remotely.
- Recording speech or vocal properties are good candidates in this regard.
- Research has shown that approximately 90% of people with PD exhibit some form of vocal impairment.

Tsanas et al. (2010)



- Whether dysphonic features extracted from speech signals recorded at home can be used as surrogate to study PD severity and progression?
- That is to find a statistical model between speech properties and UPDRS.
- The data under our consideration contains observations on p speech features.
- The goal is to select an optimally reduced subset of the p features, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, that are predictive of the UPDRS, the response variable, Y, leading to a clinically useful model.

Histogram of Y (UPDRS)



Notice the bimodality of the histogram !

- Simple exploration indicates that a single regression model is unlikely to be able to accommodate the multimodality nature of the data.
- Some of the speech features are highly correlated.
- Not all the covariates $x_1, x_2, \ldots, x_p$ are significant in explaining the random behaviour of the UPDRS.
- In general, as sample size increases, more features will likely be extracted from additional algorithms in the hope that greater prediction accuracy can be achieved. That is, p increases as n increases.

- From statistical standpoint, the challenging issue in analyzing the PD data falls into the general category of *feature or variable selection* in a regression model.

- We will use an FMR model to analyze the data.

The focus of this talk

- We discuss the problem of feature selection in *Finite Mixture of Regression (*FMR*)* Models.

- Response variable: $Y \in \mathcal{Y} \subset \mathbb{R}$

  Covariates: $\boldsymbol{x}^\top = (x_1, x_2, \ldots, x_p) \in \mathcal{X} \subset \mathbb{R}^p$.

- In a GLM model the Y and $\boldsymbol{x}$ are related through a density function

$$h(y; \theta(\boldsymbol{x}), \phi)$$

  with a known link function $\theta(\boldsymbol{x}) = \mathcal{L}(\beta_{0k} + \boldsymbol{x}^\top \boldsymbol{\beta}_k)$.

- The model claims that the response y is a function of $\boldsymbol{x}$ through a known link function $\mathcal{L}(\cdot)$.

- Let $\mathcal{H} = \{h(y; \theta, \phi); (\theta, \phi) \in \Theta \times \Phi \subset \mathbb{R} \times \mathbb{R}^+\}$ be a parametric family of density functions for Y with respect to a σ-finite measure μ, and ϕ is a dispersion parameter.

- The conditional density function of Y given $\boldsymbol{x}$ in a FMR model with K components is:

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \; h(y; \theta_k(\boldsymbol{x}), \phi_k)$$

with a known link function $\theta_k(\boldsymbol{x}) = \mathcal{L}(\beta_{0k} + \boldsymbol{x}^\top \boldsymbol{\beta}_k)$, and

$$\sum_{k=1}^{K} \pi_k = 1 \; , \; \pi_k > 0.$$

- In some of the recent feature selection problems the dimension of the model is large and may grow with the sample size.

- In this talk we allow the dimension p of the feature vector $\boldsymbol{x}$ to increase with the sample size n, that is:

  $\boldsymbol{x}_n^\top = (x_1, x_2, \ldots, x_{p_n}) \in \mathcal{X} \subset \mathbb{R}^{p_n}$, $p_n \to \infty$ as $n \to \infty$.
- The component-wise regression vectors: $\boldsymbol{\beta}_{nk}$
- The diverging vector of parameters:

$$\boldsymbol{\Psi}_n = \{(\beta_{0k}, \boldsymbol{\beta}_{nk}, \pi_k, \phi_k); k = 1, 2, \ldots, K\}$$

- Naturally, when there are a large number of covariates, there may be only a small subset of covariates that are significant within each subpopulation.
- That is, for each k, many of the β_{kj}'s are zero.
- The goal is to identify those x_j's for which $\beta_{kj} \neq 0$.
- In such situations the FMR model of interest is called:

SPARSE FMR MODEL

- Let s be a subset of $\{1, 2, \ldots, p_n\}$. We denote $\boldsymbol{x}_n[s]$ as the sub-vector of $\boldsymbol{x}_n$ with elements in s.
- Let $\beta_{nk}[s]$ be the subvector of β_{nk} with $\beta_{kj} = 0, j \notin s$.
- For any $s_1, s_2, \ldots, s_K$, we get an FMR submodel:

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}_n, s_1, s_2, \ldots, s_K) = \sum_{k=1}^{K} \pi_k \; h(y; \theta_k(\boldsymbol{x}_n[s_k]), \phi_k)$$

  with $\theta_k(\boldsymbol{x}_n[s_k]) = \mathcal{L}(\beta_{0k} + \boldsymbol{x}_n^\top[s_k]\beta_{nk}[s_k])$.

- A feature selection method aims at selecting s_k's such that the resulting FMR submodel best balances the model parsimony and goodness of fit of the data.

- All-subset selection methods such as AIC or BIC are not feasible even for moderate values of p and K.

  e.g. If $K = 3$ and $p = 20$, there are: $2^{20} \times 2^{20} \times 2^{20}$ possible submodels to be test by AIC or BIC.

- Khalili and Chen (2007) proposed a computationally efficient penalized likelihood approach for feature selection in FMR models.

- Statistical properties of their proposed method:

  (1) $\sqrt{n}$-CONSISTENCY,
  (2) ASYMPTOTIC NORMALITY,
  (3) SPARSITY.

- In this work the dimension p of the feature space is assumed fixed with respect to the sample size n.

- We consider the problem of feature selection in FMR models when $p = p_n$, and $p_n \to \infty$ as $n \to \infty$.
- Huber (1973), Fan and Peng (2004).

- Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ be a random sample of observations from the FMR model.
- The (conditional) log-likelihood function is given by

$$l_n(\boldsymbol{\Psi}_n) = \sum_{i=1}^{n} \log\{f(y_i; \boldsymbol{x}_i, \boldsymbol{\Psi}_n)\}.$$

- Maximum likelihood estimation (MLE) of $\boldsymbol{\Psi}_n$:

$$\tilde{\boldsymbol{\Psi}}_n = argmax_{\boldsymbol{\Psi}_n}\, l_n(\boldsymbol{\Psi}_n).$$

- But if $\beta_{kj}^0 = 0$, the MLE $\tilde{\beta}_{kj}$ is not necessarily zero !

- We propose to estimate $\boldsymbol{\Psi}_n$ through the regularized log-likelihood function

$$pl_n(\boldsymbol{\Psi}_n) = l_n(\boldsymbol{\Psi}_n) - p_n(\boldsymbol{\Psi}_n)$$

with the regularization function

$$p_n(\boldsymbol{\Psi}_n) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p_n} r_n(\beta_{kj}; \lambda_{nk}) + \frac{\tau_n}{2} \sum_{k=1}^{K} \sum_{j=1}^{p_n} \beta_{kj}^2.$$

Why the penalty function $r_n(\beta; \lambda_{nk})$?

- For each k, let $\beta^0_{nk} = (\beta^0_{1k,n}, \beta^0_{2k,n})^\tau$ such that $\beta^0_{2k,n}$ contains the 0 effects.

- *Consistency in feature selection (sparsity)*:

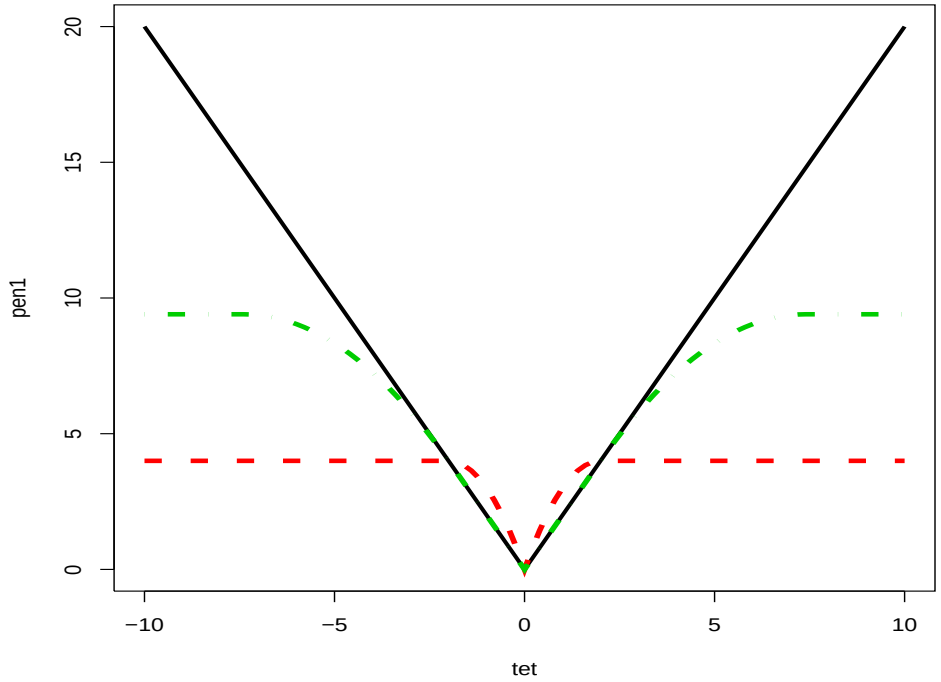
  The penalty function $r_n(\beta_{kj}; \lambda_{nk})$ is designed such that

  $$P(\hat{\beta}_{2k,n} = 0) \to 1 \quad , \quad k = 1, 2, \ldots, K, \quad as \quad n \to \infty.$$

- By having an estimator with the above property, we are in fact performing *feature selection* !

Examples of $r_n(\beta; \lambda_{nk})$

- Most common choices are: LASSO, SCAD and HARD.



- *Controlling the variance of* $\hat{\beta}_{1k,n}$:

  Similar to the ridge regression of Hoeral and Kennard (1970), the quadratic terms β_{kj}^2 are to prevent wild estimates of the true nonzero β_{kj}^0's corresponding to the highly correlated features x_j's.

- See also Hastie, Tibhsirani and Friedman (2009).

Performance of the proposed feature selection method

- The performance of the new method is studied:
  - Theoretically:
    consistency, asymptotic normality, and sparsity.
  - An extensive simulation study.
  - The PD data is also analyzed to demonstrate the use of the new method in real applications.

- Let: $\hat{\boldsymbol{\Psi}}_n = argmax_{\boldsymbol{\Psi}_n}\, pl_n(\boldsymbol{\Psi}_n)$
- Consider the joint density function $f(z_i; \boldsymbol{\Psi}_n)$ of the random variable $\boldsymbol{Z}_i = (\boldsymbol{x}_i, Y_i)$.
- Consider the quantities:
  - $q_{1n} = \max_{k,j}\{r_n(\beta^0_{kj}; \lambda_{nk})/\sqrt{n} : \beta^0_{kj} \neq 0\}$
  - $q_{2n} = \max_{k,j}\{|r'_n(\beta^0_{kj}; \lambda_{nk})|/\sqrt{n} : \beta^0_{kj} \neq 0\}$
  - $q_{3n} = \max_{k,j}\{|r''_n(\beta^0_{kj}; \lambda_{nk})|/n : \beta^0_{kj} \neq 0\}$

($\mathcal{P}_0$). For all n and λ_{nk}, $r_n(0;\lambda_{nk}) = 0$, and $r_n(\beta;\lambda_{nk})$ is symmetric and non-negative. It is non-decreasing and twice differentiable for all β in $(0,\infty)$ with at most a few exceptions.

In addition, there exists constants A_1 and A_2 such that when $\beta_1 > A_1\lambda_{nk}, \beta_2 > A_1\lambda_{nk}$, then:

$$\frac{1}{n}|r_n''(\beta_1;\lambda_{nk}) - r_n''(\beta_2;\lambda_{nk})| \leq A_2|\beta_1 - \beta_2|.$$

($\mathcal{P}_1$). As $n \to \infty$,

$$\sqrt{p_n}q_{1n} = o(1+q_{2n}), q_{2n} = o(p_n^{-1/2}), q_{3n} = o(1).$$

($\mathcal{P}_2$). As $n \to \infty$,

$$\frac{\tau_n}{\sqrt{n}} \max_{k,j} |\beta_{kj}^0| = o(1 + q_{2n}); \tau_n = o(n).$$

($\mathcal{P}_3$). For $T_n = \{\beta; 0 < \beta \leq \sqrt{\frac{p_n}{n}} \log n\}$,

$$\lim_{n\to\infty} \inf_{\beta \in T_n} \frac{r_n'(\beta; \lambda_{nk})}{\sqrt{np_n}} = \infty.$$

- Let $\boldsymbol{Z}_i = (\boldsymbol{x}_i, Y_i), i = 1, 2, \ldots, n$, be a random sample from a density function $f(\boldsymbol{z}; \boldsymbol{\Psi}_n)$ that satisfies certain regularity Conditions.

- Assume that the function $r_n(\beta; \lambda_{nk})$ satisfies Conditions $\mathcal{P}_0$ and $\mathcal{P}_1$, and the tuning parameter τ_n in ridge penalty satisfies Condition $\mathcal{P}_2$.

  If $\frac{p_n^2}{\sqrt{n}} \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\Psi}}_n$ of the function $pl_n(\boldsymbol{\Psi}_n)$ for which

$$\|\hat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}_n^0\| = O_p\{\sqrt{\frac{p_n}{n}}(1 + q_{2n})\}.$$

Theorem 2: Sparsity and asymptotic normality

- Assume conditions in Theorem 1 are fulfilled, and also assume that the regularization function satisfy Conditions $\mathcal{P}_0$-$\mathcal{P}_3$.

  If $\frac{p_n^{2.5}}{\sqrt{n}} \to 0$, then for any $\sqrt{n/p_n}$-consistent maximum regularized likelihood estimator $\hat{\boldsymbol{\Psi}}_n$, as $n \to \infty$:

  - SPARSITY: $P(\hat{\boldsymbol{\beta}}_{2k,n} = \mathbf{0}) \to 1, k = 1, 2, \ldots, K.$
  - ASYMPTOTIC NORMALITY:

$$\sqrt{n}\,\mathbf{B}_n\, I_1^{-1/2}(\boldsymbol{\Psi}_{01})\left\{\left[I_1(\boldsymbol{\Psi}_{01}) - \frac{p_n''(\boldsymbol{\Psi}_{01})}{n}\right](\hat{\boldsymbol{\Psi}}_{n1} - \boldsymbol{\Psi}_{01}) + \frac{p_n'(\boldsymbol{\Psi}_{01})}{n}\right\}$$

$$\longrightarrow^{d} N(\mathbf{0}, \mathbf{G})$$

  where $\mathbf{B}_n\mathbf{B}_n^{\tau} \to G$.

1. The estimator $\hat{\boldsymbol{\Psi}}_n$ is $\sqrt{n/p_n}$-consistent if $q_{2n} = O(1)$.
   - Proper choices of the tuning parameter λ_{nk} in LASSO and SCAD will lead to the desired consistency rate.
   - Regarding the ridge penalty, if for example $\tau_n = \log n$, then Condition $\mathcal{P}_2$ is also guaranteed.
2. To have consistency in feature selection, the λ_{nk} needs to be chosen such that $\sqrt{\frac{n}{p_n}}\lambda_{nk} \to \infty$.
   - The properties 1 and 2 cannot be achieved simultaneously by LASSO. Achievable by SCAD!

- The data are the results of a clinical trial.
- Information on the speech signals can be extracted using both linear and nonlinear algorithms to characterize clinically relevant properties.
- Some features may address the ability of the voice folds to sustain simple vibration whereas some others may be able to characterize the extent of turbulent noise in the speech signals.
- $n = 5875$ observations on 16 speech characteristics, leading to a 5875×16 design matrix.

Final selected model

- We fitted an FMR model with $K = 2$.
- The final selected model is

$$0.57\ \phi(y; \hat{\mu}_1(\boldsymbol{x}), 4.36^2) + 0.43\ \phi(y; \hat{\mu}_2(\boldsymbol{x}), 4.36^2)$$

$$\begin{aligned} \hat{\mu}_1(\boldsymbol{x}) &= 15.87 + \boldsymbol{x}^\tau[s_1]\hat{\beta}_1[s_1] \\ \hat{\mu}_2(\boldsymbol{x}) &= 28.70 + \boldsymbol{x}^\tau[s_2]\hat{\beta}_1[s_2]. \end{aligned}$$

$$\begin{aligned} s_1 &= \{9, 10, 13, 15, 16\} \\ s_2 &= \{2, 3, 4, 16\}. \end{aligned}$$

Fitted FMR model to the PD data

Voice features	Mixture components	
	Com_1	Com_2
Intercept	$15.87_{(.10)}$	$28.70_{(.12)}$
MDVP:Jitter (%)	–	–
MDVP:Jitter(Abs)	–	$-3.65_{(.13)}$
MDVP: RAP	–	$3.77_{(.12)}$
MDVP: PPQ	–	$-1.85_{(.13)}$
Jitter:DDP	–	–
MDVP:Shimmer	–	–
MDVP:Shimmer (dB)	–	–
Shimmer:APQ3	–	–
Shimmer:APQ5	$-1.92_{(.12)}$	–
MDVP: APQ	$1.73_{(.12)}$	–
Shimmer:DDA	–	–
NHR	–	–
HNR	$-1.16_{(.12)}$	–
RPDE	–	–
DFA	$-1.89_{(.10)}$	–
PPE	$1.50_{(.12)}$	$2.16_{(.12)}$

Interpretation of the fitted model

- For those patients whose PD symptoms are mild (smaller UPDRS scores), dysphonic measures such as Shimmer:APQ5, MDVP:APQ, HNR, DFA, and PPE are important features for monitoring disease progression.

- For those with more severe symptoms (larger UPDRS scores), a different set of features, MDVP:Jitter (Abs), MDVP:RAP, MDVP:PPQ, appear to be more relevant in monitoring the progression of PD.

Interpretation of the fitted model

- Although all 16 features represent dysphonic measures, and multiple measures may characterize some similar aspect of the speech, still subtle difference in different x_j's may have a profound implication in PD research.

- For instance, fundamental frequency variations (MDVP:Jitter (Abs)) and variations in signal amplitude (Shimmer:APQ5) both capture symptoms manifested in vocal fold vibration and lung efficiency, but the former is predictive of severity in advanced PD patients whereas the latter is predictive of severity in mild PD patients.

- Also our results suggest that UPDRS is affected by harmonics to noise ratio (HNR), but only in mild PD patients, with decreasing HNR leading to progression of the disease. However, as PD progresses to an advance stage, HNR may reach such a low level that it may no longer be useful for further severity prediction.

References

- Huber, P. J. (Ann. Stat., 1973). Robust regression: Asymptotic, conjectures and Monte Carlo.
- Fan and Peng (Ann. Stat., 2004). Nonconcave penalized likelihood with a diverging number of parameters.
- Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomedical Engineering*, **56**, 1015-1022.
- Little, M.A., Costello, D.A.E., Harries, M.L. (2011). Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures. *Journal of Voice*, (in press).
- Khalili and Chen (JASA, 2007). Variable selection in finite mixture of regression models.
- Khalili, A. (Can. J. Statist, 2010). New estimation and feature selection methods in mixture-of-experts models.
- Khalili, Chen and Lin (Biostatistics, 2011). Feature selection in mixture of sparse normal linear models in high dimensional feature space.
- Khalili, A., Lin, S. (2011). Regularization in finite mixture of regression models with diverging number of parameters (submitted).
- Tsanas, A., Little, M. A., McSharry, P. E., Ramig, L.O. (2010). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, **57**, 884-893.

Thank you for your attention!