FOLDED CONCAVE PENALIZED LEARNING IN IDENTIFYING MULTIMODAL MRI MARKER FOR PARKINSON'S DISEASE

Hongcheng Liu¹, PhD, Guangwei Du², MD, PhD, Lijun Zhang³, PhD, Mechelle M. Lewis^{2,4}, PhD, Xue Wang¹, MSc, Tao Yao¹, PhD, Runze Li⁵, PhD, Xuemei Huang^{2,4,6,7,8}, MD, PhD

Department of Industrial and Manufacturing Engineering¹, Department of Statistics⁵, The Pennsylvania State University, University Park, PA 16802

Departments of Neurology², Biochemistry and Molecular Biology³, Pharmacology⁴, Radiology⁶, Neurosurgery⁷, and Kinesiology⁸, Milton S. Hershey College of Medicine, The Pennsylvania State University- Hershey, PA 17033

Co-Corresponding Authors:

Xuemei Huang, MD, PhD

Professor and Vice Chair of Neurology

500 University Dr.

The Pennsylvania State University-Milton S. Hershey Medical Center

Hershey, PA 17033

Voice: 717-531-0003, ext. 287082

Email: Xuemei@psu.edu

Runze Li, PhD

Verne M. Willaman Professor of Statistics

Department of Statistics

The Pennsylvania State University, University Park, PA 16802-2111

Voice: 814-865-1555

Email: rzli@psu.edu

Tao Yao, PhD

Associate Professor of Industrial and Manufacturing Engineering,

349 Leonhard Building

The Pennsylvania State University, University Park, PA 16802

Voice: 814-865-8040

Email: taoyao@psu.edu

Word Count:

Title	54
Abstract	250 (≤ 250)
Text words	2966
Table / Figures	8
References	32

Keywords: [max 6] Parkinson's disease; Biomarker discovery; Penalized learning; Magnetic resonance imaging; Diffusion tensor imaging; R2*.

ABSTRACT

Background: Brain MRI holds promise to gauge different aspects of Parkinson’s disease (PD)-related pathological changes. Its analysis, however, is hindered by the high-dimensional nature of the data.

New method: This study introduces folded concave penalized (FCP) sparse logistic regression to identify biomarkers for PD from a large number of potential factors. The proposed statistical procedures target the challenges of high-dimensionality with limited data samples acquired. The maximization problem associated with the sparse logistic regression model is solved by local linear approximation. The proposed procedures then are applied to the empirical analysis of multimodal MRI data.

Results: From 45 features, the proposed approach identified 15 MRI markers and the UPSIT, which are known to be clinically relevant to PD. By combining the MRI and clinical markers, we can enhance substantially the specificity and sensitivity of the model, as indicated by the ROC curves.

Comparison to existing methods: We compare the folded concave penalized learning scheme with both the Lasso penalized scheme and the principle component analysis-based feature selection (PCA) in the Parkinson’s biomarker identification problem that takes into account both the clinical features and MRI markers. We show that the folded concave penalty method demonstrates a substantially better clinical potential than both the Lasso and PCA in terms of specificity and sensitivity.

Conclusions: For the first time, we applied the FCP learning method to MRI biomarker discovery in PD. The proposed approach successfully identified MRI markers that are clinically relevant. Combining these biomarkers with clinical features can substantially enhance performance.

1. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder and is marked pathologically by dopaminergic neuronal loss and presence of Lewy bodies in the substantia nigra of the basal ganglia (BG) and diffuse areas outside of the BG (Politis 2014). *In vivo* biomarker(s) that can reflect PD-related cell loss and associated pathoetiological/physiological processes will greatly enhance research into its etiology and clinical trials to test disease-modifying therapies. Brain magnetic resonance imaging (MRI) holds promise to gauge different aspects of PD-related pathological changes. Identifying PD-related biomarkers is hindered, however, by relatively small sample sizes and the high dimensional nature of the data collected in PD studies.

The statistical challenge for analyzing high dimensional data with a relatively small sample size is common not only in PD biomarker identification but also many other diseases (see examples in Casanova et al., 2011; Yasui et al., 2002; and Fan et al., 2008). Thus, effective statistical procedures for analyzing such data will be of considerable importance to the biomedical research field in general. Due to the relatively high cost in collecting clinical and MRI data in patients, the number of training samples often is smaller than the number of biomarkers to be included in the empirical analysis. This renders traditional statistical methods, such as logistic regression, inapplicable. Penalized logistic regression with folded concave penalty (FCP) has been shown to be the most effective remedy for this type of data and is introduced to identify important biomarkers related to PD in this paper.

The literature refers to the learning problem with a sample size less than the number of input (explanatory) variables as the high dimensional learning problem. Analysis of high-dimensional data poses a non-trivial challenge to most traditional approaches in statistical

learning. The current state-of-the-art is the L1-regularization, a.k.a., Lasso. It (or its variation) not only has been successfully applied to biomarker identification problems by Ghosh and Chinnaiyan (2005), Wu et al. (2012), and Gu et al. (2013), to name only a few, but has become the gold standard for high-dimensional learning (see discussions in Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009; Zhang and Huang, 2008; van de Geer, 2008). A rich literature is devoted to expose conditions under which the Lasso entails a theoretical guarantee of performance (e.g., Bickel et al., 2009; Bunea et al., 2007; Cai et al., 2010; Candès and Tao, 2007, Wainwright, 2009; see van de Geer and Bühlmann, 2009, for a comprehensive review on these conditions.) An advantage of Lasso is that its global optimal solution is tractably computable. Lasso introduces extra estimation bias due to its penalty, however, and requires a strong irrepresentable condition to yield a theoretical guarantee of solution quality (Fan et al., 2014). The FCP method (Fan and Lv, 2011) is a recent approach that enjoys desirable theoretical properties such as the unbiasedness and strong oracle property for high-dimensional sparse estimation. It has been shown to require weaker conditions and entail better statistical properties than Lasso (Zou, 2006; Meinshausen and Bühlmann, 2006; Fan et al., 2014). Two mainstream FCP functions are the clipped absolute deviation (SCAD, Fan and Li, 2001) and minimax concave penalty (MCP, Zhang, 2010). In this paper, we propose to employ the FCP to the biomarker identification problem in PD for the first time.

2. Materials and Methods

2.1 SUBJECTS

Thirty-four PD patients recruited from a tertiary movement disorder clinic were included. PD diagnosis was confirmed by a movement disorder specialist according to the UK brain bank clinical diagnosis criteria (Hughes et al., 1992). No patient had a history of neurological or

psychiatric disease other than PD. Thirty-one healthy control subjects closely matched to the PD patients for age and gender also were selected. All controls were free of any neurological or psychiatric disease. The Hamilton Depression Scale (HDS) was used to assess depression level and the University of Pennsylvania Smell Identification Test (UPSIT) was used to evaluate olfactory function in all subjects. All participants gave written informed consent that was reviewed and approved by the Penn State Hershey Institutional Review Board, and was consistent with the Declaration of Helsinki.

2.2 MRI ACQUISITION

Subjects were scanned with a 3.0-Tesla MRI system (Trio; Siemens Magnetom; Erlangen, Germany) with an 8-channel phased-array head coil. The MRI scan included T1-weighted (T1WI), T2-weighted (T2WI), multi-gradient-echo (for $R2^*$), and diffusion tensor (DTI) imaging sequences. An MPRAGE sequence was used to obtain T1WI with TR/TE=1540/2.34, FOV=256 mm $\times$ 256 mm, matrix=256 $\times$ 256, slice thickness=1 mm (with no gap), and slice number=176. T2WIs were acquired using a fast-spin-echo sequence with TR/TE=2500/316, FOV=256 mm $\times$ 256 mm, matrix=256 $\times$ 256, slice thickness=1 mm (with no gap), and slice number=176. A multi-gradient-echo sequence was used to estimate the transverse relaxation rate, $R2^*$ ($R2^*=1/T2^*$). Six echoes with TE ranging from 7 to 47 ms and an interval of 8 ms were acquired with TR=54 ms, flip angle=20°, FOV=256 mm $\times$ 256 mm, matrix=256 $\times$ 256, slice thickness=1 mm (with no gap), and slice number=64. For DTI, acquisition parameters were as follows: TR/TE=8300/82 ms, b value=1000 s/mm², diffusion gradient directions=42 and 7 b=0 scans, FOV=256 mm $\times$ 256 mm, matrix=128 $\times$ 128, slice thickness=2 mm (with no gap), and slice number=65.

2.3 IMAGE DATA PROCESSING AND FEATURE SELECTION

Each MRI dataset contains millions of voxels. A fully automatic atlas-based segmentation algorithm [AutoSeg v2.9, University of North Carolina Neuro Image Analysis Laboratory (Chapel Hill, NC)] was used to parcel the whole brain volume into regions of interest (ROI) as an initial level of dimensionality reduction (Gouttard et al., 2007; Wang et al., 2014). T1WI and T2WI were segmented using AutoSeg to obtain the following regions-of-interest (ROI): 1) total gray matter (GM), total white matter (WM), and lateral ventricles (LV) at the global level; 2) substantia nigra (SN), putamen (PUT), globus pallidus (GP), caudate nucleus (CN), red nucleus (RN), hippocampus (HIP), amygdala (AMY), and dentate nucleus (DN) at the subcortical level. Volume measures from each ROI then were scaled by total intracranial volume (TIV, sum of GM, WM, and total cerebral spinal fluid) for features used by the classification models. The reason for choosing these regions is to include both global measures (total GM, total WM) and regional measures (subcortical nuclei). The subcortical nuclei include basal ganglia, limbic, and cerebellar related structures, which are potentially relevant to PD-related pathology (Braak et al, 2004; Jellinger, 2008; Mormina, 2015).

Raw DTI images were processed using DTIPrep (Neuro Image Research and Analysis Laboratory, University of North Carolina, Chapel Hill, NC; Oguz et al., 2014). In DTIPrep, a thorough quality control for diffusion weighted images was done by eddy-currents, head motion correction, as well as slice-wise and gradient-wise inconsistency checking. Then, a weighted least squares estimation was used for diffusion tensor estimation. Fractional anisotropy (FA) and mean diffusivity (MD) maps then were generated for subsequent analysis. FA is a scalar value ranging from 0-1. It measures the diffusion of water in a system, with 0 representing unrestricted diffusion and 1 representing perfectly aligned diffusion (e.g., along axons).

Multi-gradient-echo images were used to generate $R2^*$ maps by employing a voxel-wise non-linear Levenberg-Marquardt algorithm to fit a mono-exponential function with free baseline using an in-house MATLAB (The Mathworks, Inc., Natick, MA) tool.

For DTI and $R2^*$ images, the ROIs generated from T1WI space were co-registered to DTI and $R2^*$ maps via a multi-step co-registration pipeline using ANTS (Avants et al, 2008). The mean FA and $R2^*$ values for each ROI then were calculated using an in-house MATLAB tool and later used as features for the statistical classification models. All features used for classification are listed in Table 1.

Table 1: The full list of potential biomarkers.

Predictor	Predictor	Predictor
Gender	WM.r2s	Pallidus.fa
Age	GM.r2s	Putamen.fa
UPSIT	Amygdala.r2s	SubstantiaNigra.fa
HDS	Caudate.r2s	RedNucleus.fa
WM.vol	Hippocampus.r2s	Dentate.fa
GM.vol	Pallidus.r2s	WM.md
Amygdala.vol	Putamen.r2s	GM.md
Caudate.vol	SubstantiaNigra.r2s	Amygdala.md
Hippocampus.vol	RedNucleus.r2s	Caudate.md
LatVentricle.vol	Dentate.r2s	Hippocampus.md
Pallidus.vol	WM.fa	Pallidus.md
Putamen.vol	GM.fa	Putamen.md
SubstantiaNigra.vol	Amygdala.fa	SubstantiaNigra.md
RedNucleus.vol	Caudate.fa	RedNucleus.md
Dentate.vol	Hippocampus.fa	Dentate.md

2.4 Folded concave penalized learning vs. traditional methods

In our empirical analysis, a logistic regression model was applied to study Parkinson's biomarkers. Logistic regression has been popular for modeling the relationship between a binary response variable and a set of input (explanatory) variables. It further can be applied for the prediction of binary responses based on the newly observed input variables. Logistic regression has been applied to the study of disease diagnosis by Alkan et al. (2005) and Kennedy et al. (1998), among others. This study is distinguished from the above research, however, in the sense that we deal with the issue of high dimensionality, given a large number of input variables but a small number of samples. Liao and Chin (2007) considered an under-sampled logistic regression and proposed a parametric bootstrap to reduce the model prediction error. In contrast, this study presents a substantially different statistical learning approach to handle the issue of high dimensionality.

Given a set of random samples, $(x_i, y_i): i = 1, \dots, n$, where x_i is a p -dimensional vector consisting of features (explanatory variables) and n is the sample size, logistic regression minimizes its negative logarithm of the likelihood function

$$l(\beta) = \sum_{i=1}^n -y_i x_i^T \beta + h(x_i^T \beta) \quad (1)$$

with respect to $\beta = (\beta_j: j = 1, \dots, p)$, where $h(t) = \log(1 + \exp t)$ is the canonical link function. Traditional logistic regression will break down here because the corresponding minimization problem becomes ill-posed when $p > n$. In high-dimensional data analysis, it is common that only a few of the explanatory variables actually may have an impact on the response variable(s), whereas the many other features may be irrelevant. Following this convention, we employed the FCP method, that is, we propose to minimize the following penalized logistic regression:

$$\min_{\beta=(\beta_j: j=1,\dots,p)} \frac{1}{n} l(\beta) + \sum_{j=1}^p P_\lambda(|\beta_j|), \quad (2)$$

where $P_\lambda(\cdot)$ is a penalty function with a tuning parameter λ . Two popular choices of FCPs are the smoothly clipped absolute deviation penalty (SCAD, Fan and Li, 2001) and the minimax concave penalty (MCP, Zhang, 2010). In this research, we will focus on the latter since the results of the SCAD are similar to the MCP. The specific form of the MCP is

$$P_\lambda(t) = \int_0^t \frac{(a\lambda - t)_+}{a} dt \quad (3)$$

for some pre-specified $a > 0$ and tuning parameter $\lambda > 0$. We employed the local linear approximation (LLA) algorithm (Zou and Li, 2008) to search for the solution of the penalized logistic regression. It has been shown by Fan et al. (2014) that the LLA algorithm generates a statistically desirable solution, the oracle solution, with overwhelming probability. The oracle solution here is the solution that correctly predicts the significant features and estimates the coefficients for the significant features as if no regularization term is added. The pseudo-code for LLA is given in Figure 1. The output of the LLA algorithm is a model for diagnosing PD. This model is written as:

$$Prob_{PD}(x) = \frac{1}{1 + \exp(-x^T \beta)}. \quad (6)$$

For the observed features x of a new subject, the value $Prob_{PD}(x)$ gives the probability for this subject to be diagnosed with PD. The LLA automatically determines the coefficients denoted β .

Computing Procedures of Local Linear Approximation (LLA)

Step 1. For some tuning parameter $\lambda_{lasso} > 0$, solve a Lasso problem of the form:

$$\beta^0 = \underset{\beta=(\beta_j: j=1,\dots,p)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \{-y_i x_i^T \beta + h(x_i^T \beta)\} + \lambda_{lasso} |\beta_j| \quad (4)$$

Step 2. Solve the following weighted Lasso problem:

$$\beta^{m+1} = \underset{\beta=(\beta_j: j=1,\dots,p)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \{-y_i x_i^T \beta + h(x_i^T \beta)\} + \sum_{j=1}^p P_\lambda(|\beta_j^m|) \quad (5)$$

Step 3. The algorithm terminates, if $\|\beta^m - \beta^{m+1}\| < \delta$, for a predetermined $\delta > 0$. Otherwise, let $m = m + 1$ and go to Step 2.

FIGURE 1: Pseudo-code for LLA.

3. RESULTS

3.1 BIOMARKER IDENTIFICATION

Recall that Table 1 presents a list of the 45 potential biomarkers, among which four markers easily can be obtained from clinical assessment. The rest of them are all MRI biomarkers including volume (ending with “.vol”), DTI (ending with “.fa” or “.md”), and R2* (ending with “.r2s”). We employ the proposed approach to analyze the data using all samples. The selected features and their estimated coefficients β in Eq. (6) are listed in Table 2. The estimated coefficients for unselected features in Table 1 equal 0. As explained in Section 4, the selected variables have strong clinical implications. To evaluate how sensitive the selected model was to the change of data, we randomly selected 36 out of the 65 total subjects to fit the penalized logistic regression. We replicated this procedure 50 times. In the column labeled ‘Prop’ in Table 2, we report the proportion of times each variable was excluded from the model

generated by the LLA over the 50 random replications. All variables except for GM.r2s and RedNucleus.r2s were very robust. GM.r2s and RedNucleus.r2s were excluded 32% and 16% of the time in the LLA replication model, respectively. This might be caused by potential collinearity. To see this, we further examined the correlation among the 45 variables listed in Table 1. We found that the correlation between GM.r2s and WM.r2s was 0.875, the correlation between GM.r2s and Dentate.r2s was 0.776, and the correlation between RedNucleus.r2s and SubstantiaNigra.r2s was 0.627. These correlations were much higher than other correlations among the 45 variables. This indicates that GM.r2s and RedNucleus.r2s are highly correlated with some variables being excluded from the final selected model. The instability of these two variables likely was due to the near collinearity of these two variables with others. It seems that the MCP method cannot deal well with collinearity. This can be viewed as a limitation of the method.

To study the contribution of each variable to the model, we conducted a forward stepwise procedure: Starting from the null model, we added variables listed in Table 2 one by one to the model. For a given submodel that includes k variables in Step k , we added a variable from the other $15 - k$ variables that were not in the submodel. That newly added variable was selected to yield the maximal reduction in deviance compared to the remaining $14 - k$ variables. We carried out this procedure with $k=1, 2, \dots, 14$. Table 2 shows the order of each predictor being added to the model and the deviance of the submodel in each step. The deviance of the null model (i.e. model with intercept only) was 93.74. From Table 2, it can be seen that the clinical variable UPSIT contributed the most to the model, followed by the Caudate.fa and GM.md. The biomarkers WM.vol, WM.fa, Hippocampus.fa, RedNucleus.fa, and Pallidus.fa contributed the

least to the model, but they were very stable over the 50 random replications for inclusion in the selected model. Thus, it was appropriate to include them in the selected model.

Table 2: Estimated coefficients of selected biomarkers by the folded concave penalized logistic regression analysis. “Prop” stands for the proportion of times the variable was excluded over the 50 random replications, “Order” stands for the order of the variable being included in the forward stepwise procedure, and “Deviance” in the row with “Order” equal to k stands for the deviance of the model that includes all the predictors whose values of “Order” are smaller than or equal to k (e.g., for GM.md, the “Order” equals 3. Then the corresponding “Deviance” stands for the deviance of the model that consists of all predictors with “Order” smaller than 3, namely, the model with predictors: UPSIT, Caudate.fa, and GM.md.).

Predictor	$\hat{\beta}$	Prop	Order	Deviance	Predictor	$\hat{\beta}$	Prop	Order	Deviance
UPSIT	-8.4186	0%	1	25.58	Amygdala.r2s	-2.3745	0%	9	0.56
Caudate.fa	-0.9646	0%	2	15.27	GM.r2s	1.3589	32%	10	0.49
GM.md	-1.8920	0%	3	7.65	WM.vol	0.8242	0%	11	0.45
SubstantiaNigra.vol	-0.6852	0%	4	5.19	WM.fa	-1.6549	0%	12	0.44
Pallidus.r2s	-2.267	4%	5	3.09	Hippocampus.fa	1.3284	0%	13	0.43
RedNucleus.r2s	0.8803	16%	6	1.75	RedNucleus.fa	-0.6324	0%	14	0.41
RedNucleus.md	0.6048	0%	7	0.96	Pallidus.fa	-0.3893	4%	15	0.39
SubstantiaNigra.r2s	1.2625	0%	8	0.70	Null model				93.74

3.2 MODEL COMPARISON

In this section, we compare the basic model that contains only the clinical features and the basic+MRI model that takes into account both variables in the basic model and the MRI biomarkers. The primary comparison criteria are sensitivity and specificity. Involved in the comparison are four different learning techniques: (i) the basic model, (ii) the folded concave penalized logistic regression, referred to as “FCP: basic+MRI”, (iii) the Lasso penalized logistic regression, referred to as “Lasso: basic+MRI”, and (iv) a PCA-based logistic regression, which first invokes PCA to screen the number of features to equal fifteen (which is equivalent to the number of selected features generated by our proposed approach) utilizing all of the samples and then using the ordinary logistic regression to determine the coefficients for the features using the

training set data (randomly subsampled as described subsequently). Notice that the PCA yielded a set of features that captures 93.83% of the variability. To compare the above approaches, we randomly selected 36 out of the 65 total subjects as the training set, whereas the rest of the subjects were used as the validation set. We used the training set to train the parameters of the four different models. We then inserted each subject from the validation set into the models. For each validation subject, the models generated the probability indicating the likelihood of the corresponding subject being diagnosed with PD. We then calculated sensitivity, specificity, and correctness using different choices for the decision thresholds. The above process was repeated 50 times and the results are summarized in Figures 2-5. Figures 2 and 3 present the comparisons related to sensitivity and specificity, with the error bars indicating the 95% confidence intervals calculated via the 50 random replications. Figure 4 compares the total error rate and Figure 5 presents the ROC curves. It can be seen from these four figures that the combination of the basic model and MRI biomarkers substantially improved the accuracy in diagnosis compared to the basic model alone, using either the FCP or Lasso regularization method. The ROC curves (Figure 5) indicated that the basic+MRI model provided a better trade-off than the basic model. Between the two high-dimensional learning schemes, FCP outperformed the Lasso method substantially. Among all four approaches, the PCA-based approach appeared to perform the worst. We also present in Table 3 the standard deviations for the area under the curve (AUC) out of the 50 random replications.

Table 3: AUCs and their standard deviations.

Model	Mean	Std.
Basic	0.934	0.04
Lasso: Basic+MRI	0.979	0.02
FCP: Basic+MRI	0.997	0.01
PCA: Basic+MRI	0.778	0.09

FIGURE 2: Comparison between the basic model and three basic+MRI models in terms of sensitivity. The horizontal coordinates are different choices of threshold values.

FIGURE 3: Comparison between the basic model and three basic+MRI models in terms of specificity. The horizontal coordinates are different choices of threshold values.

FIGURE 4: Comparison between the basic model and three basic+MRI models in terms of correctness. The horizontal coordinates are different choices of threshold values.

FIGURE 5: Comparison of the ROC curves.

4. Discussion

PD presents clinically with both motor and non-motor symptoms, and pathologically it is known to involve structures both within and outside of BG and cerebellar regions. Based on the prior knowledge of potential brain structures involved in PD, we collected a total of 41 MRI features within and outside BG and cerebellar structures that potentially could discriminate PD from controls (see Table 1). To ensure that the proposed procedure worked properly, we also added four clinical features, among which two are known to contribute to the discrimination of PD from healthy controls.

Our model identified sense of smell (measured using the UPSIT) as the strongest feature that predicts controls, confirming that our model worked properly. MRI features, however, also enhanced the accuracy (see the improved ROC curve, Figure 5).

In general, higher FA values and a higher volume in a given structure are related to better structural organization in the brain, whereas higher $R2^*$ values are associated with increased neurodegenerative processes and higher iron deposition. The computed directionalities of the correlations observed in the current study are consistent with our understanding of the clinical and biological meanings of the measurements. For example, all clinical and MRI features known to be linked positively to the disease (in the sense that a higher value may indicate a lower likelihood of getting the disease) primarily were identified as having positive associations with controls. Meanwhile, the features known to be negatively associated with disease were identified as associations with PD. Exceptions were that the computed model predicted that higher hippocampal FA values were associated with PD, and higher $R2^*$ values in the amygdala and globus pallidus were associated with control status. Those predictions have not yet been clinically verified. In fact, the hippocampus, amygdala, and globus pallidus are not conventionally the primary pathological sites of PD. Thus, it is possible that the MRI changes we

detected may be related to compensatory processes and this possibility shall be investigated in future studies.

The FCP method has been proposed in the theoretical statistical literature and has been shown to enjoy beneficial theoretical properties, such as the oracle property (Fan and Li, 2001, Fan and Lv, 2011). The FCP method can reduce substantially the model complexity and yield a parsimonious model that provides for a better model interpretation and likely leads to better prediction power. Compared with the traditional variable selection methods, such as stepwise regression and the best subset selection method with a certain selection criterion (e.g., the commonly-used AIC and BIC), the FCP method removes irrelevant variables by estimating their coefficients to be zero. Furthermore, the FCP method identifies important variables and estimates their coefficients simultaneously. Compared with the Lasso method, the FCP approach can reduce the estimation bias caused by the Lasso penalty. Due to the bias reduction, the FCP method typically yields a more parsimonious model than the Lasso method, and therefore may reduce false positives and provide better prediction power. Indeed, we have demonstrated by empirical analyses that the FCP approach may perform better than the Lasso method. To our knowledge, this study is the first to employ the FCP method to identify important biomarkers related to PD. The same approach easily is extendable to the automatic diagnosis of many other diseases. Therefore, it is of great interest to explore the potential developments and further applications of the FCP method in facilitating the diagnosis and treatment of PD or other diseases.

5. CONCLUSION

The Parkinson's biomarker identification problem, together with some other biomarker development problems, suffers from the issue of high-dimensionality: the sample size is small

relative to the number of the potential biomarkers. High dimensionality renders traditional statistical learning approaches invalid. In this paper, we implement the FCP method for biomarker discovery in PD for the first time. We demonstrate that it substantially outperforms some benchmark high-dimensional learning schemes, Lasso or PCA-based methods, in terms of specificity and sensitivity. The proposed high-dimensional learning scheme successfully identifies clinical and MRI features that are clinically relevant. Further studies are needed to validate this method in larger PD datasets and determine its applicability for biomarker discovery in other diseases.

Acknowledgements

Dr. Li's research was supported by NIDA, NIH grants P50 DA036107, P50 DA039838 and NSF grant DMS 1512422. This work also was supported by NINDS (NS060722 and NS085121), and by Penn State Grace Woodward Collaborative Engineering/Medicine Research Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIDA, NIH or NSF.

References:

- Alkan, A, Koklukaya, E, Subasi, A. Automatic seizure detection in EEG using logistic regression and artificial neural network. *Journal of Neuroscience Methods* 2005;148:167-176.
- Avants BB, Epstein CL, Grossman M., and Gee, J.C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal.* 2008;12:26-41
- Bickel, P.J., Ritov, Y., and Tsybakov, A.B. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics.* 2009;37:1705-1732
- Braak H, Ghebremedhin E, Rüb U, Bratzke H, Del Tredici K. Stages in the development of Parkinson's disease-related pathology. *Cell Tissue Res.* 2004;318(1):121-34. Bunea, F., Tsybakov, A., and Wegkamp, M. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* 2007;1:169-194
- Cai, T., Wang, L., and Xu, G., 2010. Stable recovery of sparse signals and an oracle inequality. *IEEE Transactions on Information Theory* 2010;56:3516-3522
- Candès, E., and Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics* 2007;35:2313-2351
- Casanova, R., Whitlow, C.T., Wagner, B., Williamson, J., Shumaker, S.A., Maldjian, J.A., and Espeland, M.A. High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization. *Frontiers in Neuroinformatics* 2011;5:22
- Fan, J., Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistician Association* 2001;96:1348-1360.

- Fan, J., Lv, J. Non-concave penalized likelihood with NP-dimensionality. *IEEE Information Theory*, 2011:57:5467-5484.
- Fan, J., Xue, L., Zou, H. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics* 2014:42:819-849.
- Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C., Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study. *NeuroImage* 2008:41:277-285.
- Ghosh, D., Chinnaiyan, A.M. Classification and selection of biomarkers in genomic data using Lasso. *J. Biomed. Biotechnol.* 2005: 2005(2): 147-154.
- Gouttard, S., Styner, M., Joshi, S., Smith, R. G., Hazlett, H. C., and Gerig, G. Subcortical structure segmentation using probabilistic atlas priors. *Medical Image Computing and Computer Assisted Intervention Workshop* 2007:37-46.
- Gu, X., Yin, G., and Lee, J.J. Bayesian two-step lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary Clinical Trials* 2013:36:642-650.
- Hughes AJ, Ben Shlomo Y, Daniel SE et al. What features improve the accuracy of clinical diagnosis in Parkinson's disease: a clinicopathologic study. *Neurology*. 1992;42:1142-1146
- Jellinger KA. A critical reappraisal of current staging of Lewy-related pathology in human brain. *Acta Neuropathol.* 2008;116(1):1-16. doi: 10.1007/s00401-008-0406-y.
- Kennedy, R. L., Burton, A. M., Fraser, H. S., McStay, L. N., and Harrison, R. F. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation:

- derivation and evaluation of logistic regression models. *European heart journal* 1996;17:1181-1191.
- Liao, J. G., and Khew-Voon Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics* 2007;23:1945-1951.
- Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 2006;34:1436-1462.
- Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* 2009;37:246-270.
- Mormina, E, Arrigo, A, Calamuneri, A, Granata, F, Quartarone A, Ghilardi, MF, Inglese, M, Di Rocco, A, Milardi, D, Anastasi, GP, Gaeta, M, Diffusion tensor imaging parameters' changes of cerebellar hemispheres in Parkinson's disease. *Neuroradiology* 2015;57:327-334
- Politis, M. Neuroimaging in Parkinson disease: from research setting to clinical practice. *Nat. Rev. Neurol.* 2014;10:708-722.
- Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., Johnson, H.J., Styner, M. DTIPrep: quality control of diffusion-weighted images. *Front Neuroinform.* 2014;8;4.
- Wainwright. M. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1 -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* 2009;55:2183-2202.
- Wang J., Vachet C., Rumple A., Rumple A., Gouttard S., Ouziel C., Perrot E., Du G., Huang X., Gerig G., and Styner M. Multi-atlas segmentation of subcortical brain structures via the AutoSeg software pipeline. *Front Neuroinform.* 2014; 8:7. doi: 10.3389/fninf.2014.00007.

- Wu, M.-Y., Dai, D.-Q., Shi, Y., and Yan, H., 2012. Biomarker identification and cancer classification based on microarray data using Laplace naïve Bayes model with mean shrinkage. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2012:9:1649-1662.
- van de Geer, S.A., High-dimensional generalized linear models and the lasso. *The Annals of Statistics* 2008:36:614-645.
- van de Geer, S.A., and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* 2009:1360-1392.
- Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.-L., Wright Jr., G.L., Qu, Y., Potter, J.D., Winget, M., Thornquist, M., and Feng, Z., A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2004:4:449-463.
- Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 2010:38:894-942.
- Zhang, C.-H., and Huang, J., The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 2008:36:1567-1594.
- Zou, H, Li, R. One-step sparse estimation in non-concave penalized likelihood method. *Annals of Statistics*, 2008:36:1509-1533.

Graphical Abstract (for review)

MRI Biomarkers (T1-weighted)

Folded concave penalized high-dimensional learning