Penalized least squares regression methods and applications to neuroimaging

Florentina Bunea a, Yiyan She a, Hernando Ombao b,⁎, Assawin Gongvatana c, Kate Devlin c, Ronald Cohen c

a Florida State University, Department of Statistics, USA
b Brown University, Biostatistics Section, USA
c Brown University, School of Medicine, USA

A B S T R A C T

The goals of this paper are to review the most popular methods of predictor selection in regression models, to explain why some fail when the number of explanatory variables exceeds the number of participants, and to discuss alternative statistical methods that can be employed in this case. We focus on penalized least squares methods in regression models, and discuss in detail two such methods that are well established in the statistical literature, the LASSO and Elastic Net. We introduce bootstrap enhancements of these methods, the BE-LASSO and BE-Enet, that allow the user to attach a measure of uncertainty to each variable selected. Our work is motivated by a multimodal neuroimaging dataset that consists of morphometric measures (volumes at several anatomical regions of interest), white matter integrity measures from diffusion weighted data (fractional anisotropy, mean diffusivity, axial diffusivity and radial diffusivity) and clinical and demographic variables (age, education, alcohol and drug history). In this dataset, the number of explanatory variables exceeds the number of participants. We use the BE-LASSO and BE-Enet to provide the first statistical analysis that allows the assessment of neurocognitive performance from high dimensional neuroimaging and clinical predictors, including their interactions. The major novelty of this analysis is that biomarker selection and dimension reduction are accomplished with a view towards obtaining good predictions for the outcome of interest (i.e., the neurocognitive indices), unlike principal component analysis that are performed only on the predictors’ space independently of the outcome of interest.

© 2010 Elsevier Inc. All rights reserved.

Introduction

Biomarker discovery has become a major focus of clinical research directed at a large number of different diseases, including neurological disorders, such as Alzheimer’s disease and Stroke. The general strategy of these research efforts is to determine whether particular clinical or laboratory measures can be used to predict clinical outcome, including beneficial response to specific treatments. A common problem in this type of research is that there are often many more potential clinical measures that need to be considered than there are clinical cases in the study, confounding statistical modeling efforts. This is often the case of neuroimaging studies, in which multimodal neuroimaging data from a large number of brain regions may exist, and the goal is to use these data to predict a clinical outcome. Here, the number of potential explanatory variables can be much larger than the number of participants in the data set (i.e., \( P \gg N \)), and the application of standard statistical methodology becomes problematic. For instance, one common approach is to select first a manageable number of variables by performing stepwise forward selection or backward elimination, or a combination of the two, in regression models. However, it is well known that such procedures are greedy and may not recover the relevant set of predictors accurately. This drawback is particularly severe when \( P \) is large (Tibshirani, 1996).

Another common approach is to perform first dimension reduction on the space of predictors via principal components analysis (PCA), by selecting those components that explain most of the variance, and then fit a regression model of the response on the selected components. There are two major drawbacks of this approach: (i) The principal components are, by construction, linear combinations of the original predictors. Therefore, they may not necessarily yield easily interpretable results. Moreover, if one of the goals of the analysis is to identify a smaller set of predictors to be used in future studies, PCA cannot be used for this purpose: the principal components are, by definition, linear combinations of all \( P \) original predictors and thus require the collection of all the predictors. (ii) Another major issue, perhaps the most important, remains: most relevant components are selected without any regard for the outcome of interest. For instance, in our application, there is no guarantee that the selected PCs are those relevant for neurocognitive performance because no information on the neurocognitive indices is directly used in reducing the dimension of the space of explanatory variables.

In this article we discuss alternative approaches, based on recent statistical methods tailored specifically to the analysis of large

⁎ Corresponding author. Fax: +1 401 863 9182.
E-mail address: ombao@stat.brown.edu (H. Ombao).

1053-8119/ See front matter © 2010 Elsevier Inc. All rights reserved.
dimensional \( P \gg N \) data. We focus on penalized least squares (PLS) methods. The PLS estimators minimize the residual sum of squares plus the penalty term. The aim of PLS is to select, from a large list of possible candidate models, the model that achieves the best trade-off between goodness of fit and model complexity. Here, models that use a smaller number of predictors are said to be of lower complexity than those that use a larger number of predictors. Following standard statistical terminology, models with a low number of predictors are also called sparse. Our goal in this paper is to discuss a number of specific methods that identify sparse models without sacrificing their goodness of fit and prediction accuracy. We will focus on computationally efficient methods. The type of penalty term one employs, convex or non-convex, is responsible for the computational complexity of the resulting PLS method. Since convex penalties have lower computational cost, we will focus on them in this article, with particular emphasis on the LASSO type penalty (Tibshirani, 1996) and its close variant, the Elastic Net (Zou and Hastie, 2005). One important aim of our work is to introduce some variants of these methods (Limitations of the LASSO and E-Net methods section) and to illustrate how they can be used for identifying neuroimaging and clinical biomarkers that best predict neurocognitive performance.

Penalized least squares methods are not new to neuroscience — although they have been previously employed to problems different than the one we treat here. For instance, one of the early applications to neuroimaging is the estimation of electrophysiological sources from electroencephalogram data. We refer the reader to Scherg and Voncramon (1986); Mosher and Leahy (1998); Uutela et al. (1999); Hämmäläinen and Ilmoniemi (1994); Silvaa et al. (2004); Valdés-Sosa et al. (2009) and Pascual-Marqui et al. (1994); Pascual-Marqui (2002) among many others. More recent applications of the PLS, in particular of the LASSO and E-Net methods, include Gunn et al. (2002), for modeling the dynamics of radiotracers, Carroll et al. (2009), for voxel selection in fMRI analysis, Vounou et al. (2010) in multivariate neuroimaging and genetics problems, Ryali et al. (2010) for classification problems.

The recent advent of the LASSO-type methods in many scientific areas and their increased usage by the neuroscience community at large motivated the structure of this work, with contributions outlined below:

(i) We present a detailed overview of the statistical properties of those penalized least squares (PLS) methods that can be used for variable selection and have high prediction accuracy. Our focus is on those methods that are computationally efficient when \( P > N \). In this case, we discuss why the traditional methods based on testing are no longer reliable. Moreover, some of the more standard PLS methods, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) based methods, fail computationally when \( P \geq 20 \). We explain why LASSO-type methods provide a computationally attractive alternative, with sound theoretical foundation. This is presented in the A review: the LASSO and the Elastic Net methods section and Comparison of the LASSO-type methods with other variable selection methods sections. We also discuss some limitations of these methods in Limitations of the LASSO and E-Net methods section.

(ii) We introduce variants of the LASSO and Elastic Net: the Bootstrap Enhanced Lasso (BE-LASSO) and the Bootstrap Enhanced Elastic Net (BE-Enet). Since predictor selection, by any method, is accompanied by a certain amount of uncertainty, we propose to give a measure for it. We use each method, the LASSO and the Elastic Net, to select a subset of predictors. We then re-sample from the data, repeat the selection process and record the selected set. We repeat this a number of times and summarize the whole selection process by giving the percentage of times each predictor was selected — which we call the variable inclusion probability (VIP). Predictors with high VIP will be investigated further. Full details on the implementation of this procedure are given in Proposed method: bootstrap enhanced LASSO and E-Net section.

(iii) We use the BE-LASSO and BE-Enet to provide the first statistical analysis that allows the assessment of neurocognitive performance from high dimensional neuroimaging and clinical predictors. The major novelty is that dimension reduction is performed with a view towards obtaining good predictions for the outcomes of interest (i.e., the neurocognitive indices). Prior analyses of the neurocognitive performance were either done in a lower dimensional setting, for instance via AIC/BIC, or involved dimension reduction techniques like PCA, with drawbacks discussed above.

Our analysis revealed that fractional anisotropy at the internal capsule appears as an important predictor. This finding is consistent with other studies (e.g., Shenkin et al., 2003) that show that higher FA in the anterior limb of the internal capsule and improved response time in a visual target-detection task. However, it is interesting that the BE-LASSO and BE-Enet methods indicate the importance of the quadratic term which suggests a complex non-linear relationships between white matter injury and neurocognitive performance, which would likely not have been identified via the traditional regression methods. The data description, analysis and discussion of the results are the content of Application to neuroimaging data section. We present an overall summary of this work in the Conclusion section.

Statistical methodology

A review: the LASSO and the Elastic Net methods

The LASSO method is a particular penalized least squares method. The method consists in computing \( \hat{\beta} \) that minimizes the following criterion

\[
\sum_{i=1}^{N} \left[ Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right]^2 + \lambda \sum_{j=1}^{p} |\beta_j| ,
\]

where \( Y_i \) is the dependent variable for subject \( i \); \( X_{ij} \) denotes a measurement on a predictor \( X_j \) for subject \( i \); and \( \lambda \) is the tuning parameter of the method whose role is to provide a balance between prediction accuracy and sparsity. In the context of neuroimaging, the independent variable \( Y_i \) is one of the neurocognitive assessments in the study (e.g., Grooved Pegboard, Trail Making A, Trail Making B, and so on) and the predictors \( X_j \)'s include clinical, demographic, volumetric and diffusion tensor imaging variables. Moreover, to anticipate possible non-linearity in the relationship between neurocognitive assessment and the independent variables, \( X_j \)'s can include quadratic terms of predictors (e.g., squared values of fractional anisotropy in the internal capsule) and interactions between predictors (e.g., fractional anisotropy in the internal capsule \( \times \) mean diffusivity in the corpus callosum). We use the notation \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_P) \), where \( \hat{\beta}_j \) is the estimated coefficient of a generic predictor \( X_j \).

Advantages of the LASSO method

The appeal and wide-spread usage of the LASSO for variable selection is justified by the following properties: (i) One remarkable property of the LASSO method is that it yields a sparse estimate \( \hat{\beta} \), with some components \( \hat{\beta}_j \) exactly equal to zero. The predictors \( X_j \) with zero coefficient estimates are discarded from the model, and only the rest are kept. This justifies the usage of the LASSO as a subset selection method; (ii) Another remarkable property of the LASSO is the scope of the method: it can be applied in the challenging situations when one has measured more predictors \( P \) than one has subjects \( N \) in the study. The theoretical validity of these estimates can be established via probabilistic arguments that are valid for any observed number of predictors \( P \) and sample size \( N \). In particular, Bunea et al. (2007a,b); Zhang and Huang
(2008) and Meinshausen and Yu (2009) show how the LASSO estimators behave for any given N and P, and even when P>N (iii) LASSO corresponds to a convex minimization procedure. Then, any algorithm that finds this minimum comes with the mathematical guarantee that it detects the global, overall, minimum. For this reason, there exist fast algorithms for computing the LASSO estimators, for instance, the homotopy methods (Osborne et al., 2000), the LARS (Efron et al., 2004), interior point methods (Kim et al., 2007), and an iterative thresholding algorithm (Daubechies et al., 2004; Friedman et al., 2007) among others. Friedman et al. (2007) provide an R package, glmnet, which is freely available online. We note that these algorithms are fast even when P is much larger than N, for instance P can be of the order N² or even larger.

Advantages of the Elastic Net method

When the predictors are highly correlated a further immediate improvement of the LASSO is the Elastic Net (Zou and Hastie, 2005). This corresponds to adding to the LASSO penalty a quadratic penalty term, designed to compensate for the correlation between predictors. Specifically, for given tuning parameters λ and µ, the Elastic Net (E-Net) estimator is the vector \( \hat{\beta} \) which minimizes the criterion

\[
\sum_{i=1}^{N} \left[ Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right]^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \mu \sum_{j=1}^{p} \beta_j^2.
\]

The minimization problem is still convex, and computationally optimal algorithms exist — see, e.g., the R package elasticnet. The E-Net estimator \( \hat{\beta} \) shares the sparsity properties of the LASSO estimator \( \hat{\beta} \). When predictors have high correlations, \( \hat{\beta} \) leads to more accurate predictions of the response than \( \hat{\beta} \). The theoretical properties of the E-Net estimate are also well understood, see e.g. Bunea (2008), who suggests caution in the choice of the tuning parameters of the method. In particular, if µ is too large relative to λ, then \( \hat{\beta} \) behaves essentially as the so called ridge regression estimator, \( \hat{\beta} \), the minimizer of

\[
\sum_{i=1}^{N} \left[ Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right]^2 + \mu \sum_{j=1}^{p} \beta_j^2.
\]

The ridge estimator \( \hat{\beta} \) has no components exactly to zero and therefore cannot be used for model selection. However, if both λ and µ in Eq. (2) are chosen by cross validation over a carefully selected grid, as described in the Proposed method: bootstrap enhanced LASSO and E-Net section below, the E-Net estimator \( \hat{\beta} \) will have the desirable sparsity property. The selection of the two tuning parameters is computationally more complex, but it is a price worth paying for models with highly correlated predictors. In practice, the E-Net estimator is a useful companion of the LASSO estimator and typically one fits both and chooses the one that yields the smallest mean squared error of the fit. We discuss the practical implementation of these estimators, together with our suggested modification in the Proposed method: bootstrap enhanced LASSO and E-Net section below.

Comparison of the LASSO-type methods with other variable selection methods

Before we discuss the practical implementation of the LASSO based methods for our analysis, we first contrast them with other existing variable selection methods, with emphasis on the “Large P, small N” case.

Variable selection methods based on testing are problematic when P>N

It is well agreed upon in the statistical community that the standard variable selection methods based on hypotheses testing are not appropriate when P>N. To give a brief overview of these methods, we recall that the traditional selection methods based on hypothesis testing make use of p-values. To compute the p-values, the practitioner typically needs to: (i) either make exact distributional assumptions on the error terms in the model and perform what is referred to as exact tests — the F-test for comparison of two nested models and the t-test are examples of such tests; (ii) or use asymptotic results that lead to approximate distributions on which testing is based. To date, the latter is valid only when \( P \leq N \). When \( P > N \) such tests cannot be performed, as they involve computing least squares estimators, which are not defined in this case. Therefore, choosing among a large number of candidate models, possibly non-nested, poses a challenge. Even worse, even if approaches (i) and (ii) or their variants can be justified for particular cases, when \( P > N \), the open problem of choosing the cut-off for multiple p-values remains: it is highly subjective what one declares as a significant predictor. There exist a number of corrections for the choice of the cut-off of the p-values, the so called False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) type methods. However, for models with correlated predictors, they are only theoretically valid when \( P \leq N \), sometimes much smaller, see, e.g. Bunea et al. (2006) and the references therein. Also, we add a note of caution on using ad-hoc procedures that would involve, generically, the following steps: (1) Fitting models that contain subsets of the original P variables that have size smaller than N and (2) Computing p-values to assess the importance of predictors in the arbitrarily chosen smaller models. Such procedures cannot be used reliably to select relevant variables. The relative ordering of the p-values depends crucially on which predictors where included in the original model on which the p-values were based. For instance, a generic predictor \( X_i \) may appear significant in a model where only, say, two other predictors, \( X_2 \) and \( X_3 \), were included, but loose significance if \( X_2 \) and \( X_3 \) are added and the model is re-fitted. This is particularly visible if the predictors are correlated.

Other variable selection methods based on penalized least squares

Whereas the LASSO and E-Net penalties are convex, and can be computed efficiently even when P is very large, there exists a large family of the PLS methods that employ non-convex penalties, which may incurs a much higher computational cost. An extreme case corresponds to the classic BIC (Schwarz, 1978), AIC (Akaike, 1974), Mallow’s \( C_p \) (Mallows, 1973) selection criteria and their variants. All these criteria involve fitting all possible \( 2^p \) regression models corresponding to all possible subsets of predictors. For each fitted model one computes a criterion that equals the mean squared error of the fitted model plus a term proportional with the number of predictors in that model. Then one selects the final model as the one with the smallest value of this criterion. This class of estimators has attractive theoretical properties see e.g. Bunea et al. (2007) for an overview. The computation of these estimators involves either an exhaustive search over the space of all \( 2^p \) possible models, or approximate forward or backward stepwise procedures to produce a series of candidate models to minimize the criterion. Despite their success in selecting relevant predictors when P is much smaller than N, they fail computationally as soon as P>20, and it has been shown that the problem is NP hard computationally: no combinatorial algorithm can solve this problem in polynomial time.

To address this major computational issue, a number of other non-convex penalties and optimization techniques have been introduced lately, see for instance Antoniadis and Fan (2001) for the SCAD penalty, Zou and Li (2008) for the LLA optimization, She (2009) for the hybrid hard-ridge penalty and the TISP, Zhang (2010) for firm shrinkage and MCP, and the references therein. All these newly developed methods can be implemented efficiently even if P is larger than N. These penalties and the resulting estimators are especially useful for models in which the predictors are very highly correlated. There are many subtle theoretical differences between model selection realized via convex
penalties as the LASSO and the above mentioned methods corresponding to these non-convex penalties. However, there is also a major difference. If one employs a non-convex penalty, the criterion function can have many local minima. The research on choosing the best local minimum is still under development. In addition, the computational cost is higher than that of the convex methods as LASSO. It is for this reason that in this paper we only focus and describe further in detail the concrete implementation of the LASSO type methods.

Limitations of the LASSO and E-Net methods

The performance of the LASSO and E-Net methods depends crucially on the choice of the respective tuning parameters $\lambda$ and $\mu$ of each method. It is known practically (see, e.g. Shi et al. (2007)) and theoretically (see, e.g. Bunea (2008) and Bunea and Barbuc (2009)) that if the tuning parameters are chosen via cross-validation, as it will be described in detail in the next section, then the LASSO and E-Net will select a subset of predictors that will predict the response with high accuracy. However, this may not be the most parsimonious subset of predictors one can select: one may include, along with the predictors strongly associated with the response, some that are only weakly correlated, and may in fact be redundant. To understand how this can happen, consider the extreme example of having only two biomarkers, say $X_1$ and $X_2$, associated with the response, although many more biomarkers, say 100, are available for the analysis. Then, if $X_1$ and $X_2$ are not highly correlated, the two methods will, with high probability, select $X_1$ and $X_2$, but may also include in the model other irrelevant variables, say $X_3$, $X_4$, and $X_5$. This will likely not inflate the overall prediction of the response by $\hat{\beta}_1X_1 + \hat{\beta}_2X_2 + \hat{\beta}_3X_3 + \hat{\beta}_4X_4 + \hat{\beta}_5X_5$, as the LASSO estimated effect sizes, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, are typically very small. In fact, since some of these estimated coefficients may have negative signs, and if all predictors have positive values, it is possible that this prediction may be slightly better than predicting the response by using only the truly associated variables $X_1$ and $X_2$. Therefore choosing tuning parameters in order to obtain best prediction may not yield the smallest useful model. This phenomenon is especially pronounced if the sample size is small and the predictors not associated with the response have medium to high correlation. Moreover, if the truly associated variables, $X_1$ and $X_2$, are almost collinear then, by the construction of the LASSO estimator, only one of them will be selected, with high probability, if the sample size is small. The E-Net attempts to correct these drawbacks, but has only marginal success: if the truly associated predictors have medium correlation, then the E-Net will help include all of them in the model, but will still include extra variables, only fewer than the LASSO would.

Another important factor that may preclude the selection via LASSO or E-Net of predictors associated with the response is the effect size of these predictors, sometimes referred to as the strength of association. It is known practically (see, e.g. Shi et al. (2007)) and theoretically (see, e.g. Bunea (2008)) that in this phenomenon, shared by any model selection method, see e.g. Candès and Plan (2009).

In summary, if the LASSO and E-Net with parameters chosen via cross validation are applied to regression problems with very highly correlated predictors, and the sample size is small relative to $P$, then typically they will select a model that includes a few extra variables that are not associated with the response, and may also miss a few relevant ones. Note that the latter happens only when those predictors have very weak association with the response, or if some predictors are almost collinear, in which case one may find this feature of the LASSO to be positive, as possibly redundant predictors, that measure essentially the same thing, are eliminated. It is worth mentioning that, if the predictors are only mildly correlated, there exist theoretical choices of the tuning parameters of these two methods that would yield models containing only the predictors truly associated with the response, see e.g. Bunea (2008). However, there is very little guidance as to how to choose these parameters in practice: cross validation will not work, as explained above, and the research on this topic is still open, see e.g. Wasserman and Roeder (2009) for a discussion.

For all these reasons we opted for the procedure outlined in the following section, where we complement the LASSO step by a bootstrap resampling step (She, 2009). This will allow us to measure the frequency with which each predictor is chosen by the LASSO procedure: predictors chosen more frequently will be investigated further. We describe the specific details below.

Proposed method: bootstrap enhanced LASSO and E-Net

We describe our algorithm below.

Step 1. [Model fitting] For each value of $\lambda$, perform the following procedure to obtain the (bias-corrected) LASSO estimate $\hat{\beta}(\lambda)$.

1a) [LASSO] Fit the LASSO on all standardized predictors.

1b) [Bias correction] Fit a restricted OLS model on the selected predictors by the LASSO estimate from 1a), the corresponding estimate denoted by $\hat{\beta}(\lambda_0)$.

Step 2. [Parameter tuning] Apply K-fold cross-validation to find the appropriate value $\lambda_0$ (see Remark 3 below for a description). Record the corresponding selected predictors determined by the nonzero components of $\hat{\beta}(\lambda_0)$.

Step 3. [Bootstrap] Draw B bootstrap samples from the data. For each bootstrap sample repeat Steps 1–2.

Step 4. [Summarizing] Record the frequency with which each variable is selected by the LASSO in each bootstrap sample.

Remarks

(1) All predictors are required to be mean-centered and then scaled such that the design matrix (without intercept) has all column-means equal to 0 and column-variances equal to 1. This allows for a fair comparison of the relative predictor importance across all explanatory variables.

(2) The bias-correction step is based on the idea of ‘LARS-OLS hybrid’ (Efron et al., 2004). The LASSO estimators have shrunken values, relative to the value of the ordinary least squares (OLS) estimators, and are therefore biased. The refitting step, after dimension reduction, corrects this.

(3) To avoid overfitting we use K-fold cross-validation to tune the regularization parameter $\lambda$. To begin with, consider a fine grid $G$ of possible values of the tuning parameter. For any $\lambda \in G$ we then repeat the following procedure. For a given value of $K$, typically 5 or 10, we randomly split the data into $K$ subsets of approximately equal size. For each value of $k$, $1 \leq k \leq K$, run Step 1 on the data without the $k$-th subset. This step yields a sequence of estimates $\hat{\beta}^{(-k)}(\lambda)$. Then, for each $1 \leq k \leq K$, compute the prediction errors, $\text{Err}^{(k)}(\lambda) \equiv \sum_i (Y_i - \sum_{j=1}^p \hat{\beta}^{(-k)}(\lambda)X_{ij})^2$; note that these are evaluated on the $k$-th subset of the data that had been left out for this purpose. Finally, compute the summarized cross-validation error curve $\text{CV} - \text{Err}(\lambda) = \sum_{k=1}^K \text{Err}^{(-k)}(\lambda)$. The optimal parameter $\lambda_0 \in G$ is then given by the value at which $\text{CV} - \text{Err}(\lambda)$ achieves its minimum. Cross-validation is perhaps the most popular way for parameter tuning in the literature see, for instance, Tibshirani (1996); Zou and...
variables of interest based on previous literature prior to their inclusion variables and their interactions. This typically necessitates a selection of especially when allowing for the nonlinear transformations of the analysis of such data can involve a very large number of variables, physiological, and metabolic measures that need to be considered these efforts, resulting in large data sets of structural, functional, and cognitive measures that need to be considered relative to one another. This poses a methodological challenge, as the analysis of such data can involve a very large number of variables, especially when allowing for the nonlinear transformations of the variables and their interactions. This typically necessitates a selection of variables of interest based on previous literature prior to their inclusion in statistical models, at the expense of ignoring potentially significant variables. Penalized regression enables the analysis of a large number of explanatory variables even in cases where the number of explanatory variables exceeds the sample size. To demonstrate such issues, we focus here on multimodal neuroimaging, cognitive and clinical data obtained from a cohort of HIV-infected individuals.

**HIV-associated brain dysfunction**

HIV infection is often accompanied by neurocognitive dysfunction, typically involving impairment in attention, speed of information processing, motor abilities, executive function, and learning and memory. Neuroimaging studies have demonstrated volume reduction in various brain structures, most notably the basal ganglia, and cerebral white matter abnormalities (Stout et al., 1998 and McArthur et al., 2005). To demonstrate the utility of the penalized least squares method, we applied this set of analytical methods on a sample of 62 HIV-infected individuals who were recruited as part of a larger longitudinal NIH-sponsored study of HIV-associated brain dysfunction.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>44 ± 10.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education (years)</td>
<td>13 ± 2.04</td>
</tr>
<tr>
<td>% Male</td>
<td>68</td>
</tr>
<tr>
<td>% Caucasian</td>
<td>60</td>
</tr>
<tr>
<td>HIV infection duration (years)</td>
<td>12 ± 6.97</td>
</tr>
<tr>
<td>Nadir CD4 (cells/μl)</td>
<td>182 ± 171.03</td>
</tr>
<tr>
<td>Current CD4 (cells/μl)</td>
<td>431 ± 208.41</td>
</tr>
<tr>
<td>% with undetectable plasma HIV RNA</td>
<td>71</td>
</tr>
<tr>
<td>% with current hepatitis C</td>
<td>37</td>
</tr>
<tr>
<td>% with alcohol history</td>
<td>53</td>
</tr>
<tr>
<td>% with cocaine history</td>
<td>47</td>
</tr>
<tr>
<td>% with opiate history</td>
<td>16</td>
</tr>
</tbody>
</table>

Data example

Description of participants

Sixty-two HIV-infected (HIV+) participants were included in this study. Participants were excluded for history of 1) head injury with loss of consciousness > 10 min; 2) history of neurological conditions including dementia, seizure disorder, stroke, and opportunistic infection of the brain; 3) severe psychiatric illness that may impact brain function, e.g., schizophrenia; and 4) current active use of alcohol, cocaine, opiates, or illicit stimulants or sedatives. A significant proportion (39%) of participants had current HCV infection (HCV+).

Table 1 shows demographic and HIV clinical information, along with alcohol and substance use history.

**Neurocognitive assessment**

Five neurocognitive domains previously shown to be most affected in HIV infection were assessed including 1) attention/working memory (WAIS-III Digit Span and Letter–Number Sequencing), 2) speed of information processing (WAIS-III Digit Symbol and Symbol Search, Trail Making A), 3) psychomotor abilities (Grooved Pegboard), 4) executive function (Trail making B, Controlled Oral Word Association Test), and 5) learning and memory (Hopkins Verbal Learning Test — Revised, Brief Visuospatial Memory Test — Revised). All measures are widely used standardized neuropsychological tests with strong reliability and validity (Table 2). Individual test scores were converted to demographically corrected T-score using the most updated normative data for each test.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WAIS-III Symbol Search</td>
<td>48</td>
</tr>
<tr>
<td>WAIS-III Digit Symbol</td>
<td>45</td>
</tr>
<tr>
<td>WAIS-III Letter Number Sequencing</td>
<td>47</td>
</tr>
<tr>
<td>Grooved Pegboard worse hand</td>
<td>42</td>
</tr>
<tr>
<td>Trail Making A</td>
<td>48</td>
</tr>
<tr>
<td>Trail Making B</td>
<td>50</td>
</tr>
<tr>
<td>COWAT</td>
<td>48</td>
</tr>
<tr>
<td>Animal Fluency</td>
<td>52</td>
</tr>
<tr>
<td>HVLT-R total free recall</td>
<td>40</td>
</tr>
<tr>
<td>BVMT-R total free recall</td>
<td>41</td>
</tr>
<tr>
<td>HVLT-R delayed recall</td>
<td>39</td>
</tr>
<tr>
<td>BVMT-R delayed recall</td>
<td>44</td>
</tr>
</tbody>
</table>
Neuroimaging data acquisition

All neuroimaging was performed on one Siemens Tim Trio 3-Tesla MRI imager located at Brown University MRI Research Facility. Diffusion weighted images (DWI) were acquired axially for the whole brain with TE/TR = 103/10,060 ms, inplane resolution = 1.77 mm × 1.77 mm, slice thickness = 1.8 mm, b-value = 1000 s/mm². One DWI was acquired in each of 64 diffusion gradient directions, in addition to 10 images with no diffusion encoding (b0). Structural MRI images were acquired sagittally using T1 MPRAGE with TE/TR = 3.06/2250 ms, isotropic 0.86-mm voxel size.

Brain segmentation and volumetric measures

Automated brain segmentation was performed on the T1 MPRAGE image using tools from Freesurfer (Fischl et al., 2004). This process produces volumetric measures of the cortical grey matter, white matter, caudate, putamen, pallidum, thalamus, hippocampus, amygdala, and corpus callosum subregions. For analysis of white matter integrity with DWI, two mask images were extracted, containing 1) cerebral white matter, and 2) subregions of the corpus callosum. Segmented T1 images were transformed to DWI space by applying transformation matrices derived from registering average b0 and T1 images. Means and standard deviations are reported in Table 3.

Diffusion tensor estimation

All image registrations were performed using FSL FLIRT tool (Smith et al., 2004). The 10 b0 images were coregistered using rigid body registration to correct for movement. Registered images were then averaged. Each of the 64 DWIs was then registered to the average b0 image using affine registration to account for movement and eddy current artifacts. Each diffusion gradient direction was then rotated according to the corresponding affine transformation. Diffusion tensor estimation was performed using AFNI 3dDWItoDT tool (Cox, 1996), with the average b0 serving as the normalization image, yielding the 3 principal eigenvectors and associated eigenvalues characterizing the diffusion ellipsoid. FA, MD, AD, and RD were then computed using principal eigenvectors and associated eigenvalues characterizing the image using tools from Freesurfer (Fischl et al., 2004). This process

Brain segmentation and volumetric measures

Regions of interest (ROI)

One isotropic 1 cm³ was placed on a standard T1 MNI152 template in each of the frontal and parietal lobes bilaterally. The mask image containing these 4 ROIs was registered to individual brains in DWI space by applying 1) affine transformations between MNI and T1 images, and 2) rigid-body transformations between T1 and average b0 image. Non-white matter voxels were removed using the white matter segmenta-
Both methods were tuned by 5-fold cross-validation (CV). The CV running time for the E-Net method, which is the most computationally involved of the two methods, was approximately 12 min on an Intel Xeon 3.0 GHz server with 8 GB of RAM, which is acceptable. The mean squared errors (MSEs) of the two PLS methods are computed. We set aside 30% of the data as a separate test set, and used the remaining to fit the model and tune the regularization parameters. The median test error over the 13 neurocognitive assessments, computed on the test dataset, is 146.3 for the LASSO, and 105.1 for the E-Net. This improvement over the LASSO is due to the additional ridge penalty in the E-Net. Our experience is that even small \( \mu \) values in Eq. (2) can lead to important prediction improvements.

We also fitted a random forest model (Breiman, 2001) to the data. Using bagging (an averaging technique) together with random selection of candidate predictors for splitting, the random forest algorithms construct a combination of trees. Since the thus constructed random forest captures various nonlinear relations and interactions among predictors in an automatic fashion, by construction, it predicts the response well. We used the randomForest package in R for our data and obtained a test error of 120.8. It is important to note that the output of such packages also contains a measure of variable importance. However, as noted in Hastie et al. (2009, page 593), although the random forest is a powerful predictive tool, it is not ideal if one is also interested in understanding which predictors should be included in the model. Therefore, it is not appropriate to estimate the probability of including a predictor in the model via the variable importance measure provided by random forest packages, as this may be misleading. Indeed, for our data set, random forest analysis did not select Education and hcv_current among the top important predictors. This finding is contrary to an established body of literature that have clearly demonstrated the strong association between education and neurocognitive assessments. Moreover, there is a growing body of literature that support the notion that HCV can injure the brain, as the HCV can infect cells in the central nervous system (CNS). As demonstrated by magnetic resonance spectroscopy, HCV-infected individuals have elevated choline-to-creatine ratios in the basal ganglia and white matter suggesting neuronal loss (see, e.g., Letendre (2008)). Moreover, the E-Net, which is the method we recommend, has smaller test error (105.1) than the random forest (120.8), and therefore gives better prediction, and also can be reliably used for predictor selection.

As described earlier, for each response variable, we computed its bootstrap frequency as the estimate of that variable’s inclusion probability. We then created a heatmap to display the inclusion probability of each predictor for the 13 neurocognitive assessments. In Fig. 1, the predictors have been reordered based on the median of occurring frequencies for all neurocognitive assessments. The detailed important variables, with bootstrapping frequency >50%, are given in Tables 5 and 6.

**Discussion of results**

The variables selected by both BE-LASSO and BE-ENet are similar, indicating that the results are consistent across the two methods. However, we note that the BE-ENet yielded models with more DTI and volumetric biomarkers than the BE-LASSO. Since the predictors in our fitted model are correlated, and since the BE-ENet is especially tailored for such situations, we advocate its usage for further analyses of this type. The results of our analysis, via both methods, suggests that the models for most of the neurocognitive assessment are sparse – the optimal set of predictors is not large – although the actual number of most important variables differ across the different tests. This suggests redundancy in information across potential explanatory variables.

Among the demographic and clinical biomarkers, the most important predictors are co-infection with Hepatitis-C and education. This finding is consistent with results from other analyses (Gongvatana et al., 2011). After the biomarker selection procedure, we fitted regression models that contain as explanatory variables only those that were selected. As expected, performance level increases with education and Hepatitis-C co-infection is linked to lower performance in the neurocognitive function.

The bootstrap-enhanced selection methods indicate that explanatory variables derived from neuroimaging data are good predictors for general neurocognitive assessments. However, the degree of importance of neuroimaging biomarkers varies across the different assessments. We note that fractional anisotropy at the corpus callosum and internal capsule and mean diffusivity at the corpus callosum appeared are important predictors for psychomotor skills and executive function, but not for learning and memory and attention/working memory domains. It is clear that while some neuroimaging variables appeared as linear, other explanatory variables enter as quadratic effects. This result underlines the potential impact of PLS in biomarker selection. Due to the “Large P small N” constraints, it would not have been possible to fit models containing interactions and non-linear effects using standard variable selection techniques. By considering such more complex

---

**Table 5** Predictors with bootstrapping frequency >50% based on the bootstrapped enhanced LASSO.

<table>
<thead>
<tr>
<th>Responses</th>
<th>Important predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>HVLT_sum_T</td>
<td>kmsk_cocopi, Education</td>
</tr>
<tr>
<td>HVLT_delay_T</td>
<td>Education, hcv_current, kmsk_cocopi, kmsk_alc</td>
</tr>
<tr>
<td>BVMT_sum_T</td>
<td>kmsk_cocopi, Education, fa_cc1-squared</td>
</tr>
<tr>
<td>BVMT_delay_T</td>
<td>Education, kmsk_cocopi, fa_cc1-squared</td>
</tr>
<tr>
<td>WAIS_LNS_T</td>
<td>Education</td>
</tr>
<tr>
<td>WAIS_DigSym_T</td>
<td>kmsk_cocopi, kmsk_alc, fa_cc5-squared, md_cc234 × cortex</td>
</tr>
<tr>
<td>WAIS_SymSrch_T</td>
<td>kmsk_cocopi, md_cc1-squared, hcv_current</td>
</tr>
<tr>
<td>GPeg_dom_T</td>
<td>fa_ic2-squared, kmsk_alc, hcv_current, md_cc234-squared, fa_cc1</td>
</tr>
<tr>
<td>GPeg_nondom_T</td>
<td>fa_ic2-squared</td>
</tr>
<tr>
<td>Trail_A_T</td>
<td>kmsk_cocopi, md_cc234-squared, kmsk_alc</td>
</tr>
<tr>
<td>Trail_B_T</td>
<td>md_cc1-squared</td>
</tr>
<tr>
<td>COWAT_T</td>
<td>hcv_current</td>
</tr>
<tr>
<td>Animal_T</td>
<td>Education, md_ic1-squared</td>
</tr>
</tbody>
</table>
models, our analysis revealed that larger values of fractional anisotropy and lower values of mean diffusivity at the corpus callosum and the internal capsule are linked to higher performance in the neurocognitive assessment. These results confirm the importance of keeping the integrity of white matter in these areas as crucial to maintaining a high level of functioning in HIV+ patients.

It is also worth pointing out that some morphometric variables play a role in predicting certain neurocognitive functions in the BE-E-Net selection. They appear as interactions (MD at the posterior corpus callosum × genu volume; and MD at the posterior corpus callosum × putamen; MD at the posterior corpus callosum × cortex; and MD at the posterior corpus callosum × pallidum). Moreover, it enters non-linearly (quadratic effect of the corpus callosum volume) for predicting the posterior corpus callosum × pallidum). Moreover, it enters non-linearly (quadratic effect of the corpus callosum volume) for predicting the posterior corpus callosum × pallidum).

To conclude, we provide some new directions on statistical research in this area. Very often, researchers might be interested in analyzing several dependent (response) variables simultaneously rather than conducting separate individual regression models. For example, one might be interested in studying learning and memory as a single domain rather than analyzing each of the neurocognitive tests separately.

To address this desiderata, we will develop dimension reduction methods for multivariate response regression models in our future research. By refining the method in Bunea et al. (2011), we will develop procedures that select the best biomarker predictors of a collection of, possibly correlated, outcome variables.

**Acknowledgments**

This paper was supported in part by grants from the National Science Foundation (Bunea, DMS 10007444) and (Ombao, DMS 0806106) and the National Health Institutes (Cohen, RO1 MH074368).

**Appendix A. Supplementary data**

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2010.12.028.

**References**


Bunea, F., 2008. Honest variable selection in linear and logistic models via \( \ell_1 \) and \( \ell_1/\ell_2 \) penalization. Electronic Journal of Statistics 2, 1153–1194.


**Table 6** Predictors with bootstrapping frequency >50% based on the bootstrapped-enhanced E-Net.

<table>
<thead>
<tr>
<th>Responses</th>
<th>Important predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>HVL_T</td>
<td>kmsk_cocopi, Education, md_cc234 × cc_genu, kmsk_alc</td>
</tr>
<tr>
<td>HTLV_delay_T</td>
<td>Education, hcv_current, kmsk_cocopi, kmsk_alc</td>
</tr>
<tr>
<td>BVM_T</td>
<td>Education, kmsk_cocopi, fa_cc1-squared, hcv_current</td>
</tr>
<tr>
<td>BVMT_delay_T</td>
<td>Education, kmsk_cocopi, fa_cc1-squared, hcv_current, md_cc234 × putamen</td>
</tr>
<tr>
<td>WAIS_LNS_T</td>
<td>Education, fa_cc234-squared</td>
</tr>
<tr>
<td>WAIS_Sym5Sym_T</td>
<td>kmsk_cocopi, kmsk_alc, md_cc5-squared, Education, fa_cc5-squared</td>
</tr>
<tr>
<td>WAIS_Sym5Sym_T</td>
<td>fa_ic2-squared, kmsk_cocopi</td>
</tr>
<tr>
<td>GPe_memom_T</td>
<td>hcv_current, fa_ic2-squared, kmsk_alc, md_cc234-squared, fa_cc1</td>
</tr>
<tr>
<td>GPe_monomom_T</td>
<td>fa_ic2-squared, kmsk_cocopi</td>
</tr>
<tr>
<td>Trail1_A_T</td>
<td>kmsk_cocopi, md_cc234-squared, kmsk_alc, hcv_current, md_cc234 × cortex, md_cc234 × pallidum</td>
</tr>
<tr>
<td>Trail2_B_T</td>
<td>md_cc1-squared, fa_ic2-squared</td>
</tr>
<tr>
<td>COWAT_T</td>
<td>hcv_current</td>
</tr>
<tr>
<td>Animal_T</td>
<td>Education, fa_ic2-squared, cc_body-squared</td>
</tr>
</tbody>
</table>

The reduced set will be typically much smaller than the sample size, one can then conduct focused analyses and perform tests of hypotheses on the selected reduced model. Moreover, PLS can guide the researcher on the conduct of future clinical studies by collecting data only using the most useful predictors and on formulating more precise tests of hypotheses.

To conclude, we provide some new directions on statistical research in this area. Very often, researchers might be interested in analyzing several dependent (response) variables simultaneously rather than conducting separate individual regression models. For example, one might be interested in studying learning and memory as a single domain rather than analyzing each of the neurocognitive tests separately.

As an overall remark, we note that this general statistical approach is best used as a preliminary step in analyzing a data set having a large number of predictors. The methods discussed here are especially useful when there is no compelling a priori basis for selecting a small set of these variable to be utilized for further focused analyses, in which one could have employed more traditional statistical methods. Then, the bootstrap-enhanced PLS methods can be used reliable to select a reduced set of explanatory variables with high predictive power. Since

---

**Note:** Table 6 has been formatted to fit within the given text boundaries, and some abbreviations have been explained to provide a clearer understanding. The table represents predictors with bootstrapping frequency greater than 50% based on the bootstrapped-enhanced E-Net.


