Discovering Predictors of Mental Health Service Utilization with k-support Regularized Logistic Regression

Hakim Sidahmed, Elena Prokofyeva, Matthew Blaschko

Discovering Predictors of Mental Health Service Utilization with *k*-support Regularized Logistic Regression

Hakim Sidahmed[a,b], Elena Prokofyeva[c,d,e], Matthew B. Blaschko[f,a]

[a] *CentraleSupélec, Grande Voie des Vignes, 92295 Châtenay-Malabry, France*
[b] *Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
[c] *Inserm, U1018, Centre for Research in Epidemiology & Population Health (CESP), Epidemiology of Occupational and Social Determinants of Health, Villejuif, France*
[d] *University of Versailles Saint-Quentin, UMRS 1018, Villejuif, France*
[e] *Northern State Medical University, Troitsky av. 51, Arkhangelsk, Russia*
[f] *Inria Saclay, Campus de l'École Polytechnique, 91120 Palaiseau, France*

Abstract

Many epidemiological studies are undertaken with a use of large epidemiological databases, which involves the simultaneous evaluation of a large number of variables. Epidemiologists face a number of problems when dealing with large data sets: multicolinearity (when variables are correlated to each other), confounding factors (when risk factor is correlated with both exposure and outcome variable), and interactions (when the direction or magnitude of an association between two variables differs due to the effect of a third variable). Correct variable selection helps to address these issues and helps to obtain unbiased results. Selection of relevant variables is a complicated and a time consuming task. Flawed variable selection methods still prevail in the scientific literature; there is a need to demonstrate the usability of new algorithms using real data. In this paper we propose to use a novel machine learning method, *k*-support regularized logistic regression, for discovering predictors of mental health service utilization in the National Epidemiologic Survey for Alcohol and Related Conditions (NESARC). We show that *k*-support regularized logistic regression yields better prediction accuracy than ℓ_1 or ℓ_2 regularized logistic regression as well as several baseline methods on this task, and we qualitatively evaluate the top weighted variates. The selected variables are supported by related epidemiological research, and give important cues for public policy.

Keywords:
variable selection, discarding variables, k-support regularized logistic regression, epidemiological data

1. Introduction

Statistical analysis, such as logistic regression, is regularly applied to problems in epidemiology, and other problems where associations must be tested in the context of confounding factors. Often, such models are applied to data consisting of few variables hand-selected from a large, unstructured database. The validity of such studies may therefore depend on scientists carefully identifying all possible confounding factors. Furthermore, important related factors to a condition of interest may be excluded from analysis erroneously. It is therefore of interest to consider strategies that may include a large amount of available data, and select relevant variables automatically. This paper proposes a novel method for this kind of analysis, *k*-support regularized logistic regression, and we demonstrate its applicability on the challenging task of identifying predictors of mental health service utilization for people with substance abuse.

One problem that arises when using statistical learning on large databases is that spurious correlation with the signal is more likely with a large number of variables in a given analysis. A strategy to counteract this problem is to make use of a sparsity regularizer, such as the lasso, or ℓ_1 penalty [1, 2, 3]. While the lasso is beneficial in setting many coefficients to zero, and therefore reducing the risk that the model rely on irrelevant variables, a problem arises when there are correlated signals [4]. If two or more variables are discriminative with respect to the target of interest, the lasso has a tendency to select only one. This can be detremental in epidemiological studies for several reasons: (i) if the selected variable is a confounding factor, the causal factor will be suppressed, and (ii) if the selected variable is

a causal factor, confounding factors may be suppressed and the effect of the causal factor will be over-estimated. The solution to this problem is to use a correlated sparsity regularizer.

Correlated sparsity is a property achieved by a subset of structured sparsity regularizers [5, 6, 7]. This family of regularizers incorporates those that favor sparse coefficient vectors, but those that have specific patterns of sparsity. In this work, we consider specifically the k-support norm [8], which achieves a lower degree of sparsity than the lasso, but favors configurations of the model in which variables with correlated signals are allowed to be activated simulatneously. We consider the novel application of k-support norm regularization to logistic regression. The k-support norm is related mathematically to the elastic net [5], but has a more principled derivation as the convexification of a norm that applies an ℓ_2 penalty to a subset of k-selected variables.

1.1. Current Practices for Variable Selection in Epidemiological Research

The main aim of epidemiological studies is to identify and to quantify risk factors of diseases. Modern epidemiological research is often based on a large data sets such as the National Epidemiologic Survey for Alcohol and Related Conditions (NESARC) [9, 10, 11], the National Health and Nutritional Examination Survey (NHANES) [12]; the CONSTANCES cohort [13], and the Whitehall cohort study [14, 15], etc. Problems that scientists face when dealing with such data sets include: multicolinearity (when variables are correlated to each other), confounding factors (when a risk factor is correlated with both exposure and outcome variable), interactions (when the direction or magnitude of an association between two variables differs due to the effect of a third variable), the sample size (the study can be too large with meaningful associations being declared or too small to detect important associations), and the number of factors being studied (the higher the number of factors the higher is the probability to find interactions due to chance alone) [16]. Correct variable selection helps to overcome these problems and to obtain unbiased results. In particular, it is used for confounder control in etiologic research [17] and for unbiased estimation of probabilities in prediction research [18].

It is often difficult to determine which variables are important and should be included in the analysis and which should be disregarded. Selection of relevant variables depends on number of factors: the research question of the study, study design, and the sample size [17]. Prior scientific knowledge was earlier used as the main criteria to identify covariates that should be included in the analysis, but it is not always available for all research questions asked [18, 19, 17]. The first step of any epidemiological study planning is the review of the prior scientific knowledge, in particular, other epidemiological studies that were performed on a similar subject, but also those for related pathologies. This is done to identify different factors that can influence the outcome of the study along with the risk factor of interest. Although this strategy of identifying covariates is an obligatory step for planning for any epidemiological study, it may not be suitable for all studies, for example, for those with a research question that has not previously been studied.

A number of statistical techniques were developed to enable the selection of covariates based on relations of the data under study: change in the effect estimate, stepwise selection, shrinkage and penalized regression, and other techniques. Recent research on contemporary epidemiological analysis showed that effect estimate change and stepwise selection techniques represent the majority of the all methods used for variable selection [17]. Despite the fact that they allow the automatic selection of important predictor variables for inclusion in the model through the forward, backward or stepwise selection process, they have number of limitations. Firstly, these methods assume independence of all variables studied; secondly, there are many different ways to decide which model suits this data the best, and thirdly these methods are time consuming and often require greater input from the researcher [16]. Indeed, for experiments on the scale considered here, forward and backward selection is computationally infeasible (see Section 3.3.1). In the light of the dominance of variable selection methods that are criticized by leading epidemiologists as flawed, there is a need to demonstrate the usability of new algorithms in real data instead of simulation studies [17].

Alcohol abuse and dependence remain one of the main public health problems. It is closely linked to car crashes [20], domestic violence [21], fetal alcohol syndrome [22], neurophysiological impairment [23], and psychiatric comorbidities [24]. Despite the obvious treatment benefits such as reduction of the severity and harm associated with alcohol use disorder, only 7% of individuals with lifetime alcohol use disorder receive treatment [25]. Therefore it is essential to study predictors of umet need for treatment for alcohol use disorder.

The aim of this article is to demonstrate the application of k-support regularized logistic regression to discover predictors of treatment or the perceived unmet need for treatment among respondents did not receive treatment for alcohol use disorder.

2. *k*-support Regularized Logistic Regression

In the sequel we will use the notation $\beta \in \mathbb{R}^d$ to specify a coefficient vector of a linear discriminant function, $\lambda \geq 0$ to specify a scalar regularization parameter, $\Omega : \mathbb{R}^d \mapsto \mathbb{R}_+$ a regularizer for a linear function class. We assume a training set of data $\{(x_i, y_i)\}_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ which are encoded in a matrix $X \in \mathbb{R}^{d \times n}$ and a matrix $Y \in \{-1, +1\}^{c \times n}$ where c is the number of classes. We assume i.i.d. sampling of this training set. The ith column of X is equal to x_i and the ith column of Y encodes the class of y_i by a binary vector with exactly one positive entry in the place of the corresponding class. For the special case of binary classification, i.e. $c = 2$, we will use the vector $y \in \{-1, +1\}^n$ to encode the class labels.

2.1. Regularized Logistic Regression

In the case of binary classification, logistic regression can be written as a regularized risk minimization with logistic loss. In this case, the objective to be minimized is

$$\hat{\beta} = \arg\min_{\beta} \lambda\Omega(\beta) + f_{\log}(\beta, X, y) \tag{1}$$

with

$$f_{\log}(\beta, X, y) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right) \tag{2}$$

where λ is a scalar parameter controlling the degree of regularization by the regularizer Ω and $\langle \cdot, \cdot \rangle$ is the canonical inner product in $\mathbb{R}^d$. The minimization of logistic loss is equivalent to classicial logistic regression with two classes in the case that it is unregularized [26].

In the multiclass setting, we specify one coefficient vector per class. With $|\mathcal{Y}| = c$ classes, we write our joint probability

$$p(\beta)p(Y|X;\beta) = \left(\frac{1}{Z_{\lambda\Omega}} e^{-\lambda\Omega(\beta)}\right) \cdot \prod_{i=1}^{n} \frac{e^{\langle \beta_{y_i}, x_i \rangle}}{\sum_{l=1}^{c} e^{\langle \beta_l, x_i \rangle}} \tag{3}$$

where $Z_{\lambda\Omega}$ is a constant normalizing $e^{-\lambda\Omega(\beta)}$ as a distribution. Taking negative logarithms we arrive at

$$-\log p(\beta)p(Y|X;\beta) = \log Z_{\lambda\Omega} + \lambda\Omega(\beta) + \sum_{i=1}^{n} \left(-\langle \beta_{y_i}, x_i \rangle + \log \sum_{l=1}^{c} e^{\langle \beta_l, x_i \rangle}\right) \tag{4}$$

We ignore the term $\log Z_{\lambda\Omega}$ as there is no dependence on β, and we arrive at the loss function

$$f_{\log}(\beta, X, Y) = \sum_{i=1}^{n} \left(-\langle \beta_{y_i}, x_i \rangle + \log \sum_{l=1}^{c} e^{\langle \beta_l, x_i \rangle}\right) \tag{5}$$

where we have abused the notation $f_{\log}$. When the matrix Y is used as an argument to $f_{\log}$, we signify that we are using the definition in Equation (5), while the use of the vector y indicates that we are in the binary setting and using Equation (2)

In the multi-class setting here we have written our joint probability slightly differently than the classical presentation of logistic regression [26, Equation (4.18)]. In that presentation, one of the classes had its class energy arbitrarily constrained to 1 as the problem is overparametrized when performing maximum likelihood learning on the probability simplex:

$$p(y = i|x;\beta) = \frac{e^{\langle \beta_i, x \rangle}}{1 + \sum_{l=1}^{c-1} e^{\langle \beta_l, x \rangle}} \quad \forall i \neq c \tag{6}$$

$$p(y = c|x;\beta) = \frac{1}{1 + \sum_{l=1}^{c-1} e^{\langle \beta_l, x \rangle}} \tag{7}$$

In the case of maximum *a posteriori* (MAP) estimation, we would like our structured-sparsity prior over functions to be symmetric with respect to the classes so that each will be regularized equally:

$$p(y = i|x;\beta) = \frac{e^{\langle \beta_i, x\rangle}}{\sum_{l=1}^{c} e^{\langle \beta_l, x\rangle}} \quad \forall i \tag{8}$$

This formulation is ill posed in the maximum likelihood setting, but the use of a regularizer yields a well-posed symmetric problem for MAP estimation.

2.2. k-support Regularization

We propose the use of a correlated sparsity regularizer, the k-support norm [8]. The k-support norm is the gauge function associated with the convex set

$$\text{conv}\{\beta \mid ||\beta||_0 \leq k, ||\beta||_2 \leq 1\}. \tag{9}$$

This set is the convex hull of all $\binom{d}{k}$ axis aligned unit ℓ_2 balls of dimensionality k. k is a parameter that will be chosen by model selection or prior information.

The k-support norm can be computed as

$$||\beta||_k^{sp} = \left(\sum_{i=1}^{k-r-1} (|\beta|_i^{\downarrow})^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{d} |\beta|_i^{\downarrow} \right)^2 \right)^{\frac{1}{2}} \tag{10}$$

where $|\beta|_i^{\downarrow}$ is the ith largest element of the vector and r is the unique integer in $\{0, \ldots, k-1\}$ satisfying

$$|\beta|_{k-r-1}^{\downarrow} > \frac{1}{r+1} \sum_{i=k-r}^{d} |\beta|_i^{\downarrow} \geq |\beta|_{k-r}^{\downarrow}. \tag{11}$$

One may observe that the k-support norm computes a weighted sum of the ℓ_2 norm on the largest components of the vector (leftmost sum in Equation (10)) and the ℓ_1 norm on the smallest components of the vector (rightmost sum in Equation (10)). The integer r is dependent on k, but can increase the number of non-zero coefficients to be larger than k depending on where on the k-support ball the vector β is located.

2.3. Relationship to Other Methods

2.3.1. Relationship to ℓ_1 and ℓ_2 Regularization

In the case that $k = 1$ the k-support norm is exactly equivalent to the ℓ_1 norm. In the case that $k = d$, where $\beta \in \mathbb{R}^d$, the k-support norm is equivalent to the ℓ_2 norm.

We note that for an objective

$$\min_{\beta} \lambda ||\beta||_k^{sp} + f(\beta, X, y) \tag{12}$$

with some loss function $f(\cdot, \cdot, \cdot)$, when $k = d$, this is equivalent to

$$\min_{\beta} \lambda ||\beta||_2 + f(\beta, X, y) \tag{13}$$

rather than the familiar squared ℓ_2 regularizer. However, for any λ there exists some $\tilde{\lambda}$ such that

$$\arg\min_{\beta} \lambda ||\beta||_2 + f(\beta, X, y) = \arg\min_{\beta} \tilde{\lambda} ||\beta||_2^2 + f(\beta, X, y). \tag{14}$$

This can be easily seen by noting that the objectives are the Lagrangians of constrained minimization problems that minimize f subject to the equivalent constraints $||\beta||_2 \leq B$ and $||\beta||_2^2 \leq B^2$, respectively, for some $B \in \mathbb{R}_+$.

We note that ℓ_1 and ℓ_2 reglarized logistic regression are precisely those methods considered in [27] and that our proposed method therefore specializes to these algorithms when $k = 1$ or $k = d$, respectively.

2.3.2. Relationship to Group Lasso

A prominent family of structured sparsity regularizers are those based on the group lasso with overlaps [28, 7]. These methods generalize the ℓ_1 regularizer to an ℓ_1 norm of a vector of ℓ_2 norms computed on potentially overlapping groups of variables. As such, one may view the k-support norm regularization as a special case of the group lasso in which groups are defined over all overlapping $\binom{d}{k}$ groups of size k. The group lasso with non-overlapping groups for logistic regression has been proposed in [29]. The method proposed here can be viewed as a special case of group lasso with overlapping groups and logistic loss. The main benefits of this special case are that the k-support norm has a beneficial interpretation in terms of correlated sparsity [8] and can be computed in $O(d \log d)$ time, while optimization strategies for arbitrarily defined groups will scale with the exponential number of groups.

2.4. Optimization

Our method for optimizing k-support regularized logistic regression essentiallly follows the strategy suggested for k-support regularized least squares in [8]. The strategy is based on Nesterov's accelerated method for the linear combination of smooth and non-smooth functions, where $f_{\log}$ is the smooth component and $\|\cdot\|_k^{sp}$ non-smooth. Accelerated methods were first proposed for smooth functions [30, 31] and later applied to non-smooth [32] and composite functions [33, 34]. An overview with a unified analysis is presented in [35].

The optimization strategy requires as inputs a function that gives oracle access to the smooth component, a function that gives the gradient of the smooth function, and a Lipschitz constant of the gradient:

$$f_{\log}(\beta, X, y) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \langle \beta, x_i \rangle}\right) \tag{15}$$

$$\frac{\partial f_{\log}}{\partial \beta} = -\sum_{i=1}^{n} \frac{e^{-y_i \langle \beta, x_i \rangle}}{1 + e^{-y_i \langle \beta, x_i \rangle}} y_i x_i \tag{16}$$

$$L_{\log} = \frac{\gamma}{4} \tag{17}$$

where the Lipschitz constant has a factor $\frac{1}{4}$ from the Lipschitz constant of the sigmoid in $\frac{\partial f_{\log}}{\partial \beta}$, and γ is the largest eigenvalue of XX^T. To minimize $J(\beta) = \lambda \|\beta\|_k^{sp} + f_{\log}(\beta, X, y)$, convergence of accelerated methods is such that at the mth iteration

$$J\left(\beta^{(m)}\right) - J\left(\beta^*\right) \leq O\left(\frac{1}{t_{\min}(m+1)^2}\right) \tag{18}$$

where we have suppressed dimensionality dependent terms in the notation $O(1)$, $\beta^{(m)}$ is the estimate of the coefficient vector after m steps of the optimization algorithm, β^* is the optimal coefficient vector

$$\beta^* := \arg\min_{\beta} J(\beta) \tag{19}$$

and

$$t_{\min} = \min\left\{1, \frac{C}{L_{\log}}\right\} \tag{20}$$

for a method dependent constant $0 < C \leq 1$. The convergence therefore scales proportionately with the Lipschitz constant of the gradient of the loss $L_{\log}$. The Lipschitz constant of the gradient of the loss is largely dependent on the largest eigenvalue of XX^T, and so the convergence matches the usual dependency on the condition number of the system matrix for second order optimization methods [36]. The optimization for the multi-class variant follows analogously. An open source implementation of k-support regularized logistic regression is available for download from `https://github.com/blaschko/ksupport`.

3. Results

3.1. The Epidemiological Data Set

The NESARC is a nationally representative study of the US population that was conducted among non-institutionalized adults ($\geq$ 18 years of age) residing in households and group quarters designed and conducted by the National Institute

Table 1: NESARC variables referred to in the article. The first column specifies the name of the variable used in the text for brevity.

Variable	NESARC Code	Full Name
A	S2CQ4A	EVER THOUGHT SHOULD SEEK HELP WITH DRINKING BUT DIDN'T GO
B	ALCABDEP12DX	ALCOHOL ABUSE/DEPENDENCE IN LAST 12 MONTHS
C	ALCABDEPP12DX	ALCOHOL ABUSE/DEPENDENCE PRIOR TO THE LAST 12 MONTHS

on Alcohol Abuse and Alcoholism (NIAAA). Wave 1 NESARC data on which this study is based were collected during 2001–2002 through computer-assisted personal interviews (CAPI) in face-to-face household settings. The sample included 43,093 respondents ages 18 and older, representing the civilian, non-institutionalized adult population in the United States, including all 50 States and the District of Columbia. Military personnel living off-base and residents in non-institutionalized group quarters housing, such as boarding houses, shelters, and dormitories, were also included. All participants were interviewed at home by experienced lay interviewers who received extensive training and supervision. All procedures, including informed consent, received full ethical review and approval from the U.S. Census Bureau and U.S. Office of Management and Budget.

We have applied k-support regularized logistic regression to the problem of predicting whether a given NESARC participant had an unmet need for alcohol addiction treatment. This condition was identified based on the subject's response to the question S2CQ4A in the NESARC survey. In the sequel, we will refer to this response as variable A. This convention (and that for other NESARC variables referred to in the text) is specified in Table 1.

3.2. Data Processing

It is primarily of interest to determine potential causes for not receiving treatment for alcoholism for those subjects that have ever had any alcohol dependence. This information is encoded in two variables in the NESARC database, one which encodes whether the respondent had any alcohol abuse or dependence in the last 12 months, or whether they had any alcohol or dependence prior to the last 12 months. We therefore filter respondents to only those that answered yes to variable B or variable C (cf. Table 1). Out of the original 43093 respondents, 4068 had suffered from alcoholism at some point prior to the administration of the survey.

Wave 1 of NESARC contains 2991 variables in total. By using a standard epidemiological approach based on prior scientific knowledge of the factors that may have an influence on help seeking for alcohol use disorder we were able to reduce our data set to 112 potentially relevant variables. This process consisted of performing a literature review of relevant variables included in previous studies and then including *all* variables in the relevant NESARC section, giving a substantially larger number of variables to the statistical learning process than previous studies. Finally, we removed variables corresponding to questions only relevant to people who do not receive treatment for alcohol dependence or abuse. Inclusion of these variables in the regression would effectively give the learning algorithm access to the label for a large number of respondents and render the output invalid. This pre-selection procedure is therefore essential to maintain the validity of the result, and has the added effect of making use of epidemiological prior knowledge. We then whiten the resulting variables to have zero mean and unit variance. In this way, the variable selection procedure encoded in the k-support regularization will not be biased towards any particular variables.

As is the case with surveys with a very large sample size and many questions, the data are incomplete. Many questions give the respondent the option to answer 'Unknown,' meaning that the respondent doesn't know the answer to the question (e.g. doesn't remember) and/or 'Blank,' when the respondent did not answer. These two cases were treated as distinct by encoding a binary variable indicating whether either event occurred.

Finally, we encode the categorical variables as follows: If a feature is binary (Yes(1)-No(0) answer), it is replaced by -1 (No) or +1 (Yes). If a feature is multivariate, it is expanded into several columns (corresponding to its number of classes) of 0's and 1's. If a feature takes values a, b, and c for instance, then it is expanded into 3 columns: a is replaced by vector $[1, 0, 0]^T$, b by $[0, 1, 0]^T$ and c by $[0, 0, 1]^T$. This encoding scheme results in a statistical learning problem with 314 dimensions.

The source code for data processing is available from `https://github.com/hakimsd9/predUsingksup/`.

3.3. Quantitative Experiments

We performed multiple experiments with random splits of the data into training (2,000 respondents), validation (1,000 respondents) and test (1,068 respondents) sets. For each experiment, there are two intrinsic parameters to choose: the regularization parameter λ and the value of k in the k-support norm. These parameters are selected using model selection, and the best performing configuration (as measured by area under the curve–AUC) on the validation set is used to train a discriminant function that is evaluated only once on the test set. We have used AUC as a selection criterion due to its better ability to distinguish between two methods when the accuracy is tied. Furthermore, recent theoretical and empirical results indicate that AUC optimization leads to good classification performance with tight regret bounds, further validating this choice of selection criterion [37].

Each experiment we run to apply logistic regression to the data set has to go through several iterations in order to choose the parameter k and the regularization parameter λ giving the best area under curve (AUC). We use a grid search over these variables to select the best configuration. The parameter k is allowed to take values in the set $\{2, 4, 8, 16, 32, 64, 128, 256\}$, while λ is in the set $\{10^{-14}, \ldots, 10^{14}\}$ in powers of 10. Both parameter ranges are therefore sampled logarithmically.

3.3.1. Comparison to baseline methods

We have first compared k-support regularized logistic regression to two baselines: ridge regression, and the selection of the single best variable. The ridge regression regularization parameter was selected from the same range of values of λ as for the logistic regression experiments. The single best variable was selected by finding the variable (or its negation) that gave the best AUC on the validation set. ROC curves for k-support regularized logistic regression, ridge regression, and the single best variable are shown in Figure 1.

We have performed statistical significance testing to formally validate the difference in performance between methods. First, performed 50 different trials to estimate the accuracy for each method. A Pearson χ^2 goodness of fit test was applied to evaluate whether the performances across trials were Gaussian distributed. This test showed that the data were not normally distributed ($p \approx 10^{-3}$). Consequently, we have applied non-parametric significance tests here and in subsequent sections. We performed pairwise Wilcoxon signed rank tests, which showed significance for all pairs of variables (k-support vs. ridge regression $p \approx 2 \times 10^{-4}$, k-support vs. single best variable $p \approx 5 \times 10^{-7}$). In addition to pairwise significance testing, we have also applied a Kruskal-Wallis test on all three methods simultaneously, showing a significant difference with $p \approx 3 \times 10^{-15}$..

Forward and backward selection methods, although interesting potential baseline methods, are not feasible to apply in the context of this problem. This is because a model must be trained $\frac{d(d+1)}{2} - 1$ times, where d is the number of dimensions to be considered. As our problem contains 314 dimensions (see Section 3.2), this would require training a model 49,454 times, which is not possible in a reasonable amount of time. By contrast, sparsity regularization methods need only be trained once, resulting in feasible computation even for problems with a large number of variables. Recent theoretical results indicate that the conditions for forward selection to provide good variable selection are very close to those for ℓ_1 regularization [38]. Consequently, we provide a comparison to ℓ_1 regularized logistic regression in the next section, which functions as a computationally tractable surrogate for forward or backward selection methods.

3.3.2. Comparison to ℓ_1 and ℓ_2 regularization

In order to compare the quality of the predictions obtained with the k-support regularizer to ℓ_1 and ℓ_2 regularizers, we perform Wilcoxon signed-rank tests as well as a Kruskal-Wallis test as in Section 3.3.1. The pairwise results are presented in Figure 2. The Kruskal-Wallis test showed significance with $p \approx 2 \times 10^{-4}$. To illustrate the performance of the k-support regularized classifier, we plot its receiver operating characteristic curve (ROC curve). ROC curves for k-support regularization, ℓ_1 regularization, and ℓ_2 regularization are shown in Figure 3.

The variables with highest weight in the model learned by k-support regularization are summarized in Table 2. The obtained list of variables contained covariates that fell into five major groups listed in the order of relevance: 1) lifetime drug use disorder (covariates describing presents or absence of abuse/dependence on opioids, sedatives, tranquilizers, amphetamine, cocaine, inhalants/solvents, hallucinogens, cannabis, and heroin), 2) lifetime mental health problems (anxiety disorders: panic disorder with or without agoraphobia, agoraphobia with no history of panic disorder, generalized anxiety; mood disorders: manic disorder, dysthymia, hypomanic disorder, social phobia), and lifetime

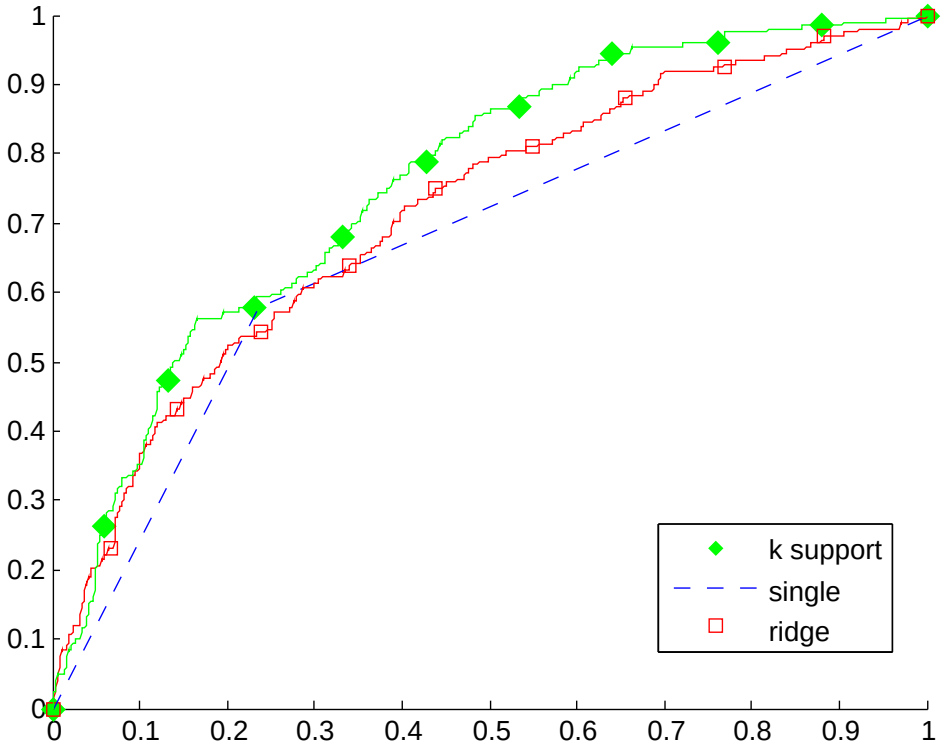

Figure 1: ROC curves for k-support norm regularized logistic regression, ridge regression, and the curve obtained by selecting the best performing single variable. The ROC curve for k-support norm regularized logistic regression is substantially higher than the other two curves. The curve for the single best variable is piecewise linear due to the selection of a categorical variable.

$$\begin{pmatrix} & k\text{-sup} & \ell_1 & \ell_2 \\ k\text{-sup} & 1 & 0.2059 & 1.9209e-06 \\ \ell_1 & 0.2059 & 1 & 1.3536e-04 \\ \ell_2 & 1.9209e-06 & 1.3536e-04 & 1 \end{pmatrix}$$

$$\begin{pmatrix} & k\text{-sup} & \ell_1 & \ell_2 \\ \text{mean} & 0.7315 & 0.7265 & 0.6922 \\ \text{std} & 0.0255 & 0.0328 & 0.0098 \end{pmatrix}$$

Figure 2: Wilcoxon signed rank test performed pairwise between ℓ_1, ℓ_2 and k-support regularized variants of logistic regression. Both k-support and ℓ_1 regularization perform significantly better than ℓ_2 regularization. k-support regularization has a higher AUC than ℓ_1 regularization on average.

Table 2: The highest weighted variables selected from the NESARC database by k-support regularized logistic regression.

	NESARC variable name
1	EVER SOUGHT HELP BECAUSE OF DRINKING
2	ANY FULL SISTERS EVER ALCOHOLICS OR PROBLEM DRINKERS
3	SEDATIVE ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
4	TRANQUILIZER ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
5	OPIOID ABUSE DEPENDENCE PRIOR TO LAST THE 12 MONTHS
6	HEROIN ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
7	SEDATIVE ABUSE DEPENDENCE IN LAST 12 MONTHS
8	ADOPTIVE FATHER EVER AN ALCOHOLIC OR PROBLEM DRINKER
9	INHALANT SOLVENT ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
10	PANIC DISORDER WITH AGORAPHOBIA PRIOR TO THE LAST 12 MONTHS ILLNESS INDUCED
11	AGORAPHOBIA WITH NO HISTORY OF PANIC DISORDER IN LAST 12 MONTHS ILLNESS INDUCED
12	INHALANT SOLVENT ABUSE DEPENDENCE IN LAST 12 MONTHS
13	HEROIN ABUSE DEPENDENCE IN LAST 12 MONTHS
14	AGORAPHOBIA WITH NO HISTORY OF PANIC DISORDER PRIOR TO THE LAST 12 MONTHS
15	COCAINE ABUSE DEPENDENCE IN LAST 12 MONTHS
16	TRANQUILIZER ABUSE DEPENDENCE IN LAST 12 MONTHS
17	ADOPTIVE MOTHER EVER AN ALCOHOLIC OR PROBLEM DRINKER
18	AMPHETAMINE ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
19	AMPHETAMINE ABUSE DEPENDENCE IN LAST 12 MONTHS
20	HALLUCINOGEN ABUSE DEPENDENCE IN LAST 12 MONTHS
21	HALLUCINOGEN ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
22	PANIC DISORDER WITH AGORAPHOBIA IN LAST 12 MONTHS ILLNESS INDUCED AND
23	OPIOID ABUSE DEPENDENCE IN LAST 12 MONTHS
24	MANIC DISORDER IN LAST 12 MONTHS ILLNESS INDUCED AND SUBSTANCE INDUCED
25	COCAINE ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
26	WENT TO COUNSELOR THERAPIST DOCTOR OTHER PERSON FOR HELP TO IMPROVE MOOD
27	STAYED OVERNIGHT IN HOSPITAL BECAUSE OF DYSTHYMIA
28	WENT TO EMERGENCY ROOM FOR HELP BECAUSE OF DYSTHYMIA
29	DOCTOR PRESCRIBED MEDICINE DRUG TO IMPROVE MOOD MAKE YOU FEEL BETTER
30	DYSTHYMIA IN LAST 12 MONTHS ILLNESS INDUCED AND SUBSTANCE INDUCED RULED OUT
31	MANIC DISORDER PRIOR TO THE LAST 12 MONTHS ILLNESS INDUCED AND
32	GENERALIZED ANXIETY IN LAST 12 MONTHS ILLNESS INDUCED AND SUBSTANCE INDUCED
33	BLOOD NATURAL MOTHER EVER AN ALCOHOLIC OR PROBLEM DRINKER
34	CANNABIS ABUSE DEPENDENCE IN LAST 12 MONTHS
35	BLOOD NATURAL MOTHER EVER HAD PROBLEMS WITH DRUGS
36	ALCOHOL ABUSE DEPENDENCE PRIOR TO THE LAST 12 MONTHS
37	DYSTHYMIA PRIOR TO THE LAST 12 MONTHS ILLNESS INDUCED AND SUBSTANCE INDUCED
38	HYPOMANIC DISORDER IN LAST 12 MONTHS ILLNESS INDUCED AND SUBSTANCE INDUCED
39	ANY FULL BROTHERS EVER ALCOHOLICS OR PROBLEM DRINKERS
40	WENT TO COUNSELOR THERAPIST DOCTOR OTHER PERSON FOR HELP WITH PANIC ATTACKS
41	WENT TO EMERGENCY ROOM FOR HELP BECAUSE OF PANIC ATTACKS
42	STAYED OVERNIGHT IN HOSPITAL BECAUSE OF PANIC ATTACKS
43	DOCTOR PRESCRIBE MEDICINE DRUG FOR PANIC ATTACKS
44	NATURAL GRANDMOTHER ON MOTHER S SIDE EVER AN ALCOHOLIC OR PROBLEM DRINKER
45	DOCTOR EVER PRESCRIBE MEDICINE DRUG FOR GENERALIZED ANXIETY

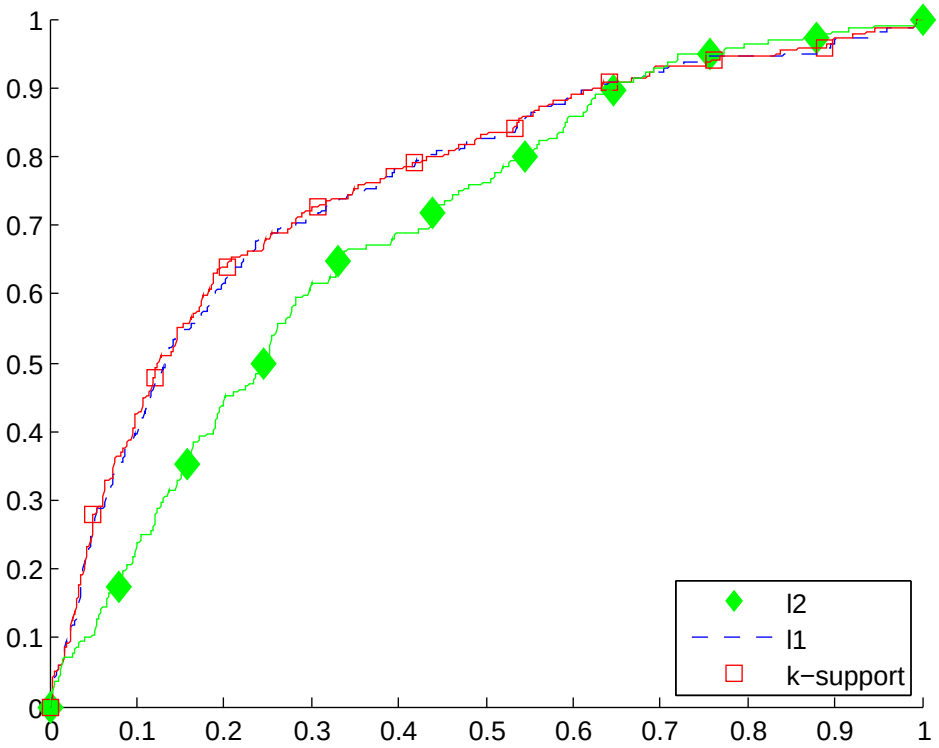

Figure 3: ROC curves for k-support norm regularization, ℓ_1 regularization, and ℓ_2 regularization of logistic regression. k-support and ℓ_1 regularization perform substantially better than ℓ_2 regularization. k-support regularization performs slightly better than ℓ_1 regularization on average, but without statistical significance (see Figure 2). However, k-support regularization is to be favored due to its better handling of correlated variables (see Section 3.3.4).

alcohol disorder, 3) family history of alcohol use disorder (full siblings, adoptive parents, natural mother, and natural grandmother) and family history of drug use disorders (natural mother, full siblings, and grandparent) and family history of mental health problems (natural mother, full sister), 4) type of treatment received for alcohol related problems (outpatient, inpatient, detoxification, rehabilitation, social services), 5) treatment received for general anxiety disorder

3.3.3. Convergence timing

We evaluate here the empirical performance of the optimization scheme described in Section 2.4 applied to k-support regularized logistic regression. After fixing λ according to the best performing setting found in the previous, we have plotted the difference in primal objective over time for varying k (Figure 4). The special case of $k = 1$ corresponds to FISTA optimization [34] of ℓ_1 regularized logistic regression, giving a strong state-of-the-art baseline method for the relative comparison of empirical convergence. We see that the convergence slows as k is increased, as has previously been noted by [8], but does not exceed one order of magnitude over the ℓ_1 optimization even for high values of k. In practice, we are likely to be interested in low values of k, which exclude irrelevant variables and have given the best generalization performance in these experiments. All computation is feasible to perform in a short amount of time on a single core, and the convergence matches the theoretical bounds given in Equation (18).

3.3.4. Empirical correlation of selected variables

In order to test whether k-support regularized logistic regression is able to automatically select multiple correlated variables, we have computed the empirical correlation between selected variables and display this matrix in graphical form in Figure 5. The variable names correspond to the NESARC Codebook available from `http://pubs.niaaa.nih.gov/publications/AA70/AA70.htm`. We first note that most pairwise correlations are close to zero, as is expected from a sparsity regularizer. However, for key subsets of related variables, there is high off-diagonal correlation. This gives strong support that the proposed correlated sparsity regularizer has the desired be-

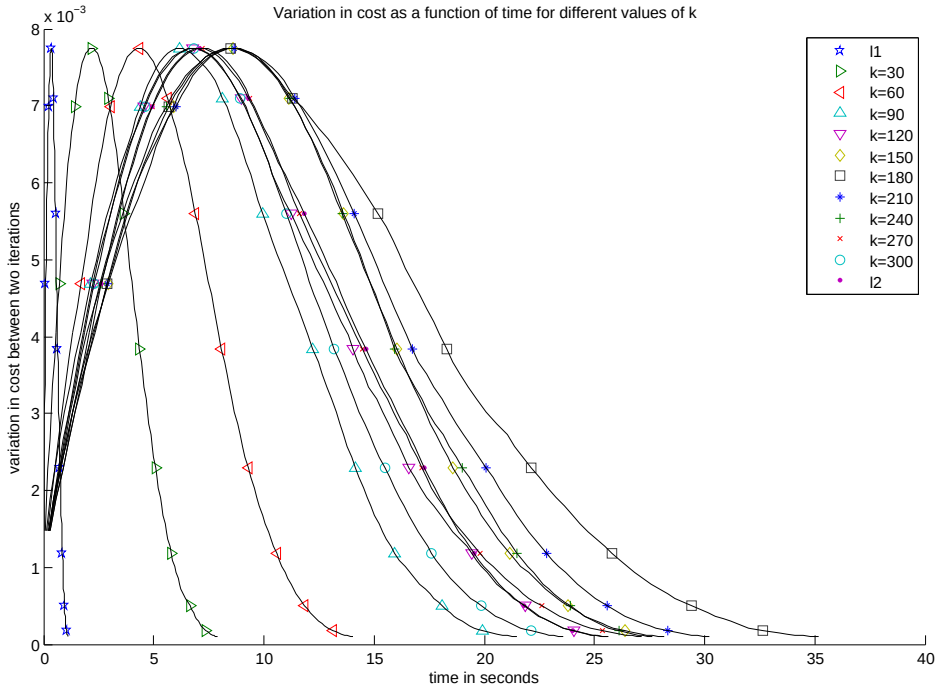

Figure 4: Empirical convergence of k-support regularized logistic regression for varying values of k. ℓ_1 regularization ($k = 1$) corresponds to FISTA optimization (see Section 3.3.3 for details). We use the same k-support code to compute ℓ_2 regularization ($k = d$), although in practice one could make use of optimization methods for twice-differentiable objective functions in this special case. There is a dependence between k and the speed of optimization, due to the longer computation required per iteration for larger values of k [8]. All computation times are feasible even on a single core for standard commercial hardware, with convergence ranging from 1–35 seconds depending on the value of k.

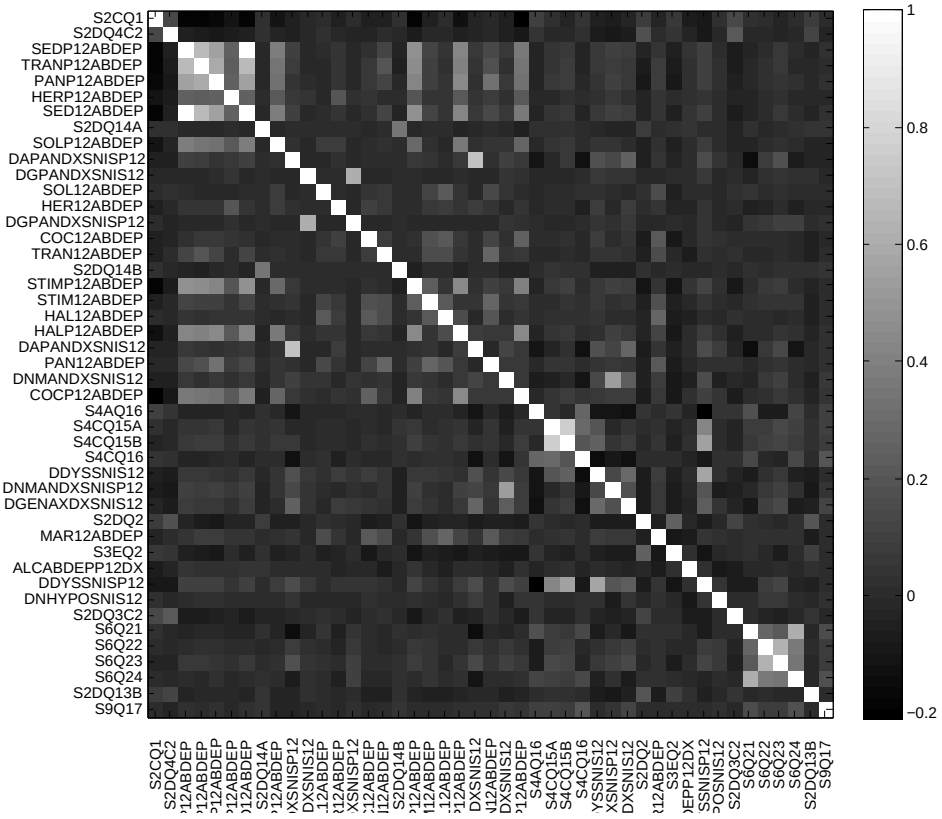

Figure 5: Empirical correlations between the top variables selected by k-support regularized logistic regression. Variable names are from the NESARC Codebook. We note that although there is a high degree of independence between variables, there are also significant correlations between some selected variables, a key benefit of the proposed method. The empirical behavior of the selected variables therefore matches the expectation of what should be achieved by a properly functioning correlated sparsity regularizer. See Section 3.3.4 for a discussion of these selected variables.

havior and is able to select multiple related variables. Furthermore, the variables for which there is a high degree of correlation indicate that the method has been able to automatically account for confounding variables.

We observe high correlations between such pairs of selected variables as for example: dysthymia prior to the last 12 months (illness-induced and substance-induced ruled out) (DDYSSNISP12) and emergency room treatment because of dysthymia (S4CQ15B); hospitalization due to dysthymia (S4CQ15A) and emergency room visit due to dysthymia (S4CQ15B). The level of comorbidity of alcohol use disorders and mood disorders such as dysthymia is high: persons mood disorders have a 2- to 3-fold increased risk of alcohol use disorders [39, 25]. Secondary alcohol use disorders may result from self-medication of symptoms of dysthymia with alcohol [40], while dysthymia may be consequences of alcohol intoxication and/or withdrawal in primary alcohol use disorders [41]. Individuals who simultaneously suffer from alcohol use disorder and dysthymia may prefer to seek treatment for their dysthymic disorder, because of lesser stigma, easier access, and simpler reimbursement scheme. This will result in an increase of unmet need for treatment of their alcohol related problems [11]. Therefore, dysthymia can be considered as a known confounder.

Treatment in the emergency room because of dysthymia (S4CQ15B) may be considered as a potential cause of unmet need for alcohol treatment, due to the absence of an integrated treatment of comorbid psychiatric and addictive conditions [11]. Alcohol use disorder is often not treated by mental health services not specialized in addiction, therefore individuals with such comorbid conditions may have the impression that their alcohol use problem is not serious enough to require treatment or that adequate treatment is not available.

A second association is represented by two different types of treatment for dysthymia: hospitalization due to dysthymia (S4CQ15A) and emergency room visit due to dysthymia (S4CQ15B), in particular its more acute forms.

Indeed, inclusion of these two variables instead of only one can help us to understand how types of treatment of dysthymia influence the unmet treatment for the alcohol use disorder. For the purpose of epidemiological analysis, these two variables will often be grouped together to represent a variable for treatment of more acute forms of dysthymia.

These observations confirm one of the key hypotheses of the application of k-support norm regularization in an epidemiological task: that confounders and other correlated variables can be automatically selected, strengthening the epidemiological analysis.

4. Discussion

4.1. Epidemiological Discussion

Our results obtained with k-support regularized logistic regression and based on a nationally representative sample of the US adult population show that the three most relevant groups of variables were represented by lifetime alcohol use disorder, as a major inclusion criteria for the study, lifetime drug use disorder, and lifetime mental health problems. This is in line with recent epidemiological studies that suggest that alcohol use disorder is in 74.9% co-occurrence with illegal drug use disorder and that comorbid alcohol use disorder is associated with a two-fold increase in the likelihood of perceived unmet need for illegal drug use disorder [11]. Furthermore, lifetime alcohol use disorder was shown to be strongly and significantly associated with lifetime psychiatric disorders [25]. In addition, people with co-occurring substance and mental health problems tend to prefer to use mental health services instead of specialized substance use programs [11].

Family history of substance (alcohol, drugs) use disorders and behavioral problems was the fourth most relevant group of variables. The heritability of alcohol consumption is estimated at 35% to 40% in twin studies [42]. Adoptive parents with alcohol use disorder increase the probability of their adoptive children to have the same disorder [43]. Furthermore, the aggregation of drug dependence and alcohol dependence within some families suggests common mechanisms for these disorders [44].

The type of treatment for alcohol use disorder is the fourth most relevant group of variables selected. From an epidemiological point of view, it is important to evaluate the current state of treatment in the population with alcohol use disorders in order to give recommendations for the improvement of access to care for these group of people.

Additionally, our model includes variables on treatment receipt for generalized anxiety disorder. The co-occurrence of generalized anxiety disorder (GAD) and alcohol use disorder (AUD) disorder with the rate of comorbidity ranging from 8.3% to 56.2% is well documented [45]. GAD was also found to be a relevant factor among individuals with alcohol use disorder seeking outpatient substance abuse treatment [46].

Overall, variable selection performed with the help of the k-support Regularized Logistic is soundly supported by the scientific evidence. Selected variables can be used in the study that can lead to recommendation for an improvement of access to treatment among population with alcohol use disorder.

From an epidemiological point of view, variable selection using such a method is useful for identifying the most important risk factors and disregarding variables that do not add additional information in a reliable and time efficient way, therefore limiting the number of variables required for statistical analyses and improving the reliability of the final results. The selected variables may provide a suggestion for data collection for future epidemiological studies.

4.2. Statistical Discussion

Our quantitative evaluation supports the use of k-support norm regularized logistic regression over the compared methods, ℓ_1 and ℓ_2 regularized logistic regression. The use of sparsity regularization is firmly associated with significant improved empirical performance over ℓ_2 regularization and baseline methods, and the k-support norm performed better on average than ℓ_1 regularization, though a Wilcoxon signed rank test was not able to reject the null hypothesis in the comparison between k-support and ℓ_1 regularization on this data. Nevertheless, the k-support norm is to be preferred over ℓ_1 regularization even when statistically tied due to its improved handling of confounding factors and correlated signals (see Section 3.3.4). This is of particular interest when considering data such as those from NESARC in which there are a large number of variables incorporating potential confounders and comorbidities.

5. Conclusions

In this paper, we introduced a novel multi-class generalization of regularized logistic regression using the k-support norm. We have shown its application to the discovery of predictors of mental health service utilization on a large scale epidemiological database, NESARC. The results were both qualitatively and quantitatively analyzed, showing statistically significant improvement of k-support norm regularized logistic regression over ℓ_2 regularized logistic regression and baseline methods, and a statistical tie between k-support regularization and ℓ_1 regularization, but with better average performance and improved statistical characteristics. Qualitative analysis of the ranked variables showed a clear trend elucidating the relationship between covariates of the primary variable of interest. Discovering factors for not obtaining help with drinking is a vital public health question, with high societal and public policy impacts, and k-support regularized logistic regression has shown to be a powerful tool for analyzing this problem.

Acknowledgements

This work was supported in part by the European Commission through ERC Grant 259112, and FP7-MC-CIG 334380. The funding agencies had no involvement in the decision to submit the article for publication.

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society (Series B) 58 (1) (1996) 267–288.

[2] F. Sha, Y. A. Park, L. K. Saul, Multiplicative updates for l1–regularized linear and logistic regression, in: M. R. Berthold, J. Shawe-Taylor, N. Lavrač (Eds.), Advances in Intelligent Data Analysis VII, Vol. 4723 of Lecture Notes in Computer Science, Springer, 2007, pp. 13–24.

[3] G.-X. Yuan, C.-H. Ho, C.-J. Lin, An improved GLMNET for l1-regularized logistic regression, Journal of Machine Learning Research 13 (1) (2012) 1999–2030.

[4] T. Manninen, H. Huttunen, P. Ruusuvuori, M. Nykter, Leukemia prediction using sparse logistic regression, PLoS ONE 8 (8) (2013) e72932.

[5] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[6] J. Huang, T. Zhang, D. Metaxas, Learning with structured sparsity, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 417–424.

[7] R. Jenatton, J.-Y. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms, Journal of Machine Learning Research 12 (2011) 2777–2824.

[8] A. Argyriou, R. Foygel, N. Srebro, Sparse prediction with the k-support norm, in: Advances in Neural Information Processing Systems, 2012, pp. 1466–1474.

[9] C. M. Chen, H. Yi, D. E. Falk, F. S. Stinson, D. A. Dawson, B. F. Grant, Alcohol use and alcohol use disorders in the United States: Main findings from the 2001–2002 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), Tech. Rep. 05-5737, National Institute on Alcohol Abuse and Alcoholism, Bethesda, MD (2006).

[10] E. Prokofyeva, S. S. Martins, N. Younès, P. J. Surkan, M. Melchior, The role of family history in mental health service utilization for major depression, Journal of Affective Disorders 151 (2) (2013) 461–466.

[11] M. Melchior, E. Prokofyeva, N. Younès, P. Surkan, S. Martins, Treatment for illegal drug use disorders: the role of comorbid mood and anxiety disorders, BMC Psychiatry 14 (1) (2014) 89.

[12] S. Chakravorty, N. Jackson, N. Chaudhary, P. J. Kozak, M. L. Perlis, H. R. Shue, M. A. Grandner, Daytime sleepiness: Associations with alcohol use and sleep duration in Americans, Sleep Disorders (2014) 7 pages.

[13] M. Zins, S. Bonenfant, M. Carton, M. Coeuret-Pellicer, A. Gueguen, J. Gourmelen, M. Nachtigal, A. Ozguler, A. Quesnot, C. Ribet, G. Rodrigues, A. Serrano, R. Sitta, A. Brigand, J. Henny, M. Goldberg, The CONSTANCES cohort: an open epidemiological laboratory, BMC Public Health 10 (1) (2010) 479.

[14] M. G. Marmot, G. D. Smith, S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, A. Feeney, Health inequalities among British civil servants: the Whitehall II study, The Lancet 337 (8754) (1991) 1387–1393.

[15] G.-M. Hinnouho, S. Czernichow, A. Dugravot, H. Nabi, E. J. Brunner, M. Kivimaki, A. Singh-Manoux, Metabolically healthy obesity and the risk of cardiovascular disease and type 2 diabetes: The Whitehall II cohort study, European Heart Journal 36 (9) (2014) 551–559.

[16] I. R. Dohoo, C. Ducrot, C. Fourichon, A. Donald, D. Hurnik, An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies., Preventive Veterinary Medicine 29 (3) (1997) 221–239.

[17] S. Walter, H. Tiemeier, Variable selection: Current practice in epidemiological studies, European Journal of Epidemiology 24 (12) (2009) 733–736.

[18] S. Greenland, Invited commentary: Variable selection versus shrinkage in the control of multiple confounders, American Journal of Epidemiology 167 (5) (2008) 523–529.

[19] M. A. Hernán, S. Hernández-Díaz, M. M. Werler, A. A. Mitchell, Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology, American Journal of Epidemiology 155 (2) (2002) 176–184.

[20] S. P. Chou, D. A. Dawson, F. S. Stinson, B. Huang, R. P. Pickering, Y. Zhou, B. F. Grant, The prevalence of drinking and driving in the United States, 2001–2002: Results from the national epidemiological survey on alcohol and related conditions, Drug and Alcohol Dependence 83 (2) (2006) 137–146.

[21] R. Caetano, S. Nelson, C. Cunradi, Intimate partner violence, dependence symptoms and social consequences from drinking among white, black and hispanic couples in the United States, The American Journal on Addictions 10 (s1) (2001) s60–s69.
[22] P. Lemoine, H. Harousseau, J. P. Borteyru, J. C. Menuet, Children of alcoholic parents–observed anomalies: Discussion of 127 cases, Theraputic Drug Monitoring 25 (2) (2003) 132–136.
[23] M. E. Bates, S. C. Bowden, D. Barry, Neurocognitive impairment associated with alcohol use disorders: Implications for treatment, Experimental and Clinical Psychopharmacology 10 (3) (2002) 193.
[24] B. F. Grant, F. S. Stinson, D. A. Dawson, S. P. Chou, M. C. Dufour, W. Compton, R. P. Pickering, K. Kaplan, Prevalence and co-occurrence of substance use disorders and independentmood and anxiety disorders: Results from the national epidemiologic survey on alcohol and related conditions, Archives of General Psychiatry 61 (8) (2004) 807–816.
[25] D. S. Hasin, F. S. Stinson, E. Ogburn, B. F. Grant, Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: Results from the national epidemiologic survey on alcohol and related conditions, Archives of General Psychiatry 64 (7) (2007) 830–842.
[26] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer Series in Statistics, Springer, 2009.
[27] A. Y. Ng, Feature selection, l_1 vs. l_2 regularization, and rotational invariance, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, pp. 78–85.
[28] L. Jacob, G. Obozinski, J.-P. Vert, Group lasso with overlap and graph lasso, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 433–440.
[29] L. Meier, S. van de Geer, P. Bühlmann, The group lasso for logistic regression, Journal of the Royal Statistical Society, Series B 70 (1) (2008) 53–71.
[30] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Soviet Mathematics Doklady 27 (1983) 372–376.
[31] Y. Nesterov, On an approach to the construction of optimal methods of minimization of smooth convex functions, Ekonomika i Matematicheskie Metody 24 (1988) 509–517.
[32] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical Programming 103 (1) (2005) 127–152.
[33] Y. Nesterov, Gradient methods for minimizing composite objective function, CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE) (2007).
[34] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences 2 (1) (2009) 183–202.
[35] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, Submitted to SIAM Journal on Optimization.
[36] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, NY, USA, 2004.
[37] H. Narasimhan, S. Agarwal, On the relationship between binary classification, bipartite ranking, and binary class probability estimation, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 2913–2921.
[38] T. Zhang, On the consistency of feature selection using greedy least squares regression, Journal of Machine Learning Research 10 (2009) 555–568.
[39] R. de Graaf, R. V. Bijl, J. Spijker, A. T. F. Beekman, W. A. M. Vollebergh, Temporal sequencing of lifetime mood disorders in relation to comorbid anxiety and substance use disorders, Social Psychiatry and Psychiatric Epidemiology 38 (1) (2003) 1–11.
[40] J. M. Bolton, J. Robinson, J. Sareen, Self-medication of mood disorders with alcohol and drugs in the National Epidemiologic Survey on Alcohol and Related Conditions, Journal of Affective Disorders 115 (3) (2009) 367–375.
[41] D. E. Falk, H.-Y. Yi, M. E. Hilton, Age of onset and temporal sequencing of lifetime DSM-IV alcohol use disorders relative to comorbid mood and anxiety disorders, Drug and Alcohol Dependence 94 (13) (2008) 234–245.
[42] C. A. Clifford, J. L. Hopper, D. W. Fulker, R. M. Murray, A genetic and environmental analysis of a twin family study of alcohol use, anxiety, and depression, Genetic Epidemiology 1 (1) (1984) 63–79.
[43] N. S. Cotton, The familial incidence of alcoholism: A review, Journal of Studies on Alcohol and Drugs 40 (01) (1979) 89.
[44] J. Nurnberger, J. I., R. Wiegand, K. Bucholz, S. O'Connor, E. T. Meyer, T. Reich, J. Rice, M. Schuckit, L. King, T. Petti, L. Bierut, A. L. Hinrichs, S. Kuperman, V. Hesselbrock, B. Porjesz, A family study of alcohol dependence: Coaggregation of multiple disorders in relatives of alcohol-dependent probands, Archives of General Psychiatry 61 (12) (2004) 1246–1256.
[45] M. G. Kushner, K. J. Sher, B. D. Beitman, The relation between alcohol problems and the anxiety disorders, The American Journal of Psychiatry 147 (6) (1990) 685–695.
[46] J. P. Smith, S. W. Book, Comorbidity of generalized anxiety disorder and alcohol use disorders among individuals seeking outpatient substance abuse treatment, Addictive Behaviors 35 (1) (2010) 42–45.