CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

QUESTÕES METODOLÓGICAS
METHODOLOGICAL ISSUES

Strategic procedure in three stages for the selection of variables to obtain balanced results in public health research

Procedimiento estratégico en tres fases para la selección de variables, con el fin de obtener resultados equilibrados en investigación sobre salud pública

Procedimento estratégico em três estágios de seleção de variáveis para a obtenção de resultados equilibrados na pesquisa em saúde pública

Manuel Lozano [1]
Lara Manyes [1]
Juanjo Peiró [2]
Adina Iftimi [2,3]
José María Ramada [4,5]

doi: 10.1590/0102-311X00174017

Abstract

Multidisciplinary research in public health is approached using methods from many scientific disciplines. One of the main characteristics of this type of research is dealing with large data sets. Classic statistical variable selection methods, known as "screen and clean", and used in a single-step, select the variables with greater explanatory weight in the model. These methods, commonly used in public health research, may induce masking and multicollinearity, excluding relevant variables for the experts in each discipline and skewing the result. Some specific techniques are used to solve this problem, such as penalized regressions and Bayesian statistics, they offer more balanced results among subsets of variables, but with less restrictive selection thresholds. Using a combination of classical methods, a three-step procedure is proposed in this manuscript, capturing the relevant variables of each scientific discipline, minimizing the selection of variables in each of them and obtaining a balanced distribution that explains most of the variability. This procedure was applied on a dataset from a public health research. Comparing the results with the single-step methods, the proposed method shows a greater reduction in the number of variables, as well as a balanced distribution among the scientific disciplines associated with the response variable. We propose an innovative procedure for variable selection and apply it to our dataset. Furthermore, we compare the new method with the classic single-step procedures.

Statistics as Topic; Methods; Interdisciplinary Research

Correspondence
M. Lozano
Universitat de València.
Vicent Andrés Estellés s/n, Burjassot / Valencia – 46100, España.
manuel.lozano@uv.es

[1] *Departament de Medicina Preventiva i Salut Pública, Ciències de l'Alimentació, Toxicologia i Medicina Legal, Universitat de València, Valencia, España.*
[2] *Departament d'Estadística i Investigació Operativa, Universitat de València, Valencia, España.*
[3] *Department of Biosciences and Nutrition. Karolinska Institutet, Huddinge, Sweden.*
[4] *Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona, España.*
[5] *CIBER de Epidemiología y Salud Pública, Madrid, España.*

Introduction

The high number of possible variables involved and the different scientific disciplines to which they belong is a characteristic of research in public health. The World Health Organization (WHO) broadly described the determinants of health and the systems established to deal with illness in 2008 [1]. The determinants were grouped in sets of variables, ranging from those related to general socioeconomic, cultural and environmental conditions to those most proximal to the individual, such as living and work conditions, health, education and lifestyle factors, and individual factors such as age, sex and hereditary factors. Most of these factors are determined by the geographical area and affect, directly or indirectly, the health and well-being of people. Therefore, public policies must adapt their conceptual frameworks considering the most important variables for their analysis without losing relevant information.

The WHO encourages stakeholders to measure all significant variables involved in action assessment, to broaden the knowledge base, to develop the trained workforce in the social determinants of health, and to raise awareness on all determinants of health. Consequently, public health research studies should include all relevant variables in the analyses, properly selecting those that are actually significant.

Frequently, these sets of variables must be reduced to obtain the simplest statistical models with the minimum loss of information. For such, many well-known statistical techniques for variable selection exist. The problem lies in the different statistical weight of each variable depending on its stronger or weaker direct association with the assessed response variable, masking each other. This causes many relevant variables to be eliminated during the process, as well as possible interactions among the variables that are not addressed correctly.

The classical statistical techniques of variable selection, known as "screen and clean" and used in a single step, select the variables with greater explanatory weight in the model in an increasingly efficient way [2,3,4]. However, the results may be biased depending on the expert's point of view of each discipline involved, due to the omission of complete subsets of explanatory variables by masking. This bias occurs when the relation between the discarded variables and the response variable were previously justified in abundant scientific literature. Social and health variables involved in public health research are typical examples [5,6]. For this reason, the choice of the variable selection strategy is particularly important, as it not only improves the prediction accuracy but also provides a clear interpretation of the most informative variables [7,8]. Also, from a statistical point of view, identifying the most important variables in each research discipline that affect a response variable is necessary to reduce its number with the minimal loss of information and to respect the causality explained by each discipline, thus, avoiding multicollinearity [9,10] and masking effects [11].

All predictive variables can affect, to a greater or lesser extent, and in different ways, the response variable [12]. However, the well-known procedures for selecting variables in a single step, either efficiently select the most relevant and neglect those with less explanatory power or try to achieve a more balanced model, sacrificing selective capacity. To deal with this fact, this article proposes a three-step statistical procedure resulting from applying several variable selection methods that are especially designed for large and heterogeneous sets of variables. Using this procedure, selecting a smaller number of variables of each disciplinary subset and obtaining a balanced final set of variables that can explain most of the data variability is possible.

To do this, in the First Step we conduct a variable pre-selection on each discipline obtaining a specific contribution index of each variable to the data variability. In the Second Step, we check the prediction capability of the pre-selected variables. In the Third Step, we apply different variable selection methods to assess the contribution of the pre-selected variables in the model using a classical method (linear regression using classical variable selection methods), a Bayesian approach (Bayesian variable selection methods using Integrated Nested Laplace Approximation – INLA), and penalized regressions (such as Least Absolute Shrinkage and Selection Operator – Lasso; Adaptive Lasso – ALasso; or Elastic Net).

The paper structure is as follows. In *Methods* section, we explain the procedure step by step, specifying the statistical methodology used for each of the steps. In *Results* section, we introduce a case study applied to public health. We analyze the results of the implementation of the procedure on the dataset to investigate its predictive ability to reduce the number of variables in high-dimensional regression. Then, we discuss the results and show the conclusions of our study in the *Discussion* section.

Methods

In this section, we explain the details of the proposed three-step statistical procedure. This method is used to analyze data from multidisciplinary research in public health when the interest is on the variability of each scientific discipline. All procedures were performed using the R statistic software, version 3.1.2 (The R Foundation for Statistical Computing; http://www.r-project.org).

First Step

The first objective is to reduce the dimensionality in each subset of variables that belong to different scientific disciplines. Reducing the dimensionality is a central problem in multivariate data analysis. When accurately describing the information within a set of p variables through a small subset of size r is possible, the dimension of the problem is reduced little information is lost [12]. This allows the identification of the variables that are generating data variability. Thus, considering the sample size (n), a set of p predictive variables and a response variable y, we try to transform the original array $X_{(n,p)}$ in a new array $Z_{(n,r)}$ (where $r < p$).

For this, in the different subgroups of predictive variables, we used the principal component analysis (PCA) when dealing with subsets of quantitative variables, or multiple correspondence analysis (MCA) when dealing with subsets of categorical variables. Usually, these methods are used to replace the variables with the main components, but in our case, we are using them to select a subset of variables. Any expected response variables from the original dataset should be excluded from these analyses and the interactions between quantitative and qualitative variables should not to be considered (since including all possible combinations of variables could produce an excessively broad set of them). Therefore, we seek to obtain the most relevant predictive variables of each subset considering the contribution of each variable in each principal component, through a Contribution Index (CI). The R package *FactoMine* was used to perform these analyses.

- **A Contribution Index**

When using PCA and MCA as dimensionality reduction techniques, we are not looking for a replacement of the original set of variables with a smaller set of principal components, but our objective is to reduce the original set of variables. For such, we define an index for each variable X_i that measures its contribution to the main components in which the variable is involved. We define this contribution index as a weighted average of its contribution to each component with the explained variance of the components as weight,

$$CI_i = \sum_{k=1}^{l} c_{i_k} v_k \qquad \text{(Equation 1)}$$

where c_{ik} is the contribution of variable X_i to the component k and v_k is the explained variance for this component. The values of c_{ik} are provided by PCA or MCA. Where c_{ik} are the eigenvalues that represent the percentage of explained variance in each component and v_k are the proportions of contribution of each variable to each component.

The *estim_ncp* function was used to obtain the best number of dimensions to use. The analysis of the top five main components was sufficient to obtain the representativeness of data variability, so l = 5 in Equation 1. Establishing a criterion of homogeneity for the number of variables selected according to the CI is advisable, checking the minimum number of variables that explain the maximum of variability in a regression model that considers all the selected variables together. Thus, the simplest model with maximum predictive capability should be chosen to continue on the next step. Based on this method, the five variables with highest contribution of each block are chosen, since this number represents a balance between the loss of information and an effective reduction in the number of variables.

Second Step

When we reduce variables in multidisciplinary research, we try to explain most of the behavior of a response variable using the smallest number of variables from a large initial set of variables. Therefore, once the variables that explain the variability in each subset have been pre-selected, the predictive quality of each of them must be evaluated in the presence of the others. Through this process, their inclusion in the statistical model can be justified.

This is done in two ways, one consists of applying a new PCA on these predictive variables and studying the behavior of the principal components regarding the response variable. Note that quantitative and qualitative variables may coexist in the pre-selected variables, and to apply the PCA again, the latter should be transformed into as many dummy variables as categories minus one. The other way to validate the selection uses a specific regression to assess the predictive capacity of the model, according to the nature of the response variable. If the predictive capability of the pre-selected variables does not reach the expectations of the study, this procedure would not be advisable. The R packages *FactoMine* (PCA and graphics) and *nnet* (multinomial regressions) were used to perform these analyses.

Third Step

Our objective in the Third Step is to further reduce the number of variables and identify those that are the most relevant of a saturated model containing all the pre-selected and assessed variables of each scientific discipline.

This is done by applying different methods to assess the contribution of these pre-selected variables in the regression model, either using a classical method (Generalized Linear Model – GLM – using classical variable selection methods), a Bayesian method (Bayesian variable selection methods using INLA), or a penalized regressions method, such as Lasso, ALasso or Elastic Net.

• GLM and classical variable selection methods

Different methods of variable selection can be applied from a classical point of view. One of the most used procedures is the stepwise method, it does not guarantee the best regression equation but provides models that are usually close to the optimum. at the purpose of this method is finding the variables that better fit the model and comparing the obtained models by adding one more variable to finally select the model with less Akaike Information Criterion (AIC) or Deviance criterion [12].

Two drawbacks are known when using the stepwise method: in both the forward selection and in the backward elimination versions, the method considers at most $p + (p - 1) + ... + 1 = p(p + 1)/2$ subsets from 2^p possible, and this makes it difficult to find the optimal model. The second drawback is the masking of variables with a true effect on the response factor due to the correlation of other variables that were selected in the model. On the other hand, the classical convergence problems in regressions with large number of variables are solved due to the reduction in the number of variables done in the First Step. The R package *stats* was used to perform these analyses.

• INLA and Bayesian variable selection methods

The INLA procedure [13] was recently implemented for Bayesian inference in several statistical models, it is faster computationally than the Markov chain Monte Carlo (MCMC) methods. MCMC methods [14] simulate samples with some form of dependence that converge on the distribution of interest, in which information about the expected a priori behavior was incorporated based on a previous professional knowledge. This method seeks samples for a posterior distribution $\boldsymbol{\pi}(\theta|y)$, constructing a Markov chain to do the Monte Carlo approximation.

The INLA alternative offers reliable approximations to the marginals a posteriori in a short computational time and also provides the Deviance Information Criterion (DIC), which is useful to select the most appropriate model [15] and is equivalent to the classic AIC [12]. The R packages *INLA* and *BayesVarSel* were used to perform these analyses.

- **Other Bayesian variable selection methods**

There are other Bayesian methods for behavior diagnosis of the regression model, and to choose the variables that best explain the variability of data when the set of variables is large: (a) we can calculate Bayes factors in linear models and then provide a formal Bayesian answer to solve the variable selection problems [16]. There are libraries in R packages that solve the problem of obtaining a posteriori probabilities of all the possible linear models resulting from the different combinations of explanatory variables; (b) an interesting alternative is the Gibbs sampling [17], which is a specific MCMC algorithm [16] to obtain a sequence of observations, and is considered a general framework for sampling a large set of variables by sampling each variable. The R packages *BayesFactor* and *MCMCpack* were used to perform these analyses.

- **Penalized logistic regression models**

Methods that use shrinkage estimators are an alternative to the methods described above. These methods reduce the variance of the estimators with lower predictive error and a variable selection that is not as arbitrary. The most common are the Ridge regression [17] and the Lasso regression [18].

We are interested in including these methods because they are usually used in large and heterogeneous datasets to implement a variable selection in a single step and are useful for dealing with multicollinearity and masking. These methods are known as regularization or shrinkage methods because they contract the regression coefficients to stabilize the estimation. This regularization means that the size of the parameter vector is restricted to a certain range, causing highly correlated explanatory variables to produce very unstable minimum quadratic estimates or simply allowing single estimates to occur (when there is collinearity, or the number of variables exceeds the number of observations). These methods are typically used for the regression of a dependent variable and an array of high-dimensional X, with highly correlated variables.

We think that the heart of the matter is the penalty. To avoid the over-shooting due to the large number of predictor variables, the method imposes a penalty on large fluctuations in the estimated parameters. Therefore, choosing the penalty parameter (λ) is essential and a procedure to estimate the parameter λ value from the data is needed.

- **The Lasso regression**

The Lasso regression combines shrinkage and variable selection by imposing a penalty on the regression coefficients, thus, for high values of the penalization parameter some of these coefficients are set to zero. Lasso imposes the L_1 norm on the least squares problem and shrinks the coefficients towards zero. This difference in the penalization may seem marginal but it has big consequences. The use of L_2 norm in other methods causes the pleasant effect of producing a linear estimator in y of the parameters vector β, but in return uses all the predictive variables in the final regression model because higher λ values contract the coefficients towards zero, usually this value is not reached. On the other hand, Lasso, through the L_1 norm, does not produce a linear estimator in y and a formula for its expression is not obtained. In this case the solution must be found through an optimization algorithm. However, depending on the choice of the complexity parameter, the L_1 penalization produces some regression coefficients equal to zero. The advantage is that in the final model only some of the variables are considered, being a method of estimation and variable selection at the same time [18,19,20].

Some authors [18,19] compared the prediction performance of the Lasso with other penalized regression methods and found that none of them uniformly dominates the other. However, given that variable selection is becoming increasingly important in modern data analysis, the Lasso is much more appealing due to its sparse representation [21]. However, Lasso presents an important limitation when considered in multidisciplinary research, if there is a group of variables among which the pairwise correlations are very high, then the method tends to select only one variable from the group and does not care which is selected. The R package *glmnet* was used to perform these analyses.

- **The Elastic Net regression**

Elastic Net [21] is a method that combines Ridge [17] and Lasso to overcome the Lasso limitations. Like the Lasso, this method is a regularization technique, however, it performs automatic variable selection and continuous shrinkage simultaneously and can select groups of correlated variables. The Elastic Net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. This method is particularly useful when the number of predictors (*p*) is much bigger than the number of observations (*n*).

In a way, the Elastic Net regression combines the strengths of Lasso and Ridge. The L_1 part of the penalty generates a sparse model and the quadratic part of the penalty removes the limitation on the number of selected variables, encourages a grouping effect and stabilizes the L_1 regularization path [21]. The R package *elasticnet* was used to perform these analyses.

- **The ALasso regression**

The ALasso is a Lasso generalization that allows different penalties to be applied to the variables by assigning different weights, which depend on the data. This generalization can impose greater penalties on the variables with lower relevance and small penalties on the most relevant [22]. The R package *parcor* was used to perform these analyses.

Three-step statistical procedure applied in a real dataset of a public health research

We applied this procedure on a sample of 299 disabled older adults who live in the Eastern region of Spain. People of both sexes aged 65 and over were included (only those who lacked the cognitive ability to follow the interview were excluded). 203 participants live in 4 metropolitan municipalities (with populations between 20,000-70,000 inhabitants), and 151 participants live in 11 rural municipalities of countryside regions (with populations between 500-5,000 inhabitants).

This sample includes all the subsets of variables considered as health determinants: anthropometrics, social (cohabitation unit, social care services, and demographics), and health variables (chronic diseases, drugs consumption, diet, functional limitations, autonomy level and disabilities), comprising up to 131 variables (the complete variable set it is available online in http://pages.uv.es/malore2/summary_variables.pdf).

Data on dietary intake, dietary habits, anthropometrics and demographics were collected from all participants through interviews. The interviews were performed by social workers from each collaborating municipality, previously trained by the project managers at the University of Valencia, Valencia, Spain). Data on food consumption were obtained strictly during spring to guarantee the same food seasonality. A validated food-frequency questionnaire was used [23]. Dietary data obtained through questionnaires were analysed by the DIAL software, version 2.12 (Alce Ingeniería; https://www.alceingenieria.net/infodial.htm). The daily intake of energy, macronutrients, micronutrients, alcohol and water of each participant was calculated.

Information about chronic diseases, drugs consumption, functional limitations, level of autonomy, family unit and social care services received were simultaneously collected by the social-service workers. This data was extracted from each participant's individual file managed by the social services.

Therefore, among the predictive variables we can highlight morbidity, polypharmacy, diet, functional limitations, level of autonomy, disabilities, access to welfare and the family unit. The response variable is the geographical profile (categorical factor with two categories: metropolitan and rural profile) that define the trends in the health determinants of each region [24], usually polarized between rural and metropolitan environment [25,26].

Furthermore, the traditional lifestyle in Spain, especially in the East, is considered as similar to the rest of the developed countries of the Mediterranean area [27,28,29]. For this reason, the research on the factors involved in public health could be easily extrapolated to the rest of these countries [30].

The studied models were built using a training set containing 70% of the sample, randomly chosen (209 participants). The performance of the model is evaluated with the validation set which constitutes the remaining 30% of the sample (90 participants) thorugh their prediction tables in each applied method. To assure that results are not biased by the random sampling, the complete procedure was repeated 50 times with new, randomly chosen, samples.

Ethics approval and consent to participate

The protocol H133534717755 of this study was approved by the Ethics Committee of the University of Valencia and respects all the principles of the *Declaration of Helsinki* and the Spanish legal regulations on protection of personal data. Study participants were informed of the objectives and the scope of the study, they signed an informed consent form for their participation.

Results

First Step applied in a real dataset of a public health research

The complete set of 131 predictive variables was divided into disciplinary subsets (dietary, basic nutrients, total nutrients, pharmacological, pathological and disability) and submitted to a preliminary analysis to reduce the number of variables of each subset with the minimal loss of information provided by each one. Each subset was analyzed according to its categorical or quantitative nature but including the geographic factor only as illustrative to observe its behavior regarding the PCA or MCA, without taking part in them. The first five principal components obtained through PCA and MCA for each subset were used to pre-select the five original variables with higher CI (Equation 1). This procedure identified the main explicative variables in each of them, considering their different causalities. Table 1 shows this pre-selection of variables.

Second Step applied in a real dataset of a public health research

To design the statistic model, the next step was to verify the association of these pre-selected variables with the response variable, the geographical factor in this case. For such, we performed two procedures. First, we associated them graphically, showing the pre-selected predictive variables and the response variable in the same graph through their respective score for the principal components that bear the highest variance of data. Therefore, as shown in Figure 1, the metropolitan profile is associated with a greater number of chronic diseases, drug consumption, traumatic diseases and severe disorders. Also, the rural profile tends towards a better health status without the need for home assistance, as well as an increased caloric intake from lipids, carbohydrates and occasional food. For the second procedure, we validated this pre-selected model implementing a logistic regression and assessing the predictive ability of the model through a prediction table. Table 2 shows that the metropolitan and rural predictions are good (93.1% and 75.8% of success, respectively).

Third Step applied in a real dataset of a public health research

Finally, we applied different methods to assess the contribution of the pre-selected variables in the regression model, either from the classical approach, from a Bayesian approach, and through penalized regressions. The predictive ability of the selected models obtained from different methods was evaluated through prediction tables. To validate the procedure, we implemented the obtained models on the data from the remaining 30% of the sample, then we contrasted the prediction tables from each method applied.

These methods are usually applied in a single step in the datasets. We further demonstrate the convenience of our three-steps strategy to achieve more balanced and efficient results in multidisciplinary research.

Table 1

Explained variability and pre-selected variables in each subset obtained through their Contribution Index – CI (First Step).

Subset	% explained variability	Pre-selected variables	CI
Dietary	47.46	Liquid food	454.71
		Water intake	408.61
		Occasional food	318.79
		Oily fish	304.76
		Seafood	292.30
Basic nutrients	89.64	Caloric intake	622.13
		Difference caloric intake/EER	616.80
		Total fat	610.90
		Carbohydrates	598.02
		Saturated fatty acids	597.33
Total nutrients	80.28	Caloric intake	254.52
		Total fat	247.55
		Iron	244.48
		Saturated fatty acids	244.43
		Carbohydrates	243.25
Pharmacological	45.97	Total drugs	452.16
		CNS drugs	394.31
		Digestive drugs	314.47
		Drugs that increase appetite	312.77
		Drugs that cause dysgeusia	311.02
Pathological	41.83	Total diseases	518.53
		Nutritional diseases	310.55
		Ocular diseases	304.72
		Traumatic diseases	296.79
		Musculoskeletal diseases	292.33
Disability	46.00	Severe confusion	158.54
		Severe conduct disorder	156.22
		Severe Alzheimer or dementia	136.90
		Absence of dependence for shopping and managements	134.66
		Absence of home assistance	128.02

CNS: central nervous system; EER: energy efficiency ratio.

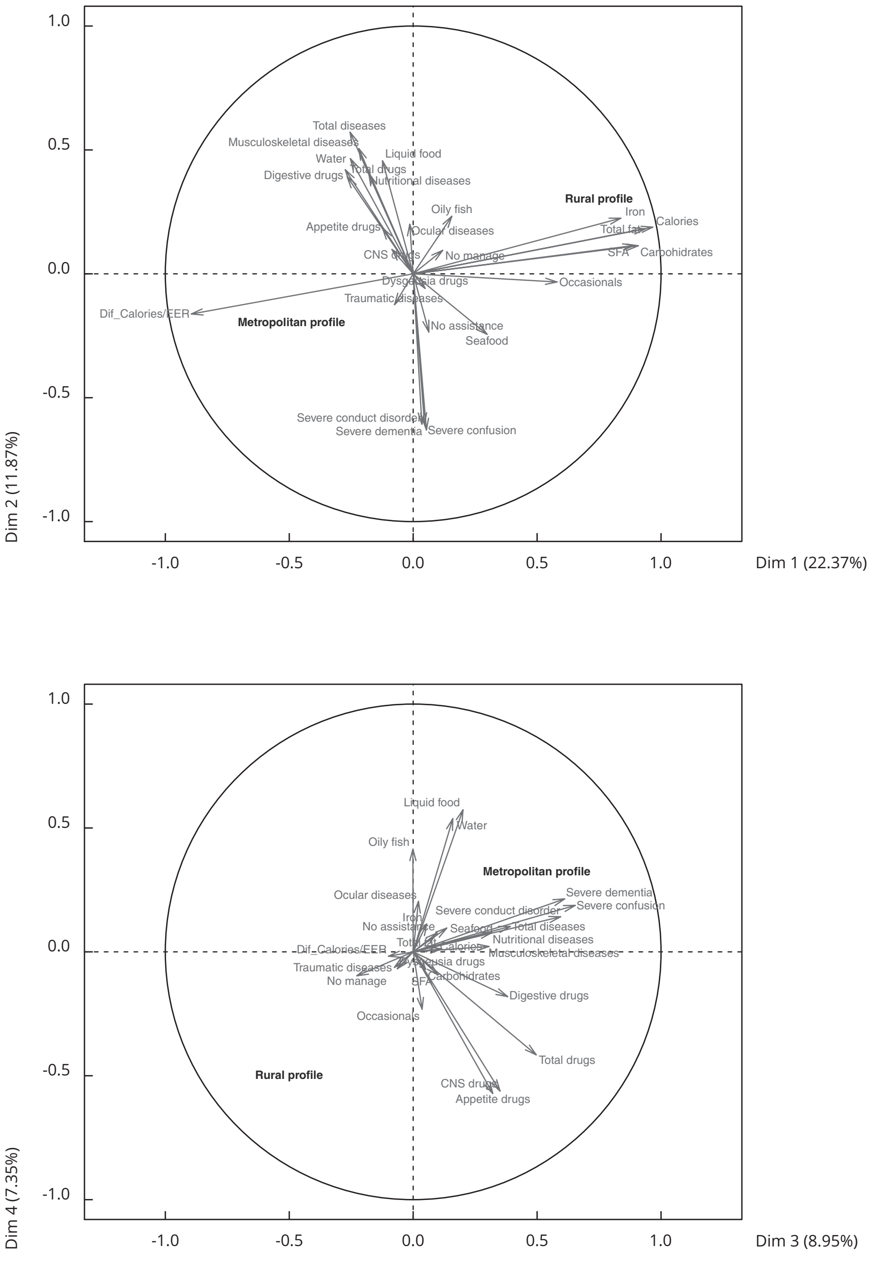
Implementation of variable selection procedures in a real dataset of a public health research

We implemented three different forms of variable selection methods according to their methodology. First, the classical ones, using GLM and a stepwise variable selection method. Next, we used Bayesian methods, using the INLA and a variable selection method based in the Bayes Factor. Finally, we implemented different penalized regressions based on the penalty procedure: Lasso, ALasso, and Elastic Net regressions. Table 3 shows that most the methods select the same ten variables (in bold): carbohydrates, water intake, occasional food, absence of home assistance, independence for managements, the severe conduct disorder, consumption of drugs that cause dysgeusia, total diseases, nutritional or endocrine diseases and traumatic diseases.

The prediction capacity of each variable selection method is assessed and shown in Table 4. As can be seen, all the different models predicted the response variable very well. Therefore, we must choose the simplest model with the highest success in prediction. In this sense, the GLM and the ALasso regression provided the best prediction results with a model in which only nine predictive variables were involved in both cases (seven were coincident). On the other hand, Bayesian automatic

Figure 1

Representation of the distribution of the predictor variables and the geographic factor (in bold), regarding the main components of the principal component analysis (PCA).



CNS: central nervous system; Dif_Calories: difference calories; DIM: dimension; EER: energy efficiency ratio; SFA: saturated fatty acids.

Table 2

Prediction table of the logistic regression model (Second Step).

Observed profile	Predicted profile		% sucess
	Metropolitan	Rural	
Metropolitan	136	10	96.1
Rural	15	48	75.8

Table 3

Coefficients of selected variables through the applied variable selection methods (Third Step).

Pre-selected variables (from PCA/MCA)	GLM (p-value < 0.05)	Stepwise	INLA with priors	Bayes Factor (BVS, HPM)	Lasso regression	Elastic Net regression	ALasso regression
Calories	-0.102	-0.014	-0.022				
Difference calories/EER							
Total fat							
Carbohydrates	0.839	0.049	0.087	0.009	0.004	0.005	0.004
Iron	0.915	0.836	1.268			0.018	
SFA	0.309	0.202	0.406				
Water intake		-0.047		-0.003	-0.019	-0.019	-0.018
Liquid food					-0.009	-0.017	
Oily fish							
Sea food							
Occasional food	0.073	0.064	0.084	0.005	0.036	0.045	0.034
Absence of home assistance	-4.406	-4.092	-4.911	-0.315	-1.922	-2.556	-2.060
Independence for managements	4.717	4.694	5.498	0.513	2.944	3.472	3.120
Severe confusion							
Severe dementia/ Alzheimer							
Severe conduct disorder		-2.325		-0.255	-1.471	-2.059	-1.190
Total drugs intake							
CNS drugs					-0.044	-0.128	
Drugs that increase the appetite							
Digestive drugs		0.533			0.217	0.395	
Drugs that cause dysgeusia	3.197	2.333	1.937		0.525	1.261	
Total diseases	-0.483	-0.505	-0.695	-0.052	-0.209	-0.330	-0.202
Nutritional diseases	1.109	1.183	1.303	0.074	0.124	0.471	0.026
Musculoskeletal diseases							
Ocular diseases						-0.063	
Traumatic diseases	2.952	2.992	1.767	0.222	1.309	2.012	0.516

ALasso: Adaptive Lasso; BVS: Bayesian variable selection; CNS: central nervous system; EER: energy efficiency ratio; GLM: Generalized Linear Model; HPM: highest probability model; INLA: Integrated Nested Laplace Approximation; Lasso: Least Absolute Shrinkage and Selection Operator; MCA: multiple correspondence analysis; PCA: principal component analysis; SFA: saturated fatty acids.

Table 4

Prediction table of the applied regressions and validation on the remaining 30% of the sample.

	Success rate				
	Metropolitan profile		Rural profile		Involved variables
	70% training set	30% validation set	70% training set	30% validation set	
Logistic regression	0.901	0.942	0.758	0.912	26
GLM	0.947	0.917	0.963	0.986	9
INLA with priors	0.947	0.923	0.954	0.951	11
Lasso	0.947	0.942	0.945	0.943	13
Elastic Net	0.934	0.942	0.953	0.947	15
ALasso	0.921	0.942	0.930	0.939	9

ALasso: Adaptive Lasso; GLM: Generalized Linear Model; INLA: Integrated Nested Laplace Approximation; Lasso: Least Absolute Shrinkage and Selection Operator.

selection of variables was the method that determined the least number of predictive variables in the final model, only eight (Table 3). These eight variables were also selected in the ALasso. The repetition of the procedure (50 times) showed no significant differences in the selected variables.

As shown in Table 3, in our Third Step, all these well-known methods reduced the number of selected variables considerably. Nevertheless, we checked if the results would be the same if we had applied these methods in a single step from the beginning, since the answer to this question justifies this article.

Justification of the three-step procedure in a real dataset of a public health research

The small selection of variables obtained, which also showed a homogeneity in the results in each of the methods used, comprise the most representative variables of the different scientific disciplines that affect the response factor and explain most of the variability of the data. This could not be achieved through a selection method directly applied on the complete dataset (p = 131, although considering that the categories of the respondents were incorporated to the analysis as dummy variables, the 176 predictive variables were reached).

We observed two undesirable consequences when the methods of variable selection were applied in a single step. The first consequence was that variables with less weight in the explanation of the variability were masked by those selected (from the expert's point of view, these omitted variables also affected the variable response when considered in isolation, this fact is widely supported in the scientific literature). The second consequence was that too many variables were selected. We tested this assumption by applying the different selection methods without the pre-selection implemented in each scientific subset. The complete results of this test can be consulted in http://pages.uv.es/malore2 and a summary of them are shown in Table 5. In this table, we compare the results obtained by applying the methods from the third step of our procedure directly in the complete data set (p = 176), with the results obtained through our three-steps procedure. As can be seen, only the ALasso selected variables of each subset due to the adaptive weights that were used for penalizing different coefficients in the L_1 penalty, as we explained above, despite selecting a greater number of variables, up to 40. All the other methods did not select any variable in some of the subsets and selected too many variables from other subsets.

Table 5

Comparison of the results between the variable selection methods directly applied on the complete dataset with the results obtained through our three-step procedure

	Number of selected variables on each subset					
	Dietary	**Nutrients**	**Pharmacologic**	**Pathologic**	**Disability (categories)**	**Total**
Stepwise on logistic regression	0	0	15	19	63	97
Stepwise on GLM	0	0	15	19	63	97
BVS	1	0	0	0	5	6
Lasso regression	0	0	0	0	10	10
Elastic Net regression	3	0	0	1	14	18
ALasso regression	8	13	1	5	13	40
Assessed three-step procedure	**1**	**2**	**1**	**3**	**3**	**10**

ALasso: Adaptive Lasso; BVS: Bayesian variable selection; GLM: Generalized Linear Model; Lasso: Least Absolute Shrinkage and Selection Operator.

Discussion

The three-step procedure for the selection of variables in large and heterogeneous data sets in public health proposed in this article, which respects the causality explained by each discipline, obtained more balanced results, a greater reduction of variables and a better prediction capacity of resulting models than other methods of variable selection applied to the dataset in a single step.

We found no research in the area of public health that considered this multidisciplinary perspective in the selection of variables, however, models that assume that the observations come from a heterogeneous population, which is a mixture of a finite number of sub-populations, have been assessed from a Bayesian point of view with a successful results [31]. Several studies have developed new metrics or algorithms to improve variable selection [4,32,33], but most of the research is based on the improvement of the existing statistical methods, often applied in real data sets [19,34].

Studies that resemble what we propose are those that evaluate the different variable selection methods by comparison to determine which is the most appropriate to analyse their data [35,36,37], or those that apply the existing methods in a specific discipline to identify the most representative variables to focus their research [38]. However, all the studies mentioned above are similar since they implement the methods of variable selection in a single step and tend to use a Bayesian approach [7,20,31,32,33,35,36,39] and/or penalized regressions [4,7,20,38] to achieve their objectives.

The variable selection strategy we propose is particularly promising, since it not only improves the prediction accuracy but also provides a clear interpretation of the most informative variables [7]. Through this method, variable analysis in public health research can be improved by selecting only the variables that strongly affect the response variable and focus on it [6]. This may optimize the process and save resources to the development of new projects, especially those based on a multidisciplinary point of view involving several potential variables.

Contributors

M. Lozano collected and processed the data, he also wrote the drafts. L. Manyes processed the data and revised the methodology. J. Peiró performed the statistical tests and revised the drafts. A. Iftimi collaborated with the statistical works and revised the English language of the drafts. J. M. Ramada led the research and revised all the drafts.

Conflict of interests

The authors declare that they have no competing interests.

References

1. Commission on Social Determinants of Health. CSDH final report: closing the gap in a generation: health equity through action on the social determinants of health. Geneva: World Health Organization; 2008.
2. Greve B, Pigeot I, Huybrechts I, Pala V, Börnhorst C. A comparison of heuristic and model-based clustering methods for dietary pattern analysis. Public Health Nutr 2016; 19:255-64.
3. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol 2008; 70:849-911.
4. Wang ZX, He QP, Wang J. Comparison of variable selection methods for PLS-based soft sensor modeling. J Process Control 2015; 26:56-72.
5. Lutomski JE, van den Broeck J, Harrington J, Shiely F, Perry IJ. Sociodemographic, lifestyle, mental health and dietary factors associated with direction of misreporting of energy intake. Public Health Nutr 2011; 14:532-41.
6. Peng W, Goldsmith R, Berry EM. Demographic and lifestyle factors associated with adherence to the Mediterranean diet in relation to overweight/obesity among Israeli adolescents: findings from the Mabat Israeli national youth health and nutrition survey. Public Health Nutr 2017; 20:883-92.
7. Chen T, Martin E. Bayesian linear regression and variable selection for spectroscopic calibration. Anal Chim Acta 2009; 631:13-21.
8. Berrendero JR, Cuevas A, Torrecilla JL. The mRMR variable selection method: a comparative study for functional data. J Stat Comput Simul 2015; 86:891-907.
9. Jadhav NH, Kashid DN, Kulkarni SR. Subset selection in multiple linear regression in the presence of outlier and multicollinearity. Stat Methodol 2014; 19:44-59.
10. Shahriari S, Faria S, Gonçąlves AM. Variable selection methods in high-dimensional regression-a simulation study. Commun Stat Simul Comput 2015; 44:2548-61.
11. Brusco MJ. Clustering binary data in the presence of masking variables. Psychol Methods 2004; 9:510-23.
12. Peña D. Análisis de datos multivariantes. Madrid: McGraw-Hill Interamerica de España; 2002.
13 Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B Stat Methodol 2009; 71:319-92.
14. Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Mach Learn 2003; 50:5-43.
15. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol 2002; 64:583-616.

16. García-Donato G, Martínez-Beneito MA. On sampling strategies in Bayesian variable selection problems with large model spaces. J Am Stat Assoc 2013; 108:340-52.
17. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 1984; PAMI-6:721-41.
18. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics 1970; 12:69-82.
19. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Ser B Stat Methodol 2011; 73:273-82.
20. Lykou A, Ntzoufras I. On Bayesian lasso variable selection and the specification of the shrinkage parameter. Stat Comput 2013; 23:361-90.
21. Fu WJ. Penalized regressions: the bridge versus the lasso? J Comput Graph Stat 1998; 7:397-416.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005; 67:301-20.
23. Fernández-Ballart JD, Piñol JL, Zazpe I, Corella D, Carrasco P, Toledo E, et al. Relative validity of a semi-quantitative food-frequency questionnaire in an elderly Mediterranean population of Spain. Br J Nutr 2010; 103:1808-16.
24. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006; 101:1418-29.
25. Abellán A, Esparza C, Castejón P, Pérez J. Epidemiology of disability and dependency in old age in Spain. Gac Sanit 2011; 25 Suppl 2:5-11.
26. Sánchez-Rodríguez MA, Santiago E, Arronte-Rosales A, Vargas-Guadarrama LA, Mendoza-Núñez VM. Relationship between oxidative stress and cognitive impairment in the elderly of rural vs. urban communities. Life Sci 2006; 78:1682-7.
27. Böell JEW, da Silva DMGV, Hegadoren KM. Sociodemographic factors and health conditions associated with the resilience of people with chronic diseases: a cross sectional study. Rev Latinoam Enferm (Online) 2016; 24:e2786.
28. Irz X, Fratiglioni L, Kuosmanen N, Mazzocchi M, Modugno L, Nocella G, et al. Sociodemographic determinants of diet quality of the EU elderly: a comparative analysis in four countries. Public Health Nutr 2014; 17:1177-89.
29. Öztürk A, Şimşek TT, Yümin ET, Sertel M, Yümin M. The relationship between physical, functional capacity and quality of life (QoL) among elderly people with a chronic disease. Arch Gerontol Geriatr 2011; 53:278-83.
30. Bamia C, Trichopoulos D, Ferrari P, Overvad K, Bjerregaard L, Tjønneland A, et al. Dietary patterns and survival of older Europeans: The EPIC-Elderly Study (European Prospective Investigation into Cancer and Nutrition). Public Health Nutr 2007; 10:590-8.
31. Lee K, Chen R, Wu YN. Bayesian variable selection for finite mixture model of linear regressions. Comput Stat Data Anal 2016; 95:1-16.
32. Chen Z, Tang M, Gao W, Shi N. New robust variable selection methods for linear regression models. Scand J Stat 2014; 41:725-41.
33. Ordonez C, Garcia-Alvarado C, Baladandayuthapani V. Bayesian variable selection in linear regression in one pass for large datasets. ACM Trans Knowl Discov Data 2014; 9:1-14.
34. Lin L, Wang Q, Sadek AW. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transp Res Part C Emerg Technol 2015; 55:444-59.
35. Ju H, Brasier AR, Kurosky A, Xu B, Reyes VE, Graham DY. Diagnostics for statistical variable selection methods for prediction of peptic ulcer disease in *Helicobacter pylori* infection. J Proteomics Bioinform 2014; 7:95-101.
36. Rentsch C, Bebu I, Guest JL, Rimland D, Agan BK, Marconi V. Combining epidemiologic and biostatistical tools to enhance variable selection in HIV cohort analyses. PLoS One 2014; 9:e87352.
37. Kujala M, Nevalainen J. A case study of normalization, missing data and variable selection methods in lipidomics. Stat Med 2015; 34:59-73.
38. Berger S, Pérez-Rodríguez P, Veturi Y, Simianer H, de los Campos G. Effectiveness of shrinkage and variable selection methods for the prediction of complex human traits using data from distantly related individuals. Ann Hum Genet 2015; 79:122-35.
39. Healy BC, Engler D. Modeling disease-state transition heterogeneity through Bayesian variable selection. Stat Med 2009; 28:1353-68.

Resumen

*La investigación multidisciplinaria en salud pública se enfoca usando métodos de muchas disciplinas científicas. Una de las principales características de este tipo de investigación es lidiar con conjuntos voluminosos de datos. Los métodos clásicos estadísticos de selección de variables, conocidos como "*screen and clean*", y utilizados en un solo paso, seleccionan las variables con mayor peso explicativo en su modelo. Estos métodos, comúnmente usados en investigación pública en salud, pueden inducir a enmascarar la multicolinealidad, excluyendo variables relevantes para los expertos en cada disciplina y sesgando el resultado. Se usan algunas técnicas específicas para resolver este problema, como las regresiones penalizadas y estadísticas bayesianas, que ofrecen resultados más equilibrados entre subconjuntos de variables, pero con umbrales menos restrictivos de selección. Usando la combinación de métodos clásicos, se propone en este trabajo un tercer paso en el procedimiento, recogiendo variables relevantes de cada disciplina científica, minimizando la selección de variables en cada una de ellas y obteniendo una distribución equilibrada que explica la mayor parte de la variabilidad. Este procedimiento fue aplicado en un conjunto de datos de una investigación en salud pública. Comparando los resultados con los métodos de un solo paso, el método propuesto expone una gran reducción en el número de variables, así como la distribución equilibrada entre las disciplinas científicas asociadas con la variable de respuesta. Proponemos un procedimiento innovador para la selección de variables y aplicarlo a nuestro conjunto de datos. Asimismo, comparamos el nuevo método con los procedimientos clásicos de un solo paso.*

Estadística como Asunto; Métodos; Investigación Interdisciplinaria

Resumo

*A pesquisa multidisciplinar em saúde pública emprega métodos provenientes de diversas disciplinas científicas. Uma das principais características desse tipo de pesquisa é o fato de lidar com conjuntos de dados grandes. Os métodos clássicos de seleção de variáveis estatísticas, conhecidos como "*screen and clean*" (filtrar e limpar), e aplicados a partir de um passo único, selecionam as variáveis com o maior peso explanatório no modelo. Esses métodos, amplamente disseminados na pesquisa em saúde pública, podem induzir ao mascaramento e à multi-colinearidade, excluindo variáveis que seriam relevantes para os especialistas em cada disciplina e enviesando os resultados. Algumas técnicas específicas usadas para resolver esse problema, como regressões penalizadas e estatísticas Bayesianas, oferecem resultados mais equilibrados entre subconjuntos de variáveis, porém com limiares de seleção menos restritivos. O artigo propõe um procedimento com três passos, usando uma combinação de métodos clássicos, captando as variáveis relevantes de cada disciplina científica, minimizando a seleção de variáveis em cada disciplina e obtendo uma distribuição equilibrada que explica a maior parte da variabilidade. O procedimento foi aplicado a um conjunto de dados de uma pesquisa em saúde pública. Ao comparar os resultados com os métodos que utilizam um único passo, o método proposto demonstra maior redução no número de variáveis, assim como, uma distribuição equilibrada entre as disciplinas científicas relacionadas à variável dependente. Propomos um procedimento inovador para a seleção de variáveis, que aplicamos depois ao nosso conjunto de dados. Além disso, comparamos o método novo com os procedimentos clássicos de apenas um estágio.*

Estatística como Assunto; Métodos; Pesquisa Interdisciplinar

Submitted on 05/Oct/2017
Final version resubmitted on 16/Apr/2018
Approved on 03/May/2018