A Review on Variable Selection in Regression Analysis

Loann David Denis Desboulets

Review

A Review on Variable Selection in Regression Analysis

Loann. D. Desboulets [1,†,‡]

[1] Aix-Marseille University, CNRS, EHESS, Centrale Marseille, AMSE; loann.DESBOULETS@univ-amu.fr
† Current address: 5-9 Boulevard Maurice Bourdet, 13001 Marseille

Version May 31, 2018 submitted to Econometrics

Abstract: In this paper, we investigate on 39 Variable Selection procedures to give an overview of the existing literature for practitioners. "Let the data speak for themselves" has become the motto of many applied researchers since the amount of data has significantly grew. Automatic model selection have been raised by the search for data-driven theories for quite a long time now. However while great extensions have been made on the theoretical side still basic procedures are used in most empirical work, eg. Stepwise Regression. Some reviews are already available in the literature for variable selection, but always focus on a specific topic like linear regression, groups of variables or smoothly varying coefficients. Here we provide a review of main methods and state-of-the art extensions as well as a topology of them over a wide range of model structures (linear, grouped, additive, partially linear and non-parametric). We provide explanations for which methods to use for different model purposes and what are key differences among them. We also review two methods for improving variable selection in the general sense.

Keywords: Variable Selection; Automatic Modelling; Sparse Models

JEL Classification: C50,C59

---

1. Introduction

When building a statistical model the question of which variables to include arise very often. In practice is it true almost all the time. This can come from ignorance, competing theories, or whatever. Practitioners have now at their disposal a wide range of technologies to solve this issue. Literature on this topic started with Stepwise Regression (Breaux 1967) and Autometrics (Hendry et al. 1987), moving to more advanced procedures from which the most famous are the Non Negative Garrotte (Breiman 1995), the Least Angle and Shrinkage Selection Operator (hereafter LASSO, Tibshirani (1996)) and the Sure Independence Screening (Fan and Zhang 2008). Many papers are available for empiricists to get an overview of the existing methods. Fan and Lv (2010) reviews most of the literature on linear and generalized models. A large part is devoted to penalized methods and algorithmic solutions, also the optimal choice of the parameter penalty is discussed. Breheny and Huang (2009) and Huang et al. (2012) gave a complete review of selection procedures in grouped variables models with great technical comparisons, especially in terms of rate of convergence. Castle et al. (2011) compared Autometrics to a wide range of other methods (Stepwise, Akaike Information Criterion, LASSO, etc. [1]) in terms of prediction accuracy under orthogonality of the regressors, with a particular attention given to dynamic models. In the same spirit as our paper Park et al. (2015) gave a very recent review of variable selection procedures but dealing only with varying-coefficient models. Fan and Lv (2017) provided a comprehensive review in the context of Sure Independence Screening major improvements. We can also cite more focusing papers like Fu (1998) who compared the Bridge and the LASSO theoretically but also empirically both through simulation and real data.

[1] Some of them are not presented in this paper either because they are out of its scope, eg. bayesian framework, or because they are special cases of other ones.

Epprecht et al. (2013) that compared Autometrics and the LASSO according to prediction and selection accuracy.

The contribution of this paper is threefold. First, 39 procedures are considered, these are listed and clearly classified. Secondly, we establish a topology of procedures under different model structures. We consider major ones: Linear Models, Grouped Variables Models, Additive Models, Partial Linear Models and Non-parametric Models. Thirdly, we describe and compare state-of-the-art papers in the literature. We give contexts where each procedure should be used, to which specific problem they answer and compare them on this ground. In this sense any practitioner with enough knowledge in Statistics can refer to our paper as a methodological guide for doing variable selection. It gives a wider view of existing technologies than the other reviews we mentioned.

The paper is organized as follows, in section 2 we introduce the three main categories of variable selection procedures and we provide a typological table of these ones on the ground of model structures. Descriptions as well as comparisons are discussed along the sections 3 to 7. Each of these sections focuses on a particular model structure. Section 8 is devoted to two general methods for improving model selection, both can be applied on all the procedures presented accross the paper. In the final section we make few critics on actual procedures and give insight on future area of research.

2. Typology of Procedures

In this section we propose a typology of state-of-the-art selection methods in many different frameworks, there are many types of models that can be considered. For this aim, Table 1 provides the classification of statistical methods available in the literature and that will be discussed in the paper. From the latter we have determined 3 main categories of algorithms:

- Tests-based
- Penalty-based
- Screening-based

Originally, the first developed are based on statistical tests. The work was to automate standard tests in Econometrics (like testing residuals for normality, t-tests etc.) for choosing among candidate variables. It includes Stepwise Regression and Autometrics for example.
Then there are Penalty-based procedures. Imposing a constraint on parameters directly inside estimation encourages sparsity among them. For instance LASSO and Ridge belongs to this category.
The lasts are Screening procedures, they are not all designed to do selection intrinsically but rather ranking variables by importance. The main advantage is that it applies more easily to very large dimensional problems, when number of regressors is diverging with the number of observations (eg. cases with $p >> n$). This is mainly true because it considers additive models (linear or not) and so variables can be treated separately.
We discuss this distinction more deeply in subsections below and give brief description of their main features.

2.1. Tests-Based

This is the most classical way of handling variable selection in statistical models. It was also the first attempt of variable selection. Everything started with Stepwise Regression (Breaux 1967), one of the latest of this kind is Autometrics (Hendry et al. 1987) [2]. We focus on Stepwise Regression and Autometrics for two reasons. The first is that Stepwise Regression is the most well-known and the most widespread method for choosing variables in a model. Despite it dates back to 1967 many empiricists still practice it. The second is that Autometrics has integrated many features of Econometrics to achieve the highest degree of completeness for an automatic procedure. Authors have considered endogeneity, non-linearities, unit-roots and many others, trying to overcome most issues a statistician can face.
Stepwise Regression is the most simple and most straightforward way of doing model selection by just retrieving

[2] Even though it started in 1987 there are still improvements nowadays.

insignificant variables (backward approach) or adding significant ones (forward approach) based on some statistical criterion. Therefore it is pretty easy to use it empirically because implementation is straightforward. However in several situations this does not ensure consistent selection. Its selection properties have been investigated in Wang (2009). On the contrary Autometrics is a complete philosophy of modelling, but comes at the cost of a quite complex algorithm and many tuning parameters are required, making its use more difficult for non-expert.

2.2. Penalty-Based

Thanks to the work of Tibshirani (1996) it became a quite common strategy in empirics. This kind of methods involves applying a penalty on parameters to encourage sparsity (i.e. some are set exactly to zero). Sparsity is a necessary condition for situations of unidentifiability ie. where $p > n$. Such a problem can be solved using penalties on parameters to make inference possible. These parameters can come from parametric models or from non-parametric models, so penalty based method can be applied on both structures. This kind of procedure started with the Non Negative Garrote (NNG of Breiman (1995)) in an Ordinary Least Squares framework up now to much complex model structure like varying coefficients and two-ways interactions ANOVA non-parametric models. The idea of producing sparse models is a convenient way of integrating a test inside the estimation. Inference of such models requires the prior assumption that some variables are not relevant, this is the test part, and penalty-based methods helps estimating the coefficients, this is the inference part. So both procedures are merged in an unified framework giving rise to a novel conception of statistical modelling. Maybe the most famous in this category is the LASSO of Tibshirani (1996).

2.3. Screening-Based

Screening is actually the most effective way of dealing with very high dimensional features (large p). Few other selection methods can be as computationally efficient as these ones. However Screening often does not perform model selection itself, it rather ranks variables. To do so they have to be mixed up with other procedures, in the literature they are mainly penalty-based. Even if it does not select variables reducing the candidate set is an important aspect of the variable selection and screening methods are powerful in this task. In this respect it is worth mentioning the Sure Independence Screening (hereafter SIS, Fan and Lv (2008)) that is the first of this kind.

Screening makes the use of a ranking measure, either linear or not so it can be applied in both frameworks. Some may rely on specific models (like a linear model) while others are model-free. The major differences among procedures in this category relies on the choice of the ranking measure. Correlation coefficients are the first coming to mind, these are mainly used. One limitation in screening is that they usually treats variables by pairs to compute their measure of association, so every effect is considered as additive and does not correct for the presence of interactions effects. This is not necessarily true, especially in the non-parametric settings. Sophisticated correlations such as distance correlation or canonical kernel correlation are employed in a multivariate framework and account for such interactions even if they do not model them explicitly. However in this case they may loose their computational efficiency compared to independence screening ones. As said before, a brief review of some SIS methods can be found in Fan and Lv (2017).

Table 1. Topology of Variable Selection Methods

	Screening	Penalty	Testing
Linear	SIS SFR CASE FA-CAR	SparseStep LASSO Ridge BRidge SCAD MCP NNG SHIM	Stepwise Autometrics
Group		gLASSO gBridge gSCAD gMCP ElasticNet	
Additive	NIS CR-SIS	SpAM penGAM	
Partial Linear		kernelLASSO adaSVC DPLSE PSA PEPS	SP-GLRT
Non-Parametric	DC-SIS HSIC-SIS KCCA-SIS Gcorr MDI MDA RODEO	VANISH COSSO	MARS

3. Linear Models

We began with the first model structure: the linear model. It is described as:

$$y = X\beta + \varepsilon \tag{1}$$

The variable to be explained y (sometimes also called the output, the response or the dependent variable) is a one dimensional vector of length n, corresponding to the number of observations. The matrix X contains the explanatory variables (sometimes also called the inputs, the regressors or the independent variables) of length n and dimension p which is the number of candidate variables. Therefore the one dimensional vector β of length k contains the parameters of interest. The residuals (sometimes called also the error term) are denoted ε, even if it could be of interest we do not solely focus on their properties and consequences on variable selection in this paper. This notation will be held constant throughout the paper. Notice that all of the three methodologies are able to handle linear models, while this is not necessarily true for other structures (e.g. additive models).

3.1. Testing

Stepwise Regression (Breaux 1967) is the first model selection procedure. This approach have been motivated when statisticians started to consider model uncertainty. This means that among p variables we can possibly construct 2^p models, so we should maybe take them all into account. To test all possibilities we have to compute "all-subsets". This cannot be achieved for large p. In order to overcome this problem and reduce the search, stepwise regression investigates only a subset of all possible regressions with the hope to end with the true model. There exist two approaches: Backward and Forward. Either the process starts from a null model (only an intercept) and introduces variables one by one, this is the forward step. Or it starts from the full model (all variables) and deletes them one by one, this is the backward step. One improvement is also to

consider both. Usually the selection within each step is made according to some criterion. One consider all one-variable increments from the actual model and choose the best move according to this criterion, it might be the lowest p-value, highest adjusted R^2, lowest Mallow's C_p, lowest AIC, lowest prediction error, leave-one-out cross validation, etc.
You can imagine any criterion to perform this job, but the main issue arising from Stepwise Regression does not come from the choice of the criterion. Interesting critics (Doornik 2009) arise from the developers of Autometrics. The main one is the lack of search. Stepwise regression proceeds step by step along a single path. Then, there is no backtesting. That is the procedure never considers testing again variables that have been removed after each step. Such an idea is present using the forward backward combination but it is restricted to the previous step only. Obviously they are not the only one to express admonitions about Stepwise Regression. We can mention many papers from Hurvich and Tsai (1990), Steyerberg et al. (1999), Whittingham et al. (2006) or Flom and Cassell (2007) where they all prove biased estimation and inconsistent selection of Stepwise Regression.
However even if used as a selection method it behaves poorly, used a screening method it showed better results. This has been developed by Wang (2009) and is detailed in the next subsection.

On the other side is Autometrics, an algorithm for model selection developed by Hendry et al. (1987) under the famous Theory of Encompassing and the LSE (London School of Economics) type of Econometrics. This method has been created as early as 1987 and is still under development. The basis of its methodology is the General-to-Specific approach. Theory of Encompassing states that the researcher should start from a very large model (called the GUM: General Unrestricted Model) encompassing all other possible models and then reduce it to a simpler but congruent specification. This idea is somehow related to the backward specification in Stepwise Regression. His work is an automation of standard way of testing for relevance in Econometrics such as t-tests and F-tests and major concerns deal with power of tests, repeated testing but also outliers detection, non-linearities, high dimensional features (with $p > n$) and parameter invariance.
Tests come with some hyperparameters specifying the size of the battery of tests (t-tests, F-tests, normality checks, etc.).
Repeated testing occurs when a variable that has been deleted under a certain specification that has now changed is reintroduced and tested again. The absence of such a thing in Stepwise Regression is a severe drawback and the main reason why it fails pretty often.
Non-linearities are handled using Principal Components Analysis (see Castle and Hendry (2010)) that makes the design matrix orthogonal. Such a decomposition allows to introduce squares and cubics of the transformed variables which are linear combination of the original ones. Orthogonality limits the number of non-linear terms since it already accounts for interaction using components. In simple words a polynomial of degree d with p variable results in $\binom{d+p}{d} - 1$ terms, while their methods reduces to $d \times p$ which is very much less. It is advocated that it can reproduce non-linear functions often met in Economics and Social Sciences. However the class of functions that it can reproduce may be restricted compared to standard non-parametric methods [3].
High dimensional features and non-identifiability ($p > n$) of the GUM is solved in a very simple way called "Block Search". Regressors are divided in different blocks until the size of each block is lower than p. Then tests are applied in each block, some variables are discarded, the remaining blocks are merged and the process continues. This idea is based on the fact that the methodology is still consistent under separability. This idea is quite similar to the Split-and-Conquer methodology of Chen and Xie (2014) to solve ultra-high dimensional problems.
Outliers can be detected using the Impulse Indicator Saturated Selection (IIS) developed by Hendry et al. (2006). This is in the same spirit as the Block-Search (or Split-and-Conquer) approach defined previously. A set of indicator is added to the GUM for every observation, and tests are applied in a Block-Search manner to remove observations that are not consistent with the model, identified as outliers.

[3] This should be investigated more deeply, to the best of our knowledge no papers have tried to compare their non-linear regression to the very well-known non-parametric procedures like Kernels or Splines. An obvious link can be made with Projection Pursuit Regression (PPR), in this respect we claim that Autometrics may be a special case of PPR.

Stepwise Regression and Autometrics are serial procedures where selection and estimation are performed sequentially. In some sense Penalty-Based methods aim at performing both at the same time. One can view penalty-based procedures as the direct implementation of tests inside inference.

3.2. Penalty

Penalty-based methods can be divided in two categories: penalties on the norm and concave ones. The shape of the penalty may have a great influence on the selected set of variables and their estimates. Sparse model is achieved because we reduce nearly zero coefficients to zero in estimation. The penalty parameter plays the role of a threshold but in a non-orthogonal framework. To understand better the origins of these penalty one should refer to threshold methods in Kowalski (2014). For that reason the penalty also introduces shrinkage of the coefficients, making them biased. The literature is focused on the choice of the penalty in terms of selection consistency and bias properties.

3.2.1. Norm Penalties

There are almost as many methods as there are norms, but generally the objective is to solve:

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_\gamma^2 \tag{2}$$

Each methods applies to different L_γ norms.

- SparseStep: $\gamma = 0$
- LASSO : $\gamma = 1$
- Ridge : $\gamma = 2$

This methodology is gathered in the more general Bridge estimator (Frank and Friedman 1993) that considers any value for γ, but the authors did not say how to solve the problem. The advantage of Ridge (Hoerl and Kennard 1970) is that it has an analytical solution. However the solution is not sparse so it does not select variables (only shrinkage). The Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani (1996)) does because the L_1 norm is singular at the origin. However both give bias estimates because they apply shrinkage to the coefficients. The zero norm used in SparseStep (van den Burg et al. 2017) is the counting norm, they penalize directly the number of non-zero elements in β, not their values (no shrinkage). Usually constraints on the number of non-zero elements require high computational costs (exhaustive search over the model space). Here they use an easy even though very precise continuous approximation from de Rooi and Eilers (2011) and that turns the problem into something computationally tractable.
Meinshausen and Bühlmann (2006) shown that LASSO tends to select noise variables using a penalty parameter optimally chosen for prediction. For this reason Zou (2006) developed AdaLASSO (Adaptive LASSO). His paper proved that the optimal estimation rate is not compatible with consistent selection. Moreover even sacrificing the estimation rate does not ensure that the LASSO will select the right variables with positive probability. This phenomenon is highlighted through a necessary condition on the covariance matrix of the regressors that cannot always be satisfied using the LASSO with a single penalty parameter. Therefore he introduced adaptive weights to the LASSO to make it consistent with variable selection.

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|w\beta\|_1 \tag{3}$$

The latest improvement on linear models is to allow for interactions terms. Even if it is possible, only adding them into a LASSO is not an efficient procedure because it greatly extends the dimensionality of the design matrix. The idea of the Strong Heredity Interaction Model (SHIM,Choi et al. (2010)) is to add interactions only if main effects are selected also (strong heredity property), this greatly reduces the search space and provides an efficient way of doing ANOVA-types of models. They consider a reparametrization of the two-ways interactions models:

$$y = X\beta + \sum_{j=1}^{p} \sum_{k \neq j} \gamma_{jk} \beta_j \beta_k x_j x_k \tag{4}$$

Introducing main effect parameters β on top of cross-effects γ ensures that the interaction will be non-zero if and only if both main effects are non-zeros. The problem is a composite LASSO of the following form:

$$\min_{\beta,\gamma} \|y - X\beta\|_2^2 + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1 \tag{5}$$

Solutions to these problems are numerous. Usually either it reduces to the LASSO and then algorithms like Least Angle Regression (LARS) of Efron et al. (2004) are employed. Otherwise iterative algorithms like the Local Quadratic Approximation Fan and Li (2001) can be used.

3.2.2. Concave Penalties

Norm penalties are very standard and easy to work with but there exists also other types of penalties. Thus we can consider penalties in a very general framework:

$$\min_{\beta} \|y - X\beta\|_2^2 + p_\lambda(\beta) \tag{6}$$

The difference will then lie in the choice of $p_\lambda(\beta)$.

- NonNegativeGarotte:

$$p_\lambda(\beta) = n\lambda \sum_{j=1}^{p} \left(1 - \frac{\lambda}{\beta_j^2}\right)_+ \tag{7}$$

- SCAD :

$$p_\lambda(\beta) = \begin{cases} \lambda, & \text{if } |\beta| \leq \lambda \\ \frac{a\lambda - |\beta|}{a-1}, & \text{if } \lambda < |\beta| < a\lambda \\ 0, & \text{if } |\beta| \geq a\lambda \end{cases} \tag{8}$$

- MCP :

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| - \frac{x^2}{2\gamma}, & \text{if } |\beta| \leq \gamma\lambda \\ 0.5\gamma\lambda^2, & \text{if } |\beta| > \gamma\lambda \end{cases} \tag{9}$$

The Non Negative Garotte (Breiman 1995) was the first penalty of this kind, but because it has bad properties (especially variables selection inconsistency) it was rapidly abandoned. SCAD (Smoothly Clipped Absolute Deviation,Fan and Li (2001)) was the first penalty method that was consistent, continuous and unbiased for large values of β. MCP (Minimax Convex penalty,Zhang et al. (2010)) has little difference with SCAD in terms of selected variables. A comparative study between them can be found in Zhang (2007). One thing with penalty method is that there are always some penalty parameters (eg λ in LASSO) that have to be chosen. Usually they are set to optimal values according to some General Cross Validation (GCV) criterion or out-of-sample predictions. This is crucial because results can be very sensitive to the choice of these parameters. SCAD is more robust to this problem thanks to a bias-free property.[4]

3.3. Screening

Another methodology in variable selection is Screening. In fact these are ranking methods that rely on some association measure between the dependent variable and the regressors. Very often this measure is taken to be bivariate allowing then an extremely fast analysis.

[4] This is true only for large values of parameters, the reader can get intuitions of this phenomenon with threshold methods (Kowalski 2014).

3.3.1. Regressor Based

The Sure Independence Screening (SIS,Fan and Lv (2008)) is the first of this kind and almost all methods are derived from it. It uses simple correlation on standardized variables : $\hat{\omega}(x_j, y) = \tilde{x}_j\tilde{y}$ and gives a ranking of the x_j. The set $\widehat{M}$ of relevant features is determined by a simple threshold:

$$\widehat{M} = \{1 \leq j \leq p : |\hat{\omega}(x_j, y)| \text{ is among the top } d \text{ largest ones}\} \tag{10}$$

This set is reduced step by step until some moment. The method in itself does not select anything in fact, it just remove the less correlated features from the set of candidates, but we are left with a candidate set where selection has to apply. SIS needs a selection procedure in the end to obtain consistent results. The main advantage of the method is that when the number of variables p is very large compared to the number of observations n usual selection procedures tend to misbehave (Fan and Lv 2008). In their paper SIS has proven to lead to a set of candidates that is manageable for LASSO and others in order to have good properties. SIS allows for ultrahigh dimensional features, ultrahigh being defined as: $\log(p) = \mathcal{O}(n^{\alpha})$ with $0 \leq \alpha \leq 1$.
In this respect the screening properties of screening of Forward Regression (Wang 2009) have been investigated and with little improvements proved to be consistent in variables selection. However it still requires a selection procedure in the end, Forward Regression is just used for the screening part that is ranking and reducing the set of candidates.
Because SIS may encounters issue for selecting weakly correlated variables (weak signal-to-noise ratio) Fan and Lv (2008) introduced Iterative conditional SIS that is applying correlation ranking but conditional on selected features. This is equivalent as looking through correlation between features and residuals from a model using primarily selected variables instead of correlation with the dependent variable. This idea can be related to former algorithms that were developed to infer the LASSO (eg. Forward Stagewise).

3.3.2. Covariance Based

The last approach is less common. The Covariate Assisted Screening Estimates (CASE, Ke et al. (2014)) is a method that looks for sparse models but in the case where signals are rare and weak. All methods presented so far work well if β is sparse (so rare) and has high values (strong signals). In this case methods like SCAD are even bias-free. But if the signals are weak on top of rare then they won't manage to perform variable selection very well. The idea in CASE is to sparsify the covariance matrix of the regressors using a linear filter and then look for models inside this sparse covariance matrix using tests and penalties. Drawbacks are the choice of the filter that is problem dependent and the power of the tests.
To improve on CASE when regressors are highly correlated, giving a very dense covariance structure, Factor Adjusted-Covariate Assisted Ranking (FA-CAR, Ke and Yang (2017)) proposes using PCA to sparsify it. This is in line with selecting appropriately the filter in CASE when the problem to solve includes strong collinearity. In fact the covariance is assumed to have a sparse structure, hidden by latent variables. These are estimated by PCA and then removed from the variables. The process does not change anything for the equation and the parameters to be estimated does not require more technology than the simple OLS on the transformed decorrelated variables. The main issue is to select the number of latent variables to be removed, this can be done via cross-validation for instance, still it remains difficult.

4. Grouped Models

Depending on the application the model can come in a group structure form of the type:

$$y = \sum_{g=1}^{G} \sum_{j \in g} \beta_j x_j + \varepsilon \tag{11}$$

which can be rewritten in matrix-grouped notation:

$$y = \sum_{g=1}^{G} X_g \beta_g + \varepsilon \tag{12}$$

Within this framework there are 2 main possibilities. One can look for which group to be selected or which variable is more relevant in which group. The former is referred to as single-level selection (sparse between group estimates) and the latter as bi-level selection (sparse between and within group estimates). Technical reviews of selection procedures with grouped variables can be found in Breheny and Huang (2009) and Huang et al. (2012).

4.1. Penalty

4.1.1. Single-level

The concept of group-penalty was introduced in Yuan and Lin (2006) (groupLASSO) in a LASSO framework. The objective is to solve a modified LASSO:

$$\min_{\beta} \|y - \sum_{g=1}^{G} \sum_{j \in g} \beta_j x_j\|_2^2 + \lambda \sum_{g=1}^{G} c_g (\beta_g' R_g \beta_g)^{1/2} \tag{13}$$

The parameters c_g are used to adjust for the group sizes in order to have selection consistency. The parameter λ controls for the penalty. The choice of R_g that weights each coefficients within the group is still challenging. A solution is to take $R_g = (X_g' X_g)/n$ the Gram matrix of the grouped variables X_g. The effect is to scale the variables within groups and so make coefficients comparable in some sense. It can be easily shown that this lead to the LASSO solution with standardization of regressors when the group is formed with only one variable, such a thing is made pretty often empirically and is even advised by the LASSO's authors.
An obvious extension is to take into account any penalty, providing the following objective:

$$\min_{\beta} \|y - \sum_{g=1}^{G} \sum_{j \in g} \beta_j x_j\|_2^2 + p(\sum_{g=1}^{G} \|\beta_g\|_{R_g}; c_g \lambda, \gamma) \tag{14}$$

Where $p(\cdot)$ can be taken to be the Bridge, the SCAD or the MCP criterion introducing then the groupBridge (Huang et al. 2009), the groupSCAD Wang et al. (2007) and the groupMCP (Breheny and Huang 2009) respectively.

4.1.2. Bi-level

Improvements have been made on norm penalties by considering mixed norms like the ElasticNet (Zou and Hastie 2005):

$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \tag{15}$$

This method overcomes the issue of collinearity because it favours selection of correlated regressors simultaneously while LASSO tends to select only on out of them. In fact the EasticNet can be solved as a LASSO using slight modification of the LARS algorithm. Since it is a mix of Ridge and LASSO, parameters can be estimated by Ridge in a first step then apply the LASSO. A small correction due to the second penalty λ_2 is required. Originally the Elastic-net was not designed explicitly for grouped structure.
Also composite penalties have been considered in Breheny and Huang (2009) using the MCP criterion at both stages (between and within).

Since there is a great literature of reviews on these method (Breheny and Huang 2009, Huang et al. 2012) we do not spend time giving more details and advise readers interested in group models to have a look at them.

5. Additive Models

A step further in model structure complexity is to consider different non-parametric functions associated with each variables. The non-parametric additive model takes the following form:

$$y = \sum_{j=1}^{p} f_j(x_j) + \varepsilon \tag{16}$$

5.1. Penalty

The Sparse Additive Model (SpAM) of Ravikumar et al. (2007) applies to this kind of models. The idea is simply to apply the LASSO to functions non-parametrically fitted with parametric coefficients coming in top of them. This is obviously the most natural extension of LASSO to the additive structure. The main program to solve is:

$$\min_{\beta, f_j} \|y - \sum_{j=1}^{p} \beta_j f_j(x_j)\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{17}$$

Even though the term $\sum_{j=1}^{p} \beta_j f_j(x_j)$ remind us the very well-known Splines where the f_j would be the basis functions, the authors claim that any non-parametric method can be used for fitting them. The solution is given in the form of a backfitting algorithm (Breiman and Friedman 1985). Another approach have been investigated by Meier et al. (2009): the penalized General Additive Model (penGAM). It applies to the same models as before but are especially designed for splines estimation. In the same spirit the individual functions are penalized, but since each function can be represented as the sum of linear combinations of basis functions. It turns out to be a groupLASSO problem.
Their contribution is also to consider not only sparsity but also smoothness in the estimation. Because complex functions require many basis functions it is common in the splines settings to construct an over complete basis and then apply shrinkage on coefficients [5] to have a smooth estimates, this is known as smoothing splines. This takes the form of a Ridge regression so it can be easily integrated inside the procedure. The main objective is to solve:

$$\min_{f} \|y - f(X)\|_2^2 + J(f) \tag{18}$$

With the sparsity-smoothness penalty being:

$$J(f) = \lambda_1 \sqrt{\|f_j\|^2 + \lambda_2 \int (f_j''(x))^2 dx} \tag{19}$$

and because we can rewrite each $f_j(x) = \sum_{k=1}^{K} \beta_{j,k} b_{j,k}(x)$ as a sum of K basis $b(\cdot)$ then the problem can be written as:

$$\min_{\beta} \|y - B\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\beta_j' B_j' B_j \beta_j + \lambda_2 \beta_j' \Omega_j \beta_j} \tag{20}$$

Ω_j composed of the inner products of the second derivatives of the basis functions.

5.2. Screening

In an equivalent manner on the screening side the Non-parametric Independence Screening procedure (NIS) has been introduced by Fan et al. (2011) as a natural extension to SIS. Instead of marginal linear correlation they use the concept of "marginal utility", already defined in Fan et al. (2009) for generalized linear models, and here

[5] Usually the Ridge because it has an analytical solution.

set this marginal utility to be the sum of squared marginal residuals resulting from a non-parametric additive model.

$$\hat{\omega}_j = \sum_{i=1}^{n} \left(y_i - \hat{f}_j(x_{i,j})\right)^2 \tag{21}$$

The latter, with $\hat{f}_j(x_{ij})$ obtained by splines [6], gives a ranking of variables in the same way as SIS:

$$\widehat{M} = \{1 \leq j \leq p : \hat{\omega}_j > \delta\} \tag{22}$$

Where δ is a predefined threshold. Usually this step does not ensures selection consistency so they rely on a external procedure, namely SpAM or penGAM. Because of the problem of weak signals Iterative Conditional SIS has been discussed exactly the same as Iterative Conditional SIS was for SIS, that is applying NIS on residuals, conditionally on primarily selected variables. It is worth mentionning the work of Zhang et al. (2017) who developed Correlation Ranked SIS (CR-SIS). The main purpose is to allow for any monotonic transformation of y by using its cumulative distribution as the dependent variable.

$$\begin{aligned} \omega_j &= Cov(f_j(x_j), G(y))^2 \\ G(y) &= \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y) \end{aligned} \tag{23}$$

The resulting model is less restricted allowing a non-linear response.

6. Partial Linear Models

A Partial Linear model takes the form:

$$y = X_1\beta + g(X_2) + \varepsilon \tag{24}$$

An important feature of these models is to assume two sets of variables. The X matrix is divided into X_1 and X_2 of dimension $p1$ and $p2$ respectively. The motivation behind this is to say that linearity is satisfactory enough for some variables and treating these ones non-parametrically result in a loss of efficiency. So one should divide the regressors according to their link function either it is parametric (X_1) or not (X_2). This section is divided in two parts. The first one will concern Partial Linear models in their general form. Because a great literature has focused on smoothly varying-coefficients the second part will focus only on them.

6.1. Standard

6.1.1. Penalty

The Double-Penalized Least Squares Estimator (DPLSE) of Ni et al. (2009) is a method for selection of variables and selection between parametric and non-parametric parts. A penalty is imposed on the parametric part to select variables and splines are used for non-parametric estimation. Since in the splines settings one can rewrite this function as a linear combination of basis expansion:

$$g = [J, X_2]\delta + Ba \tag{25}$$

with J the unit vector a are the parameters of the basis expansion B and δ is the overall parameter on X_2. The SCAD penalty is then applied on the vector $\beta^* = [\beta, \delta]$ This can be viewed as a composite penalty where the key idea is to write everything as linear and perform usual model selection. Partial Splines with Adaptive penalty (PSA) of Cheng et al. (2015) try to achieve a sparse parametric part while having a non-parametric part aside

---

[6] Because of low computational costs, but it can be estimated with any non-parametric regression technology.

using a combination of Adaptive LASSO on the parametric part and Penalized Splines for the non-parametric. Therefore the problem to solve is:

$$\min_{\beta,f} \|y - X_1\beta - f(X_2)\|^2 + \lambda_1 \int_0^1 (f''(X_2))^2 dX_2 + \lambda_2 \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma} \tag{26}$$

We remark the last term is exactly the penalty from the adaptive LASSO. This is in line with DPLSE, adding a smoothness penalty on top of the procedure. In this respect it is worth mentioning the Penalized Estimation with Polynomial Splines (PEPS) of Lian et al. (2015). The same objective is achieved in a quite similar fashion. The only difference is that the penalty is not adaptive:

$$\min_{\beta,f} \|y - B\beta\|^2 + n\lambda_1 \sum_{j=1}^{p} w_{1,j}\|\beta_j\|_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2,j}\|\beta_j\|_{D_j} \tag{27}$$

Basis expansion is contained in B therefore exploiting once again the linear transformation provided in splines, just like DPLSE introduced it. The whole thing is turned as a linear model on which penalties are applied to achieve sparsity $\|\beta_j\|_{A_j} = \|\sum_k \beta_{j,k} B_k(x_j)\|$ and linear parts are recovered from the smoothness penalty $\|\beta_j\|_{D_j} = \|\sum_k \beta_{j,k} B''_k(x_j)\|$.

In the end there is little difference between the 3 procedures. All exploits the linearity provided by splines. PEPS improves on DPLSE adding a smoothness penalty and PSA improves on PEPS making the penalty adaptive to achieve better selection consistency.

6.2. Varying Coefficients

Another usual structure for modelling is the semi-varying coefficient model, written as:

$$y = X_1\beta + X_2\alpha(Z) + \varepsilon \tag{28}$$

The coefficients α associated to each $x_{j\in 2}$ are supposed to vary smoothly along another variable Z. This can be seen as a particular case of previous models where $g(\cdot)$ has the specific varying coefficient form.

6.2.1. Penalty

The methods in this section do not use the semi-structure form, they work only with the varying-coefficient part.

$$y = X\beta(Z) + \varepsilon \tag{29}$$

The Kernel LASSO of Wang and Xia (2009) deals with this problem in the spirit of groupLASSO.

$$\min_{\beta} \sum_{t=1}^{n}\sum_{i=1}^{n} \{y_i - X_i\beta(Z_t)\}^2 K_h(Zt - Z_i) + \sum_{j=1}^{p} \lambda_j \|\beta_j\| \tag{30}$$

The penalty enforces the procedure to reduce estimated varying coefficients close to zero to true zeros in a single-level group fashion.
Another improvement in this setting is the Adaptive Semi-Varying Coefficients (AdaSVC) of Hu and Xia (2012). Instead of all coefficients varying smoothly one may think that some don't (hence semi-varying). To avoid the loss of efficiency introduced by non-parametric estimation when the true underlying coefficient is constant the latter have to be identified. Their method can simultaneously identify and estimate such a model. Selection is done only over constant regressors. They do not consider sparsity as in Kernel LASSO. The idea is to impose a group penalty on the estimated varying-coefficients such that the penalty enforces nearly constant coefficients to

be truly constant. Their penalty is in line with the FusedLASSO of Tibshirani et al. (2005). The main idea is that nearly constant coefficients will become constant in a grouped fashion. The objective is to solve:

$$\min_{\beta} \sum_{t=1}^{n} \sum_{i=1}^{n} \{y_i - X_i\beta(Z_t)\}^2 K_h(Zt - Z_i) + \sum_{j=1}^{p} \lambda_j \|b_j\| \tag{31}$$

with the penalty applied on a different norm than the Kernel LASSO:

$$\|b_j\| = \left\{ \sum_{t=2}^{n} (\beta_j(Z_t) - \beta_j(Z_{t-1}))^2 \right\}^{1/2} \tag{32}$$

6.2.2. Testing

The Semi-Parametric Generalized Likelihood Ratio Test (SP-GLRT) of Li and Liang (2008). It applies to semi-varying coefficients model. The purpose is both to identify relevant variables and whether if they belong to the non-linear or the linear component. The likelihood can be written as:

$$\mathcal{L}(\alpha, \beta) = l(\alpha, \beta) - n \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \tag{33}$$

The two parts are estimated alternatively conditionally on the other. Then they introduce a novel generalized likelihood ratio test:

$$\mathcal{T}_G LR = r_K\{\mathcal{R}(H_1) - \mathcal{R}(H_0)\} \tag{34}$$

with

$$\mathcal{R}(H_1) = \mathcal{Q}(X_1\beta + X_2\alpha(Z), y) \tag{35}$$

The conditional likelihood under H_1 : at least one coefficient from the non-parametric part is non-zero.

$$\mathcal{R}(H_0) = \mathcal{Q}(X_1\beta, y) \tag{36}$$

The conditional likelihood under H_0 : the variable does not appear in the non-parametric part. where the conditional likelihood is given by:

$$\mathcal{Q}(\mu, y) = \int_{\mu}^{y} \frac{s - y}{V(s)} ds \tag{37}$$

The test is then evaluated using a Monte Carlo or Bootstrap methods to empirically estimates distribution of the statistics since the theoretical degrees of freedom tends to infinity preventing from a parametric test.
This has to be noticed because this is one of the first attempt of introducing non-parametric and therefore automatic tests inside a selection procedure. While methods like Autometrics and Stepwise Regression relies on parametric tests, SP-GLRT uses data-driven tests to construct the model. This idea of exploiting the data themselves to conduct tests is certainly not new, but it was in model selection. This idea is the core of methodologies for improving model selection in section 8.

7. Non-Parametric Models

A fully non-parametric model takes the form of:

$$y = f(X) + \varepsilon \tag{38}$$

Where $f(\cdot)$ is any multivariate function, linear or not, additive or not. This framework is very general therefore making it complicated for estimation. The most well known drawback is the Curse of Dimensionality. Briefly, it states that the number of observations required for estimation of this function grows exponentially with the dimension of X : p. It is already complicated to fit such a (maybe very non-linear) function non-parametrically in a reduced dimension, thus looking for a sparse representation is necessary when dealing with large p.

This time the different methods differ under several aspects. Testing ones like MARS shares similarities with Stepwise Regression for example, in an ANOVA Splines settings. Penalty ones uses ANOVA models also, the reason is that it limits interactions terms and gets closer to an additive model, this is indeed very common when dealing with fully non-parametric regression. The screening based ones can be divided in two categories: some make the use of generalized correlations to avoid using a model (DC-SIS,HSIC-SIS,KCCA-SIS,Gcorr) [7] while others rely on a specific model ex ante (MDI,MDA,RODEO).[8]

7.1. Penalty

The Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions (VANISH) of Radchenko and James (2010). It is very similar to the SHIM of Choi et al. (2010) but in a non-linear framework. In order to approach the complexity of the function it uses an ANOVA-type of model defined as:

$$f(X) = \sum_{j=1}^{p} f_j(x_j) + \sum_{j<k} f_{j,k}(x_j, x_k) + ... + \varepsilon \tag{39}$$

Where f_j are the main effects, $f_{j,k}$ are the two-way interactions and so on. Their approach is closely related to the penGAM of Meier et al. (2009) generalized to include interaction terms [9] but with a different penalty. The authors say that the penalty shouldn't be the same for main effect than for two-way interactions. They advocate the fact that ceteris paribus including an interaction term add more regressors than a main effect and thus that they are less interpretable. So interactions should be more penalized. Therefore this condition is a little bit different from the "strong heredity constraint" introduced in Choi et al. (2010). The objective is to solve:

$$\min_f \|y - f(X)\|_2^2 + \tau^2 J(f) \tag{40}$$

With:

$$J(f) = \lambda_1 \sum_{j=1}^{p} \left(\|f_j\|^2 + \sum_{k \neq j} \|f_{j,k}\|^2 \right)^{1/2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=j+1}^{p} \|f_{j,k}\| \tag{41}$$

The penalty is written so that the first part penalizes additional regressors while the second penalizes interactions occurring without main effects. In the SHIM there was no possibility for that. Here this constraint is released but a stronger penalty can be applied to restrict interactions without main effects, which are less interpretable. Another approach for fitting this type of models is the Component Selection and Smoothing Operator (COSSO) of Lin et al. (2006). It differs from VANISH in the penalty function. The key idea is to use a penalty term written in terms of a sum of Reproducible Kernel Hilbert Space (RHKS) norms. In a model with only two-way interactions it would be:

$$J(f) = \sum_{\alpha=1}^{p(p-1)/2} \|P^{\alpha} f\|^2 \tag{42}$$

This time the penalty is not designed to take into account the structure of the resulting model. There is no desire to limit interactions. Since the heredity constraint is not present as before the model authors of VANISH claim it has trouble with high dimensional settings. Nevertheless the heredity constraint can obviously be inadequate in some applications where only interactions matter, in this type of settings COSSO is more advisable than VANISH.

---

7 Respectively Distance Correlation-SIS, Hilbert Schmidt Independence Criterion-SIS, Kernel Canonical Correlation Analysis and the Generalized Correlation.

8 Respectively Mean Decrease Impurity, Mean Decrease Accuracy and the Regularization Of Derivative Expectation Operator.

9 They also introduce it as SpIn (SpAM with INteractions) in their paper but claim that interactions would then not be treated efficiently.

7.2. Testing

Introduced by Friedman (1991) the Multivariate Adaptive Regression Splines is a method for building non-parametric fully non-linear ANOVA sparse models (39). The model is written in terms of splines as:

$$\hat{f}(x) = \sum_{k=1}^{K} c_k B_k(x) \tag{43}$$

The basis functions B_k are taken to be hinge functions. The form of these functions makes the model piecewise linear.

$$B_k(x, \alpha, \beta) = \beta max(0, \alpha + x) \tag{44}$$

Therefore α can be considered as "knots" like in standard splines. The β are parameters on which selection will occur through a pretty complicated algorithm. The building process is quite comparable to the one of usual Regression Trees and Stepwise Regression. Starting from a null model a forward step search over all possible variables and determines by least squares the parameter β (thus it creates a new hinge function) and over all possible values where to add a knot α that reduces best the residuals sum of squares [10] . This process goes until some stopping criterion is met. All combinations have to be taken into account, therefore it is computationally intractable for high interactions effects. Friedman advises to limit the number of interactions m to a small value like 2 such that the model can be build in a reasonable time. Selection of variables is part of the building process. If using a fit based criterion like the sum of squares residuals, variables are selected only if they bring enough explanatory power during the search. The same thing applies for Regression Trees on non-parametric models. In this sense MARS is closely related to Stepwise Regression. Also MARS is available with a backward approach, and a combination of both. This method is mainly used to fit high dimensional non-linear functions because since it is piecewise linear, it does not suffer much from the Curse of Dimensionality. However its selection consistency can be directly linked to the way variables are selected in trees, this is discussed in the next subsections. Used directly MARS is more like a non-linear version of Stepwise Regression using piecewise functions.

7.3. Screening

7.3.1. Model-free

In the screening literature of non-parametric methods we find a bunch of papers that deals with the same core idea. They all define some association measure that generalizes usual linear correlation. Here is the list of them as well as the criteria they use. In fact these methods are quite nested within each other. Considering which one is the best is a question of computational complexity rather than in which case they apply. Otherwise it seems that the last one (KCCA) should be selected.

- DC-SIS (Li et al. 2012)
  The Distance Correlation is a generalization of the Pearson Correlation Coefficient in terms of norm distances. It can be written as:
  $$\omega_j = \frac{dcov(x,y)}{\sqrt{dcov(x,x)dcov(y,y)}} \tag{45}$$
  Where:
  $$\begin{aligned} dcov(x,y)^2 = {} & \mathbb{E}[\|X - X'\|\|Y - Y'\|] \\ & + \mathbb{E}[\|X - X'\|]\mathbb{E}[\|Y - Y'\|] \\ & - 2\mathbb{E}[\mathbb{E}[\|X - X'\|]\mathbb{E}[\|Y - Y'\|]] \end{aligned} \tag{46}$$

[10] This is known as "Greedy Algorithms" where the optimal global solution is sought by taking optimal local solutions.

- HSIC-SIS (Balasubramanian et al. 2013)
  The Hilbert Schmidt Independence Criterion generalizes the previous one as it defines a maximum distance metric in a RKHS space:

$$\begin{aligned}\omega^2_{(k)} &= \mathbb{E}[k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')] \\ &+\mathbb{E}[k_{\mathcal{X}}(X,X')]\mathbb{E}[k_{\mathcal{Y}}(Y,Y')] \\ &-2\mathbb{E}[\mathbb{E}[k_{\mathcal{X}}(X,X')]\mathbb{E}[k_{\mathcal{Y}}(Y,Y')]]\end{aligned} \tag{47}$$

  We recognize again the form of the usual correlation but this time written in terms of kernels. In order to avoid the choice of the bandwidths in kernels, they decided to use the sup of the criterion over a family of Kernel $\mathcal{K}$.

$$\gamma = \sup\left\{\omega_{(k)} : k \in \mathcal{K}\right\} \tag{48}$$

  Empirically the ranking measure is simpler to compute:

$$\hat{\omega} = \frac{1}{n}\sup_{k_{\mathcal{X}},k_{\mathcal{Y}}}\sqrt{trace(K_{\mathcal{X}}HK_{\mathcal{Y}}H)} \tag{49}$$

  with $H = I - (1/n)JJ'$, I being the $n \times n$ unit matrix and J the $n \times 1$ unit vector.
- KCCA-SIS Liu et al. (2016)
  The Kernel Canonical Correlation Analysis is the last improvement in the field of Non-parametric Screening. It encompasses SIS as it can handle non-linearities. Unlike DC-SIS it is scale-free and does not rely on the Gaussian assumption. However even if it shares many aspects of the HSIC-SIS it differs in one aspect: HSIC is based on maximum covariance between the transformations of two variables, while KCCA uses the maximum correlation between the transformations by removing the marginal variations. Their measure is defined as:

$$\mathcal{R}_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2} \tag{50}$$

  Because the covariance matrices may not be invertible they introduce a ridge penalty ε:

$$\mathcal{R}_{YX} = (\Sigma_{YY} + \varepsilon I)^{-1/2}\Sigma_{YX}(\Sigma_{XX} + \varepsilon I)^{-1/2} \tag{51}$$

  The correlation measure is then defined as the norm of the correlation operator:

$$\omega(\varepsilon)_j = \|\mathcal{R}_{YX_j}\| \tag{52}$$

  Empirical estimates of covariance matrices Σ are obtained after singular decomposition of kernel matrices (the latter being the same as in HSIC). While bandwidths in kernels can be chosen optimally ex ante, ε has to be estimated via GCV over a grid of values.

For each one the variables are ranked along marginal association measures $\hat{\omega}_j$ between y and x_j and one defines the set of relevant features after applying a threshold. The latter's value differs among them.

$$\widehat{M} = \{1 \leq j \leq p : \hat{\omega}_j \geq \delta\} \tag{53}$$

- DC-SIS: $\delta = cn^{-k}$
- HSIC-SIS: $\delta = cn^{-k}$
- KCCA-SIS: $\delta = cn^{-k}\varepsilon^{-3/2}$

with $0 \leq k \leq 1/2$.

Another of the same kind is the Generalized Correlation Screening (Gcorr) of Hall and Miller (2009) that was introduced as a more general method than NIS. The general correlation coefficient is used as the measure of non-linear relationship. It can be defined as:

$$\hat{\omega}_j = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^{n}\{h(x_{i,j}) - \bar{h}_j\}(y_i - \bar{y})}{\sqrt{n\sum_{i=1}^{n}\{h(x_{i,j})^2 - \bar{h}_j^2\}}} \tag{54}$$

Then these estimates are tested using bootstrap confidence interval instead of threshold like the others usually do. Finally significant ones are ranked. Even though their method seems very general, empirically $h(\cdot)$ are chosen to be polynomial functions. This can be restrictive in some situations and less non-parametric in some sense.

7.3.2. Model Based

The Regularization Of Derivative Expectation Operator (RODEO) of Lafferty et al. (2008), named in reference to the LASSO, applies in the framework of Multivariate Kernel Methods. In kernel regression a specific attention is given to the choice of the bandwidth. We recall that this hyperparameter defines the width of the support for the regression, the lower it is the less observations enter the local regression, leading to less bias but more variance and conversely for a high bandwidth. The authors here state that for variables that are important in the model the derivative of the estimated function with respect to the bandwidth h is higher than for useless variables. A change in bandwidth affects the estimation if the variable intervenes in the model, it affects the bias-variance trade-off. For an irrelevant variable a change in bandwidth has no effect since more or less observations does not change the fitted curve. For a Gaussian kernel we have:

$$\begin{aligned}
\frac{\partial f_h(x)}{\partial h_j} &= e'(X'WX)^{-1}X'\frac{\partial W}{\partial h_j}(y - X\hat{\beta}) \\
\frac{\partial W}{\partial h_j} &= WL_j \\
L_j &= \frac{1}{h_j^3} diag((x_{1,j} - \bar{x}_j)^2, ..., (x_{n,j} - \bar{x}_j)^2)
\end{aligned} \tag{55}$$

Note that it refers to a specific point in the sample $\bar{x}$. The derivative is not computed over the whole sample. The authors propose an extension of local RODEO to a global procedure where the derivative is computed in every point and then averaged.

The idea is to exploit this derivative iteratively, starting from a high bandwidth value and adapted in each step according to a certain rate of decay. Important variables should have low bandwidth, so the derivative is greater and the bandwidth reduces more quickly. Variables then can be ranked according to the final value of their bandwidth. One can apply some threshold on these to end up with a sparse solution. In this respect RODEO can be classified as a screening procedure. RODEO is based on a full estimation via Kernel, therefore it suffers the Curse of Dimensionality mentioned earlier. RODEO may not be able to deal with high dimensional feature space.

A large part of the literature focuses on a quite restricted set of regression methods for doing selection such as Ordinary Least Squares for linear models, Splines and Kernels for non-linear ones. However there exists other ways for doing regression from which model selection procedures intuitively arise. In a Bayesian framework [11] one will consider a collection of models called an Ensemble. There is a distribution of them and we are uncertain on which one is the truth [12]. Still we can exploit this distribution accross these different models to assign probabilities to each variables since they may not all appear in every models. This idea has also been developed in the frequentist approach by Breiman (2001) who introduced Random Forest. From an Ensemble of Regression Trees (called a Forest) he derived two types of variables importance measures : Mean Decrease

[11] Which is out of the scope of this paper but still very important.
[12] This relates obviously to the problem raised when discussing Stepwise Regression. Here the Ensemble is a subset of the model space.

Impurity (MDI) and Mean Decrease Accuracy (MDA). We recall briefly that a tree is constructed as a recursive partitioning over the sample space. Simple Regression Trees allows for constant estimation in subregions, this is closely related to the Nadaraya-Watson local constant kernel estimator. Splits are chosen according to an impurity criterion that describes the degree of similarity [13] of the data in the partition.

$$MDI(x_j) = \frac{1}{N_t}\sum_{T}\sum_{t\in T}\frac{N_t}{N}\left(i(t) - \frac{N_{t_{left}}}{N_t}i(t_{left}) - \frac{N_{t_{right}}}{N_t}i(t_{right})\right) \tag{56}$$

The importance of variable j is computed as the average decrease in impurity among each node t in tree T. The idea is to show the decrease in impurity caused by a split in this variable. It is computed as the impurity in the node minus the sum of impurity in the child nodes weighted by their respective sizes. This gain is weighted in the end by the number of observations entering the node. MDI can be easily extended to an Ensemble of Trees (i.e. a Forest).

The second measure relies on the predictive power of the model instead of the impurity inside nodes. From a statistical point of view it is equivalent as focusing on out-of-sample fit rather than in-sample fit. Since it does not rely on an inside criterion it is only defined for a tree and therefore applies only for an ensemble of them.

$$VI^T(x_j) = \frac{\sum_{i\in\mathcal{B}^{(T)}} I(y_i = y_i^{(T)})}{|\mathcal{B}^{(T)}|} - \frac{\sum_{i\in\mathcal{B}^{(t)}} I(y_i = y_{i,\pi_j}^{(T)})}{|\mathcal{B}^{(T)}|} \tag{57}$$

$$MDA(x_j) = \frac{\sum_{T\in F} VI^T(x_j)}{N_T} \tag{58}$$

The importance of variable j is computed as the average decrease in accuracy among each tree T in the forest F. The idea is that if a variable is uninformative then the prediction accuracy should be unchanged under permutation. The difference between actual prediction and permuted prediction give sthe decrease in accuracy for each variable and the whole is a weighted average of each tree in the forest.

8. Improving on Variable Selection

This last section is devoted to general methodologies designed for improving model selection procedures. Based on bootstrap or resampling, the core idea is to exploit randomness to account for uncertainty in the modelling. Usual model selection procedures may suffer from inconsistency under some conditions. For example we remember the LASSO where the regularization parameter λ can not be chosen optimally [14] so that it ensures correct identification. This has lead to the adaptive LASSO (Zou 2006), but this problem can also be solved using these procedures.

8.1. Stability Selection

The Stability Selection (Stabsel) has been introduced by Meinshausen and Bühlmann (2010) to improve on selection. Given a specific selection procedure a variable is said to be stable if its selection probability under subsampling [15] (number of times it has been selected among the random samples) exceeds a specified threshold δ. The selection probabilities for a variable j to belong to the set S^{λ} of selected variables for a given regularization parameter λ is:

$$\Pi_j^{\lambda} = \mathbb{P}(j \subseteq S^{\lambda}) \tag{59}$$

The set of stable variables is then:

$$S^{Stable} = \{j : \max_{\lambda\in\Lambda}\Pi_j^{\lambda} \geq \delta\} \tag{60}$$

[13] In case of a regression: How well the subregion can be approximated by a constant.
[14] Both from an estimation and a predictive point of view
[15] Without replacement, random samples have to be non-overlapping.

This is given by the underlying selection procedure, it can be the LASSO or whatever, but the methodology aims at improving a procedure, not being one itself.
Another way for randomness that is almost equivalent is to divide the sample in two non-overlapping parts of sizes $\lfloor n/2 \rfloor$ and look for variables that are selected simultaneously in both. This is more computationally efficient. The threshold can be selected appropriately so that the expected number of false inclusion V is bounded.

$$\mathbb{E}[V] \leq \frac{1}{2\delta - 1}\frac{q_\lambda^2}{p} \tag{61}$$

Thus one will ensure $\mathbb{P}(V > 0) \leq \alpha$ by setting for example:

$$\begin{aligned} \delta &= 0.9 \\ q_\lambda &= \sqrt{0.8\alpha p} \end{aligned} \tag{62}$$

The results are then presented as stability paths: Π_j^λ as a function of λ. This is in contrast to regularization paths of LASSO: β_j as a function of λ.
Extensions to Stabsel are proposed in Bach (2008) and Shah and Samworth (2013). The first uses bootstrap with replacement instead of resampling without while the latter uses subsampling of complementary pairs.

8.2. Ranking-Based Variable Selection

The Ranking-Based Variable Selection (RBVS) of Baranowski and Fryzlewicz (2016) is a screening procedure based on bootstrap and permutation tests. Contrary to Stabsel it does not rely on any threshold nor any assumptions.
Given a metric to assess the strength of the relationship denoted ω and then using the m-out-of-n bootstrap of Bickel et al. (2012) they construct a permutation ranking $\mathcal{R}$.

$$\mathcal{R} = (\mathcal{R}_1, ..., \mathcal{R}_p) \text{ satisfying } \omega_{\mathcal{R}_1} \geq ... \geq \omega_{\mathcal{R}_p} \tag{63}$$

The metric can be anything like the Pearson Correlation, the LASSO coefficients, etc. The probability of the set of the k top-ranked variables $\mathcal{A}_k$ is defined as:

$$\pi(\mathcal{A}_k) = \mathbb{P}(\{\mathcal{R}_1, ..., \mathcal{R}_k\} = \mathcal{A}) \tag{64}$$

This value is approximated with using the m-out-of-n bootstrap procedure involving random draws without replacements of the observations.
In fact selection can be performed on the set of top-ranked variables $\mathcal{A}$ from which the number of terms k^* can be determined automatically without threshold. The idea is not to look for a threshold δ that would cut in the ranking of ω. As an alternative they try to estimate k^* as:

$$k^* = argmin_{k=0,...,p-1} \frac{\pi(\mathcal{A}_{k+1,m})}{\pi(\mathcal{A}_{k,m})} \tag{65}$$

That is the number of terms for which the differences among the $\pi(\mathcal{A})$ is the greater. This is equivalent to look for a threshold that best separates assuming there are two sets: the relevant and the irrelevant. It has the advantage of being totally non-parametric. Just like the SIS has its iterative counterpart they introduce the Iterative RBVS that accounts for marginally related variables with low Signal-to-Noise and for the multicollinearity problem.

9. Discussion

In this article, we provide a review for 39 state-of-the-art procedures to perform variable selection over a wide range of model structures going from the simple linear one to the complex non-parametric one. Procedures have been classified in three groups: Tests-Based, Penalty-Based and Screening-Based. They have been described and compared on the ground of model structures.

The main difference consists of modelling purposes and objectives rather than their strength as oracles. In an empirical work the choice between two strategies should rely on the form of the model, data specificities (collinearity,groups, etc..) and objectives (in other words understandability).

Selection consistency for widely used methods in empirical work have been discussed and several improvements were presented. Far beyond Stepwise Regression and the LASSO, empiricists have access to more advanced technologies that we claim are not much more complicated than the basic ones. The limits in main methods (LASSO, Stepwise Regression) are now well understood and various answers have come to light.

The area of model selection is still very investigated, much more now that amounts of data have become available. Nevertheless, methods for handling large number of variables are restricted in terms of model complexity. This is mainly due to the Curse of Dimensionality and it prevents from looking for very complex models in high dimensions. Sure Independence Screening is a powerful tool in linear models but have lower dataset capacities when it comes to non-linearities. Also, the literature is lacking from very complete algorithmic solutions. To the best of our knowledge, no statistical procedure have been developed to reach the level of completeness of Autometrics. Other methods are only parts of the statistical work and do not cover as many problems as Autometrics do.

Conflicts of Interest: The authors declare no conflict of interest..

Bach, Francis R. 2008. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pp. 33–40. ACM.

Balasubramanian, Krishnakumar, Bharath Sriperumbudur, and Guy Lebanon. 2013. Ultrahigh dimensional feature screening via rkhs embeddings. In *Artificial Intelligence and Statistics*, pp. 126–134.

Baranowski, Rafal and Piotr Fryzlewicz. 2016. Ranking-based variable selection for high-dimensional data.

Breaux, Harold J. 1967. On stepwise multiple linear regression. Technical report, Army Ballistic Research Lab Aberdeen Proving Ground MD.

Breheny, Patrick and Jian Huang. 2009. Penalized methods for bi-level variable selection. *Statistics and its interface 2*(3), 369.

Breiman, Leo. 1995. Better subset regression using the nonnegative garrote. *Technometrics 37*(4), 373–384.

Breiman, Leo. 2001. Random forests. *Machine learning 45*(1), 5–32.

Campos, Julia, Neil R Ericsson, and David F Hendry. 2005. General-to-specific modeling: an overview and selected bibliography.

Candes, Emmanuel, Terence Tao, et al.. 2007. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Castle, Jennifer L, Jurgen A Doornik, and David F Hendry. 2011. Evaluating automatic model selection. *Journal of Time Series Econometrics 3*(1).

Castle, Jennifer L and David F Hendry. 2010. A low-dimension portmanteau test for non-linearity. *Journal of Econometrics 158*(2), 231–245.

Chen, Xueying and Min-ge Xie. 2014. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 1655–1684.

Cheng, Guang, Hao Helen Zhang, and Zuofeng Shang. 2015. Sparse and efficient estimation for partial spline models with increasing dimension. *Annals of the Institute of Statistical Mathematics 67*(1), 93–127.

Choi, Nam Hee, William Li, and Ji Zhu. 2010. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association 105*(489), 354–364.

Cleveland, William S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association 74*(368), 829–836.

Cleveland, William S and Eric Grosse. 1991. Computational methods for local regression. *Statistics and Computing 1*(1), 47–62.

Doornik, Jurgen A. 2009. Econometric model selection with more variables than observations. *Unpublished paper). Economics Department, University of Oxford*.

Doornik, Jurgen A and David F Hendry. 2015. Statistical model selection with “big data”. *Cogent Economics & Finance 3*(1), 1045216.

Doornik, Jurgen A, David F Hendry, and Felix Pretis. 2013. Step-indicator saturation. *University of Oxford, Department of Economics Discussion Paper Series 658*.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al.. 2004. Least angle regression. *The Annals of statistics 32*(2), 407–499.

Efroymson, MA. 1960. Multiple regression analysis. *Mathematical methods for digital computers*, 191–203.

Epprecht, Camila, Dominique Guegan, and Alvaro Veiga. 2013. Comparing variable selection techniques for linear regression: Lasso and autometrics.

Fan, Jianqing, Yang Feng, and Rui Song. 2011. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association 106*(494), 544–557.

Fan, Jianqing and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, Jianqing and Jinchi Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, Jianqing and Jinchi Lv. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101.

Fan, Jianqing and Jinchi Lv. 2017. Sure independence screening.

Fan, Jianqing, Yunbei Ma, and Wei Dai. 2014. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association 109*(507), 1270–1284.

Fan, Jianqing, Richard Samworth, and Yichao Wu. 2009. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of machine learning research 10*(Sep), 2013–2038.

Fan, Jianqing and Wenyang Zhang. 2008. Statistical methods with varying coefficient models. *Statistics and its Interface 1*(1), 179.

Flom, Peter L and David L Cassell. 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland*.

Frank, LLdiko E and Jerome H Friedman. 1993. A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Friedman, Jerome H. 1991. Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

Fu, Wenjiang J. 1998. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics 7*(3), 397–416.

Hall, Peter and Hugh Miller. 2009. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics 18*(3), 533–550.

Hardle, Wolfgang, Hua Liang, and Jiti Gao. 2012. *Partially linear models*. Springer Science & Business Media.

Hastie, Trevor and Robert Tibshirani. 1993. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hendry, David F, Soren Johansen, and Carlos Santos. 2006. Selecting a regression saturated by indicators.

Hendry, David F and Hans-Martin Krolzig. 1999. Improving on 'data mining reconsidered'by kd hoover and sj perez. *The econometrics journal 2*(2), 202–219.

Hendry, David F and Hans-Martin Krolzig. 2005. The properties of automatic gets modelling. *The Economic Journal 115*(502).

Hendry, David F, J-F Richard, et al.. 1987. Recent developments in the theory of encompassing. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Hoerl, Arthur E and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hoover, Kevin D and Stephen J Perez. 1999. Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The econometrics journal 2*(2), 167–191.

Hu, Tao and Yingcun Xia. 2012. Adaptive semi-varying coefficient model selection. *Statistica Sinica*, 575–599.

Huang, Jian, Patrick Breheny, and Shuangge Ma. 2012. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics 27*(4).

Huang, Jian, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang. 2009. A group bridge approach for variable selection. *Biometrika 96*(2), 339–355.

Hurvich, Clifford M and Chih—Ling Tsai. 1990. The impact of model selection on inference in linear regression. *The American Statistician 44*(3), 214–217.

Kai, Bo, Runze Li, and Hui Zou. 2011. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of statistics 39*(1), 305.

Kazemi, M, D Shahsavani, and M Arashi. 2017. A sure independence screening procedure for ultra-high dimensional partially linear additive models. *arXiv preprint arXiv:1708.08604*.

Ke, Tracy, Jiashun Jin, and Jianqing Fan. 2014. Covariate assisted screening and estimation. *Annals of statistics 42*(6), 2202.

Ke, Zheng Tracy and Fan Yang. 2017. Covariate assisted variable ranking. *arXiv preprint arXiv:1705.10370*.

Kim, Yongdai, Hosik Choi, and Hee-Seok Oh. 2008. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association 103*(484), 1665–1673.

Kong, Efang and Yingcun Xia. 2007. Variable selection for the single-index model. *Biometrika 94*(1), 217–229.

Kowalski, Matthieu. 2014. Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study. pp. 4151–4155.

Krolzig, Hans-Martin and David F Hendry. 2001. Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control 25*(6-7), 831–866.

Lafferty, John, Larry Wasserman, et al.. 2008. Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics 36*(1), 28–63.

Li, Runze and Hua Liang. 2008. Variable selection in semiparametric regression modeling. *Annals of statistics 36*(1), 261.

Li, Runze, Wei Zhong, and Liping Zhu. 2012. Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Lian, Heng, Hua Liang, and David Ruppert. 2015. Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statistica Sinica*, 591–607.

Lin, Yi, Hao Helen Zhang, et al.. 2006. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics 34*(5), 2272–2297.

Liu, Tianqi, Kuang-Yao Lee, and Hongyu Zhao. 2016. Ultrahigh dimensional feature selection via kernel canonical correlation analysis. *arXiv preprint arXiv:1604.07354*.

Meier, Lukas, Sara Van de Geer, Peter Buhlmann, et al.. 2009. High-dimensional additive modeling. *The Annals of Statistics 37*(6B), 3779–3821.

Meinshausen, Nicolai and Peter Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436–1462.

Meinshausen, Nicolai and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Ni, Xiao, Hao Helen Zhang, and Daowen Zhang. 2009. Automatic model selection for partially linear models. *Journal of multivariate Analysis 100*(9), 2100–2111.

Noh, Hoh Suk and Byeong U Park. 2010. Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, 1183–1202.

Park, Byeong U, Enno Mammen, Young K Lee, and Eun Ryung Lee. 2015. Varying coefficient regression models: a review and new developments. *International Statistical Review 83*(1), 36–64.

Racine, Jeffrey S. 2017. A primer on regression splines. *CRAN. R-Project http://cran. rproject. org/web/packages/crs/vignettes/spline primer. pdf*.

Radchenko, Peter and Gareth M James. 2010. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association 105*(492), 1541–1553.

Ravikumar, Pradeep, Han Liu, John Lafferty, and Larry Wasserman. 2007. Spam: Sparse additive models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1201–1208. Curran Associates Inc.

Santos, Carlos, David F Hendry, and Soren Johansen. 2008. Automatic selection of indicators in a fully saturated regression. *Computational Statistics 23*(2), 317–335.

Shah, Rajen D and Richard J Samworth. 2013. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 55–80.

Steyerberg, Ewout W, Marinus JC Eijkemans, and J Dik F Habbema. 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of clinical epidemiology 52*(10), 935–942.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

van den Burg, Gerrit JJ, Patrick JF Groenen, and Andreas Alfons. 2017. Sparsestep: Approximating the counting norm for sparse regularization. *arXiv preprint arXiv:1701.06967*.

Wang, Hansheng. 2009. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Wang, Hansheng and Yingcun Xia. 2009. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*(486), 747–757.

Wang, Lifeng, Guang Chen, and Hongzhe Li. 2007. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics 23*(12), 1486–1494.

Whittingham, Mark J, Philip A Stephens, Richard B Bradbury, and Robert P Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of animal ecology 75*(5), 1182–1189.

Yan, Xiaodong, Niangsheng Tang, and Xingqiu Zhao. 2017. The spearman rank correlation screening for ultrahigh dimensional censored data. *arXiv preprint arXiv:1702.02708*.

Yuan, Ming and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, Cun Hui. 2007. Penalized linear unbiased selection. *Department of Statistics and Bioinformatics, Rutgers University 3*.

Zhang, Cun-Hui et al.. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*(2), 894–942.

Zhang, Jing, Yanyan Liu, and Yuanshan Wu. 2017. Correlation rank screening for ultrahigh-dimensional survival data. *Computational Statistics & Data Analysis 108*, 121–132.

Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Zou, Hui and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.

Zou, Hui and Runze Li. 2008. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics 36*(4), 1509.

Zou, Hui and Hao Helen Zhang. 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics 37*(4), 1733.