Desboulets, Loann David Denis

Article

A review on variable selection in regression analysis

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Desboulets, Loann David Denis (2018) : A review on variable selection in regression analysis, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 6, Iss. 4, pp. 1-27, https://doi.org/10.3390/econometrics6040045

econometrics

MDPI

Review

A Review on Variable Selection in Regression Analysis

Loann David Denis Desboulets

CNRS, EHESS, Centrale Marseille, AMSE, Aix-Marseille University, 5-9 Boulevard Maurice Bourdet, 13001 Marseille, France; loann.DESBOULETS@univ-amu.fr

Received: 31 May 2018; Accepted: 20 November 2018; Published: 23 November 2018

Abstract: In this paper, we investigate several 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 number of data has significantly grown. Automatic model selection has been promoted to search for data-driven theories for quite a long time now. However, while great extensions have been made on the theoretical side, basic procedures are still used in most empirical work, e.g., stepwise regression. 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) and available software resources for implemented methods so that practitioners can easily access them. We provide explanations for which methods to use for different model purposes and their key differences. 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 often arises. 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 and Richard 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 (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 (2010a) reviewed 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 (2010b) provided a comprehensive review in the context of sure

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

independence screening major improvements. We can also cite more focusing papers, e.g., Fu (1998) compared the Bridge and LASSO theoretically as well as empirically using both simulation and real data, and Epprecht et al. (2017) compared autometrics and LASSO according to prediction and selection accuracy.

The contribution of this paper is threefold. First, many procedures are considered, which are listed and clearly classified. 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. Secondly, 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. Thirdly, we present appropriate software resources implementing the described methods, whenever they are available, on three different software packages: R, Matlab and SAS. 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 mentioned reviews.

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 in Sections 3–7. 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 across the paper. In Section 9, 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 are discussed in this paper. From the latter we have determined three main categories of algorithms:

- Test-based
- Penalty-based
- Screening-based

Originally, the first developed are based on statistical tests. The work is to automate standard tests in econometrics (e.g., testing residuals for normality, t-tests, etc.) to choose among candidate variables. It includes, for example, stepwise regression and autometrics.

Then, there are penalty-based procedures. Imposing a constraint on parameters directly inside estimation encourages sparsity among them. For instance, LASSO and ridge belong to this category.

The last includes screening procedures; they are not all designed to do selection intrinsically but rather ranking variables by importance. The main advantage is that they apply more easily to very large dimensional problems, when number of regressors is diverging with the number of observations (e.g., 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. Meanwhile, it is important to notice that in other fields such as computer science there exists a classification of "features selection". They distinguish among "filter", "wrapper" and "embedded" methods. Filter methods can be directly linked to statistical methods using thresholds for selection. Wrapper methods are iterative procedures involving filter methods and can be seen as special cases of "testing" ones. Embedded methods try to gather both in a unified way, leading to methods such as LASSO. Good reviews for this classification were reported by Blum and Langley (1997) and Saeys et al. (2007). More recent reviews, including higher connections to econometrical methods, were published by Mehmood et al. (2012) and Jović et al. (2015).

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

2.1. Test-Based

This is the most classical way of handling variable selection in statistical models. It is also the first attempt of variable selection. Everything started with stepwise regression (Breaux 1967); one of the latest of this type is autometrics (Hendry and Richard 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 dating 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 simplest and most straightforward way of doing model selection by just retrieving 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] Even though it started in 1987, there are still ongoing improvements being reported.

2.2. Penalty-Based

Thanks to the work of Tibshirani (1996), it this has become 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, i.e., 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 more complex model structures such as varying coefficients and two-way interaction 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, which is the test part, and penalty-based methods help estimate the coefficients, which is the inference part. Thus, both procedures are merged into a unified framework giving rise to a novel conception of statistical modelling. Perhaps the most famous in this category is LASSO of Tibshirani (1996).

2.3. Screening-Based

Screening is actually the most effective way of dealing with very high dimensional features ($\log(p) = \mathcal{O}(n^{\alpha})$ with $0 \leq \alpha \leq 1$). Few other selection methods can be as computationally efficient as these ones. However, screening often does not perform model selection itself, but rather ranks variables. To do so, they have to be combined 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 (SIS, Fan and Lv (2008)) that is the first of this kind.

Screening uses a ranking measure, either linear or not, so it can be applied in both frameworks. Some may rely on specific models (e.g., a linear model) while others are model-free. The major difference among procedures in this category is the choice of the ranking measure. Correlation coefficients are the first that come to mind, as they 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 lose their computational efficiency compared to independence screening ones. As said before, a brief review of some SIS methods was reported by Fan and Lv (2010b).

3. Linear Models

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

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

The variable to be explained $\mathbf{y}$ (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 $\mathbf{X}$ 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 $\boldsymbol{\beta}$ of length p contains the parameters of interest. The residuals (sometimes also called the error term) are denoted $\boldsymbol{\varepsilon}$; 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 is held constant throughout the paper. Notice that all three methodologies can handle linear models, while

this is not necessarily true for other structures (e.g., additive models). Software resources in this section exceed the ones for methods dealing with additive, partially linear and fully non-parametric models.

3.1. Testing

Stepwise regression (Breaux 1967) is the first model selection procedure. This approach was developed when statisticians started to consider model uncertainty. This means that, among p variables, we can possibly construct 2^p models, so we should possibly take them all into account. To test all possibilities, we have to compute "all-subsets". This cannot be achieved for large p. 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 (the forward step) or it starts from the full model (all variables) and deletes them one by one (the backward step). One improvement is also to consider both. Usually, the selection within each step is made according to some criterion. One considers all one-variable increments from the actual model and chooses the best move according to this criterion, which might be the lowest p-value; highest adjusted R^2; lowest Mallow's C_p (Mallows 1973); lowest AIC (Akaike 1973), AICc (Hurvich and Tsai 1989), BIC (Schwarz 1978) or HQIC (Hannan and Quinn 1979); lowest prediction error; leave-one-out cross validation; etc. Stepwise regression is available in almost any statistical software. In R, the package **leaps** Lumley (2017) provides this method with a wide range of criteria such as the ones cited above. In Matlab, the function `stepwiselm` is available from the "Statistical and Machine Learning Toolbox". In SAS, stepwise regression is implemented through the procedure `PROC GLMSELECT`.

One can choose any criterion to perform this task, but the main issue arising from stepwise regression does not come from the choice of the criterion. Interesting criticisms (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 ones to express admonitions about stepwise regression. We can mention many papers by Hurvich and Tsai (1990), Steyerberg et al. (1999), Whittingham et al. (2006) or Flom and Cassell (2007), who all proved biased estimation and inconsistent selection of stepwise regression.

If used as a selection method, it behaves poorly, but, used as it screening method, it shows 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 and Richard (1987) under the famous theory of encompassing and the LSE (London School of Economics) type of econometrics. This method was created in 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 the 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, 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 it often fails.

Non-linearities are handled using principal components analysis (see Castle and Hendry (2010)) that makes the design matrix orthogonal. Such a decomposition allows introducing 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 into different blocks until the size of each block is lower than p. Then, tests are applied to 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 Santos et al. (2008). 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. Autometrics was originally developed as a package in OxMetrics; there also exists a R package implementation named **gets** (Pretis et al. 2018).

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 into 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 it reduces 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 penalties, one should refer to threshold methods presented by 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:

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

Each methods applies to different L_γ norms[4].

- 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

---

[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 such as 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.

[4] For the purpose of illustration, one can refer to Fan and Li (2001) and Yuan and Lin (2006).

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). 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.[5] The zero norm used in SparseStep (van den Burg et al. 2017) is the counting norm; it penalizes 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 but very precise continuous approximation from de Rooi and Eilers (2011) and that turns the problem into something computationally tractable. It is worth noting that criteria such as the Akaike Information criterion (AIC, Akaike (1973); AICc, Hurvich and Tsai (1989)), Bayesian Information Criterion (BIC, Schwarz (1978)), Hannan-pQuinn Information Criterion (HQIC, Hannan and Quinn (1979)), Mallow's C_p (Mallows 1973) or the adjusted R-squared are also L_0 penalties. The difference is that AIC, BIC and HQIC apply to the likelihood which can employ a different loss function than the usual sum of squares residuals. Therefore, they represent other approaches to the L_0 penalty. It allows for variable selection into a broader class of parametric models than the simple linear model. However, the computational complexity of these criteria (NP complexity) makes them infeasible in contrast to SparseStep.

Meinshausen and Bühlmann (2006) showed that LASSO tends to select noise variables using a penalty parameter optimally chosen for prediction. For this reason, Zou (2006) developed AdaLASSO (Adaptive LASSO). He proved that the optimal estimation rate is not compatible with consistent selection. Moreover, even sacrificing the estimation rate does not ensure that 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 LASSO with a single penalty parameter. Therefore, he introduced adaptive weights $\mathbf{w}$ to LASSO to make it consistent with variable selection.

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

The vector of weights $\boldsymbol{w}$ is used to adjust the size of the penalty to each estimates. Hence, it acts similar to LASSO with p penalty parameters. The latest improvement on linear models is to allow for interactions terms. Even if it is possible, only adding them into 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 also selected (strong heredity property), which greatly reduces the search space and provides an efficient way of doing ANOVA-types models. Its considers a reparameterization of the two-way interactions models:

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

Introducing main effect parameters $\boldsymbol{\beta}$ on top of cross-effects $\boldsymbol{\gamma}$ 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:

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

Solutions to these problems are numerous. Usually, it reduces to LASSO and then algorithms such as the shooting algorithm (Fu 1998), the local quadratic approximation (Fan and Li 2001), the Least Angle Regression (LARS, Efron et al. (2004)) or the coordinate descent (Wu and Lange 2008) are employed. These methods are rather well-implemented in most software packages. In R, the package

[5] One can get insights of how to connect LASSO to stepwise regression (Efron et al. 2004) via the forward stagewise method of Weisberg (2005).

glmnet (Friedman et al. 2010) has options for LASSO, AdaLASSO and ridge. Sparsestep can be found in the package **SparseStep** (van den Burg et al. 2017). However, LASSO can be solved in many different ways; for each possible algorithm, there exists a corresponding package. The package **lars** (Hastie and Efron 2013) implements the LARS algorithm as well as the forward Stagewise and the stepwise approach but the shooting algorithm is only available in the **lassoshooting** package Abenius (2012). In Matlab, the "Statistics and Machine Learning Toolbox" implements only LASSO and ridge through functions of the same name. Matlab uses the ADMM (Alternating Direction Method of Multipliers of Boyd et al. (2011)). If one wants to use a specific algorithm to solve for LASSO, then there exists different resources available on Matlab Central. In addition, there exists a more complete toolbox named **penalized** (McIlhagga 2016) which has almost the same features as the R package **glmnet**. It implements the AdaLASSO which is missing in the "Statistics and Machine Learning Toolbox". In SAS, LASSO and AdaLASSO are available via `PROC GLMSELECT` using the LARS algorithm, and ridge via `PROC REG`. The SHIM model is only available in R via a github repository (https://github.com/sahirbhatnagar/shim).

3.2.2. Concave Penalties

Norm penalties are very standard and easy to work with but other types of penalties also exist. Thus, we can consider penalties[6] in a very general framework:

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

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

- Non-negative garotte:

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

- 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:

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

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 (2010)) has little difference with SCAD in terms of selected variables. A comparative study between them can be found in Zhang (2007). Both can be solved using the local quadratic approximation algorithm developed by (Fan and Li 2001).

One thing with penalty method is that there are always some penalty parameters (e.g., λ in LASSO) that have to be chosen. 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[7].

---

[6] For the purpose of illustration, one can refer to Fan and Li (2001) and Kim et al. (2008).

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

Usually, hyperparameter tuning seeks an optimal value according to some external criteria. The general way to proceed is to use Leave-One-Out Cross-Validation (LOOCV), Fold Cross-Validation (FCV) or General Cross-Validation (GCV). Only some data are used to fit the model and the Mean Squared Error (MSE) of predictions over the unused part is minimized to determine the best value of the hyperparameter. However, doing so often implies a bias in selection, as shown by Varma and Simon (2006) and Cawley and Talbot (2010). Several attempts have been made to overcome this issue such as "nested cross-validation", "weighted mean correction" and "Tibshirani's procedure", for which detailed explanations can be found in the work of Ding et al. (2014). In R, the package **ncvreg** (Breheny and Huang 2011) implements SCAD as well as MCP criteria while non-negative garotte is implemented in the package **lqa** (Ulbricht 2012). In Matlab, SCAD and MCP can be found in the toolbox **SparseReg** available on github (http://hua-zhou.github.io/SparseReg/) while some attempts to implement non-negative garotte can be found on Matlab Central. Only SCAD is available in SAS via the command `PROC SCADLS`.

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.

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, $\hat{\omega}(x_j, y) = \tilde{\mathbf{x}}_\mathbf{j}\tilde{\mathbf{y}}$, and gives a ranking of the $\mathbf{x_j}$. 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 removes 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 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 only used for the screening part that is ranking and reducing the set of candidates.

Because SIS may encounter the issue of selecting weakly correlated variables (weak signal-to-noise ratio), Fan and Lv (2008) introduced iterative conditional SIS, which applies correlation ranking but conditional on selected features. This is equivalent to 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 are developed to infer LASSO (e.g., forward stagewise). The SIS method is available in R using the package **SIS** (Saldana and Feng 2018) or in SAS via the `PROC GLMSELECT` command.

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 $\boldsymbol{\beta}$ is sparse (i.e., rare) and has high values

(strong signals). In this case, methods such as SCAD are even bias-free. However, if the signals are weak as well as rare, then they do not 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 do not require more technology than the simple OLS on the transformed decorrelated variables. The main issue is that selecting the number of latent variables to be removed, which can be done via cross-validation for instance, remains difficult.

4. Grouped Models

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

$$\mathbf{y} = \sum_{g=1}^{G} \mathbf{X_g}\boldsymbol{\beta}_g + \boldsymbol{\varepsilon}. \tag{11}$$

where the regressors are divided into G non-overlapping groups. Within this framework, there are two 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 LASSO framework. The objective is to solve a modified LASSO:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X_g}\boldsymbol{\beta}_g\|_2^2 + \lambda \sum_{g=1}^{G} c_g(\boldsymbol{\beta}_g' \mathbf{R_g}\boldsymbol{\beta}_g)^{1/2}. \tag{12}$$

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 thus make coefficients comparable in some sense. It can be easily shown that this leads to LASSO solution with standardization of regressors; when the group is formed with only one variable, it is often done empirically and is even advised by LASSO's authors.

An obvious extension is to take into account any penalty, providing the following objective:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \sum_{g=1}^{G} \sum_{j \in g} \beta_j \mathbf{x_j}\|_2^2 + p(\sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_{R_g}; c_g\lambda, \gamma) \tag{13}$$

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

4.1.2. Bi-Level

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

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2. \tag{14}$$

This method overcomes the issue of collinearity because it favours selection of correlated regressors simultaneously while LASSO tends to select only one out of them. In fact, the EasticNet can be solved as 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 applying LASSO. A small correction due to the second penalty λ_2 is required. Originally, the Elastic-net was not designed explicitly for grouped structure.

In addition, 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.

All group-related methods can be found in the R package **grpreg** (Breheny and Huang 2015) except the ElasticNet, which is available in **glmnet**. In SAS, `PROC GLMSELECT` has all the previously cited features available (i.e., LASSO, AdaLASSO, etc.) for group selection and also includes the ElasticNet. In Matlab, however, only the ElasticNet is provided in the "Statistics and Machine Learning Toolbox" using the function `lasso`. Implementation for grouped structure can be found on Matlab Central.

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:

$$\mathbf{y} = \sum_{j=1}^{p} f_j(\mathbf{x_j}) + \boldsymbol{\varepsilon}. \tag{15}$$

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 LASSO to functions non-parametrically fitted with parametric coefficients coming on top of them. This is obviously the most natural extension of LASSO to the additive structure. The main program to solve is:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \sum_{j=1}^{p} \beta_j f_j(\mathbf{x_j})\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{16}$$

Even though the term $\sum_{j=1}^{p} \beta_j f_j(\mathbf{x_j})$ might remind us of the very well-known Splines where the f_j would be the basis functions, the authors claimed 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 was investigated by Meier et al. (2009): the penalized General Additive Model (penGAM). It applies to the same models as before but is 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[8] to have a smooth estimates, which 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:

$$\min_{f} \|\mathbf{y} - f(\mathbf{X})\|_2^2 + J(f) \tag{17}$$

with the sparsity-smoothness penalty being:

$$J(f) = \lambda_1 \sqrt{\|f_j\|^2 + \lambda_2 \int (f_j''(x))^2 dx} \tag{18}$$

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 $b(\cdot)$, the problem can be written as:

$$\min_{\beta} \|y - B\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_j' \mathbf{B}_j' \mathbf{B}_j \boldsymbol{\beta}_j + \lambda_2 \boldsymbol{\beta}_j' \boldsymbol{\Omega}_j \boldsymbol{\beta}_j} \tag{19}$$

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

Both methods have been implemented in R. SpAM is available in the package **SAM** (Zhao et al. 2014), while penGAM can be found in the package **hgam** (Eugster et al. 2013).

5.2. Screening

In an equivalent manner on the screening side ,the Non-parametric Independence Screening procedure (NIS) was introduced by Fan et al. (2011) as a natural extension to SIS. Instead of marginal linear correlation, they used the concept of "marginal utility", already defined by Fan et al. (2009) for generalized linear models, and here set this marginal utility to be the sum of squared marginal residuals resulting from a non-parametric additive model:

$$\hat{\omega}_j = \sum_{i=1}^{n} \left(\mathbf{y_i} - \hat{f}_j(\mathbf{x_{i,j}}) \right)^2. \tag{20}$$

The latter, with $\hat{f}_j(\mathbf{x_{i,j}})$ obtained by splines[9], gives a ranking of variables in the same way as SIS:

$$\widehat{M} = \{1 \leq j \leq p : \hat{\omega}_j > \delta\}. \tag{21}$$

where δ is a predefined threshold. Usually, this step does not ensure selection consistency so they relied on a external procedure, namely SpAM or penGAM. The problem of weak signals in iterative conditional SIS is exactly the same as it is for SIS, i.e., applying NIS on residuals, conditionally on primarily selected variables. It is worth mentioning 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.

$$\omega_j = Cov(f_j(\mathbf{x_j}), G(\mathbf{y}))^2 \quad \text{with} \quad G(\mathbf{y}) \quad = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{y_i} \leq \mathbf{y}). \tag{22}$$

---

[8] Usually the ridge because it has an analytical solution.

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

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

6. Partial Linear Models

A Partial Linear model takes the form:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta} + g(\mathbf{X_2}) + \boldsymbol{\varepsilon}. \tag{23}$$

An important feature of these models is to assume two sets of variables. The $\mathbf{X}$ matrix is divided into $\mathbf{X_1}$ and $\mathbf{X_2}$ of dimension $p1$ and $p2$, respectively. The motivation behind this is to say that linearity may be satisfactory enough for some variables and treating these ones non-parametrically would result in a loss of efficiency. This notation is held throughout the whole section. This section is divided in two parts. The first one concerns partial linear models in their general form. Because a great literature has focused on smoothly varying-coefficients, the second part focuses only on them.

6.1. Standard

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. In the splines settings, one can rewrite this function as a linear combination of basis expansion:

$$\mathbf{g} = [\mathbf{J}, \mathbf{X_2}]\delta + \mathbf{Ba} \tag{24}$$

with $\mathbf{J}$ the unit vector $\mathbf{a}$ are the parameters of the basis expansion $\mathbf{B}$ and δ is the overall parameter on $\mathbf{X_2}$. The SCAD penalty is then applied on the vector $\boldsymbol{\beta}^* = [\boldsymbol{\beta}, \delta]$. This can be viewed as a composite penalty where the key idea is to take advantage of the linear rewriting of the problem provided by the splines representation. This allows for usual linear model selection in a non-linear setting. Partial Splines with Adaptive penalty (PSA) of Cheng et al. (2015) tries to achieve a sparse parametric part while having a non-parametric part aside using a combination of adaptive LASSO on the parametric part and penalized splines for the non-parametric. Therefore, the problem to solve is:

$$\min_{\boldsymbol{\beta},\mathbf{f}} \|\mathbf{y} - \mathbf{X_1}\boldsymbol{\beta} - f(\mathbf{X_2})\|^2 + \lambda_1 \int_0^1 (f''(\mathbf{X_2}))^2 dX_2 + \lambda_2 \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}. \tag{25}$$

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 also 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:

$$\min_{\boldsymbol{\beta},\mathbf{f}} \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|^2 + n\lambda_1 \sum_{j=1}^{p} w_{1,j}\|\beta_j\|_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2,j}\|\beta_j\|_{D_j}. \tag{26}$$

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

In the end, there is little difference among the three procedures. All exploit 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:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta} + \mathbf{X_2}\alpha(\mathbf{Z}) + \boldsymbol{\varepsilon}. \tag{27}$$

The coefficients α associated to each $\mathbf{x}_{j\in 2}$ are supposed to vary smoothly along another variable $\mathbf{Z}$. 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 in this form:

$$\mathbf{y} = \mathbf{X}\beta(\mathbf{Z}) + \boldsymbol{\varepsilon}. \tag{28}$$

The kernel LASSO of Wang and Xia (2009) deals with this problem in the spirit of groupLASSO (see Section 4.1.1). The varying coefficients are fitted using kernel regression (Nadaraya (1964); Watson (1964)). The problem to solve is:

$$\min_{\beta} \sum_{t=1}^{n}\sum_{i=1}^{n} \{\mathbf{y_i} - \mathbf{X_i}\beta(\mathbf{Z_t})\}^2 \mathbf{K_h}(\mathbf{Z_t} - \mathbf{Z_i}) + \sum_{j=1}^{p} \lambda_j \|\beta_j\|. \tag{29}$$

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 do not (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:

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

with the penalty applied on a different norm from the kernel LASSO:

$$\|\mathbf{b_j}\| = \left\{ \sum_{t=2}^{n} (\beta_j(\mathbf{Z_t}) - \beta_j(\mathbf{Z_{t-1}}))^2 \right\}^{1/2}. \tag{31}$$

6.2.2. Testing/Penalty

The Semi-Parametric Generalized Likelihood Ratio Test (SP-GLRT) of Li and Liang (2008) 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:

$$\mathcal{L}(\alpha, \beta) = l(\alpha, \beta) - n \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|). \tag{32}$$

The two parts are estimated alternatively conditionally on the other. One can note that it relies on the underlying choice of the penalty, making it a mixture of testing and penalty. Then, they introduce a novel generalized likelihood ratio test:

$$\mathcal{T}_{GLR} = r_K\{\mathcal{R}(H_1) - \mathcal{R}(H_0)\} \tag{33}$$

with

$$\mathcal{R}(H_1) = \mathcal{Q}(X_1\beta + X_2\alpha(Z), y). \tag{34}$$

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

$$\mathcal{R}(H_0) = \mathcal{Q}(X_1\beta, y) \tag{35}$$

The conditional likelihood under H_0: *the variable does not appear in the non-parametric part*.

$$\mathcal{Q}(\mu, y) = \int_{\mu}^{y} \frac{s-y}{V(s)} ds. \tag{36}$$

The test is then evaluated using a Monte Carlo or bootstrap method to empirically estimate the distribution of the statistics since the theoretical degrees of freedom tends to infinity, preventing a parametric test.

This has to be noticed because this is one of the first attempts to introduce non-parametric and therefore automatic tests inside a selection procedure. While methods such as autometrics and stepwise regression rely 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:

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} \tag{37}$$

where $f(\cdot)$ is any multivariate function, linear or not, additive or not. This framework is very general, therefore making it more complicated to estimate. 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 $\mathbf{X}$: p. It is already complicated to fit such a (perhaps 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 in several aspects. Testing ones such as MARS shares similarities with stepwise regression for example, in an ANOVA splines settings. Penalty ones use ANOVA models as well because they limit interaction terms and get closer to an additive model, which is indeed very common when dealing with fully non-parametric regression. The screening based ones can be divided into two categories: some make the use of generalized correlations to avoid using a model (DC-SIS, HSIC-SIS, KCCA-SIS, and Gcorr)[10], while others rely on a specific model ex ante (MDI, MDA, RODEO)[11].

---

[10] Distance Correlation-SIS, Hilbert Schmidt Independence Criterion-SIS, Kernel Canonical Correlation Analysis and the Generalized Correlation, respectively.

[11] Mean Decrease Impurity, Mean Decrease Accuracy and the Regularization Of Derivative Expectation Operator, respectively.

7.1. Testing

Introduced by Friedman (1991), multivariate adaptive regression splines is a method for building non-parametric fully non-linear ANOVA sparse models (Equation (40)). The model is written in terms of splines as:

$$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K} c_k B_k(\mathbf{x}). \tag{38}$$

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

$$B_k(x, \alpha, \beta) = \beta max(0, \alpha + x). \tag{39}$$

Therefore, α can be considered as "knots" such as 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[12]. 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 advised to limit the number of interactions m to a small value such as two so the model can be build in a reasonable time. Selection of variables is part of the building process. If using a fit based criterion such as 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. In addition, 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 it is piecewise linear, thus 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, which is discussed in the next subsections. Using directly MARS is more similar to a non-linear version of stepwise regression using piecewise functions. This procedure is implemented in most statistical softwares. In R, the package **earth** (Milborrow 2018) has implemented the MARS procedure. In SAS, it is available in the `ADAPTIVEREG` procedure. In Matlab, the toolbox **ARESlab** provides an implementation. This toolbox can be downloaded from http://www.cs.rtu.lv/jekabsons/regression.html.

7.2. Penalty

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

$$f(\mathbf{X}) = \sum_{j=1}^{p} f_j(\mathbf{x_j}) + \sum_{j<k} f_{j,k}(\mathbf{x_j}, \mathbf{x_k}) + ... + \boldsymbol{\varepsilon}. \tag{40}$$

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[13] but with a different penalty. The authors stated that the penalty should not be the same for main effect as for two-way interactions. They advocated the fact that ceteris paribus including an interaction term adds more regressors than a main effect and thus they are less interpretable. Thus, interactions should be

---

[12] These are known as "greedy algorithms" where the optimal global solution is sought by taking optimal local solutions.

[13] They also introduced it as SpIn (SpAM with INteractions) in their paper but claimed that interactions would then not be treated efficiently.

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:

$$\min_f \|\mathbf{y} - f(\mathbf{X})\|_2^2 + \tau^2 J(f) \tag{41}$$

with

$$J(f) = \lambda_1 \sum_{j=1}^{p} \left(\|f_j\|^2 + \sum_{k \neq j} \|f_{j,k}\|^2 \right)^{1/2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=j+1}^{p} \|f_{j,k}\|. \tag{42}$$

The penalty is written so that the first part penalizes additional regressors while the second penalizes interactions occurring without main effects. In SHIM, there is 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 and Zhang (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:

$$J(f) = \sum_{\alpha=1}^{p(p-1)/2} \|P^\alpha f\|^2. \tag{43}$$

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 claimed 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. COSSO has been implemented in R through the package **cosso** (Zhang and Lin 2013).

7.3. Screening

7.3.1. Model-Free

In the screening literature of non-parametric methods, we found many papers that deal with the same core idea. They all define some association measure that generalizes usual linear correlation. The following is a 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 (DC) is a generalization of the Pearson Correlation Coefficient in terms of norm distances. It can be written as:

$$\omega_j = \frac{dcov(\mathbf{x}, \mathbf{y})}{\sqrt{dcov(\mathbf{x}, \mathbf{x}) dcov(\mathbf{y}, \mathbf{y})}} \tag{44}$$

  where

$$\begin{aligned} dcov(\mathbf{x}, \mathbf{y})^2 = &\mathbb{E}[\|\mathbf{X} - \mathbf{X}'\| \|\mathbf{Y} - \mathbf{Y}'\|] \\ &+ \mathbb{E}[\|\mathbf{X} - \mathbf{X}'\|] \mathbb{E}[\|\mathbf{Y} - \mathbf{Y}'\|] \\ &- 2\mathbb{E}[\mathbb{E}[\|\mathbf{X} - \mathbf{X}'\|] \mathbb{E}[\|\mathbf{Y} - \mathbf{Y}'\|]]. \end{aligned} \tag{45}$$

- HSIC-SIS (Balasubramanian et al. 2013)

The Hilbert–Schmidt Independence Criterion (HSIC) generalizes the previous one as it defines a maximum distance metric in a RKHS space:

$$\begin{aligned}\omega^2_{(k)} &= \mathbb{E}[k_{\mathcal{X}}(\mathbf{X},\mathbf{X}')k_{\mathcal{Y}}(\mathbf{Y},\mathbf{Y}')] \\ &+\mathbb{E}[k_{\mathcal{X}}(\mathbf{X},\mathbf{X}')]\mathbb{E}[k_{\mathcal{Y}}(\mathbf{Y},\mathbf{Y}')] \\ &-2\mathbb{E}[\mathbb{E}[k_{\mathcal{X}}(\mathbf{X},\mathbf{X}')]\mathbb{E}[k_{\mathcal{Y}}(\mathbf{Y},\mathbf{Y}')]].\end{aligned} \tag{46}$$

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

$$\gamma = \sup\left\{\omega_{(k)} : k \in \mathcal{K}\right\}. \tag{47}$$

Empirically, the ranking measure is simpler to compute:

$$\hat{\omega} = \frac{1}{n}\sup_{k_{\mathcal{X}},k_{\mathcal{Y}}}\sqrt{trace(\boldsymbol{K}_{\boldsymbol{\mathcal{X}}}\boldsymbol{H}\boldsymbol{K}_{\boldsymbol{\mathcal{Y}}}\boldsymbol{H})} \tag{48}$$

with $\mathbf{H} = \mathbf{I} - (1/n)\mathbf{J}\mathbf{J}'$, I being the $n \times n$ unit matrix and $\mathbf{J}$ the $n \times 1$ unit vector.

- KCCA-SIS Liu et al. (2016)

The Kernel Canonical Correlation Analysis (KCCA) 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 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:

$$\boldsymbol{\mathcal{R}}_{YX} = \boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2}. \tag{49}$$

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

$$\boldsymbol{\mathcal{R}}_{YX} = (\boldsymbol{\Sigma}_{YY} + \epsilon\mathbf{I})^{-1/2}\boldsymbol{\Sigma}_{YX}(\boldsymbol{\Sigma}_{XX} + \epsilon\mathbf{I})^{-1/2}. \tag{50}$$

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

$$\omega(\epsilon)_j = \|\boldsymbol{\mathcal{R}}_{YX_j}\|. \tag{51}$$

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.

$$\widehat{M} = \{1 \leq j \leq p : \hat{\omega}_j \geq \delta\} \tag{52}$$

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

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

Another of the same kind is the Generalized Correlation Screening (Gcorr) of Hall and Miller (2009), introduced as a more general method than NIS (see Section 5.2). This coefficient is used as the measure of non-linear relationship. It can be defined as:

$$\hat{\omega}_j = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^{n} \{h(\mathbf{x_{i,j}}) - \bar{h}_j\}(\mathbf{y_i} - \bar{\mathbf{y}})}{\sqrt{n \sum_{i=1}^{n} \{h(\mathbf{x_{i,j}})^2 - \bar{h}_j^2\}}}. \tag{53}$$

Then, these estimates are tested using bootstrap confidence interval instead of threshold as 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. Only DC-SIS and Gcorr have been implemented in R. The former can be found in the packages **VariableScreening** (Li et al. 2018) and **cdcsis** (Wen et al. 2014). The latter is available from the author as a supplementary file to their paper. This file contains R code for the Generalized Correlation Coefficient and it can be found at https://www.tandfonline.com/doi/suppl/10.1198/jcgs.2009.08041?scroll=top.

7.3.2. Model Based

The Regularization of Derivative Expectation Operator (RODEO) of Lafferty and Wasserman (2008), named in reference to LASSO, applies in the framework of multivariate kernel methods. In kernel regression, 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 fewer observations enter the local regression, leading to less bias but more variance and conversely for a high bandwidth. The authors states 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 fewer observations do not change the fitted curve. For a Gaussian kernel we have:

$$\begin{aligned} \frac{\partial f_h(\mathbf{x})}{\partial h_j} &= \mathbf{e}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\frac{\partial \mathbf{W}}{\partial h_j}(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ \frac{\partial \mathbf{W}}{\partial h_j} &= \mathbf{W}\mathbf{L}_j \\ \mathbf{L}_j &= \frac{1}{h_j^3} diag((\mathbf{x}_{1,j} - \bar{\mathbf{x}}_j)^2, ..., (\mathbf{x}_{n,j} - \bar{\mathbf{x}}_j)^2). \end{aligned} \tag{54}$$

Note that it refers to a specific point in the sample $\bar{x}$. The derivative is not computed over the whole sample. The authors proposed 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 lernel, therefore it suffers form the curse of dimensionality. 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, and splines and kernels for non-linear ones. However, other ways for doing regression exist, from which model selection procedures intuitively

arise. In a Bayesian framework[14], 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[15]. However, we can exploit this distribution across these different models to assign probabilities to each variables, since they may not all appear in every model. 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 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 allow for constant estimation in subregions, whcih 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 [16] of the data in the partition. The mean decrease impurity is defined as:

$$MDI(x_j) = \frac{1}{N_t} \sum_T \sum_{t \in T} \frac{N_t}{N} \left(i(t) - \frac{N_{t_{left}}}{N_t} i(t_{left}) - \frac{N_{t_{right}}}{N_t} i(t_{right}) \right). \tag{55}$$

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 to 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. The mean decrease accuracy is defined as:

$$MDA(x_j) = \frac{\sum_{T \in F} VI^T(x_j)}{N_T} \tag{56}$$

with

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

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 gives the decrease in accuracy for each variable and the whole is a weighted average of each tree in the forest.

Both measures can be found in software after fitting decision trees or random forests. These methods are available in the R package **randomForest** (Liaw and Wiener 2002), in the Matlab "Statistics and Machine Learning Toolbox" under the function `predictorImportance` and in SAS using the `PROC HPFOREST` command.

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

---

[14] Which is out of the scope of this paper but still very important.
[15] This relates obviously to the problem raised when discussing stepwise regression. Here, the ensemble is a subset of the model space.
[16] In the case of a regression, it is how well the subregion can be approximated by a constant.

under some conditions. For example, we remember LASSO where the regularization parameter λ cannot be chosen optimally from an estimation and a predictive point of view so that it ensures correct identification. This led to adaptive LASSO (Zou 2006), but this problem can also be solved using these procedures.

8.1. Stability Selection

Stability Selection (Stabsel) was 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[17] (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:

$$\Pi_j^\lambda = \mathbb{P}(j \subseteq S^\lambda). \tag{58}$$

The set of stable variables is then:

$$S^{Stable} = \{j : \max_{\lambda \in \Lambda} \Pi_j^\lambda \geq \delta\}. \tag{59}$$

This is given by the underlying selection procedure, be it LASSO or another procedure, 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 into 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.

$$\mathbb{E}[V] \leq \frac{1}{2\delta - 1} \frac{q_\lambda^2}{p}. \tag{60}$$

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

$$\begin{aligned} \delta &= 0.9 \\ q_\lambda &= \sqrt{0.8 \alpha p}. \end{aligned} \tag{61}$$

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 were proposed by 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. StabSel has been implemented in R through the package **stabs** (Hofner and Hothorn 2017).

8.2. Ranking-Based Variable Selection

The Ranking-Based Variable Selection (RBVS) of Baranowski et al. (2018) is a screening procedure based on bootstrap and permutation tests. Contrary to Stabsel, it does not rely on any threshold or 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), it constructs a permutation ranking $\mathcal{R}$:

$$\mathcal{R} = (\mathcal{R}_1, ..., \mathcal{R}_p) \text{ satisfying } \omega_{\mathcal{R}_1} \geq ... \geq \omega_{\mathcal{R}_p}. \tag{62}$$

[17] Without replacement, random samples have to be non-overlapping.

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

$$\pi(\mathcal{A}_k) = \mathbb{P}(\{\mathcal{R}_1, ..., \mathcal{R}_k\} = \mathcal{A}). \tag{63}$$

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. An alternative is trying to estimate k^* as:

$$k^* = argmin_{k=0,...,p-1} \frac{\pi(\mathcal{A}_{k+1,m})}{\pi(\mathcal{A}_{k,m})}. \tag{64}$$

Instead of an absolute threshold, they rather used a relative threshold. The optimal number of included variables is reached when the sequence of $\pi(\mathcal{A})$ declines the most. This is equivalent to looking for the value of k that best separates assuming there are two sets: the relevant and the irrelevant. Similar to SIS having its iterative counterpart, they introduced the iterative RBVS that accounts for marginally related variables with low signal-to-noise and for the multicollinearity problem. RBVS is available in the R package **rbvs** (Baranowski et al. 2015) developed by the authors.

9. Discussion

In this article, we review numerous 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 into three groups: test-based, penalty-based and screening-based (see Table 1). They have been described and compared on the grounds 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.

Selection consistency for widely used methods in empirical work has been discussed and several improvements were presented. Far beyond Stepwise regression and 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 and 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 many 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 seeking very complex models in high dimensions. Sure independence screening is a powerful tool in linear models but has lower dataset capacities when it comes to non-linearities. In addition, the literature is lacking very complete algorithmic solutions. To the best of our knowledge, no statistical procedures 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.

Funding: This research received no external funding.

Acknowledgments: I thank Costin Protopopescu, Maître de conférences, AMSE, Aix-Marseille Université, Faculté d'économie et de gestion (FEG) and Emmanuel Flachaire, Professeur des universités, AMSE, Aix-Marseille Université, Faculté d'économie et de gestion (FEG) for comments that greatly improved the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

Abenius, Tobias. 2012. Lassoshooting: L1 Regularized Regression (Lasso) Solver Using the Cyclic Coordinate Descent algorithm aka Lasso Shooting. R Package Version 0.1.5-1. Available online: https://CRAN.R-project.org/package=lassoshooting (accessed on 15 November 2018).

Akaike, Hirotugu. 1973. Information Theory and an Extension of Maximum Likelihood Principle. Paper presented at 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, September 2–8, pp. 267–81.

Bach, Francis R. 2008. Bolasso: Model Consistent Lasso Estimation through the Bootstrap. Paper presented at 25th International Conference on Machine Learning, Helsinki, Finland, July 5–9, pp. 33–40.

Balasubramanian, Krishnakumar, Bharath Sriperumbudur, and Guy Lebanon. 2013. Ultrahigh dimensional feature screening via rkhs embeddings. *Artificial Intelligence and Statistics* 31: 126–34.

Baranowski, Rafal, Patrick Breheny, and Isaac Turner. 2015. rbvs: Ranking-Based Variable Selection. R Package Version 1.0.2. Available online: https://CRAN.R-project.org/package=rbvs (accessed on 15 November 2018).

Baranowski, Rafal, Yining Chen, and Piotr Fryzlewicz. 2018. Ranking-based variable selection for high-dimensional data. *Statistica Sinica*, in press. [CrossRef]

Bickel, Peter J., Friedrich Götze, and Willem R. van Zwet. 2012. *Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses*. New York: Springer, pp. 267–97.

Blum, Avrim L., and Pat Langley. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97: 245–71. [CrossRef]

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3: 1–122. [CrossRef]

Breaux, Harold J. 1967. *On Stepwise Multiple Linear Regression*. Technical Report. Aberdeen: 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: 369. [CrossRef] [PubMed]

Breheny, Patrick, and Jian Huang. 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* 5: 232–53. [CrossRef] [PubMed]

Breheny, Patrick, and Jian Huang. 2015. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing* 25: 173–87. [CrossRef] [PubMed]

Breiman, Leo, and Jerome H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80: 580–98. [CrossRef]

Breiman, Leo. 1995. Better subset regression using the nonnegative garrote. *Technometrics* 37: 373–84. [CrossRef]

Breiman, Leo. 2001. Random forests. *Machine Learning* 45: 5–32. [CrossRef]

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

Castle, Jennifer L., and David F. Hendry. 2010. A low-dimension portmanteau test for non-linearity. *Journal of Econometrics* 158: 231–45. [CrossRef]

Cawley, Gavin C., and Nicola L. C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11: 2079–107.

Chen, Xueying, and Min-Ge Xie. 2014. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* 24: 1655–84.

Cheng, Guang, Hao H. Zhang, and Zuofeng Shang. 2015. Sparse and efficient estimation for partial spline models with increasing dimension. *Annals of the Institute of Statistical Mathematics* 67: 93–127. [CrossRef] [PubMed]

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: 354–64. [CrossRef]

Ding, Ying, Shaowu Tang, Serena G. Liao, Jia Jia, Steffi Oesterreich, Yan Lin, and George C. Tseng. 2014. Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics* 30: 3152–58. [CrossRef] [PubMed]

Doornik, Jurgen A. 2009. *Econometric Model Selection with More Variables Than Observations*. Oxford: Economics Department, University of Oxford. Unpublished Work.

de Rooi, Johan and Paul Eilers. 2011. Deconvolution of pulse trains with the L0 penalty. *Analytica Chimica Acta* 705: 218–26. [CrossRef] [PubMed]

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32: 407–99.

Epprecht, Camila, Dominique Guegan, Álvaro Veiga, and Joel Correa da Rosa. 2017. *Variable Selection and Forecasting via Automated Methods for Linear Models: Lasso/adalasso and Autometrics*. Documents de travail du Centre d'Economie de la Sorbonne 2013.80. Paris: Centre d'Economie de la Sorbonne.

Eugster, Manuel, Torsten Hothorn, The Students of the 'Advanced R Programming Course' Hannah Frick, Ivan Kondofersky, Oliver S. Kuehnle, Christian Lindenlaub, Georg Pfundstein, Matthias Speidel, Martin Spindler, Ariane Straub, and et al. 2013. hgam: High-Dimensional Additive Modelling. R Package Version 0.1-2. Available online: https://CRAN.R-project.org/package=hgam (accessed on 15 November 2018).

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: 544–57. [CrossRef] [PubMed]

Fan, Jianqing, and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348–60. [CrossRef]

Fan, Jianqing, and Jinchi Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* 70: 849–911. [CrossRef] [PubMed]

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

Fan, Jianqing, and Jinchi Lv. 2010b. Sure Independence Screening. *R Package Version*. Available online: https://cran.r-project.org/web/packages/SIS/SIS.pdf (accessed on 15 November 2018).

Fan, Jianqing, Richard Samworth, and Yichao Wu. 2009. Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research* 10: 2013–38. [PubMed]

Fan, Jianqing, and Wenyang Zhang. 2008. Statistical methods with varying coefficient models. *Statistics and Its Interface* 1: 179. [CrossRef] [PubMed]

Flom, Peter L., and David L. Cassell. 2007. Stopping Stepwise: Why Stepwise and Similar Selection Methods Are Bad, and What You Should Use. Paper presented at NorthEast SAS Users Group Inc 20th Annual Conference, Baltimore, MD, USA, November 11–14.

Frank, Ildiko, and Jerome H. Friedman. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35: 109–35. [CrossRef]

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33: 1. [CrossRef] [PubMed]

Friedman, Jerome H. 1991. Multivariate adaptive regression splines. *The Annals of Statistics* 19: 1–67. [CrossRef]

Fu, Wenjiang J. 1998. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7: 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: 533–50. [CrossRef]

Hannan, Edward J., and Barry G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B* 41: 190–95.

Hastie, Trevor, and Bradley Efron. 2013. Lars: Least Angle Regression, Lasso and Forward Stagewise. R Package Version 1.2. Available online: https://CRAN.R-project.org/package=lars (accessed on 15 November 2018).

Hendry, David F., and Jean-Francois Richard. 1987. *Recent Developments in the Theory of Encompassing*. Technical Report. Louvain-la-Neuve: 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: 55–67. [CrossRef]

Hofner, Benjamin, and Torsten Hothorn. 2017. Stabs: Stability Selection with Error Control. R Package Version 0.6-3. Available online: https://CRAN.R-project.org/package=stabs (accessed on 15 November 2018).

Hu, Tao, and Yingcun Xia. 2012. Adaptive semi-varying coefficient model selection. *Statistica Sinica* 22: 575–99. [CrossRef]

Huang, Jian, Patrick Breheny, and Shuange Ma. 2012. A selective review of group selection in high-dimensional models. *Statistical Science* 27. [CrossRef] [PubMed]

Huang, Jian, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang. 2009. A group bridge approach for variable selection. *Biometrika* 96: 339–55. [CrossRef] [PubMed]

Hurvich, Clifford M., and Chih-Ling Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76: 297–307. [CrossRef]

Hurvich, Clifford M., and Chih-Ling Tsai. 1990. The impact of model selection on inference in linear regression. *The American Statistician* 44: 214–17.

Jović, Alan, Karla Brkić, and Nikola Bogunović. 2015. A Review of Feature Selection Methods with Applications. Paper presented at 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, May 25–29, pp. 1200–5.

Ke, Tracy, Jiashun Jin, and Jianqing Fan. 2014. Covariate assisted screening and estimation. *The Annals of Statistics* 42: 2202. [CrossRef] [PubMed]

Ke, Tracy, and Fan Yang. 2017. Covariate assisted variable ranking. *arXiv*. 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: 1665–73. [CrossRef]

Kowalski, Matthieu. 2014. Thresholding Rules and Iterative Shrinkage/Thresholding Algorithm: A Convergence Study. Paper presented at 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, October 27–30, pp. 4151–55.

Lafferty, John, and Larry Wasserman. 2008. Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics* 36: 28–63. [CrossRef]

Li, Runze, Liying Huang, and John Dziak. 2018. VariableScreening: High-Dimensional Screening for Semiparametric Longitudinal Regression. R Package Version 0.2.0. Available online: https://CRAN.R-project.org/package=VariableScreening (accessed on 15 November 2018).

Li, Runze, and Hua Liang. 2008. Variable selection in semiparametric regression modeling. *The Annals of Statistics* 36: 261. [CrossRef] [PubMed]

Li, Runze, Wei Zhong, and Liping Zhu. 2012. Feature screening via distance correlation learning. *Journal of the American Statistical Association* 107: 1129–39. [CrossRef] [PubMed]

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* 25: 591–607.

Liaw, Andy, and Matthew Wiener. 2002. Classification and regression by randomforest. *R News* 2: 18–22.

Lin, Yi, and Hao H. Zhang. 2006. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* 34: 2272–97. [CrossRef]

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

Lumley, Thomas. 2017. Leaps: Regression Subset Selection. R Package Version 3.0. Available online: https://CRAN.R-project.org/package=leaps (accessed on 15 November 2018).

Mallows, Colin L. 1973. Some comments on cp. *Technometrics* 15: 661–75.

McIlhagga, William H. 2016. Penalized: A matlab toolbox for fitting generalized linear models with penalties. *Journal of Statistical Software* 72. [CrossRef]

Mehmood, Tahir, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. 2012. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 118: 62–69. [CrossRef]

Meier, Lukas, Sara Van de Geer, and Peter Buhlmann. 2009. High-dimensional additive modeling. *The Annals of Statistics* 37: 3779–821. [CrossRef]

Meinshausen, Nicolai, and Peter Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34: 1436–62. [CrossRef]

Meinshausen, Nicolai, and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B* 72: 417–73. [CrossRef]

Milborrow, Stephen. 2018. Earth: Multivariate Adaptive Regression Splines. R Package Version 4.6.2. Available online: https://CRAN.R-project.org/package=earth (accessed on 15 November 2018).

Nadaraya, Elizbar A. 1964. On estimating regression. *Theory of Probability & Its Applications* 9: 141–42.

Ni, Xiao, Hao H. Zhang, and Daowen Zhang. 2009. Automatic model selection for partially linear models. *Journal of Multivariate Analysis* 100: 2100–11. [CrossRef] [PubMed]

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: 36–64. [CrossRef]

Pretis, Felix, J. James Reade, and Genaro Sucarrat. 2018. Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software* 86: 1–44. [CrossRef]

Radchenko, Peter, and Gareth M. James. 2010. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association* 105: 1541–53. [CrossRef]

Ravikumar, Pradeep, Han Liu, John Lafferty, and Larry Wasserman. 2007. Spam: Sparse Additive Models. Paper presented at 20th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, December 3–6. Red Hook: Curran Associates Inc., pp. 1201–8.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507–17. [CrossRef] [PubMed]

Saldana, Diego Franco, and Yang Feng. 2018. Sis: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software* 83: 1–25. [CrossRef]

Santos, Carlos, David F. Hendry, and Soren Johansen. 2008. Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 23: 317–35. [CrossRef]

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6: 461–64. [CrossRef]

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* 75: 55–80. [CrossRef]

Steyerberg, Ewout W., Marinus J. C. 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: 935–42. [CrossRef]

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* 58: 267–88.

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* 67: 91–108. [CrossRef]

Ulbricht, Jan. 2012. lqa: Penalized Likelihood Inference for GLMs. R Package Version 1.0-3. Available online: https://CRAN.R-project.org/package=lqa (accessed on 15 November 2018).

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

Varma, Sudhir, and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *Bioinformatics* 7: 91. [PubMed]

Wang, Hansheng. 2009. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* 104: 1512–24. [CrossRef]

Wang, Hansheng, and Yingcun Xia. 2009. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* 104: 747–57. [CrossRef]

Wang, Lifeng, Guang Chen, and Hongzhe Li. 2007. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics* 23: 1486–94. [CrossRef] [PubMed]

Watson, Geoffrey S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 26: 359–72.

Weisberg, Sanford. 2005. *Applied Linear Regression*. Hoboken: John Wiley & Sons, vol. 528.

Wen, Canhong, Wenliang Pan, Mian Huang, and Xueqin Wang. 2014. cdcsis: Conditional Distance Correlation and Its Related Feature Screening Method. R Package Version 1.0. Available online: https://CRAN.R-project.org/package=cdcsis (accessed on 15 November 2018).

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: 1182–89. [CrossRef] [PubMed]

Wu, Tong Tong, and Kenneth Lange. 2008. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* 2: 224–44. [CrossRef]

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

Zhang, Cun-Hui. 2007. *Penalized Linear Unbiased Selection*. Camden: Rutgers University.

Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38: 894–942. [CrossRef]

Zhang, Hao H., and Chen-Yen Lin. 2013. cosso: Fit Regularized Nonparametric Regression Models Using COSSO Penalty. R Package Version 2.1-1. Available online: https://CRAN.R-project.org/package=cosso (accessed on 15 November 2018).

Zhang, Jing, Yanyan Liu, and Yuanshan Wu. 2017. Correlation rank screening for ultrahigh-dimensional survival data. *Computational Statistics & Data Analysis* 108: 121–32.

Zhao, Tuo, Xingguo Li, Han Liu, and Kathryn Roeder. 2014. SAM: Sparse Additive Modelling. R Package Version 1.0.5. Available online: https://CRAN.R-project.org/package=SAM (accessed on 15 November 2018).

Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101: 1418–29. [CrossRef]

Zou, Hui, and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* 67: 301–20. [CrossRef]