

Facing Multicollinearity in Data Mining

Data Mining e Multicollinearità

Isabella Morlini

Sezione di Statistica, Dipartimento di Economia, Università di Parma
isabella.morlini@unipr.it

Riassunto: Il presente studio riguarda il problema della scelta di un modello di regressione non lineare che si presenta nel *data mining* quando la funzione che lega una variabile dipendente ad un pluralità di variabili esplicative non è nota ma deve essere desunta dai dati. Viene mostrato come, in presenza di multicollinearità, la scelta del modello non possa essere basata unicamente sull'errore quadratico od indici ad esso collegati (ad esempio, AIC, BIC/SBC), in quanto alcuni modelli che utilizzano l'algoritmo di *backfitting* sono soggetti a grande instabilità ed arbitrarietà nella scelta delle funzioni di base. Il comportamento dei più noti metodi non lineari basati sia sulla *subset selection* sia sulla proiezione delle variabili, in presenza di multicollinearità, viene illustrato attraverso un esempio numerico.

Keywords: Data Mining, Nonlinear Projection Methods, Nonlinear Selection Methods.

1. Introduction

In regression problems, when the form of the relationship between a dependent variable and multiple predictors is not known *a priori*, non parametric models are often applied in order to extract knowledge from data and to adaptively build a regression function. One of the aim of data mining is to search for the best final model. Usually, model selection is based on the sum of squares error (SSE) or related indexes like BIC/SBC or AIC, and mathematical convenience, regardless of the problem at hand. When the predictors are highly collinear, however, nonlinear models like *General Additive Models* (GAMs), and *Multivariate Adaptive Regression Spline* (MARS) based on the backfitting algorithm (a stepwise procedure of systematically adding and removing basis functions) may present great instability and arbitrariness in the selection process, unlike linear models in which the original coordinate system is a meaningful one. For this reason, even if these tools produce simpler and more understandable models of a response to an arbitrary function of a set of predictor variables, projection methods like *Projection Pursuit Regression* (PPR), the *Multi-Layer Perceptron* (MLP) and *Radial Basis Function Networks* (RBFNs) or selection methods based on a different subset selection, like *classification and Regression Trees* (CART), may appear more suited when the data matrix shows high collinearity. It is the goal of this paper to understand how nonlinear methods based on the backfitting algorithm are affected by multicollinearity and to show that projection methods manage better with this problem. The structure of the paper is as follows. Section 2 briefly reviews the backfitting algorithm in GAMs and MARS. Section 3 focuses on a numerical example in order to understand further how these methods work and to compare them with other selection and projection tools. Section 4 provides concluding remarks.

2. The backfitting algorithm in GAMs and MARS

In the following we briefly review the backfitting algorithm in GAMs and MARS. We assume that the readers are familiar with these models and, due to space limitation, we do not describe them but we refer the readers to descriptions in Ripley (1996, cap. IV). The *backfitting* algorithm (Hastie and Tibshirani, 1990) adaptively builds a set of basis functions by forward selection. This technique works in the original coordinate system and finds linear and nonlinear combinations of these coordinates.

In GAMs the forward procedure holds all but one of the additive terms constant, removes that term and fits a smooth term to the residuals. The fitting is applied a variable at a time until the process converges. It is this procedure that makes GAMs vulnerable to collinearities between the dependent and the independent variables. If the first variables are correlated with the response, and the smooth term is flexible enough, then the partial residuals result in small values and the algorithm may converge before processing all variables. So the final model may depend on the order in which variables are presented. In a less extreme case, all predictors are selected as basis functions, but the degree of freedom of each basis function may arbitrarily depend on the order of the variables.

In MARS a tree structure is present and interaction between variables is explicitly allowed. The forward procedure is somehow different from GAMs. For each predictor and every possible value of these predictors (knot), MARS divides the data into two parts, one on either side of the knot. MARS selects the knot and variable pair which give the best fit, and to each part it fits the response using a pair of linear functions. If two variables are correlated, at same stage of the tree construction MARS may be forced to choose between placing a knot on one of these predictors. If both predictors result in roughly the same penalized residual sum of the squares, then the selection may be somehow arbitrarily and in the final set of basis function only one of these variables may be represented. In an extreme case it may happen that the choice of one variable at the current step may have a great impact on the choice of all further variables and knot selections and thus on the final model as well. The backward step, which follows the forward phase and aims to produce a model with comparable performance but fewer terms, is also vulnerable to multicollinearity, especially in the additive case (when no interaction is allowed) since over-fitting is avoided by reducing the number of knots rather than via a smoothness penalty.

In conclusion, MARS and GAMs are affected by multicollinearity in that they select the basis functions in some arbitrarily manner, since this choice has no impact on the SSE when a set of variables is highly correlated with each others and with the response. In addition, in many applications a subset selection of the predictors may not be the optimal choice since a weighted average of the input variables may be preferable to the single one with the highest correlation with the response (for example, in quality control, a weighted average of sensors may be preferable to a single one).

3. Numerical Example

To understand further how these methods work, it is useful to look at their behavior on a synthetic example. Let $\mathbf{p}$ be a (6×1) dimensional vector and $\mathbf{U}$ a (6×5) matrix, such that $\mathbf{p}'\mathbf{p}=1$, $\mathbf{U}'\mathbf{U}=\mathbf{I}$, $\mathbf{U}\mathbf{U}'=\mathbf{I}-\mathbf{p}\mathbf{p}'$, $\mathbf{p}'\mathbf{U}=\mathbf{0}$. Define the (100×6) matrix $\mathbf{X}$ of the dependent variables as $\mathbf{X}=\mathbf{z}\mathbf{p}'+\mathbf{V}\mathbf{U}'$ with $\mathbf{z}$ of dimension (100×1) and components generated from a

$N(0, 256)$, $\mathbf{V}$ of dimension (100×5) with the first column generated from a $N(0, 49)$, the second one from a $N(0, 0.0121)$ and the last three columns from a $N(0, 0.005)$. Define the vector $\mathbf{y}$ of the dependent variable as $\mathbf{y}=\mathbf{z}+\mathbf{e}$ with $\mathbf{e}$ generated from a $N(0, 0.1225)$. From a geometric point of view, $\mathbf{y}$ belongs to the space spanned by $\mathbf{z}$ and following the definition of Naes and Helland (1993) this space is *strongly relevant* to the prediction of $\mathbf{y}$. The correlation matrix $\mathbf{C}$ between predictive and dependent variables is:

$$\mathbf{C} = \begin{bmatrix} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & y \\ x_1 & 1 & & & & & & \\ x_2 & 0.96 & 1 & & & & & \\ x_3 & 0.85 & 0.96 & 1 & & & & \\ x_4 & -0.1 & 0.26 & 0.5 & 1 & & & \\ x_5 & 0.19 & 0.45 & 0.67 & 0.98 & 1 & & \\ x_6 & 0.32 & 0.57 & 0.76 & 0.94 & 0.99 & 1 & \\ y & 0.41 & 0.65 & 0.82 & 0.90 & 0.97 & 0.99 & 1 \end{bmatrix}$$

and is highly collinear. The peculiarity of this data set is that there are two underlying factors that give rise to the variables. This is clear from the pattern of the high and low correlations in $\mathbf{C}$ such that the variables in a particular subset have high correlations among themselves but low correlations with all the other variables. Nevertheless, only one of this factor is important for the prediction of $\mathbf{y}$, and this is less evident from $\mathbf{C}$, since all correlations between the predictors and the dependent variables are significant ($\alpha=0.01$). Although the subspace of dimension 1, relevant to the prediction of $\mathbf{y}$, is linear, methods like Principal Components Regression (PCR) and Partial Least Squares (PLS) extract 2 components (this is obvious for PCR, in which the data are transformed in new orthogonal variables before regression, but less obvious for PLS in which the projection and the reduction steps are faced simultaneously). For what concern nonlinear models, results are as follows. CARTs show stable behaviors: if we consider the set of only the last three variables and the full data set, in both cases, given the same parameters choice, they select variables x_5 and x_6 , which are the most correlated with the dependent variables, with the same model complexity (8 terminal nodes). MARS behaves in a different way: considering the subset and the full data set and restricting the interaction order to 2, we find that at the second knot placement the algorithm has to choose between x_5 and x_4 and the variable selected depends on the set of predictors used to build the model. Unfortunately, the choice between x_5 and x_4 results in approximately the same cross validated error, but in different choices in the subsequent stages of the hierarchy and in the pruning step. Models built have similar SSEs but different degrees of freedom (df) and different knots placement. GAMs, given the same parameters choice, result in different models, depending arbitrarily on the order in which the variables are presented. Since all predictors are correlated with output, the backfitting algorithm converges after having fit the linear part of the first two terms. For example, with smoothing splines, for variables presented in their original order ($x_1, \dots, x_6$) predictors with a parametric df are x_1 and x_2 ; for variables in the reverse order, predictors with a parametric df are x_6 and x_5 (Table 1). The two models have approximately the same cross validated SSE and equal number of total df. However, basis functions selected are different and quite arbitrary, since they depend on the order of the variables. The problem of knots placement in presence of multicollinearity is more understandable when using natural cubic or B-splines, since the GAMs reduce in a

parametric fit. Table 2 shows the parametric df for variables presented in the original and in the reverse order, with B-splines of degree 2 and 3.

Table 1: *Some results for GAMs with smoothing splines.*

	Degrees of the Smoothing Splines				Variables with a Parametric Degree of Freedom
	1	2	3	4	
Variables	Cross Validated SSE				
original order	2.36	3.47	5.63	8.75	x_1 and x_2
reverse order	2.35	3.66	6.05	8.78	x_5 and x_6

Table 2: *Degrees of Freedom for GAMs with B-splines (degrees 2 and 3).*

Variables	intercept	x_1	x_2	X_3	x_4	x_5	x_6
original order	1	3	3	2	1	0	0
reverse order	1	0	0	1	2	3	3

For what concern projection methods, having fit a PPR with a $\mathcal{R}^6$ projection space and considering the magnitude of the coefficients in the linear combination, we see that PPR correctly finds a $\mathcal{R}^1$ (linear) projection space, since the last 5 coefficients are really smaller than the first one. A sigmoidal MLP trained with a number of hidden units ranging from 1 to 30 gives the best cross-validated results with 1 hidden unit (that is, a $\mathcal{R}^1$ nonlinear projection space). In RBFNs the gaussian transformations over the Euclidean distances computed by the hidden units maintain the multicollinearity if the widths of the gaussians are too wide. Therefore, we use a RBFN with multi quadratic functions obtaining the best result with 2 basis functions.

Conclusion

Nonlinear selection models based on the backfitting algorithm are often liked better than non parametric projection methods, since they build simpler and more understandable models. However, they are affected by multicollinearity in that they select the knots placement in some arbitrarily manner, when this choice has not impact on the SSE. Hence, they may not be the optimal alternative in model building in presence of multicollinearity. Nonparametric methods like PPR and MLP are shown to find the correct dimension of the projection space relevant for prediction. RBFNs are found to give rise to numerical problems using the gaussian transformation. With a non localized function, they are shown to identify a projection space not far from the dimension of the relevant subspace.

Main References

- Hastie T.J., Tibshirani L.J. (1990) *Generalized Additive Models*, Chapman, London.
Naes T., Helland I. S.(1993) Relevant components in regression, *Scandinavian Journal of Statistics*, 20, 239-250.
Ripley B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK.