[Kostov, 2010]. This is a postprint of a research article. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning B: Planning and Design, volume 37, issue 3, pages* 533–549, *2010*, http://dx.doi.org/10.1068/b35137

Model boosting for spatial weighting matrix selection in spatial lag models

Philip Kostov
Lancashire Business School
University of Central Lancashire

Abstract

The spatial lag specification is often used in spatial econometrics. The choice of an appropriate spatial weighting matrix is an important outstanding methodological problem in the quantitative spatial dependence literature. This paper proposes applying a component-wise model boosting algorithm to deal with the issue of the choice of a spatial weighting matrix amongst a predetermined set of alternatives. The resulting procedure is computationally simple and easy to implement. We present an empirical application of the proposed methodology. Some possible extensions to a more general setting are discussed.

Keywords: spatial lag, spatial weighting matrix, model boosting

Introduction

The issues of social interaction are gaining prominence in economic literature. Examples of theoretical models explicitly considering such issues include the models of increasing returns, path dependence and imperfect competition that underline much of the new economic geography literature (see Fujita *et al*. 1999), neighbourhood spillover effects (Durlauf, 1994; Borjas, 1995; Glaeser *et al*., 1996) and the macroeconomic interaction models developed by Aoki,1996 and Durlauf, 1997. It is difficult to provide a consistent overview of this area, but any formal or informal analysis employing concepts such as social norms, social capital, neighbourhood effects, peer group effects, strategic interaction, reference behaviour or yardstick competition falls into this category. A common implication of this type of models is that they imply certain type of interaction that links together economic agents. This interaction can be represented as a variation over space, where 'space' may not necessary be defined in geographical sense, but may be based on other metrics, such as 'economic' or 'social' distances.

Brueckner, 2003 presents two theoretical frameworks for such interaction. In the first one known as the spillover model, the decisions taken by other economic agents enter directly into the objective function of the economic agent. The other framework is known as the resource flow model. In this case the objective function of an agent is only indirectly

[Kostov, 2010]. This is a postprint of a research article. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning B: Planning and Design, volume 37, issue 3, pages* 533–549, *2010*, http://dx.doi.org/10.1068/b35137

affected by the decisions taken by other agents. Typically this is due to the fact that this objective function includes the value of some 'resource' the availability of which depends on the decisions taken by the other agents. Both the spillover and the resource flow models lead to the spatial lag specification. Statistically the spatial lag specification can be expressed as follows:

$$y = \lambda W y + X\beta + u \qquad (1)$$

where the classical linear regression model is augmented by the inclusion of the spatially lagged dependent variable Wy. The spatial lag is represented by the spatial weighting matrix W which needs to be specified.

The need to pre-specify the spatial weighting matrix presents a serious challenge to empirical modelling, particularly since often there is very little guidance about how exactly to do it. This paper suggests using component-wise boosting to choose the appropriate spatial weighting matrix amongst a set of pre-determined alternatives. It shows how the spatial weighting matrix selection problem can be reformulated as a variable selection problem and thus standard variable selection techniques would be available. The paper is organised as follows. First we discuss the issues surrounding spatial dependence and the formulation of the spatial weighting matrix. Then some existing approaches dealing with the arbitrariness of the choice of spatial weighting are discussed. The general idea of our proposal is outlined and the proposed methodology is briefly compared to some of the more popular alternatives within the variable selection literature. After a detailed description of the underlying algorithms, we present an empirical example using agricultural land sales data from Northern Ireland. Finally, some possible extensions of the proposed framework are discussed.

Spatial dependence and spatial weighting matrix

Although we will only be considering the spatial lag specification, it would be important to note that this is not the only form of spatial dependence. An alternative is the so called spatial error representation. It would be useful to distinguish these two forms of spatial dependence. In principle the exact causes of spatial dependence determine whether it is spatial lag or spatial error. Spatial dependence may for example arise simply because economic agent independently adopt similar behaviour, because e.g. of the underlying institutional arrangements. If so, the spatial dependence observed in our data does not

reflect a truly spatial process, but merely spatial clustering of the sources of the behavior in question, e.g. of the underlying institutional arrangements. This type of spatial clustering, known as spatial error model, produces (spatial) heterogeneity in the error terms. Hence ignoring this form of spatial dependence has the same implications as the violation of the homoscedasticity assumption in regression models. The parameter estimates remain consistent, but owing to the spatial heteroscedasticity, the estimated standard errors are biased downwards and this increases the occurrence of Type 1 errors when these standard errors are used for statistical inference.

Alternatively spatial dependence may be produced by diffusion process, when spatial spillovers cause spatial correlation. As we have already discussed any such spatial spillovers lead to the spatial lag model. Having spatial lag model implies genuine spatial spillovers and has much more serious implications for estimation. These are essentially the same as omitting a significant explanatory variable. The resulting estimates are generally biased and inconsistent. Thus the consequences of ignoring spatial lag are much more serious than these resulting from ignoring spatial error. Furthermore, the sources of spatial lag dependence are much more 'interesting' in the sense that they can be nested in one of the underlying theoretical frameworks.

To complicate things further, the spatial lag and spatial error specifications can be difficult to distinguish, since the spatial error representation can be viewed as a restriction on the more general spatial lag one, something that is popularly referred to as the spatial Durbin model. This however provides us with the possibility to explicitly test within a given spatial lag representation whether then spatial error restriction holds or not and further enhance our understanding of the substantive sources of spatial dependence present in the data. Therefore methods dealing with the spatial lag specification could be useful even if the suspected forms of spatial dependence was this of spatial error since they can be employed as first step in a more general modeling strategy.

The specification of spatial dependence via a spatial weighting matrix is a convenient way to describe theoretical or a priori knowledge and understanding of the underlying structure generating the 'spatial' dependence between different economic agents and units of analysis. In simple words defining a spatial weighting matrix involves two choices, namely a neighbourhood scheme and spatial weights. The neighbourhood scheme involves determination of which units of analysis are linked and which are not. When units are economic agents this means the decisions of which agents are to be included in the

objective functions of other agents. A social network structure could for example be used to infer the neighbourhood scheme. The weighting scheme on the other hand defines the strength of these links. The weighting scheme is based on some distance metrics, which could be spatial, economic distance, or in the case of the social network example a social distance (e.g. family, close friends, acquaintances etc.). The weighting scheme takes the distance metrics and combines it in order to derive the strength of the impact each unit has on another unit.

For identification purposes the spatial weighting matrix needs to be exogenous (Manski, 1993). One reason for the popularity of spatial weighting matrices based on geographical distances is the fact that their exogeneity is automatically ensured.

In practice the spatial weighting matrix carries out a spatial smoothing over the dependent variable, thus incorporating part (given by the spatial weights) of the values at the neighbouring observations. For logical and identification purposes some structure is imposed on the spatial weighting matrix. The first assumption is to set its diagonal elements to zero. This reflects that one is not a 'neighbour' to itself in that spillovers from itself are not allowed. This assumption is facilitating interpretation of the results. Furthermore the spatial lag coefficient λ is usually assumed to be in the (-1,1) interval. This is needed to provide a comparative perspective and to interpret this coefficient as the strength of the spatial diffusion process. Such an interpretation would not however be possible if the spatial weighting matrix is not normalised. The weights need to be normalised because different spatial weighting matrices can define the same diffusion process up to a factor of proportionality, meaning that just by scaling up or down a spatial weighting matrix one can represent the same structure. A convenient normalisation is to produce a row standardised spatial weighting matrix. This amounts to setting the sum of each row to add up to 1. This yields a unique spatial weighting matrix for a given weighting scheme. Furthermore this standardisation ensures that the spatial lag coefficient λ can be viewed as strength of the diffusion process and should logically be restricted to the interval (-1, 1) to avoid an explosive type of spatial diffusion process. Finally the spatial filtering matrix (i.e. $I - \lambda W$, where I is a unity matrix with an appropriate dimension) is assumed non-singular for estimation purposes.

Very often spatial distances may reasonably well approximate the underlying 'true' metrics, which may be unobservable or unavailable. For example often spatial distance can approximate the strength of social relationships. Therefore in the absence of direct

measurement of the underlying relationship, the spatial distances could be used. Note however that in such an approximation process even if one knows the exact form of the linkages, as expressed in the underlying unavailable metrics, translation into spatial distances (or any other alternative metrics system) changes matters. The translation may effectively break down the theoretical spillover definition. Hence the uncertainty about what the spatial distances measure introduces additional uncertainty in the process of specifying an appropriate spatial weighting matrix.

Choosing the spatial weighting matrix

In some applications some of the choices underlying the spatial weighting matrix (i.e. neighbourhood definition and weighting scheme) may be logically predetermined, e.g. the nature of the problem may suggest the neighbourhood scheme and/or equal weights could be a logical choice. In most cases however this choice is far from trivial. The choice of spatial weighting matrix in empirical applications has been usually subject to some arbitrariness. This arbitrariness presents a serious problem to the inference in such models since estimation results have been shown to critically depend on the choice of spatial weighting matrix (Anselin, 2002; Fingleton, 2003).

Popular weighting schemes are inverse distances (raised to some power), lengths of shared borders (divided by the perimeter), n^{th} nearest neighbour distance, ranked distances, constrained weights for an observation equal to some (predetermined) constant, all observations within a given distance. And the search for appropriate specification does not seem to stop. Some proposals include the bandwidth distance decay (Fotheringham *et al*., 1996), Gaussian distance decline (LeSage, 2003); the tri-cube distance decline function (McMillen and McDonald, 2003); the 'local statistics model' (Getis and Aldstadt, 2001, 2002), the 'optimize bandwidth' approach (Fotheringham *et al.*, 2002) and the AMOEBA (Aldstadt and Getis, 2003). Other approaches try to relax the neighbourhood definition. These include the moving windows regression, geographically weighted regression (Brunsdon *et al*, 1996) and locally weighted regression (McMillen, 1996). The general idea of these approaches is to substitute a 'sliding neighbourhood' for the predefined neighbourhood boundaries.

The issue of spatial weighting matrix have been outstanding for considerable amount of time. Kooijman, 1976 proposed to choose the spatial weighting matrix by maximizing

Moran's coefficient. In a more general vein this has led to the practice of choosing spatial weighting matrix maximising alternative spatial dependence statistics. Research into reducing the degree of arbitrariness in spatial weighting matrix choice has been particularly active in recent years. One could classify this strand of research into two main types. First, new and more flexible ways to specify the neighbourhood and/or the weighting schemes have been proposed. The above mentioned approaches fall into this category. The second type of proposals deals with essentially selecting the spatial weighting matrix either implicitly or explicitly from a pre-defined set of candidates. Bhattacharjee and Jensen-Butler, 2005 proposed estimating spatial weighting matrix consistent with the data distribution, but their approach only applies to the spatial error model. Lima and Macedo, 1999 proposed an interesting procedure dealing with estimating the weights decay and thus the spatial weights matrix with a predefined 'soft' neighbourhood (soft in the sense that the weight decay can exclude some observations from the neighbourhood definition). When we have an explicit set of competing spatial weighting matrices, LeSage and Parent, 2007 proposed a Bayesian model averaging procedure for spatial model which incorporates the uncertainty about the correct spatial weighting matrix. Holloway and Lapar, 2007 used a Bayesian marginal likelihood approach to select a neighbourhood definition (cut-off points for the neighbourhood), but one can consider their approach as a general model selection approach, which could be applied to any other set of competing models. Finally Kelejian, 2008 proposed a formal statistical test to distinguish between non-nested spatial specifications.

Our proposal lies within the model selection approaches, i.e. selecting amongst a predefined set of models. In this case we are primarily interested in models with alternative spatial weighting matrices. A common drawback of the model selection approaches is that the competing models need to be estimated, either explicitly (e.g. in Holloway and Lapar, 2007), or implicitly as a part of the testing procedure (e.g. in Kelejian, 2008). Despite the huge advances in computing technology, computationally simpler approaches are still beneficial. In this paper we suggest using component-wise model boosting as a computationally simple model selection procedure to alleviate the arbitrariness of spatial weighting matrix choice. Although the approach suggested here can be used for general specification search (see e.g. Florax *et al.*, 2003, 2006 and Hendry, 2006) for simplicity here we will implicitly assume correct specification and will focus specifically on choosing the appropriate spatial weighting matrix.

Conceptual framework

The spatial lag specification includes the spatially lagged dependent variable Wy on the right hand side. This results in endogeneity with the dependent variable. In such a setting conventional estimators are inconsistent. There are two main types of estimators for the spatial lag model that deal with the endogeneity issue and have been extensively studied and used in the literature. These are the maximum likelihood or quasi maximum likelihood estimator (see e.g. Anselin, 1988) and the generalized method of moment estimator (see Kelejian and Prucha 1998, 1999). We propose using the spatial two-stage least squares approach of Kelejian and Prucha, 1998 (which can be viewed as a type of generalised method of moments estimator) to transform the spatial weighting matrix choice into a variable selection one. The spatial two-stage least squares amounts to using the spatially lagged independent variables as instruments for the spatially lagged dependent variable. Thus we can simply project the spatially lagged dependent variable in the vector space of the instruments and use the transformed in this way variable instead of the original one. This can be done by direct matrix manipulation or by running an auxiliary regression (of the spatially weighted dependent variable on the spatially weighted independent ones) and using the residuals from this regression in the second estimation step. In simple words this means that we can run separate auxiliary regressions for each potential spatial weighting matrix. These will provide us with the corresponding transformed variable to include in the 'second stage'. Thus the question of whether a given spatial weighting matrix needs to be included gets translated into the one which of the created transformed variables need to be included in the main regression model. This is a typical variable selection problem and there are many different methods to perform variable selection. Here we suggests using a component-wise boosting algorithm.

There are many different methods for variable selection in linear models. The best known approaches are forward selection and backward elimination. The combination of these two approaches is usually referred to as a stepwise regression (see e.g. Miller 2002). Alternative approaches for subset selection in linear models which are closely related to each other are LASSO (Least Absolute Sum of Squares Operator, see Tibshirani, 1996), forward stagewise regression and LARS (Least Angle Regression, see Efron et al., 2004)), boosting approaches (Bühlmann, 2006), the elastic net (Zou and Hastie, 2005) and the Dantzig selector (Candes and Tao, 2007). We will not discuss the Bayesian variable selection

methods here. Most non-Bayesian model selection methods are essentially based on penalised estimation criteria. Other penalised methods are the nonnegative garrote (Breiman, 1994), the bridge estimator (Frank and Friedman, 1993; Fu, 1998), SCAD (smoothly clipped absolute deviation, Fan and Li, 2001). A comprehensive overview of penalised methods is available in Fan and Li (2006).

With such a wide range of available methods, how does one choose the one appropriate to the problem in hand. In the case of choosing an appropriate spatial weighting matrix, there is large number of alternatives. Therefore we need a method that can handle well high-dimensional problems and is relatively fast in terms of computational time. We do not however strictly require the 'oracle' property in the sense of Fan and Li, 2001. The oracle property requires that the asymptotic distribution of the non-zero coefficients in the estimated model is the same as when the zero coefficients are known in advance. It is useful when the method is used for both model selection and estimation. Note however that since the underlying two –step estimation requires adjustments to the standard error estimates, it is impractical to use the variable selection method also for estimation. Therefore only consistency with regard to the variable selection is necessary.

In terms of computational burden, some of the variable selection methods are relatively more expensive than others. Step-wise regression is amongst the more demanding methods, particularly when the number of covariates is large. The computational burden for most penalized estimators arises from the nature of the used penalty term. For example the SCAD penalty (Fan and Li, 2001) involves non-convex optimization and thus can be computationally expensive. The LASSO estimator uses $L1$ (absolute deviations) penalty and can also be relatively demanding. A fast estimator is the LARS (Efron et al., 2004). The computational requirements of the LARS algorithm are similar to this of a least squares fitting. Furthermore, in addition to its speed, it provides an illuminating overview of the linkages amongst different variable selection algorithms. In particular LARS can be modified to yield either the LASSO solutions or that of a forward stagewise fitting (Efron et al.2004). Forward stage fitting on the other hand can be viewed as a simplified version of boosting with a small fixed step size (Hastie et al., 2001). Thus LARS, LASSO and boosting (*L2* boosting) are somewhat 'related'. This does not mean that they are equivalent. Their equivalence can only be established for orthogonal predictors and the difficult to verify case of monotone paths, but even in general they often produce similar solutions. Thus when one of these algorithms is impractical to implement the others could

be used instead. It could sometimes be prohibitively expensive to solve LASSO for a large number of candidate spatial weighting matrices with general loss functions for many regularisation parameters via quadratic programming. The LARS algorithm is very efficient computationally for least squares problems when the number of predictors is small. It does not however deal with other loss functions and is not adequate with a large number of predictors. Boosting on the other hand can use different loss functions and works well with large number of predictors. Furthermore even for smaller number of predictors component-wise boosting is about 3 times faster than LASSO. A major advantage of the component-wise boosting algorithms is that it can fit models with negative degrees of freedom, i.e. when the number of predictors exceeds the number of observations. Since choosing a spatial weighting matrix can involve too many alternatives, such a property is highly desirable. Bühlmann and Yu, 2003 provide an empirical illustration of the advantages of boosting for models with high-dimensional predictors. In more classical settings with smaller number of predictors alternative method performs similarly.

The R statistical system (R Development Core Team, 2008) contains extensive selection of ready to use regularisation methods code, contained in different packages. The lasso2 and lars packages implement LASSO and LARS estimators, package grplasso provides groupwise LASSO (simultaneous updates for predefined groups of parameters. Other useful methods exist in the packages glmpath, elasticnet, glmnet, penalized and relaxo. There are various implementations of boosting algorithms in R, contained in packages such as gbm, boost and GAMBoost. The methods discussed in this paper were implemented using the mboost package (Hothorn and Bühlmann, 2008) which provides extendable framework for a wide range of models. All code underlying the R system and all official (i.e. available from the Comprehensive R Archive Network (CRAN) sites) packages is publicly available and could be modified with no restrictions. This allows one to combine ease of implementation and flexibility.

Methodology

Boosting itself is a vast area. There is a growing number of different boosting algorithms and approaches and it would be far beyond the scope of this paper to review them. To further complicate matters originally boosting was conceived from a machine learning perspective as a combination of ‘weak’ learners. Here we will present the alternative statistical perspective on boosting. We will present a generic overview to the boosting

algorithm, demonstrating its generality. Where applicable the specifics of our implementation would be described.

We will consider regression model where the response *y* is an additive function of the predictors. Thus we can denote

$$y = \eta(x) + \xi = \beta_0 + \sum_{i=1}^{k} f_i(x) + \xi \qquad (2)$$

Recently Bühlmann, 2006 suggested component-wise boosting to specifically deal with the issues of variable and model selection. We will briefly introduce the idea of the boosting algorithm. Then the component-wise version of boosting will be discussed in relation to the components used in this application.

From statistical perspective boosting can be viewed as a functional gradient descent method that minimises the constrained empirical risk function

$$\sum_{i=1}^{n} w_i \rho(y_i, \eta(x_i))$$

where w_i are some weights, and $\rho(.)$ is some suitable (in practice this means convex and differentiable) loss function. To simplify the discussion, from now on we will implicitly assume equal weights. Typical examples for loss function would be the log-likelihood function or the *L*2 norm (sum of squared residuals). Note that classical estimators essentially solve the same optimisation problem. The main difference is that they apply a specific algorithm, that is typically applicable only to a given class of models specified by the underlying functions $f(x)$. Therefore we may think of the boosting approach as providing a general approach to model estimation. The general idea of the boosting algorithm is to minimise the empirical risk with regard to η.

To explain the boosting algorithm, let us assume a given type of underlying function (base learner) *f*. In this particular case we will only consider linear base learners, i.e. $f(x) = f_{linear}(x) = x\beta$, but the approach is generalisable to a wider range of alternative functions (see Kneib *et al*., 2009 for more details).

Lets us further simplify matter and assume the *L*2 norm for the empirical risk function $\rho(y, \eta) = (y - \eta)^2$.

[Kostov, 2010]. This is a postprint of a research article. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning B: Planning and Design, volume 37, issue 3, pages* 533–549, *2010*, http://dx.doi.org/10.1068/b35137

The boosting algorithm is initialised by an initial value for η, e.g. η_0. This implies an initial evaluation for the underlying function $\hat{f}_0$. Typically we start with an offset set to the unconditional mean of the response variable.

Then it iteratively goes through the following steps:

1. Compute the negative gradient of the empirical risk function evaluated at the current function estimate (η_m for every step from m=1, …)

$$u_i = -\left.\frac{\partial \rho(y_i, \eta)}{\partial \eta}\right|_{\eta=\hat{\eta}_{m-1}(x_i)} \quad \text{for } i = 1,2,\ldots,n.$$

2. Use the above calculated negative gradients (sometimes called 'residuals', because with $L2$ norm empirical risk and linear model they do coincide with the current regression residuals) to fit the underlying function $\hat{g}_m(.)$ Here $\hat{g}_m(.)$ is the fitted to the current residuals value of the used function at iteration m.

3. Update $\hat{f}_m = \hat{f}_{m-1} + \nu \hat{g}_m(.)$ for a given step size ν.

The algorithm iterates between steps 1-3 until a maximum number of iterations is reached.

As the generic description of the algorithm demonstrates the boosting algorithm constructs iteratively η (i.e. all functions $f_i(x)$) by pursuing iterative approximate steepest descent in function space, calculated using the adopted empirical risk function.

It is a simple algorithm. With an $L2$ empirical risk function it essentially does an iterative least squares fitting of the residuals for a linear models. The approach is also flexible, because it can be applied to a wide range of alternative loss functions. This could be of concern when 'robust' versions are required (see Lutz *et al*., 2008).

Here we will consider the component-wise version of the algorithm which can be used for variable and model selection purposes. In contrast to the general boosting algorithm, it fits a single component at each iteration. This is achieved by the following slight modification of the general boosting algorithm. In step 2 we simply chose the best fitting component-wise learner

$j^* = \arg\min_{1 \le j \le k} \sum_{i=1}^{n} \left(u_i - g_j(x_i) \right)$ which leads to this particular base learner being the only one updated in step 3, i.e. $\hat{f}_{j^*m} = \hat{f}_{j^*,m-1} + \nu \hat{g}_{j^*,m}(.)$, while $\hat{f}_{jm} = \hat{f}_{j,m-1}$ for $\forall\ j \neq j^*$, where the first subscript denotes the base learner and the second one is the iteration counter.

In simple terms we fit base learners (typically consisting of one covariate). In the component-wise boosting only one of the different learners is selected for updating at each step. If functional forms are given as in this case, selecting a base learner corresponds to selecting a covariate. In this case the selected covariate is the one which gives the smallest residual sum of squares, i.e. the variable that gives the largest contribution to the fit. After the algorithm has run for the maximum number of iterations, some of the base learners may have never been selected for updating, which means their final evaluations are zero. In this way the algorithm may be used for variable selection. Bühlmann, 2006 provides detailed discussion of *L2* boosting and component-wise linear fitting.

For practical implementation, we need to select an updating factor ν (also referred to as step-length factor or shrinkage factor). A value of 1 seems like a natural choice but following Friedman, 2001 most applications use smaller values. Friedman, 2001 showed empirically that that small values of ν perform better and that the boosting procedure is not very sensitive to a whole range of small values of ν. Here we will use ν=0.2, which is within the 'standard' range of values between 0.1 and 0.5 typically used in boosting applications.

Finally we have to choose an optimal number of iterations for stopping the algorithm. This can be estimated via cross-validation (see e.g. Bühlmann and Hothorn, 2007a).

Note however that cross-validation can be time consuming, particularly when we are doing variable search in high dimensions. In such cases a more standard models selection criterion such as the Akaike Information Criterion (AIC) could be used instead. For linear models the alternative 'corrected' AIC (Hurvich et al., 1998) could be implemented. Bühlmann and Yu (2006) have shown that a data driven compromise between AIC and BIC, namely the g-prior minimum description length (gMDL) introduced in Hansen and Yu (2001) can be successful in boosting for variable selection problems.

Since boosting essentially provides a unified approach to estimating a wide range of statistical models and given the initial transformation of the spatially weighted dependent variables, this approach provides us with a boosting equivalent to the spatial two-stage least

squares estimator. Thus we can use the component-wise boosting algorithm to 'estimate' a model with several competing weighting matrices. The main advantage of the proposed procedure is that it is computationally simple. Since at every iteration the boosting algorithm essentially does univariate least squares fitting, the computational cost is very low. The algorithm is not guaranteed to choose a single spatial weighting matrix amongst the available alternatives. It can nevertheless considerably reduce the universe of potential candidates, so that some of the other model selection methods, as described above, could be used in a subsequent analysis, if a single alternative is desired. Alternatively if several spatial weighting matrices are used, the selected combination of such matrices may be used to characterise a more complex spatial diffusion process.

Data

To illustrate the proposed methodology we will use the well known Boston housing dataset. It is one of the first medium size housing datasets. The corrected version we will use can be obtained at: http://lib.stat.cmu.edu/datasets/boston_corrected.txt. It consists of 506 observations. The original Boston housing data is due to the Harrison and Rubinfeld 1978. Gilley and Pace 1996 discuss the corrections for some minor errors and augmented the data with the latitude and longitude of the observations. The spatial information has been shown to improve estimates (Pace and Gilley 1997). It is a very popular dataset routinely employed in data mining and machine learning. One could say that this is one of the most popular datasets that have stimulated a whole 'industry' emerging over the years that have used this and some other datasets to examine and compare alternative statistical methods.

Briefly this dataset contains one observation for each census tract in the Boston Standard Metropolitan Statistical Area. The variables comprise of proxies for pollution, crime, distance to employment centres, geographical features, accessibility, housing size, age, race, status, tax burden, educational quality, zoning, and industrial externalities. A detailed description of the variables, to be used in this study is presented in table 1.

[Kostov, 2010]. This is a postprint of a research article. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning B: Planning and Design, volume 37, issue 3, pages* 533–549, *2010,* http://dx.doi.org/10.1068/b35137

Table 1 Variable description

Variable	Description
MEDV	Median values of owner-occupied housing in thousands of USD
LON	Tract point longitude in decimal degrees
LAT	Tract point latitude in decimal degrees
CRIM	Per capita crime
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft per town
INDUS	Proportion of non-retail business acres per town
CHAS	An indicator: 1 if tract borders Charles River; 0 otherwise
NOX	Nitric oxides concentration (parts per 10 million) per town
RM	Average number of rooms per dwelling
AGE	Proportions of owner-occupied units built prior to 1940
DIS	Weighted distance to five Boston employment centres
RAD	Index of accessibility to radial highways per town
TAX	Property-tax rate per USD 10,000 per town
PTRATIO	Pupil-teacher ratio per town
B	Calculated as 1000*(Bk - 0.63)^2 where Bk is the proportion of blacks
LSTAT	Percentage of lower status population

The basic model we will implement is as follows:

log(MEDV)= *f* {CRIM, ZN, INDUS, CHAS, NOX^2, RM^2, AGE, log(DIS), log(RAD), TAX, PTRATIO, B, log(LSTAT)}

We will consider linear functional form and will augment it with alternative candidate spatial weighting matrices, constructed using the longitude and latitude information. The main reason for applying logarithms and squares is to capture some of the underlying nonlinearities.

[Kostov, 2010]. This is a postprint of a research article. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning B: Planning and Design, volume 37, issue 3, pages* 533–549, *2010*, http://dx.doi.org/10.1068/b35137

Some descriptive statistics for the transformed) variables used in the model are presented in table 2

Table 2. Descriptive statistics

	Mean	Standard deviation	Minimum	Maximum
Log(MEDV)	3.035	0.409	1.609	3.912
CRIM	3.614	8.602	0.006	88.976
ZN	11.364	23.322	0.000	100.000
INDUS	11.137	6.860	0.460	27.740
CHAS	1.069	0.254	1.000	2.000
NOX^2	0.321	0.139	0.148	0.759
RM^2	39.989	9.080	12.681	77.088
AGE	68.575	28.149	2.900	100.000
log(DIS)	1.188	0.540	0.122	2.495
log(RAD)	1.868	0.875	0.000	3.178
TAX	408.237	168.537	187.000	711.000
PTRATIO	18.456	2.165	12.600	22.000
B	356.674	91.295	0.320	396.900
log(LSTAT)	2.371	0.601	0.548	3.637

Study design and results

For the problem in hand, spatial spillovers could ensue from neighbouring sales. There are several natural candidates for this how to construct potential spatial weighting matrices for this problem. First, the n^{th} nearest observations and all observations within a predefined distance are logical candidates for realistic representation of potential spillovers. Additionally if one obtains the actual boundaries of the tracts to which the observations pertain, spatial weighting matrices based on contiginuity and the length of the common border look like a reasonable choice, but we do not have this information here. For illustrative purposes here we only apply the n^{th} nearest neighbours criteria. Employing only one of the above two criteria leads to nested in each other neighbourhood specifications. Additionally we need to specify the weighting scheme. A popular choice for weighting scheme is the one based on inverse distance raised to some power. Hereafter we will call

the latter a (spatial) weighting parameter. For example inverse squared distances correspond to weighting parameter of 2, while inverse distanced to a parameter of 1.

Note however that the universe of potential alternative spatial weighting matrices is very large. Assuming the above definitions for neighbourhood and weighting scheme, we still have a very large number of alternatives. The proposed model boosting approach can be very useful in dealing with such large number of alternatives. In particular in this particular case the possible values of the number of neighbours range from 1 to 505. Here however it does not look reasonable to expect a very large number of 'neighbours' and we restrict the considered neighbourhood definition to maximum of 50 neighbours. The value of the weighting scheme parameter *w* (which is the inverse power of the weight decay) can be evaluated on a regular grid over a suitably defined interval. With a more detailed grid this will result in a large number of values.

Using each of these alternative spatial weighting matrices we create the corresponding variables for inclusion in the model selection procedure. From now on we will use a name combining the codes for the neighbourhood definition and the weighting scheme to refer to the corresponding spatial weighting matrix and the resulting additional variable to be included in the boosting model. All these variables are named using the following convention: *nxwy*, where *x* is the number of neighbours and *y* is the weighting parameter. For example the spatial weighting matrix with the nearest 50 observations as neighbours and inverse squared distance weights as well as the resulting transformed variable will be denoted as *n*50*w*2.

We employ all values for number of neighbours from 1 to 50 and evaluate *w* in the interval [0.4, 4] using increments of 0.1. In simple words this means we are combining 50 possible neighbourhood definitions with 37 alternatives for the weighting parameter resulting in 1,850 alternative spatial weighting matrices to be considered simultaneously. Most alternative methods will struggle with this task, since in this design we already have negative degrees of freedom (with 506 observations) so the most penalty based methods will not be applicable. Component-wise boosting however fits a single component at a time and thus could estimate the model even if it had negative degrees of freedom. In this case nevertheless we expect that most of the alternative spatial weighting matrices will never be selected.

In simple words we create the spatially weighed dependent and independent variables for each of the alternative spatial weighting matrices and by projecting the spatially weighted dependent variable into the column vector space of the spatially weighted independent variables, which could be done either by direct matrix manipulation or by taking the fitted values from a least-squares regression. This results in the transformed variables. Our model is then augmented by these transformed variables and we use component-wise boosting to estimate it. The boosting procedure will only select these transformed variables which contribute to the model fit. If some of the additional variables is not selected, this means that the corresponding spatial weighting matrix is inappropriate for the model and thus has to be rejected. We run the boosting algorithm for 5,000 iterations and employ seven different stopping rule criteria, namely the Akaike Information Criterion (AIC), the corrected Akaike Information Criterion(cAIC), the g-prior Minimum Description Length (gMDL) criterion, 10-fold cross validation (10fCV), 8-fold cross-validation (8fCV), cross-validation with a 25 bootstrap replications used to select the folds (25bCV) and the latter with 100 bootsrap replications (100bCV).

We ran the boosting algorithm with six different values of the updating step size ν (as defined in step 3 of the algorithm). Table 3 presents details on the number of iterations required to stop the algorithm, according to the different stopping criteria. We ran the boosting algorithm for 5,000 iterations and calculated the stopping criteria, except for a step size of 0.1 where 10,000 iterations were used. Where the stopping criteria chose the last iteration, the required number of iterations is probably larger than 5,000. In such cases table 4 shows the number of spatial weighting matrices present in the model at the end of the last iteration. In these cases larger number of iterations would be needed to properly assess the required stopping criteria and the associated with it selected variables. Since with exception to the cAIC, this only occurred once, the 5,000 iterations run was considered to be sufficient for larger values of the updating parameter ν.

Details on the typical computational time involved in each calculation will be presented later. One can clearly see that there is a trade-off between the step size and the number of stopping iterations. Larger step size requires less iterations, which in some cases could speed up the process. Overall however the number of selected variables is relatively insensitive to the choice of the step size. In principle larger step sizes introduce some sparseness and thus should in general lead to less variables being selected. In practice

however, due to the similarity of the alternative spatial weighting matrices and the linear specification, this effect is insignificant.

Table 3. Details on selection of spatial weighting matrices

Criteria	Step size, ν	Stopping iterations	Number of selected spatial weighting matrices	Step size, ν	Stopping iterations	Number of selected spatial weighting matrices
cAIC	0.1	>10000	11	0.4	4997	19
gMDL		6212	9		1641	11
10fCV		7089	10		1469	11
8fCV		5177	9		1119	9
25bCV		9999	11		3989	17
100bCV		9999	11		3556	16
cAIC	0.2	4999	11	0.5	>5000	19
gMDL		3029	9		996	9
10fCV		3358	10		1042	10
8fCV		2303	9		766	9
25bCV		>5000	11		2756	16
100bCV		>5000	11		2803	16
cAIC	0.3	>5000	16	0.6	>5000	22
gMDL		1923	9		825	10
10fCV		2172	10		800	10
8fCV		1486	9		678	9
25bCV		4952	16		2421	18
100bCV		4955	16		1992	15

The cAIC seems to overfit the model, selecting larger number of stopping iterations and thus over-specified models. The classical version of the AIC produces similar results. The 10-fold 8-fold cross-validation (10fCV and 8fCV) perform satisfactory, with the 8-fold cross-validation selecting slightly more parsimonious models. Interestingly cross-validation, based on 25 bootstrap samples (25bCV) does seem to select too many variables for lower values of the step size. Increasing the number of bootstrap replications does reduce the number of selected variables (see the 100 bootstrap samples (100bCV) results above), but this reduction is rather slow and comes at significant additional computational cost. The gMDL criterion performs extremely well. Its results are very similar to those obtained via 10 and 8-fold cross-validation, but at a fraction of the computational cost.

Given its low computational cost (comparable with e.g. calculating the AIC) it would be preferable to employ it in selecting the number of stopping iterations. Table 4 presents the computational time for the boosting estimation and the computation of some early stopping criteria, undertaken on Intel Core 2 PC with 2.13GHz clock speed. The early stopping criteria are calculated over the 5000 iterations of the boosting algorithm run on the dataset that includes the 1850 alternative spatial weighting matrices.

Table 4. Computational costs

Computation	Time (seconds)
5000 iterations of the boosting algorithm	24.89
Calculation of AIC	6.86
Calculation of corrected AIC	6.53
Calculation of gMDL	6.55
Calculation of 10 fold cross validation	246.75
Calculation of 8 fold cross validation	197.13
Calculation of cross-validation based on 25 bootstrap samples	617.20
Calculation of cross-validation based on 100 bootstrap samples	2463.41

As it is to be expected the cross-validation procedures have relatively high computational costs, particularly for the bootstrap-based versions. The information criteria calculation is very fast. The boosting algorithm itself is very fast, mainly due to its compiled code implementation. Hence using the gMDL criterion for early stopping of the algorithm is however advantageous in combining good results (comparable with multifold cross-validation) with very low computational cost. Therefore it is advisable to use the gMDL where applicable. If the selected model was characterised by negative degrees of freedom, then information criteria could not have been calculated and multi-fold cross-validation would have been the only reasonable alternative. Hence the proposed methodology would be applicable even in such (however rare) cases, although at higher computational costs compared to using the GMDL criterion.

The selected spatial weighting matrices do not depend on the step size and are consistent amongst the comparable criteria, i.e. the same 9 or 10 spatial weighting matrices are selected across step sizes and different main criteria (i.e gMDL, 10fCV and 8fCV). Hence we will only examine the gMDL results. The spatial weighting matrices selected by the

application of the gMDL criterion are as follows: n3w1.1, n3w1.2, n6w0.4, n6w0.5, n6w0.6, n6w0.7, n6w0.8, n6w0.9, and n6w1. Except n6w0.4, none of these variables lies on the boundary of the used parameter space (thus no other spatial weighting matrices with 1 or 50 neighbours or alternatively with weight parameter equal to 0.4 or 4 are selected). This reduces the probability of misspecification due to inappropriate choice of grid over which the considered in this application spatial weighting matrices are constructed. The variable n6w0.4 features, but it is in a block of variables representing spatial weighting matrices with 6 neighbours and spatial weight parameter ranging from 0.4 to 1. Careful examination of the order of updating and magnitude of the corresponding coefficients suggests that if we needed a single spatial weighting matrix to approximate the underlying process n6w0.7 would be a reasonable choice. What the result state, is that the spatial spilovers are only defined over a small neighbourhood. There is however some uncertainty about the weighting scheme, which suggests that an alternative, probably more complicated weighting scheme could be appropriate. Given the nature of the problem it looks like the results are supportive to a spatial weighting scheme based on the length of the common boundary for the tracts. Therefore even when the 'true' spatial weighting matrix is not present amongst the alternatives supplied to the boosting algorithm, the obtained results could be indicative of what other alternatives could be worth exploring.

The boosting algorithm does not provide confidence intervals for the estimates. In some cases such as likelihood based boosting these could be straightforward to estimate, but fo most forms of boosting, it is relatively difficult task. Consequently we estimate the chosen model by the spatial two-stage least squares method of Kelejian and Prucha, 1998. The results from this estimation are presented in table 5. The boosting results suggest that if we needed a single spatial weighting matrix, one based on *n*6*w*0.7 would provide reasonable approximation. In table 5 below we present the results from estimating this model.

[Kostov, 2010]. This is a postprint of a research article. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning B: Planning and Design, volume 37, issue 3, pages* 533–549, *2010,* http://dx.doi.org/10.1068/b35137

Table 5. Final model estimation results

	Estimate	P level
Intercept	2.390	0.000
Spatial lag	0.461	0.000
CRIM	-0.008	0.000
ZN	0.000	0.307
INDUS	0.001	0.728
CHAS	0.019	0.483
NOX^2	-0.299	0.002
RM^2	0.007	0.000
AGE	0.000	0.544
log(DIS)	-0.166	0.000
log(RAD)	0.075	0.000
TAX	0.000	0.000
PTRATIO	-0.011	0.010
B	0.000	0.001
log(LSTAT)	-0.253	0.000

There is strong and significant spatial dependence. Interestingly if we use some of the other spatial weighting matrices (e.g. n5w0.4) the results for the spatial lag and the other coefficients do not change significantly.

There are also some scaling issues. The significant coefficients for TAX and B are correspondingly -0.00034763 and 0.00028863, but appear as zeros in table 5 above due to rounding. Otherwise the results are as expected. In particular crime, pollution (NOX^2), distance to employment centres, less teachers availability (i.e. higher PTRATIO), higher tax and greater low status population all decrease the housing value.

Conclusions and possible extensions

This paper proposes using component-wise model boosting for selection of spatial weighting matrix in the context of the spatial lag econometric model. It is a computationally simple procedure that avoids estimating the models implied by the alternative weighting

schemes. Therefore we can significantly reduce the arbitrariness of the spatial weighting matrix choice that is often cited as one of the main disadvantages of the lattice approach in spatial statistics. We present an illustrative application of the proposed methodology to a well known dataset of house prices. In this we demonstrate

The proposed approach is a general variable and model selection approach that can have much wider application that the one we presented here. The main attractiveness of the usage of model boosting for spatial weights matrix in the linear spatial lag econometric model is its low computational cost. Numerous extensions of the proposed approach are possible, but these generally involve some additional computational costs.

In this application we find that the possible spatial spillovers are restricted to a relatively small neighbourhood, i.e. only sales in the closest tracts influence the price of housing.

An interesting extension would be to employ alternative empirical risk functions, e.g. least absolute deviations or the Huber function. This could help produce "robust" model selection approach, although as discussed above it will involve some additional computational cost. Another possibility would be to combine in the same model the spatially lagged dependent variable (i.e. the spatial lag model) with spatial effects modelled as in the geo-additive modelling approach and thus implicitly test one against the other. Furthermore the proposed methodology is readily applicable to more general models with non-parametric effects, in which case we can relax the linearity assumption. The latter case would of course be much more computationally demanding that its present application, but it is a viable alternative to other non-parametric model selection strategies.

References:

Anselin L, 1988, *Spatial Econometrics: Methods and Models*, Kluwer Academic Press.

Anselin L, 2002, "Under the hood: Issues in the specification and interpretation of spatial regression models" *Agricultural Economics* **27** 247-267.

Aoki M, 1996 *New Approaches to Macroeconomic Modelling* (Cambridge University Press, Cambridge).

Bhattacharjee A, Jensen-Butler C, 2005, "Estimation of Spatial Weights Matrix in a Spatial Error Model, with an Application to Diffusion in Housing Demand", University of St.

Andrews, Centre for Research into Industry, Enterprise, Finance and the Firm (CRIEFF), Discussion Paper 0519.

Borjas G J, 1995, "Ethnicity, neighborhoods, and human-capital externalities" *American Economic Review* **85** 365–390.

Breiman L, 1995, "Better subset regression using the nonnegative garrotte" *Technometrics* **37** 373-384.

Brueckner J K, 2003, "Strategic interaction among governments: An overview of empirical studies" *International Regional Science Review*, **26 (2)** 175-188.

Brunsdon C, Fotheringham A S, Charlton M E, 1996, "Geographically weighted regression: a method for exploring spatial nonstationarity" *Geographical Analysis* **28** 281-298

Bühlmann P, 2006, "Boosting for high-dimensional linear models" *The Annals of Statistics* **34** 559–583.

Bühlmann P, Hothorn T, 2007a, "Boosting algorithms: Regularization, prediction and model fitting", *Statistical Science*, 22(4):477-505.

Bühlmann P, Hothorn T, 2007b, "Rejoinder: Boosting algorithms: Regularization, prediction and model fitting" *Statistical Science* **22(4)** 516-522.

Bühlmann P, Yu B, 2003, "Boosting with L2 loss: Regression and classification" *Journal of the American Statistical Association* **98** 324– 338.

Bühlmann P, Yu B, 2006, "Sparse Boosting" *Journal of Machine Learning Research* **7** 1001–1024.

Candes E, Tao T, 2007, "The Dantzig selector: Statistical estimation when *p* is much larger than *n" Annals of Statistics* **35(6)** 2313-2351.

Durlauf S N, 1994, "Spillovers, stratification and inequality" *European Economic Review* **38** 836–845.

Durlauf S N, 1997, "Statistical mechanics approaches to socioeconomic behaviour", in *The Economy as an Evolving Complex System II*, Eds B W Arthur, S N Durlauf and D A Lane (Addison-Wesley, Reading, MA) pp. 81–104..

Efron B, Hastie T, Johnstone I, Tibshirani R., 2004, "Least angle regression" *Annals of Statistics* **32** 407–451.

Fan J, Li R, 2001, "Variable selection via nonconcave penalized likelihood and its oracle properties" *Journal of the American Statistical Association* **96** 1348–1360.

Fan J, Li R, 2006, "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery", in M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, (eds.) *Proceedings of the International Congress of Mathematicians*, Vol. III, 595-622.

Fingleton B, 2003, "Externalities, economic geography and spatial econometrics: conceptual and modeling developments" *International Regional Science Review* **26** 197-207.

Florax R J G M, Folmer H, Rey S J, 2003, "Specification searches in spatial econometrics: the relevance of Hendry's methodology" *Regional Science and Urban Economics* **33** 557–579.

Florax R J G M, Folmer H, Rey S J, 2006, "A comment on specification searches in spatial econometrics: the relevance of Hendry's methodology: a reply" *Regional Science and Urban Economics* **36** 300–308.

Fotheringham A S, Brunsdon C, Charlton M, 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (John Wiley, Chichester)

Fotheringham A S, Charlton M E, Brunsdon C, 1996, "The Geography of Parameter Space: An Investigation into Spatial Nonstationarity" *International Journal of GIS* **10** 605–27.

Frank I E, Friedman J H, 1993, "A statistical view of some chemometrics regression tools (with discussion)" *Technometrics* **35** 109–148.

Fu W J, 1998, "Penalized regressions: The bridge versus the Lasso" *Journal of Computational and Graphical. Statististics* **7** 397–416.

Fujita M, Krugman P, Venables A, 1999 *The Spatial Economy: Cities, Regions and International Trade* (MIT Press, Cambridge, MA)

Gilley O W, Pace R K, 1996, "On the Harrison and Rubinfeld Data" *Journal of Environmental Economics and Management* **31** 403-405.

Glaeser E L, Sacerdote B, Scheinkman J, 1996 "Crime and social interactions" *Quaterly Journal of Economics* **111** 507–548.

Hansen M, Yu B, 2001, Model selection and minimum description length principle" *Journal of the American Statistical Association* **96** 746–774.

Harrison D, Rubinfeld D L, 1978, "Hedonic Housing Prices and the Demand for Clean Air" *Journal of Environmental Economics and Management* **5** 81-102.

Hastie T, Tibshirani R, Friedman J, 2001, *The Elements of Statistical Learning: Data mining, Inference and Prediction*, Springer Verlag, New York.

Hendry D F, 2006, "A comment on specification searches in spatial econometrics: the relevance of Hendry's methodology" *Regional Science and Urban Economics* **36** 309–312.

Holloway G, Lapar M L A, 2007, "How Big is Your Neighbourhood? Spatial Implications of Market Participation Among Filipino Smallholders" *Journal of Agricultural Economics* **58(1)** 37-60.

Hothorn T, Buhlmann P, 2006, "mboost: Gradient Boosting for Fitting Generalized Linear, Additive and Interaction Models. R package version 0.4-8", available at: http://CRAN.R-project.org

Hothorn T, Bühlmann P, 2006, "Model-based boosting in high dimensions" *Bioinformatics* **22(2)** 2828-2829.

Hurvich C, Simonoff J, Tsai C-L, 1998, "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion" *Journal of the Royal Statististical Society Series B* **60(2)** 271–293.

Kelejian H, 2008, "A spatial J-test for model specification against a single or a set of non-nested alternatives" *Letters in Spatial and Resource Sciences* **1(1)** 3-11.

Kelejian H, Prucha I R, 1998, "A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances" *Journal of Real Estate Finance and Economics* **17** 99-121.

Kelejian H, Prucha I R, 1999, "A generalized moments estimator for the autoregressive parameter in a spatial model" *International Economic Review* **40** 509-533.

Kneib T, Hothorn T, Tutz G, 2009 forthcoming, "Variable selection and model choice in geoadditive regression models" *Biometrics*.

Kooijman S A L M, 1976, "Some Remarks on the Statistical Analysis of Grids Especially with Respect to Ecology" *Annals of Systems Research* **5** 113-132.

LeSage J P, 2003, "A Family of Geographically Weighted Regression Models" in *Advances in Spatial Econometrics: Methodology, Tools and Applications* Eds. L Anselin, R J G M Florax and S J Rey (Heidelberg: Springer).

LeSage J P, Parent O, 2007, "Bayesian Model Averaging for Spatial Econometric Models" *Geographical Analysis* **39** 241–267.

Lima E C R, Macedo P B R, 1999, "Estimation of a weights matrix for determining spatial effects", IPEA Discussion paper 672, Rio de Janeiro.

Lutz R W, Kalisch M, Bühlmann P, 2008, "Robustified L2 boosting" *Computational Statistics & Data Analysis* **52** 3331-3341.

Manski C F, 1993, "Identification of endogenous social effects: The reflexion problem" *Review of Economic Studies* **60** 531–542.

McMillen D P, McDonald J F, 2003, "Locally Weighted Maximum-Likelihood Estimation: Monte Carlo Evidence and an Application" in *Advances in Spatial Econometrics: Methodology, Tools and Applications* Eds L Anselin, R J G M Florax and S J Rey (Heidelberg, Springer).

Miller A., 2002, *Subset Selection in Regression*, Chapman and Hall, Boca Raton.

Pace R K, Gilley O W, 1997, "Using the Spatial Configuration of the Data to Improve Estimation" *Journal of the Real Estate Finance and Economics* **14** 333-340.

R Development Core Team 2007, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Tibshirani R, 1996, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society Series B* **58** 267-288.

Tutz G, Binder H, 2006, "Generalized additive modelling with implicit variable selection by likelihood based boosting" *Biometrics* **62** 961–971.

Zou H, Hastie T, 2005, "Regularization and variable selection via the elastic net" *Journal of the Royal Statistical Society: Series B* **67** 301-320.