Neurocomputing 124 (2014) 81–88

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semiparametric spatial effects kernel minimum squared error model for predicting housing sales prices

Jooyong Shim [a], Okmyung Bin [b], Changha Hwang [c,*]

[a] Institute of Statistical Information, Department of Data Science, Inje University, Kyungnam 621-749, South Korea
[b] Department of Economics, East Carolina University, Greenville, NC 27858-4353, USA
[c] Department of Statistics, Dankook University, Gyeonggido 448-160, South Korea

ARTICLE INFO

Article history:
Received 27 January 2013
Received in revised form
30 May 2013
Accepted 25 July 2013
Communicated by T. Heskes
Available online 17 August 2013

Keywords:
Housing sale price
Kernel minimum squared error
Least squares support vector machine
Prediction
Semiparametric
Spatial effect

ABSTRACT

Semiparametric regression models have been extensively used to predict housing sales prices, but semiparametric kernel machines with spatial effect have not been studied yet. This paper proposes the semiparametric spatial effect kernel minimum squared error model (SSEKMSEM) and the semiparametric spatial effect least squares support vector machine (SSELS-SVM) for estimating a hedonic price function and compares the price prediction performance with the conventional parametric models and a semiparametric generalized additive model (GAM). This paper utilizes two data sets. One is a large data set representing 5966 single-family residential home sales between July 2000 and August 2008 from Pitt County, North Carolina. The other is a data set of residential property sales records from September 2000 to September 2004 in Carteret County, North Carolina. The results show that the SSEKMSEM and SSELS-SVM outperform the parametric counterparts and the semiparametric GAM in both in-sample and out-of-sample price predictions, indicating that these kernel machines can be useful for measurement and prediction of housing sales prices.

1. Introduction

The hedonic method has been frequently used to estimate the functional relationship between housing prices and the attributes such as the age of the house, the number of rooms, and the neighborhood and environmental characteristics [1,14,31,24]. Regression analysis of the hedonic price models allows the researcher to construct a house price index and to predict the sales price given a set of housing attributes. The housing price indices are used for multiple purposes, such as analyses of real estate portfolios or mortgage-lending decisions by major financial institutions. The estimates generated by the hedonic models are used as a basis for property taxes by cities and counties. Inaccurate appraisal of house values may result in substantial property tax adjustments. The hedonic price function is often estimated by the ordinary least squares regression.

A frequent concern in this literature is the adequacy of commonly assumed parametric specifications as hedonic price functions. This specification problem arises quite naturally from the inability of theory on the proper functional form between house price and its attributes [8,11]. Recognizing the potentially serious consequences of functional misspecification, researchers have attempted to estimate hedonic price models by specifying more flexible regression models. Most of these attempts have concentrated on parametric specifications ranging from simple data transformations, including the model introduced by [3] and its variants and approximations based on second order Taylor-expansions to flexible nonlinear models such as those introduced by [33].

While the majority of the literature has concentrated on the parametric hedonic price models, several recent studies have applied nonparametric and semiparametric models which can improve the drawbacks of parametric models. Nonparametric regression models are very flexible in that regressions are allowed to belong to a vastly broader class of functions than that in parametric models. However, a fully unrestricted nonparametric regression is undesirable in a multivariate setting because of slow convergence rates of the estimators and diminished confidence on inference, also known as the curse of dimensionality identified by [10]. Semiparametric models combine the benefits of the parametric and nonparametric estimations, permitting the function to be linear, convex, or concave of curvilinear and seeking for the best fitted model. Ref. [18] applied the kernel nonparametric regression technique to hedonic price models. Ref. [15] reported that nonparametric techniques produce more robust results of the house price model. Ref. [19] illustrated GAMs can outperform naive parametric and polynomial models in out-of-sample predictive behavior. Ref. [5] combined local polynomial regression with a kriging model to better measure spatial variation

* Corresponding author.
E-mail addresses: ds1631@hanmail.net (J. Shim), BINO@ecu.edu (O. Bin), chwang@dankook.ac.kr (C. Hwang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.07.035

in house prices. Ref. [1] extended the approach of [12] using the additive semiparametric models with GIS and found it can be useful for measurement and prediction of housing sales prices.

The economic value of housing is very much affected by its geographical location. There have been many studies conducted to examine the influence of spatial effects on house prices [9,20–22]. Ref. [4] distinguished two levels of spatial effects: neighborhood effects and adjacency effects. Neighborhood effects are the locational attributes. Adjacency effects are externalities associated with the relative geographical position of houses. The conventional hedonic model utilizes the neighborhood effects by including a set of attributes which account for the socioeconomic and physical make-up of the neighborhood and the accessibility of houses to urban amenities. Adjacency effects also contribute to similarities among neighboring houses. We expect that adjacency effects will reduce after introducing neighborhood effects into the hedonic model by employing spatial variables. However, if any remaining systematic spatial pattern is observed, it is likely that potentially valuable information has not been retrieved from the process [9]. The use of techniques such as kriging and spatial autocorrelation models is then recommended. In this paper we use a set of town dummy variables to take into consideration adjacency effects as in many studies.

Closer inspection of this literature, however, suggests that the estimation methods of hedonic price models can be further improved. In this paper, we propose semiparametric kernel machines with spatial effects which are based on the least squares support vector machine (LS-SVM) and the kernel minimum squared error model (KMSEM). The support vector machine (SVM), first developed by Vapnik [32] and his group at AT&T Bell Laboratories, solves the weak point of neural network such as the existence of local minima in the area of statistical learning theory and structural risk minimization. SVM has been successfully applied to a number of real world problems related to classification and regression problems. One of its prominent advantages is the idea of using kernels to realize the nonlinear transformations without knowing the detailed transformations. Based on this idea, researchers proposed a class of kernel-based algorithms, such as kernel Fisher discriminant analysis (KFDA) [17], LS-SVM [30], kernel ridge regression (KRR) [26] and KMSEM [34]. LS-SVM is a modified version of SVM in a least squares sense. KMSEM is a generalization of the conventional minimum squared error model to yield a new type of nonlinear model, which was devised by using the theory of reproducing kernels and adding different penalty terms. Ref. [34] argued that LS-SVM can be viewed as a special case of KMSEM. Ref. [29] developed a semiparametric SVM which is useful in the case where domain knowledge exists about the function to be estimated or emphasis is put onto the understandability of the model. Ref. [27] proposed a semiparametric LS-SVM for the accelerated failure time model. Extensive empirical comparisons showed that LS-SVM and KMSEM obtain good performance on various classification and regression problems. To the best of our knowledge, however, these semiparametric kernel machines have not been utilized for housing price prediction.

The rest of this paper is organized as follows. Section 2 provides a brief discussion of semiparametric specifications for the hedonic price function and the estimation procedures. Section 3 proposes the SSELS-SVM and SSEKMSEM for predicting housing sales price. Sections 4 and 5 present numerical studies and conclusion, respectively.

2. Semiparametric house price models

In this section we provide a brief discussion of semiparametric specifications for the hedonic price function and the estimation procedures. Ref. [31] found out that the semiparametric estimators combined the benefits of the parametric and nonparametric estimations. The semiparametric models permitted the function to be linear, convex, or concave of curvilinear and sought for the best fitted model. Semiparametric method can act as a compromise between parametric and nonparametric methods, and obtain better estimators. Semiparametric regression models are often used in situations where the fully nonparametric model may not perform well.

Suppose that each residential house is viewed by economic agents as a bundle of different amounts of a vector of attributes. All these attributes are observed by the economic agents when making their choices. Hereafter, we assume that $\boldsymbol{x} \in R^p$ and $\boldsymbol{z} \in R^d$ represent vectors of dichotomous and nondichotomous attributes of the house, respectively. The hedonic price function specifies how the log of sale price, denoted by y, varies as the attributes vary, i.e.

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{z}_i) + \epsilon_i, \tag{1}$$

where i indexes houses and ϵ_i is the i.i.d. noise term. Eq. (1) is referred to as the hedonic price function.

Selecting an appropriate functional form for Eq. (1) has been a frequent issue. Given that an incorrect choice of functional form may result in inconsistent estimates, earlier studies have attempted to estimate hedonic price models with more flexible functional forms. Most of these attempts have concentrated on parametric specifications such as the Box–Cox transformation, which includes several popular functional forms as special cases. Box–Cox model in the specification of hedonic price model is expressed as follows:

$$y_i = b_0 + \boldsymbol{\beta}^t \boldsymbol{x}_i + \sum_{j=1}^{d} \gamma_j z_{ji}^{(\lambda)} + \epsilon_i, \tag{2}$$

where $z_{ji}^{(\lambda)} = ((z_{ji}^{\lambda} - 1)/\lambda)$ if $\lambda \neq 0$, and $z_{ji}^{(\lambda)} = \ln(z_{ji})$ if $\lambda = 0$. Eq. (2) is estimated using a maximum likelihood estimator. The Box–Cox transformation includes the semi-log ($\lambda = 1$) and the double-log ($\lambda = 0$) models as special cases depending on the transformation parameter λ. Only the continuous variables are subject to the transformation.

We face with the curse of dimensionality when the vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ involve a large number of attributes. Indeed, unconstrained nonparametric estimates of the unknown function $f(\cdot)$ deteriorate rapidly as $p+d$ increases. It is hence necessary to impose restrictions on $f(\cdot)$. One possibility is to allow $f(\cdot)$ to be nonparametric only in $\boldsymbol{z}$ and to specify a parametric form for $\boldsymbol{x}$, as illustrated in [25]. In this paper we assume a semiparametric model given by

$$y_i = b_0 + \boldsymbol{\beta}^t \boldsymbol{x}_i + f(\boldsymbol{z}_i) + \epsilon_i. \tag{3}$$

Ref. [1] used generalized additive model (GAM) well explained in [12] as semiparametric model, which is given by

$$y_i = b_0 + \boldsymbol{\beta}^t \boldsymbol{x}_i + \sum_{j=1}^{d} m_j(z_{ji}) + \epsilon_i. \tag{4}$$

The functions $m_j(z_{ji})$ in Eq. (4) are estimated using the iterative procedure known as the backfitting estimator, which reduces multivariate regression to successive simple regressions. In this paper we use the backfitting procedure explained in [1]. GAM is a powerful and flexible tool to explore large data.

3. SSELS-SVM and SSEKMSEM

The awareness of the limitations of conventional hedonic price models to account for spatial effects has led to the development and use of spatial statistical and econometric methods in real estate applications [20–22]. In this section we propose SSELS-SVM and SSEKMSEM for predicting housing sale prices and for exploring spatial effects on them, which combine parametric and