University of Victoria
Department of Economics

Econometrics Working Paper EWP0903
ISSN 1485-6441

# Bayesian Fuzzy Regression Analysis and Model Selection: Theory and Evidence

**Hui Feng[a], David E. Giles[b,*]**

[a] *Department of Economics, Business & Mathematics,*
*King's University College at the University of Western Ontario,*
*266 Epworth Avenue, London, ON N6A 2M3, Canada*

[b] *Department of Economics, University of Victoria,*
*P.O. Box 1700, STN CSC, Victoria, B.C., Canada V8W 2Y2*

**Revised, June 2009**

## Abstract

In this study we suggest a Bayesian approach to fuzzy clustering analysis – the Bayesian fuzzy regression. Bayesian Posterior Odds analysis is employed to select the correct number of clusters for the fuzzy regression analysis. In this study, we use a natural conjugate prior for the parameters, and we find that the Bayesian Posterior Odds provide a very powerful tool for choosing the number of clusters. The results from a Monte Carlo experiment and two real data applications of Bayesian fuzzy regression are very encouraging.

**Keywords:** Bayesian posterior odds, model selection, fuzzy regression, fuzzy clustering

[*] Corresponding Author: Tel: +1 250 721 8540, fax: +1 250 721 6214

*Email addresses*: hfeng8@uwo.ca (Hui Feng), dgiles@uvic.ca (David Giles)

# 1. Introduction

Recent developments in econometric modelling have emphasized a variety of non-linear specifications. Parametric examples of these include various regime-switching models (the threshold and Markov switching autoregressive models): threshold autoregressive (TAR) models, self-exciting threshold autoregressive (SETAR) models, smoothing threshold autoregressive (STAR) models, and various others. In addition, non-parametric and semi-parametric models are widely used, though the well-known "curse of dimensionality" can place some limitations on their use with multivariate data. The use of fuzzy clustering analysis in the context of econometric modelling is a rather new approach within the class of nonlinear econometric models [13]. Fuzzy clustering analysis is a very flexible and powerful technique that is used widely in the pattern recognition literature, and elsewhere, and has recently been applied in the area of modelling and forecasting economic variables [11, 13, 15, 16, 23].

In fuzzy regression modelling, we use the explanatory variables to partition the sample into a pre-assigned number of clusters. The membership functions are calculated using some algorithm for each of the observations, and these provide weights between zero and one which indicate the degree to which each observation belongs to each cluster. The fuzzy clustering regression is then obtained by taking the weighted average of the regression results from each of the clusters, using the membership functions as weights. To date, researchers have treated the number of fuzzy clusters as being pre-determined, and no formal procedures have been used to determine the "optimal" number of clusters. In practice, the number of clusters is set to be between one and four [12, 13, 14] with one simply being the usual case where the full sample is used (standard regression).

We observe that the choice of a particular value for the number of clusters implies the choice of a particular model specification. Altering the number of clusters changes the number of "sub-models" that are fitted to the data and subsequently combined into a final result. We approach this model-selection problem from a Bayesian perspective, and propose the use of the Bayesian posterior odds to determine the number of fuzzy clusters to be used in the fuzzy regression analysis. In fuzzy regression analysis, the models fitted to each cluster can be estimated by any appropriate technique. Ordinary Least Squares is a typical choice. Nonlinearity is modeled successfully because the fuzzy combination of the regression results from each cluster involves weights that vary continuously from one data-point to another. Here, however, in order to be consistent in our overall approach we use Bayesian estimation for each cluster's sub-model. More specifically, we adapt the standard Bayesian regression estimator based on the natural-conjugate prior density function to our clustering context. As a result, we refer to this overall modelling methodology as Bayesian Fuzzy Regression analysis.

This paper is organized as follows. The next section introduces fuzzy clustering analysis. Section 3 discusses some basic concepts associated with Bayesian inference, and sets up the Bayesian Posterior Odds analysis in a general model selection framework.  Section 4 derives the Bayesian Posterior Odds under the natural conjugate prior for choosing the number of fuzzy clusters. The design and results of an extensive Monte Carlo experiment are presented in section 5, and two simple applications using real data are discussed in section 6. The last section offers our conclusions and some further research suggestions.

## 2. Fuzzy clustering analysis

A classical set can be viewed as a "crisp" set, in the sense that it has a clearly defined boundary. For example, the set *E* could be defined as any integer that is greater than 10. The membership of a "crisp" set requires the individual element either be a member or not - the "degree of membership" is either unity or zero:

$$Mu(x) = \{\begin{matrix} 1 \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} x \in E \\ x \notin E \end{matrix} \quad . \tag{1}$$

On the other hand, a fuzzy set is just as the name implies: "without a crisp boundary".  The difference between a fuzzy set and a classical set lies in the nature of the membership for each element. Zadeh [26] defined the meaning of the membership for fuzzy sets to be a continuous number between zero and one. This means that any element can be associated with one or more clusters. Further, all of these membership values added together should equal unity. Generally, this association involves different degrees of membership with each of the fuzzy sets. Just as this makes the boundaries of the sets fuzzy, it makes the location of the centroid of the set fuzzy as well.

The “Fuzzy c-Means” (FCM) algorithm, which was developed and improved by Bezdek [2], Dunn [9, 10] and Ruspini [21], is frequently used in pattern recognition. The objective is to partition the data into fuzzy sets or clusters, to locate these clusters, and to quantify the degree of membership of every data-point with every cluster. It is based on minimization of the following functional:

$$J(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^{m}(d_{ik})^{2} \quad , \qquad (2)$$

where $u_{ik}$ is the “degree of membership” of data-point $k$ in cluster $i$, and $d_{ik}$ is the distance between data $x_k$ and the $i$-$th$ cluster center $v_i$. $c$ is the number of clusters presumed, and $m$ is any real number greater than 1, which measures the degree of the fuzziness. In the limit, as $m$ approaches unity, the membership becomes crisp. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership $u_{ik}$ and the cluster centers $v_i$ by:

$$u_{ik}=1/\{\sum_{j=1}^{n}[(d_{ik})^{2}/(d_{jk})^{2}]^{1/(m-1)}\} \qquad (3)$$

and,

$$v_{i}=[\sum_{k=1}^{n}(u_{ik})^{m}x_{k}]/[\sum_{k=1}^{n}(u_{ik})^{m}] \ ; i = 1, 2, \ldots, c. \qquad (4)$$

Over the years, there have been numerous developments in fuzzy set studies following the research by Zadeh [26]. Many applications can be found in the areas of computer science, systems analysis, electrical and electronic engineering, pattern recognition and psychology. In recent years, various applications of fuzzy sets in the area of econometrics have been led by Giles and his co-workers [4, 8, 12, 13, 14, 15, 16, 17]. In these applications, “fuzzy regression”, analysis is introduced. After clustering the data on the basis of the explanatory variables, separate regressions are fitted over each of the clusters. The overall fuzzy regression is obtained by combining the results from each cluster using the membership functions as the weights. As these weights vary continuously over the sample, even if linear models are fitted to each cluster, the resulting model can capture complex non-linearities in the data. See [13, 17] for full details. We have written code for commands in the SHAZAM econometrics package [22] for the analysis in this paper.

For the fuzzy clustering analysis, the number of clusters *c* and the fuzziness parameter *m* have to be assigned before the actual regression analysis. Usually, for researchers, the common choice of *m* is 2. In previous fuzzy regression applications involving economic data (e.g., [13, 15, 17, 21]) the experience has been that the appropriate number of fuzzy clusters is usually four or less, and it is uncommon for *c* to exceed 6. However, in the past no formal procedures have been introduced for the choice of these parameters. Different values of *c* and *m* imply that different numbers of clusters and different degrees of fuzziness will be used for the regression analysis, which means effectively that different fuzzy models are going to be employed. This suggests a model-selection problem. What are the optimal choices for *c* and *m* for a given sample of data when applying fuzzy regression? In order to simplify the analysis, in this paper we fix the value of *m* to be 2, and view the selection of *c* as the formal model-selection problem.[1] Here, we adopt a Bayesian model-selection approach, and use Bayesian posterior odds analysis to select the value of *c*.

## 3. Bayesian posterior odds analysis

### *3.1 Bayesian model selection*

An excellent discussion of Bayesian model selection is provided by Zellner [27, pp.292-298]. In order to discuss model selection *via* Bayesian posterior odds analysis, we first need to introduce some basic notation and concepts. Let $M$ denote the model space, and $M_i$ denote the $i^{\text{th}}$ candidate model. The number of models is $N$, a finite or countably infinite value. So, $M = \{M_i\}$ $i$=1, 2, 3,..., $N$. There are two sets of prior information that we need to consider in our Bayesian analysis here: the prior information about the models and the prior information about the parameters of those models.

First, assigning a prior mass function over $M$, let $p(M_i)$ denote the prior probability that the $i^{\text{th}}$ model is the 'true' model, where

$$0 \leq p(M_i) \leq 1$$

and

$$\sum_{i=1}^{N} p(M_i) = 1 \ .$$

Suppose that the $i^{th}$ model has a vector of parameters, $\theta_i$, belonging to a parameter space, $\Omega_i$, $i$ = 1, 2, …., $N$. Let $p(\theta_i \mid M_i)$ denote the prior p.d.f. for the parameters of the $i^{\text{th}}$ model. For each

[1] Our analysis could be extended to the case where both *m* and *c* are to be selected, though the fact that the former parameter can take a continuum of values would complicate matters.

model in $M$ the joint data density is given by $p(y \mid \theta_i, M_i)$, and viewed as a function of the parameters this is the usual likelihood function, $l(\theta_i \mid y, M_i)$. The conditional (on the model) data density is obtained as

$$p(y \mid M_i) = \int_{\Omega_i} p(y \mid \theta_i, M_i) p(\theta_i \mid M_i) d\theta_i , \tag{5}$$

and the marginal data density is

$$p(y) = \sum_{j=1}^{m^*} p(M_j) p(y \mid M_j) . \tag{6}$$

Applying Bayes' Theorem, the posterior probability of model $M_i$ is

$$p(M_i \mid y) = \frac{p(M_i) p(y \mid M_i)}{p(y)} \propto p(M_i) p(y \mid M_i) \qquad . \tag{7}$$

Note that the posterior probability of $M_i$ will be calculated incorrectly if the calculation of $p(y)$ is incorrect, and this will be the case in the (likely) event that the model space, $M$, is not fully specified. So, we use the Bayesian posterior odds (BPO) to select the preferred because, even if $M$ is incompletely specified, the BPO will always be correct, for the following reason.

Let $[p(M_i) / p(M_j)]$ be the prior odds between model $i$ and $j$. Then

$$BPO_{ij} = \frac{p(M_i \mid y)}{p(M_j \mid y)} = \{\frac{p(M_i) p(y \mid M_i) / p(y)}{p(M_j) p(y \mid M_j) / p(y)}\} = [\frac{p(M_i)}{p(M_j)}][\frac{p(y \mid M_i)}{p(y \mid M_j)}]. \tag{8}$$

That is:

Bayesian Posterior Odds = [Prior Odds] × [Bayes Factor].

These posterior odds are independent of $p(y)$, and hence of the specification of $M$. Of course, if $M$ is complete, we could use the BPO to calculate individual posterior probabilities which reflect our beliefs about each model being a true model, given the sample information. In the context of the model selection problem, after obtaining the BPO, often we will still want to come to a conclusion of rejecting or accepting one model compared with some other competing model. This is a *two-action* problem [27, pp.291].

In general terms, let us consider two competing hypotheses – a true hypothesis, $H_0$ and a false hypothesis, $H_1$. Let $\hat{H}_i$ denote the action of choosing the $i^{th}$ hypothesis ($i$ = 1, 2). If we accept the true hypothesis or we reject the false hypothesis, we will incur zero loss. However, our loss would be the positive amount $L(H_1 , \hat{H}_0)$ if we accept the false null hypothesis, and our loss would be $L(H_0, \hat{H}_1)$ if we reject the true hypothesis. Which model will be accepted depends on which model minimizes posterior expected loss:

If $E(L \mid \hat{H}_0) < E(L \mid \hat{H}_1)$, Accept $H_0$.

If $E(L \mid \hat{H}_1) < E(L \mid \hat{H}_0)$, Accept $H_1$.

It is well-known that under any symmetric loss function[2], $H_0$ will be accepted only if

$$p(H_0 \mid y) > p(H_1 \mid y), \text{ or } \frac{p(H_0 \mid y)}{p(H_1 \mid y)} > 1.$$

Again, it should be emphasized that this determination of the rankings of the models through the BPO does not require the calculation of the individual posterior probabilities for the individual models, and so it is not affected if the model space is under-specified.

### *3.2 Bayesian estimation*

In the above discussion, each model was parameterized by a vector of parameters, about which prior information was needed in order to construct the BPO. It is natural, therefore, to approach the estimation of these parameters from a Bayesian perspective. Such estimation is necessary in order to complete our fuzzy regression analysis.

The Bayes and Minimum Expected Loss (MEL) estimators coincide if the posterior expected loss is finite, so the discussion here focuses on MEL estimation, strictly speaking. Let us suppose that model $M_i$ has been selected. Applying Bayes' theorem, we first obtain the posterior p.d.f. for $\theta_i$ as:

$$p(\theta_i \mid y, M_i) \propto p(\theta_i \mid M_i) p(y \mid \theta_i, M_i) . \tag{9}$$

Then the MEL estimator of $\theta_i$ is the $\hat{\theta}_i$ that minimizes the posterior expected loss,

[2] We could assume that the loss is asymmetric. This would simply alter the “threshold” value for the BPO – i.e., the value at which the odds lead us to “flip” from choosing one to model to choosing the other.

$$\int_{\Omega_i} L(\theta_i, \hat{\theta}_i) p(\theta_i \mid y, M_i) \, d\theta_i \,. \qquad (10)$$

It is well known that if the loss function is quadratic, then $\hat{\theta}_i$ is the mean of the distribution described by $p(\theta_i \mid y, M_i)$; for an absolute error loss this MEL estimator is the median of this distribution; and $\hat{\theta}_i$ is the mode of this distribution in the case of a zero-one loss function [3, pp. 308-309]. In what follows in the next section our choice of prior p.d.f. for the parameters results in a posterior density whose characteristics ensure that the same MEL estimator arises under each of these three particular loss functions.

## 4. Bayesian fuzzy regression analysis

As was indicated in section 1, different choices for the number of fuzzy clusters, *c*, generate different fuzzy regression models. For example, letting the value of *c* run from 1 to 4 implies that there are four possible fuzzy models. The first model is the one-cluster fuzzy model (i.e., all of the data are used, and we have a conventional regression situation); the second model is based on two fuzzy clusters over the sample; and the third and fourth models assume there are respectively three and four fuzzy clusters over the sample range. The regressions that we fit over each cluster can be of any kind, depending on the nature of the data. In our case they will be linear multiple regressions, with normally distributed errors.

The assumption of normally distributed errors is quite standard, and actually is far less restrictive than might be imagined. This assumption implies that if we estimate these regressions by least squares, we are actually using the maximum likelihood estimator (MLE), or the Bayes estimator with a diffuse prior. It is well known that the equivalence between least squares and MLE also holds under various other distributional assumptions for the errors, such as multivariate Student-t. In fact the situation is even more general than this. Kariya and Eaton [18] and King [19] have shown that any scale-invariant statistic based on least squares regression has the same distribution if the error distribution is a member of the elliptically symmetric family (e.g., Chmielewski [5]). So, our assumption of normally distributed errors can be relaxed to let the error distribution be any member of the elliptically symmetric family without substantively affecting our results based on least squares estimation, or Bayesian estimation with a relatively uninformative natural conjugate prior.

Now, let

$$
\begin{aligned}
M_1: \quad & y_{11} = X_{11}\beta_{11} + u_{11} && ;\ u_{11} \sim \mathrm{N}(0,\ \sigma_{11}\mathrm{I}_{n11}) \\
M_2: \quad & y_{21} = X_{21}\beta_{21} + u_{21} && ;\ u_{21} \sim \mathrm{N}(0,\ \sigma_{21}\mathrm{I}_{n21}) \\
& y_{22} = X_{22}\beta_{22} + u_{22} && ;\ u_{22} \sim \mathrm{N}(0,\ \sigma_{22}\mathrm{I}_{n22}) \\
M_3: \quad & y_{31} = X_{31}\beta_{31} + u_{31} && ;\ u_{31} \sim \mathrm{N}(0,\ \sigma_{31}\mathrm{I}_{n31}) \\
& y_{32} = X_{32}\beta_{32} + u_{32} && ;\ u_{32} \sim \mathrm{N}(0,\ \sigma_{32}\mathrm{I}_{n32}) \\
& y_{33} = X_{33}\beta_{33} + u_{33} && ;\ u_{33} \sim \mathrm{N}(0,\ \sigma_{33}\mathrm{I}_{n33}) \\
M_4: \quad & y_{41} = X_{41}\beta_{41} + u_{41} && ;\ u_{41} \sim \mathrm{N}(0,\ \sigma_{41}\mathrm{I}_{n41}) \\
& y_{42} = X_{42}\beta_{42} + u_{42} && ;\ u_{42} \sim \mathrm{N}(0,\ \sigma_{42}\mathrm{I}_{n42}) \\
& y_{43} = X_{43}\beta_{43} + u_{43} && ;\ u_{43} \sim \mathrm{N}(0,\ \sigma_{43}\mathrm{I}_{n43}) \\
& y_{44} = X_{44}\beta_{44} + u_{44} && ;\ u_{44} \sim \mathrm{N}(0,\ \sigma_{44}\mathrm{I}_{n44})
\end{aligned}
$$

The $X_{ij}$'s are ($n_{ij}$ × $k$) matrices, each with rank, where $n_{ij}$ is the number of observations in the $j^{\text{th}}$ cluster for the $i^{\text{th}}$ model. The $\beta_{ij}$ are each ($k$ × 1) coefficient vectors, and the $u_{ij}$ are ($n_{ij}$ × 1) vectors of random error terms. $\beta_{ij}$ and $\sigma_{ij}$ are the coefficient vector and error standard deviation for the $j^{th}$ cluster of model $i$, for $i, j$ = 1, 2, 3, 4.

The prior probabilities associated with each model are denoted $p(M_1)$, $p(M_2)$, $p(M_3)$ and $p(M_4)$. The prior p.d.f.'s for the parameters $\beta_{ij}$ and $\sigma_{ij}$ ($i$, $j$ = 1, 2, 3, 4) are taken to be the natural conjugate prior densities. Obviously, other possibilities could be considered, but this choice will suffice to illustrate the methodology and it provides an interesting (and mathematically tractable) benchmark. The following methodology would be unaltered if alternative prior p.d.f.'s were used, though the specific results would change, of course. In our case, we have:

$$
p(\beta_{ij}, \sigma_{ij}) = p(\beta_{ij} \mid \sigma_{ij})\, p(\sigma_{ij})
$$

$$
p(\beta_{ij} \mid \sigma_{ij}) = \frac{\left|C_{ij}\right|^{1/2}}{(2\pi)^{k_{ij}/2}\sigma_i^{k_{ij}}} \exp[-\frac{1}{2\sigma_{ij}^{\ 2}}(\beta_{ij} - \overline{\beta}_{ij})' C_{ij} (\beta_{ij} - \overline{\beta}_{ij})] \qquad (11)
$$

$$p(\sigma_{ij}) = \frac{K_{ij}}{\sigma_{ij}^{q_{ij}+1}} \exp(-\frac{q_{ij}\bar{s}_{ij}^2}{2\sigma_{ij}^2}) \ , \quad i, j = 1, 2, 3, 4$$

where the normalizing constant is $K_{ij} = 2(q_{ij}\bar{s}_{ij}^2 / 2)^{q_{ij}/2} / \Gamma(q_{ij} / 2)$. That is, the conditional prior p.d.f. for $\beta_{ij}$ given $\sigma_{ij}$ is multivariate normal with prior mean vector $\bar{\beta}_{ij}$ and covariance matrix $\sigma_{ij}^2 C_{ij}^{-1}$. The marginal prior information for $\sigma_{ij}$ is represented by an inverted gamma density, with parameters $q_{ij}$ and $\bar{s}_{ij}^2$ to be assigned values by the investigator. For this marginal prior to be proper, we need $0 < q_{ij}, \ \bar{s}_{ij}^2 < \infty$; $i, j = 1, 2, 3, 4$.

We can now proceed with the Bayesian posterior odds calculations. Equation (8) gives us the formula for the BPO. However, in order to get the BPO we need to derive the conditional data densities for each of the models. Given the assumption of normal errors in all of the sub-models, the likelihood function for Model 1, for example, is:

$$p(y_{11} \mid \beta_{11}, \sigma_{11}, M_1) = \frac{1}{(2\pi)^{n_{11}/2}} \frac{1}{\sigma_{11}{}^{n_{11}}} \exp\{\frac{1}{2\sigma_{11}^2}[v_{11}s_{11}^2 + (\beta_{11} - \hat{\beta}_{11})' X_{11}{}' X_{11}(\beta_{11} - \hat{\beta}_{11})]\} \tag{12}$$

where $\hat{\beta}_{11} = (X_{11}{}' X_{11})^{-1} X_{11}{}' y_{11}$, $v_{11} = n_{11} - k$ and $v_{11}s_{11}^2 = (y_{11} - X_{11}\hat{\beta}_{11})'(y_{11} - X_{11}\hat{\beta}_{11})$.

The likelihoods associated with the various clusters for the multi-cluster models follow in an obvious manner.

From (5), for the first model, which has one cluster, the conditional data density is:

$$p(y \mid M_1) = \iint p(y_{11} \mid \beta_{11}, \sigma_{11}, M_1) p(\beta_{11} \mid \sigma_{11}) p(\sigma_{11}) d\beta_{11} d\sigma_{11} \qquad .$$

To consider the models based on two or more clusters we assume that the clusters are generated independently of each other. In many cases this is a very reasonable assumption, especially with cross-section data. For example, the clusters might implicitly relate to different countries at one point in time, with no spatial autocorrelation. In other situations the independence assumption may be less innocuous, but in such cases any underlying dependencies between the clusters would have to be known, or would need to be modeled in order to proceed with the analysis. We do not consider such extensions in this study. So, making this assumption of independence, the conditional data density for Model 2, with two clusters, is:

$$p(y|M_2)=\iint p(y_{21}|\beta_{21},\sigma_{21},M_2)p(\beta_{21}|\sigma_{21})p(\sigma_{21})\,d\beta_{21}d\sigma_{21}\iint p(y_{22}|\beta_{22},\sigma_{22},M_2)p(\beta_{22}|\sigma_{22})p(\sigma_{22})\,d\beta_{22}d\sigma_{22} \tag{13}$$

For Model 3 with three clusters this density is:

$$\begin{aligned}p(y|M_3)=&\iint p(y_{31}|\beta_{31},\sigma_{31},M_3)p(\beta_{31}|\sigma_{31})p(\sigma_{31})\,d\beta_{31}d\sigma_{31}\iint p(y_{32}|\beta_{32},\sigma_{32},M_3)p(\beta_{32}|\sigma_{32})p(\sigma_{32})\,d\beta_{32}d\sigma_{32}\\ &\times\iint p(y_{33}|\beta_{33},\sigma_{33},M_3)p(\beta_{33}|\sigma_{33})p(\sigma_{33})\,d\beta_{33}d\sigma_{33}\end{aligned}$$

For Model 4 with four clusters it is:

$$\begin{aligned}p(y|M_4)=&\iint p_{41}(y|\beta_{41},\sigma_{41},M_4)p(\beta_{41}|\sigma_{41})p(\sigma_{41})\,d\beta_{41}d\sigma_{41}\iint p(y_{42}|\beta_{42},\sigma_{42},M_4)p(\beta_{42}|\sigma_{42})p(\sigma_{42})\,d\beta_{42}d\sigma_{42}\\ &\times\iint p(y_{43}|\beta_{43},\sigma_{43},M_4)p(\beta_{43}|\sigma_{43})p(\sigma_{43})\,d\beta_{43}d\sigma_{43}\iint p(y_{44}|\beta_{44},\sigma_{44},M_4)p(\beta_{44}|\sigma_{44})p(\sigma_{44})\,d\beta_{44}d\sigma_{44}\end{aligned} \tag{14}$$

where, as before, the $\beta_{ij}$ and $\sigma_{ij}$ are the parameters for the $j^{th}$ cluster of model *i*, for *i, j* = 1, 2, 3, 4. Using the results of Zellner [27, pp. 306-312]:

$$\begin{aligned}p(y|M_1)&=\iint p(y_{11}|\beta_{11},\sigma_{11},M_1)p(\beta_{11}|\sigma_{11})p(\sigma_{11})\,d\beta_{11}d\sigma_{11}\\ &=\iint\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma_{11}{}^{n}}\exp\{\frac{1}{2\sigma_{11}^2}[v_{11}s_{11}^2+(\beta_{11}-\hat{\beta}_{11})'X_{11}{}'X_{11}(\beta_{11}-\hat{\beta}_{11})]\}\\ &\quad\times\frac{\left|C_{11}\right|^{1/2}}{(2\pi)^{k_{11}/2}\sigma_{11}{}^{k_{11}}}\exp[-\frac{1}{2\sigma_{11}{}^{2}}(\beta_{11}-\overline{\beta}_{11})'C_{11}(\beta_{11}-\overline{\beta}_{11})]\\ &\quad\times\frac{K_{11}}{\sigma_{11}^{q_{11}+1}}\exp(-\frac{q_{11}\overline{s}_{11}^2}{2\sigma_{11}^2})\,d\beta_{11}d\sigma_{11}\\ &=\frac{1}{2}(2\pi)^{-n/2}K_{11}\left|\frac{\left|C_{11}\right|}{\left|A_{11}\right|}\right|^{1/2}2^{\frac{n_{11}+q_{11}}{2}}\Gamma(\frac{n_{11}+q_{11}}{2})(q_{11}\overline{s}_{11}^2+v_{11}s_{11}^2+Q_{11a}+Q_{11b})^{-\frac{n_{11}+q_{11}}{2}}\end{aligned} \tag{15}$$

where $A_{11}=C_{11}+X_{11}{}'X_{11}$ ; $\tilde{\beta}_{11}=A_{11}^{-1}(C_{11}\overline{\beta}_{11}+X_{11}{}'X_{11}\hat{\beta}_{11})$

$$Q_{11a}=(\overline{\beta}_{11}-\tilde{\beta}_{11})'C_{11}(\overline{\beta}_{11}-\tilde{\beta}_{11})\,;\qquad Q_{11b}=(\hat{\beta}_{11}-\tilde{\beta}_{11})X_{11}{}'X_{11}(\hat{\beta}_{11}-\tilde{\beta}_{11}) \tag{16}$$

$$K_{11}=\frac{2(q_{11}\overline{s}_{11}^2/2)^{q_{11}/2}}{\Gamma(q_{11}/2)}$$

Similar operations are used to determine the conditional data densities for the other three models. With independent clusters, the results for the other models with more than one cluster can be written as the product of the integrals for each of the clusters. So, using notation that is an obvious generalization of that in (16):

$$p(y\,|\,M_2) = \frac{1}{2}K_{21}\left\|\frac{|C_{21}|}{|A_{21}|}\right\|^{1/2} 2^{\frac{n_{21}+q_{21}}{2}}\Gamma(\frac{n_{21}+q_{21}}{2})(q_{21}\bar{s}_{21}^2 + v_{21}s_{21}^2 + Q_{21a} + Q_{21b})^{-\frac{n_{21}+q_{21}}{2}}$$

$$\times\frac{1}{2}K_{22}\left\|\frac{|C_{22}|}{|A_{22}|}\right\|^{1/2} 2^{\frac{n_{22}+q_{22}}{2}}\Gamma(\frac{n_{22}+q_{22}}{2})(q_{22}\bar{s}_{22}^2 + v_{22}s_{22}^2 + Q_{22a} + Q_{22b})^{-\frac{n_{22}+q_{22}}{2}}$$

$$= \frac{1}{4}(2\pi)^{-n/2}K_{21}K_{22}\left[\frac{|C_{21}\|C_{22}|}{|A_{21}\|A_{22}|}\right]^{1/2}$$

$$\times\frac{(q_{21}\bar{s}_{21}^2 + v_{21}s_{21}^2 + Q_{21a} + Q_{21b})^{-\frac{n_{21}+q_{21}}{2}}(q_{22}\bar{s}_{22}^2 + v_{22}s_{22}^2 + Q_{22a} + Q_{22b})^{-\frac{n_{22}+q_{22}}{2}}}{2^{\frac{n_{21}+q_{21}}{2}}\Gamma(\frac{n_{21}+q_{21}}{2})2^{\frac{n_{22}+q_{22}}{2}}\Gamma(\frac{n_{22}+q_{22}}{2})}$$

(17)

Then, using these results in (8), the BPO between Models 1 and 2 are:

$$BPO_{12} = [\frac{p(M_1)}{p(M_2)}]\times[\frac{p(y\,|\,M_1)}{p(y\,|\,M_2)}]$$

$$= 2\times[\frac{p(M_1)}{p(M_2)}]\times\frac{K_{11}}{K_{21}K_{22}}\left[\frac{|C_{11}\|A_{21}\|A_{22}|}{|A_{11}\|C_{21}\|C_{22}|}\right]^{1/2}\frac{2^{\frac{n_{11}+q_{11}}{2}}\Gamma(\frac{n_{11}+q_{11}}{2})}{2^{\frac{n_{21}+q_{21}}{2}}\Gamma(\frac{n_{21}+q_{21}}{2})2^{\frac{n_{22}+q_{22}}{2}}\Gamma(\frac{n_{22}+q_{22}}{2})}$$

$$\times\frac{(q_{21}\bar{s}_{21}^2 + v_{21}s_{21}^2 + Q_{21a} + Q_{21b})^{-\frac{n_{21}+q_{21}}{2}}(q_{22}\bar{s}_{22}^2 + v_{22}s_{22}^2 + Q_{22a} + Q_{22b})^{-\frac{n_{22}+q_{22}}{2}}}{(q_{11}\bar{s}_{11}^2 + v_{11}s_{11}^2 + Q_{11a} + Q_{11b})^{-\frac{n_{11}+q_{11}}{2}}}$$

$$= 2\times[\frac{p(M_1)}{p(M_2)}]\times\left[\frac{|C_{11}\|A_{21}\|A_{22}|}{|A_{11}\|C_{21}\|C_{22}|}\right]^{1/2}\times(\frac{\delta_{11}{}^{n_{21}}}{\delta_{21}{}^{n_{21}}\delta_{22}{}^{n_{22}}})^{-1/2} \qquad (18)$$

$$\times\frac{\bar{s}_{11}{}^2/\delta_{11}}{(\bar{s}_{21}{}^2/\delta_{21})(\bar{s}_{22}{}^2/\delta_{22})}\frac{f_{q_{11},n_{11}}(\bar{s}_{11}{}^2/\delta_{11})}{f_{q_{21},n_{21}}(\bar{s}_{21}{}^2/\delta_{21})f_{q_{22},n_{22}}(\bar{s}_{22}{}^2/\delta_{22})}$$

where

$$K_{ij} = \frac{2(q_{ij}\bar{s}_{ij}^2/2)^{q_{ij}/2}}{\Gamma(q_{ij}/2)}$$

$$\delta_{ij} = (vs_{ij}{}^{2} + Q_{ija} + Q_{ijb}) / n_{ij} \, ,$$

and $f_{q_{ij},n_{ij}}(\bar{s}_1{}^{2} / \delta_{ij})$ denotes the ordinate of the p.d.f. of the *F* distribution with $q_{ij}$ and $n_{ij}$ degrees of freedom. The other notation in this BPO formula is again an obvious generalization of that used in (16).

Similarly, the BPO between Models 1 and 3 are:

$$\begin{aligned}
BPO_{13} &= [\frac{p(M_1)}{p(M_3)}] \times [\frac{p(y \mid M_1)}{p(y \mid M_3)}] \\
&= [\frac{p(M_1)}{p(M_3)}] \times 2^{3-1} \frac{K_{11}}{K_{31}K_{32}K_{33}} \left[ \frac{|C_{11}\|A_{31}\|A_{32}\|A_{33}|}{|A_{11}\|C_{31}\|C_{32}\|C_{33}|} \right]^{1/2} \\
&\quad \times \frac{2^{\frac{n_{11}+q_{11}}{2}} \Gamma(\frac{n_{11}+q_{11}}{2})}{2^{\frac{n_{31}+q_{31}}{2}} \Gamma(\frac{n_{31}+q_{31}}{2}) 2^{\frac{n_{32}+q_{32}}{2}} \Gamma(\frac{n_{32}+q_{32}}{2}) 2^{\frac{n_{33}+q_{33}}{2}} \Gamma(\frac{n_{33}+q_{33}}{2})} \\
&\quad \times (q_{31}\bar{s}_{31}^2 + v_{31}s_{31}^2 + Q_{31a} + Q_{31b})^{-\frac{n_{31}+q_{31}}{2}} \\
&\quad \times \frac{(q_{22}\bar{s}_{32}^2 + v_{22}s_{32}^2 + Q_{32a} + Q_{32b})^{-\frac{n_{32}+q_{32}}{2}} (q_{33}\bar{s}_{33}^2 + v_{33}s_{33}^2 + Q_{33a} + Q_{33b})^{-\frac{n_{33}+q_{33}}{2}}}{(q_{11}\bar{s}_{11}^2 + v_{11}s_{11}^2 + Q_{11a} + Q_{11b})^{-\frac{n_{11}+q_{11}}{2}}} \\
&= 2 \times [\frac{p(M_1)}{p(M_3)}] \times \left[ \frac{|C_{11}\|A_{31}\|A_{32}\|A_{33}|}{|A_{11}\|C_{31}\|C_{32}\|C_{33}|} \right]^{1/2} \times (\frac{\delta_{11}}{\delta_{31}\delta_{32}\delta_{33}})^{-n/2} \times \frac{\bar{s}_{11}{}^{2} / \delta_{11}}{(\bar{s}_{31}{}^{2} / \delta_{31})(\bar{s}_{32}{}^{2} / \delta_{32})(\bar{s}_{33}{}^{2} / \delta_{33})} \\
&\quad \times \frac{f_{q_{11},n_{11}}(\frac{\bar{s}_{11}{}^{2}}{\delta_{11}})}{f_{q_{31},n_{31}}(\frac{\bar{s}_{31}{}^{2}}{\delta_{31}}) f_{q_{32},n_{32}}(\frac{\bar{s}_{32}{}^{2}}{\delta_{32}}) f_{q_{33},n_{33}}(\frac{\bar{s}_{33}{}^{2}}{\delta_{33}})} \qquad (19)
\end{aligned}$$

The BPO between Models 2 and 3 are:

$$
\begin{aligned}
BPO_{23} &= [\frac{p(M_2)}{p(M_3)}]\times[\frac{p(y\,|\,M_2)}{p(y\,|\,M_3)}] \\
&= 2^{3-2}\times[\frac{p(M_2)}{p(M_3)}]\times\frac{K_{21}K_{22}}{K_{31}K_{32}K_{33}}\left[\frac{|C_{21}\|C_{22}\|A_{31}\|A_{32}\|A_{33}|}{|A_{21}\|A_{22}\|C_{31}\|C_{32}\|C_{33}|}\right]^{1/2} \\
&\quad\times\frac{2^{\frac{n_{21}+q_{21}}{2}}\Gamma(\frac{n_{21}+q_{21}}{2})2^{\frac{n_{22}+q_{22}}{2}}\Gamma(\frac{n_{22}+q_{22}}{2})}{2^{\frac{n_{31}+q_{31}}{2}}\Gamma(\frac{n_{31}+q_{31}}{2})2^{\frac{n_{32}+q_{32}}{2}}\Gamma(\frac{n_{32}+q_{32}}{2})2^{\frac{n_{33}+q_{33}}{2}}\Gamma(\frac{n_{33}+q_{33}}{2})} \\
&\quad\times(q_{31}\bar{s}_{31}^2+v_{31}s_{31}^2+Q_{31a}+Q_{31b})^{-\frac{n_{31}+q_{31}}{2}} \\
&\quad\times\frac{(q_{22}\bar{s}_{32}^2+v_{22}s_{32}^2+Q_{32a}+Q_{32b})^{-\frac{n_{32}+q_{32}}{2}}(q_{33}\bar{s}_{33}^2+v_{33}s_{33}^2+Q_{33a}+Q_{33b})^{-\frac{n_{33}+q_{33}}{2}}}{(q_{21}\bar{s}_{21}^2+v_{21}s_{21}^2+Q_{21a}+Q_{21b})^{-\frac{n_{21}+q_{21}}{2}}(q_{22}\bar{s}_{22}^2+v_{22}s_{22}^2+Q_{22a}+Q_{22b})^{-\frac{n_{22}+q_{22}}{2}}} \\
&= [\frac{p(M_2)}{p(M_3)}]\times\left[\frac{|C_{21}\|C_{22}\|A_{31}\|A_{32}\|A_{33}|}{|A_{21}\|A_{22}\|C_{31}\|C_{32}\|C_{33}|}\right]^{1/2}\times(\frac{\delta_{21}{}^{n_{21}}\delta_{22}{}^{n_{22}}}{\delta_{31}{}^{n_{31}}\delta_{32}{}^{n_{32}}\delta_{33}{}^{n_{33}}})^{-1/2} \\
&\quad\times\frac{(\bar{s}_{21}{}^2/\delta_{21})(\bar{s}_{22}{}^2/\delta_{22})}{(\bar{s}_{31}{}^2/\delta_{31})(\bar{s}_{32}{}^2/\delta_{32})(\bar{s}_{33}{}^2/\delta_{33})}\frac{f_{q_{21},n_{21}}(\frac{\bar{s}_{21}{}^2}{\delta_{21}})f_{q_{22},n_{22}}(\frac{\bar{s}_{22}{}^2}{\delta_{22}})}{f_{q_{31},n_{31}}(\frac{\bar{s}_{31}{}^2}{\delta_{31}})f_{q_{32},n_{32}}(\frac{\bar{s}_{32}{}^2}{\delta_{32}})f_{q_{33},n_{33}}(\frac{\bar{s}_{33}{}^2}{\delta_{33}})}
\end{aligned}
\tag{20}
$$

Further details relating to the four-cluster model are available from the authors on request, to conserve space.

After we calculate the Bayesian posterior odds, under a symmetric loss function, if the odds are greater than 1, say $\text{BPO}_{12} > 1$, then Model 1 is preferred to Model 2, etc. The prior information becomes diffuse in this natural conjugate prior case as $|C_{ij}| \rightarrow 0$ and $q_{11} = q_{21} = q_{22} \rightarrow 0$. Under these condition, the Bayesian estimator collapses to the OLS estimator. In this case it is readily shown that under some mild conditions, and with prior odds for the model of unity, choosing Model 1 if $\text{BPO}_{12}$ exceeds unity (as would be the case under a symmetric loss function) is equivalent to choosing the model with the higher coefficient of determination ($R^2$). This last result, and a comprehensive discussion of the roles of the various terms in $\text{BPO}_{12}$ can be found in Zellner [27, pp. 310-311]. It is also well known that the BPO become indeterminate in the case of a totally diffuse, a point that has been taken into account in the next two sections.

In summary, our Bayesian fuzzy regression analysis involves clustering the data into *c* fuzzy clusters, where the optimal value of *c* is determined by BPO analysis. Then, Bayesian regression models are fitted over each of the *c* clusters, and the results are combined using the membership functions as weights. Of course, this procedure could be generalized still further by undertaking a form of Bayesian "model averaging". Specifically, separate Bayesian fuzzy regression models can be obtained for each value of *c* under consideration. Then, rather than selecting just one model on the basis of the BPO, a weighted average of all of the models' predictions can be formed, using the posterior probabilities for the model space as the weights. This presumes that the model space is well specified. An example of this Bayesian model averaging is provided in section 6.

## 5. Monte Carlo experiment

### *5.1 Experimental design*

We have undertaken a Monte Carlo experiment to assess the performance of the above BPO analysis in selecting the appropriate number of clusters to use in fuzzy regression analysis. There are four exercises in this experiment. The first exercise involves a one-cluster data generating process (DGP). The other three exercises in the experiment involve two-cluster, three-cluster and four-cluster DGPs. Ideally, the BPO should favor a fuzzy model with the true number of clusters all of the time. The programming for the Monte Carlo experiment has been done with SHAZAM [22] code.

The sample sizes that have been considered are $n$ = 24, 60, 120, 240, 480 and 1200 for every exercise of the experiment. The number of Monte Carlo repetitions was 1000, which was found to be more than sufficient to ensure the stability of the results. Further discussion of the accuracy associated with this number of repetitions is provided in association with Table 1 below. The DGP that has been used comprises a number of separate line segments, one for each cluster in the underlying process:

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij} \quad ; \; i = 1, 2, 3, \ldots, (n/c) \; ; \; j = 1, 2, \ldots, c.$$

In this experiment, $c$ runs from 1 to 4. The regressor series is $\{x_i\} = \{(i / 100): i = 1, 2, \ldots, n\}$, and the observations are assigned sequentially, and in equal numbers, to each cluster in turn according to the DGP. The intercept parameters take the values {1}, {1, 5}, {1, 5, 6} and {1, 5, 6, 9} according to the number of clusters in each exercise. Within each exercise, two cases are considered as far as the slope coefficients for each cluster in the DGP are concerned. The first (the "big slopes" case) has $\beta$ values of {1}, {1, -5}, {1, -5, 3} and {1, -5, 3, -8}. The second (the "small slope" case) has $\beta$ values of {0.1}, {0.1, -0.5}, {0.1, -0.5, 0.3} and {0.1, -0.5, 0.3, -0.8}. The error terms, $\varepsilon_{ij}$, are generated as being independent and normally distributed[3], with mean zero. In order to make the results more general, in each exercise of the experiment we let the error term standard deviation of the DGP vary between 1 and 10, which helps us to increase the DGP's degree of fuzziness and this in turn provides a significant challenge for the BPO analysis, and the upper limit of $\sigma = 10$ being sufficient for the results to stabilize.

Figures 1 and 2 show the four data DGPs when $n = 240$. In each case the standard deviation for the DGP is 3. These graphs provide an indication of the degree of fuzziness of the data that are used in the Monte Carlo experiment, and they show how difficult it would be for us in real life to determine the correct number of clusters just by simply looking at a data plot alone. These plots relate to only a small part of all of our Monte Carlo experiments.

We made the prior information for the parameters minimal by setting $q_{ij} = 0.01$ and $C_{ij} = 0.00001$ (for all $i, j$). The values for the other parameters of the prior are $\overline{\beta}_{ij} = 0.2$ and $\overline{s}_{ij}^{2} = 2$ (for all $i, j$), although the results are fully robust to these last choices. We assign equal prior probabilities for each of the competing models, so that we do not favor any particular model before the analysis.

### *5.2 Experiment results*

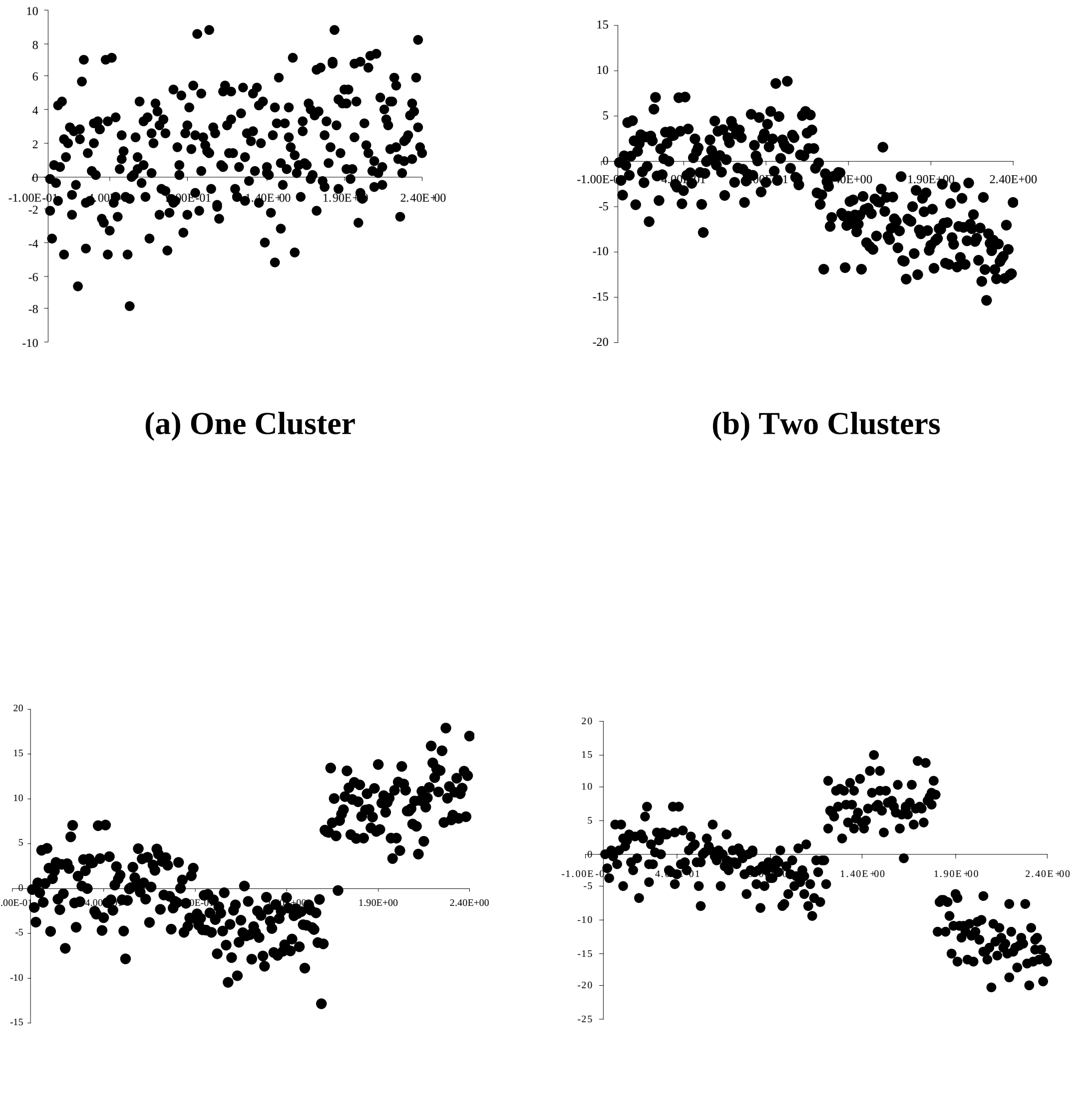
#### *5.2.1 Results with "big" slopes*

The results in Figures 3 to 5 show the probability, $P_{ij}$, that the BPO favour model $i$ over model $j$ based on the 1000 repetitions in each part of the experiment. Figure 3 summarizes the results for the three sets of experiments when the DGP runs from two clusters to four clusters with the "big" slope parameters defined in section 5.1.[4] Before discussing these figures, we use an illustrative table to explain how they are derived from the more detailed background results.[5]

[3] As noted in section 4, within each cluster an OLS model will be used to fit the data, so this assumption about the error term is quite standard but could readily be modified.

[4] We found that when the DGP involves only one cluster, the BPO identify the correct model 100% of the time, regardless of the level of the variance in the DGP.

[5] Tables for the complete set of detailed results can be downloaded from http://web/uvic.ca/~dgiles .



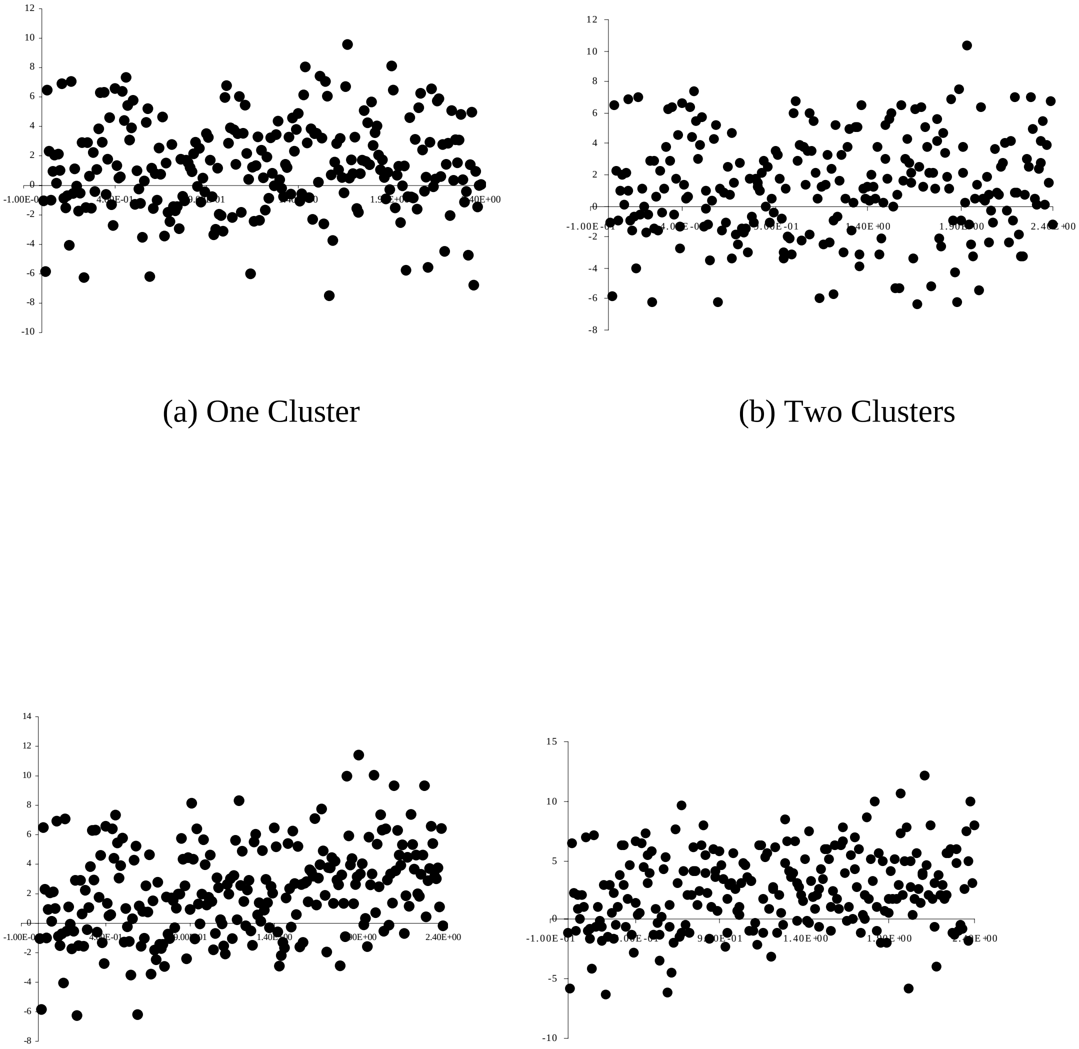
**(a) One Cluster**

**(b) Two Clusters**

(c) Three Clusters

(d) Four Clusters

**Figure 1. Data generating process – “big” slope case: *n* = 240.**



(a) One Cluster

(b) Two Clusters

(c) Three Clusters

(d) Four Clusters

**Figure 2. Data generating process – “small” slope case: *n* = 240.**

**Table 1. Probability of selecting M$_i$ over M$_j$: two-cluster DGP; *c* = 1, 2, 3, 4; *n* = 240**

| | $\sigma = 1$ | $\sigma = 2$ | $\sigma = 3$ | $\sigma = 4$ | $\sigma = 5$ | $\sigma = 6$ | $\sigma = 7$ | $\sigma = 8$ | $\sigma = 9$ | $\sigma = 10$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $P_{12}$ | 0 | 0 | 0 | 0.2 | 0.6 | 0.8 | 0.9 | 0.9 | 1 | 1 |
| $P_{13}$ | 0 | 0 | 0.7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $P_{14}$ | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $P_{23}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $P_{24}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $P_{34}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 1 relates to the case where the sample size is 240, with two clusters as the true DGP. We are choosing among four fuzzy models where the number of clusters runs from 1 to 4. Each row shows the value, P$_{ij}$, of selecting model *i* over model *j*, using the BPO. Across the columns, the standard deviation for the error term in the DGP increases from 1 to 10.

The values for the probabilities in this table are subject to simulation error, of course. The "accuracy" of these values can be measured according to the definition suggested by Kleijnen [20, p.479], and elaborated upon by Dìaz-Emparanza [6]. Specifically, with a confidence level of (1 – $\alpha$), the "accuracy" of a P$_{ij}$ in Table 1 is given by $A = Z_{(\alpha/2)}\sqrt{P_{ij}(1-P_{ij})/1000}$ , where $Z_{(\alpha/2)}$ is the (1 – $\alpha$/2) quantile for the standard normal density. So, with 90% confidence, the entries of 0.2 and 0.8 in the first row of Table 1 each have an accuracy of 0.0208; while the entry of 0.6 has an accuracy of 0.0255; etc.

When $\sigma$ = 1 in Table 1, using the notation $A \succ B$ to indicate "*A* is preferred to *B*", we have:

$M_1 \succ M_2$, with 0% probability $\Leftrightarrow M_2 \succ M_1$ with 100% probability.

$M_1 \succ M_3$, with 0% probability $\Leftrightarrow M_3 \succ M_1$ with 100% probability.

$M_1 \succ M_4$, with 0% probability $\Leftrightarrow M_4 \succ M_1$ with 100% probability.

$M_2 \succ M_3$, with 100% probability.

$M_2 \succ M_4$, with 100% probability.

$M_3 \succ M_4$, with 100% probability.

As a result, we can rank the four models in the order: $M_2 \succ M_3 \succ M_4 \succ M_1$. This is reflected in the value of the first point for the line corresponding to $n$ = 240 in Figure 3(a). The latter figure illustrates the values of just $P_{12}$ for different sample sizes and choices of $\sigma$. To conserve space, we have not presented corresponding figures for $P_{13}$, $P_{14}$, $P_{23}$, $P_{24}$ and $P_{34}$. This point and a more complete explanation of Figure 3(a) are discussed further below.

In the illustrative case being discussed here, the BPO analysis correctly selects the model with two fuzzy clusters with 100% probability. We find a similar ranking among the models when the standard deviation of the error term in the DGP is 2. When the standard deviation increases to 3, the two-cluster fuzzy model still is chosen correctly 70% of the time among the four models, though the ranking of the three-cluster and four-cluster fuzzy models is now reversed, and the three-cluster fuzzy model is chosen over the one-cluster fuzzy model with 30% probability. Nevertheless, the result that the two-cluster fuzzy model is the best among the four models still holds. As the standard deviation increases to 4, the probability that the two-cluster fuzzy model is chosen correctly over the one-cluster fuzzy model decreases to 80%, and overall the two-cluster fuzzy model is still the best. As the degree of fuzziness in the DGP increases further (with the standard deviation greater than or equal to 5), the one-cluster fuzzy model dominates the other three models, including the true two-cluster model. Perhaps not surprisingly, when the data are extremely diffuse, there is a tendency for the model selection procedure to eventually infer that there is just a single cluster of data.

One important result we find in this research is that the BPO tend to favour models based on few clusters. For example, in the above case where the true model is two-cluster fuzzy model, the three-cluster and four-cluster fuzzy models are never chosen over the true model, even when the standard deviation of the error term in the DGP increases to 10. As this standard deviation changes, the relative rankings of the three-cluster and four-cluster models also change, but the BPO analysis still always selects the two-cluster fuzzy model over the others. This is also true for the other cases where the true model involves either three or four clusters[6]. As a result, little would be added by providing detailed model selection results for cases involving fuzzy models that have more clusters than the true model. Hence, the BPO model selection results in Figures 3,

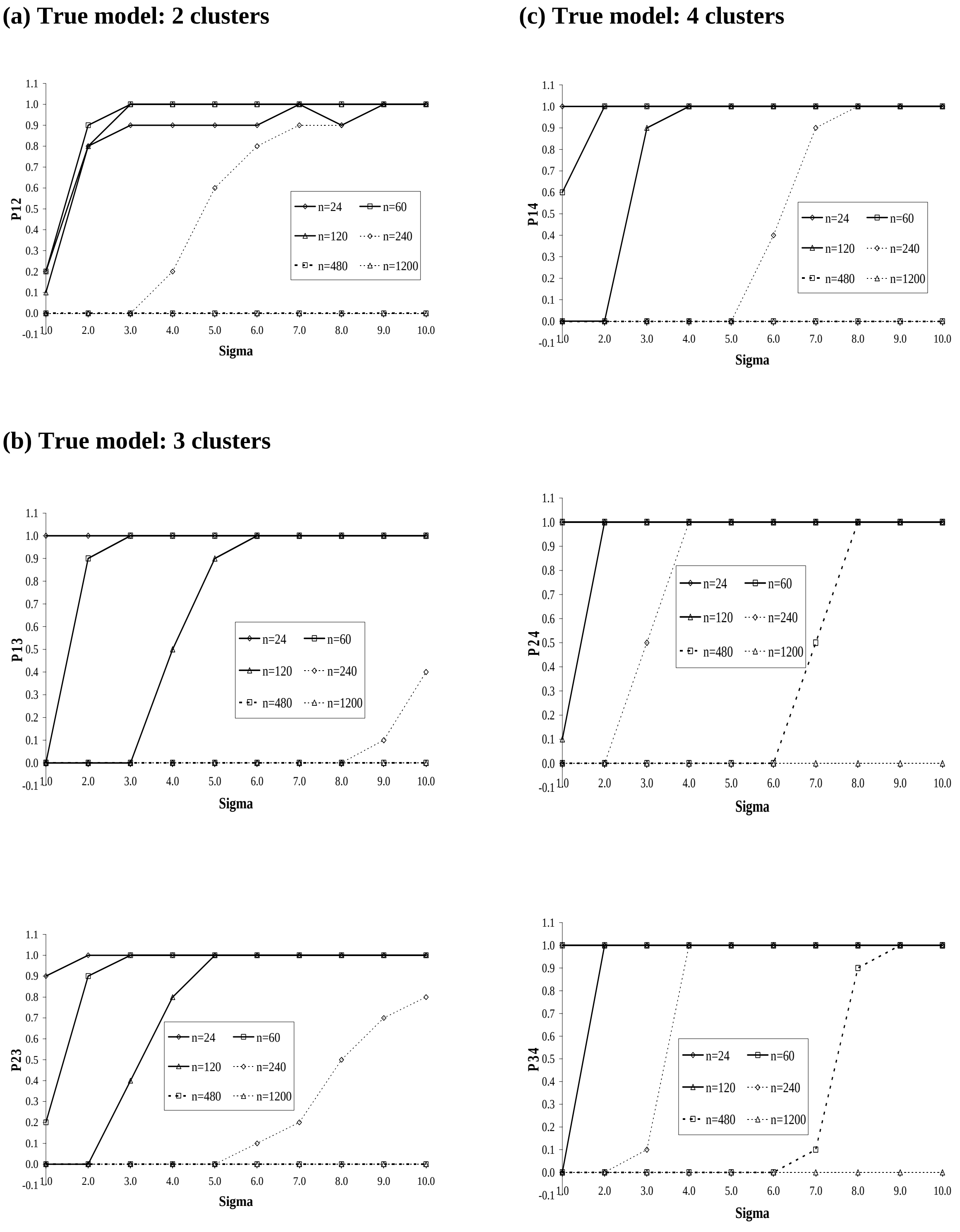
[6] When the true DGP has one cluster, the BPO analysis correctly chooses the one-cluster fuzzy model over the other three fuzzy models 100% of the time, even when the standard deviation increases to 10. For this reason, we did not provide detailed results for this case. The same is true for the “small slope” part of the Monte Carlo experiment where the true model has one cluster.

4 and 5 summarize $P_{ij}$ values between the true model and other fuzzy models for which the number of clusters is less than in the true model. For each sample size, these probabilities are plotted against the standard deviation of the error term in the DGP. Referring back to Table 1, we can see the line for $n$ = 240 in Figure 3 (a). When $\sigma \leq 3$, we see that this line coincides with the horizontal axis, indicating that (the true) $M_2$ is preferred to $M_1$, and from the above discussion it also dominates $M_3$ and $M_4$. When the standard deviation increases from 5 to 10, we see that (the false) $M_1$ is preferred to $M_2$ with increasing probability. Eventually, $M_1$ is chosen over the true model ($M_2$) 100% of the time, and also dominates models 3 and 4.

Figure 3 (a) shows the model selection results between one cluster and two clusters fuzzy model when the true DGP is two-cluster. The six lines represent six sample sizes between 24 and 1200. We have discussed the case where sample size is 240 above. If the sample size is smaller than 240, we can see that the correct model is selected if the standard deviation is smaller than 2, and the one-cluster model is chosen when the standard deviation is higher than 2. However, when the sample size is over 240, we see that the true model dominates the other three models even in the fuzziest case considered, with 100% probability (the lines for both cases are always on the horizontal axis). Similar results emerge in Figures 3 (b) and (c) where the true DGP has three and four clusters respectively ($P_{13}$, $P_{23}$, $P_{14}$, $P_{24}$ and $P_{34}$ for both cases are always on zeros).

In Figure 3, we see that the ability of the BPO analysis to select the true model increases as the sample size increases. The results are poor when $n$ = 24. However, for $n$ > 60 we begin to see some improvement, and when the sample size exceeds 240 we find that the BPO select the true model with very high frequency when the degree of the fuzziness is moderate. The value of the variance associated with the DGP also affects the results, and as the data become fuzzier the probability of the BPO picking the true model decreases.

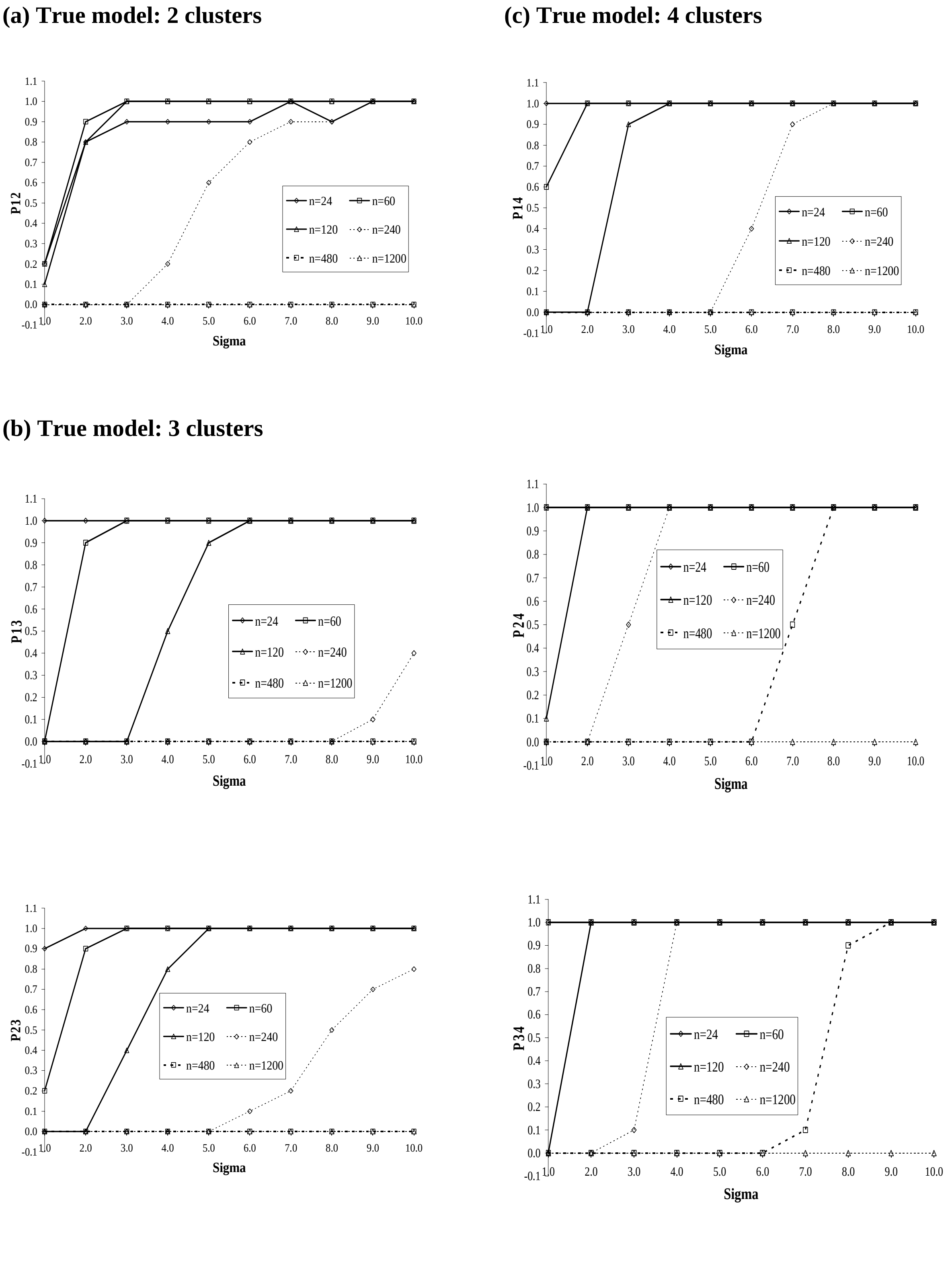
In practice, sometimes when we look at a plot of the data, we may get a sense that there should be more than one cluster and as a result, the one-cluster model can be eliminated from the outset. Overall, the one-cluster model suggests that there is no nonlinearity for the data. So, the next part of the Monte Carlo experiment relates to the case where the data have been generated with $c \geq 2$. In this case, we decrease the model space to three models – those based on two, three and four fuzzy clusters. The other aspects of the experimental design are unaltered.


(a) True model: 2 clusters

(c) True model: 4 clusters

(b) True model: 3 clusters

**Figure 3. Probabilities of selecting M*ᵢ* over M*ⱼ*: c = 1, 2, 3, 4; with "big slopes"**

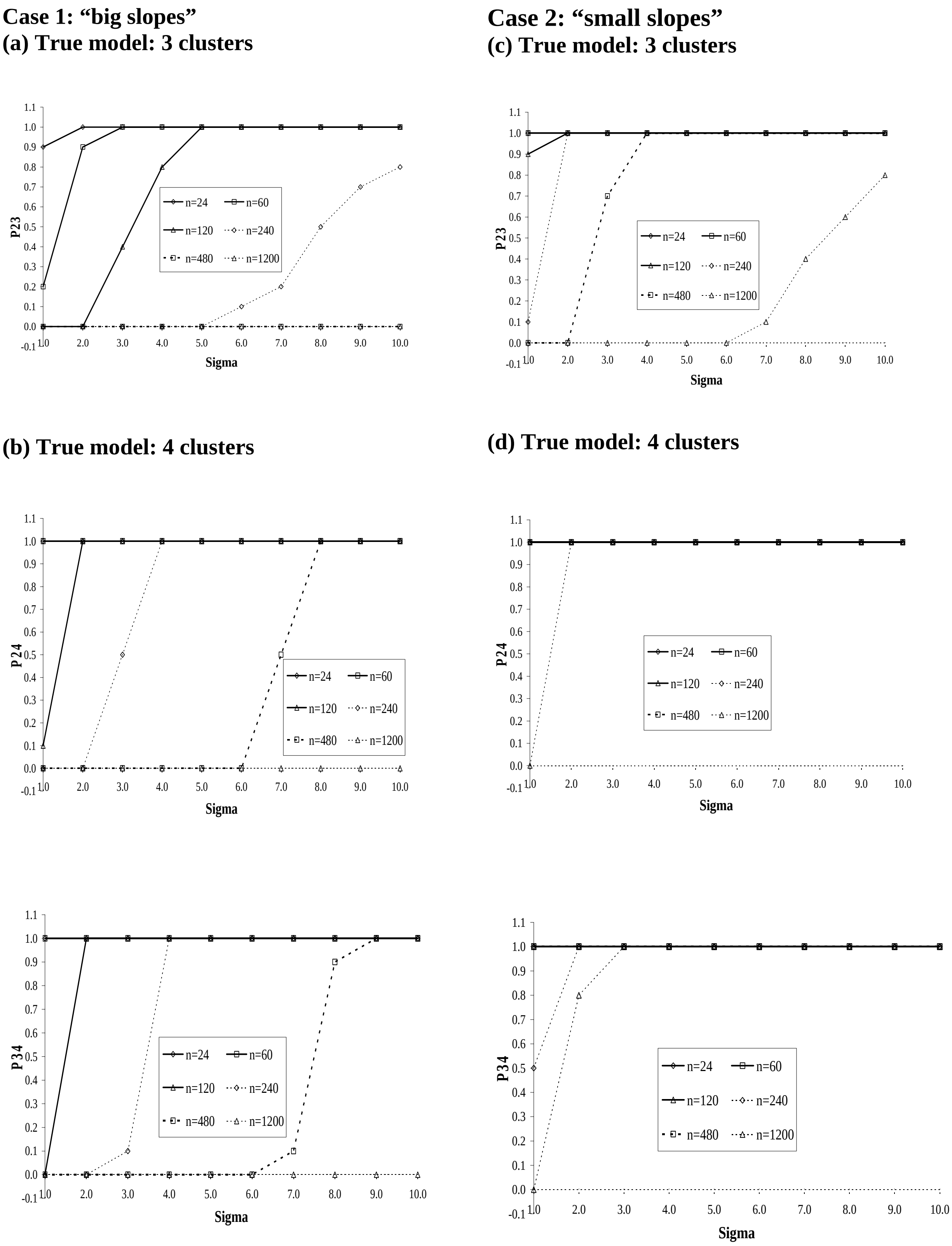


(a) True model: 2 clusters

(c) True model: 4 clusters

(b) True model: 3 clusters

**Figure 4. Probabilities of selecting M$_i$ over M$_j$: c = 1, 2, 3, 4; with “small slopes”**



Case 1: “big slopes”

(a) True model: 3 clusters

Case 2: “small slopes”

(c) True model: 3 clusters

(b) True model: 4 clusters

(d) True model: 4 clusters

Figure 5. Probabilities of selecting M$_i$ over M$_j$: c = 2, 3, 4

Figures 5 (a) and (b) provide the results for this part of the experiment. We see the BPO's ability to select the true model has been enhanced. All of the other features noted in Figure 3 also hold here: the BPO select the true (two cluster) model with 100% probability.[7] In addition, the sample size only needs to exceed 120 for the true model to be selected with higher probability than was the case in Figure 3. In Figures 5(a) and (b), we see that when the sample size exceeds 240, the BPO select the correct model all of the time. (The lines associated with $P_{23}$, $P_{24}$ and $P_{34}$ coincide with the horizontal axis for all values of the standard deviation of the error term in the DGP.)

### *5.2.2 Results with "small" slopes*

In this part of the Monte Carlo experiment we have used the "small slope" coefficient vector for the DGP, as defined in section 5.1. From the DGP, we can see from Figure 2 that the degree of fuzziness added to the sample comes from two sources: the variances of the error terms that we assign for the DGP, and the values of the slope parameters. As we see from Figure 2(a) to (d), with the "small slopes" the data become increasingly "cloudy" and it is very difficult to determine the correct number of clusters by a casual visual inspection. This part of the Monte Carlo experiment provides an even more stringent test of the BPO analysis. The standard deviation for the error term in the DGP again runs from 1 to 10, and the sample size runs from 24 to 1200.

Figure 4 provides the probabilities that the BPO favor the first model over the second model in the 1000 repetitions. All four fuzzy models are considered. As before, the DGP in Figure 4 runs from two cluster to four clusters[8]. In Figure 4, the results are similar to those in the "big slopes" case — the performance of the BPO increases as the sample size increases, and decreases as the value of the standard deviation for the error term in the DGP increases. Figures 5(c) and (d) give the results for the BPO analysis when we reduce the model space by dropping the model with one cluster. Again, the results are similar to those described in section 5.2.1.

The main difference between the results for the "big slopes" and "small slopes" parts of the Monte Carlo experiment is the understandable reduction in the ability of the BPO analysis to choose the true model in the latter case. Overall, as the number of observations increases, the BPO's ability to select the true model also increases. It is a "consistent" selection procedure.

[7] For this reason, in order to save the space we do not provide this result here since all of the lines ($P_{23}$, $P_{24}$) will be horizontal at a value of unity.

[8] Again, regardless of the level of the variation in the DGP with one cluster, the model picked by the BPO is *always* the correct one.

Some experimentation with the specification of the prior information about the models' parameters indicated that our results are quite robust in this respect.

## 6. Applications

In this section, we apply the BPO analysis/Bayesian fuzzy regression analysis to two real data sets that exhibit a range of characteristics. The first application involves cross-section data, while the second involves time-series data. In each case, the fuzzy c-means algorithm is applied with $m = 2$, and our Bayesian fuzzy regression results are compared with those obtained using nonparametric kernel regression. In each case the Bayesian analysis assigns equal prior probabilities across the model space, and uses relatively uninformative (but proper) natural conjugate prior p.d.f.'s for the associated parameters. The kernel regressions use Silverman's [24, p.45] approximately optimal bandwidth and a normal kernel, as implemented in the SHAZAM [22] package.

### *6.1 Journal subscriptions*

Our first application relates to library subscriptions for 180 economics journals for the year 2000, as reported by Bergstrom [1]. Our objective is to fit a regression to explain the number of subscriptions to a journal as a function of its price. Based on the four possible fuzzy models, the BPO analysis results in the posterior probabilities and root mean squared errors (RMSE's) shown in Table 2. These clearly favor the fuzzy regression model based on four fuzzy clusters, but it is interesting to note that the fuzzy model based on only three clusters also out-performs the kernel regression model in terms of RMSE. In Figure 6, we plot both the preferred fuzzy regression and a nonparametric kernel regression. The preferred fuzzy regression model outperforms the kernel regression model in the left tail and also the middle range of the data.

### *6.1 Phillips curve*

Our second example involves a very simple "Phillips curve", in which the annual rate of inflation is explained by the unemployment rate in Canada. The sample comprises 196 monthly observations for the period January 1993 to April 2009, and the seasonally adjusted data are constructed from the series v2062815 and v41690914 from the Statistics Canada CANSIM database [25]. Again, we see from Table 2 that the fuzzy regression model based on four fuzzy clusters is preferred over the other fuzzy models in terms of both the posterior probabilities and RMSE, and in terms of the latter measure it also dominates the kernel regression model. The data, and the fitted kernel and (preferred) fuzzy regression models are shown in Figure 7.

As the posterior probabilities in Table 2 for this data-set do not discriminate as sharply between the competing models as compared with the first example, we have also used Bayesian model averaging to combine the four fuzzy regression models, using the posterior probabilities as weights, as was discussed at the end of section 4. The results are also shown in Figure 7 and Table 2, and we can see that in this application the model averaging produces results that are slightly inferior to those for the four-cluster model, but still superior to the nonparametric kernel regression results.

### *6.3 Unemployment rate*

Finally, we estimate first-order autoregressive models (with drift) for the monthly unemployment rate data considered in section 6.2. In this case the prior for the slope coefficient is constrained to have a mean of zero and negligible prior density outside the stationary region.[9] The stationarity of the time-series is confirmed by applying the Dickey-Fuller [7] test for a unit root. As can be seen in Table 2, in this case the posterior probabilities over the model space favour a fuzzy regression model with one cluster (i.e., the full sample). Table 2 and Figure 8 also confirm that the fuzzy regression model out-performs the kernel regression model, especially between July 2005 and January 2009. Interestingly, in this application, although the BPO convincingly favour the one-cluster model over the four-cluster model, the latter fits the data marginally better than the former in terms of RMSE. Unsurprisingly, these two different criteria can lead to different outcomes.

## 7. Conclusions

In this study we have applied standard Bayesian model selection methods to the problem of choosing the number of fuzzy clusters to be used in the context of the recently developed fuzzy regression analysis. Using the Bayesian posterior odds to select the number of clusters to be used enhances the fuzzy regression analysis in an important way. The use of a Bayesian approach to both model selection and regression estimation illustrates two of the merits of Bayesian inference – namely its unity, and the flexibility with which prior information about the model space and the parameter space can both be incorporated into the analysis.

[9] This is achieved by choosing the parameters ($q_{ij}$ and $\bar{\bar{s}}_{ij}^2$) for the marginal prior for $\sigma_{ij}$ in (11) so as to take account of the fact that the mode of the inverted gamma p.d.f. occurs at $\sigma_{ij} = \bar{s}_{ij}[q_{ij}/(q_{ij}+1)]$, and that the 0.5 and 99.5 percentiles of the standard normal density occur at $\mp 2.57$.

**Table 2. Summary statistics for real data applications**

| | Journal Subscriptions | | Phillips Curve | | Unemployment Rate | |
|---|---|---|---|---|---|---|
| *c* | Posterior Probability | RMSE | Posterior Probability | RMSE | Posterior Probability | RMSE |
| 1 | 0.000 | 193.781 | 0.092 | 0.799 | 0.999 | 0.1800 |
| 2 | 0.000 | 198.298 | 0.004 | 0.781 | 0.001 | 0.1801 |
| 3 | 0.000 | 175.451 | 0.207 | 0.752 | 0.000 | 0.1806 |
| 4 | 1.000 | 174.326 | 0.697 | 0.700 | 0.000 | 0.1793 |
| Kernel regression | | 178.906 | | 0.747 | | 0.2483 |
| Bayesian model averaging | | | | 0.710 | | |

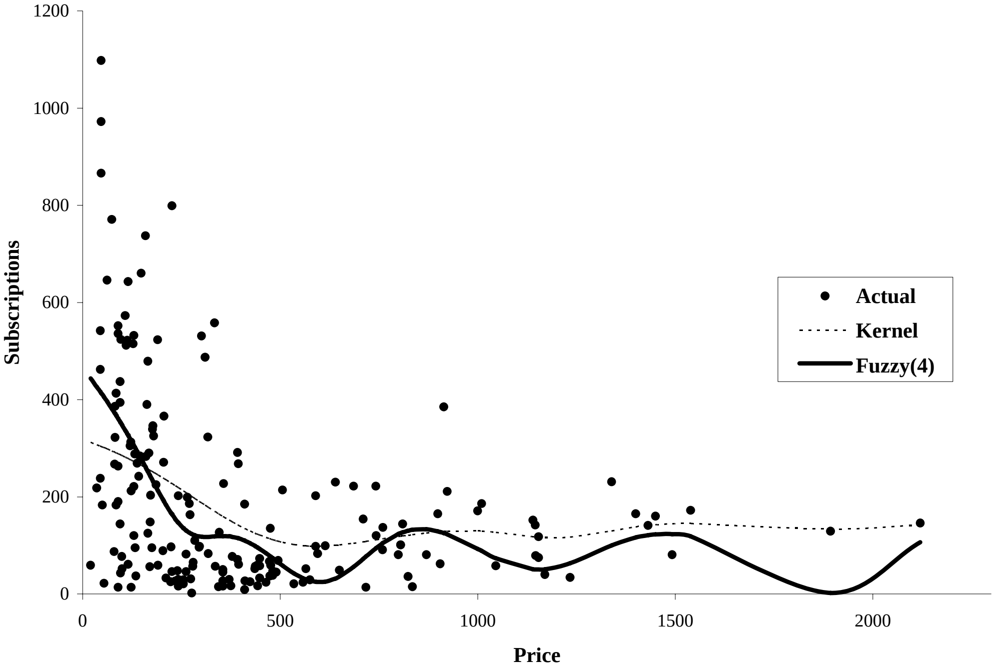


**Figure 6. Comparison of preferred Bayesian fuzzy regression models and non-parametric kernel regression model for journal subscriptions.**

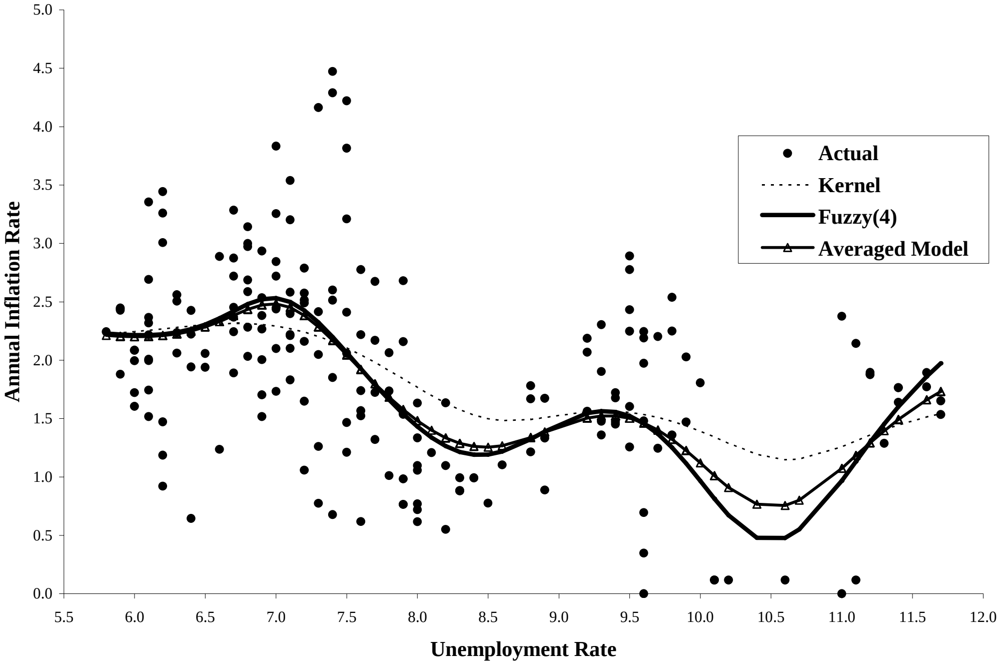


**Figure 7. Comparison of Bayesian fuzzy regression models and non-parametric kernel regression model for Phillips curve.**

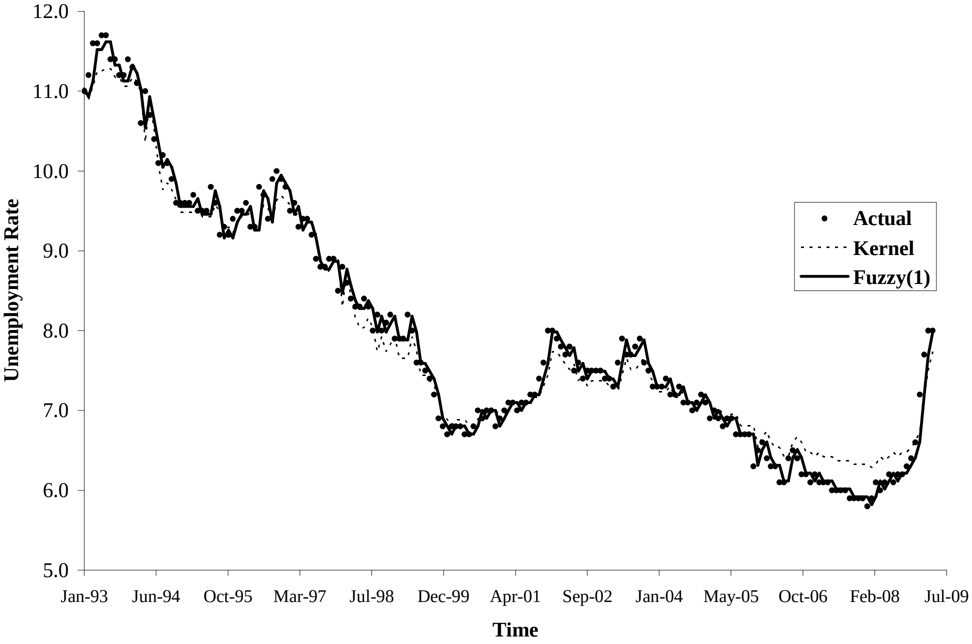


**Figure 8. Comparison of Bayesian fuzzy regression models and non-parametric kernel regression model for unemployment.**

Our Monte Carlo experiments show how powerful this approach can be, especially with moderate to large samples. Even when the sample data visually appear to be generated from one regime, the Bayesian model selection analysis is very successful in determining the true number of underlying clusters. Three brief applications with economic data are also extremely encouraging. Compared with standard non-parametric kernel regression, the Bayesian fuzzy regression captures the nonlinearity of the data extremely well. The sample sizes in these two applications are less than 200, illustrating that in practice Bayesian fuzzy regression can perform very satisfactorily even when the sample size is modest.

Of course, there are some limitations to the illustrative applications of this Bayesian fuzzy regression methodology, as is evident in the above results. Although our theoretical results are established for the general multivariate case, the applications considered in this study focus only on the univariate case. The practical difficulties associated with the construction of the prior pd.f.'s for the models' parameters in more complex models should not be under-stated. Nonetheless, the evidence provided in this research lends credibility to Bayesian fuzzy regression analysis, and especially to the use of Bayesian posterior odds to select the number of fuzzy clusters that are to be used. Other recent studies (e.g., [13]) have shown that the fuzzy regression methodology performs well when a frequentist approach to inference is taken in the multivariate case. In the latter case the Bayesian analysis that we have introduced in this paper becomes only marginally more burdensome to implement, provided that natural conjugate priors are used, and there is every reason to suppose that it will have an advantage over the frequentist approach with regard to the crucial problem of determining the number of fuzzy clusters. Work in progress explores this issue.

## Acknowledgments

We are grateful to an anonymous referee for constructive comments and suggestions on an earlier version of this paper. The first author also thanks King's College at the University of Western Ontario for financial support for this research.

## References

[1] T. C. Bergstrom, Free labor for costly journals? Journal of Economic Perspectives, 15 (2001) 183-198.

[2] J. C. Bezbek, Pattern Recognition With Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[3] G. E. P. Box , G. C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading MA, 1973.

[4] J. Chen, D. E. A. Giles, Gender convergence in crime: evidence from Canadian adult offense charge data, Journal of Criminal Justice, 32 (2004) 593-606.

[5] M. A. Chmielewski, Elliptically symmetric distributions: a review and bibliography, International Statistical Review, 49 (1981) 67-74.

[6] I. Dìaz-Emparanza, Is a small Monte Carlo analysis a good idea? Checking the size, power and consistency of a simulation-based test, Statistical Papers, 43 (2002) 567-577.

[7] D. A. Dickey, W. A. Fuller, Distribution of the estimators for autoregressive time series with a unit root, Journal of the American Statistical Association, 74 (1979) 427-431.

[8] R. Draeseke, D. E. A. Giles, Modelling the New Zealand underground economy using fuzzy logic techniques, Mathematics and Computers in Simulation, 59 (2002) 115-123.

[9] J. C. Dunn, Well separated clusters and optimal fuzzy partitions, Journal of Cybernetics 4 (1974) 95-104.

[10] J. C. Dunn, Indices of partition fuzziness and the detection of clusters in large data sets, in: M. Gupta, G. Seridis (eds.), Fuzzy Automata and Decision Processes, Elsevier, New York, 1977.

[11] H. Feng, Forecasting comparison between two nonlinear models: fuzzy regression *vs*. SETAR, mimeo., 2007.

[12] D. E. A. Giles, Output convergence and international trade: time-series and fuzzy clustering evidence for New Zealand and her trading partners, 1950-1992, Journal of International Trade and Economic Development, 14 (2005) 93-114.

[13] D. E. A. Giles, R. Draeseke, Econometric modelling using fuzzy pattern recognition via the fuzzy c-means algorithm, in: D. E. A. Giles (ed.), Computer Aided Econometrics, Marcel Dekker, New York, 2003, 407-450.

[14] D. E. A. Giles, H. Feng, Output and well-being in industrialized nations in the second half of the 20th century: testing for convergence using fuzzy clustering analysis, Structural Change and Economic Dynamics, 16 (2005) 285-308.

[15] D. E. A. Giles, C. A. Mosk, A long-run environmental Kuznets curve for enteric $CH_4$ emissions in New Zealand: a fuzzy regression analysis, Econometrics Working Paper EWP0307, Department of Economics, University of Victoria, 2003.

[16] D. E. A. Giles, C. Stroomer, Identifying the cycle of a macroeconomic time-series using fuzzy filtering, Econometrics Working Paper EWP0406, Department of Economics, University of Victoria, 2004.

[17] D. E. A. Giles, C. Stroomer, Does trade openness affect the speed of output convergence? Some empirical evidence, Empirical Economics, 31 (2006) 883-903.

[18] T. Kariya, M. L. Eaton, Robust tests for spherical symmetry, Annals of Statistics, 5 (1977) 206-215.

[19] M. L. King, Robust tests for spherical symmetry and their applications to least squares regression, Annals of Statistics 8 (1980), 1265-1271.

[20] J. P. C. Kleijnen, Statistical Tools for Simulation Practitioners, Marcel Dekker, New York, 1987.

[21] E. Ruspini, Numerical methods for fuzzy clustering, Information Science 2 (1970) 319-350.

[22] SHAZAM, SHAZAM Econometrics Package, User’s Guide, Version 9, Northwest Econometrics, Vancouver, B.C., 2001.

[23] D. Shepherd, F. K. C. Shi, Economic modelling with fuzzy logic, paper presented at the CEFES ’98 Conference, Cambridge, U.K., 1998.

[24] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[25] Statistics Canada, CANSIM Multidimensional, http://dc2.chass.utoronto.ca/cansimdim/ .

[26] L. A. Zadeh,  Fuzzy sets, Information and Control 8 (1965) 338-353.

[27] A. Zellner, An Introduction to Bayesian Inference in Econometrics, Wiley, New York, 1971.