한국통계학회논문집
2012, 19권, 6호, 809–817

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.809

Negative Binomial Varying Coefficient Partially Linear Models

Young-Ju Kim [1,a]

[a]Department of Statistics, Kangwon National University

Abstract

We propose a semiparametric inference for a generalized varying coefficient partially linear model(VCPLM) for negative binomial data. The VCPLM is useful to model real data in that varying coefficients are a special type of interaction between explanatory variables and partially linear models fit both parametric and nonparametric terms. The negative binomial distribution often arise in modelling count data which usually are overdispersed. The varying coefficient function estimators and regression parameters in generalized VCPLM are obtained by formulating a penalized likelihood through smoothing splines for negative binomial data when the shape parameter is known. The performance of the proposed method is then evaluated by simulations.

Keywords: Negative binomial, penalized likelihood, semiparametric, smoothing parameter, smoothing spline, varying coefficients.

1. Introduction

A classical linear regression approach has assumed that the explanatory variables are linearly associated with the response and the fitted model has been interpreted by regression parameter estimates. However, the ordinary linear regression model was often found to be too simple to fit to interpret the various types of data. An alternative method is to use a nonparametric regression approach that can to relieve the parametric restriction by allowing the estimator to be infinite-dimensional. Recently, a semiparametric approach (which involves both linear terms and nonparametric function terms) has been studied in a constructive way to avoid the 'curse of dimensionality'. Such structured models include generalized additive models and varying coefficient models (Hastie and Tibshirani, 1993).

The generalized varying coefficient partially linear model(VCPLM) is one of the semiparametric models in which some of coefficients are varying and others are constant. The VCPLM includes many simpler models; classical nonparametric models, partially linear models, generalized additive models, and varying coefficient models. The details can be found at Hastie and Tibshirani (1993), Fan and Zhang (1999), Fan *et al.* (2003), Fan and Huang (2005), Senturk and Muller (2008) and references therein.

Previous studies on semiparametric models have focused mostly on Gaussian data (Zhang *et al.*, 2002; Xia *et al.*, 2004; Fan and Huang 2005; Ahmad *et al.*, 2010) and binomial or Poisson data (Lu, 2008). The negative binomial distribution is widely used for modelling the discrete data where it is believed that the variance of the response variable is larger than its mean; subsequently, it is often considered as an

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2010-0021695).

[1] Associate Professor, Department of Statistics Kangwon National University, Chuncheon 200-090, Republic of Korea. Email: ykim7stat@kangwon.ac.kr

overdispersed Poisson distribution. In this paper, we propose an inference for a generalized VCPLM for negative binomial data.

The paper is organized as follows. Section 2 describes the penalized likelihood method for negative binomial data in generalized VCPLM. Section 3 presents the computation method and its algorithm. The smoothing parameter selection method is described in Section 4. Section 5 reports the numerical results of simulated examples and conclusions are drawn in Section 6.

2. Penalized Likelihood

The penalized likelihood estimates f in the ordinary nonparametric estimation is a minimizer of a penalized likelihood functional,

$$\frac{1}{n}\sum_{i=1}^{n} l(\eta(u_i)|y_i) + \lambda J(\eta), \tag{2.1}$$

where the first term is the minus log likelihood, $J(\eta)$ is a roughness penalty functional. The smoothing parameter λ controls the trade-off between the lack of fit and the roughness of $\eta(u)$ and thus plays an important role to determine the performance of the estimator. For y be a response from negative binomial distribution describing the number of failures before the α^{th} success in Bernoulli trials with a success probability p, $\mu = E(y) = \alpha(1-p)/p$ and $\mathrm{Var}(y) = \alpha(1-p)/p^2$. The minus log likelihood is $l(p_i|y_i) = \log\Gamma(\alpha) - \log\Gamma(\alpha+y_i) - \alpha\log(p_i) - y_i\log(1-p_i) + C(y_i)$, where α is the known shape parameter and $C(y)$ is a term involving only y. For fitting the negative binomial model for y in a regression setting there are several link functions available and a common approach to negative binomial regression model is to take log link $\log(\mu) = f$ (Thurston *et al.*, 2000). In this paper, we take a logit link $\log(p/(1-p)) = \eta$ as in logistic regression and focus on estimating η. For given α, $l(\eta(u_i)|y_i) = (\alpha+y_i)\log(1+\exp(\eta(u_i))) - \alpha\eta(u_i) + D(\alpha, y_i)$, where $D(\alpha, y)$ is a term involving both α and y.

The nonparametric estimate of η is obtained by minimizing the penalized likelihood functional in (2.1) in a space $\mathcal{H} \subseteq \{f : J(f) < \infty\}$ of functions on the domain $\mathcal{T}$. In fact, the minimizer of (2.1) is in infinite dimensional space $\mathcal{H} \subseteq \{f : J(f) < \infty\}$. Specifically, the minimizer of (2.1) lies in $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where $\mathcal{N}_J = \{f : J(f) = 0\}$ be the null space of $J(\eta)$ and the space $\mathcal{H}_J$ is an reproducing kernel Hilbert space(RKHS) with $J(\eta)$ as the square norm. Note that a space $\mathcal{H}$ in which the evaluation functional $[x]f = f(x)$ is continuous is called an RKHS possessing a reproducing kernel(RK) $R(\cdot,\cdot)$, a non-negative definite function satisfying $\langle R(x,\cdot), f(\cdot)\rangle = f(x)$, $\forall\eta \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$. Letting $J(f) = \int_0^1 \ddot{f}^2 dx$ on $\mathcal{T} = [0,1]$, one gets the popular (smoothing) cubic splines with $\mathcal{N}_J = \mathrm{span}\{1, k_1(t)\}$, where $k_1(t) = t - 0.5$. In $\mathcal{H}_J = \{f : \int_0^1 f\,dt = \int_0^1 \dot{f}\,dt = 0, J(f) < \infty\}$ with $J(f)$ as the square norm, one has the RK $R_J(t_1,t_2) = k_2(t_1)k_2(t_2) - k_4(t_1-t_2)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials (see Gu, 2002).

Estimating η in infinite dimensional space $\mathcal{H}$ is challenging in practice. As remedy for this problem, a data-adaptive lower-dimensional approximation can be used in penalized likelihood methods. Gu and Kim (2002) showed for regression that the minimizer of the penalized likelihood functional in $\mathcal{H}$ shared the same convergence rates as one in the lower dimensional function space $\mathcal{H}_q = \mathcal{N}_J \oplus \mathrm{span}\{R_J(w_j,\cdot), j = 1,\ldots,q\}$, where $\{w_j\}$ are random subsets of $\{u_i, i = 1,\ldots,N\}$, as long as $q \asymp n^{2/(vr+1)+\epsilon}$, for some $v \in [1,2]$, $r > 1$, $\epsilon > 0$ is arbitrary. Here, v represents how smooth the true function is and $v = 2$ is used under the assumption that the true function is smooth enough. For the cubic spline, $r = 4$ is used.

The generalized VCPLM is

$$\text{logit}(p) = \eta(u)^T x + \beta^T z, \tag{2.2}$$

where $p = p(u, x, z)$, $(x^T, z^T) \in R^{p_1} \times R^{p_2}$, $\eta(u) = (\eta_1(u), \ldots, \eta_{p_1}(u))^T$ is a p_1-dimensional vector of smooth functions of covariates u, and $\beta = (\beta_1, \ldots, \beta_{p_2})^T$ is a vector of unknown parameters. The varying coefficient function η's can be interpreted the same way as the parameters for interactions in a classical multiple regression. In this paper, we are interested in estimating both η and β in (2.2).

3. Computation

In this study, we propose an estimating algorithm for both varying coefficient functions and regression parameters iteratively. Assume that the shape parameter is known. The full algorithm consists of two alternating algorithms; an algorithm for estimation of η and β and a backfitting-type algorithm for estimation of multiple η's. The alternating algorithm for η and β is as follows. (i) First, get a starting value for β. (ii) Given β, compute $\hat{\eta}$ by using weighted penalized least squares in the lower-dimensional approximating space $\mathcal{H}_q$. (iii) Using $\hat{\eta}$, get the estimates $\hat{\beta} = \text{argmax}_\beta l(\beta|\hat{\eta})$, where $l(\beta|\hat{\eta})$ is log likelihood for β given $\hat{\eta}$. Repeat steps (ii) and (iii) until convergence.

More specifically, penalized likelihood for VCPLM is

$$\frac{1}{n}\sum_{i=1}^{n} l\left(\eta(u_i)^T x_i + \beta^T z_i\right) + \lambda J(\eta). \tag{3.1}$$

Assume that β is given. The estimating algorithm of η consists of two nested loops, which the inner loop computes the minimizer of the penalized likelihood for fixed smoothing parameters and the outer loop computes the optimal smoothing parameters. For fixed smoothing parameters, (3.1) is strictly convex and thus the estimator of η can be computed by Newton iteration. Since the profile likelihood $l(\eta|y) \propto \alpha(\eta^T x + \beta^T z) - (\alpha + y)\log(1 + e^{\eta^T x + \beta^T z})$, $\tilde{u}_i = \partial l/\partial\eta|_{\tilde{\eta}(u_i)} = (\alpha + y_i)p_i x_i - \alpha x_i$ and $\tilde{w}_i = \partial^2/\partial\eta^2|_{\tilde{\eta}(u_i)} = (\alpha + y_i)p_i(1 - p_i)x_i^2$, where $p_i = \exp(\hat{\eta}^T x + \beta^T z)/(1 + \exp(\hat{\eta}^T x + \beta^T z))$. The quadratic approximation of $l(\eta|y)$ at $\tilde{\eta}$ yields the weighted penalized least squares

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i \left(\tilde{y}_i - \eta(u_i)\right)^2 + \lambda J(\eta), \tag{3.2}$$

where $\tilde{y}_i = \tilde{\eta}(u_i) - \tilde{u}_i/\tilde{w}_i$.

Incorporating the expression of varying coefficients η in $\mathcal{H}_q$ and substituting into (3.2), one calculate the minimizer of (3.1) with respect to η through the minimization of the weighted penalized least square functional. Then the resulting normal equation for the solution can be solved by a Cholesky decomposition followed by forward and backward substitutions. Note that for the dimension of $\mathcal{H}_q$, we take Kim and Gu (2004)'s suggestion; set $q = 10n^{2/9}$.

If there are multiple varying coefficient functions η's to be estimated, a backfitting-type algorithm is used as follows. (i) Let $\hat{\eta}_j$ be some starting values, for $j = 1, \ldots, p_1$. (ii) For j, calculate η_j by using profile penalized likelihood functional given η_l, $l \neq j$, $j = 1, \ldots, p_1$. (iii) Repeat the step (ii) until convergence.

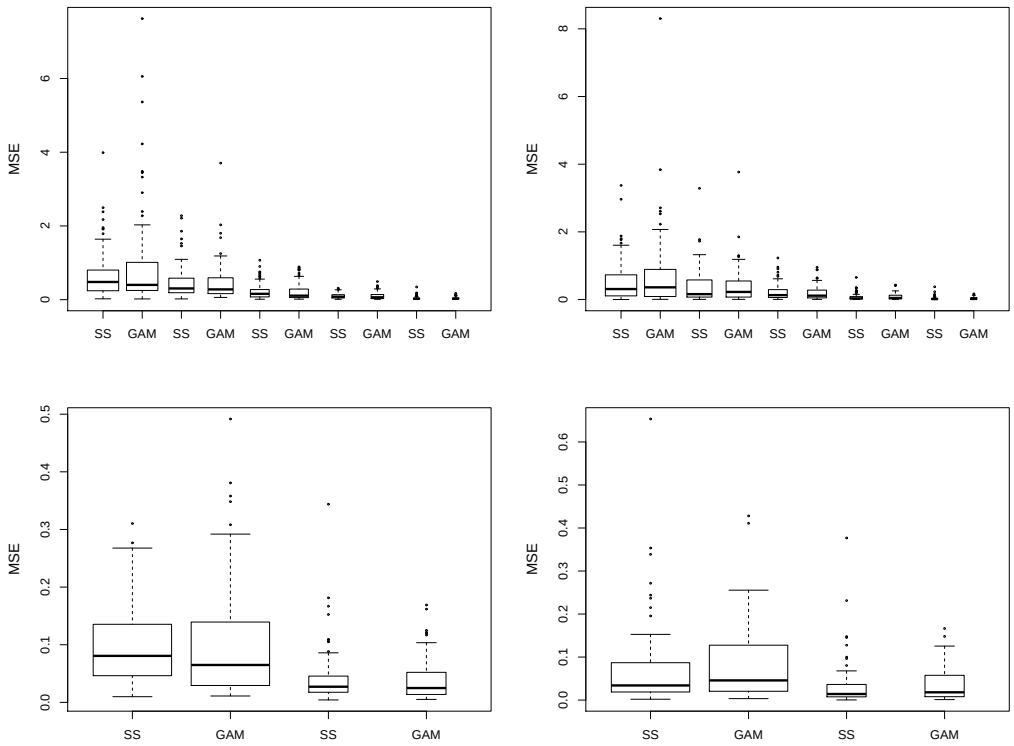

Figure 1: *The MSE losses for varying coefficients estimates under VCM. First row: each pairs of boxplots of two methods is for* $n = 30, 50, 100, 200,$ *and* 500 *from left to right. Second row: boxplots for* $n = 200$ *(fat) and* $n = 500$ *(thin) with better resolution. Left column: plots for* η_1*. Right column: plots for* η_2*. SS stands for smoothing splines (our method) and GAM stands for* gam *in* mgcv *R package.*

4. Smoothing Parameter Selection

The performance of the estimators η_λ of η is determined by the selected smoothing parameters in smoothing splines. Several methods to select the optimal smoothing parameters have been used in the literature as follows: the generalized cross-validation score (Wahba, 1985), the indirect cross-validation score of (Gu, 2002)(also called performance-oriented iteration), and the direct cross-validation score of Gu and Xiang (2001) and Xiang and Wahba (1996). We adapt the score of Gu and Xiang (2001), which is a modified version of the score of Xiang and Wahba (1996). Specifically, the alternative generalized approximate cross-validation(AGACV) in penalized likelihood regression settings can be derived based on the Kullback-Leibler distance between the true function η and the minimizer η_λ of the penalized likelihood functional as follows:

$$V(\lambda) = \frac{1}{n}\sum_{i=1}^{n} l\left(\eta(u_i)_\lambda^T x_i + \beta^T z_i\right) + \frac{\mathrm{tr}\left(A_w W^{-1}\right)}{n - \mathrm{tr}\, A_w}\frac{1}{n}\sum_{i=1}^{n} y_i p_i \tilde{u}_i, \tag{4.1}$$

here A_w is the smoothing matrix.

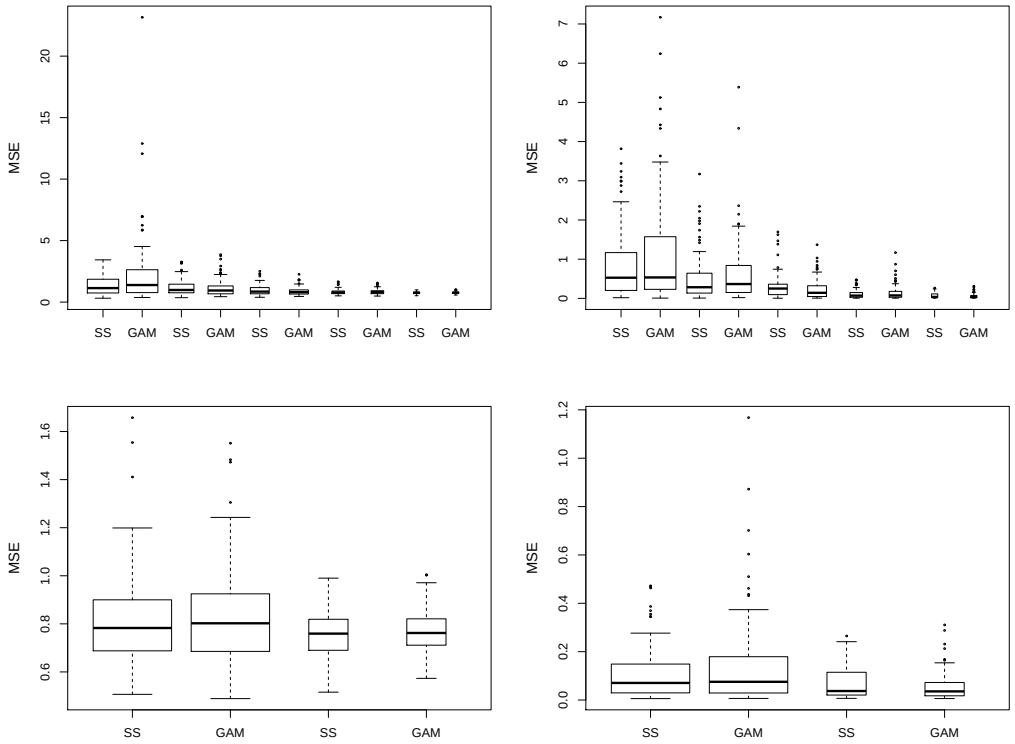

Figure 2: *The MSE losses for varying coefficients estimates under VCPLM. First row: each pairs of boxplots of two methods is for* $n = 30$*,* 50*,* 100*,* 200*, and* 500 *from left to right. Second row: boxplots for* $n = 200$ *(fat) and* $n = 500$ *(thin) with better resolution. Left column: plots for* η_1*. Right column: plots for* η_2*. SS stands for smoothing splines (our method) and GAM stands for* gam *in* mgcv *R package.*

5. Simulations

In this section, we conducted two sets of simulations to evaluate the performances of the proposed methods for negative binomial data. Assume that the shape parameter is known. Two sets of simulations were conducted; for the model with varying coefficients only (varying coefficient models; VCM) and the varying coefficients model with partially linear terms(VCPLM). For VCM, the data are generated from the negative binomial with $\eta_1(u_{\langle 1\rangle}) = \log(2\sin(2\pi u_{\langle 1\rangle}) + u_{\langle 1\rangle}^2 + 2.1)$, $\eta_2(u_{\langle 2\rangle}) = -(u_{\langle 2\rangle} - 0.5)^2$, and $\beta^T = (0, \ldots, 0)$ in (2.2), and the shape parameter $\alpha = 2$. The covariates $u_{\langle 1\rangle}$, $u_{\langle 2\rangle}$ are generated from the uniform distribution on $[0, 1]$ and $x_{\langle 1\rangle}$, $x_{\langle 2\rangle}$ are generated from the normal distribution with mean 1 and standard deviation 0.2. For sample size $n = 30, 50, 100, 200$ and 500, 100 replicates were generated and cubic smoothing splines for varying coefficient functions. The performance of the proposed method was evaluated by mean squared error(MSE) for varying coefficient function estimates. Note that the computations of varying coefficient functions for $n = 30$ and $n = 50$ were conducted in $\mathcal{H}_n$ as the sample sizes were relatively small and the effect of the lower-dimensional approximation may be minimal.

The performances of the proposed method were compared with the generalized additive model(GAM) of Wood (2008) and Wood (2011) for negative binomial estimation. Wood (2008) and Wood (2011)'s meth-

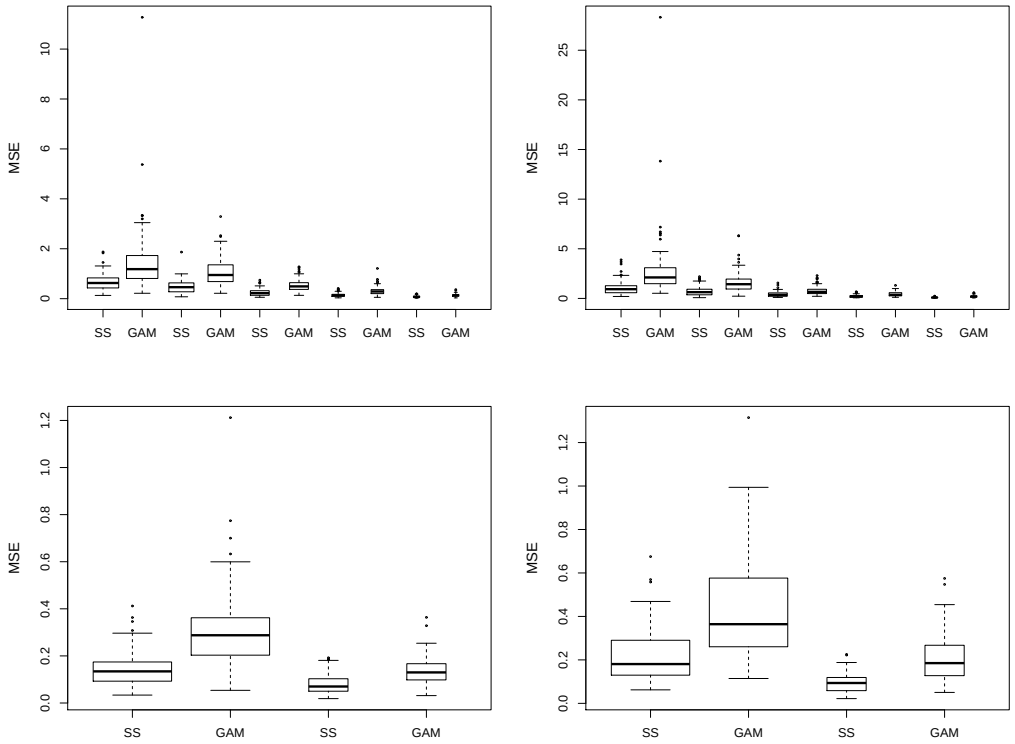

Figure 3: *The MSE losses for the estimates of μ. First row: each pairs of boxplots of two methods is for n = 30, 50, 100, 200, and 500 from left to right. Second row: boxplots for n = 200 (fat) and n = 500 (thin) with better resolution. Left column: plots under VCM. Right column: plots under VCPLM. SS stands for smoothing splines (our method) and GAM stands for* gam *in* mgcv *R package.*

ods were implemented via gam function in mgcv R package. For comparison, we used his REML score for smoothing parameter selection which Wood (2011) derived for restricted maximum likelihood(REML) based on direct optimization of GCV for generalized linear model. For negative binomial data, Wood (2011) used log link. Note that the gam function in the package only computed the estimates of $\eta_i(u_i)x_i$ and we had to divide the estimates by x_i to get the estimates of varying coefficient functions $\eta_i(u_i)$ for comparison.

For VCPLM, the data are generated from the negative binomial with the same test functions $\eta_1(u_{\langle 1\rangle})$, $\eta_2(u_{\langle 2\rangle})$, and $\beta_1 = 1, \beta_2 = -1$ in (2.2) and $z_{\langle 1\rangle}$, $z_{\langle 2\rangle}$ are generated from the normal distribution with with mean 1 and standard deviation 0.2. The performances of the proposed method for the VCM and for the VCPLM respectively and comparisons with GAM of Wood (2011) were summarized in Figure 1~Figure 5. Figure 1 and Figure 2 showed boxplots of MSE to show the performances of the varying coefficient function estimators for each sample size n under VCM and VCPLM respectively. The box width in boxplots in the first row in both figures gradually decreased as n increased. Our method estimated both η_1 and η_2 slightly better than GAM in terms of smaller ranges even though there were some larger outliers in our method. Figure 3 showed boxplots of MSEs of the mean functions. In comparing the mean function estimates, the

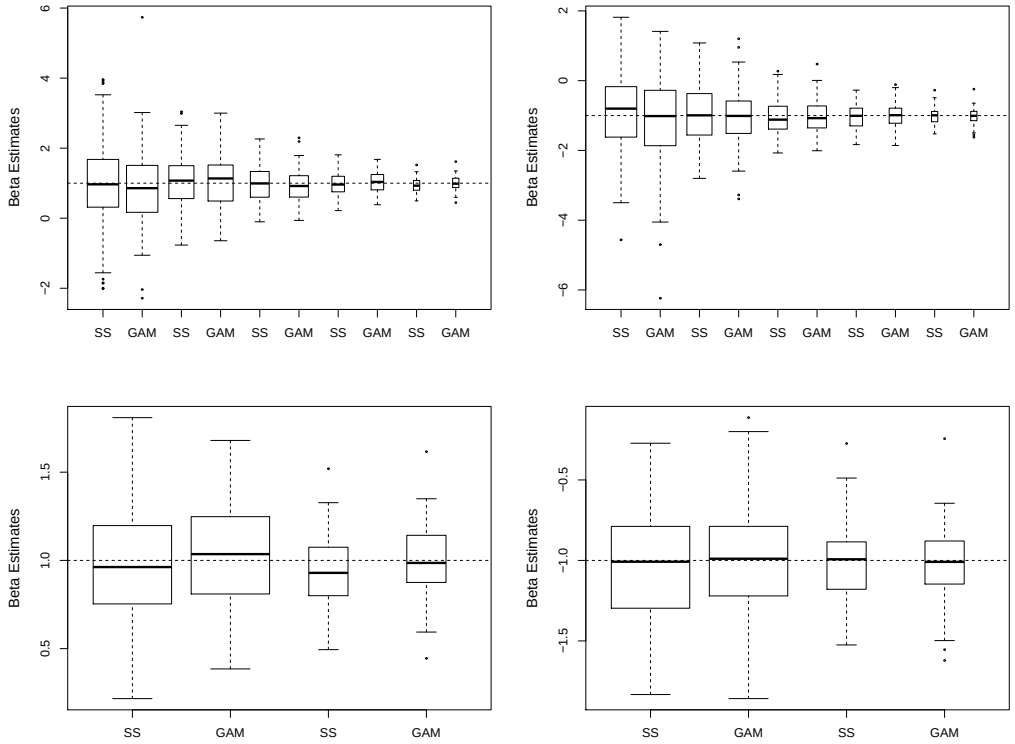

Figure 4: *Performance of regression parameter β estimates. First row: each pairs of boxplots of two methods is for n = 30, 50, 100, 200, and 500 from left to right. Second row: boxplots for n = 200 (fat) and n = 500 (thin) with better resolution. Left column: plots for β_1. Right column: plots for β_2. SS stands for smoothing splines (our method) and GAM stands for* gam *in* mgcv *R package.*

MSE differences of two methods were larger than the varying coefficient function estimates because of the different ranges of the target functions being estimated. For small sample sizes, say $n = 30$ and $n = 50$, our method was no better than GAM. However, the performance of our method became better as the sample size increased. It also confirmed that our method performed equivalently good as GAM for large sample sizes. Figure 4 showed boxplots of regression parameter β estimates and two methods showed similar performances. Figure 5 showed the varying coefficient function estimates obtained from a sample data with $n = 200$ using our method and GAM of Wood (2011) under the VCM and the VCPLM respectively. It also confirmed that our method performed better to estimate the varying coefficient functions in both VCM and VCPLM.

6. Conclusions

In this paper, we proposed a semiparametric estimation method for varying coefficient partially linear models for negative binomial data through smoothing spline approach. The performance of the proposed method was then compared with a well-known existing semiparametric method.

When the shape parameter is unknown, the distribution no longer belongs to exponential family and a

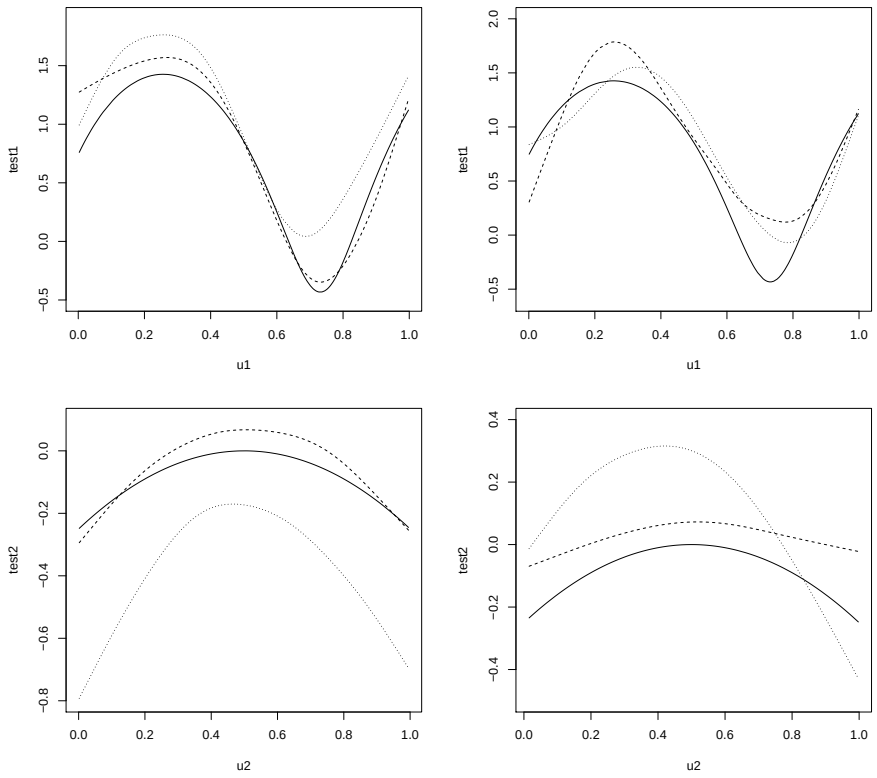

Figure 5: *Each varying coefficient function estimates from a sample with $n = 200$. The solid line is the true coefficient functions, the dashed line is our estimates and the dotted line is GAM estimates for η_1 (first row) and η_2 (second row) under VCM (left column) and VCPLM (right column).*

different approach on the estimation of varying coefficient functions and regression parameters along with the shape parameter estimation is needed. The algorithmic developments in this case are under way.

References

Ahmad, I., Leelahanon, S. and Li, Q. (2010). Efficient estimation of a semiparametric partially linear varying coefficient model, *The Annals of Statistics*, **33**, 258–283.

Fan, J. and Huang, T. (2005). Profile likelihood inference on semiparametric varying-coefficient partially linear models, *Bernoulli*, **11**, 1031–1057.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models, *Journal of the Royal Statistical Society Series B*, **65**, 57–80.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models, *The Annals of Statistics*, **27**, 1491–1518.

Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer-Verlag.

Gu, C. and Kim, Y.-J. (2002). Penalized likelihood regression: General formulation and efficient approximation, *Canadian Journal of Statistics*, **30**, 619–628.

Gu, C. and Xiang, D. (2001). Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited, *Journal of Computational and Graphical Statistics*, **10**, 581–591.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models, *Journal of the Royal Statistical Society Series B*, **55**, 757–796.

Kim, Y.-J. and Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation, *Journal of the Royal Statistical Society Series B*, **66**, 337–356.

Lu, Y. (2008). Generalized partially linear varying-coefficient models, *Journal of Statistical Planning and Inference*, **138**, 901–914.

Senturk, D. and Muller, H.-G. (2008). Generalized varying coefficient models for longitudinal data, *Biometricka*, **95**, 653–666.

Thurston, S. W., Wand, M. P. and Wiencke, J. K. (2000). Negative binomial additive models, *Biometrics*, **56**, 139–144.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem, *The Annals of Statistics*, **13**, 1378–1402.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models, *Journal of the Royal Statistical Society Series B*, **70**, 495–518.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *Journal of the Royal Statistical Society Series B*, **73**, 3–36.

Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models, *Biometrika*, **91**, 661–681.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data, *Statistica Sinica*, **6**, 675–692.

Zhang, W., Lee, S. and Song, X. (2002). Local polynomial fitting semivarying coefficient model, *Journal of Multivariate Analysis*, **82**, 166–188.

2012년 9월 13일 접수; 2012년 10월 12일 수정; 2012년 10월 29일 채택