Shen, Xiangjin; Li, Shiliang; Tsurumi, Hiroki

Working Paper

Comparison of parametric and semi-parametric binary response models

Working Papers, Department of Economics, Rutgers, The State University of New Jersey, No. 2013-08

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Shen, Xiangjin; Li, Shiliang; Tsurumi, Hiroki (2013) : Comparison of parametric and semi-parametric binary response models, Working Papers, Department of Economics, Rutgers, The State University of New Jersey, No. 2013-08

This Version is available at:
http://hdl.handle.net/10419/94225

Terms of use:
Documents in EconStor may be saved and copied for your personal and scholarly purposes.
You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.
Comparison of Parametric and Semi-Parametric Binary Response Models
Xiangjin Shen, Shiliang Li, and Hiroki Tsurumi∗

April 30, 2013

Abstract

A Bayesian semi-parametric estimation of the binary response model using Markov Chain Monte Carlo algorithms is proposed. The performances of the parametric and semi-parametric models are presented. The mean squared errors, receiver operating characteristic curve, and the marginal effect are used as the model selection criteria. Simulated data and Monte Carlo experiments show that unless the binary data is extremely unbalanced the semi-parametric and parametric models perform equally well. However, if the data is extremely unbalanced the maximum likelihood estimation does not converge whereas the Bayesian algorithms do. An application is also presented.

Key Words: Parametric and Semi-parametric binary response models, Markov Chain Monte Carlo algorithms, Kernel densities, Standard bandwidth, Optimal bandwidth, Mean squared errors, Receiver operating characteristic curve, Marginal effect, graphic processing unit

∗Address correspondence to Xiangjin Shen, Shiliang Li and Hiroki Tsurumi, Department of Economics, Rutgers University, NJ 08854, USA. Email: xshen@econ.rutgers.edu, shiliangli1@gmail.com and tsurumi@rci.rutgers.edu.
1 Introduction

The use of generalized linear models (GLMs) (Nelder and Wedderburn(1972)) for the quantitative analysis of social science data has increased appreciably in the past four decades. Especially logit and probit parametric models have been widely used. Amemiya(1981), Aldrich and Nelson (1984) present comprehensive discussions of the GLM binary choice models. Since the distributions of binary responses are not known and are often not estimable, the persistent question is whether semi-parametric models are better than logit or probit parametric models.

In this paper we propose a Bayesian semi-parametric binary choice model using the quasi-likelihood function as the likelihood part of the posterior distribution. We compare the performances of the Bayesian semi-parametric model with the sample theory semi-parametric model. Also we compare the semi-parametric models with probit and logit parametric models. The comparisons are based on simulated data and Monte Carlo experiments. As the criteria of comparison we use the marginal effect, mean squared error (MSE) and receiver operating characteristic (ROC) curve. What we find are (i) when the data is balanced the performances of the semi-parametric models are indistinguishable from the performances of the parametric models (i.e. probit and logit models.) (ii) However, when the data is extremely unbalanced (the yes response rate being less than 3%), the maximum likelihood estimation of the semi-parametric as well as the parametric models does not converge, whereas the Bayesian estimation converges.

After the simulated data and Monte Carlo experiments, we use the Panel Study of Income Dynamics (PSID) data, and test the robustness of the Bayesian semi-parametric binary choice model and other binary choice models.

The organization of the paper is as follows. In Section 2, the Bayesian binary choice model and estimation method are presented. In Section 3 we compare different estimators using simulated data. In Section 4 Monte Carlo experiments are presented. An empirical application is presented in Section 5. Concluding remarks are given in Section 6.
2 Bayesian Semi-Parametric Binary Response Model

We have a sample of binary responses, \( y_1, \ldots, y_n \), where

\[
y_i = \begin{cases} 
1 & \text{if yes with probability } p_i \\
0 & \text{otherwise with probability } 1 - p_i 
\end{cases}
\]

and \( p_i \) is given by \( p_i = F(x_i \beta) \), where \( x_i = (x_{i1}, \ldots, x_{ik}) \) and \( \beta = (\beta_0, \beta_1, \ldots, \beta_k)' \).

Since \( F(\cdot) \) is not known, we use the quasi-likelihood function:

\[
\ell(\beta | f_i, x_i = 1, \ldots, n) = \prod_{i=1}^{n} \hat{p}_i^{f_i} \times \prod_{i=1}^{n} (1 - \hat{p}_i)^{1-f_i}.
\]

We follow the single index-parametric model of Klein and Spady (1993) and Klein and Vella (2009) among others, and obtain \( \hat{p}_i \) by

\[
\hat{p}_i = Pr[Y_i = 1 | V_i(\beta)] = \frac{p(Y = 1)\hat{g}(V | Y = 1)}{\hat{g}(V)} = \frac{\frac{n_1}{n} \sum_{i=1}^{n} \frac{1}{h_n} K(\frac{t - v_i}{h_n}) (\frac{y_i}{n_1})}{\sum_{i=1}^{n} \frac{1}{h_n} K(\frac{t - v_i}{h_n}) \frac{n_1}{n}}.
\]

\[
V_i(\beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.
\]

\[
\hat{g}(t) = \sum_{i=1}^{n} \frac{1}{h_n} K(\frac{t - v_i}{h_n}) \frac{n_1}{n}
\]

is a non-parametric kernel density estimation function, where

\[
K(\frac{t - v_i}{h_n})
\]

is the kernel function satisfying \( \int K(x)dx = 1 \) and \( K(x) \geq 0 \), and \( h_n \) is the kernel density window size or bandwidth.
$V_i(\beta)$ in equation (3) is the index. Given the linearity of $V_i$ in equation (3) we may write:

$$X_i\beta = \beta_1(X_{i1} + \theta_2X_{i2} + \cdots + \theta_kX_{ik}) + \beta_0,$$

where $\theta_i = \beta_i/\beta_1$, $\beta_1 \neq 0$. From equation (5) we get the new single index as

$$V_i(\theta) = X_{i1} + \theta_2X_{i2} + \cdots + \theta_kX_{ik}.$$  

Because the probability of the linear transformation of the index is the same as the probability of the original index, equation (2) will have the following property:

$$Pr(Y = 1 \mid V = v(\beta)) = Pr(Y = 1 \mid V = v(\theta)).$$  

Rather than maximizing the quasi-likelihood function, we propose a Bayesian semi-parametric estimation algorithm by using the quasi-likelihood function (1) as the likelihood to obtain the posterior distribution of $\theta = (\theta_2, \ldots, \theta_k)$:

$$p(\theta) \propto \pi(\theta)\ell(\theta \mid \text{data}),$$

where $\pi(\theta)$ is the prior and $\ell(\cdot)$ is the quasi-likelihood.\(^1\) We use MCMC algorithms with the Metropolis-Hastings criterion.

The MCMC algorithms are carried out as follows: let $\theta^{(i)}$ be the $i$-th draw of $\theta$.

**Step 1** Choose an initial value $\theta^{(0)}$. We use the OLS estimates of the standardized transformed model of equation (6).\(^2\)

$$y_i = x_{i1} + \theta_2x_{i2} + \cdots + \theta_kx_{ik}.$$  

---

\(^1\)Zhang, Silvapulley and Papaspirouz (2009) also use the quasi-likelihood in Bayesian inference, but they set priors for both $\theta$ and the bandwidth and get the posterior with both $\theta$ and bandwidth.

\(^2\)In the maximum likelihood estimation of the semi-parametric model, the covariate $x_{ij}$ is standardized as $x_{ij}/s_j$, where $s_j$ is the standard deviation of $x_{ij}$’s. This standardization of the covariates is done to make the convergence of the MLE procedure easier and get rid of the large variances influences among different types of variables.
**Step 2** We use a random walk draw:

\[ \theta^{(i)} = \theta^{(i-1)} + \varepsilon_i \]

where \( \varepsilon_i \) is normal with mean 0 and variance \( c(X'X)^{-1} \). We set \( c = 1 \).

**Step 3** Set \( \theta^{(i)} = \theta^{(i)} \) if \( u < \alpha \). Otherwise set \( \theta^{(i)} = \theta^{(i-1)} \), where \( u \) is drawn from Uniform(0, 1) and \( \alpha \) is given by

\[ \alpha = \min \left\{ 1, \frac{p(\theta^{(i)} \mid \text{data})}{p(\theta^{(i-1)} \mid \text{data})} \right\}. \]

\( p(\cdot \mid \text{data}) \) is the posterior pdf of \( \theta \).

**Step 4** Repeat **Step 2** and **Step 3** for \( i = 1, 2, \ldots, M \).

In estimating the semi-parametric model the kernel density of equation (4) is used for both the MLE and Bayesian estimation. In the case of the Bayesian MCMC algorithm the kernel density is estimated for each draw of \( \theta^{(i)} \). In the case of the MLE the kernel density is estimated for each iteration until convergence is attained.

The kernel density is dependent on the choice of the kernel, \( K(\cdot) \) and the bandwidth, \( h \). Li (2001) shows that the choice of the bandwidth is more important than the choice of the kernel. Keeping the normal distribution as the kernel, we use two bandwidths to see if the choice of the bandwidth makes the difference in the MLE as well as in the Bayesian estimation. The first bandwidth we use is Silverman’s estimation (1998):

\[ h = \left( \frac{4}{3n} \right)^{2/3} \sigma. \]  

(9)

We call this bandwidth the usual bandwidth. The second bandwidth we use is the optimal bandwidth given by

\[ h_{\text{optimal}}^* = \left( \frac{R(K)}{\left( \int x^2 K(x) dx \right)^2 \widetilde{R}(\hat{g}''(x; p(h)))} \right)^2. \]  

(10)

The optimal bandwidth \( h_{\text{optimal}}^* \) is explained in the appendix. The optimal bandwidth tends to trace sharp modes of a density better than the usual bandwidth. This is illustrated in Figures 1
and 2 where 15 Gaussian mixture densities are presented.

*Figures 1 and 2 Here.*

In Figures 1 and 2 the solid black lines are the true Gaussian mixture densities, whereas the
lines in red are kernel densities. In Figure 1 the kernel densities are obtained using the usual
bandwidth while in Figure 2 they are obtained by the optimal bandwidth. We see that the usual
bandwidth in Figure 1 misses the sharp modes of the true densities but the optimal bandwidth
in Figure 1 traces the sharp modes fairly accurately as vividly illustrated by the multimodal
claw distribution in the center of Figures 1 and 2.

The computation of the optimal bandwidth is time consuming and thus we use a graphic
processing unit (GPU). GPU computation has been used more and more in Bayesian estimation
of many models.

3 **Comparing the Performances of Parametric and Semi-parametric
Binary Response Models**

Let us compare the performances of the parametric and semi-parametric binary response mod-
els using the Bayesian and maximum likelihood estimators. We choose the probit and logit
models as the parametric models. For the semi-parametric model we use two bandwidths: the
usual bandwidth of equation (9) and the optimal bandwidth of equation (10). In summary the
estimators and models we compare are:

\[
\text{Bayesian} \begin{cases} 
\text{Probit} \\
\text{Logit} \\
\text{Semi-parametric with the usual bandwidth} \\
\text{Semi-parametric with the optimal bandwidth}
\end{cases}
\]
MLE \left\{ \begin{array}{l}
\text{Probit} \\
\text{Logit} \\
\text{Semi-parametric with the usual bandwidth} \\
\text{Semi-parametric with the optimal bandwidth}
\end{array} \right.

As given in equation (5) in the semi-parametric model the regression coefficients, \( \beta_i \)'s, are transformed into \( \theta_i \)'s. This makes it difficult to compare the regression coefficient estimates from a parametric model to those from the semi-parametric model. Therefore let us use three model selection criteria: the marginal effects, mean squared errors and the receiver operating characteristic (ROC) curve.

The marginal effect is a popular statistic for the binary response model. When the distribution is known or the model is parametric, the generalized form of the true marginal effect of \( X_k \) for models with known density distribution is:

\[
\frac{\partial F(x_i \beta)}{\partial x_k} = \frac{\partial F(x_i \beta)}{\partial x_i \beta} \frac{\partial x_i \beta}{\partial x_k} = F'(x_i \beta) \beta_k = f(x_i \beta) \beta_k.
\]

Within the semi-parametric model the marginal effect needs to be defined differently: we use the predicted probability, \( \hat{p} \), and define the estimated marginal effect as \( \hat{p}(x + \Delta x) - \hat{p}(x) \), in which \( \hat{p}(.) \) is given in equation (2) and \( \Delta x \) is an increment of the \( x \). In order to capture the entire distribution of the \( X \), we will consider \( \Delta x = \{ \text{std}(x), 2 \times \text{std}(x), 3 \times \text{std}(x) \} \).

In regressions, one way to select model is to choose the model with the smallest unweighted MSE or the normal MSE\(^{4}\), which is calculated by

\[
\sum_{i=1}^{n} \frac{(y_i - \hat{P}_i)^2}{n - k}.
\]

where \( y_i = 1 \) or \( 0 \); \( \hat{P}_i \) is the computed probability \( F(x_i \hat{\beta}) \) for the case of a parametric model or the equation (2) for the case of the semi-parametric model.

\(^{3}\text{We may also use the quantile of } X \text{ as } \Delta x.\)

\(^{4}\text{There is the weighted mean squared errors. Amemiya (1981) argues for the use of the weighted mean squared errors, but as shown in Chen and Tsurumi (2010) the unweighted MSE is a better model selection criterion than the weighted MSE.}\)
The receiver operating characteristic (ROC) curve is one of the best choices (McNeil and Hanley (1982, 1984), Swets et al. (2000), Fawcett (2006), etc.) to select the binary response model. We compare the area under the ROC curve (Alonzo (2002), Agresti (2007)). The bigger is the area, the better the predictive power of the binary response model is. We will use the algorithm from Fawcett (2006) to plot ROC curve and calculate the area under the ROC curve.

Let us compare the performances of the different estimation methods and models by a simulated data. We specify the binary choice model to be

\[ Y_i^* = \beta_0 + \beta_1 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \]  

where \( X_{i3} \) is a zero-one dummy variable to represent a discrete covariate and \( X_{i2} \) is drawn from a uniform distribution, \( U(0, a) \). The continuous regressor, \( X_{i2} \) is included since the large sample properties of the semi-parametric estimator requires that at least one regressor is a continuous variable. The values of the parameters \( (\beta_0, \beta_1, \beta_2, a) \) are chosen to control the percentage of \( Y_i = 1 \) to represent a balanced or unbalanced data. The observed binary values, \( Y_i \), are set as

\[ Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}. \]

The sample size \( n \) is set at 1,000 (\( n = 1,000 \)).

Before we compare the performances of the sample theory and Bayesian estimates of the binary response models, let us see how well the discretized marginal effect:

\[ p(x + \Delta x) - p(x) \]

is estimated by the sample theory and Bayesian semi-parametric models. We generated a sample of size 1,000 (\( n = 1,000 \)) from the logistic distribution setting \( \beta = \{\beta_0, \beta_1, \beta_2\} \). The percentage of \( Y = 1 \) is 9.25\%. \( \Delta x \) is set at one, two, and three standard deviations.
From the table above we see that the discretized marginal effects are reasonably well estimated by the sample theory and Bayesian models.

The error term in equation (11), $\varepsilon_i$, is also drawn from 16 distributions that are given in Table 1. Distributions #1–15 are Gaussian mixture densities from Marron and Wand (1992) and distribution #16 is a skew logistic distribution with the distribution function given by

$$\Pr(y_i = 1) = \frac{1}{(1 + e^{-x_i \beta})^\theta}.$$

The first 15 distributions are presented in Figures 1 and 2. Some of these distributions, especially, trimodal, claw, and comb distributions may seldom occur in real data, but these distributions are different from the probit or logit distributions and thus the semi-parametric models may perform better than the parametric models.

Table 1: Distributions of the Error Terms of the Binary Response Models

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Gaussian</td>
<td>9 Trimodal</td>
</tr>
<tr>
<td>2 Skewed unimodal</td>
<td>10 Claw</td>
</tr>
<tr>
<td>3 Strongly skewed</td>
<td>11 Double Claw</td>
</tr>
<tr>
<td>4 Kurtotic unimodal</td>
<td>12 Asymmetric Claw</td>
</tr>
<tr>
<td>5 Outlier</td>
<td>13 Asymmetric Double Claw</td>
</tr>
<tr>
<td>6 Bimodal</td>
<td>14 Smooth Comb</td>
</tr>
<tr>
<td>7 Separated bimodal</td>
<td>15 Discrete Comb</td>
</tr>
<tr>
<td>8 Skewed bimodal</td>
<td>16 Skew Logistic</td>
</tr>
</tbody>
</table>

Table 2 and Table 3 present the ROC areas and MSE’s of different estimators based on simulated data. Although we have obtained results for all of the 16 distributions the results are
quite similar to those given in Tables 2 and 3. Table 2 is for the balanced cases in which the percentage of $Y = 1$ ranges from 21.4% to 66.7% while Table 3 is for extremely unbalanced cases in which the percentage of $Y = 1$ ranges from 6% to 2.5%.

From Table 2 and 3 we conclude that judged by the ROC area and MSE, we cannot discriminate among the different models and estimation procedures except in the cases of extremely unbalanced data as given in Table 3: all the MLE estimation procedures failed to converge, whereas all the Bayesian MCMC algorithms attain convergence. Hence we conclude that when the data is extremely unbalanced the Bayesian MCMC algorithms may be preferred to the MLE algorithms. Comparing the bandwidths, we see that the use of the optimal bandwidth does not have a visible advantage over the standard bandwidth.

4 Monte Carlo Experiments

In the previous section based on one sample draw we compared the performances of the different models and estimation procedures. In this section I conduct Monte Carlo (MC) experiments to compare the performances of the different models and estimation procedures.

In the literature MC results using the optimal bandwidth is few because of the heavy computational burden in searching the optimal bandwidth. The most difficult part of MC simulation is to estimate the optimal bandwidth efficiently without smoothing techniques and specific bounds, and to run Bayesian MCMC simulations quickly. By using GPU (graphics processing unit) computing with C/C++ in Matlab(Li(2011)), which is more than 400 times faster than the regular computing method, we are able to make the MC simulations effectively.

The number of Monte Carlo replications is 500. The Monte Carlo simulation results are consistent with results obtained in the previous section for the simulated data. Therefore, we only present two examples for balanced cases and unbalances cases.
In Table 4, the first part is for the balanced case with claw distribution and the second part is for unbalanced case of the model with skewed log distribution; only Bayesian MC results can be presented for the unbalanced case because not all 500 replications yield convergence when the MLE is used. Clearly the results from ROC and MSE are very similar among different models: either parametric or Semiparametric models by either MLE or Bayesian methods. However, the marginal effects from both MLE or Bayesian Semiparametric methods are smaller than parametric methods. In the balanced case we see that the mean of the Bayes Semi optimal bandwidth and the mean of the MLE Semi optimal bandwidth are .3579 and .3323, respectively, and they are close to each other, but the standard deviation of the Bayes Semi optimal bandwidth is much smaller than that of the MLE Semi optimal bandwidth.

5 Analysis of employment status

Let us present an application to the employment status using the PSID data. The Panel Study of Income Dynamics (PSID) is a longitudinal survey that collects economic and demographic information of U.S. families Since 1968. PSID data has been frequently used to investigate the employment status of the U.S. in the literature (James and Audrey (1992), David and Ann (1999), Lawrence et.al. (2006)).

Our data is extracted from PSID family data for those individuals who are the family head in the labor force market. There are 8002 family heads in labor force in 2005 survey. After dropping the missing values there are 4034 observations. We set

\[ Y = \begin{cases} 
1 & \text{if head of family is unemployed} \\
0 & \text{otherwise.} 
\end{cases} \]

According to the Bureau of Labor Statistics, the annual unemployment rate in year 2005 is 6.1%. In our data, the unemployment rate \( P(Y=1) \) is 6.17%, and we may say the data is
not extremely unbalanced. Three covariates or regressors are continuous variable: they are age, years of education and years of the working experience. Four covariates are categorical variables: sex (male or female, ‘1’ or ‘0’), race (white or non-white, ‘1’ or ‘0’), marital status (married or not, ‘1’ or ‘0’), and city size (6 levels, ‘1’ to ‘6’, bigger number means smaller city). Summary statistics of all variables are in Table 5. The covariates characterizing the family head are age and years of work. The mean and median age of the head of family is around 41 with the standard deviation 12.33. The mean and median years of working is around 10 years with the standard deviation 9.12. Judged by the quantiles all the variables distributed symmetric.

Table 5 Here.

The parameter estimates and standard deviations are given in Table 6. The estimated parameters of the logit and probit models yield the same signs except the constant term. The MLE and Bayes estimates (posterior means) are similar. The signs of the estimated parameters of the semi-parametric models are in general opposite of the signs of the estimated parameters of the logit and probit models. This is because the parameters of the semi-parametric model is normalized by the parameter of age, \( \beta_1 \): \( \theta_i = \frac{\beta_i - 1}{\beta_1} \).

Table 6 Here.

The MSE and ROC of the parametric and semi-parametric models are similar but a careful examination shows that the ROC curves of the semi-parametric models are higher than those of the parametric models as shown in Figure 3.

Figure 3 Here.

The marginal effects analysis is in Table 7. The marginal effects are more informative measures than the parameters in a binary response model.

Table 7 Here.
It is clear that the marginal effects among each model are close, except Citysize and YRWORK for the MLE estimate of the semi-parametric model. For Age and YRWORK variables, the marginal effects are computed using three $\Delta(x)$: 1 std, 2 std, and 3 std where std is the standard deviation, because these variables have large ranges: 18 to 83 for Age and 0 to 51 for YRWORK. The negative sign indicates that older people are less likely to be unemployed than younger people. This is consistent with the fact that in general older people have more working experience and are easy to find a job. Similarly, all the models show that white, high-educated, married people have a less probability of losing jobs. All the models except the MLE of the semi-parametric model give the result that people living in a large city have a lower chance of getting unemployed. As to the years of working experience (YRWORK), all the models except the MLE semi-parametric model, show that a longer working experience increases the probability of unemployment.

Although the differences among models are small, generally the semi-parametric model estimated by either the MLE or MCMC yields a better ROC. Based on ROC we may say that the semi-parametric model is a better model for the PSID data.

6 Concluding Remarks

We first presented a Bayesian semi-parametric binary response model based on the quasi-likelihood function that is based on the kernel density estimate. The major difference between our Bayesian semi-parametric binary response model and the sample theoretic semi-parametric binary response model of Klein and Spady (1993) is that we use the Markov Chain Monte Carlo (MCMC) algorithm with the Metropolis-Hastings criterion rather than the maximum likelihood estimator. We used the normal kernel and employed two bandwidths: the usual bandwidth and the optimal bandwidth.

Using simulated data we compared the performances of the semi-parametric models to those of the logit and probit models. We used the MLE and MCMC algorithms. The error term of the regression model is generated from 16 different distributions. The comparison of
performances is based on the mean squared errors (MSE), the receiver operating characteristic curve (ROC) and the marginal effect. We find that the performances of the parametric and semi-parametric models are virtually indistinguishable if they estimated by the MLE or MCMC procedures except when the data is extremely unbalanced ( % of \(Y = 1\) < 3%). In the extremely unbalanced cases the MCMC procedure work but the MLE does not converge.

Although the optimal bandwidth traces sharp modes better than the usual bandwidth as shown in Figures 1 and 2, the quasi-likelihood function produced by the kernel density with the optimal bandwidth is not much different than the one produced by the usual bandwidth. Consequently, the semi-parametric models based on the optimal bandwidth yield virtually the same results as the semi-parametric models based on the usual bandwidth do.

As an application we estimated the binary response model using the PSID data. We set the unemployed head of family as 1, and the employed head of family as 0. All the parametric and semi-parametric models yield similar estimates except the city size and years of work variables. Judged by the ROC curves, the semi-parametric models are better than the parametric models.

There are Bayesian semi-parametric qualitative choice models. One model is based on the B-splines to approximate the link function using Laplace transform of the normal distribution (Fahrmeir and Lang (2001), Antoniadis and Ian (2004), Fahrmeir and Raach (2007)). The second model uses the binary response version of the median regression model (Newton and Chappell (1996), Kottas and Gelfand (2001)). Both of these methods need link functions subject to identifiability. It will be interesting to compare our Bayesian semi-parametric model to the models by these authors.
Appendix Optimal bandwidth

Wand and Jones (1995) and Silverman (1986) show that we can obtain the optimal bandwidth \( h \) by minimizing the Mean Integrated Squared Error (MISE):

\[
MISE \{ \hat{g}(x;lh), g(x) \} = E \left[ \int (\hat{g}(x;h) - g(x))^2 dx \right],
\]

where \( g(\cdot) \) is the non-parametric kernel density estimation function. It is clear that integration needs to be made on the whole real line, \( x \in (-\infty, \infty) \) instead of a finite discrete set. Li (2011) shows that the choice of the kernel function \( K(x) \) is not as important as the choice of the bandwidth. Hence, we will use the standard normal distribution for \( K(\cdot) = \Phi(\cdot) \) and will find the optimal bandwidth.

By applying the Central Limit Theorem (CLT), we get an approximation of MISE called Asymptotic Mean Integrated Squared Error (AMISE):

\[
AMISE \{ \hat{g}(x;h), g(x) \} = (Nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(g),
\]

where \( R(K) = \int K(x)^2 dx \) and \( \mu_2(K) = \int x^2K(x) dx \). The AMISE is a monotonic function of the optimal bandwidth \( h \) and the optimal \( h \) is generally defined as:

\[
h_{optimal} = \left[ \frac{R(K)}{(\int x^2K)^2 R(g''N)} \right]^{\frac{1}{5}}
\]

This optimal bandwidth cannot be calculated directly because \( R(g'') \) is a function of the second order derivative of the true density function \( g \) which is unknown.

When the data set is Gaussian or asymptotically Gaussian with standard deviation, we will get the regular optimal bandwidth in the literature:

\[
h_{optimal} = \left[ \frac{4}{3N} \right]^{\frac{1}{5}} \sigma
\]

(12)

When data is not normal, this optimal bandwidth may not fit into the real data, and we may
use the most popular solve-the-equation plug-in approach and get the optimal bandwidth as

$$h_{optimal}^* = \left[ \frac{R(K)}{\left( \int x^2 K(x) dx \right)^2 R(\hat{g}''(x; p(h)))} \right]^{\frac{1}{2}}$$

(13)

Here $p(h) = \left[ \frac{-2K^{(4)}(0)\mu_2(K)\hat{\psi}_4}{R(K)\hat{\psi}_6} \right]^{\frac{1}{2}} h^2$ is the optimal pilot bandwidth and $\hat{\psi}_r = \frac{1}{N} \sum_{i=1}^{N} \hat{g}^{(r)}(x_i; p^{(r)})$, where $p^{(r)}$ is the pilot bandwidth to estimate the $r$th derivative of the density $g^{(r)}$.

Equation (12) is the most popular simple optimal bandwidth and it is only optimal for Gaussian data. If data is not Gaussian, we should use equation (13), which requires multiple complex computations and it is extremely time consuming. This is one of the reason that many people use different estimation methods to estimate bandwidth such as Zhang, Silvapulley and Papaspirouz (2009), or build different smooth factors (Chan Shen, Klein (2010)) with specific bound to minimize the bias in estimating bandwidth. The computation in equation (13) can be realized efficiently by using graphic processing unit (GPU) computing with C/C++ in Matlab (Li (2011)), and its speed is about 400 time faster than the regular computing method such as in Gauss or Matlab itself. Therefore, we can estimate much more accurate optimal bandwidth because we will consider all real numbers without any arbitrary lower or upper bound.
References


Table 2: ROC areas and MSE’s: Balanced Cases

<table>
<thead>
<tr>
<th></th>
<th>Strongly skewed</th>
<th>Sparated bimodal</th>
<th>Claw</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y=1 is 66.7%</td>
<td>Y=1 is 21.4%</td>
<td>Y=1 is 21.4%</td>
</tr>
<tr>
<td>ROC area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>probit</td>
<td>0.67</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>logit</td>
<td>0.67</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>semi</td>
<td>0.67</td>
<td>0.75</td>
<td>0.88</td>
</tr>
<tr>
<td>semi-opt</td>
<td>0.67</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>probit</td>
<td>0.2</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>logit</td>
<td>0.2</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>semi</td>
<td>0.19</td>
<td>0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>semi-opt</td>
<td>0.19</td>
<td>0.14</td>
<td>0.03</td>
</tr>
</tbody>
</table>

|               |                  |                   |               |
| ROC area      |                  |                   |               |
| MSE           |                  |                   |               |
| Bayes         |                  |                   |               |
| probit        | 0.27            | 0.76              | 0.88          |
| logit         | 0.27            | 0.76              | 0.88          |
| semi          | 0.67            | 0.76              | 0.88          |
| semi-opt      | 0.67            | 0.76              | 0.88          |
| probit        | 0.2             | 0.14              | 0.04          |
| logit         | 0.2             | 0.15              | 0.04          |
| semi          | 0.19            | 0.14              | 0.04          |
| semi-opt      | 0.19            | 0.14              | 0.03          |

Notes: semi = semi parametric with the bandwidth $h$ in equation (9)
semi-opt = semi parametric with the optimal $h_{optimal}$ in equation (10)
Table 3: ROC areas and MSE’s: Extremely unbalanced Cases

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Skewed logistic ($\theta = .1$) Y=1 is 2.5%</th>
<th>Outlier Y=1 is .6%</th>
<th>Kurtotic unimodal Y=1 is 1.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROC area</td>
<td>MSE</td>
<td>ROC area</td>
</tr>
<tr>
<td>Bayes</td>
<td>Bayes</td>
<td>Bayes</td>
<td>Bayes</td>
</tr>
<tr>
<td>probit</td>
<td>.96</td>
<td>.02</td>
<td>.85</td>
</tr>
<tr>
<td>logit</td>
<td>.96</td>
<td>.02</td>
<td>.85</td>
</tr>
<tr>
<td>semi</td>
<td>.96</td>
<td>.02</td>
<td>.87</td>
</tr>
<tr>
<td>semi-opt</td>
<td>.96</td>
<td>.02</td>
<td>.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MLE</th>
<th>MLE</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>probit</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>logit</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>semi</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>semi-opt</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
</tr>
</tbody>
</table>

Notes: semi = semi parametric with the bandwidth $h$ in equation (9)
semi-opt = semi parametric with the optimal $h_{optimal}$ in equation (10)
NC = not converged
### Table 4: Monte Carlo Experiment Result

Claw distribution, around 20% of ‘Y=1’  
Replications = 500

<table>
<thead>
<tr>
<th>ROC area</th>
<th>Marginal Effect</th>
<th>Binary Model Evaluation Criteria</th>
<th>MSE MEAN STD.</th>
<th>MSE MEAN STD.</th>
<th>MSE MEAN STD.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Probit</td>
<td></td>
<td>Bayes Probit</td>
<td>0.0370 0.0042</td>
<td>0.0702 0.0121</td>
<td></td>
</tr>
<tr>
<td>Bayes Logit</td>
<td></td>
<td>Bayes Logit</td>
<td>0.0370 0.0042</td>
<td>0.0739 0.0122</td>
<td></td>
</tr>
<tr>
<td>Bayes Semi</td>
<td></td>
<td>Bayes Semi</td>
<td>0.0365 0.0042</td>
<td>0.0519 0.0084</td>
<td></td>
</tr>
<tr>
<td>Bayes Semiopt</td>
<td></td>
<td>Bayes Semiopt</td>
<td>0.0365 0.0042</td>
<td>0.0516 0.0086</td>
<td></td>
</tr>
<tr>
<td>MLE Probit</td>
<td></td>
<td>MLE Probit</td>
<td>0.0370 0.0042</td>
<td>0.0704 0.0121</td>
<td></td>
</tr>
<tr>
<td>MLE Logit</td>
<td></td>
<td>MLE Logit</td>
<td>0.0370 0.0042</td>
<td>0.0739 0.0122</td>
<td></td>
</tr>
<tr>
<td>MLE Semi</td>
<td></td>
<td>MLE Semi</td>
<td>0.0363 0.0042</td>
<td>0.0576 0.0103</td>
<td></td>
</tr>
<tr>
<td>MLE Semiopt</td>
<td></td>
<td>MLE Semiopt</td>
<td>0.0364 0.0042</td>
<td>0.0552 0.0113</td>
<td></td>
</tr>
</tbody>
</table>

### Optimal Bandwidth

<table>
<thead>
<tr>
<th>BayesSemi optimal bandwidth</th>
<th>MLE Semi optimal bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3579 0.0198</td>
<td>0.3323 0.0480</td>
</tr>
</tbody>
</table>

### Skewed log alpha=0.25, around 2.5% of ‘Y=1’  
Replications = 500

<table>
<thead>
<tr>
<th>ROC area</th>
<th>Marginal Effect</th>
<th>Binary Model Evaluation Criteria</th>
<th>MSE MEAN STD.</th>
<th>MSE MEAN STD.</th>
<th>MSE MEAN STD.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Probit</td>
<td></td>
<td>Bayes Probit</td>
<td>0.0108 0.0019</td>
<td>0.0890 0.0154</td>
<td></td>
</tr>
<tr>
<td>Bayes Logit</td>
<td></td>
<td>Bayes Logit</td>
<td>0.0108 0.0020</td>
<td>0.0907 0.0151</td>
<td></td>
</tr>
<tr>
<td>Bayes Semi</td>
<td></td>
<td>Bayes Semi</td>
<td>0.0109 0.0019</td>
<td>0.0752 0.0153</td>
<td></td>
</tr>
<tr>
<td>Bayes Semiopt</td>
<td></td>
<td>Bayes Semiopt</td>
<td>0.0109 0.0019</td>
<td>0.0712 0.0144</td>
<td></td>
</tr>
<tr>
<td>BayesSemi optimal bandwidth</td>
<td></td>
<td>0.3225 0.0019</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5: Summary Statistics of the 2005 PSID data

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>unemployment</td>
<td>4034</td>
<td>0.06173</td>
<td>0.24069</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>racial05</td>
<td>4034</td>
<td>0.42712</td>
<td>0.49472</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>sex</td>
<td>4034</td>
<td>0.76301</td>
<td>0.42529</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>age05</td>
<td>4034</td>
<td>41.556</td>
<td>12.332</td>
<td>18</td>
<td>31</td>
<td>42</td>
<td>51</td>
<td>83</td>
</tr>
<tr>
<td>educ05</td>
<td>4034</td>
<td>13.272</td>
<td>2.469</td>
<td>0</td>
<td>12</td>
<td>13</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>marital05</td>
<td>4034</td>
<td>0.66262</td>
<td>0.47288</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>citysize05</td>
<td>4034</td>
<td>3.3508</td>
<td>1.7707</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>YRWORK05</td>
<td>4034</td>
<td>10.944</td>
<td>9.12</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>51</td>
</tr>
</tbody>
</table>

Notes: Q1 and Q3 are 25 and 75 percentiles, respectively.
Table 6: Estimations for the 2005 PSID data (Y=1.0 is 6.17%)

<table>
<thead>
<tr>
<th>variable</th>
<th>MLE</th>
<th>Probit</th>
<th>Bayes</th>
<th>Semi ($\theta$)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Logistic</td>
<td>Probit</td>
<td>Logistic</td>
<td>Probit</td>
</tr>
<tr>
<td>c Constant</td>
<td>0.4148</td>
<td>-0.0183</td>
<td>0.3866</td>
<td>-0.0303</td>
</tr>
<tr>
<td>\beta_1 Age</td>
<td>-0.0118</td>
<td>-0.0054</td>
<td>-0.01211</td>
<td>-0.0050</td>
</tr>
<tr>
<td>\beta_2 Racial05</td>
<td>-0.1464</td>
<td>-0.0657</td>
<td>-0.1369</td>
<td>-0.0633</td>
</tr>
<tr>
<td>\beta_3 Sex</td>
<td>-0.4312</td>
<td>-0.2121</td>
<td>-0.4379</td>
<td>-0.2012</td>
</tr>
<tr>
<td>\beta_4 Educ05</td>
<td>-0.1487</td>
<td>-0.0723</td>
<td>-0.1471</td>
<td>-0.0722</td>
</tr>
<tr>
<td>\beta_5 Marital05</td>
<td>-0.4441</td>
<td>-0.2098</td>
<td>-0.4282</td>
<td>-0.2219</td>
</tr>
<tr>
<td>\beta_6 Citysize05</td>
<td>-0.0457</td>
<td>-0.0261</td>
<td>-0.0488</td>
<td>-0.0276</td>
</tr>
<tr>
<td>\beta_7 YRWORK05</td>
<td>0.0023</td>
<td>0.0018</td>
<td>0.0028</td>
<td>0.0020</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0566</td>
<td>0.0566</td>
<td>0.0567</td>
<td>0.0567</td>
</tr>
<tr>
<td>ROC area</td>
<td>0.6641</td>
<td>0.6646</td>
<td>0.6643</td>
<td>0.6645</td>
</tr>
</tbody>
</table>

Notes: 1. First row is MLE, second row is standard error.
2. First row and second row are posterior mean and standard error, respectively.
Table 7: Marginal Effects of the 2005 PSID data

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Marginal Effects</th>
<th>MLE Logistic</th>
<th>MLE Probit</th>
<th>Bayes Logistic</th>
<th>Bayes Probit</th>
<th>MLE-Semi</th>
<th>Bayes-Semi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>$\Delta x = 1std$</td>
<td>-0.0077</td>
<td>-0.0074</td>
<td>-0.0080</td>
<td>-0.0067</td>
<td>-0.0041</td>
<td>-0.0075</td>
</tr>
<tr>
<td></td>
<td>$\Delta x = 2std$</td>
<td>-0.0146</td>
<td>-0.0141</td>
<td>-0.0151</td>
<td>-0.0129</td>
<td>-0.0077</td>
<td>-0.0142</td>
</tr>
<tr>
<td></td>
<td>$\Delta x = 3std$</td>
<td>-0.0207</td>
<td>-0.0201</td>
<td>-0.0212</td>
<td>-0.0185</td>
<td>-0.0111</td>
<td>-0.0200</td>
</tr>
<tr>
<td>Racial05</td>
<td>dummy</td>
<td>-0.0034</td>
<td>-0.0032</td>
<td>-0.0034</td>
<td>-0.0031</td>
<td>-0.00128</td>
<td>-0.00099</td>
</tr>
<tr>
<td>Sex</td>
<td>dummy</td>
<td>-0.0185</td>
<td>-0.0191</td>
<td>-0.0256</td>
<td>-0.0186</td>
<td>-0.01502</td>
<td>-0.0227</td>
</tr>
<tr>
<td>Educ05</td>
<td>$\Delta x = 1std$</td>
<td>-0.0178</td>
<td>-0.0182</td>
<td>-0.0176</td>
<td>-0.0174</td>
<td>-0.0157</td>
<td>-0.0171</td>
</tr>
<tr>
<td>Marital05</td>
<td>dummy</td>
<td>-0.0156</td>
<td>-0.0155</td>
<td>-0.0156</td>
<td>-0.0148</td>
<td>-0.0135</td>
<td>-0.0138</td>
</tr>
<tr>
<td>Citysize05</td>
<td>$\Delta x = 1std$</td>
<td>-0.0044</td>
<td>-0.0052</td>
<td>-0.0044</td>
<td>-0.0046</td>
<td>0.001755</td>
<td>-0.00264</td>
</tr>
<tr>
<td>YRWORK05</td>
<td>$\Delta x = 1std$</td>
<td>0.0012</td>
<td>0.0020</td>
<td>0.001468</td>
<td>0.0017</td>
<td>-0.0048</td>
<td>0.0031</td>
</tr>
<tr>
<td></td>
<td>$\Delta x = 2std$</td>
<td>0.0025</td>
<td>0.0041</td>
<td>0.0029</td>
<td>0.0035</td>
<td>-0.0091</td>
<td>0.0066</td>
</tr>
<tr>
<td></td>
<td>$\Delta x = 3std$</td>
<td>0.0038</td>
<td>0.0062</td>
<td>0.0045</td>
<td>0.0054</td>
<td>-0.0129</td>
<td>0.0105</td>
</tr>
</tbody>
</table>

Notes: $std =$ standard deviation.
Figure 1: Various distributions and kernel densities using the usual bandwidth.

- Gaussian
- Skewed unimodal
- Strongly skewed
- Kurtotic unimodal
- Outlier
- Bimodal
- Separated bimodal
- Skewed bimodal
- Trimodal
- Double Claw
- Asymmetric Claw
- Asymmetric Double Claw
- Smooth Comb
- Discrete Comb

Legend:
- Black: Actual Density
- Red: Estimated Density
Figure 2: Various distributions and kernel densities using the optimal bandwidth.
The ROC curves above show that at the same level of the 'False Positive Rate' \( P(\hat{y} = 1 | y = 0) \), Semiparametric model gives a greater value of the 'True positive rate' \( P(\hat{y} = 1 | y = 1) \).

For example, when 'False Positive Rate(FP) = 0.5', 'True positive rate (TP) of Semi=0.71' > 'True positive rate of Probit&Logit=0.65'. The Fp&Tp rates of the Semiparametric model are calculated by a specific cut-off \( \pi_0 \) from the following classification table:

<table>
<thead>
<tr>
<th>Actual</th>
<th>Prediction, set ( \pi_0 = 0.64 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y = 1 )</td>
<td>( \hat{y} = 1 )</td>
</tr>
<tr>
<td>( y = 0 )</td>
<td>( a )</td>
</tr>
<tr>
<td>( y = 0 )</td>
<td>( c )</td>
</tr>
</tbody>
</table>

The predicted \( \hat{y} \)'s in this table get from setting \( \pi_0 = 0.64 \) in the example.

So the TP rate = \( P(\hat{y} = 1 | y = 1) = a/(a+b) = 0.71 \); and FP rate = \( P(\hat{y} = 1 | y = 0) = c/(c+d) = 0.5 \).

This is only for one point on the ROC curve according to one cut-off \( \pi_0 \). For each model, we can calculate N points according to N values of \( \pi_0 \), linking these N points will be our ROC curve for the Semiparametric model. The ROC curve for logit&probit model are plotted by the same way.