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 $p_i = \mathrm{F}(x_i\beta)$, where $x_i = (x_{i1}, \cdots, x_{ik})$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$.

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

$$\ell(\beta \mid f_i, x_i = 1, \ldots, n) = \prod_{i=1}^{n} \hat{p}_i^{f_i} \times \prod_{i=1}^{n} (1 - \hat{p}_i)^{1-f_i}. \tag{1}$$

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

$$\begin{aligned} \hat{p}_i &= Pr[Y_i = 1 \mid V_i(\beta)] \\ &= \frac{p(Y=1)\hat{g}(V \mid Y=1)}{\hat{g}(V)} = \frac{\frac{n_1}{n} \sum_{i=1}^{n} \frac{1}{h_n} K(\frac{t-v_i}{h_n})(\frac{y_i}{n_1})}{\sum_{i=1}^{n} \frac{1}{h_n} \frac{K(\frac{t-v_i}{h_n})}{n}}. \end{aligned} \tag{2}$$

$$V_i(\beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \tag{3}$$

$$\hat{g}(t) = \sum_{i=1}^{n} \frac{1}{h_n} \frac{K(\frac{t-v_i}{h_n})}{n} \tag{4}$$

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:

$$X_i\beta = \beta_1(X_{i1} + \theta_2 X_{i2} + \cdots + \theta_k X_{ik}) + \beta_0, \tag{5}$$

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

$$V_i(\theta) = X_{i1} + \theta_2 X_{i2} + \cdots + \theta_k X_{ik}. \tag{6}$$

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:

$$Pr(Y = 1 \mid V = v(\beta)) = Pr(Y = 1 \mid V = v(\theta)). \tag{7}$$

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 $\theta = (\theta_2, \ldots, \theta_k)$:

$$p(\theta) \;\propto\; \pi(\theta)\ell(\theta \mid \text{data}), \tag{8}$$

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 θ.

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

$$y_i = x_{i1} + \theta_2 x_{i2} + \cdots + \theta_k x_{ik}.$$

[1] Zhang, Silvapulley and Papaspirouz (2009) also use the quasi-likelihood in Bayesian inference, but they set priors for both θ and the bandwidth and get the posterior with both θ 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 ε_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 α 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 θ.

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):

$$h = \left(\frac{4}{3n}\right)^{.2} \sigma. \tag{9}$$

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

$$h^*_{optimal} = \left(\frac{R(K)}{\left(\int x^2 K(x)dx\right)^2 R(\hat{g}''(x; p(h)))}\right)^{.2}. \tag{10}$$

The optimal bandwidth $h^*_{optimal}$ 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 models 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:

$$\text{Bayesian} \begin{cases} \text{Probit} \\ \text{Logit} \\ \text{Semi-parametric with the usual bandwidth} \\ \text{Semi-parametric with the optimal bandwidth} \end{cases}$$

$$\text{MLE} \begin{cases} \text{Probit} \\ \text{Logit} \\ \text{Semi-parametric with the usual bandwidth} \\ \text{Semi-parametric with the optimal bandwidth} \end{cases}$$

As given in equation (5) in the semi-parametric model the regression coefficients, β_i's, are transformed into θ_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, $\widehat{p}$, 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 Δx is an increment of the x. In order to capture the entire distribution of the X, we will consider $\Delta x = \{std(x), 2 \times std(x), 3 \times std(x)\}$.[3]

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] We may also use the quantile of X as Δx.

[4] 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

$$Y_i^* = \beta_0 + \beta_1 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \tag{11}$$

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.

ΔX	true marginal effect	semi-parametric model	
		MLE	Bayesian
one std(x_1)	.0109	.0159	.0149
two std(x_1)	.0233	.0273	.0266
three std(x_1)	.0370	.0302	.0307

Notes: std=standard deviation; Bayesian=posterior mean; The usual bandwidth is used.

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), ε_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

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

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

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

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%.

Tables 2 and 3 Here.

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.

Table 4 Here.

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, β_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):

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

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):

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

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

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

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:

$$h_{optimal} = \left[\frac{4}{3N}\right]^{\frac{1}{5}} \sigma \tag{12}$$

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

$$h^*_{optimal} = \left[\frac{R(K)}{\left(\int x^2 K(x) dx \right)^2 R(\hat{g}''(x; p(h)))} \right]^{\frac{1}{5}} \tag{13}$$

Here $p(h) = \left[\frac{-2K^{(4)}(0)\mu_2(K)\hat{\psi}_4}{R(K)\hat{\psi}_6} \right]^{\frac{1}{7}} h^{\frac{5}{7}}$ 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 rth 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

Agresti, A.(2007). An introduction to categorical data analysis. Wiley (2007), 2nd edition, chp5.

Amemiya, T.(1981) ”Qualitative response models. A survey,” Journal of Eco-nomic Literature”. Vol. 19, 1483-1536

Antonidis, G. and I. Jan (2004). ”Bayesian estimation in single-index models”, Statistica Sinica, 14, 1147-1164.

Aldrich, J.H. and F.D. Nelson (1984). ”Linear probability, logit, and probit models”. Beverly Hills: Sage Publications, Inc. 1984.

Alonzo,T. (2002). ”Distribution-free ROC analysis using binary regression techniques”. Biostatistics (2002), 3, pp. 421-432.

Bester H. and Petrakis E. (2003). ”Wages and productivity growth in a competitive industry”. Journal of Economic Theory, 109 (1), 52-69

Chen, G. and H. Tsurumi (2010). ”Probit and logit model selection”. Commu-nications in Statistics

David A. J. and Ann H.S. (1999). ”Is Job Stability in the United States Falling?”.Journal of Labor Economics, Vol. 17 (1999), Issue 4, pp. S1-28.

Fahrmeir,L and Raach ,A. (2007). ”A Bayesian Semiparametric Latent Variable Model for Mixed Responses”. Psychometrika, Vol.72, No. 3, 327-346.

Fahrmeir, L. and S. Lang (2001). ”Bayesian semiparametric regression analysis of multi-categorical time-space data”. Annals of the Institute of Statistical Math-ematics, 53, 10-30.

Fawcett,T. (2006). ”An introduction to ROC analysis”. Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition archive Volume 27 Issue 8.

Feldstein M .(2008). ”Did wages reflect growth in productivity?”. Journal of Policy Modeling, 30 (4), 591-594.

James N. B. and Audrey L. (1992). ”Interpreting Panel Data on Job Tenure”.Journal of Labor Economics, Vol. 10, No. 3 (Jul., 1992), pp. 219-257.

Harrison P (2009), ”Median Wages and Productivity Growth in Canada and the United States”. Center for the Study of Living Standards Research Note, 2009-2, URL http://www.csls.ca/no
2.pdf

Horowitz, J.L. and N.E. Savin (2004). ”Binary Response Models: Logits, Probits and Semiparametrics”. Journal of Economic Perspectives-Volume 15, No. 4, Fall 2001, 43-56.

Hutchens R. (1989). ”Seniority, Wages and Productivity: A Turbulent Decade”. The Journal of Economic Perspectives, Vol. 3, No. 4 (Autumn, 1989), pp. 49-64.

Kottas, A. and A.E. Gelfand (2001). "Bayesian Semiparametric Median Regression Modeling". Journal of the American Statistical Association. December 1, 2001, 96(456): 1458-1468.

Klein, R. and R. Spady (1993). "An efficient semi-parametric estimator for the binary response model". Econometrica, 61, 387-421.

Klein, R. and F. Vella (2009). "A semi-parametric model for binary response and continuous outcomes under index heterogeneity". Journal of Applied Econo-metrics, 24, 735-762

Lawrence M., Jared B., Sylvia A. (2006). The State of Working America, 2006/2007. ILR Press; 10 edition (December 20, 2006).

Li, S. (2011). "Combining MATLAB with C/C++ and graphic processing unit: an example of matrix multiplication and kernel density estimation". Chapter 2 of an unpublished dissertation, Rutgers University.

Marron, J.S. and M.P. Wand (1992). "Exact mean integrated square errors". Annals of Statistics, 20 (2), 712-736.

McNeil, B. and J. Hanley (1984). "Statistical approaches to the analysis of receiver operating characteristic (ROC) curves". Med Decis Making. 1984;4(2):137-50.

Nelder, J. ; R. Wedderburn (1972). "Generalized linear models". Journal of the Royal Statistical Society Series A (Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3) 135 (3): 370-384.

Newton, C. and M.Chappell (1996). "Bayesian inference for semiparametric binary regression". Journal of the American Statistical Association, 91 (433), 142- 153.

Silverman, B. W. (1996). Density estimation for statistics and data analysis, Chap-man & Hall, New York.

Swets, J.A., Dawes, R.M., Monahan, J. (2000). "Better decisions through science". Scientific American 283, 82-87.

Wand, M.P. and M.C.Jones (1995). Kernel smoothing, Chapman & Hall, New York.

Zhang, Silvapulley and Papaspiruiz (2009) "Bayesian estimation of bandwid-dth and parameters for the semi-parametric single index binary response model". mimeograph .

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

	Strongly skewed Y=1 is 66.7%		**Sparated bimodal Y=1 is 21.4%**		**Claw Y=1 is 21.4%**	
	ROC area	**MSE**	**ROC area**	**MSE**	**ROC area**	**MSE**
	Bayes		**Bayes**		**Bayes**	
probit	0.67	0.2	0.76	0.15	0.88	0.04
logit	0.67	0.2	0.76	0.15	0.88	0.04
semi	0.67	0.19	0.75	0.14	0.88	0.03
semi-opt	0.67	0.19	0.76	0.14	0.88	0.03
	MLE		**MLE**		**MLE**	
probit	0.67	0.2	0.76	0.14	0.88	0.04
logit	0.27	0.2	0.76	0.15	0.88	0.04
semi	0.67	0.19	0.76	0.14	0.88	0.04
semi-opt	0.67	0.19	0.76	0.14	0.88	0.03

Notes: semi = semi parametric with the bandwidth h in equation (9)
semi-opt = semi prametric with he optimal , $h^*_{optimal}$ in equation (10)

Table 3: ROC areas and MSE's: Extremely unbalanced Cases

	Skewed logitic (θ = .1) Y=1 is 2.5%		**Outlier Y=1 is .6%**		**Kurtotic unimodal Y=1 is 1.5%**	
	ROC area	**MSE**	**ROC area**	**MSE**	**ROC area**	**MSE**
	Bayes		**Bayes**		**Bayes**	
probit	.96	.02	.85	.01	.96	.04
logit	.96	.02	.85	.01	.95	.04
semi	.96	.02	.87	.01	.98	.01
semi-opt	.96	.02	.87	.01	.98	.01
	MLE		**MLE**		**MLE**	
probit	NC		NC		NC	
logit	NC		NC		NC	
semi	NC		NC		NC	
semi-opt	NC		NC		NC	

Notes: semi = semi parametric with the bandwidth h in equation (9)
semi-opt = semi prametric with he optimal , $h^*_{optimal}$ in equation (10)
NC = not converged

Table 4: Monte Carlo Experiment Result

Claw distribution, around 20% of 'Y=1' Replications = 500							
Binary Model Evaluation Criteria						Marginal Effect	
ROC area	**MEAN**	**STD.**	**MSE**	**MEAN**	**STD.**	**MEAN**	**STD.**
Bayes Probit	0.8357	0.0283	Bayes Probit	0.0370	0.0042	0.0702	0.0121
Bayes Logit	0.8356	0.0284	Bayes Logit	0.0370	0.0042	0.0739	0.0122
Bayes Semi	0.8387	0.0270	Bayes Semi	0.0365	0.0042	0.0519	0.0084
Bayes Semiopt	0.8384	0.0273	Bayes Semiopt	0.0365	0.0042	0.0516	0.0086
MLE Probit	0.8357	0.0283	MLE Probit	0.0370	0.0042	0.0704	0.0121
MLE Logit	0.8355	0.0283	MLE Logit	0.0370	0.0042	0.0739	0.0122
MLE Semi	0.8428	0.0261	MLE Semi	0.0363	0.0042	0.0576	0.0103
MLE Semiopt	0.8427	0.0258	MLE Semiopt	0.0364	0.0042	0.0552	0.0113
Optimal Bandwidth							
	MEAN	**STD.**		**MEAN**	**STD.**		
BayesSemi optimal bandwidth	0.3579	0.0198	MLE Semi optimal bandwidth	0.3323	0.0480		
Skewed log alpha=0.25, around 2.5% of 'Y=1' Replications = 500							
Binary Model Evaluation Criteria						Marginal Effect	
ROC area	**MEAN**	**STD.**	**MSE**	**MEAN**	**STD.**	**MEAN**	**STD.**
Bayes Probit	0.9790	0.0075	Bayes Probit	0.0108	0.0019	0.0890	0.0154
Bayes Logit	0.9788	0.0076	Bayes Logit	0.0108	0.0020	0.0907	0.0151
Bayes Semi	0.9788	0.0071	Bayes Semi	0.0109	0.0019	0.0752	0.0153
Bayes Semiopt	0.9789	0.0071	Bayes Semiopt	0.0109	0.0019	0.0712	0.0144
BayesSemi optimal bandwidth	0.3225	0.0019					

Table 5: Summary Statistics of the 2005 PSID data

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

Notes: Q1 and Q3 are 25 and 75 percentiles, respectively.

Table 6: Estimations for the 2005 PSID data (**Y=1.0 is 6.17%**)

	variable	**MLE**		**Bayes**[1]		**Semi (θ)**[2]	
		Logistic	**Probit**	**Logistic**	**Probit**	**MLE**	**Bayes**
c	Constant	0.4148 0.4150	-0.0183 0.2097	0.3866 0.4254	-0.0303 0.2094	/	/
β_1	Age	-0.0118 0.0075	-0.0054 0.0036	-0.01211 0.0074	-0.0050 0.00368	/	/
β_2	Racial05	-0.1464 0.1363	-0.0657 0.0652	-0.1369 0.1343	-0.0633 0.0661	7.4515 5.6099	5.3118 8.9994
β_3	Sex	-0.4312 0.1802	-0.2121 0.08715	-0.4379 0.1908	-0.2012 0.0897	40.3179 13.762	44.8558 10.468
β_4	Educ05	-0.1487 0.0250	-0.0723 0.0127	-0.1471 0.0245	-0.0722 0.0131	22.6273 6.7067	14.8074 1.8031
β_5	Marital05	-0.4441 0.1759	-0.2098 0.0829	-0.4282 0.1878	-0.2219 0.0835	42.6703 14.3294	34.9547 9.4151
β_6	Citysize05	-0.0457 0.0376	-0.0261 0.0180	-0.0488 0.0380	-0.0276 0.0179	-2.478 1.5281	3.18461 2.5144
β_7	YRWORK05	0.0023 0.0102	0.0018 0.0048	0.0028 0.0099	0.0020 0.0049	1.6060 0.7997	-0.6088 0.4881
	MSE **ROC area**	0.0566 0.6641	0.0566 0.6646	0.0567 0.6643	0.0567 0.6645	0.0557 0.6875	0.0565 0.6711

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

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

Notes: *std*=standard deviation.

Figure 1: Various distributions and kernel densities using the usual bandwidth

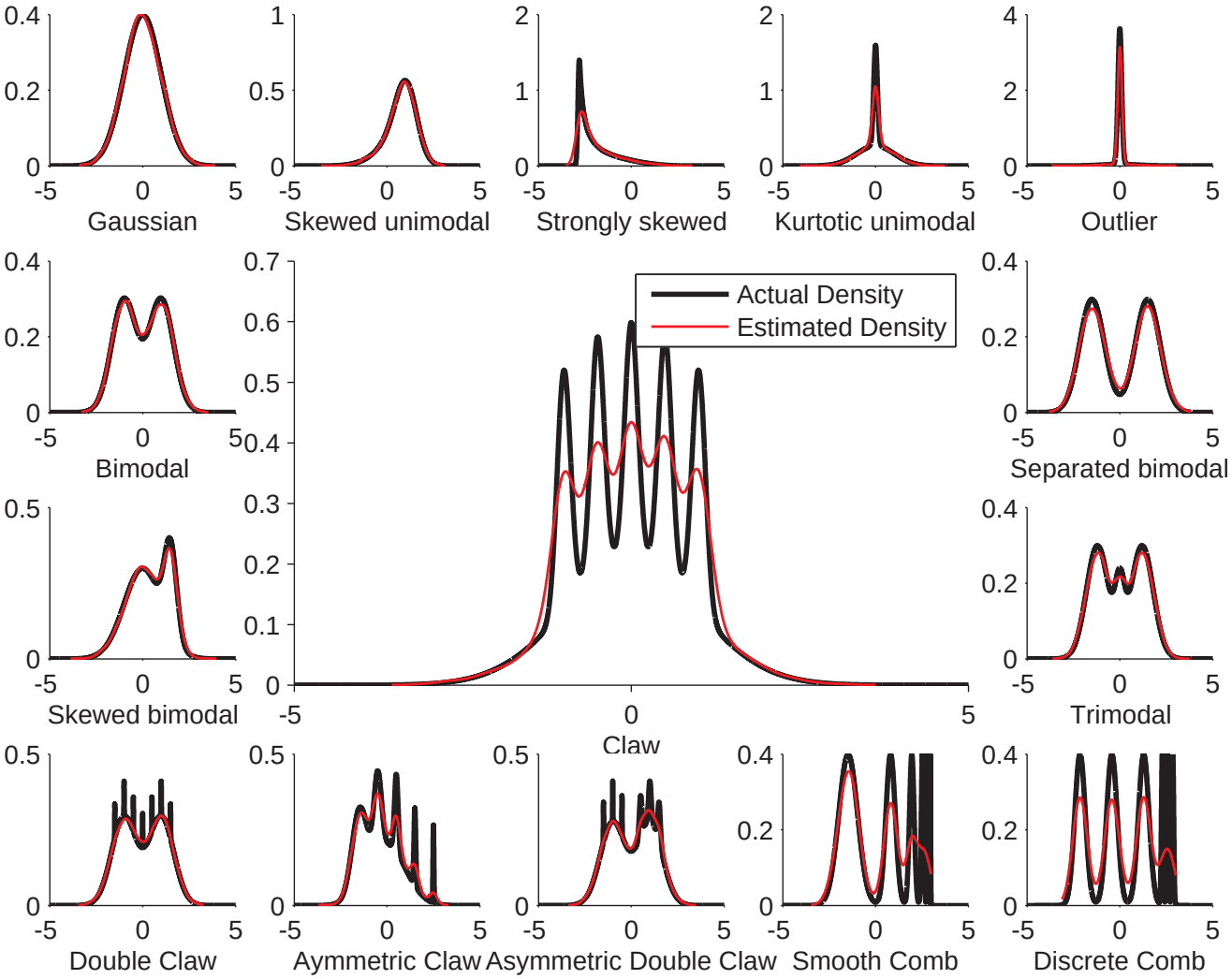


Figure 2: Various distributions and kernel densities using the optimal bandwidth

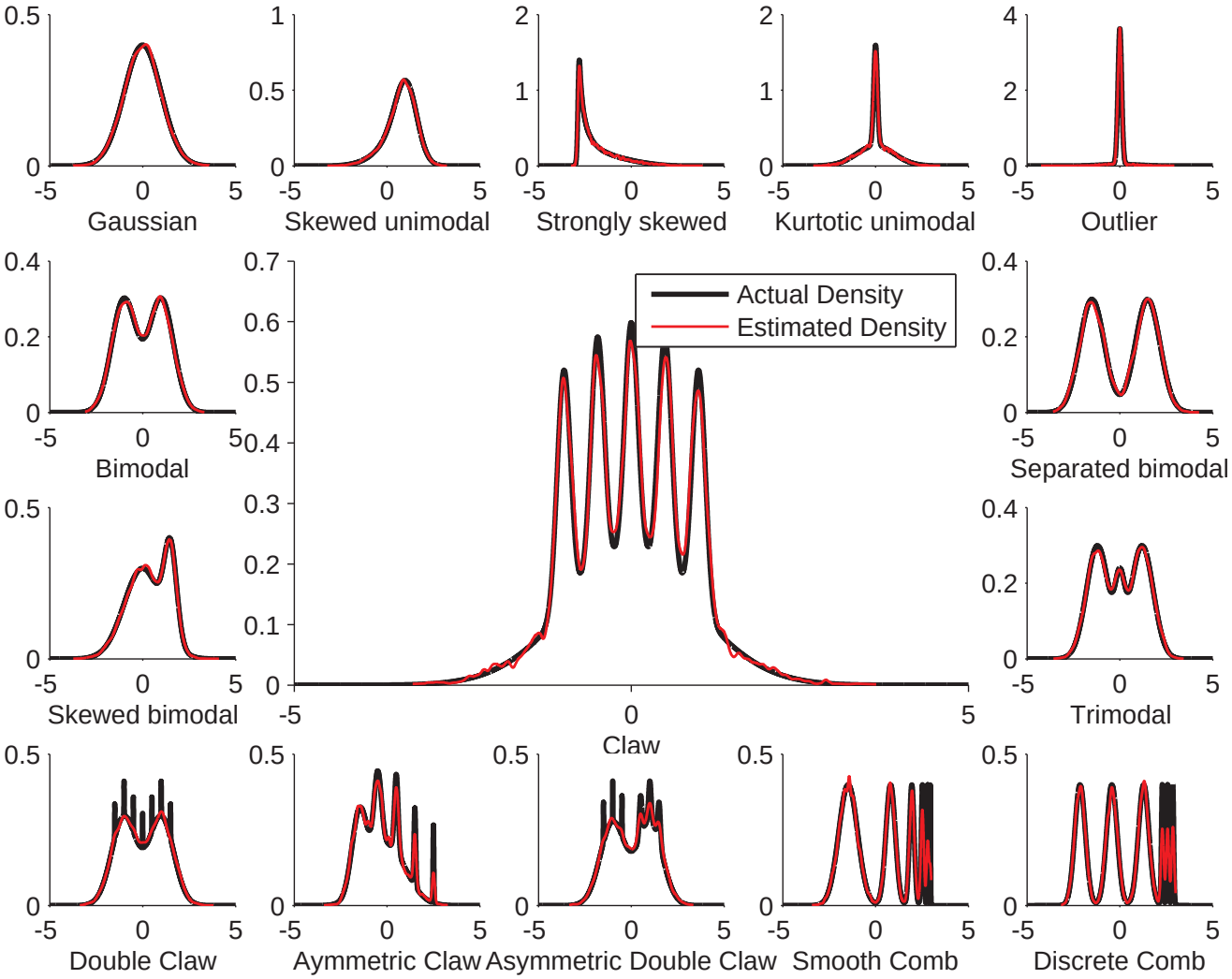
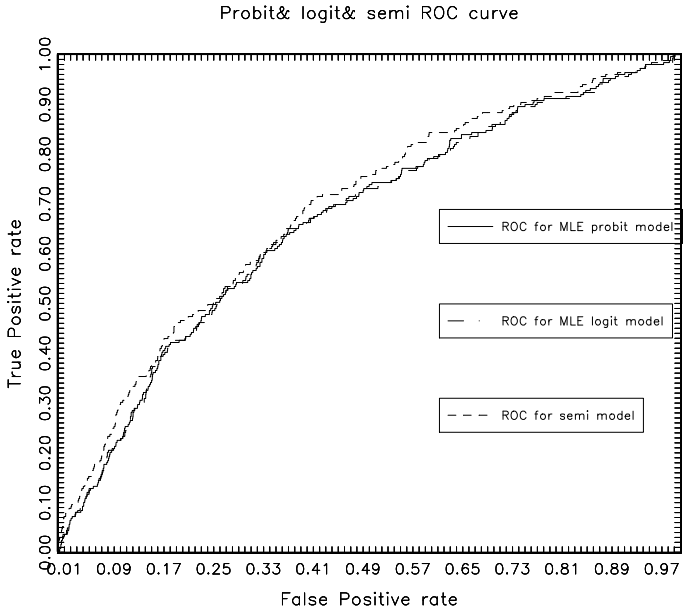


Figure 3: ROC curve analysis for the PSID data (4034 observations)



The ROC curves above show that at the same level of the 'False Positive Rate' ($P(\hat{y}=1 \mid y=0)$), Semiparametric model gives a greater value of the 'True positive rate'($P(\hat{y}=1 \mid y=1)$).

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 π_0 from the following classification table:

	Prediction, set π_0=0.64	
Actual	$\hat{y}=1$	$\hat{y}=0$
y= 1	a	b
y = 0	c	d

The predicted $\hat{y}$'s in this table get from setting π_0=0.64 in the example.

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

This is only for one point on the ROC curve according to one cut-off π_0 . For each model, we can calculate N points according to N values of π_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.