行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※

※ 含闕失數據下廣義線性模型的研究（2/2） ※

※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　□整合型計畫

計畫編號：NSC89－2118－M－002－013－

執行期間：　89 年 08 月 01 日至 90 年 09 月 30 日

計畫主持人：陳宏教授

共同主持人：

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：台灣大學數學系

中　華　民　國　90　年　1　月　31　日

Kernel Weighted GMM Estimation for Continuous-Time Interest Rate Models

Hong Chen, Mao-wei Hung and Chiou-ming Shiau *

July 2001

Abstract

In this paper, we adopt the kernel weighted GMM approach to estimate the parameters of drift and diffusion for continuous-time interest rate models. In this framework, we find that our estimates have typical large-sample properties, such as consistency, asymptotic normality, and robustness. Moreover, after using the local linear regression smoother, we can see that at low and high levels of interest rates the implied volatility is higher than that at mid-level interest rates. This supports the evidence for a volatility smile.

*Chen is from the College of Science, National Taiwan University, Taipei, Taiwan. Hung and Shiau are from the College of Management, National Taiwan University, Taipei, Taiwan. Correspondence to: Professor Mao-wei Hung, College of Management, National Taiwan University, 50, Lane 144, Keelung Road, Section 4, Taipei, Taiwan. Email: hung@mba.ntu.edu.tw.

1 Introduction

In the financial field, we usually use some stochastic processes to study the behavior of interest rates. Many models have been proposed in the literature. These models use diffusion processes to describe the dynamic of interest rates. In other words, if $r(t)$ denotes the instantaneous interest rate at time t, and it follows a diffusion process, then by Itô's formula, we have the following stochastic differential equation (SDE)

$$dr(t) = \mu(t, r(t))dt + \sigma(t, r(t))dB(t), \tag{1}$$

where, $\{B(t), t \geq 0\}$ is standard Brownian motion; and $\mu(\cdot, \cdot)$ is called the drift parameter of the process or instantaneous mean; $\sigma(\cdot, \cdot)$ is called the diffusion parameter of the process or instantaneous variance. Under such specifications, we can deduce the term structure of interest rate and price interest rate derivatives.

In the literature, many econometric methods have been developed to estimate equation (1). The maximum likelihood estimation (MLE) and generalized method of moments (GMM) are two methods which are most commonly used. For example, Chen and Scott (1993), Pearson and Sun (1994) use the maximum likelihood estimation. On the other hand, Chan, Karolyi, Longstaff and Sanders (1992) use the generalized method of moments. However, these two methods have some limitations. Specifically, we have to make appropriate assumptions when we use these methods. For example, in using the MLE, we assume that the likelihood function of the change of interest rate is known, but, unfortunately, this is impossible. There are also misspecification problems with the likelihood function. White (1982) studies the misspecification of likelihood and introduces a method to correct this problem. Moreover, the pseudo maximum likelihood estimation is used in Gourierous, Monfort and Trognon (1984) to correct this error.

Although the GMM is more robust than the MLE, it still has some problems. In a recent paper, Eom (1998) shows that if the moment conditions are not exact, then the estimates are wrong when using the GMM. Moreover, if too many moment conditions (overidentification) or too few (underidentification) are used, it will result in no solution (infinitely many solutions). Since the estimation of parameters is sensitive to moment conditions, the GMM is not a robust estimator. Facing such a problem, Gallant and Tauchen (1995) and Andersen and Lund (1997) propose the efficient method of moment method (EMM) to solve this problem.

In our paper, we assume that

$$\mu(t, r(t)) = \mu(r(t)) \quad \text{and} \quad \sigma(t, r(t)) = \sigma(r(t)) \tag{2}$$

Table 1 lists some parametric models which have been proposed in the literature.

Because it is impossible to guess the true model of interest rate, some approach of nonparametric models or semi-parametric models have been proposed in the literature. Table 2 is a list of some nonparametric models.

Although nonparametric models are more robust than parametric models, nonparametric model estimations cannot catch the features of serial correlation or persistence. In contrast, the GMM will do a much better job than nonparametric models in this case. In addition, since the kernel estimate is a special form of weighted scheme, which counts both sides in a bandwidth of each set of data. Data near the boundary only has weight on one side and this lacks weight on the other. It will result in estimation bias, is the so-called boundary effect. Therefore, to reduce the boundary effect we have to correct the specifications of the kernel function. Table 3 represents some simulating results by comparing the parametric models estimation and the nonparametric models estimation. It says that the nonparametric model estimation is more robust than the parametric model estimation. Morever, under suitable assumptions, we may find that such a correction does,in fact, reduce the boundary effect. The main argument is based on Hansen and Scheinkman (1995), in which they inset a kernel function into each moment condition. Using this method, we can estimate the drift and diffusion in equation (1), and discuss the properties of their estimations, respectively.

The remainder of the paper is organized as follows. Section 2 presents the nonparametric estimation method incorporating the boundary effect. Section 3 discusses how we implement the kernel weighted GMM method. The empirical results are given in Section 4. The last section concludes the paper.

2 Nonparametric Estimation and the Boundary Effect

2.1 Identification

Generally, the interest rate is thought of as a continuous-time process, however, on observation, only discrete data can be accessed. Weekly, daily, even hourly observations are available, but they are not continuous- time observations. Therefore, when we use this discrete-time data to approach the continuous-time interest rate, there will be some bias. This will result in the ability to catch some features of the data. Unfortunately, we cannot abbreviate the time interval when we observe it. If we could, there would be two advantages: one is that the effect of the limitation would be enhanced. Second, we could increase the number of observations.

Unfortunately, abbreviating the observed time interval will cause the market micro-structure problem[1]. As a result, if we want to increase the number of observations but not abbreviate the time interval observed, the only thing we can do is to increase the amount of time observed.

On the identification of parameters, Aït-Sahalia (1996a) proposed that $\left(\mu, \sigma^2\right)'$ and $\left(a\mu, a\sigma^2\right)'$, $a \neq 0$ are distinguishable if we do not put any restrictions on the parameters. Given this, while including the information of the interest rate observed data, he also made some specification on the parameters. He assumed that $\mu(r) = \beta\cdot(\alpha - r)$, and the diffusion was no restriction. Given this, the model of Aït-Sahalia is given by

$$dr(t) = \beta\cdot(\alpha - r(t))dt + \sigma(r(t))dB(t) \tag{3}$$

Under this specification, we may estimate the drift and diffusion. In the next section, we will introduce the estimation of Aït-Sahalia.

2.2 Estimation of Drift Parameter

According to equation (3) and Klenbaner (1998), we have

$$\begin{aligned} r(t) &= \exp(-\beta t)\cdot\left\{r(0) + \int_0^t \alpha\beta\exp(\beta s)\, ds + \int_0^t \exp(\beta s)\sigma(r(s))\, dB(s)\right\} \\ &= \alpha + (r(0) - \alpha)\exp(-\beta t) + \int_0^t \exp\left(-\beta(t-s)\right)\sigma(r(s))\, dB(s) \\ &= \alpha + (r(t-\triangle) - \alpha)\exp(-\beta\triangle) \\ &\quad + \int_{t-\triangle}^t \exp\left(-\beta(t-s)\right)\sigma(r(s))\, dB(s), \end{aligned} \tag{4}$$

such that, $\forall\ \triangle > 0$,

$$E\left[r(t)\ \mid r(t-\triangle)\right] = \alpha + (r(t-\triangle) - \alpha)\exp(-\beta\triangle) \tag{5}$$

Hence, when we use the AR(1) model: $r(t) = \delta_0 + \rho\cdot r(t-\triangle) + \varepsilon_t$, we can get $\hat{\delta}_0$ and $\hat{\rho}$. Now use the one-to-one mapping of $(\delta_0, \rho)'$ and $(\alpha, \beta)'$, and we have, $\hat{\alpha}$ and $\hat{\beta}$. Then

[1] In Aït-Sahalia(1996a), he proposed the market micro-structure problem including the bid-ask spread, the discreteness of the prices observed and the irregulariy of the intra-day sampling interval three kinds to affect the estimation of parameter.

the estimator of drift parameter is obtained. The relation is given by

$$\begin{aligned}\hat{\alpha} &= \frac{\hat{\delta}_0}{(1-\hat{\rho})}, && (6)\\ \hat{\beta} &= -\frac{1}{\Delta}\cdot\ln(\hat{\rho}), && (7)\\ \hat{\mu}(r) &= \hat{\beta}\cdot(\hat{\alpha}-r)\,; && (8)\end{aligned}$$

such that, we can indirectly find the estimator of drift by using the estimator $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}\right)'$ of $\theta = (\alpha, \beta)'$.

2.3 Estimation of Diffusion Parameters

Since the diffusion is hidden behind the information in the data, it is unobservable, so we cannot use the observed data to estimate the diffusion parameter. This means that we have to use another method of estimation or use some proxy. Later, we will introduce a feasible method to estimate the diffusion. This method uses the marginal probability density function and the relation of drift and diffusion.

2.3.1 Transition p.d.f. and Marginal p.d.f.

Let $\mathcal{P}\left(t, \cdot \mid s, x\right)$ denote the transition probability density function of a stochastic process $\{\ r(t), t \geq 0\ \}$ from $r(s) = x$ to time $t (t > s)$, and $\pi^t(\cdot)$ denote the marginal probability density function of it at time t, giving us the following relation

$$\int_{\mathbb{R}} \mathcal{P}\left(t, y \mid s, x\right)\cdot\pi^s(x)\ dx = \pi^t(y)\ ,\ \forall y \in \mathbb{R} \qquad (9)$$

Consider the Kolmogorov Forward Equation

$$\frac{\partial}{\partial t}\mathcal{P}\left(t, y \mid s, x\right) = -\frac{\partial}{\partial y}\left[\mu(y;\theta)\cdot\mathcal{P}\left(t, y \mid s, x\right)\right] + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y)\cdot\mathcal{P}\left(t, y \mid s, x\right)\right]. \qquad (10)$$

Combining these two equations, we have

$$\begin{aligned}\frac{d}{dt}\pi^t(y) &= \frac{d}{dt}\int_{\mathbb{R}} \mathcal{P}\left(t, y \mid s, x\right)\cdot\pi^s(x)\ dx\\ &= \int_{\mathbb{R}} \frac{\partial}{\partial t}\left[\mathcal{P}\left(t, y \mid s, x\right)\cdot\pi^s(x)\right]\ dx\\ &= \int_{\mathbb{R}}\left\{-\frac{\partial}{\partial y}\left[\mu(y;\theta)\cdot\mathcal{P}\left(t, y \mid s, x\right)\right]\right.\end{aligned}$$

$$
\begin{aligned}
& +\frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y)\cdot\mathcal{P}(t,y \mid s,x)\right]\Big\}\cdot\pi^s(x)\ dx \\
= & \int_{\mathbb{R}} -\frac{\partial}{\partial y}\left[\mu(y;\theta)\cdot\mathcal{P}(t,y \mid s,x)\cdot\pi^s(x)\right]\ dx \\
& +\frac{1}{2}\int_{\mathbb{R}}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y)\cdot\mathcal{P}(t,y \mid s,x)\cdot\pi^s(x)\right]\ dx \\
= & -\frac{\partial}{\partial y}\left[\mu(y;\theta)\cdot\int_{\mathbb{R}}\mathcal{P}(t,y \mid s,x)\cdot\pi^s(x)\ dx\right] \\
& +\frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y)\cdot\int_{\mathbb{R}}\mathcal{P}(t,y \mid s,x)\cdot\pi^s(x)\ dx\right] \\
= & -\frac{\partial}{\partial y}\left[\mu(y;\theta)\cdot\pi^t(y)\right]+\frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y)\cdot\pi^t(y)\right]. \qquad (11)
\end{aligned}
$$

And, supposing that the transition probability density function is stationary, there then exists a $\pi(\cdot)$, such that

$$\pi^t(y) = \pi(y)\ ,\ \forall y \in \mathbb{R}\ ,\ \forall t \geq 0. \qquad (12)$$

Therefore,

$$\frac{d}{dt}\pi^t(y) = 0\ ,\ \forall y \in \mathbb{R}. \qquad (13)$$

So that,

$$
\begin{aligned}
& \frac{\partial}{\partial y}\left[\mu(y;\theta)\cdot\pi(y)\right] = \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[\sigma^2(y)\cdot\pi(y)\right] \\
\Longrightarrow\ & \sigma^2(y) = \frac{2}{\pi(y)}\cdot\int_{(-\infty,y]}\mu(\nu;\theta)\cdot\pi(\nu)\ d\nu\ ,\ \forall y \in \mathbb{R}. \qquad (14)
\end{aligned}
$$

2.3.2 Find the Estimator of Diffusion by Marginal p.d.f.

Above all, if we know the form of the marginal probability density function and the estimator of drift, we can easily find the estimator of diffusion. It is shown as follows:

$$\hat{\sigma}^2(y) = \frac{2}{\pi(y)}\cdot\int_{(-\infty,y]}\mu(\nu;\hat{\theta})\cdot\pi(\nu)\ d\nu\ ,\ \forall y \in \mathbb{R}, \qquad (15)$$

where, $\pi(\cdot)$ is a known probability density function.

However, we do not usually know the form of $\pi(\cdot)$, so we must use some estimator to replace it. This time, nonparametric estimation will offer us a strong tool to help us to find it. Following Cheng and Li (1978), we have

$$\hat{f}_{\lambda_T}(r) \overset{\text{def}}{=} \frac{1}{T\cdot\lambda_T}\sum_{t=1}^{T}\mathcal{K}\left(\frac{r-r(t)}{\lambda_T}\right), \qquad (16)$$

where, $\{r(t)\}_{t=1}^{T}$ is the data; T is the number of observations; $\mathcal{K}(\cdot)$ is the kernel function; λ_T is the bandwidth of $\mathcal{K}$. Generally speaking, the nonpararmetric estimation of the probability of each point is of special weighted form. The shorter the data ,the more weight the data have. And the weight is counted by the kernel function. In Härdle (1989), we can conclude the following five properties of the kernel function:

1. $\text{Support}\,(\mathcal{K}) = [-1, 1]$.
2. $\int_{-1}^{1} \mathcal{K}(u)\; du = 1$.
3. $\int_{-1}^{1} u\cdot\mathcal{K}(u)\; du = 0$.
4. $\int_{-1}^{1} u^2\cdot\mathcal{K}(u)\; du = \eta \neq 0$.
5. $\mathcal{K} \in \mathcal{C}^1\left([-1, 1]\right)$.

In the empirical discussion, we will use the following kernel function to estimate the marginal probability density function.

$$\begin{aligned} \mathcal{K}(u) &= \frac{2}{\pi}\cdot\frac{1}{1+u^2}\cdot\mathbf{1}_{[-1,1]}(u) \\ &= \begin{cases} \dfrac{2}{\pi}\cdot\dfrac{1}{1+u^2} & ,\text{if } \mid u \mid \leq 1; \\ 0 & ,\text{if } \mid u \mid > 1. \end{cases} \end{aligned} \tag{17}$$

When we replace $\pi(\cdot)$ for the above kernel function, we have

$$\hat{\sigma}^2_{\lambda_T}(y) = \frac{2}{\hat{f}_{\lambda_T}(y)}\cdot\int_{(-\infty,y]} \mu(\nu;\hat{\theta})\cdot\hat{f}_{\lambda_T}(\nu)\; d\nu\;,\; \forall y \in I\!R. \tag{18}$$

2.3.3 Boundary Effect and Correction

As our estimator of the parameters are of the symmetric weighted form, we do not need to correct the kernel function if the boundary effect is not significant, or the near boundary data are negligible, because the correction is just of advantage in reducing the boundary effect. Moreover, since the computation of the boundary kernel function is more complex and difficult than on the original kernel function, there therefore must be a trade-off in selecting the kernel function when we decide to account the boundary effect in, or we may get an unreasonable outcome in pricing some securities.

In the above discussion, we have shown the importance of the marginal probability function in estimating diffusion parameters, and have pointed out the form of the kernel function. Here, the major argument is with regard to results of the boundary effect and correction. According to Härdle (1989), a boundary kernel function $\mathcal{K}_b(\cdot)$ must satisfy the following properties:

1. $\int_{-1}^{b} \mathcal{K}_b(u)\ du = 1.$

2. $\int_{-1}^{b} u\cdot\mathcal{K}_b(u)\ du = 0.$

3. $\int_{-1}^{b} u^2\cdot\mathcal{K}_b(u)\ du = \eta_b \neq 0.$

4. $\int_{-1}^{b} \{\mathcal{K}_b(u)\}^2\ du < +\infty.$

5. $\mathcal{K}_b \in \mathcal{C}^1\left([-1,b]\right).$

6. $\lim_{b\uparrow 1} \mathcal{K}_b(\cdot) = \mathcal{K}(\cdot).$

And in Section 2.3.2, the kernel function we adapted is

$$\mathcal{K}(u) = \frac{2}{\pi}\cdot\frac{1}{1+u^2}\cdot\mathbf{1}_{[-1,1]}(u),$$

such that, when we consider the boundary effect, its boundary kernel function $\mathcal{K}_b(\cdot)$ is given by

$$\mathcal{K}_b(u) = [h_1(b)\cdot u + h_2(b)\cdot\mathcal{K}(u)]\cdot\mathbf{1}_{[-1,1]}(u) \tag{19}$$

where, $h_1(b) = \dfrac{6\left[\ln 2 - \ln(1+b^2)\right]}{4(1+b^3)\left[\arctan(b)+\frac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]}$ and, $h_2(b) = \dfrac{2\pi(1+b^3)}{4(1+b^3)\left[\arctan(b)+\frac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]}$.

Proposition 1: $\mathcal{K}_b(\cdot)$ must satisfy the six properties in Härdle (1989).

For the proof, please see Appendix A.

In accordance with the above derivation, we may see that the original kernel function $\mathcal{K}(\cdot)$ can be replaced by $\mathcal{K}_b(\cdot)$ when the boundary effect is significant enough to reduce in a bias. In other words,

$$\hat{f}_{\lambda_T}(r;b) \overset{\text{def}}{=} \frac{1}{T\cdot\lambda_T}\sum_{t=1}^{T}\mathcal{K}_b\left(\frac{r-r(t)}{\lambda_T}\right) \tag{20}$$

$$\hat{\sigma}^2_{\lambda_T}(y;b) \quad = \quad \frac{2}{\hat{f}_{\lambda_T}(y;b)} \cdot \int_{(-\infty,y]} \mu(\nu;\hat{\theta}) \cdot \hat{f}_{\lambda_T}(\nu;b)\, d\nu \ , \ \forall y \in I\!R. \quad (21)$$

Figure 1 displays the effect of the correction of the boundary effect. Here, given $\mu(r;\theta) = 0.25 \cdot (0.055 - r)$, the boundary-effect is significant. Therefore, we can replace the original kernel function with the kernel function adjusted for the boundary effect.

3 Inserting the Kernel Weight to GMM

In Aït-Sahalia (1996a), we may find the semi-parametric estimates of both drift and diffusion, although the method used to get the estimate of diffusion is more difficult than that of drift. This method takes into account both the estimate of drift and the marginal probability density function[2], and the Kolmogorov Forward Equation and stationary assumption.

Next, we may ask whether we can find the nonparametric estimate of diffusion by using another method different from Aït-Sahalia (1996a). The answer is "yes!" In Stanton (1998), he uses the same kernel function to estimate both drift and diffusion, such that the estimate of these two parameters is of nonparametric form. However, the nonparametric estimation is more generalized than the parametric estimation, and it has faced much criticism, such as the selection of kernel function, the size of bandwidth, the convergence rate as it is affected by the dimension of variables, boundary effect, etc. In addition, when we apply the nonparametric estimation to time series data, for example, to variables of economics or finance, it is usually accused of having disregarded the persistent and serial correlation which are always important features of time series data.

As there are so much criticism for the implication of nonparametric estimation to time series data, should we abandon this method and go back to the parametric estimation to correct some specification on it to avoid the occurrence of the above problems? In fact, the ability of nonparametric estimation in catching the features of persistence and serial correlation of data is weaker than that of parametric estimation in nature. Fortunately, the moment conditions do a good job in catching the features. Given this, in order to compensate for the blots of nonparametric estimation, we propose a new estimation model: inset the kernel weight into the moment conditions. This model not only preserves the robustness of nonparametric estimation, but also has the advantage of moment conditions with respect to catching the features of persistence and serial correlation. Here, we will

[2] Regardless of whether or not the boundary effect is considered.

develop such a new estimation model and study the properties of estimates which this method provided.

3.1 Derivation of the Moment Conditions

In this section, we derive the moment conditions of the following specific interest rate process:

$$\begin{aligned}
&dr(t) = \mu\left(r(t);\vec{\theta}\right) dt + \sigma\left(r(t)\right) dB(t) \\
&\mu\left(r(t);\vec{\theta}\right) = \beta\cdot(\alpha - r(t)), \\
&\vec{\theta} = \begin{pmatrix}\alpha \\ \beta\end{pmatrix} \\
&\sigma^2\left(r(t)\right) = \sigma^2\cdot v\left(r(t)\right)
\end{aligned}$$

By equation (5), we have, $\forall t > \Delta > 0$,

$$\begin{aligned}
& E[\ r(t) \mid r(t-\Delta)\ ] \\
= \;& \alpha + [r(t-\Delta) - \alpha]\exp(-\beta\Delta) \\
= \;& r(t-\Delta) + (1-\exp(-\beta\Delta))\cdot(\alpha - r(t-\Delta)) \\
= \;& \alpha\,[1-\exp(-\beta)] - (1-\exp(-\beta))\cdot r(t-\Delta), \qquad (22)
\end{aligned}$$

such that, by defining

$$\varepsilon(t) \overset{\text{def}}{=} r(t) - r(t-\Delta) - E[\ r(t) - r(t-\Delta) \mid r(t-\Delta)\ ], \qquad (23)$$

we can find that $E[\ \varepsilon(t) \mid r(r-\Delta)\ ] = 0$. Moreover, by using the iterated expectation rule, we have,

$$E[\ \varepsilon(t)\ \ ] = E[\ E[\ \varepsilon(t) \mid r(r-\Delta)\ ]\ ] = 0. \qquad (24)$$

So that, the first moment condition, C1, is given by

$$\sum_{j=0}^{T-1}\{[r(T-j) - r(T-j-1)] - (\alpha\cdot[1-\exp(-\beta\cdot j)] - [1-\exp(-\beta\cdot j)]\cdot r(T-j-1))\} = 0 \qquad (25)$$

Moreover, if we define $\Delta = 1$, then, since

$$E[\ r(t) - r(t-1) \mid r(t-1)\ ] = \alpha\cdot[1-\exp(-\beta)] + \exp(-\beta)\cdot r(t-1)$$

and by the iterated expectation rule, we know that

$$E[\ r(t)-r(t-1)\ ]=\alpha\cdot[1-\exp(-\beta)]+\exp(-\beta)\cdot E[\ r(t-1)\ ]\ ,\ t=1,2,3,\ldots,T. \quad (26)$$

Equation (26) shows us how to find the expected value of the interest rate at time t conditional to the $(t-1)$-information. By the iterative law, we have the following proposition:

Proposition 2:

$$E[\ r(t)-r(t-1)\ ]=\alpha\cdot[1-\exp(-\beta)]\,t+\exp(-\beta\cdot t)\cdot E[\ r(0)\ ]\ ,\ t=1,2,3,\ldots,T \quad (27)$$

Moreover, in Jiang (1998), when define the difference of interest rate as follows:

$$\Delta r_t \overset{\text{def}}{=} r(t)-r(t-\Delta)\ ,\ t>\Delta,$$

then $\forall\tau>\Delta>0$,

$$\mathrm{Cov}\left(\Delta r_t,\Delta r_{t-\tau}\right)=\exp\left(-\beta\cdot(\tau-\Delta)\right)[1-\exp(-\beta\Delta)]^2\sigma^2, \quad (28)$$

such that, we may show the following proposition

Proposition 3: If we suppose that $\Delta=1$, then a τ-lagged correlation coefficient is given by

$$\mathrm{Corr}\left(\Delta r_t,\Delta r_{t-\tau}\right)=-\frac{1}{2}\exp\left(-\beta\cdot(\tau-1)\right)[1-\exp(-\beta)]. \quad (29)$$

Therefore, by $r(t-1)=r(0)+\sum_{\tau=1}^{t-1}\Delta r_{t-\tau}$ we find that

$$\mathrm{Cov}\left(\Delta r_t,r(t-1)\right)=\sum_{\tau=1}^{t-1}\exp\left(-\beta\cdot(\tau-1)\right)[1-\exp(-\beta)]^2\sigma^2. \quad (30)$$

Hence, the second moment condition, C2, has been shown as follows:

$$\begin{aligned}&\sum_{j=0}^{T-\tau}\{[r(T-j)-r(T-j-1)]\cdot[r(T-j-1)-r(T-j-\tau)]\}\\ =\ &-\frac{1}{2}\exp\left(-\beta\cdot(\tau-1)\right)[1-\exp(-\beta)]\ ,\ \tau\geq 3. \quad (31)\end{aligned}$$

Giving these two moment conditions, we can get the estimates for α and β, respectively. To find an estimate for drift, you can substitute the estimates of α and β into the specification of drift. In other words, by solving the system of equations C1 and C2, we may find $\hat{\alpha}$ and $\hat{\beta}$, and substitute them into the specification of μ, such that,

$$\hat{\mu}_{\text{GMM}} = \mu(r;\hat{\theta}) = \hat{\beta}\cdot(\hat{\alpha} - r) \tag{32}$$

3.2 The Estimation of Moment Conditions After Insertion of Kernel Weight

Using the above derivation, we find not only the two moment conditions, but also find the estimate of drift parameters by solving the system of equations C1 and C2. However, since the estimates provided by the GMM are not robust estimator[3], we have to make some modifications to correct this problem.

First, we insert the kernel weight into each moment condition. And by the following equation

$$\begin{aligned} & \sum_{j=0}^{T-1} \{r(T-j) - \alpha\cdot[1-\exp(-\beta\cdot j] - \exp(-\beta\cdot j)\cdot r(T-j-1)\} \\ = & \sum_{j=0}^{T-1} \{[r(T-j) - r(T-j-1)] - \alpha\cdot[1-\exp(-\beta\cdot j] \\ & \qquad + [1-\exp(-\beta\cdot j)]\cdot r(T-j-1)\}\,. \end{aligned}$$

Hence, C1 can be replaced for C1' as follows:

$$\begin{aligned} 0 = & \ \textstyle\sum_{j=0}^{T-1} \ \left\{\left[r^{(T-j)} - r^{(T-j-1)}\right] - a(\theta)\cdot[1-\exp(-\beta\cdot j)]\cdot\left[\alpha - r^{(T-j-1)}\right]\right\} \\ \times & \quad \frac{1}{T\cdot\lambda}\int_{s(T-j-1)}^{s(T-j)} \mathcal{K}\left(\frac{u - r^{(T-j-1)}}{\lambda}\right)\, du \end{aligned} \tag{35}$$

where, $s(t) \overset{\text{def}}{=} \dfrac{r^{(t)} + r^{(t-1)}}{2}$, $t = 1, 2, 3, \cdots, T$, and $r^{(t)} \overset{\text{def}}{=}$ the t-th ordered statistic of $r(1), r(2), r(3), \cdots, r(T)$.

In C1', we put a nonrandomized function $a(\theta)$. This makes the estimation more robust and comprehensive by taking on a special form. For example,

1. When we take $a(\theta) = 1$, and $\mathcal{K}(u) = \dfrac{1}{T}$, then $\mathbf{C1'} \equiv \mathbf{C1}$, and our estimates will be identical to that of the GMM estimation.

[3] It is sensitive to what the moment conditions are and the number of moment conditions.

2. When we take $a(\theta) = 0$, and define

$$\hat{\mu}_{\mathcal{K}} \overset{\text{def}}{=} \frac{1}{T\cdot\lambda}\sum_{t=0}^{T-1}\left[r^{(t+1)} - r^{(t)}\right] \times \int_{s(t)}^{s(t+1)} \mathcal{K}\left(\frac{u - r^{(t)}}{\lambda}\right)\, du, \tag{34}$$

then **C1**$'$ gives us estimates just the same as in Stanton (1998). which are of nonparametric form.

Hence, we may change C1' into C1"

$$\begin{aligned} 0 = \hat{\mu}_{\mathcal{K}} \quad - & \sum_{j=0}^{T-1}\left\{a(\theta)\cdot[1 - \exp(-\beta\cdot j)]\cdot\left[\alpha - r^{(T-j-1)}\right]\right\} \\ & \times \frac{1}{T\cdot\lambda}\int_{s(T-j-1)}^{s(T-j)} \mathcal{K}\left(\frac{u - r^{(T-j-1)}}{\lambda}\right)\, du. \end{aligned} \tag{35}$$

Then, by solving the system of equations C1" and C2 and adjusting the value of $a\left(\tilde{\theta}_{\mathcal{K}}\right)$, we may obtain the optimal estimating result.

Under such an estimation, we may find the estimate of drift

$$\tilde{\mu}_{\mathcal{K}} = \mu\left(r; \tilde{\theta}_{\mathcal{K}}\right), \tag{36}$$

and by equation(16) and (18)[4], we can find the estimate of diffusion

$$\tilde{\sigma}_{\mathcal{K}} = \sigma\left(r; \tilde{\theta}_{\mathcal{K}}\right), \tag{37}$$

Also, from Härdle (1990), we have the following proposition,

Proposition 4: $\tilde{\sigma}_{\mathcal{K}}$ has the following large-sample properties:

1. $\mathcal{P}\lim_{n\to+\infty} \tilde{\sigma}^2_{\mathcal{K}} = \sigma^2$, in other words, it is consistent with the true diffusion parameter.
2. $\sqrt{n}\cdot\left(\tilde{\sigma}^2_{\mathcal{K}} - \sigma^2\right) \overset{\text{asymp.}}{\sim} \mathcal{N}(0, \Sigma)$, in other words, it is asymptotic normal with the asymptotic variance Σ.

The proof of Proposition 4 is shown in Appendix B.

[4] Or, you may use the equation(20) and (21) which have considered the boundary effect.

4 Empirical Results and Analysis

In our research, the data used are the U.S. FEDERAL FUNDS (EFFECTIVE) -MIDDLE RATE from 1/1/1990 to 25/8/1999. We compute and list the descriptive statistics of the data in Table 5, such as mean, variance, skewness, kurtosis, etc.

Before discussing the empirical results of our study, we solve the system of equations C1" and C2 to find the estimate of drift parameters. Also, we use equation (18) or (21) to find the parameter of diffusion, so that, in the stochastic differential equation of interest rates

$$dr(t) = \hat{\mu}\,(r(t)) + \hat{\sigma}\,(r(t))\, d\mathcal{B}(t)$$

we may find the term structure of interest rates.

4.1 Marginal Density Estimation

Figure 2 is the graph of the nonparametric estimate of the marginal probability density function. The real line is the curve of the line which accounts for the boundary effect; the dash line does not. This shows that the marginal probability density function is unimodal but asymmetric to the mean, and the boundary marginal probability density (solid line) has a fatter tail than the other.

4.2 Drift Estimation

Table 4 compares the estimates of drift parameter. The 2^{nd}-step FGLS estimates were described in Aït-Sahalia (1996a). The optimal bandwidth[5] is given by

$$h_n = \frac{c}{\ln(n)} \cdot n^{-1/(2r+1)} \approx 0.0369727$$

It is oblivious that the boundary effect is significant, so the estimate with the original kernel should be correct.

4.3 Diffusion Estimation

Figure 3 displays the graph of the estimate of diffusion parameter. As mentioned before, we must use the estimating results of the drift parameter and the marginal probability density function to estimate the diffusion parameter. Hence, its bias maybe larger than

[5] The admissible bandwidth choices are described in Aït-Sahalia (1996a) Assumption A5.

the parametric estimate if the specification of the diffusion parameter is correct. In Figure 3, we see that all the three curves have a "smile-like" U-shape. But, in Aït-Sahalia (1996a), he had supposed that

$$\lim_{r \downarrow 0} \sigma(r) = 0,$$

such that the graph of the estimate of diffusion using the model in Aït-Sahalia (1996a) will be more likely to pull the diffusion to 0 when the interest rate is low. This contradicts the "volatility smile" feature. Since at low levels of interest rates, the investors have no willingness to hold such a security, if they do not possess the security, they will not buy it. On the other hand, if he already holds a low level interest rate bond, he will short it, so the liquidation of a low level interest rate bond will be high. At mid-level interest rates, investors will hold securities for a longer time than in the case of low level interest rates, so the level of liquidation is lower than that at low of level interest rates. And the dashed line[6]is too flat to look like a line. In contrast, the real line[7] seems to be a U-shape. As the above discussion, this is result by the boundary effect.

5 Conclusion

In this paper, we find that the mean-reverting property of interest rates proposed in the Vasicek (1977), CIR (1985), Courtadon (1982), Chan (1992) and Duffie-Kan (1993), etc. is not evident. Locally, we can see when the interest rate goes up, the drift will pull it back to the trend level and conversely, when the interest rate goes down, the drift will pull it up to the trend level. Overall, we can say that the term structure of interest rate has a local mean-reverting property.

Secondly, in Figure 3, we see that the implied volatility is smile-shaped. At low levels of interest rates, the liquidation of Treasury bonds is higher than that at mid-level interest rates. Hence the implied volatility at low level interest rates is higher than that at mid-level interest rates. This is contradict to the result in Aït-Sahalia (1996a).

[6] The curve of the estimate of diffusion which is estimated by using the original kernel function.

[7] The curve of the estimate of diffusion which is estimated by using the boundary kernel function.

Appendix A—Proof of Proposition 1

Lemma 1: $\lim\limits_{b\uparrow 1} h_1(b) = 0.$

[proof:]

$$\begin{aligned}
\lim_{b\uparrow 1} h_1(b) &= \lim_{b\uparrow 1}\left\{\frac{6\,[\ln 2 - \ln(1+b^2)]}{4(1+b^3)\left[\arctan(b)+\frac{\pi}{4}\right] - 3(1-b^2)\,[\ln 2 - \ln(1+b^2)]}\right\} \\
&= \frac{\lim\limits_{b\uparrow 1}\left\{6\left[\ln 2 - \ln(1+b^2)\right]\right\}}{\lim\limits_{b\uparrow 1}\left\{4(1+b^3)\left[\arctan(b)+\frac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]\right\}} \\
&= \frac{0}{4\cdot\pi - 0} \\
&= 0
\end{aligned}$$

Lemma 2: $\lim\limits_{b\uparrow 1} h_2(b) = 1.$

[proof:]

$$\begin{aligned}
\lim_{b\uparrow 1} h_2(b) &= \lim_{b\uparrow 1}\left\{\frac{2\pi(1+b^3)}{4(1+b^3)\left[\arctan(b)+\frac{\pi}{4}\right] - 3(1-b^2)\,[\ln 2 - \ln(1+b^2)]}\right\} \\
&= \frac{\lim\limits_{b\uparrow 1}\left\{2\pi(1+b^3)\right\}}{\lim\limits_{b\uparrow 1}\left\{4(1+b^3)\left[\arctan(b)+\frac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]\right\}} \\
&= \frac{4\pi}{4\cdot 2\cdot\frac{\pi}{2} - 0} \\
&= 1
\end{aligned}$$

We now will verify the six properties described in Härdle (1989) of $\mathcal{K}_b(\cdot)$ one by one, so the Proposition 1 has been shown below.

1.

$$\begin{aligned}
\int_{-1}^{b} \mathcal{K}_b(u)\,du &= \int_{-1}^{b} [h_1(b)\cdot u + h_2(b)\cdot\mathcal{K}(u)]\;du \\
&= \int_{-1}^{b} [h_1(b)\cdot u]\;du + \int_{-1}^{b} [h_2(b)\cdot\mathcal{K}(u)]\;du \\
&= \frac{b^2-1}{2}\cdot h_1(b) + \frac{2}{\pi}h_2(b)\left[\arctan(b)+\frac{\pi}{4}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{6\left[\ln 2 - \ln(1+b^2)\right]\cdot\dfrac{b^2-1}{2}}{4(1+b^3)\left[\arctan(b)+\dfrac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]} \\
&\quad + \frac{2\pi(1+b^3)\cdot\dfrac{2}{\pi}\left[\arctan(b)+\dfrac{\pi}{4}\right]}{4(1+b^3)\left[\arctan(b)+\dfrac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]} \\
&= 1
\end{aligned}
$$

2.

$$
\begin{aligned}
\int_{-1}^{b} u\cdot\mathcal{K}_b(u)\,du &= \int_{-1}^{b} u\cdot\left[h_1(b)\cdot u + h_2(b)\cdot\mathcal{K}(u)\right]\,du \\
&= \int_{-1}^{b}\left[h_1(b)\cdot u^2\right]\,du + \int_{-1}^{b}\left[h_2(b)\cdot u\cdot\mathcal{K}(u)\right]\,du \\
&= \frac{1}{3}(b^3+1)h_1(b) + \frac{1}{\pi}h_2(b)\left[\ln(1+b^2) - \ln 2\right] \\
&= \frac{6\left[\ln 2 - \ln(1+b^2)\right]\cdot\dfrac{b^3+1}{3}}{4(1+b^3)\left[\arctan(b)+\dfrac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]} \\
&\quad + \frac{2\pi(1+b^3)\cdot\dfrac{1}{\pi}\left[\ln(1+b^2) - \ln 2\right]}{4(1+b^3)\left[\arctan(b)+\dfrac{\pi}{4}\right] - 3(1-b^2)\left[\ln 2 - \ln(1+b^2)\right]} \\
&= 0
\end{aligned}
$$

3.

$$
\begin{aligned}
\int_{-1}^{b} u^2\cdot\mathcal{K}_b(u)\,du &= \int_{-1}^{b} u^2\cdot\left[h_1(b)\cdot u + h_2(b)\cdot\mathcal{K}(u)\right]\,du \\
&= \int_{-1}^{b}\left[h_1(b)\cdot u^3\right]\,du + \int_{-1}^{b}\left[h_2(b)\cdot u^2\cdot\mathcal{K}(u)\right]\,du \\
&= \frac{1}{4}(b^4-1)h_1(b) + \frac{2}{\pi}h_2(b)\left\{(b+1) - \left[\arctan(b)+\frac{\pi}{4}\right]\right\} \\
&\overset{\text{def}}{=} \eta_b \ \ (\neq 0\ ,\ \forall\,|b|\le 1)
\end{aligned}
$$

4.

$$
\int_{-1}^{b}\{\mathcal{K}_b(u)\}^2\,du = \int_{-1}^{b}\left\{h_1^2(b)\cdot u^2 + 2h_1(b)h_2(b)u\mathcal{K}(u) + h_2^2(b)\mathcal{K}^2(u)\right\}\,du
$$

$$
\begin{aligned}
&\leq \frac{1}{3}(b^3+1)h_1^2(b) + 2h_1(b)h_2(b)\cdot\int_{-1}^{1} u\mathcal{K}(u)\ du \\
&\qquad + h_2^2(b)\cdot\int_{-1}^{1} \mathcal{K}^2(u)\ du \\
&= \frac{1}{3}(b^3+1)h_1^2(b) + h_2^2(b)\cdot\int_{-1}^{1} \mathcal{K}^2(u)\ du \\
&< +\infty \quad \text{(Since } \mathcal{K}(\cdot) \text{ is a kernel function.)}
\end{aligned}
$$

5. It is clear that $\mathcal{K}_b \in \mathcal{C}^1([-1,b])$ does hold by the definition of $\mathcal{K}_b(\cdot)$

6. To show $\lim_{b\uparrow 1}\mathcal{K}_b(\cdot) = \mathcal{K}(\cdot)$, we have to show $\lim_{b\uparrow 1}\ h_1(b) = 0$ and $\lim_{b\uparrow 1}\ h_2(b) = 1$ first. By Lemma 1 and Lemma 2 we see that this property does hold.

References

Aït-Sahalia Y., 1996a, Nonparametric Pricing of Interest Rate Derivative Securities, Econometrica, 64, 527-560.

Aït-Sahalia Y., 1996b, Testing Continuous-Time Models of the Spot Interest Rate, The Review of Financial Studies, 9, 385-426.

Aït-Sahalia Y., and A.W. Lo, 1998, Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices, Journal of Finance, Vol.LIII, 499-547.

Banon G., 1978, Noparametric Identification for Diffusion Process, S.I.A.M. Journal of Control and Optimization,16, 380-395.

Black F., and M. Schole, 1973, The Pricing of Options and Corporate Liabilities, Journal of Political Economy, 3, 133-155.

Brennan M. J.,and E. S. Schwartz, 1979, A Continuous-Time Approach to the Pricing Bond, Journal of Banking and Finance, 3, 133-155.

Brown S. J.,and P. H. Dybvig, 1986, The Empirical Implications of the Cox, Ingersoll, Ross Theory of the Term Structure of Interest Rates, Journal of Finance, 41, 617-630.

Chan K. C., G. A. Karolyi, F. A. Longstaff, and A.B. Sanders, 1992, An Empirical Comparison of Alternative Models of the Short-Term Interest Rate, Journal of Finance, 47, 1209-1227.

Chen R. R.,and L. Scott, 1993, Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates, Journal of Fixed Income, 3, 14-31.

Cheng K. C., and P. E. Lin, 1981, Nonparametric Estimation of a Regression Function, Z.Wahrsch verw. Gebiete, 57, 223-233.

Conley T. G.,L. P. Hansen,and W. F. Liu, 1997, Bootstrapping the Long Run, Macroeconomic Dynamics, 1, 297-311.

Constantinides G.M., 1992, A Theory of the Nominal Term Structure of Interest Rates, The Review of Financial Studies, 5, 531-552.

Courtadon G., 1982, The Pricing of Options on Default-Free Bonds, Journal of Financial and Quantitative Analysis, 17, 75-100.

Cox J. C., E. Ingersoll, and S. A. Ross, 1985, A Theory of the Term Structure of Interest Rates, 1985, Econometrica, 53, 385-407.

Dothan L. U., 1978, On the Term Structure of Interest Rates, Journal of Financial Economics, Vol.6, 59-69.

Duffie D., and R. Kan, 1993, A Yield Factor Model of Interest Rates, Stanford University Mimeo.

Eom Y. H., 1998, On Efficient GMM Estimation of Continuous-Time Asset Dynamics: Implications for the Term Structure of Interest Rates, Working paper, Federal Reserve Bank of New York.

Gallant A. R., and G. Tauchen, 1995, Estimation of Continuous-Time Models for Stocks Returns and Interest Rates, working paper, University of North Carolina at Chapel Hill.

_______, 1996, Which Moments to Match, Econometric Theory, 12, 657-681.

Gibbons M. R., and K. Ramaswamy, 1993, A Test of the Cox, Ingersoll and Ross model of the Term Structure, The Review of Financial Studies, 6, 619-658.

Gouriérous C., A. Monfort, and A. Trognon, 1984, Peudo Maximum Likelihood Methods: Theory, Econometrica, 52, 681-700.

Hansen L. P., 1982, Large Sample Properties of Generalized Method of Moments Estimators, Econometrica, 50, 1029-1054.

Hansen L. P., and J. A. Scheinkman, 1995, Back to the Future: Generating Moments Implications for Continuous Time Markov Process, Econometrica, 63, 767-804.

Härdle W., 1980, Applied Nonparametic Regression, Cambridge University Press,Cambridge.

Harvey A. C., 1990, The Econometric Analysis of Time Series, Second Edition, MIT Press.

Jiang G. J., 1998, Nonparametric Modeling of U.S. Interest Rate Term Structure Dynamics and Implications on the Prices of Derivative Securities, Journal of Financial Quantitative Analysis, 33, 465-487.

Klebaner F. C., 1998, Introduction to Stochastic Calculus with Applications, Imperial College Press.

Marsh T. A., and E.R. Rosenfeld, 1983, Stochastic Processes for Interest Rates and Equilibrium Bond Price, Journal of Finance, 38, 635-646.

Merton R. C., 1973, Theory of Rational Option Pricing, Bell Journal of Economics and Management Science, 4, 141-183.

Pearson N. D., and T. Sun, 1994, Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll and Ross Model, Journal of Finance, 49, 1279-1304.

Pritsker M. G., 1998, Nonparametric Density Estimation and Test of Continuous Time Interest Rate Models, The Review of Financial Studies, 11, 449-487.

Robinson P.M., 1983, Nonparametric Estimators for Time Series, Journal of Time Series Analysis, 4, 185-207.

Stanton R., 1997, A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk, Journal of Finance, 52, 1973-2002.

Siddique A. R., 1994, Noparametric Estimation of Mean and Variance and Pricing of Securities, working paper, Georgetown University.

Vasicek O., 1977, An Equilibrium Charaterization of the Term Structure, Journal of Financial Economics, 5, 177-188.

White H., 1982, Maximum Likelihood Estimation of Misspecified Models, Econometrica, 50, 1-25.

Table 1: Some parametric models $dr = \mu(r)dt + \sigma(r)dB(t)$

$\mu(\mathrm{r})$	$\sigma(r)$	Implied Transition	Reference
$\beta\cdot(\alpha - r)$	σ	Normal	Vasicek (1977)
$\beta\cdot(\alpha - r)$	$\sigma\cdot r^{1/2}$	Non-central χ^2	Cox et al. (1985)
$\beta\cdot(\alpha - r)$	$\sigma\cdot r$	Lognormal	Courtadon (1982)
$\beta\cdot(\alpha - r)$	$\sigma\cdot r^{\nu}$	Gamma	Chan et al. (1992)
$\beta\cdot(\alpha - r)$	$\sqrt{\sigma + \nu\cdot r}$	Unknown	Duffie-Kan (1993)
$\beta\cdot r\cdot[\alpha - \ln(r)]$	$\sigma\cdot r$	Locally Lognormal	Brennan-Schwartz (1979)
$\beta\cdot r + \alpha\cdot r^{-(1-\delta)}$	$\sigma\cdot r^{\delta/2}$	Unknown	Marsh-Rosenfeld (1983)
$\alpha + \beta\cdot r + \nu\cdot r^2$	$\sigma + \nu\cdot r$	Unknown	Constantinides (1992)
β	σ	Normal	Merton (1973)
0	$\sigma\cdot r$	Lognormal	Dothan (1978)
0	$\sigma\cdot r^{3/2}$	Inverse-Gamma	Cox(1975) Cox et al. (1980)

Table 2: Some Nonparametric Model Estimations

Estimation	Specification	Reference
Neutral Network	learning curve and training	Hutchinson al. et. (1994)
Semi-parametric	$\mu(r) = \beta\cdot(\alpha - r)$, σ : nonparametric	Aït-Sahalia(1996a)

Nonparametric	μ and σ are nonparametric	Siddique(1994)
		Stanton(1997)
		Jiang (1998)

Table 3: Comparison the Estimation Results of Parametric model and Noparametric Model

This table represents the simulating results of two parametric model estimation and noparametric model estimation. The base-line model the term structure of interest rate followed as Chan, Karolyi, Longstaff and Sanders (1992) and spcifying $\alpha = 0.05$, $\beta = 1.25$, $\sigma = 0.064$, and $\nu = 0.75$. There are 100 samples of size 1000. And the standard errors of each estimate of parameters are in parentheses.

Models	Specification	Results
Base-line	$\mu(r) = 1.25(0.05 - r)$, $\sigma(r) = 0.064r^{0.75}$	
Vasicek	$\mu(r) = \beta(\alpha - r)$, $\sigma(r) = \sigma$	$\alpha = 0.083$, $\beta = 1.017$
		(0.041), (1.184)
CIR	$\mu(r) = \beta(\alpha - r)$, $\sigma(r) = \sigma r^{1/2}$	$\alpha = 0.076$, $\beta = 1.003$
		(0.058), (1.547)
Nonparametric	None	$\alpha = 0.056$, $\beta = 1.145$
		(0.033), (1.009)

Table 4: Comparison between Different Estimation Models

	$\mathcal{K}_b(u)$	$\mathcal{K}(u)$	Aït-Sahalia	FGLS	CIR	Vasicek
α	0.0621	0.0989	0.0611	0.0745	0.0668	0.0657
β	1.0014	1.9905	1.0388	0.9693	1.1003	1.0192

Table 5: Descriptive Statistics of U.S. FEDERAL FUNDS(EFFECTIVE)-MIDDLE RATE

In our research, the data used are the U.S. FEDERAL FUNDS (EFFECTIVE) - MIDDLE RATE from 1/1/1990 to 25/8/1999. We compute and list the descriptive statistics of the data, such as mean, variance, skewness, kurtosis, etc.

mean	median	range	variance	st. dev.	kurtosis	skewness
5.29	5.34	3.73	0.14	0.37	3.76	0.40

Bias Issue on The EM Algorithm *

Hung Chen and Wei-Yann Tsai

Department of Mathematics, Department of Biostatistics

National Taiwan University, Columbia University

Taipei, Taiwan; New York

July 2001

Abstract

EM algorithm has been found to be useful in finding maximum likelihood estimate based on incomplete data. In past ten years, this algorithm is found to be useful in handling regression model with missing covariate. In the E-step, it may involve a non-parametric estimate which usually introduce bias in the estimation equation or score equation. In this paper, we will discuss how it will affect the asymptotic behavior of resulting estimate.

AMS 1991 Subject Classification. Primary 62G07; secondary 62F12, 62J12.

Key Words and Phrases. Missing covariates, conditional mean imputation, curse of dimensionality.

Running title: Conditional Mean Imputation for Regression

Corresponding Author: Hung Chen, Department of Mathematics, National Taiwan University, Taipei, Taiwan 106, ROC.

Email: hchen@math.ntu.edu.tw

*This research was supported in part by Grants from the National Research Council of Republic of China.

1 Introduction

The expectation maximization, EM (See Dempster, Laird and Rubin, 1977, and the references thereof), algorithm for finding maximum likelihood estimates (MLE's) is a powerful numerical technique useful in contexts ranging from standard incomplete data problems (e.g. missing, grouping, censoring and truncation), to iteratively reweighted least squares analysis and empirical Bayes models.

The primary conceptual power of the EM algorithm lies in converting a maximization problem involving a complicated likelihood, into a sequence of **pseudo-complete** problems, where at each step the updated parameters can be obtained in a closed form (or at least in a straightforward manner). The general idea is to represent the observed data vector as the realization of some incompletely or indirectly observed data vector. The problem remained to be answer is how to do it systematically? For statistical data models including mixtures, convolutions and random effects, they are transparent. In general, it is not.

There is a growing literature on the development of least squares, maximum likelihood and Bayesian methodology for dealing with incomplete covariates in regression models. These methods are of particular interest in epidemiological studies where constraints of time, cost and technical difficulty prohibit the complete observation of an important covariate or covariates. For incomplete covariate regression models, the EM algorithm becomes a useful framework to tackle the estimation problem since the complete data and incomplete data are transparent. Most proposals in the literature involve an E-step and a nonparametric estimate of density or regression. This is a contrast to the case that both the E-step and the M-step can be carried out exactly using closed form solutions. This motivates the study of this paper. We want to study the effectiveness of EM algorithm when the E-step involves with an approximation.

In particular, we want to study

- Why can we replace $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{(v)}$ in the E-step without incurring further error in estimation?

  As a remark, EM typically leads to MLE but it introduces *error* using the above approximation.

- Suppose that we cannot get a closed form solution in E-step. Instead, some approximations such as nonparametric density estimation are introduced in E-step. Here we introduce the approximation error. As documented in the literature, the approximation error of nonparametric curve estimate in not up to the order of $n^{-1/2}$. How will it affect the effectiveness of EM algorithm to find MLE?

2 Some Useful Facts

2.1 EM Algorithm

We now describe the EM Algorithm.

- Suppose we have a model for the complete data $\mathbf{x}$, with associated density $f(\mathbf{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset R^k$ is the unknown parameter and $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathcal{X}$. Write $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{mis})$, where $\mathbf{x}_{obs}$ represents the observed part of $\mathbf{x}$ and $\mathbf{x}_{mis}$ denotes the missing values.

  The objective of the EM algorithm is to find the MLE for $\boldsymbol{\theta}$ based on the observed data $\mathbf{x}_{obs}$.

  Instead of observing $\mathbf{x}$, one observes the value of a measurable function $Y(\mathbf{x}) = \mathbf{x}_{obs} = y \in Y$.

- The EM method is only attractive in situations where finding the complete data MLE and either the observed or the expected information matrix would be straightforward, but the problem based on the incomplete data Y requires an iterative solution.

- The EM algorithm starts with an initial estimate $\boldsymbol{\theta}^{(0)}$. Letting $\boldsymbol{\theta}^{(t)}$ be the estimate of $\boldsymbol{\theta}$ at the tth iteration, iteration $(t+1)$ of EM is as follows.

  - **E step:** Find the expected complete-data log-likelihood of $\boldsymbol{\theta}$ were $\boldsymbol{\theta}^{(t)}$:

    $$\lambda^*(\mathrm{y}, \boldsymbol{\theta}|\boldsymbol{\theta}^{(v)}) = \log\{f(y|\boldsymbol{\theta})\} = \log\left\{\int_R f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}^{(v)}) d\mu(\mathbf{x})\right\}, \tag{1}$$

    where $R = \{\mathbf{x} : y(\mathbf{x}) = y\}$, and $\mu(\mathbf{x})$ is a dominating measure.

  - **M step:** Determine $\boldsymbol{\theta}^{(t+1)}$ by maximizing this expected log-likelihood:

    $$\lambda^*(y, \boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq \lambda^*(y, \boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \quad \text{for all } \boldsymbol{\theta}.$$

- Let

  $$\begin{aligned} \lambda(\mathbf{x}, \boldsymbol{\theta}) &= \log\{f(\mathbf{x}|\boldsymbol{\theta})\} \\ \lambda^*(y, \boldsymbol{\theta}) &= \log\{f(y|\boldsymbol{\theta})\} = \log\left\{\int_R f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) d\mu(\mathbf{x})\right\}, \end{aligned}$$

  Now, instead of maximizing $\boldsymbol{\theta}^*$ directly, the EM algorithm proceeds by using an initial estimate $\boldsymbol{\theta}^{(0)}$ and solving the pseudo-complete data problem:

  $$\text{maximize}_{\boldsymbol{\theta} \in \Theta} E_{\boldsymbol{\theta}^{(0)}}[\lambda(\mathbf{X}, \boldsymbol{\theta})|\mathbf{X} \in R].$$

- The maximizing value for this pseudo-complete data problem called $\boldsymbol{\theta}^{(1)}$ and the iteration is continued until $||\boldsymbol{\theta}^{(v+1)} - \boldsymbol{\theta}^{(v)}||$ is sufficiently small or some other convergence criterion is satisfied. The M step is standard maximum likelihood estimation for complete-data problems, and the E step is usually available from standard complete-data theory of conditional distributions.

2.2 Tool for Asymptotic Results

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots$ be a sequence of independent random vectors and $\mathbf{g}_i(y_i, \boldsymbol{\theta})$ be a sequence of $p \times 1$ vector functions which involve an unknown parameter $\boldsymbol{\theta}$. Denote $\mathcal{Y} = \{\mathbf{y}_i, i \geq 1\}$ and

$$G_n(\dagger, \boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \mathbf{g}_i(\mathbf{y}_i, \boldsymbol{\theta}).$$

In many statistical problems, we need to get an estimate $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ based on the observed variables $\mathbf{y}_i$ through function $G_n(\mathcal{Y}, \boldsymbol{\theta}) = 0$. They include maximum likelihood estimators, least squares and reweighted least sequares estimators.

Yuan and Jennrich (1998) gave conditions to ensure consistency and asymptotic normality of a sequence $\hat{\boldsymbol{\theta}}_n$ of roots of $G_n(\boldsymbol{\theta})$. The conditions are

- $G_n(\boldsymbol{\theta}_0) \to 0$ with probability one.
- There is a neighborhood N of $\boldsymbol{\theta}_0$ on which with probability one all $G_n(\boldsymbol{\theta})$ are continuously differentiable and the Jacobians $DG_n(\boldsymbol{\theta})$ converge uniformly to a nonstochastic limit which is nonsingular at $\boldsymbol{\theta}_0$.
- $G_n(\boldsymbol{\theta}_0) \xrightarrow{L} N(0, V)$.

Conditions 1 and 2 are for the existence and consistency of $\hat{\boldsymbol{\theta}}_n$. Condition 3 is for the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$.

Lemma 1 of Yuan and Jennrich (1998) claims that there is a continuously differentiable function $G(\boldsymbol{\theta})$ on N such that $G(\boldsymbol{\theta}_0) = 0$ and with probability one $G_n(\boldsymbol{\theta}) \to G(\boldsymbol{\theta})$ and $DG_n(\boldsymbol{\theta}) \to DG(\boldsymbol{\theta})$ uniformly on N under Conditions 1 and 2. When Conditions 1 to 3 hold,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} N(0, A^{-1}V(A^T)^{-1}),$$

where $A = DG(\boldsymbol{\theta}_0)$.

3 Examples

To motivate the derivation of likelihood function in our study, we first derive the likelihood function for random censoring on survival times. We start with parametric setting and then extend it to nonparametric setting.

3.1 Parametric Censoring

To motivate the derivation of likelihood function in our study, we first derive the likelihood function for random censoring on survival times. We first consider parametric setting and then extend to nonparametric setting. Let X have a distribution in the family $\{F(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, where Θ is an open set of R^k. Assume that the $F(x;\boldsymbol{\theta})$s are dominated by a σ-finite measure μ on a Euclidean space of which x is a generic point and possess nicely behaved densities $f_X(x;\boldsymbol{\theta})$ that are sufficiently smooth to make MLEs consistent and asymptotically normal. Write survival functions as $S_X(x,\boldsymbol{\theta})$. In survival or reliability analyses, a study to observe a random variable $X_1, \ldots, X_n$ will generally be terminated in practice before all of these random variables are able to be observed. Assume that the termination mechanism is random censoring.

Let $\{(y_i, r_i), 1 \leq i \leq n\}$ denote the observed data. According to the assumed data generation mechanism, we observe $y_j = x_j$ and $r_j = 1$ if $x_j \leq C_j$ and $y_j = c_j$ and $r_j = 0$ if $X_j > c_j$. Or, $r_j = 0$ or 1 according as the observation X_j is censored or uncensored at C_j $(j = 1, \ldots, n)$. To derive estimate of $\boldsymbol{\theta}$, we can use the method of maximum likelihood. Since $(R_1, X_1, C_1), \ldots, (R_n, X_n, C_n)$ are independent and identically distributed, $(R_1, Y_1), \ldots, (R_n, Y_n)$ are also independent and identically distributed. The likelihood function for $\boldsymbol{\theta}$ formed on the basis of $\{(r_i, y_i), 1 \leq i \leq n\}$ is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(r_i, y_i).$$

Observe that

$$\begin{aligned} P(R=1, Y=y)\mu(dy) &= P(X=y, C \geq y)dy \\ &= P(C \geq y) f_X(y)\mu(dy) = S_C(y) f_X(y)\mu(dy) \qquad (2) \end{aligned}$$

and

$$\begin{aligned} P(R=0, Y=y)\mu(dy) &= P(C=y, X \geq y)dy \\ &= P(X \geq y) f_C(y)\mu(dy) = S_X(y) f_C(y)\mu(dy). \qquad (3) \end{aligned}$$

Hence,

$$\begin{aligned} L_F(\boldsymbol{\theta}) &= \prod_{i=1}^{n} [f_X(y_i;\boldsymbol{\theta}) dy S_C(y_i)]^{r_i} [f_C(y_i) dy S_X(y_i;\boldsymbol{\theta})]^{1-r_i} \\ &= \prod_{i=1}^{n} [f_X(y_i;\boldsymbol{\theta}) dy]^{r_i} [S_X(y_i;\boldsymbol{\theta})]^{1-r_i} \cdot \prod_{i=1}^{n} [S_C(y_i)]^{r_i} [f_C(y_i) dy]^{1-r_i}. \qquad (4) \end{aligned}$$

Assume parameter distinctness defined in Rubin (1976, Biometrika). (i.e., It means that there is no *a priori* ties between data parameter $\boldsymbol{\theta}$ and incompleteness parameter associated with

C.) Hence, we can ignore the incompleteness mechanism and derive valid inference based on the following likelihood function instead.

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}[f_X(y_i;\boldsymbol{\theta})dy]^{r_i}[S_X(y_i;\boldsymbol{\theta})]^{1-r_i}. \tag{5}$$

Alternatively, we can write $L(\boldsymbol{\theta})$ as

$$\prod_{i=1}^{n}[S_X(y_i;\boldsymbol{\theta})dy]\left[\frac{f_X(y_i:\boldsymbol{\theta})}{S_X(y_i;\boldsymbol{\theta})}\right]^{r_i}.$$

In survival analysis, we consider the dynamic of the change of state *alive* to the state of *death* as time progresses. It is most natural to phrase it in terms of its dynamics. The above representation is derived under such a motivation. $S_X(y;\boldsymbol{\theta})$ refers to the studied subject is still alive at time y and the hazard function $[\lambda_X(y:\boldsymbol{\theta})]^{r_i}$ reflects that the studied subject is either dead or censored at time t. Note that the score equation is

$$\sum_{i=1}^{n}\left[r_i\frac{\partial f_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y_i;\boldsymbol{\theta})}\right]+\sum_{i=1}^{n}\left[(1-r_i)\frac{\partial S_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y_i;\boldsymbol{\theta})}\right]=0.$$

Suppose that we can observe the *ideal* data, $\{X_i, 1\le i\le n\}$. Then the score equation would be

$$\sum_{i=1}^{n}\frac{\partial f(X_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X_i;\boldsymbol{\theta})}=0.$$

Since $\sum_{i=1}^{n}[\partial f(X_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}]/f(X_i;\boldsymbol{\theta})$ cannot be observed, we might consider to estimate it by some nonlinear function of observed data. Suppose the mean squared error is the criterion to be used for determining this nonlinear function. Namely, we find $Q_n(y_1,\ldots,y_n;\boldsymbol{\theta})$ which is the minimizer of

$$E\left[\sum_{i=1}^{n}\frac{\partial f(X_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X_i;\boldsymbol{\theta})}-Q_n(y_1,\ldots,y_n;\boldsymbol{\theta})\right]^2.$$

Since no parametric form is assumed on the censoring distribution and y_j's are independent, it follows from the standard conditioning argument that

$$Q_n(y_1,\ldots,y_n;\boldsymbol{\theta})=\sum_{i=1}^{n}E\left(\frac{\partial f(X_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X_i;\boldsymbol{\theta})}\middle| y_i\right)$$

and

$$E_{\boldsymbol{\theta}}Q_n(Y_1,\ldots,Y_n;\boldsymbol{\theta})=E_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n}\frac{\partial f(X_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X_i;\boldsymbol{\theta})}\right)=0 \tag{6}$$

under regularity condition.

The above fact is the key on the proof of consistency of MLE or the minimum contrast estimates. Note that $E\left(\frac{\partial f(X_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X_i;\boldsymbol{\theta})}\middle| y_i\right)$'s are nonlinear transformation of y_i's. Also, its mean is zero by (6). It follows easily from the strong law of large numbers that

$$n^{-1}Q_n(Y_1,\ldots,Y_n;\boldsymbol{\theta})\overset{a.s.}{\to}0.$$

We just show that $n^{-1}Q_n(Y_1, \ldots, Y_n; \boldsymbol{\theta})$ satisfies Condition 1 in Section 2.2.

Fisher (1925) showed that incomplete data scores are conditional expectations (given the incomplete data) of the complete data scores. Efron, in his comments to Dempster *et al.* (1977), makes the link between Fisher's result and incomplete data methods. We will demonstrate it for this particular example. Observe that

$$\begin{aligned} E\left(\frac{\partial f(X;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X;\boldsymbol{\theta})}\middle| y_i, r_i = 1\right) &= \frac{\partial f_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y_i;\boldsymbol{\theta})}, \\ E\left(\frac{\partial f(X;\boldsymbol{\theta})/\partial}{f(X;\boldsymbol{\theta})}\middle| y_i, r_i = 0\right) &= E\left(\frac{\partial f(X;\boldsymbol{\theta})/\partial\theta}{f(X;\boldsymbol{\theta})}\middle| X > y_i\right) \\ &= \frac{1}{S_X(y_i;\boldsymbol{\theta})}\int_{y_i}^{\infty}\frac{\partial}{\partial\boldsymbol{\theta}}f_X(x;\boldsymbol{\theta})\mu(dx) = \frac{\partial\log S_X(y_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \end{aligned}$$

The last equality holds under usual assumed regularity conditions of MLE that the differentiation and integration can be interchanged.

The above derivation just reiterates Fisher's statement on the incomplete data scores. As commented by Efron, the E-step of the EM algorithm is equivalent to finding the score vector of the observed data. However, it does not address the issue on the calculation of

$$E\left(\frac{\partial f(X;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X;\boldsymbol{\theta})}\middle| X > y\right)$$

in Dempester et al.(1977). Under the above parametric setting, it can be done by numerical integration or Monte-Carlo when an analytic calculation is not apparent. However, special attention should be given on the approximation error such that it won't swamp the random error. Let $\hat{E}\left(\frac{\partial f(X;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(X;\boldsymbol{\theta})}|X > y\right)$ denote such an approximation. Set

$$\hat{G}_n(\boldsymbol{\theta}) = n^{-1}\left\{\sum_{i=1}^{n}\left[r_i\frac{\partial f_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y_i;\boldsymbol{\theta})}\right] + \sum_{i=1}^{n}\left[(1-r_i)\hat{E}\left(\frac{\partial S_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y_i;\boldsymbol{\theta})}\right)\right]\right\}.$$

Denote the root of $\hat{G}_n(\boldsymbol{\theta}) = 0$ by $\hat{\boldsymbol{\theta}}_n^{EMLE}$.

We now use the results presented in Yuan and Jennrich (1998) to study whether $\hat{\boldsymbol{\theta}}_n^{EMLE}$ will be asymptotically normal as usually expected in likelihood analysis. Write

$$\hat{G}_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}) + [\hat{G}_n(\boldsymbol{\theta}) - G_n(\boldsymbol{\theta})],$$

where

$$G_n(\boldsymbol{\theta}) = n^{-1}\left\{\sum_{i=1}^{n}\left[r_i\frac{\partial f_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y_i;\boldsymbol{\theta})}\right] + \sum_{i=1}^{n}\left\{(1-r_i)E\left(\frac{\partial S_X(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y_i;\boldsymbol{\theta})}\right)\right]\right\}.$$

Note that $\hat{G}_n(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta})$ in this paricular example.

Since Y_i are independent and identically distributed, we can use strong law of large numbers to show that

$$G_n(\boldsymbol{\theta}_0) \overset{as}{\to} E_{\boldsymbol{\theta}_0}\left(\frac{\partial f(X;\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f(X;\boldsymbol{\theta}_0)}\right) = 0.$$

We conclude that it satisfies Condition 1. We now use an alternative to show that the expectation of score vector of $L(\boldsymbol{\theta}_0)$ is zero. Let 1_X and 1_C denote the support of densities functions of X and C, respectively. By (2) and (3), we have

$$\begin{aligned} & E\left[R\frac{\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y;\boldsymbol{\theta})}+(1-R)\frac{\partial S_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y;\boldsymbol{\theta})}\right] \\ = & \int\frac{\partial f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}S_C(y)\cdot 1_C\mu(dy)-\int f_X(y;\boldsymbol{\theta})f_C(y)\cdot 1_X\mu(dy) \\ = & \int\frac{\partial f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}S_C(y)\cdot 1_C\mu(dy)+\int f_X(y;\boldsymbol{\theta})\frac{dS_C(y)}{d\mu}\cdot 1_X\mu(dy). \end{aligned}$$

If X and C have common support and $f_X(S_X^{-1}(1);\boldsymbol{\theta})=0$,

$$E\left[R\frac{\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y;\boldsymbol{\theta})}+(1-R)\frac{\partial S_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y;\boldsymbol{\theta})}\right]=0$$

by integration by parts.

To check for Conditions 2 and 3, we just show that it satisfies the usual regularity conditions assumed for likelihood analysis. Observe that

$$\begin{aligned} & E\left[R\frac{\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y;\boldsymbol{\theta})}+(1-R)\frac{\partial S_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y;\boldsymbol{\theta})}\right]^2 \\ = & \int\frac{[\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}]^2}{f_X(y;\boldsymbol{\theta})}S_C(y)\cdot 1_C\mu(dy)+\int\frac{f_X^2(y;\boldsymbol{\theta})}{S_X(y;\boldsymbol{\theta})}f_C(y)\cdot 1_X\mu(dy)>0 \end{aligned}$$

and

$$\begin{aligned} & E\left[R\frac{\partial^2\log f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}+(1-R)\frac{\partial^2\log S_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\right] \\ = & -\left\{\int\frac{[\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}]^2}{f_X(y;\boldsymbol{\theta})}S_C(y)\cdot 1_C\mu(dy)+\int\frac{f_X^2(y;\boldsymbol{\theta})}{S_X(y;\boldsymbol{\theta})}f_C(y)\cdot 1_X\mu(dy)\right\} \\ & -\left\{\int\frac{\partial f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}f_C(y)\cdot 1_C\mu(dy)-\int\frac{\partial^2 f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}S_C(y)\cdot 1_X\mu(dy)\right\} \\ = & -E\left[R\frac{\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y;\boldsymbol{\theta})}+(1-R)\frac{\partial S_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y;\boldsymbol{\theta})}\right]^2 \\ & +\left\{\int\frac{\partial f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{dS_C(y)}{d\mu}\cdot 1_C\mu(dy)+\int\frac{\partial^2 f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}S_C(y)\cdot 1_X\mu(dy)\right\}. \end{aligned}$$

If X and C have common support and $\partial f_X(S_X^{-1}(1);\boldsymbol{\theta})/\partial\boldsymbol{\theta}=0$, then the last term of the above equation is equal to zero by integration by parts. We have

$$E\left[R\frac{\partial^2\log f_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}+(1-R)\frac{\partial^2\log S_X(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\right]=-E\left[R\frac{\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y;\boldsymbol{\theta})}+(1-R)\frac{\partial S_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y;\boldsymbol{\theta})}\right]^2.$$

Condition 3 holds by the central limit theorem with $A=V$. Condition 2 holds by assuming bounded third derivatives of density function. When $\sqrt{n}\left(\hat{G}_n(\boldsymbol{\theta})-G_n(\boldsymbol{\theta})\right)=o_P(1)$, $\hat{\boldsymbol{\theta}}_n^{EMLE}$ is asymptotical normal with variance

$$\left\{E\left[R\frac{\partial f_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f_X(y;\boldsymbol{\theta})}+(1-R)\frac{\partial S_X(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{S_X(y;\boldsymbol{\theta})}\right]^2\right\}^{-1}.$$

This variance can be estimated by

$$n\left\{\sum_{i=1}^{n} R_i\left(\frac{\partial \log f_X(Y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2+\sum_{i=1}^{n}(1-R_i)\left(\frac{\partial \lambda_X(y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right\}^{-1}.$$

3.2 Estimation of Survival Function in Nonparametric Setting

Let $W_1,\ldots,W_n$ be *survival times* that are iid nonnegative random variables from a cdf F, and $C_1,\ldots,C_n$ be iid nonnegative random variables independent of W_i's. Here we assume that $F\in\mathcal{F}$ and $\mathcal{F}$ is the collection of all cdf's on R^d. Suppose that we are only able to observe the smaller of W_i and C_i and an indicator of which variables is smaller:

$$X_i=\min(W_i,C_i),\quad \delta_i=I_{(0,C_i)}(W_i),\quad i=1,\ldots,n.$$

This is called a *random censorship model*. We consider the estimation of F.

Given $W_1=w_1,\ldots,W_n=w_n$, the nonparametric likelihood function is defined to be the following functional from $\mathcal{F}$ to $[0,\infty)$:

$$\ell(G)=\prod_{i=1}^{n}P_G(\{w_i\}),\quad G\in\mathcal{F}.$$

Here P_G is the probability measure corresponding to $G\in\mathcal{F}$. Apparently, $\ell(G)=0$ if $P_G(\{w_i\})=0$ for at least one i. Kiefer and Wolfowitz (1956) shows that the empirical c.d.f. F_n is a nonparametric MLE of F. Define

$$H(p_1,\ldots,p_n,\lambda)=\prod_{i=1}^{n}p_j+\lambda(\sum_{i=1}^{n}p_i-c),$$

where λ is the Lagrange multiplier. Set

$$\frac{\partial H}{\partial\lambda}=\sum_{i=1}^{n}p_i-c=0,\quad \frac{\partial H}{\partial p_j}=p_j^{-1}\prod_{i=1}^{n}p_i+\lambda=0, j=1,\ldots,n.$$

The solution is $p_i=c/n$, $i=1,\ldots,n$, $\lambda=-(c/n)^{n-1}$. This means that $\max\ell(G)=(c/n)^n$ which is maximized at $c=1$ for any fixed n.

Recall that W_i is the time to death for the ith studied subject. Define $N_i(t)$ as the counting process which is increasing in t and takes values on $\{0,1\}$. Conditional on the death times $W_{(1)}<W_{(2)}<\cdots<W_{(n)}$, we have

$$P\left(\sum_{i=1}^{n}N_i(W_{(j)})=j,1\le j\le n\right)=\prod_{j=1}^{n}P\left(\sum_{i=1}^{n}N_i(W_{(j)})=j,\middle|\sum_{i=1}^{n}N_i(W_{(j-1)})=j-1\right).$$

Observe that

$$\begin{aligned}&P\left(\sum_{i=1}^{n}N_i(W_{(j)})=j,\middle|\sum_{i=1}^{n}N_i(W_{(j-1)})=j-1\right)\\ =\;& C(n-j+1,1)\left(\frac{p_j}{1-\sum_{k=1}^{j-1}p_k}\right)\left(\frac{1-\sum_{k=1}^{j}p_k}{1-\sum_{k=1}^{j-1}p_k}\right)^{n-j}.\end{aligned}$$

Hence,

$$P\left(\sum_{i=1}^{n} N_i(W_{(j)}) = j, 1 \le j \le n \,\middle|\, W_{(1)} < W_{(2)} < \cdots < W_{(n)}\right) = n! \prod_{i=1}^{n} p_i,$$

where $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$.

When $\delta_{(n)} = 1$, the nonparametric mle only needs to consider those F with support on the collection of $W_{(i)}$ with $\delta_{(i)} = 1$. When $\delta_{(n)} = 0$, the nonparametric mle only needs to consider those F with support on the union of the collection of $W_{(i)}$ with $\delta_{(i)} = 1$ and any point which is larger than $W_{(n)}$. Observe that

$$P\left(\sum_{i=1}^{n} N_i(C_{(j)}) = j, 1 \le j \le n\right) = \prod_{j=1}^{n} P\left(\sum_{i=1}^{n} N_i(C_{(j)}) = j, \,\middle|\, \sum_{i=1}^{n} N_i(C_{(j-1)}) = j-1\right)$$

and

$$P\left(\sum_{i=1}^{n} N_i(C_{(j)}) = j, \,\middle|\, \sum_{i=1}^{n} N_i(C_{(j-1)}) = j-1\right)$$
$$= \begin{cases} C(n-j+1,1) & \delta_{(j)} = 0 \\ C(n-j+1,1)\left(\frac{p_j}{1-\sum_{k=1}^{j-1}\delta_{(k)}p_k}\right)\left(\frac{1-\sum_{k=1}^{j} p_k\delta_{(k)}}{1-\sum_{k=1}^{j-1} p_k\delta_{(k)}}\right)^{n-j} & \delta_{(j)} = 1. \end{cases}$$

Hence, we get the following $\ell(F)$.

An *maximum empirical likelihood estimator* (MELE) of F is defined by the maximizer of

$$\ell(F) = \prod_{i=1}^{n} p_i^{\delta_{(i)}} \left(\sum_{j=i+1}^{n+1} p_j\right)^{1-\delta_{(i)}}$$

subject to $p_i = P_F(\{X_{(i)}\}) \ge 0$, $1 \le i \le n$, $p_{n+1} = 1 - F(X_{(n)}) \ge 0$, and $\sum_{i=1}^{n+1} p_i = 1$.

It can be shown that the above maximization problem is equivalent to the maximization of

$$\prod_{i=1}^{n} q_i^{\delta_{(i)}} (1-q_i)^{n-i+1-\delta_{(i)}}$$

where $q_i = p_i / \sum_{j=i}^{n+1} p_j$, $i = 1, \ldots, n$. Also, the MELE is

$$\hat{F}(t) = \sum_{i=1}^{n+1} \hat{p}_i I_{[X_{(i-1)}, X_{(i)})}(t),$$

where $X_{(0)} = 0$, $X_{(n+1)} = \infty$, $X_{(i)}$ are order statistics, and

$$\hat{p}_1 = n^{-1}, \;\; \hat{p}_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{(j)}}{n-j+1}\right), i = 2, \ldots, n, \hat{p}_{n+1} = 1 - \sum_{j=1}^{n} \hat{p}_j.$$

$\hat{F}(t)$ can also be written as

$$1 - \prod_{X_{(i)} \le t} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right),$$

which is the Kaplan-Meier product-limit estimator.

3.3 Mixture Models

A common proposal to model heterogeneous data is to consider a finite-mixture model. In the problem considered by Do and McLachlan (1984), the population of interest consists of rats from g species $G_1, \ldots, G_g$, that are consumed by owls in some unknown proportions $\pi_1, \ldots, \pi_g$. The problem is to estimate the π on the basis of the observation vector $\mathbf{W}$ containing measurements recorded on a sample of size n of rat skulls taken from owl pellets. The rats constitute part of an owl's diet, and indigestible material is regurgitated as a pellet.

We can use the argument of conditioning, the underlying population can be modeled as consisting of g distinct groups $G_1, \ldots, G_g$ in some unknown proportions $\pi_1, \ldots, \pi_g$, and where the conditional pdf of $\mathbf{W}$ given membership of the ith group G_i is $f_i(\mathbf{w})$. Let $\mathbf{y} = (w_1^T, \ldots, w_n^T)^T$ denote the observed random sample obtained from the mixture density

$$f(w; (\pi_1, \ldots, \pi_{g-1})) = \sum_{i=1}^{g} \pi f_i(w).$$

The log likelihood function for $(\pi_1, \ldots, \pi_{g-1})$ can be formed from the observed data $\mathbf{y}$ is given by

$$\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{g} \pi_j f_j(w_i) \right\}.$$

On differentiating log likelihood function with respect to π_j $(j = 1, \ldots, g-1)$, we obtain

$$\sum_{i=1}^{n} \left\{ \frac{f_j(w_i)}{f(w_i; (\pi_1, \ldots, \pi_{g-1}))} - \frac{f_g(w_i)}{f(w_i; (\pi_1, \ldots, \pi_{g-1}))} \right\} = 0,$$

for $j = 1, \ldots, g-1$. It clearly does not yield an explicit solution for $(\pi_1, \ldots, \pi_{g-1})^T$.

4 Incomplete Covariates in Regression Models

In Section 3, we consider the case that the unobservable score or estimation equation can be estimated unbiased or with error $O(n^{-1})$. In this Section, we will consider the case that those unobservable score or estimating equation may not be estimated unbiased or with error bigger than $O(n^{-1/2})$.

Suppose the regression model to be considered is of the form $f_{\beta}(Y|X, Z)$ which relates the outcome Y to X and Z. Here X and Z represent the incomplete and complete covariates, respectively. To handle such data, it requires the specification of the form of the conditional density $f(X|Z)$. Since the performance of estimate of β relies critically on the specification of the form of $f(X|Z)$. A number of semi-parametric methods have been derived which does not require specification of the form of $f(X|Z)$. Here we consider the mean-score method proposed in Reilly and Pepe (1995) and **hot-deck** used by the U.S. Census Bureau. In the following discussion, it assumes that Y and Z to be categorical.

Reilly and Pepe (1997) describe the motivation of the mean-score method as follows. If the relationship between the complete and incomplete covariates $f(X|Z)$ was fully known, the maximum likelihood estimate of $\boldsymbol{\beta}$ can be found by using the EM algorithm. Denoting the complete cases (the ***validation sample***) as V and the incomplete cases as $\bar{V}$, this would involve the iterative maximization of

$$\sum_{i\in V}\log f_{\boldsymbol{\beta}}(Y_i|X_i,Z_i)+\sum_{j\in\bar{V}}E\left[\log f_{\boldsymbol{\beta}}(Y_j|X,Z_j)|\boldsymbol{\beta},Y_j,Z_j\right]$$

or equivalently, the solution of

$$\sum_{i\in V}S_{\boldsymbol{\beta}}(Y_i|X_i,Z_i)+\sum_{j\in\bar{V}}E\left[S_{\boldsymbol{\beta}}(Y_j|X,Z_j)|\boldsymbol{\beta},Y_j,Z_j\right]=0,$$

where $S_{\boldsymbol{\beta}}(Y_i|X_i,Z_i) = (\partial/\partial\boldsymbol{\beta})f_{\boldsymbol{\beta}}(Y_i|X_i,Z_i)$. The mean-score method involves using non-parametric estimates of the conditional expected values in the second term, and hence solving the score equation

$$S_n(\boldsymbol{\beta})=\sum_{i\in V}S_{\boldsymbol{\beta}}(Y_i|X_i,Z_i)+\sum_{j\in\bar{V}}\sum_{i\in V^{Z_jY_j}}S_{\boldsymbol{\beta}}(Y_j|X,Z_j)/n_V^{Z_jY_j}=0,$$

where $V^{Z_jY_j}$ denotes those validation sample members whose $Z=Z_j$ and $Y=Y_j$, and $n_V^{Z_jY_j}$ is the number of such cases.

To make the above point precise, let R_i denote an indicator with $R_i=1$ meaning that X_i is observed and $R_i=0$ if X_i is missing. On the application of EM algorithm, it leads to the calculation of

$$\begin{aligned}&\sum_{i\in V}E\left[S_{\boldsymbol{\beta}}(Y_i|X_i,Z_i)|(Y_i,X_i,Z_i,R_i)\right]+\sum_{j\in\bar{V}}E\left[S_{\boldsymbol{\beta}}(Y_j|X,Z_j)|(Y_j,Z_j,R_j)\right]\\&=\sum_{i\in V}S_{\boldsymbol{\beta}}(Y_i|X_i,Z_i)+\sum_{j\in\bar{V}}E\left[S_{\boldsymbol{\beta}}(Y_j|X,Z_j)|(Y_j,Z_j)\right],\end{aligned}$$

if the missingness only depend on (Y,Z). Namely,

$$P(R=1|(Y,X,Z))=\pi(Y,Z).$$

Nonparametric estimates of $E\left[S_{\boldsymbol{\beta}}(Y_j|X,Z_j)|(Y_j,Z_j)\right]$, $j\in\bar{V}$, are needed. A natural estimate is $\sum_{i\in V^{Z_jY_j}}S_{\boldsymbol{\beta}}(Y_j|X,Z_j)/n_V^{Z_jY_j}$. Note that the above imputation can be viewed as conditional mean imputation based on Y and Z. This method has been reviewed in Little (1992).

A simple imputation method known as **hot-deck**, used by the U.S. Census Bureau, completes the data set by imputing (that is, filling in) for each subject with missing X, an

X which is selected at random, with replacement, from subjects who match them on the observed variables. A multiple-imputation version of hot-deck involves repeating this simple imputation step a number of times, and for each completed data set performing a standard analysis to obtain the usual regression estimate of β in the model $f_\beta(Y|X, Z)$, which we denote as $\hat{\beta}_i$ for the ith completed data set. The multiple-imputation hot-deck estimate is the average of the completed-data estimates:

$$\hat{\beta}_{HD} = \sum_{i=1}^{K} \frac{\hat{\beta}_i}{K}$$

where K is the number of imputations.

When Y and Z are categorical, the above estimate of unobservable score equation is again unbiased. Moreover, the introduced variability of EM algorithm is of order $O_P(n^{-1/2})$. Therefore, the resulting estimate will be asymptotically normal with convergence rate $n^{-1/2}$. If Y and Z are not categorical, we can apply standard stratification technique and then use the above approach. In other words, we use a nonparametric curve fitting technique which will run into curse of dimensionality. In this case, the resulting estimate will no longer be $O_P(n^{-1/2})$. This raises the bias issue on using EM algorithm.

5 Discussion

In applications, it is quite often that we cannot get the random sample. Instead, we get a biased sample. As an example, we consider the random censored data. It is well-known that we cannot just use those observed data to do usual analysis. The reason is that those samples with longer life span will be censored heavier than those samples with shorter life span. Therefore, we need to adjust data properly so that a *pseudo random sample* can be obtained.

When we get a biased sample, EM algorithm is a general recipe for getting an estimate of unobservable score or estimation equations. However, EM algorithm is being used when both the E-step and M-step can be done easily. When E-step cannot be done easily, Monte-Carlo or resampling method are being used. Usually, it is assumed that the bias caused by those algorithms is negligible. In this paper, we consider the regression with missing covariates. In this case, the E-step will involve a nonparametric estimation. Quite often, the bias can be huge. Therefore, the performance of resulting estimate can be bad. This phenomenon is being reported in the literature. However, none of these works associates it with the bias issue in E-step. In this paper, we just try to raise this issue and use the results of Yuan and Jennrich (1998) to quantify the effect caused by the magnitude of bias.

References

[1] Carroll, R.J. and Wand, M.P., 1991. Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc. B* **53**, 573-585.

[2] Dempster, A.P. Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B* **39**, 1-38.

[3] Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. *J. Am. Statist. Ass.* **58**, 13-30.

[4] Kiefer, R.J.A. and Wolfowitz, 1956. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Annals of Mathematical Statistics*, **27**, 887-906.

[5] Little, R.J.A., 1992. Regression with missing X's: *J. Am. Statist. Ass.* **86**, 1227-1237.

[6] Pepe, M.S. and Fleming, T.R., 1991. A non-parametric method for dealing with mismeasured covariate data. *J. Am. Statist. Ass.* **86**, 108-113.

[7] Reilly, M. and Pepe, M.S., 1995. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299-314.

[8] Reilly, M. and Pepe, M.S., 1997. The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* **16**, 5-19.

[9] Rubin, D.B., 1976. Inference and missing data. *Biometrika* **63**, 581-592.

[10] Sande, I.G., 1983. Hot-deck imputation procedures. In Volume 3 of *Incomplete Data in Sample Surveys* edited by Madow, W.G., Olkin, I. and Rubin, D.B., p339-352. New York: Academic Press.

[11] Yuan, K.H. and Jennrich, R.J., 1998. Asymptotics of estimating equations under natural conditions. *J. Multivariate Anal.* **65**, 245-260.