DEPARTMENT OF ECONOMICS
UNIVERSITY OF CYPRUS

Partially Adaptive Estimation via Maximum Entropy Densities

Thanasis Stengos and Ximing Wu

Discussion Paper 2005-07

P.O. Box 20537, 1678 Nicosia, CYPRUS Tel.: ++357-2-892430, Fax: ++357-2-892432
Web site: http://www.econ.ucy.ac.cy

Partially Adaptive Estimation via the Maximum Entropy Densities

Thanasis Stengos*and Ximing Wu†

October 6, 2005

Abstract

We propose a partially adaptive estimator based on information theoretic maximum entropy estimates of the error distribution. The maximum entropy (maxent) densities have simple yet flexible functional forms to nest most of the mathematical distributions. Unlike the nonparametric fully adaptive estimators, our parametric estimators do not involve choosing a bandwidth or trimming, and only require estimating a small number of nuisance parameters, which is desirable when the sample size is small. Monte Carlo simulations suggest that the proposed estimators fare well with non-normal error distributions.

*Department of Economics, University of Guelph, Guelph, Ontario, Canada, N1G 2W1; email: tstengos@uoguelph.ca.

†Department of Economics, University of Guelph, Guelph, Ontario, Canada, N1G 2W1; email: xiwu@uoguelph.ca.

1 Introduction

It is well known that the widely used least squares estimator is efficient if the errors are independent and identically normally distributed and independent of the regressors. When the error distribution is non-normal, the efficiency of the least squares estimator deteriorates. For example, when the error distribution has "fatter" tails than the normal, the least squares estimator can be inefficient relative to other estimators.

One approach to deal with non-normal error distributions is the adaptive estimation, which "adapts" to an unknown error distribution by maximizing an estimated likelihood function based on an estimate of the error distribution. The idea of an adaptive estimator was first developed by Stein (1956). Beran (1974) and Stone (1975) considered adaptive estimation in the symmetric location model. Bickel (1982) extended this to linear regression and other models for *i.i.d.* errors. Manski (1984) studied adaptive estimation in non-linear models. Steigerwald (1992) and Linton (1993) considered dependent errors; Li and Stengos (1994) looked at heterogenous errors.

Consider the classical linear regression

$$y_i = \alpha_0 + x_i \beta_0 + u_i, \quad i = 1, 2, \ldots, n.$$

Here u_i is independent of x_i and *i.i.d.* distributed according to a density $f(u, \theta)$, where θ is the shape parameter of f. Denote the likelihood function as $L(y|x, \beta, \theta)$. When the information matrix is block-diagonal, or

$$E\left[\partial \ln L(y|x, \beta, \theta) / \partial \beta \cdot \partial \ln L(y|x, \beta, \theta) / \partial \theta\right] = 0, \tag{1}$$

Bickel (1982) showed that the slope parameter of the model can be estimated adaptively, namely, one can do as well in terms of asymptotic variance as if one knew the true error distribution f.

When the density function f is known, one can obtain the maximum likelihood estimator (MLE) of β_0 by setting the average score function

$$
\begin{aligned}
s(\beta) = s(\beta; f) &= \frac{1}{n} \sum_{i=1}^{n} s_i(\beta; f) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \frac{f'(u_i(\beta))}{f(u_i(\beta))} x_i
\end{aligned} \tag{2}
$$

equal to zero. Alternatively, Bickel (1975) proposed a one step Newton-Raphson estimator

$$
\overline{\beta}_{NR} = \widetilde{\beta} + \widetilde{I}\left(\widetilde{\beta}; f\right)^{-1} s\left(\widetilde{\beta}; f\right),
$$

where $\widetilde{\beta}$ is a preliminary root-n consistent estimate of β_0 and $\widetilde{I}$ is a consistent estimate of the information matrix. This estimator is also referred to as the linearized likelihood estimator.

Since in most cases the density f is unknown, it is usually estimated from the residuals after a consistent estimate of β_0 is obtained. Within the parametric framework, Newey (1988) used a GMM approach for adaptive estimation. Alternatively, in the context of a nonparametric regression model with an unknown regression function, if the error distribution is consistently estimated using some nonparametric smoothers, the resulting fully adaptive estimator is asymptotically efficient (see Linton and Xiao (2004) and references therein).

Instead of trying to obtain an asymptotically efficient estimator using nonparametric methods, some researchers propose partially adaptive estimators based on parametric estimates of the error distribution. For example, McDonald and Newey (1988) and McDonald and White (1993) used the generalized t distribution, and Phillips (1994) used the mixture of normal distributions. As contended by Bickel (1982) and McDonald and Newey (1988), a partially adaptive estimator based on parametric estimates of the error distribution might be more practical. In particular, when the sample size is small, the partially adaptive estimator with a small number of nuisance parameters may outperform the fully adaptive estimator. The fully adaptive estimator, which estimates the score function (2) nonparametrically, depends crucially on the choice of bandwidth. Generally, the converge rate of a nonparametric estimator is different for the density function itself and its first derivative. As separate derivations of an optimal bandwidth for the density and its first derivative are rather complicated in the adaptive estimation, in practice often one single bandwidth is used for the density and its first derivative. In contrast, a root-n consistent parametric estimate of a differentiable density function retains its root-n consistency for its first derivative and does not involve bandwidth selection. Furthermore, our parametric estimator does not suffer from numerical difficulties associated with nonparametric estimation of the score function (2) when the density estimate in the denominator is close to zero.

In this study, we propose partially adaptive estimators based on the Maximum Entropy (maxent) density estimates of the error distribution. The Maximum Entropy principle is a general method of assigning values to probability distributions based on limited information such as moments. The maxent densities have simple yet flexible functional forms that nest most commonly used mathematical distributions. We propose a particular maxent density that has certain advantages over the t distribution and its generalizations used in the literature. It nests the normal distribution as a special case rather than a limiting case. Practically, it is more numerically stable, as the saddle point problem involved in the estimation of the t family of distributions can sometimes behave irregularly. The resulting partially adaptive estimators are quasi maximum likelihood estimators when the estimated maxent density approximates the unknown distribution of errors, and maximum likelihood estimators when underlying error distribution belongs to the family specified by the assumed maxent density. Our Monte Carlo simulations show that the proposed method demonstrates considerable degree of adaptiveness to different shape of error distributions and compares favorable with existing methods.

The next section reviews the maximum entropy density and introduces the particular maxent density estimator that we propose. The third section introduces the partially adaptive estimator. The fourth and fifth section report Monte Carlo simulations and an empirical application of the proposed estimator. The last section concludes.

2 Maximum Entropy Density

In this section, we introduce the Principle of Maximum Entropy and the maximum entropy densities. We then discuss the merits of the maxent densities as a practical tool for parametric density estimation. We introduce a simple but flexible maxent density specification that works well in approximating skewed and/or leptokurtic distributions. This proposed maxent density will provide the basis of obtaining partially adaptive estimators.

2.1 Background

The central concept of information theory is Shannon's Information Entropy

$$W = -\int f\left(z,\theta\right) \log f\left(z,\theta\right) dz,$$

where f is the density function for a random variable z. Entropy is a measure of disorder or uncertainty.

The celebrated Maximum Entropy (maxent) Principle states that among all the distributions that satisfy certain moment constraints, one should choose the distribution that maximizes the entropy. According to Jaynes (1957), the maxent distribution is "uniquely determined as the one which is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters."

The maxent density is obtained by maximizing the entropy subject to certain moment constraints. Let $z_1, z_2, ..., z_n$ be an *i.i.d.* random sample of size n from a distribution $f(z,\theta)$ on the real line. We maximize the entropy subject to

$$\begin{aligned}
&\int f(z,\theta)\,dz = 1, \\
&\int g_k(z)\,f(z,\theta)\,dz = \hat{\mu}_k, \quad k = 1,2,\ldots,K,
\end{aligned}$$

where $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} g_k(z_i)$, and $g_k(z)$ is generally continuously differentiable.[1] The first moment condition is imposed to render f a proper density function. The solution takes the form

$$f\left(z,\hat{\theta}\right) = \exp\left(-\hat{\theta}_0 - \sum_{k=1}^{K}\hat{\theta}_k g_k(z)\right). \tag{3}$$

To ensure $f\left(z,\hat{\theta}\right)$ integrates to one, we set

$$\hat{\theta}_0 = \log\left(\int \exp\left(-\sum_{k=1}^{K}\hat{\theta}_k g_k(z)\right)dz\right).$$

The maximized entropy $W = \hat{\theta}_0 + \sum_{k=1}^{K}\hat{\theta}_k\hat{\mu}_k$.

The maxent density is of the generalized exponential family and can be completely characterized by the moments $Eg_k(z), k = 1,2,\ldots,K$. Hence, we call these moments "characterizing moments", which are the sufficient statistics of the maxent density. A wide range of distributions belong to this family. For example, the Pearson family and its extensions described in Cobb et al. (1982), which nest the normal, beta, gamma and inverse gamma densities as special cases, are all maxent densities characterized by a few simple moments.

In general, there is no analytical solution for the maxent density, and nonlinear optimization is required (see Zellner and Highfield (1988), Wu (2003) and Wu and Perloff (forthcoming)). We use Lagrange's method to solve for this problem by iteratively

[1] This condition can be relaxed. For example, when $g(z) = |z|$, the corresponding maxent density is the Laplace distribution.

updating $\hat{\theta}$. For the $(t+1)^{th}$ stage of updating,

$$\hat{\theta}_{(t+1)} = \hat{\theta}_{(t)} - \mathbf{H}_{(t)}^{-1}\mathbf{b}_{(\mathbf{t})},$$

where $\mathbf{b} = [b_1, b_2, \ldots, b_k]'$, $b_k = \int g_k(z) f\left(z, \hat{\theta}\right) dz - \hat{\mu}_k$ and the Hessian matrix $\mathbf{H}$ takes the form

$$H_{k,j} = \int g_k(z) g_j(z) f\left(z, \hat{\theta}\right) dz, \; 0 \leq k, j \leq K. \tag{4}$$

The positive-definitiveness of the Hessian ensures the existence and uniqueness of the solution.[2] Moreover, the maxent method is equivalent to a maximum likelihood approach where the likelihood function is defined over the exponential distribution and therefore consistent and efficient.

2.2 A Flexible Maxent Density Specification

Barron and Sheu (1991) characterized the maxent density alternatively as an approximation of the log density by some basis functions, such as polynomials, trigonometric series or splines. They showed that the estimator does not depend on the choice of basis function. Denote the unknown true density f and its estimate $\hat{f}$, the Kullback-Leibler distance is defined as

$$D = \int f \log \frac{f}{\hat{f}} dz.$$

The Kullback-Leibler distance measures the discrepancy between f and $\hat{f}$. It is non-negative and takes the value zero if and only if $f = \hat{f}$ everywhere. Under some mild

[2] Let $\gamma' = [\gamma_0, \gamma_1, \ldots, \gamma_K]$ be a non-zero vector and $g_0(z) = 1$, we have

$$\begin{aligned}\gamma' \mathbf{H} \gamma &= \sum_{k=0}^{K} \sum_{j=0}^{K} \gamma_k \gamma_j \int g_k(z) g_j(z) f(z, \theta) dz \\ &= \int \left(\sum_{k=0}^{K} \gamma_k g_k(z) \right)^2 f(z, \theta) dz > 0.\end{aligned}$$

Hence, $\mathbf{H}$ is positive-definite.

regularity condition, the maxent density estimates converge to the underlying density, in terms of the Kullback-Leibler distance, if the number of moment conditions increases with sample size.

The Kullback-Leibler distance is a pseudo-metric as it is not symmetric with respect to f and $\hat{f}$. Tagliani (2003) showed that

$$V \leq 3 \left[-1 + \left(1 + \frac{4}{9} D \right)^{\frac{1}{2}} \right]^{\frac{1}{2}},$$

where $V = \int \left| f - \hat{f} \right| dz$ is the variation measure. Hence, convergence in the Kullback-Leibler distance implies convergence in the variation measure.

Theoretically, one can approximate an unknown continuous distribution arbitrarily well using the maxent density if the number of moment conditions is allowed to increase with sample size. The maximized entropy decreases monotonically with the number of moment conditions. The change in entropy measures the contribution of additional moment conditions in reducing the degree of uncertainty regarding the unknown distribution. For example, a normal distribution is a maxent density completely characterized by its first two moments. Imposing higher order moments does not change the entropy and in that sense, has zero information content.

In practice, only a few moment conditions are used since the Hessian matrix (4) quickly approaches singularity as the number of moment conditions increases. Nonetheless, one can approximate distributions with various shapes using the maxent densities subject to a small number of moment conditions. In this study, we propose a simple yet flexible maxent density for adaptive estimation:

$$f(z, \theta) = \exp\left(-\theta_0 - \theta_1 z - \theta_2 z^2 - \theta_3 \log\left(1 + z^2\right) - \theta_4 \sin(z) - \theta_5 \cos(z) \right). \qquad (5)$$

This density is normal when $\theta_3 = \theta_4 = \theta_5 = 0$. Because z^2 is the dominant term in the exponent of our maxent density, its associated shape parameter θ_2 is restricted to be positive such that the density vanishes at both ends.

The term $\log\left(1+z^2\right)$ is introduced to accommodate fat tails. Note that the fat-tailed student t distribution has a density

$$f\left(z,v\right)=\frac{\left(1+\frac{z^2}{v}\right)^{-\frac{v+1}{2}}}{B\left(.5,.5v\right)\sqrt{v}},$$

where $B(\cdot)$ is the beta function and v is a positive integer shape parameter. Apparently, the t distribution is also a maxent density with characterizing moment $\log\left(1+\frac{z^2}{v}\right)$. In practice, usually the degrees of freedom parameter v is unknown. Direct estimation of v places this unknown parameter on both sides of the moment constraint in the maxent optimization problem:

$$\int\log\left(1+z^2/v\right)\exp\left(\;-\theta_0-\theta_1\log\left(1+z^2/v\right)\right)dz=\frac{1}{n}\sum\nolimits_{i=1}^{n}\log\left(1+z_i^2/v\right),$$

resulting in a difficult saddle point problem. Instead, we use the linear combination of z^2 and $\log\left(1+z^2\right)$ to approximate $\log\left(1+\frac{z^2}{v}\right)$. When the degree of freedom is one, or the distribution is Cauchy, $\log\left(1+z^2\right)$ characterizes the density; on the other extreme, when the degrees of freedom goes to infinity, the t distribution approximates the normal distribution, so z and z^2 characterize the density.

To examine how well z^2 and $\log(1+z^2)$ approximate $\log\ (1+z^2/v)$, we use ordinary least squares to regress $\log\ (1+z^2/v)$ on z^2, $\log(1+z^2)$ and a constant term. Because all functions involved are even, we only look at z on the positive real line. In the experiment, we set z as the vector of all the integers within [1, 10,000]. For an arbitrary integer v within [1, 100], the R^2 is always larger than 0.999, indicating that $\log(1+z^2/v)$ can be well approximated by z^2 and $\log(1+z^2)$.

Compared with the generalized t distribution, our specification has two advantages: i) it nests the normal distribution as a special case rather than a limiting case; ii) it is numerically more stable, as the saddle point problem involved in the estimation of the generalized t family distribution can behave irregularly and does not always converge to a global optimum, especially in the presence of other moment constraints such as terms to capture the degree of asymmetry.

The Sine and Cosine terms are employed to capture skewness and other deviations from the bell shape of symmetric distribution, such as that of normal or t distribution.[3] These two terms introduce considerable flexibility to the density function. For example, multi-modal distributions are allowed for under this specification. The combination of low order polynomial and trigonometric series, referred to as Flexible Fourier Transforms (FFT), was first proposed by Gallant (1981) and shown to approximate curves with various shapes well. For non-periodic functions, the linear and quadratic terms reduce the number of necessary trigonometric terms considerably.

Alternatively, we can use higher order polynomials in the exponent of a maxent density. However, higher order sample moments are sensitive to outliers and consequently, so are the density estimators involved higher moments. Also, Dalén (1987) showed that the sample moment ratios, such as skewness and kurtosis, are restricted by the sample size. In what follows in order to obtain the partially adaptive estimators we will use a number of variant maxent estimators based on the density (5), depending on which terms are included on the right hand side.

3 Partially Adaptive Estimator

Consider the classical linear regression

$$y_i = \alpha_0 + x_i\beta_0 + u_i, \quad i = 1, 2, \ldots, n, \tag{6}$$

where y is the dependent variable, x is a $n \times k$ full-rank design matrix and u is an $i.i.d.$ error which is independent of the regressors.

[3] Since the error terms are generally aperiodic, the domain of the density is scaled to be within $(-\pi, \pi)$.

We first estimate the error distribution, based on the consistent OLS residuals, using the maxent density $f(\cdot,\theta)$. In the second stage, we take the estimated error density $f\left(\cdot,\hat{\theta}\right)$ as given and estimate β using the MLE

$$\tilde{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \ln f\left(y_i - \hat{\alpha} - x_i\beta, \hat{\theta}\right).$$

Under the "block-diagonal" condition that β and θ are independent, we can estimate β adaptively, that is, we can do as well in terms of asymptotic variance as if we knew the true error distribution. The OLS and the Least Absolute Deviation (LAD) estimator fit into this framework when f is the normal and double exponential (Laplace) distribution respectively. Usually, estimate of the intercept α varies with the error distribution and is not identified. Hence, we will focus on the slope vector β below.

Under the "block diagonality" property of condition (1) , there is no loss of asymptotic efficiency in using preliminary estimates of the distributional parameters $\hat{\theta}$ in the final estimation of the slope parameter. Following McDonald and Newey (1988), we denote

$$\begin{aligned}
\hat{u}_i &= y_i - \hat{\alpha} - x_i\hat{\beta}, \\
\hat{l}_{u_i} &= \frac{\partial \ln f\left(\hat{u}_i, \hat{\theta}\right)}{\partial \hat{u}_i}, \\
\hat{\upsilon}_1 &= \left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{l}_{u_i}\right)^2\right] - \left[\frac{1}{n}\sum_{i=1}^{n}\hat{l}_{u_i}\right]^2, \\
\hat{\upsilon}_2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \hat{l}_{u_i}}{\partial u_i}, \\
\hat{\upsilon} &= \frac{\hat{\upsilon}_1}{(\hat{\upsilon}_2)^2}.
\end{aligned}$$

An estimator of the asymptotic covariance matrix of the slope vector $\hat{\beta}$ is given by

$$\hat{\Omega} = \hat{\upsilon}\hat{Q}_x^{-1},$$

where $\hat{Q}_x = \left(\frac{1}{n}\sum_{i=1}^{n} x_i' x_i\right) - \bar{x}'\bar{x}$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Under some mild regularity conditions, McDonald and Newey (1988) proved the asymptotic normality of the estimated slope vector:

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \upsilon Q_x^{-1}\right),$$

where $\hat{\nu} \longrightarrow \nu$ and $\hat{Q}_x \longrightarrow Q_x$. When the error term is distributed according to $f(\cdot, \theta)$, $\hat{\beta}$ is the MLE; otherwise, it is the Quasi-Maximum Likelihood Estimator (QMLE). If the estimated error density approximates the underlying distribution well, the efficiency is expected to be close to that of the MLE.

The partially adaptive estimator offers some advantages over the fully adaptive estimators. The fully adaptive estimation requires nonparametric estimation of the score function (3), which depends crucially on the choice of bandwidth. Generally, the converge rate of a nonparametric estimator is different for the density function itself and its first derivative. Hence, an optimal bandwidth for the density function may not be optimal for its first derivative. As separate derivations of an optimal bandwidth for the density and its first derivative are rather complicated in the adaptive estimation, in practice often one single bandwidth is used for the density estimation and its first derivative. In contrast, a root-n consistent parametric estimate of the density function remains root-n consistent for its first derivative for differentiable densities. Also, nonparametric estimation of the score function (3) encounters numerical difficulties when $f(u_i(\beta))$ in the denominator is close to zero. Some trimming procedures are usually needed to restrict the behavior of this estimator. Our parametric estimator does not suffer from this difficulty and no trimming is required.

4 Simulations

In this section, we use Monte Carlo simulations to investigate the performance of the proposed partially adaptive estimator. For the error distribution, we consider the standard normal, the Laplace distribution, the t distribution with 3 degrees of freedom and the standard log-normal distribution. The Laplace and t distributions are leptokurtic, and the log-normal distribution is both skewed and leptokurtic. All distributions are standardized to have zero mean and unit variance. The explanatory variables, excluding the constant term, are generated as an $n \times k$ matrix of standard normal random variables, with $n = 50$, 100, 200 and 500, and $k = 1, 2, 3$. Altogether, we have 48 possible combinations, and each specification is repeated 5,000 times. We study the classical linear model as specified by Equation (6). Without loss of generality, we set $\alpha_0 = -1$ and $\beta_0 = 1$ for all the Monte Carlo simulations.

We consider the following estimators in our experiments:

- OLS
- LAD: the least absolute deviation estimator. Note that the LAD is the MLE for a Laplace error distribution, which is also a maxent density with a single characterizing moment $E\left|z\right|$.
- FAE: fully adaptive estimator. We use kernels to estimate the error distribution and its first derivative nonparametrically. The bandwidth is chosen according to Silverman's rule of thumb. The trimming conditions set the value of score function to zero if: i) the absolute value of residual $|u_i| > tr_1$; ii) the estimated density $\hat{f} < tr_2$; iii) the value of the 'updating step' $\left|\hat{f}'/\hat{f}\right| > tr_3$. Following Hsieh and Manski (1987), we set $tr_1 = m$, $tr_2 = \exp\left(-m^2/2\right)$ and $tr_3 = m$, where $m = 8$.
- PAE$_1$: partially adaptive estimator with the maxent error distribution

$$f_1\left(z, \theta\right) = \exp\left(-\theta_0 - \theta_1 z - \theta_2 z^2 - \theta_3 \log\left(1 + z^2\right)\right)$$

- PAE$_2$: partially adaptive estimator with the maxent error distribution

$$f_2(z,\theta) = \exp\left(-\theta_0 - \theta_1 z - \theta_2 z^2 - \theta_3 \sin(z) - \theta_4 \cos(z)\right)$$

- PAE$_3$: partially adaptive estimator with the maxent error distribution

$$f_3(z,\theta) = \exp\left(-\theta_0 - \theta_1 z - \theta_2 z^2 - \theta_3 \log\left(1+z^2\right) - \theta_4 \sin(z) - \theta_5 \cos(z)\right)$$

For the partially adaptive estimators, we also estimate the one-step estimator described in McDonald and Newey (1988). The results are very close to those obtained from the iterative estimates and therefore not reported.

Following Phillips (1994), we use the relative inefficiency measure to gauge the efficiency of an alternative estimator $\widetilde{\beta}$ relative to that of OLS estimator $\hat{\beta}$.[4] This measure is defined as

$$RIF\left(\widetilde{\beta}\right) = \mathrm{E}\left\|\widetilde{\beta} - \beta_0\right\|^2 / \mathrm{E}\left\|\hat{\beta} - \beta_0\right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. Because $\mathrm{E}\left\|\widetilde{\beta} - \beta_0\right\|^2 = \sigma_0^2 \mathrm{tr}\left[(X'X)^{-1}\right]$, where $\sigma_0^2 = \mathrm{E}\left(u_i^2\right)$ and $\mathrm{tr}(A)$ is the trace of matrix A, $RIF\left(\widetilde{\beta}\right)$ is invariant to variation of β. The lower RIF is, the more efficient the estimator is compared to the OLS.

The results for the regression with a single explanatory variable are reported in Table 1. As expected, the FAE, which estimates the score function consistently and is asymptotically efficient independent of the functional form of the error distribution, improves with sample size. Despite the fact that its error density only approximates the underlying error distribution in all cases with non-normal error distributions, the PAE$_3$ also improves with sample size, indicating its flexibility in accommodating various error distributions used in the simulations. On the other hand, the LAD, PAE$_1$ and PAE$_2$

[4] Phillips (1994) used a mixture of normal with zero mean and varying variance in the adaptive estimation, focusing on symmetric error distributions.

do not generally improve with sample size. Because of their restrictive functional forms for the error distribution, when the true error distribution differs from the assumed error distribution, those estimators can be severely mis-specified and a larger sample size does not help.

For all the experiments (except for the case of a Laplace error distribution with sample size 500), the PAE$_3$ outperforms the FAE. In most of these cases, the margin is substantial. The nonparametric score estimates by the FAE may be consistent and asymptotically efficient, but the maxent estimates of the error distribution of the PAE$_3$ appear to be flexible enough and perform quite well for small and medium sample size. The PAE$_3$ generally outperforms the PAE$_1$ and PAE$_2$ when the error distributions are non-normal. For normal error distributions, the PAE$_1$ and PAE$_2$ provide better results, but their efficiency gains over the PAE$_3$ are at best marginal.

The pattern of the comparisons varies across the error distributions. For normal errors, as expected, the OLS is efficient and outperforms all other estimators. However, the efficiency loss due to redundant nuisance parameters in the PAEs is rather small. For example, when the sample size $n = 50$, the average efficiency loss of the PAEs is about 10%. When $n = 500$, the average efficiency loss reduces to 2%. Across different sample size, the FAE is less efficient than the PAEs, probably due to the large number of nuisance parameters involved in the nonparametric estimation of the score functions. The efficiency loss is 34% for $n = 50$ and 9% for $n = 500$. The LAD has the largest efficiency loss and does not improve with sample size. While the FAE is asymptotically efficient and all the PAEs' error specifications nest the normal, the assumed error distribution of the LAD is Laplace and does not nest the normal as a special or limiting case. Therefore, the LAD is mis-specified and does not benefit from a larger sample size. Comparing the PAEs, we note that the PAE$_3$ is less efficient than the PAE$_1$ and PAE$_2$. When the underlying error distribution is normal, the PAE$_3$ has more redundant nuisance parameters than the PAE$_1$ and PAE$_2$, but our results suggest that the efficiency loss is quite small. The average efficiency loss of the PAE$_3$ relative to the PAE$_1$ and PAE$_2$ is 7% for $n = 50$, and it reduces to 1% for $n = 500$.

For the Laplace error distribution, the LAD is the Maximum Likelihood estimator and therefore efficient. All the adaptive estimators improve on the OLS except for the FAE with $n = 50$. For sample size no greater than 100, the relative efficiency of LAD compared to that of the PAE$_3$ is less than 7%.

When the errors are generated from the student t distribution, the PAE$_1$, whose assumed error distribution approximates the t distribution closely, performs best for $n = 50$. However, the PAE$_3$, which is more flexible, is more efficient than the PAE$_1$ when the sample size is larger than 50. The LAD also improves on the OLS, largely because of its resistance to outliers as a robust estimator and the fact that its assumed error distribution is more leptokurtic than the normal.

In both of the cases where the error distribution is symmetric and leptokurtic, the PAE$_1$, which is designed for fat-tailed error distributions, outperforms the PAE$_2$ considerably. The FAE improves with the sample size, but it is always less efficient than the PAE$_3$ except for the Laplace error distribution case with $n = 500$.

When the error distribution is generated from the log normal distribution, which is both skewed and leptokurtic, all the estimators improve on the OLS substantially. Across different sample sizes, the average efficiency gain of the PAE$_3$ is about 88%. The PAE$_2$, whose assumed error distribution allows for asymmetric densities, shows a 79% improvement in efficiency. The consistent FAE averages a 57% efficiency gain. On the other hand, the LAD and PAE$_2$, although assuming a symmetric error distribution, also improve on the OLS because they allow for leptokurtic error distributions.

Table 2 and 3 report the regression results with two and three explanatory variables. The general patterns resemble those with a single explanatory variable. Consistent with previous studies, the relative efficiency of all the estimators does not appear to be affected by the number of explanatory variables.

Following one of the referee's suggestions, we also investigate the performance of out-of-sample prediction of the adaptive estimators. Although it is known that the

OLS minimizes the mean square errors, we find that the adaptive estimators often have slightly smaller mean square errors for out-of-sample prediction than that of the OLS when the error distribution is non-normal. Among the adaptive estimators considered in our experiments, the PAE$_3$ is the only one that out-performs the OLS in all cases with non-normal error distribution.

5 Empirical Applications

In this section we apply the proposed partially adaptive estimator to a stochastic frontier model. Stochastic frontier models have been commonly used in the empirical study of firm efficiency and productivity. A production frontier represents the maximum amount of output that can be obtained from a given level of inputs. Similarly, cost frontiers describe the minimum level of cost given a certain output level and certain input prices. In practice, the actual output of a firm will typically fall below the maximum that is technically possible. Hence, these models typically combine two stochastic elements in the specification of the sampling model: one is a symmetric error term, corresponding to the usual measurement error, and another is the one-sided inefficiency term. Due to the presence of the inefficiency term, the distribution of the compounded error term is skewed. Therefore, estimators assuming normal error distribution are not efficient.

To account for the skewed error distribution commonly occurred in production and cost function analysis, Aigner et al. (1977) proposed the original stochastic frontier model

$$y_i = x_i\beta + v_i - |u_i| \,,$$

where v_i and u_i are normally distributed with zero means and constant variance σ_v^2 and σ_u^2. Other commonly used specifications of production frontier analysis model u_i as half normal, truncated normal or exponential. Some researchers noted the restrictive functional form assumption on the inefficiency distribution and proposed further extensions to the original model. For example, Greene (1990) proposed a normal-gamma stochastic frontier model. Although it provides a richer and more flexible parameterization of the inefficiency distribution, the normal-gamma model is practically difficult due to its complicated log likelihood function.

Instead of estimating separate error and inefficiency distributions, we use the proposed partially adaptive estimator for the model

$$y_i = x_i\beta + \varepsilon_i,$$

where $\varepsilon_i = v_i - |u_i|$. We use the maxent density to estimate the potentially non-normal distribution of the composite error ε.

We use data on the production cost of 145 American electricity generating companies from Nerlove (1963), which were also studied by Christensen and Greene (1976). The model takes the form

$$\log\left(\frac{c}{p_f}\right) = \beta_0 + \beta_1 \log(q) + \beta_2 \log^2(q) + \beta_3 \log\left(\frac{p_l}{p_f}\right) + \beta_4 \log\left(\frac{p_k}{p_f}\right) + \varepsilon,$$

where c is total cost, q is total output, p_f, p_l and p_k is the price of fuel, labor and capital respectively, and ε is an *i.i.d.* error term from an unknown distribution. We first estimate the model using the OLS. We then perform the normality test on the OLS residuals. Not surprisingly, normality is rejected decisively by both the Jarque and Bera test and the Kolmogorov-Smirnov test. Therefore, estimators with more flexible error distributions are called for.

In Table 4, we report the estimates from the OLS, the normal, half-normal model and the partially adaptive estimator. The partially adaptive estimates from the PAE$_3$ are generally very close to those from the classical normal, half-normal stochastic frontier model yet the coefficients are estimated more precisely. Compared with the OLS estimates, the partially adaptive estimator reports a larger coefficient for the linear output term (0.268 vs. 0.153) but smaller quadratic coefficient (0.043 vs. 0.051). As for the other two inputs, the coefficient for labor is essentially the same while the coefficient for capital is larger yet estimated more precisely.

6 Concluding Remarks

The classical ordinary least squares estimator is not efficient when the errors are not normally distributed. The adaptive estimation tackles this problem by adapting to the unknown error distribution and maximizing a likelihood function based on an estimate of the error distribution. When the coefficients of the model are independent of the nuisance parameters of the error distribution, one can do as well in terms of asymptotic variance as if one knew the true error distribution.

A fully adaptive estimator requires estimating the score of the likelihood function consistently. In practice, this is achieved through nonparametric estimates of the score function, which might be sensitive to the choice of bandwidth. An alternative procedure is to obtain partially adaptive estimators, which approximate the error distribution parametrically. In this study, we propose a partially adaptive estimator based on certain maximum entropy (maxent) estimates of the error distribution. The maxent densities used in this study have simple functional forms, and at the same time are flexible enough to "adapt" to various distributions. In particular, we show that the more general proposed maxent density works well with skewed and/or leptokurtic distributions, which are frequently encountered in empirical works. Our Monte Carlo simulations and empirical example demonstrate that the proposed estimator achieves a very promising small sample performance and compares favorably with existing methods.

References

Aigner, D., K. Lovell, and P. Schmidt. "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics* 6 (1977): 21-37.

Barron, A.R., and C. Sheu. "Approximation of Density Functions by Sequences of Exponential Families." *Annals of Statistics* 19 (1991): 1347-69.

Beran, R. "Asymptotically Efficient Adaptive Rank Estimates in Location Models." *Annals of Statistics* 2 (1974): 63-74.

Bickel, P.J. "One-Step Huber Estimates in the Linear Model." *Journal of American Statistical Association* 70 (1975): 428-34.

Bickel, P.J. "On Adaptive Estimation." *Annals of Statistics* 10 (1982): 647-71.

Christensen, L.R., and W.H. Greene. "Economies of Scale in U.S. Electric Power Generation." *Journal of Political Economy* 84 (1976): 655-76.

Cobb, L., P. Koppstein and N. Chen. "Estimation and Moment Recursion Relations for Multimodal Distributions of the Exponential family." *Journal of American Statistical Association*, 381-8 (1982): 124-30.

Dalén, J. "Algebraic Bounds on Standardized Sample Moments." *Statistics and Probability Letters* 5 (1987): 329-31.

Gallant, R.A. "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form." *Journal of Econometrics* 15 (1981): 221-46.

Greene, W.H. "A Gamma Distributed Stochastic Frontier Model." *Journal of Econometrics* 46 (1990): 141-64.

Hsieh, D.A., and C.F. Manski. "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression." *Annals of Statistics*, 15 (1987): 541-551.

Jaynes, E.T. "Information Theory and Statistical Mechanics." *Physics Review* 106 (1957): 620-30.

Li, Q., and T. Stengos. “Adaptive Estimation in the Panel Data Error Component Model with Heteroskedasticity of Unknown Form.” *International Economic Review* 35 (1994): 981-1000.

Linton, O.B. “Adaptive Estimation in Arch Models.” *Econometric Theory* 9 (1993): 539-69.

Linton, O.B., and Z. Xiao. “A Nonparametric Regression Estimator that Adapts to Error Distribution of Unknown Form.” Working Paper, 2004.

Manski, C.F. “Adaptive Estimation of Non-Linear Regression Models.” *Econometric Reviews* 3 (1984): 145-94.

McDonald, J.B., and W.K. Newey. “Partially Adaptive Estimation of Regression Models Via the Generalized T Distribution.” *Econometric Theory* 4 (1988): 428-57.

McDonald, J.B., and S.B. White. “A Comparison of Some Robust, Adaptive, and Partially Adaptive Estimators of Regression Models.” *Econometric Reviews* 12 (1993): 103-124.

Nerlove, M. “Returns to Scale in Electricity Supply.” In *Measurement in Economics–Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld,* edited by Carl F. Christ. Stanford University Press, 1963.

Newey, W.K. “Adaptive Estimation of Regression Models Via Moment Restrictions.” *Journal of Econometrics* 38 (1988): 301-39.

Phillips, R.F. “Partially Adaptive Estimation Via a Normal Mixture.” *Journal of Econometrics* 64 (1994): 123-44.

Steigerwald, D.G. “On the Finite Sample Behavior of Adaptive Estimators.” *Journal of Econometrics* 54 (1992): 371-400.

Stein, C. “Efficient Nonparametric Testing and Estimation.” *Proceeding of Third Berkeley Symposium on Mathematical Statistics and Probability* 1 (1956): 187-96.

Stone, C. “Adaptive Maximum Likelihood Estimators of a Location Parameter.” *An-*

nals of Statistics 3 (1984): 267-84.

Tagliani, A. “A Note on Proximity of Distributions in Terms of Coinciding Moments.” *Applied Mathematics and Computation* 145 (2003): 195-203.

Wu, X. “Calculation of Maximum Entropy Densities with Application to Income Distribution.” *Journal of Econometrics* 115 (2003): 347-54.

Wu, X. and J.M. Perloff. “GMM Estimation of Maximum Entropy Density with Interval Data.” *Journal of Econometrics* (forthcoming).

Zellner, A., and R.A. Highfield. “Calculation of Maximum Entropy Distribution and Approximation of Marginal Posterior Distributions.” *Journal of Econometrics* 37 (1988): 195-209.

Table 1: Relative efficiency for regression with one explanatory variable

n		Normal	Laplace	T-3	Log-normal
50	LAD	1.560	0.744	0.854	0.536
	FAE	1.340	0.957	0.910	0.541
	PAE1	1.056	0.793	0.740	0.536
	PAE2	1.094	0.942	0.878	0.205
	PAE3	1.148	0.795	0.762	0.154
100	LAD	1.630	0.666	0.767	0.447
	FAE	1.194	0.830	0.758	0.426
	PAE1	1.028	0.758	0.667	0.511
	PAE2	1.055	0.892	0.815	0.198
	PAE3	1.102	0.717	0.649	0.118
200	LAD	1.551	0.620	0.751	0.408
	FAE	1.130	0.751	0.685	0.359
	PAE1	1.019	0.753	0.656	0.504
	PAE2	1.028	0.889	0.825	0.201
	PAE3	1.050	0.708	0.618	0.109
500	LAD	1.624	0.573	0.643	0.382
	FAE	1.094	0.667	0.604	0.378
	PAE1	1.011	0.757	0.660	0.553
	PAE2	1.020	0.891	0.824	0.236
	PAE3	1.029	0.700	0.599	0.118

LAD: least absolute deviation estimator

FAE: fully adaptive estimator

PAE: partially adaptive estimator

Table 2: Relative efficiency for regression with two explanatory variables

n		Normal	Laplace	T-3	Log-normal
50	LAD	1.545	0.783	0.857	0.512
	FAE	1.341	0.992	0.950	0.509
	PAE1	1.055	0.812	0.736	0.525
	PAE2	1.092	0.936	0.858	0.206
	PAE3	1.132	0.807	0.761	0.167
100	LAD	1.590	0.708	0.787	0.439
	FAE	1.205	0.863	0.765	0.385
	PAE1	1.031	0.771	0.666	0.486
	PAE2	1.052	0.906	0.813	0.194
	PAE3	1.087	0.740	0.646	0.118
200	LAD	1.553	0.642	0.709	0.397
	FAE	1.125	0.775	0.668	0.304
	PAE1	1.020	0.769	0.651	0.492
	PAE2	1.034	0.902	0.816	0.202
	PAE3	1.060	0.725	0.607	0.110
500	LAD	1.557	0.573	0.694	0.376
	FAE	1.089	0.676	0.630	0.302
	PAE1	1.007	0.755	0.683	0.535
	PAE2	1.013	0.887	0.858	0.232
	PAE3	1.017	0.693	0.621	0.119

LAD: least absolute deviation estimator

FAE: fully adaptive estimator

PAE: partially adaptive estimator

Table 3: Relative efficiency for regression with three explanatory variables

n		Normal	Laplace	T-3	Log-normal
50	LAD	1.564	0.825	0.850	0.497
	FAE	1.361	1.037	0.961	0.528
	PAE1	1.055	0.821	0.742	0.546
	PAE2	1.093	0.949	0.857	0.227
	PAE3	1.152	0.840	0.773	0.194
100	LAD	1.569	0.705	0.790	0.448
	FAE	1.211	0.852	0.769	0.373
	PAE1	1.034	0.764	0.666	0.490
	PAE2	1.054	0.902	0.811	0.198
	PAE3	1.091	0.737	0.653	0.123
200	LAD	1.550	0.642	0.723	0.395
	FAE	1.149	0.784	0.690	0.275
	PAE1	1.021	0.761	0.658	0.483
	PAE2	1.030	0.896	0.824	0.199
	PAE3	1.060	0.723	0.621	0.107
500	LAD	1.574	0.572	0.669	0.370
	FAE	1.085	0.686	0.636	0.241
	PAE1	1.006	0.771	0.683	0.522
	PAE2	1.010	0.890	0.872	0.230
	PAE3	1.020	0.710	0.625	0.117

LAD: least absolute deviation estimator

FAE: fully adaptive estimator

PAE: partially adaptive estimator

Table 4. Cost function estimation

	Intercept	log(q)	log2(q)	log(pl/pf)	log(pk/pf)
OLS	-3.764	0.153	0.051	0.481	0.074
	0.702	0.062	0.005	0.161	0.150
NHN	-4.488	0.268	0.043	0.479	0.084
	0.719	0.085	0.006	0.150	0.141
PAE	-3.998	0.207	0.047	0.491	0.098
	0.559	0.073	0.006	0.120	0.109

NHN: normal, half-normal stochastic frontier model

PAE: partially adaptive estimator

Standard errors below coefficients for each estimator