Risk & Sustainable Management Group

Risk & Uncertainty Program Working Paper: R07#1

Research supported by an Australian Research Council Federation Fellowship
http://www.arc.gov.au/grant_programs/discovery_federation.htm

Estimating complex production functions: The importance of starting values

Mark Neal
University of Queensland

Schools of Economics and Political Science
University of Queensland
Brisbane, 4072
rsmg@uq.edu.au
http://www.uq.edu.au/rsmg

Estimating complex production functions: The importance of starting values

Mark Neal

Postdoctoral Research Fellow
Risk and Sustainable Management Group
University of Queensland

Room 519, School of Economics, Colin Clark Building (39)
University of Queensland, St Lucia, QLD, 4072

Email: m.neal@uq.edu.au
Ph: +61 (0) 7 3365 6601
Fax: +61 (0) 7 3365 7299
http://www.uq.edu.au/rsmg

Acknowledgements

Chris O'Donnell helpfully provided access to Gauss code that he had written for estimation of latent class models so it could be translated into Shazam. Leighton Brough (UQ), Ariel Liebman (UQ) and Tom Pechey (UMelb) helped by organising a workshop with Nimrod/enFuzion at UQ in June 2006. enFuzion (via Rok Sosic) provided a limited license for use in the project. Leighton Brough, tools coordinator for the ARC Centre for Complex Systems (ACCS), assisted greatly in setting up an enFuzion grid and collating the large volume of results. Son Nghiem provided assistance in collating and analysing data as well as in creating some of the figures.

ABSTRACT

Production functions that take into account uncertainty can be empirically estimated by taking a state contingent view of the world. Where there is no *a priori* information to allocate data amongst a small number of states, the estimation may be carried out with finite mixtures model. The complexity of the estimation almost guarantees a large number of local maxima for the likelihood function. However, it is shown, with examples, that a variation on the traditional method of finding starting values substantially improves the estimation results. One of the major benefits of the proposed method is the reliable estimation of a decision maker's ability to substitute output between states, justifying a preference for the state contingent approach over the use of a stochastic production function.

Keywords:
Production function; econometrics; starting values; state contingent production.

Abbreviations:
EM=Expectation Maximisation; P=Probability; LL=Log likelihood;

1. INTRODUCTION

Production functions are used in the analysis of the relationship between agricultural inputs and outputs. The estimation of production functions are somewhat complicated by factors such as the poor availability of data, the existence of inefficiency and the existence of a risky operating environment. In the case of inefficiency, a production frontier may be estimated assuming that deviation from this frontier are partly random error and partly due to inefficiency. A common example of this is the stochastic production frontier developed by Aigner, Lovell and Schmidt (1977) and Meeusen and van der Broeck (1977) which simultaneously estimates the efficiency of individual firms in the sample.

Chambers and Quiggin (2000) provide a substantial critique of the stochastic production function (and frontier) because it does not appropriately take into account price or production uncertainty. They envisage a world where producers face many possible states of nature, where a state of nature is associated with particular levels of production and prices. Their critique is based on the notion that a stochastic production frontier is estimated while implicitly assuming only one state of nature could occur, and so deviations are either random or due to inefficiency. This ignores the possibility that deviations from this frontier are potentially deliberate as a response to the set of possible states that may have occurred *ex ante*. In other words, a producer could rationally decide to produce less than another producer in the state that is observed *ex post* because they committed inputs to achieve higher production in a different state *ex ante*. Hence they appear inefficient, having produced less than another producer for the observed state when it may be a result of rational decisions. This notion is shown with examples by O'Donnell et al. (2006).

Although Chambers and Quiggin (2000) provide reasons for approaching production function and production frontier estimation from the state contingent framework, empirical estimation in this framework is currently very limited. One example is O'Donnell and Griffiths (2006) which uses a latent class or finite mixtures model in a Bayesian approach to estimate efficiency of Phillipine rice farmers. O'Donnell (2006) provides another example of estimating latent class models to approximate an 'unknown' production function from simulated data. This estimation procedure requires the maximisation of a complex likelihood function and is described in more detail in a following section. Due to this complexity, it is likely that there are a number of local maxima in the likelihood function and as a response the Expectation Maximisation (EM) method is employed. However, the starting point of the process may still affect the estimates of important economic parameters such as conditional means and marginal products.

Starting points in econometric procedures have been researched in various contexts. For example, Tonsor and Kastens (2006) examine the impact of starting points on the estimates of meat demand models. They find that as the econometric task becomes increasingly nonlinear, starting conditions become increasingly important. Bond et al. (2005) examine the impact of changes in initial parameter estimates in the context of a nonlinear optimisation for Phillips curve estimation. They find that the results are highly sensitive to the data set and algorithm choice and suggest the existence of local maxima as a reason. St Pierre (1998) found that the EGARCH-M model is sensitive to the choice of starting values, particularly on computers with a high degree of precision.

The purposes of this paper are four-fold. First, it aims to demonstrate the effect of different starting points for latent class estimation when using the EM algorithm. Second, it proposes a method that improves the chances of finding a high value for the likelihood function. Third, it demonstrates how checking the impact of starting points can be done quickly by using a computing grid. Fourthly, it outlines difficulties in a potential application of estimating state contingent production frontiers using latent class models.

2. LATENT CLASS ESTIMATION

The concept of latent class estimation is briefly outlined below in raletion to estimating production functions, but described in detail by McLachlan and Peel (2000) for a range of applications. As an example, a linear equation could be used to estimate a production function:

$$y_i = z_{im}\beta_m + e_i \qquad (1)$$

where y_i is the *i*th output, z_i is the matrix of *i* observations on *m* inputs, β_m is the vector of *m* coefficients, and e_i is the error term. This equation implies that the same relationship applies to all observations. However, it might be possible that there are several different underlying relationships. In this case, it may be known *ex ante* which data observations apply to each relationship. For example, if the data observations could be split into *j* groups (or classes), and a separate relationship was to be estimated for each group, equation 1 could be estimated j times using the subsets of the data in separate estimations. Alternatively, one equation could be estimated:

$$y_i = 1_j'[z_{ik}\beta_{mj}.*.t_{ij}] + e_i \qquad (2)$$

where β_{mj} is the matrix of *m* coefficients by *j* classes, t_{ij} is a matrix of *i* observations by *j* classes in which t_{ij}=1 if observation *i* belongs to class *j* and zero otherwise, .*. is the operator representing element by element (or Hadamard) multiplication and 1_j is a vector of ones of length *j*.

If there is no information available ex ante regarding which class the data observations belong to (ie t_{ij} is unknown), a latent class model can be used to simultaneously estimate the probability of a data point belonging to a class and the coefficients associated with each class.

The latent class model can be estimated by maximum likelihood methods, and the log likelihood calculation is:

$$Ln(L) = \sum_{i=1}^{N} Ln\left(\sum_{j=1}^{J} \pi_j p(y_i \mid z_i, \beta_{mj}, \sigma^2) \right) \qquad (3)$$

where π_j is the unobserved prior probability of random variable y_i belonging to the jth class. While direct maximisation of the likelihood function is possible, in practise it is difficult. This is because the likelihood function is not usually concave, having multiple local maxima, and the large number of parameters may make standard algorithms unreliable (O'Donnell, 2006). An alternative approach is to use the EM algorithm of Dempster, Laird and Rubin (1977), which is discussed in more detail in the next section.

3. EM ALGORITHM

The EM algorithm works by looping iteratively through an expectation step and a maximisation step until some stopping criterion is met. The expectation step is based on the expectation of the LL (conditional on the data), which assuming that π_j and θ_j are known, is expressed as:

$$E\{Ln(L) \mid y_1,...,y_N\} = \sum_{j=1}^{J} \sum_{i=1}^{N} \tau_{ij}[\ln \pi_j + \ln f_j(y_i \mid \theta_j)] \qquad (4)$$

where τ_{ij} is the posterior probability of random variable y_i belonging to the jth class, and θ_j is the matrix of parameters for the model, including β_{mj} and σ^2. An estimate for τ_{ij} can be found as:

$$\tau_{ij} = \frac{\pi_j^{(k)} f_j(y_i \mid \theta_j^{(k)})}{\sum_{s=1}^{J} \pi_s^{(k)} f_s(y_i \mid \theta_s^{(k)})} \qquad (5)$$

where the superscript k represents the estimates at the kth iteration of the EM algorithm. Thus an estimate of equation 4 (for the kth step of the algorithm) can be given by:

$$Ln(L)^{(k)} = \sum_{j=1}^{J} \sum_{i=1}^{N} \tau_{ij}^{(k)}[\ln \pi_j + \ln f_j(y_i \mid \theta_j)] \qquad (6)$$

The maximisation step involves maximising equation 6 with respect to π_j and θ_j in order to calculate updated values for these variables. To update the estimate for π_j:

$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^{N} \tau_{ij}^{(k)}}{N} \tag{7}$$

To update the estimate for θ_j requires a further maximisation of:

$$S = \sum_{j=1}^{J} \sum_{i=1}^{N} \tau_{ij}^{(k)} \ln f_j(y_i \mid \theta_j^{(k)}) \tag{8}$$

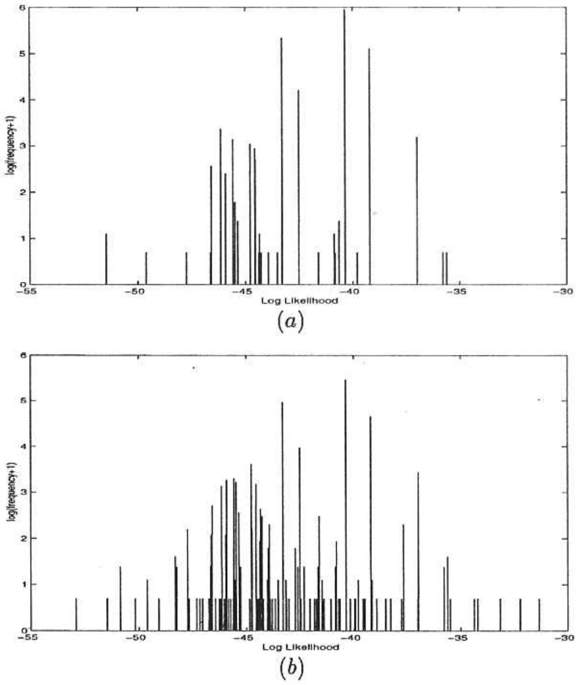
with respect to $\theta_j^{(k)}$.

The starting point for the EM algorithm can influence the results in important ways such as affecting coefficient estimates and the outcomes of likelihood ratio tests (eg Seidel, Mosler and Alker, 2000). One of the most common approaches to a starting point depends on specifying initial values for τ_{ij}, where each of the N data observations is initially allocated to one of the J classes (ie τ_{ij}=1 if observation i belongs to class j and zero otherwise).

McLachlan and Peel (2000) suggest eight methods in which this initial allocation can be done, or alterations can be made to the EM algorithm to assist recovery from a poor starting point. First, an *ad hoc* method is to plot two dimensions of the data and divide the bivariate plot into two groups. Second, a clustering algorithm or third, a hierarchical procedure could be used to initially group the data. Fourth, an allocation can be based on based on random sampling, assuming that each data observation belongs to the jth class with equal probability. Fifth, a subsample of the data is assigned randomly to the J components (as for the previous method), and the initial M-step is performed on the subsample. Sixth, for univariate mixtures, the initial mixing proportions can be determined from a quantile-quantile plot, with random assignment based on these probabilities. Seventh, a deterministic annealing EM algorithm, employing the concept of maximum entropy, can assist in recovery from a poor starting point. Finally, a stochastic EM algorithm may be used to escape convergence paths from poor starting points as it randomly assigns each observation outright to one of the classes at every E-step.

McLachlan and Peel (2000) make several points about the usefulness of the different methods. The allocation based on random sampling (equal probability for each class) can lead to the component parameters being similar, which may in turn lead to a suboptimal estimation result. They propose the subsample approach as a potential improvement, with simulation results about the frequency of finding local maxima of the LL function (Figure 1). They conclude that the latter approach found a wider range of local maxima, but also believe some additional local maxima correspond to spurious local maxima. A spurious local maximum is one where a class has a very low (but nonzero) variance or generalised variance as a result of only containing a few data points lying close together (or in a lower dimensional subspace).

Figure 1: Frequencies of the different local maxima found based on a random assignment into g groups using (a) all the data and (b) a subsample of the data



Source: McLachlan and Peel (2000)

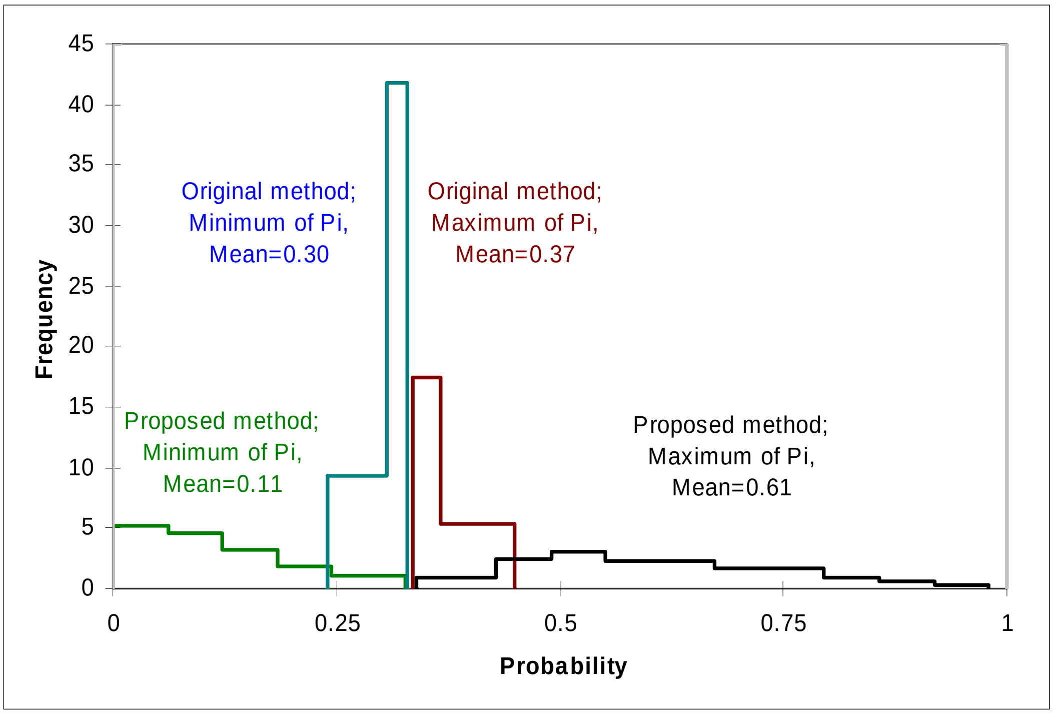
McLachlan and Peel (2000) also suggest that the two methods of altering the EM algorithm either have a tendency to find spurious maximisers or that their ability to find improved optima can be achieved more simply by running the standard EM algorithm from a few different starting points. Other methods such as the *ad hoc* and quantile-quantile method may not be appropriate for higher dimensional data or models with larger numbers of classes.

Karlis and Xedelaki (2003) test a number of different methods for assigning starting points with simple experiments with finite mixtures of normal distributions and then finite mixtures of Poisson distributions. Their general results confirm earlier work that suggests it is preferable to start from several different starting points. Two of the more successful methods they tested were a variation on the method of Finch et al. (1989) and a “best of ten” strategy. The Finch method was designed for two-component normal mixtures and estimates the mixing proportions before assigning p_j*N observations to each class J in a sequential manner. The best of ten method calculates the LL value at the first iteration for 10 starting points and uses the starting point that generates the highest LL value.

The set of all possible starting points is quite large. As explained previously, the starting point ($\tau_{ij}^{(0)}$) can be specified by an initial allocation of each observation to one of the j classes. This means there are theoretically j^N possible starting points. As an example, with 200 observations across three classes, there would be 3^{200} (ie $2*10^{95}$) possible starting points, far more than could ever be enumerated. Starting points as a mixed allocation could also be considered. For a mixed allocation, an observation is not allocated outright to one class, but with some probability to each class. If, instead of starting points with outright allocations, mixed allocations were also considered, the potential starting points are much larger, and are only finite if probabilities are measured to some limited precision.

The manner in which the allocation takes place makes a large impact on the types of starting points that are selected. Two methods of allocation are explored in this paper. The method of allocation whereby each data observation belongs to the jth class with equal probability will be referred to as the original method. A new method, where the probability of belonging to each class is determined before allocating observations based on those probabilities will be referred to as the proposed method. To illustrate the differences between these methods in establishing a starting point, 100 starting points were generated for each method, assuming 200 data observations and 3 classes. Figure 2 is a histogram outlining the distribution of the maximum and minimum probability for each method. The figure shows that the original method generates most starting points with minimum probabilities in the range of 0.23 and 0.33, and maximum probabilities in the range 0.33 to 0.45. In contrast, the proposed method generated minimum probabilities from zero to 0.33, and maximum probabilities from 0.33 to one. In other words, the proposed method is likely to generate a wider range of possible starting points. This wider range of starting points is likely to be of benefit if different starting points lead to convergence at different (potentially suboptimal) LL maxima.

Figure 2: Distribution of initial probabilities: Original and proposed methods



One potential problem with the proposed method is that it can generate an initial allocation of data points to a class that is too small to generate the parameter estimates for the equations of that class. In this case, a singular covariance matrix can occur, and no model can be estimated. Another potential problem with the proposed method is that it could lead to spurious maximisers. This could be identified with classes that have a low minimum probability (ie there are few data points in the class) and there is a suspiciously low standard error for the class (relative to the standard error for other classes).

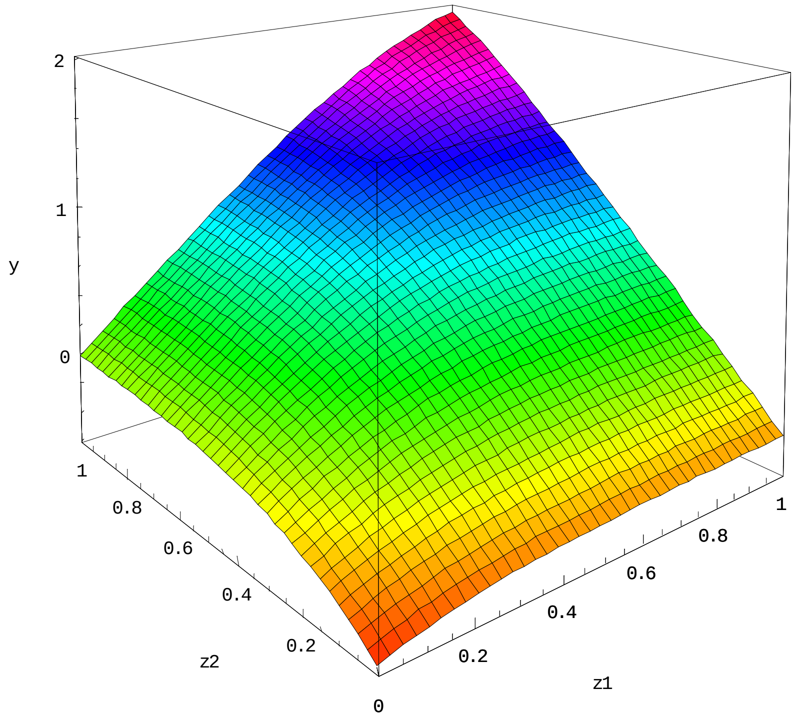
4. AN EXAMPLE

An example of the latent class estimation process was carried out on simulated data generated by O’Donnell (2006). The simulated data was based on the two input production function proposed by White (1980) shown in equation x.

$$y_i = -\frac{1}{\gamma_1}\ln\left(e^{-\gamma_1 . z_{1i}} + \gamma_1 e^{-\gamma_1 . z_{2i}}\right) \qquad (9)$$

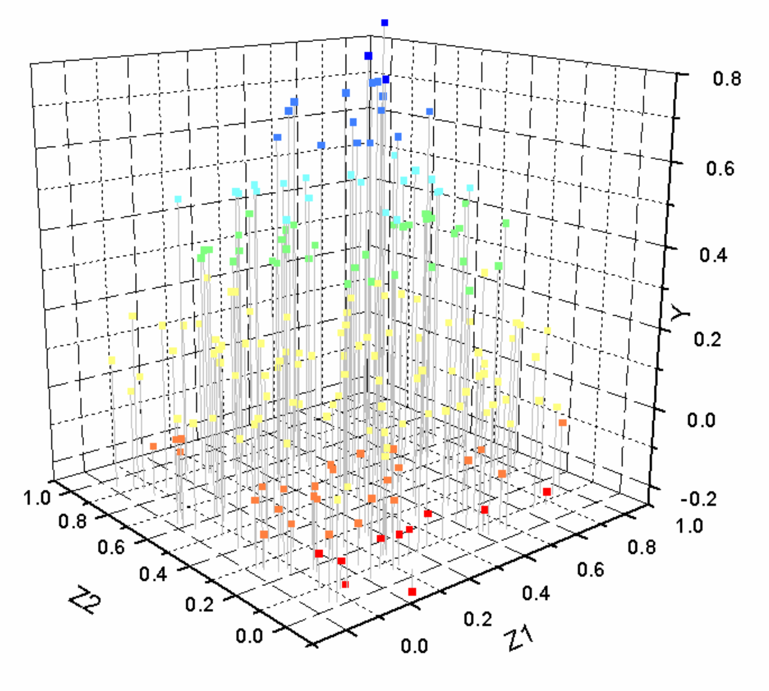
where y_i represents the *i*th observation of the log of output and z_{ki} represents the *i*th observation of the log of the *k*th input. The number of inputs *k* was set at two and the parameters of the production function, γ_1 and γ_2, were set to 5 and 2 respectively. The production surface of the two-input production function is shown in figure 3.

Figure 3: Two-input production function: Production surface



The first observation on the log of output (y_1) was generated assuming that the log of inputs (z_1 and z_2) were equal to 0.5. A further 199 observations on log output were obtained by drawing values for the log of inputs z_1 and z_2 from a uniform distribution between zero and one. This data represented observations without noise. A noise component (e_i~N(0, 0.01)) was added to these observations assuming to produce a data set that included noise. The data including the noise component is plotted in figure 4.

Figure 4: Simulated (noisy) data from the two-input production function



The latent class estimation was used to estimate the true functional form as if it was unknown, as per O’Donnell (2006). Three classes were assumed with each class being a translog functional form. The EM algorithm was used as described in the previous section. The starting points were generated by the original method of initially assigning data observations to a class randomly, and the proposed method of determining initial probabilities before assigning data observations to each class based on these initial probabilities. The former method implicitly assumes initial probabilities near 1/3 for large samples, while the latter method makes no such assumption.

The stopping criterion for the EM algorithm was based on an Aitken acceleration-based stopping criterion (as proposed by McLachlan and Peel (2000)) where the tolerance was set to $1*10^{-11}$. However, the econometric software used (Shazam) may be limited by the precision of floating point numbers, and so the actual tolerance could be closer to $1*10^{-5}$. The analysis generated 60,000 starting points for each of the four cases: Original method on data without noise, proposed method on data without noise, original method on data including noise, and proposed method on data including noise.

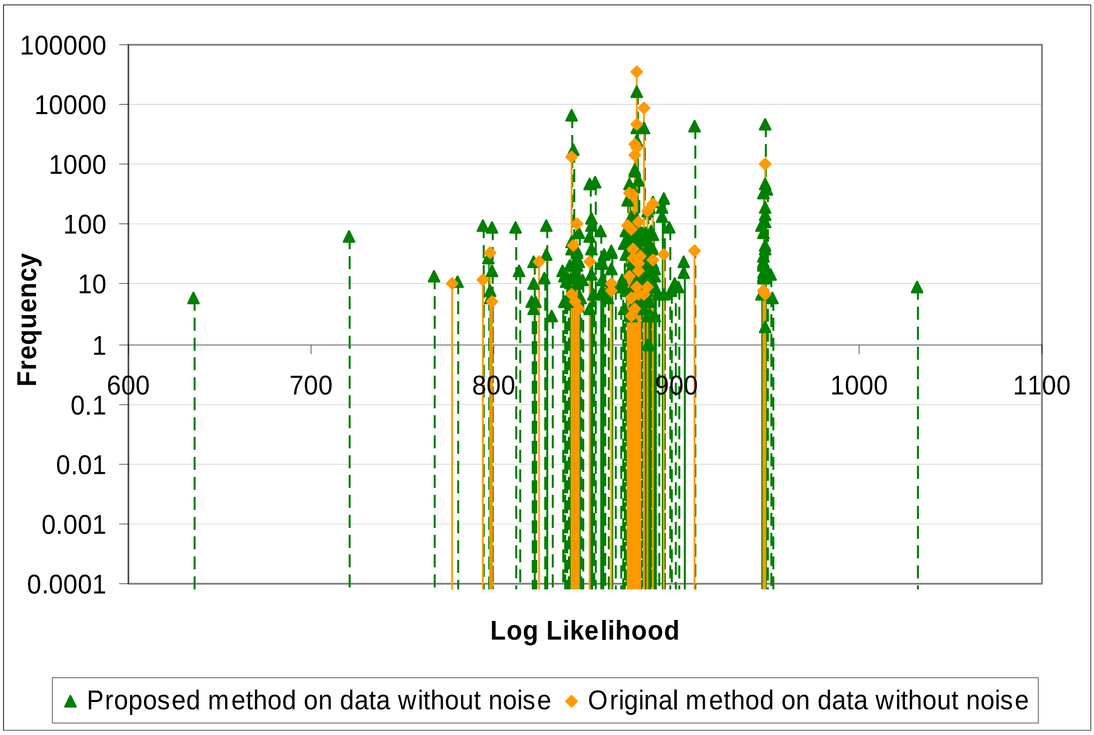
4. RESULTS

Results are presented in two parts. First, for the estimation of a latent class model on the data that does not have a noise component, and second, on the data that does include the noise component.

Data without noise

The EM algorithm converged at a number of different estimates for model parameters, regardless of whether the original or proposed method was used. The EM algorithm frequently converged at local optima for the likelihood function, regardless of the method for assigning a starting point, as shown in Figure 5. Both methods generated the same mode for the converged log likelihood function, accounting for 28% and 61% of the starting points from the proposed and original methods respectively. However, the proposed method found a higher value for the log of the likelihood function and converged at optima higher than the mode more frequently that the original method.

Figure 5: Local optima for data without noise: Original method and proposed method



Following O’Donnell (2006), the results from the estimation process were used to estimate the conditional mean *m(z)*, marginal product of input 1 $m_1(z)$ and marginal product of input 2 $m_2(z)$ for the unknown function at observation one (input means) and observation 49. Table 1 presents these estimates from the most frequently found optimum, the best found by the original method and the best overall log likelihood value. Partly due to the lack of noise in the data, the standard errors are very low. All estimates are significantly different (P<0.05) from the true values, with the proposed method having the lowest bias for estimating the marginal products at both data points. The conditional mean is best estimated (lowest bias) by the original method at data point 1, but the most frequently found optimum has the lowest at data point 49. Importantly, the estimates of the economic quantities are all statistically different from each other (P<0.05).

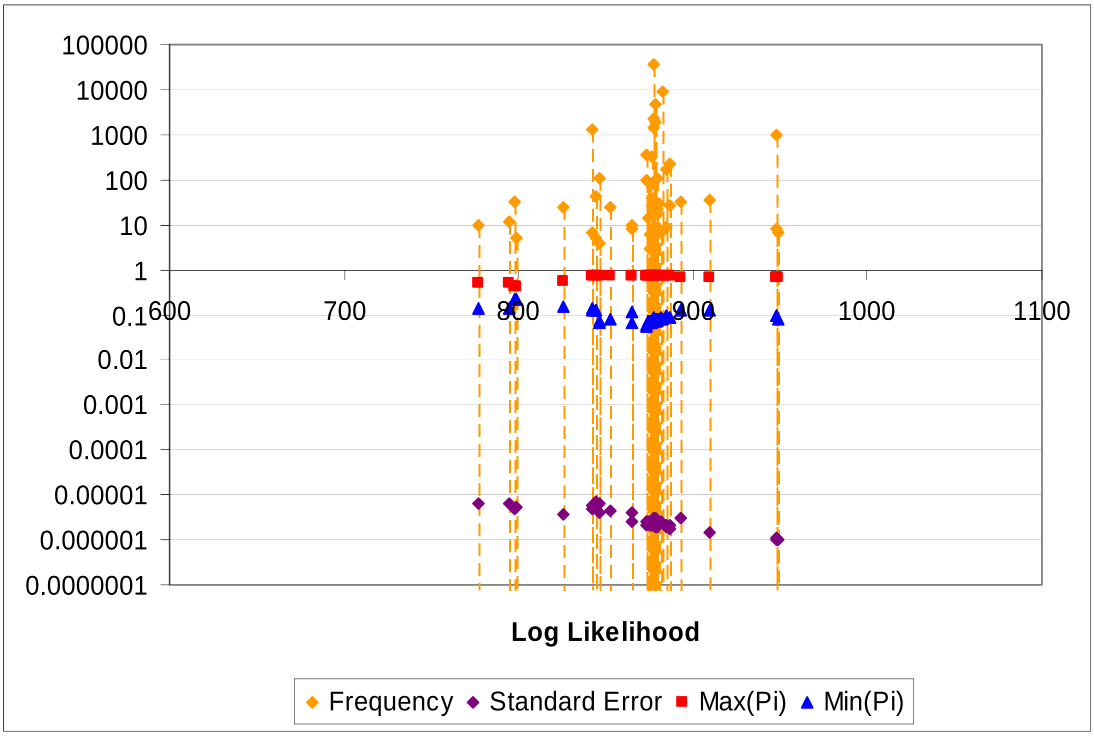
Table 1: Estimates of economic quantities at various local optima for data without noise

	Evaluated at z_1 = (0.5,0.5)								
	$m(z_1)$			$m_1(z_1)$			$m_2(z_1)$		
	True value = 0.280			True value = 0.333			True value = 0.667		
	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall
Estimate	0.266	0.285	0.298	0.364	0.297	0.327	0.630	0.706	0.673
Standard Error	*0.0001*	*0.0001*	*0.0001*	*0.0003*	*0.0002*	*0.0001*	*0.0003*	*0.0002*	*0.0001*
Bias	-0.0141	0.0046	0.0179	0.0309	-0.0359	-0.0066	-0.0365	0.0393	0.0059
RMSE	0.0141	0.0047	0.0179	0.0309	0.0359	0.0066	0.0365	0.0393	0.0059

	Evaluated at z_{49} = (0.05,0.03)								
	$m(z_{49})$			$m_1(z_{49})$			$m_2(z_{49})$		
	True value = -0.184			True value = 0.308			True value = 0.692		
	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall
Estimate	-0.183	-0.177	-0.166	0.313	0.266	0.311	0.632	0.716	0.690
Standard Error	*0.0003*	*0.0002*	*0.0001*	*0.0010*	*0.0006*	*0.0003*	*0.0011*	*0.0006*	*0.0004*
Bias	0.0006	0.0071	0.0178	0.0051	-0.0413	0.0032	-0.0602	0.0236	-0.0023
RMSE	0.0007	0.0071	0.0178	0.0052	0.0413	0.0032	0.0603	0.0236	0.0023

Using the original method, the frequency of convergence at each LL value is shown in figure 6 together with the standard error as well as the maximum and minimum probability of an observation being within each class. The standard error can be seen to decline fairly steadily with higher values of the LL function, but the maximum and minimum probability do not vary systematically with the increases in the LL value. The observation that the minimum probability does not fall far below 0.10 suggests that few of the local maximisers are spurious. This tends is confirmed by reference to the standard error of each class which are of a similar magnitude.

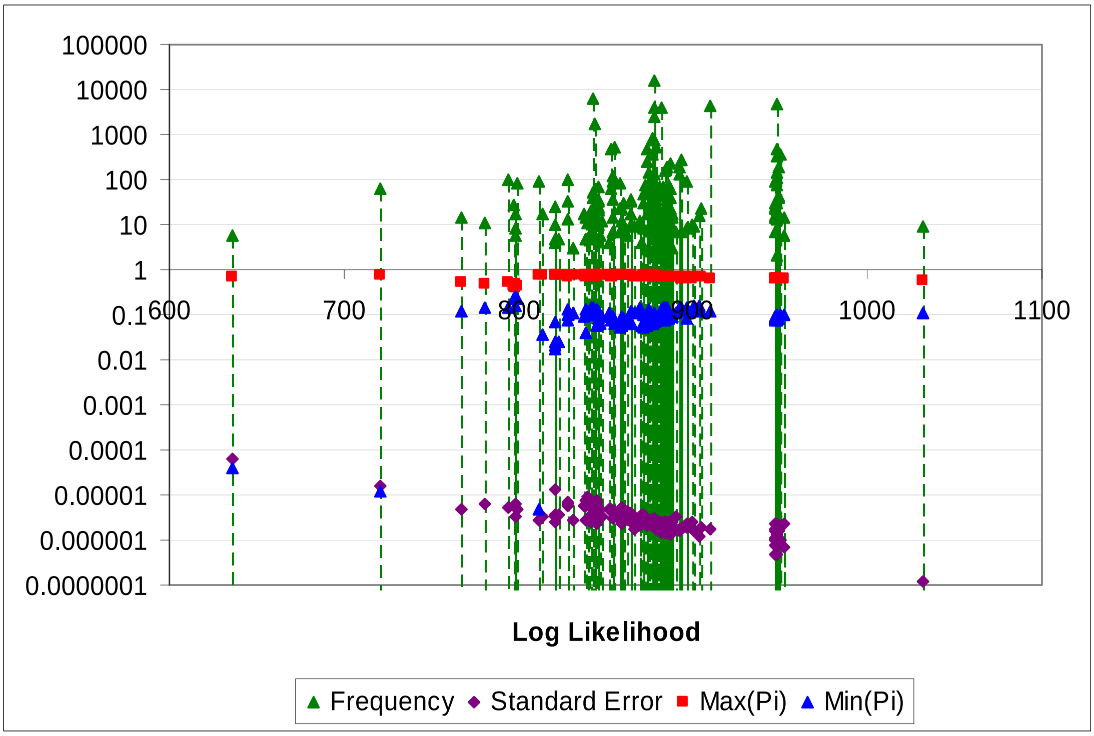
Figure 6: Original method on data without noise: Log likelihood frequency, standard error, maximum and minimum probability of local optima.



Using the proposed method, the frequency of convergence at each LL value is shown in figure 7 together with the standard error as well as the maximum and minimum probability of an observation being within each class. Comparing figure 6 with figure 7 shows that using the proposed method generates much more discrete local optima. Similarities in figure 7 with figure 6 include the tendency for a reduction in the standard error as the LL value increases and no obvious relationship between the probability and the LL value. There are some local maximisers that are likely to be spurious, as shown by the low minimum probabilities, but the local maximisers with the LL values higher than the mode do not appear to be spurious based on the criterion of low minimum probabilities. The maximum overall likelihood value was not associated with a suspiciously low minimum probability (0.1099), and the standard error for that class was similar to the standard errors for the other classes.

Furthermore, the original method had a lower minimum probability (0.0813) and a wider range between the class standard errors.

Figure 7: Proposed method on data without noise: Log likelihood frequency, standard error, maximum and minimum probability of local optima.



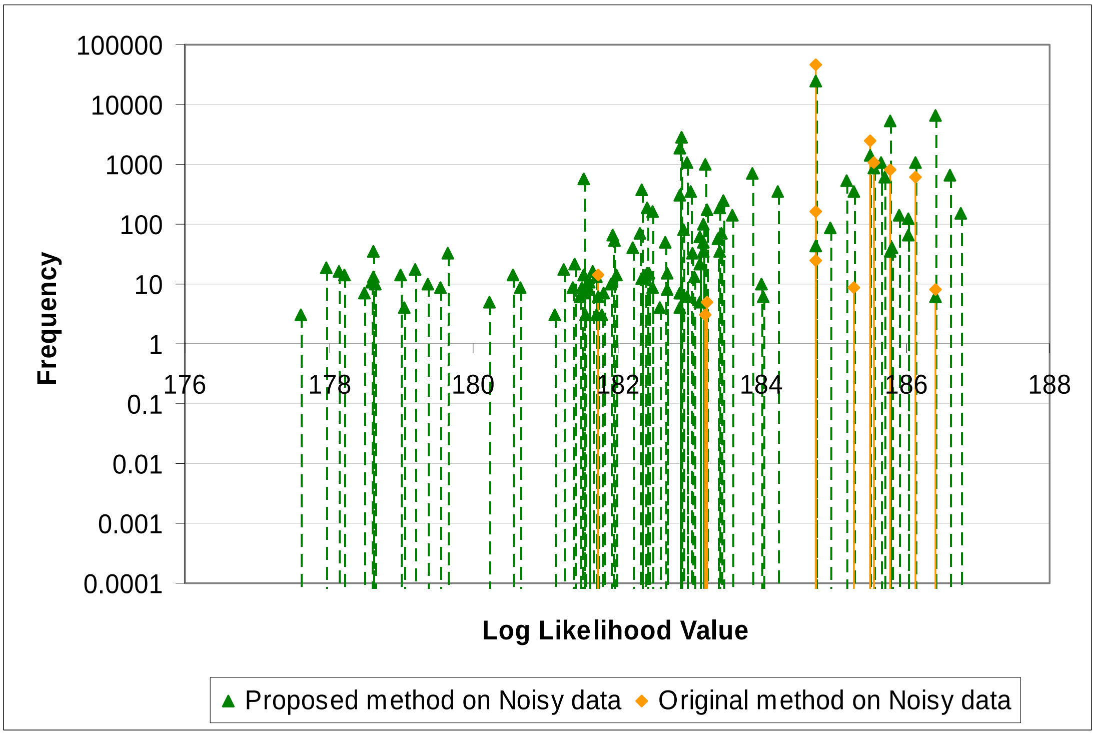
As the proposed method generated starting points that had classes with a low number of data points (due to a low minimum probability), it was more likely to fail due to encountering a singular covariance matrix, either in the first or subsequent iterations. This failure occurred using 6% of the starting points generated by the proposed method, and for none of the starting points generated by the original method.

In summary, using data without noise, the proposed method provided higher LL values, and it did not appear that the higher LL value was a spurious maximiser. This lead to lower bias in estimating the marginal products (measured at two points) but not for the conditional mean. The estimates provided by the most frequent local optimum, the highest LL optimum under the original method and the highest LL optimum under the proposed method were all significantly different from each other.

Data including noise

Similarly to when the data did not include noise, both the original or proposed method resulted in the EM algorithm converging at a number of different estimates for model parameters when the data included a noise component. The EM algorithm frequently converged at local optima for the likelihood function, regardless of the method for assigning a starting point, as shown in Figure 8. Both methods generated the same mode for the converged log likelihood function, accounting for 41% and 90% of the starting points from the proposed and original methods respectively. However, the proposed method found a higher value for the log of the likelihood function and usually converged at optima higher than the mode more frequently that the original method. The proposed method also converged at many discrete sub optima, some of which had much lower LL values than the lowest from the original method.

Figure 8: Local optima for noisy data: Original method and proposed method



From the additional local maxima associated with the proposed method, it is important to consider whether these would be spurious local maximisers. The figures A1 and A2 in appendix A show the local optima from figure 8 with the respective maximum and minimum probabilities. Both methods may have spurious maximisers, evidenced by quite low minimum probabilities (for a minority of the local maximisers). Despite this, the overall LL maximum (found with the proposed method) was not likely to be a spurious maximiser, having a minimum probability of 0.1997 and standard errors for each class that were of a similar size.

Table 2 shows that there is a large improvement in the LL value by moving from the most frequently found optimum to the highest optimum using the original method, with a smaller incremental improvement when moving from the latter optimum to the highest found with the proposed method. A similar effect is seen with the reduction in standard error. The probabilities change by 5 to 10% when moving from the most frequently found optimum to the highest optimum from the original method, with only a small difference between these latter probabilities and those found by the proposed method. An interesting observation was that the highest optimum found by the original method was found from only 0.02% of the starting points using that method, but was found by 11% of the starting points from the proposed method.

Table 2: Comparing optima on noisy data: Most frequently found, highest by original method and highest overall.

	Most frequent	**Highest by original method**	**Highest overall (proposed method)**
Log likelihood value	184.75	186.43	186.77
Frequency: Proposed	41.05%	11.15%	0.26%
Frequency: Original	90.11%	0.02%	n/a
Standard error (*10^{-3})	5.00	3.97	3.70
Min(Pi)	0.2859	0.1997	0.1977
Max(Pi)	0.3856	0.4312	0.4236

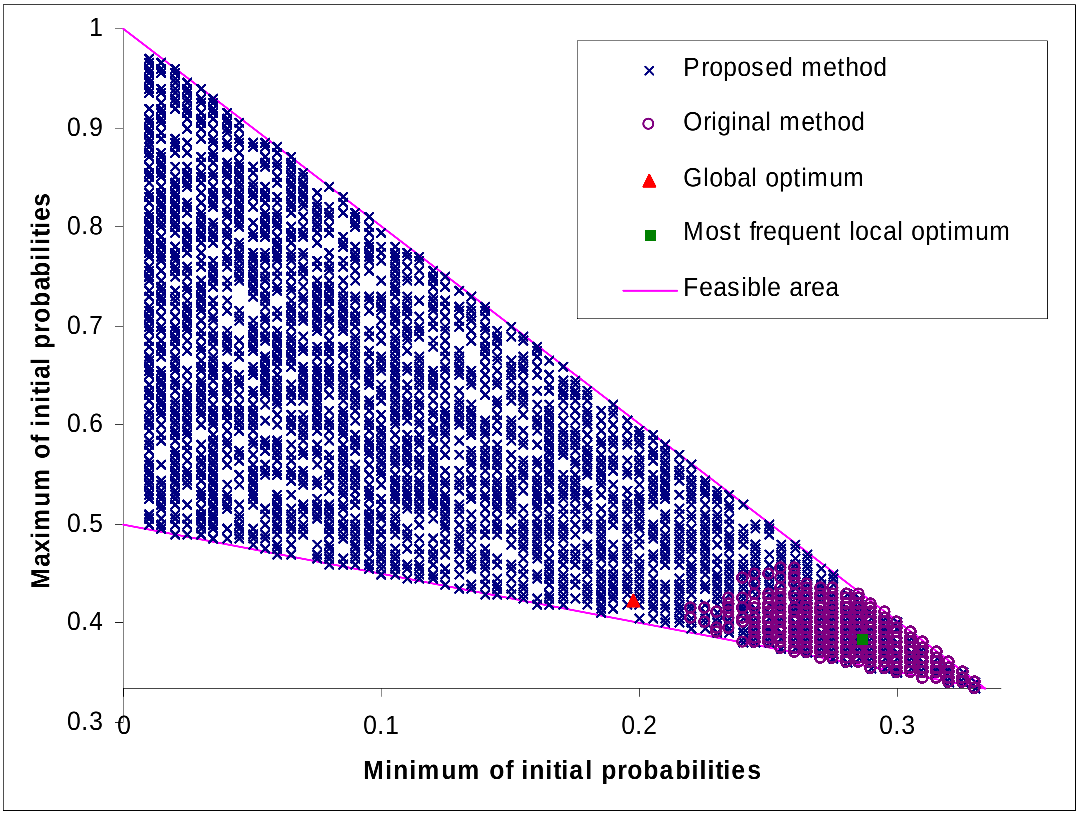
Figure 9 shows the maximum and minimum probabilities for each class at the starting point for both the original and proposed method within the feasible area of probability combinations.[1] This demonstrates that the proposed method considers a wider set of starting points. The figure also shows that the most frequently found optimum is within the range considered by the original method, but the global optimum (or at least the highest LL value found[2]) is not within the range of starting points considered

[1] The initial random probabilities are determined by a random uniform function in Shazam that operates at a discrete step of 0.005. In other words, each starting probability is some multiple of 0.005. Only those probability combinations that lead to model estimates are shown.

[2] The proposed method was run for a further 150,000 starting points and no higher LL values were found.

by the original method. This may be a potential reason why the global optimum is not found by the original method.

Figure 9: Initial maximum and minimum probabilities: Original method and proposed method.



As explained previously, the proposed method was more likely to fail to estimate a model due to encountering a singular covariance matrix, either in the first or subsequent iterations. This occurred using almost 7% of the starting points generated by the proposed method, but almost none of the starting points generated by the original method.

Table 3 presents estimates of the conditional mean *m(z)*, marginal product of input 1 $m_1(z)$ and marginal product of input 2 $m_2(z)$ for the model resulting from the most frequently found optimum, the highest LL model by the original method and the highest overall log likelihood value. Estimates are significantly different from the true values for the marginal products at data point 1, but the conditional means are not significantly different from the true value at this data point. The LL optimum that generates the lowest bias varies between each economic quantity at this data point. At data point 49, no estimates are statistically different from their true values, with the optimum under the original method having the lowest bias.

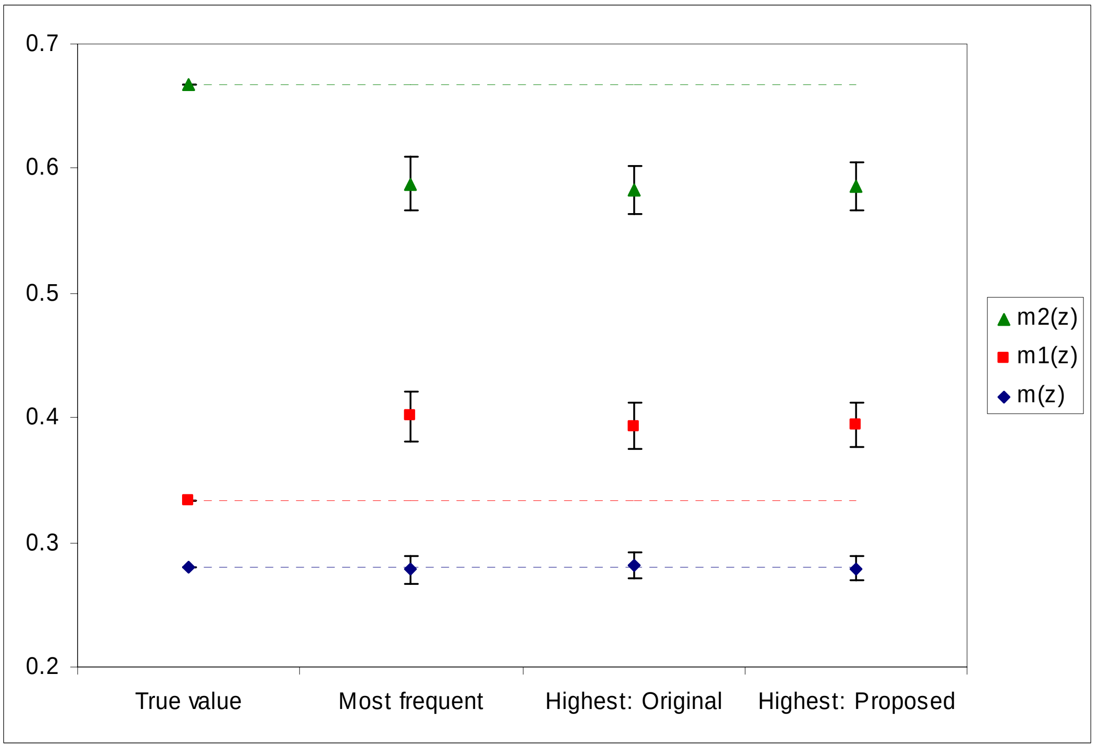
Table 3: Estimates of economic quantities at various local optima for noisy data

	Evaluated at z = (0.5,0.5)								
	$m(z_1)$			$m_1(z_1)$			$m_2(z_1)$		
	True value = 0.280			True value = 0.333			True value = 0.667		
	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall
Estimate	0.278	0.282	0.279	0.401	0.393	0.395	0.587	0.583	0.586
Standard Error	*0.006*	*0.005*	*0.005*	*0.010*	*0.009*	*0.009*	*0.011*	*0.010*	*0.010*
Bias	-0.002	0.001	-0.001	0.068	0.060	0.061	-0.079	-0.084	-0.081
RMSE	0.006	0.005	0.005	0.069	0.061	0.062	0.080	0.084	0.082

	Evaluated at z = (0.05,0.03)								
	$m(z_{49})$			$m_1(z_{49})$			$m_2(z_{49})$		
	True value = -0.184			True value = 0.308			True value = 0.692		
	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall	Most frequent	Highest-original method	Highest overall
Estimate	-0.170	-0.172	-0.168	0.325	0.307	0.289	0.623	0.675	0.659
Standard Error	*0.012*	*0.011*	*0.010*	*0.038*	*0.035*	*0.034*	*0.043*	*0.039*	*0.038*
Bias	0.014	0.012	0.016	0.017	-0.001	-0.019	-0.069	-0.018	-0.033
RMSE	0.018	0.016	0.019	0.042	0.035	0.039	0.081	0.043	0.050

Figure 10 shows the true values of the economic quantities at data point 1 together with the econometric estimates plus or minus two standard errors. While the estimates are statistically different from the true values, the estimates are not statistically different from each other, at least for this data point ($P<0.05$). At data point 49, the estimates were not statistically different from the true values, and were also not statistically different from each other ($P<0.05$).

Figure 10: True values of economic quantities compared to their estimates at z=(0.5,0.5)



In summary, the proposed method performed better at finding higher LL values with noisy data, as it did for data without noise. However, the bias was not consistently lower for the proposed method. Furthermore, there was not a significant difference between the estimates from the most frequently found optimum, or the highest LL value model by the proposed or original methods. This suggests that in some cases, such as where there is significant noise in the data, finding a local rather than global optimum may not affect the estimates of interest.

5. PARALLEL ECONOMETRIC COMPUTATION

Estimation of a latent class model is a relatively fast operation, taking only a matter of seconds with software such as Shazam. Repeating the estimation a large number of times (for example, tens of thousands) using different starting points can easily take days or weeks of computation on a single computer. Parallelisation of this process across multiple computers can reduce the delay in receiving the results almost linearly with respect to increases in computing power.

Many universities and research groups have access to a supercomputing facility for procedures that are computation intensive. These supercomputers are usually available for periods of time, but may incur a fee for use. The main features of a supercomputer include not only a large number of processors, but a fast connection between processors to allow dependent calculations to receive results quickly. These supercomputers may not be necessary or appropriate for econometricians for two main reasons. First, econometric problems such as many sampling procedures and simulations are not dependent on other calculations (referred to as “embarrassingly parallel”) and do not require the fast connection between processors. Second, supercomputing facilities frequently use a Linux or Unix operating system which may not run econometric software designed only for the Windows operating system. If these conditions hold, it may be appropriate for the econometrician to use a grid of heterogeneous computers.

The most commonly known examples of parallelising computation across a grid of heterogeneous computer resources are the SETI@Home project and the Folding@Home project.[3] Both projects involve volunteer users downloading software and data to process when the computer processor is idle and reporting results to an internet-based server.

To generate the results discussed in the previous section as quickly as possible, a small heterogeneous grid of five Microsoft Windows-based computers was employed. The grid was connected across the universities intranet and was arranged in a master-slave framework, which is where one computer (the master) controls the procedure

[3] The SETI@Home project analyses electromagnetic data from space for signs of intelligent life and the Folding@Home project analyses the way proteins fold in order to understand diseases.

and requests the other computers (slaves) to carry out calculations and return the results.[4] Each slave was requested to estimate one model (after generating a starting point) and report the results to the master together with information defining the starting point, after which it could be assigned to estimate another model with a different starting point. The computational task from the previous section would originally have taken several weeks on a single computer, but was completed in a few days using the grid framework.

The grid was managed by the enFuzion software (TurboLinux, 2000). enFuzion allows a grid procedure to be organised through a GUI using a minimum of scripting compared to more traditional grid software and works across Windows and Unix operating systems.[5] Shazam was relatively straightforward econometric software to use within this framework because it could be operated entirely with command line instructions (ie it did not require input via the GUI during the estimation procedure). Alternative technical software products such as GAMS (Brooke et al., 2005), Gauss (Aptech, 2005) and Mathematica (Wolfram, 2006) also have this capability and could be used in a similar way.[6]

An important factor to consider in a large computational task is whether the grid process is fault tolerant. Within the enFuzion software, if a computer becomes unavailable (for example due to hardware or software failure), the task will either be reported at the end of the process or run again on another machine (depending on the initial setup). Fault tolerance is enhanced enFuzion’s ability to record which jobs have been done by the master, allowing the procedure to recommence from a checkpoint should there be a temporary failure in the master computer (eg due to a power outage). It can also be helpful to use a laptop as the master as it has its own temporary battery backup.

An important factor that makes the grid very practical is the flexibility with which different computers in the grid can be used. For example, any computers that are on

[4] In this example, the master also operated as a slave.

[5] Alternative grid software for the Microsoft Windows operating system is available (eg Digipede, www.digipede.net).

[6] GAMS, Gauss and Mathematica may have benefits in terms of precision (relative to Shazam) but their multiple computer or site licenses are far more expensive.

the same network or connected to the internet can feasibly be used. This allows the grid to take advantage of computers within the same room, within the same research group or even computer labs. These resources can be set up so that being part of the grid does not affect other users who may use the computers. For example, a network may be set up to utilise only idle computers. Idle computers may be defined as computers that are on, but no user is logged, or extended to include computers where the user is logged in, but the processor is not heavily used. In the former case, if a user logs into one of the slave computers, it automatically halts any calculations. In the latter case, the priority of the grid process can be set low so that the user would not normally notice any degradation in performance.

The results from the estimation of each model were saved as separate text files. These were then stored on the master computer and backed up at regular intervals to an external hard drive, improving the fault tolerance of the procedure[7]. The output in each separate text file could have been added to a file as part of the grid procedure, but was instead merged into a summary files by use of Perl scripting, made available through ACCS. Similar procedures could have been done without scripting using GUI-based software such as Textpipe (DataMystic, 2006).

To ensure that the available processing power is used effectively, it is helpful to ensure that the communication network can handle the expected communication traffic. In the example, the communication between the master and slave consisted of relatively small text files every few seconds, well within the network's capability.

In summary, grid procedures can provide a fault-tolerant approach to quickly obtaining results from processes involving heavy computation loads. Grid techniques can make it more feasible to do complex computations and reduce the time required to check a multitude of starting points. Critical success factors for the use of a grid include access to computers, grid software, and econometric software licensed for the grid that can be run from a command line. It is helpful to have scripting experience.

[7] Carrying out periodic backup of files in a particular directory is easily managed through the "scheduled tasks" option in Windows XP. Windows XP can manage tens of thousands of files in a single directory, although depending on the storage device, it may be slow to read and take up significantly larger space on the device than indicated by the file size. This problem can be reduced by compressing files either as part of the grid procedure or in the backup process.

Most econometricians already have this experience from writing their own econometric procedures within software such as Shazam or Gauss. It is also helpful to have a computer technician available when setting up a grid. This is particularly the case if there is not easy access to the computers and communication is to be made through very secure firewalls.

6. APPLICATIONS

Estimating state contingent production functions through the application of econometric techniques (such as latent class models) is problematic for several reasons. These include the data requirements, appropriately defining states, and ensuring production functions consistent with theory.

Data requirements are an issue as a large amount may be required to establish a deterministic production function for every state of nature. This becomes more problematic as higher numbers of states and more factors of production are considered. Theoretically, one needs to know not only how much of the inputs are selected, but to which states they will affect. Latent class models can ease the data burden somewhat by using the data to determine some of these effects.

Identifying states can be problematic. For example, data may be sampled across firms at the same period in time, but within each period all firms are not facing the same state. For example, rainfall could differ between farms, even within the same area. Latent class models such as those presented by O’Donnell (2006) can be appropriate when there is no information about which data observations relate to a particular class. However, when there is information that could be used to partially determine the class, a Bayesian framework (ie with priors) may be preferred to a frequentist framework.

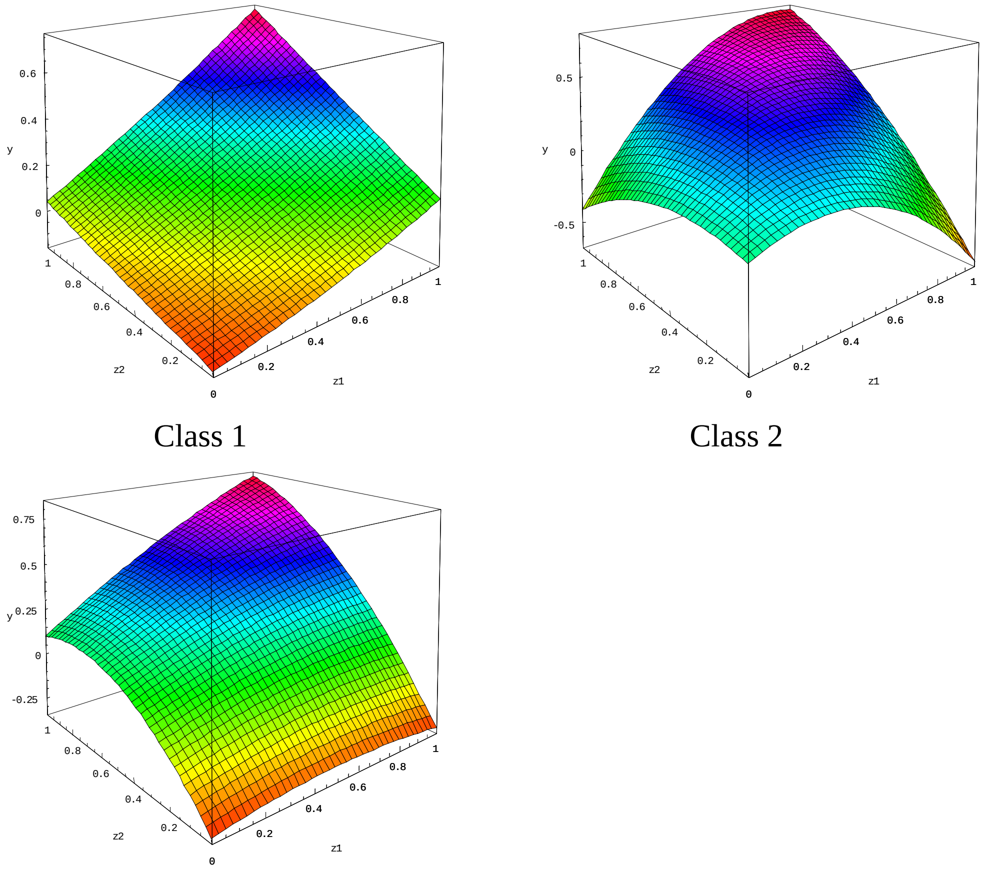
Identifying the number of states is a controversial issue, given that there could be an infinite number of possible states. Rather than identifying a large number of states to represent all possible states, Rasmussen (2004) suggested that the problem can be tackled by considering states that represent variability in some (relatively important) factors, but with each state having a stochastic production function to allow for variability in other factors. As an example, O’Donnell et al. (2005) presented an application of the state contingent approach using stochastic production frontiers with finite mixture estimation of the states for rice farming in the Philippines. It was assumed that there were three possible states that could occur, but more generally statistical tests such as the Akaike Information Criterion or the Bayesian Information Criterion may be helpful in determining whether the number of states is appropriate for the data.

Testing the appropriateness of certain specifications for production technology can be problematic. For example, the method of O'Donnell et al. (2005) is not flexible enough to statistically test the nature of the state substitution – ie is the technology output cubical or not? Chavas (2006) used a different approach with US agricultural data to test whether the level of state substitution was significant and concluded it was not. This result suggests that in some contexts, modelling the state contingent production function to allow for non-output cubical technologies may not be important.

Determining an appropriate model of state contingent production may not be solely a question of whether the model was found to have the highest (non-spurious) LL maximum. This is because the production function for each state still should maintain reasonable economic properties. For example, returns to scale and marginal products should be lie within reasonable bounds for each state. This may be approached through a Bayesian framework (incorporating priors) or restrictions (O'Donnell, 2006). For the latent class model estimated in an earlier section using noisy data, the estimated models for the three classes are shown in figure 11. If it were a production function that was estimated for three states, the model implies that the marginal product of one of the inputs is negative for a significant region of feasible inputs in two states[8]. If this was not considered to be realistic, the incorporation of Bayesian priors or other restrictions may be required.

[8] It can be seen in figure 3 that the original production function used to generate the data observations did not have negative marginal products at any part of the production surface.

Figure 11: Estimated production function for three classes



Class 1

Class 2

Class 3

7. CONCLUSIONS

It was demonstrated that different methods for choosing starting points in the estimation of a latent class model would result in different optima being found. It was shown that the single most common optimum found in a large proportion of starting points (up to 90%) was not the highest possible LL value.

A method was proposed that determined the probability of each observation belonging to a class before assigning each data point to a class, which differed from the accepted (original) method that implicitly assumed a probability of 1/*j* for a data point belonging to each of *j* classes. In an example, whether using data with or without noise, the proposed method found higher LL values than the original method. These maximum LL values were not likely to be spurious maximisers, as they did not match the criterion of a having a class with a very low probability and having a low standard error associated with that class. For the model estimated from data without a noise component, finding a higher LL maximum was particularly important because the estimated economic quantities were statistically different (P<0.05). For the model where noise was included, the estimated economic quantities of the different optima were not statistically different between LL maxima found by the two different methods. In estimating latent class models, this author recommends that 1000 starting points be used with the proposed method to check whether the quantities of interest vary between different optima such as the highest LL value and the most frequently found optimum.

It was also demonstrated how a grid of heterogeneous Microsoft Windows computers could be used to check a large number of starting points. The advantages of using a grid of computers are that results were found far more quickly than would be possible if a single computer was used. Advantages of the system such as faster results, the ability to do more complex estimations and fault tolerance may outweigh the cost of learning how to use the system, particularly if the user is familiar with scripting and uses software that can be run from the command prompt. Difficulties in setting up a grid include access to computing power, technical assistance, dealing with firewalls, and licensing costs for the software.

Difficulties in the estimation of state contingent production frontiers by methods such as latent class models is problematic due to data requirements, appropriately defining states, and ensuring production functions consistent with theory. The latter problems can be reduced by using a Bayesian framework to incorporate priors relating to state identification and the incorporation of restrictions to ensure estimated production functions consistent with expectations and theory.

8. REFERENCES

Aigner, D. J., Lovell, C. A. K. and Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics,* 6(1), pp. 21-37.

Aptech. 2005. *Gauss: A users guide*, Aptech, Black Diamond (WA).

Bond, D., Harrison, M. J., and O'Brien, E. J. 2005. Investigating Nonlinearity: A Note on the Estimation of Hamilton's Random Field Regression Model, *Studies in Nonlinear Dynamics and Econometrics,* 9(3), pp. 1-41.

Brooke, A., Kendrick., D., Meeraus, A., and Ramesh, R. 2005. *GAMS: A users guide*, GAMS Development Corporation, Washington.

Chambers, R. G. and Quiggin, J. 2001. *Uncertainty, production, choice and agency: The state-contingent approach*, Cambridge University Press, Cambridge.

Chavas, J. 2006. A Cost Approach to Economic Analysis under Production Uncertainty, Presented at the American Agricultural Economics Association meeting, July, Long Beach (CA).

DataMystic. 2006. TextPipe Manual, Available at www.datamystic.com.

Finch, S., Mendell, N., and Thode, H. 1989. Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of American Statistical Association* 84, pp. 1020-1023.

Karlis, D., and Xekalaki, E. 2003. Choosing initial values for the EM algorithm for finite mixtures, *Computational Statistics & Data Analysis*, 41, pp.577-590.

McLachlan, G. J. and Krishnan, T. 1997. *The EM algorithm and extensions*, Wiley, New York.

McLachlan, G. J. and Peel, D. 2000. *Finite mixture models*, Wiley, New York.

Meeusen, W. and van den Broeck, J. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review*, 18, pp. 435-444.

O’Donnell, C. J., Chambers, R. G., and Quiggin, J. 2006. Efficiency analysis in the presence of uncertainty, Risk and Sustainable Management Group Working Paper R06_2.

O'Donnell, C. J., and Griffiths, W. 2006. Estimating State-Contingent Production Frontiers, *American Journal of Agricultural Economics*, Vol. 88 Issue 1, pp. 249-266.

O'Donnell, C. J. 2006. Some econometric options for dealing with unknown functional form, Paper presented at the Annual Conference of the Australian Agricultural and Resource Economics Society, February 8-10, Manly.

St. Pierre, E. F. 1998. Estimating EGARCH-M Models: Science or Art?, *Quarterly Review of Economics and Finance*, Summer 1998, 38(2), pp. 167-180.

Tonsor, G., and Kastens, T. 2006. How much do starting values really matter? An empirical comparison of genetic algorithm and traditional approaches, Paper presented at the Annual Conference of the American Agricultural Economics Society, July 23-26, Long Beach.

TurboLinux. 2000. *enFuzion 6.0 User Guide*, TurboLinux, Brisbane (CA).

White, H. 1980. Using least squares to approximate unknown regression functions, *International Economic Review*, 21, pp. 149-170.

White, K. J. 1993. *Shazam User's Reference Manual, Version 7.0*, McGraw-Hill, Vancouver.

Wolfram, S. 2006. *The Mathematica book*, Wolfram Media, Champaign.

APPENDIX A

Figure A1: Proposed method on noisy data: Log likelihood frequency, standard error, maximum and minimum probability of local optima.

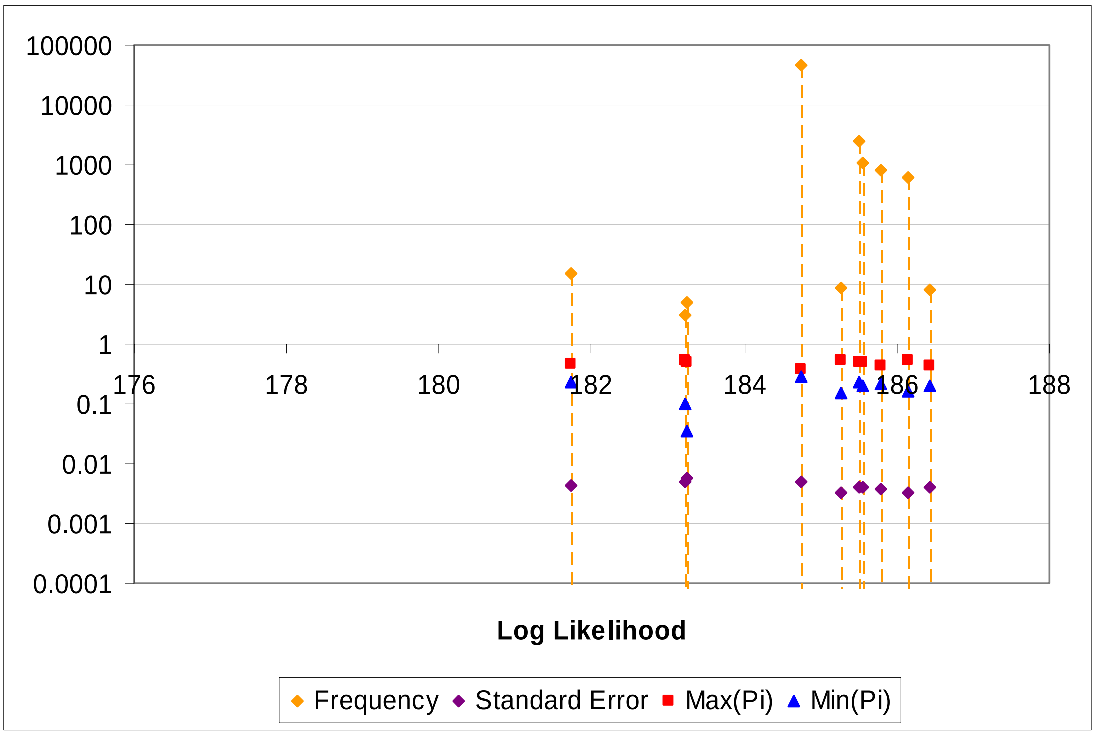
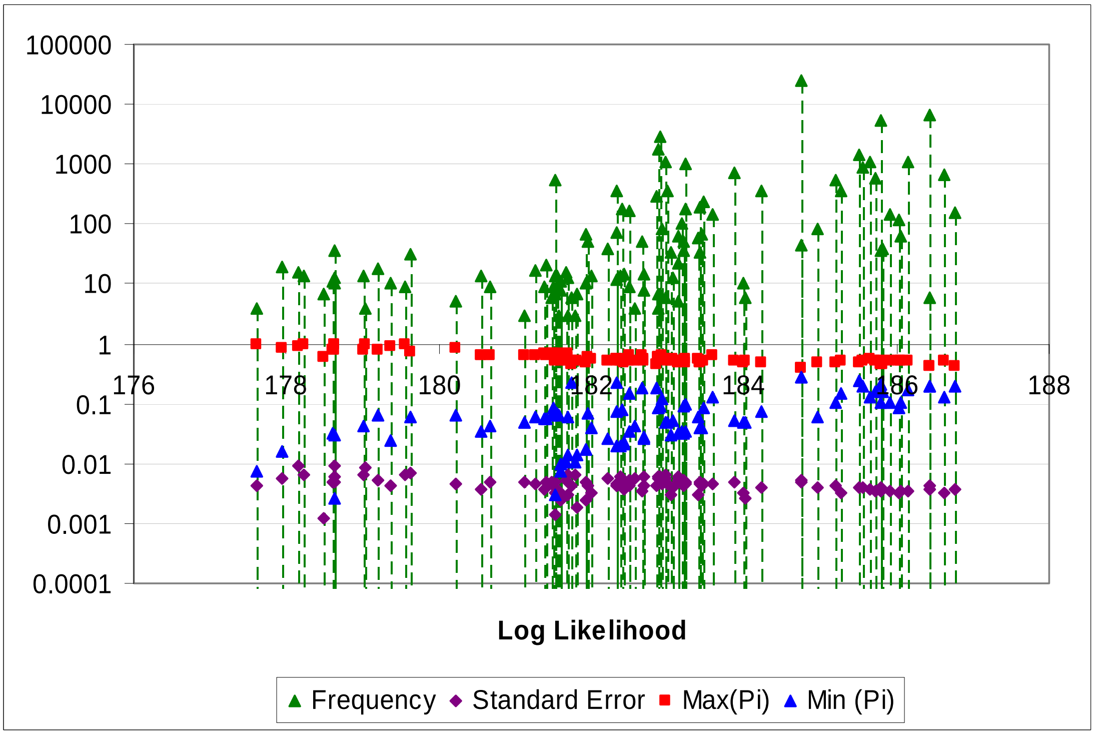


Figure A2: Proposed method on noisy data: Log likelihood frequency, standard error, maximum and minimum probability of local optima.


Log Likelihood
▲ Frequency ◆ Standard Error ■ Max(Pi) ▲ Min (Pi)