Estimating non-parametric and parametric transformation functions from survey data: An application of minimum cross-entropy

Øyvind Hoveid *

American Agricultural Economics Association Annual Meeting
August 2000, Tampa, Florida

*Senior Researcher, Norwegian Agricultural Economics Research Institute, P.O. Box 8024, N-0030 Oslo dep. E-mail: oyvind.hoveid@nilf.no

Abstract

A model of a (convex) technology of representative and non-representative firms in a heterogeneous sector is presented in non-parametric and parametric versions. The heterogeneity is specified with error terms. The models including a non-parametric distribution of the errors can be estimated with entropy econometrics from firm survey data. This requires two important modifications in the standard approach to entropy estimation of Golan, Judge and Miller: The compact support of the probability distribution should be designed to capture eventual non-zero covariance. And cross-entropy need to be redefined for cases of multiple observations.

1 Introduction

The standard way to proceed from survey data to technology is to estimate with OLS a convex constant returns to scale (CRTS) technology function, $F(x_t;\alpha) = \epsilon_t$, where x_t is the netputs (variable or fixed) of firm t, α is a vector of parameters to be found and ϵ_t is an error term. The set of points with $F(x;\alpha) = 0$ is then the technology of "the representative firm" on which further analysis is based. There are two draw-backs in this approach. First, it does not lead to a CRTS model for the non-representative firms ($\epsilon_t \neq 0$). If ϵ_t should be kept constant, that would imply globally increasing (decreasing) returns to scale for firms with $\epsilon_t < 0 (> 0)$. Second, the distribution of ϵ mirrors the heterogeneity of the firms as well as measurement errors and specification errors due to functional form. With the empirical nature of ϵ, it is not expected to be normally distributed, and the use of OLS cannot be easily rationalized as maximum likelihood. In addition, it is not at all obvious that OLS will leave behind a reasonable picture of the heterogeneity in the distribution of ϵ.

These problems have economical relevance. As is well known, a sector of heterogeneous firms need not behave as if it consisted of a single representative firm. With CRTS models for all firms and a suitable representation of their heterogeneity, one can explore to which extent sector behavior according to the representative firm model deviates from that of heterogeneous firms.

Modeling of non-representative firms has been undertaken by Howitt (1995) with the label of Positive Mathematical Programming. Recently Paris & Howitt (1998) have developed the ideas in estimating cost functions with heterogeneous parameters. Lansink (1999) have also applied the same technique. This methodology might be applied in primal mode with a transformation function with heterogeneous parameters, $F(x_t; \alpha_t) = 0$. That might solve both problems with the standard procedure, if it did not create a new one. This technology model does not reflect the situation that all firms have access to the same technological possibilities, and consequently should have some common structure across firms.

The problem of finding more than one parameter from one observation and one equation is *ill-posed*. Parameters cannot be identified without some additional information. Paris & Howitt (1998) supply this information in terms of a *prior distribution* of the parameters. They can then be identified using an estimation method based on the principle of maximum entropy developed by Golan, Judge & Miller (1996) (GJM). One might say that the technological relationship between observations is expressed in the prior distribution. However, as the prior is not contained in the economical model, only in the statistical one, this is not entirely satisfactory.

Another drawback with the Howitt-Paris approach is that it utilizes merely price information on some commodities and merely quantity information on others. Additional information on quantities and prices are in general available — if not in the firm survey itself, from other data collecting insti-

tutions. It may be objected that with CRTS, the perfect duality between prices and quantities makes information of both superfluous. However, the correspondence between prices and quantities hold for data recorded without errors. If data of the one type have errors, erroneous data of the other type will presumably have some information to be exploited.

The technology models to be developed here reflect both these points. The common possibilities are specified in the conventional manner with some convex transformation function F and a relation $F(x;\alpha) \leq 0$. where α is a vector of parameters to be determined. The error structure is unconventional though, with vectors of multiplicative commodity-wise error terms, ξ_t, π_t, and an operation $\cdot$ defined as element-wise multiplication. For estimation we have the feasibility equation:

$$F(x_t \cdot \xi_t; \alpha) = 0$$

and the first order conditions:

$$\partial F(x_t \cdot \xi_t; \alpha) = p_t \cdot \pi_t$$

These equations are main constituents of the behavioral model for non-representative firms, that is firms with $(\xi_t, \pi_t) \neq 0$. They do not determine the scale of operation though. The scale can only be determined with reference to some bundle of fixed netputs. For sake of estimation the distinction between variable and fixed netputs need not be maintained in the analysis.

All commodities have their quantities and their shadow prices on which we have more or less reliable information.

The model with errors both on quantities and prices might be estimated with an errors-in-variables approach, if there is a sufficient number of instruments available. Otherwise, the model will be ill-posed and standard econometrics is not applicable. In this paper it will be shown how entropy based econometrics (EE) can be applied in estimating the ill-posed model, that is the error terms and their non-parametric distribution and eventual parameters. GJL have provided the introduction to entropy methods and some practical hints. Otherwise, the research based on Kullback (1959) has led to several points of constructive criticism of the GJL approach.

The layout of the paper is as follows. The estimation of a non-parametric distribution by means of entropy based econometrics is considered in section 2. In section 3 it is shown how the model above can be estimated with EE when the technology is non-parametric following Varian (1984). With respect to EE estimation this is the simplest model. It turns out that there are some lessons to be learnt which are relevant for estimation of semi-parametric and parametric models considered in section 4. The paper ends with some concluding comments.

2 Estimating a non-parametric distribution with entropy

The concept of entropy was established by Boltzman in the 1870-s as a measure of disorder. The idea has proved to be extremely applicable in processing of noisy and indeterminate information. It is now heavily connected with information theory and statistics, where the basic idea is that minimum information consistent with observations should be utilized. Here, a version named the Kullback-Leibler measure of information — or in GJM's terminology cross-entropy — will be utilized.

The Kullback-Leibler measure involves two hypotheses, H_1, H_0, and two probability densities for data, $f_1(x), f_0(x)$, corresponding to the two hypotheses. The measure of information, $I(1,0)$

$$I(1,0) = \int f_1(x) \ln \left(\frac{f_1(x)}{f_0(x)} \right) dx$$

is said to measure the amount of information that supports H_1 in favor of H_0 (Kullback 1959). It can also be considered as the directed squared distance from the density f_0 to the density f_1 (Csiszar 1975). A certain test statistic based on the minimum value of $I(1,0)$ subject to certain constraints given by data, decides whether the support is sufficiently strong to reject H_0.

As a special case, minimization of the measure serves as a method of Bayesian estimation. The prior probability distribution (of errors and pa-

rameters) is then $f_0(x)$. Some additional observations are added and the posterior distribution, $f_1(x)$, is found as the one that *minimizes* the information in favor of H_1, or minimizes the squared distance from $f_0(x)$ to $f_1(x)$. If this information (distance) is sufficiently large, the posterior may serve as a prior for the next round. If not, one sticks to the old prior and wait for more information.

The current setting of technology estimation is *not* purely Bayesian, however. We have data and a model that have not previously been combined. No strong prior distribution exists. This has consequences for the way the entropy based method should be conducted. If we have data that are rich in the sense that multiple observations span the domain of the distributions, f_1, f_0, the priors should be made relatively *uninformative*. If there are few observations — at the extreme merely a single one as in Howitt and Paris' case — there is no way to an estimated distribution without an informative prior. This raises the question how EE estimation of distributions can be adapted to multiple observations and more or less informative priors.

Consider first the "natural" way of dealing with multiple independent observations. The information x has then the form of a set of independent but identically distributed, $\{x_1, \ldots, x_T\}$. The statistical task is to infer the distribution of x from observations y, in terms of a posterior density $g_1(x_t)$ based on a prior density $g_0(x_t)$. If there was a functional relationship $x_t = F(y_t; \alpha)$ with some parameters α, this task might be done with standard econometrics. The models that will be considered in this paper have

not such relationships however, but rather many-to-many correspondences, $(x_1, \ldots, x_T) \in F(y_1, \ldots, y_T)$. This creates an ill-posed model and need for entropy methods.

The information measure anyway takes the form:

$$\begin{aligned}
I(1,0) &= \int \prod_t g_1(x_t) \ln \left(\frac{\prod_t g_1(x_t)}{\prod_t g_0(x_t)} \right) dx_1 \cdots dx_T \\
&= \int \prod_t g_1(x_t) \sum_t \left(\ln g_1(x_t) - \ln g_0(x_t) \right) dx_1 \cdots dx_T \\
&= \sum_t \int \prod_s g_1(x_s) \left(\ln g_1(x_t) - \ln g_0(x_t) \right) dx_1 \cdots dx_T \\
&= \sum_t \int g_1(x_t) \left(\ln g_1(x_t) - \ln g_0(x_t) \right) dx_t \prod_{s \neq t} \int g_1(x_s) dx_s \\
&= \sum_t \int g_1(x_t) \left(\ln g_1(x_t) - \ln g_0(x_t) \right) dx_t
\end{aligned}$$

The prior and posterior densities can be parametric functions of a certain class. But the distribution representing technological heterogeneity has an empirical content that make non-parametric representation preferable. I therefore switch to the approach of GJM which considers precisely non-parametric distributions.

The densities g_1 and g_0 are now approximated with discrete distributions over a grid of support points X. To be more precise, $X \in \mathcal{R}^{J \times N}$ so that every x_t with $g_1(x_t) > 0$ or $g_0(x_t) > 0$ is contained in the convex closure of the columns of X, $X_1, \ldots, X_N$. This means that $g_1(x_t) > 0$ or $g_0(x_t) > 0$ implies the existence of $P_t \in [0,1]^N$ with $x_t = XP_t$. There are some issues to be discussed with respect to the design of this grid, but the theoretical

requirements are clear. The prior distribution is a fixed vector $Q \in [0,1]^N$. The posterior $\bar{P}$ is the expectation of $P_1, \ldots, P_T$, $\sum_{t=1}^{T} P_t/T$. In terms of these discrete distributions, the information measure $I(\bar{P}, Q)$ is called *cross-entropy* and takes the form:

$$I(\bar{P}, Q) = \sum_n \bar{P}_n \left(\ln \bar{P}_n - \ln Q_n \right)$$

The estimated $\bar{P}$ that results from minimization of cross-entropy subject to contraints can be spelled out as:

$$\bar{P} = \text{argmin}_{\bar{P}_n \geq 0} \left\{ \sum_n \bar{P}_n \left(\ln \bar{P}_n - \ln Q_n \right) \;\middle|\; \begin{array}{c} (XP_1, \ldots, XP_T) \in F(y_1, \ldots, y_T) \\ 1'P_t = 1,\ P_t \geq 0,\ \bar{P} = \sum_{t=1}^{T} P_t \end{array} \right\} \tag{1}$$

Look now at the first order condition with respect to $\bar{P}_n$. The multiplier of the last constraint is ϕ:

$$\ln(\bar{P}_n - Q_n) + 1 \leq \phi$$

The somewhat surprising situation occurs that the estimated posterior distribution $\bar{P}$ is proportional to the prior Q in the points where $\bar{P}$ is non-zero. In turn $\bar{P}_n$ is non-zero only when some P_{nt} is non-zero. This means that the more observations, the closer comes $\bar{P}$ to Q. This is a highly counterintuitive effect which suggests that the minimum cross-entropy estimator $\bar{P}$ of $g_1(x)$

is inconsistent.

At this stage it should be observed that when entropy methods is used for interpretation of noisy and indeterminate signals or images, this situation is quite acceptable — and probably the key to its success. A pattern of information in terms of a non-parametric distribution should be revealed, involving a lot of data but essentially only one observation. When a several images are stacked before processing, one should not expect to find more information than in the single image. The task of estimating a common distribution from several independent observation is therefore rather different from the tasks that entropy methods usually handle. Some revision of the standard EE seems required to make it work properly for this problem.

An alternative procedure will now be suggested: Assume that the prior distribution Q is established at time 0. Later observations have not deviated sufficiently from Q to reject it. At time T, though, it is asked if the accumulated observations, $y_1, \ldots, y_T$, can be sufficient to make the prior rejected. Obviously, one has to make use of some average of the observations in this test, but it need to be different from the one in (1). The average $\bar{P}$ minimizing the following measure of accumulated information, $\mathrm{AI}(P_T, \ldots, P_1, \bar{P}, Q)$, is proposed:

$$\mathrm{AI}(P_T, \ldots, P_1, \bar{P}, Q) = \sum_{t=1}^{T} I(P_t, \bar{P}) + I(\bar{P}, Q)$$

Thus, $\bar{P}$ is the distribution which minimizes the sum of squared distances

from the prior Q to $\bar{P}$ and from $\bar{P}$ to all P_t. In terms of the cross-entropy program, the estimator is:

$$\bar{P} = \text{argmin}_{\bar{P}_n \geq 0} \left\{ \begin{array}{c|c} \sum_{t=1}^{T} \sum_n P_{nt} \left(\ln P_{nt} - \ln \bar{P}_n \right) & (XP_1, \ldots, XP_T) \in F(y_1, \ldots, y_T) \\ + \sum_n \bar{P}_n \left(\ln \bar{P}_n - \ln Q_n \right) & 1'P_t = 1,\ P_t \geq 0,\ 1'\bar{P} = 1 \end{array} \right\} \tag{2}$$

As a first piece of justification, it should be observed that with merely one observation, the minimum of $\text{AI}(P_1, \bar{P}, Q)$ is identical to the minimum of $I(P_1, Q)$. Thus, nothing new is invented in the one observation case. Secondly, we have here the desired effect that the influence of the prior on $\bar{P}$ is lessened as more observations are added. Actually can the prior be made less informative by assigning a weight smaller than 1 to the last squared distance. In the limit the effect of the prior vanishes totally. The posterior is then according to the first order conditions of the program, simply, $\bar{P} = \sum_t P_t / T$. This is formally identical to the estimator (1), but now the prior is totally uninformative.

The estimator $\bar{P}$ defined by (2), seems to be what is needed for estimation of a non-parametric distribution from a set of independent noisy observations. There is of course research to be done on the statistical properties of this measure. At this stage it should be considered a promising application of EE to estimation of a distribution. Standard EE estimation following GJM will simply not do the job.

3 Estimating the non-parametric producer model

If netputs of production, x_t, and corresponding shadow prices, p_t, were observed without error, technology representation would be no problem. Because of CRTS would the zero profit condition, $p_t' x_t = 0$, hold for all t. Convexity of the technology function would imply the WAPM inequalities, $p_s' x_t \leq 0$ for all s and t, (Varian 1984). This case reveals no heterogeneity as any discrepancy across firms is explained with different prices.

Errors are likely, however and the WAPM relations are expected to hold only for the correct prices and quantities, $\hat{p}_t, \hat{x}_t$.

$$\hat{p}_t' \hat{x}_t = 0$$

$$\hat{p}_s' \hat{x}_t \leq 0$$

Multiplicative errors are convenient when dealing with CRTS technologies because they are independent of scale. Thus,

$$\hat{p}_t = p_t \cdot \pi_t$$

$$\hat{x}_t = x_t \cdot \xi_t$$

where $\cdot$ means element-wise multiplication. These errors need the assumption though that the sign of each element of x_t and p_t is correctly observed.

The WAPM-relations constitute the technological model. The model itself challenges the common conception that Varian's non-parametric models are inherently deterministic. It should be mentioned though, that Varian (1985) proposes a related model, and so does Chavas & Cox (1995) with a little more structure motivated by technical progress. Both contributions have errors merely on quantities and apply OLS to find the minimum errors that satisfy the WAPM-relations.

Here the errors and their probability distribution will be found by minimization of a Kullback-Leibler measure of information — the measure of accumulated information, $\mathrm{AI}(P_T, \ldots, P_1, \bar{P}, Q)$, of the previous section — or in the terminology of GJM, a certain cross-entropy. The use of EE estimation is motivated by the fact that this model is ill-posed. The error terms $(\pi_t', \xi_t')'$ are not exactly identified by the model. Compared to the model of the previous section, the error vector, (π_t, ξ_t), is the unobservable for which a non-parametric probability distribution will be estimated. (p_t, x_t) is the observable variable vector. And the combination of WAPM relations and the stochastic model:

$$(p_t \cdot \pi_t)'(x_t \cdot \xi_t) = 0, \quad (p_s \cdot \pi_s)'(x_t \cdot \xi_t) \leq 0 \tag{3}$$

defines the correspondence between observable and unobservable variables.

Following GJM a grid of support points $X \in \mathcal{R}^{2J \times N}$ is chosen for the error vector. As the elements are positive with 1 representing no error, it is convenient to convert them to logarithms. The grid should then be centered in origo, and this is one point of the grid labeled 0. The remaining points should lie sufficiently far away from origo to have all error terms identified within the interior of their convex closure. This can be accomplished with a common positive constant D which can be enlarged if required. It is presumably too much to hope for independent elements. In order to capture all dependencies within pairs of elements, all vectors of the form $\delta_{ij} \in \mathcal{R}^{2J}$ where $i \neq j$, $\delta_{iji}, \delta_{ijj} \in \{-D, D\}$ and $\delta_{ijk} = 0$ for $k \neq i, j$, should be contained in the grid. It thus consists of $(2J)^2(2J-1)^2 + 1$ points. With $J = 2$ the grid then has $3^2 4^2 + 1 = 145$ points, with $J = 10$, it amounts to $20^2 19^2 + 1 = 144401$. This structure of the grid is considerably more complex than what GJM suggest. Their grid vectors have only one element different from zero. This implies that smaller grids can be applied. With 3 points for each variable, it suffices with merely $6J$ points. This simplicity comes at a large cost, though. Dependencies between variables are out-ruled by assumption.

And then the prior distribution. First, a warning should be issued: A uniform prior with identical probability mass on each point of the grid, is not a good choice when it comes to variables which are not naturally bounded. For example should an approximation to a normal distribution have considerably more probability mass placed in the centre point than on the tails. The distribution of probability mass on centre and tails is not immediately

clear — unless one can look into the data first. One should expect that they will speak strongly on this issue, so whatever is chosen as a prior-prior in the first round, can be corrected in a second round. The important role of the prior is the distribution of mass among the tails. The natural prior is of course uniform mass on all tails. One cannot expect that data span all dimensions of the domain of the probability distribution. In dimensions where observations are scant, the prior may play a decisive role.

All prerequisites for EE-estimation of the non-parametric producer model are now in place. After an initial solve it should be checked whether all error terms $(\pi_t', \xi_t')'$ are in the interior of the convex closure of the grid. This is ensured when $P_{0t} > 0$ for all t. Eventually should the grid be enlarged. The distribution of mass between centre and tails, $\sum_t P_{0t} / \sum_t \sum_{n \neq 0} P_{nt}$ should also be computed and the prior should be brought in reasonable accordance.

After a few tentative solves, estimates of π_t, ξ_t for all t are found as XP_t. The heterogeneity of the sector is represented with the estimated distribution $\bar{P}$. This is more or less identical to $\sum_t XP_t$ with some flavor of the prior added. With these results can exercises on the behaviour of the heterogeneous sector be conducted.

Without any empirical test of the estimation method, it is nevertheless clear that errors are larger than those computed with OLS, because errors have not been minimized. What has been minimized is instead a certain cross-entropy, or a certain sum of squared distances between probability distributions, or a certain amount of information extracted from the observa-

tions.

It is also clear that this version of EE estimation can form a base for testing of hypotheses — in particular because this was the starting point of the method. During testing, however, one needs informative priors to express H_0, and the testing model should be based on minimization of the original Kullback-Leibler information measure within the constraints imposed by H_1. The informative prior can be found first by estimating the model as explained above with an uninformed prior and the constraints of H_0. It is less clear what the distribution of $I(1,0)$ actually is. The grid based method that have been followed here, is essentially non-parametric. Possibly can some asymptotic distribution be assigned, but in any case is boot-strapping possible.

4 Semi-parametric and parametric producer models

With the non-parametric model of the previous section, it may happen that the price and/or quantity information has obvious weaknesses that can be mended with additional information. This is particular the case with respect to price information, where the available information is found at market, not at the firm. For example, information on the education of the farmer may introduce price variations in the family labor applied on the farm and explain the variance of the respective error term. Information on farmer's age may likewise explain the variance of quantity errors. Agricultural land

is another factor for which additional information may be appropriate. Data on precipitation, length of growing season and average temperature indicates the quality of the land and may explain the variance of quantity errors on land. Data on acreage payments (not set-aside) may likewise come in on the price side. Of course are acreage payments expected to carry over to land rents, but this absorption need not be complete.

Farm surveys have most often rather detailed information on region. Regional dummies can then be used for prices and/or quantities to explain variation. With panel data one may even introduce farm dummies. Any such additional explanatory variable will be comprised in a vector, z_t. Their eventual explanatory power can be tested with equations

$$\hat{p}_t = p_t \cdot \exp(\ln \pi_t + \gamma z_t), \quad \hat{x}_t = x_t \cdot \exp(\ln \xi_t + \beta z_t)$$

to be incorporated in cross-entropy program. γ and β are here parameter matrices to be determined. Such modeling is parallel to the use of instrumental variables in standard econometrics.

In the GJM-approach are support points assigned also for such parameters. This seems to be waste of computational resources. When cross-entropy is minimized, the best values come up. There is no essential difference between these parameters and the parameters coming from minimization of a log-likelihood function. Only when the probability distribution of the parameters are needed, a support is required. As these parameters carry no

specific economic meaning, this part can in general be skipped.

A more relevant issue is whether these variables (or subsets thereof) actually improves the model. This is again an issue for testing as explained above. Zero parameter values against optimized values. However, with the current methodology it does not make sense to take zero values as H_0. Optimized parameters will then be brought to zero under cross-entropy-minimization, and H_0 can never be rejected. By working the other way with optimized parameters as H_0, the minimum cross-entropy will measure the minimum loss of information due to parameter restrictions. If this loss is small, can H_0 be rejected and parameters be kept zero for sake of simplicity.

Both non- and semi-parametric models can be converted to fully parametric ones by means of a parametric convex and linearly homogeneous transformation function, $F(x;\alpha)$ where α is a vector of parameters. (Unless strictly required, the reference to the parameter vector will be skipped.) Such a function satisfies the subgradient inequality, $\partial F(\hat{x}_s)\hat{x}_t \leq F(\hat{x}_t)$, and the WAPM relations can consequently be expressed in terms of the equations:

$$F(\hat{x}_t;\alpha) = 0$$

$$\partial F(\hat{x}_t;\alpha) = \hat{p}_t$$

For estimation of α in the parametric model, much the same comments applies as with the parameters γ and β of the semi-parametric model. In gen-

eral, the parameters themselves carry no specific economic meaning. Their values and their probability distributions are irrelevant. What matters are their contribution to the model when optimized. Is the model with a subset of parameters fixed to 0 (or some other default value) a significantly poorer model than the one with optimized values? Testing will show. This testing is slightly different from testing γ and β because non-zero-parameters are not associated with additional explanatory variables. Nevertheless, a non-zero parameter may extract particular information from a variable, so testing should proceed as above. Estimate first the model with parameters optimized and a relatively uninformative prior. Use the resulting estimate of the errors as an informative prior in the test-solve with zero parameters. If the loss of information is small, parameters can be kept zero.

5 Concluding comments

The problem of estimating heterogeneous CRTS technologies with a common structure has been posed and is solved with entropy econometrics. The solution is most likely not the only one, and possibly not the best one, but it seems at least consistent. A prior distribution of the heterogeneity can be specified and will be decisive when observations are few, but will be downplayed when observations are numerous. In contrast, entropy based estimation following Golan et al. (1996) is not consistent for this task. This may be part of the explanation why successful application of their methods is reported mainly

for single observation situations.

The suggested model will, if correct, constitute a significant extension of entropy econometrics into models of multiple observations. The benefit from entropy methods is in any case the ability to deal with ill-posed models. This may bring relief to researchers in applied econometrics. Lots of assumptions that are required with standard methods to get models identified, can now be dropped. With less assumptions there are less tests to make and less models to reject. This should make more time and effort available for testing of economically relevant hypotheses.

References

Chavas, J.-P. & Cox, T. L. (1995), 'On nonparametric supply response analysis', *American Journal of Agricultural Economics*.

Csiszar, I. (1975), 'I-Divergence Geometry of Probability Distributions and Minimization Problems', *Annals of Probability* **3**(1), 146–158.

Golan, A., Judge, G. & Miller, D. (1996), *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, John Wiley.

Howitt, R. E. (1995), 'Positive mathematical programming', *American Journal of Agricultural Economics* **77**(2), 329–342.

Kullback, J. (1959), *Information Theory and Statistics*, John Wiley, New York.

Lansink, A. O. (1999), 'Generalised maximum entropy estimation and heterogeneous technologies', *European Review of Agricultural Economics* **26**, 101–115.

Paris, Q. & Howitt, R. E. (1998), 'An analysis of ill-posed production problems using maximum entropy', *American Journal of Agricultural Economics* **80**(1), 124–138.

Varian, H. R. (1984), 'The nonparametric approach to production analysis', *Econometrica* **52**(3), 579–591.

Varian, H. R. (1985), 'Non-parametric analysis of optimizing behavior with measurement error', *Journal of Econometrics* **30**, 445–458.