Count Data Stochastic Frontier Models, with an application to the patents-R&D Relationship

Eduardo Fé*†
University of Manchester

Richard Hofler ‡
University of Central Florida

March 6, 2012

Abstract

This article introduces a new Count Data Stochastic Frontier model that researchers can use in order to study efficiency in production when the output variable is a count (so that its conditional distribution is discrete). We discuss parametric and nonparametric estimation of the model, and a Monte Carlo study is presented in order to evaluate the merits and applicability of the new model in small samples. Finally, we use the methods discussed in this article to estimate a production function for the number of patents awarded to a firm given expenditure on R&D.

Key Words: *Discrete Data; Stochastic Frontier Analysis; Local Maximum Likelihood; Maximum Simulated Likelihood; Halton Sequence.*

JEL Classification: *C01, C13, C14, C16, C25, C51.*

*The authors thank the associate editor and two anonymous referees for helpful comments. We also thank E. DeVuyst and Bill Greene who provided helpful advice at several stages of this work and express gratitude for the many helpful comments from participants of the North American Productivity Workshop VI, June 2010, especially Giannis Karagiannis. Finally, we thank Peiming Wang, Iain Cockburn and Martin Puterman for kindly providing the data set used in section 5.

†Health Economics. Room 4.305, Jean McFarlane Building. University of Manchester. Manchester, M13 9PL, UK. email:e.fe-rodriguez@manchester.ac.uk.

‡Department of Economics, University of Central Florida. Orlando, FL, 32816-1400. email: rhofler@bus.ucf.edu

1 Introduction.

There is a large and ever-growing literature about the estimation of Stochastic Frontier models (SFM) and efficiency scores which has origin in the seminal articles by Aigner et al. (1977) and Meeusen and van den Broeck (1977). All of the published models have been developed with a continuous dependent variable in mind. However, there are situations when the output variable of interest is a count (for example, the number of patents obtained by a firm or the number of infant deaths in a region). It is natural that a researcher facing such a count dependent variable case would desire to maintain the discrete aspects of production in the analysis, and therefore the question is whether traditional continuous data methods are still suitable in this setting. There are a number of reasons why new models might be required.

As discussed in Kumbhakar and Lovell (2003), underlying the SFM, there is a multiplicative distance function relating observed output (y), frontier output (f), and inefficiency (d), so that $y = fd$, and empirical counterparts of this model rely on the transformation $\log y = \log f + \log d$ (Kumbhakar and Lovell, 2003, pg. 64-65). If y, f, and d are discrete valued, then the multiplicative scheme is unlikely to be satisfied[1]. Furthermore, zero values are pervasive in count data sets, but log(0) is not defined. Indeed, one could drop observations with a zero output, but this might result in a very high proportion of the data being discarded. Alternatively, zero values could be replaced with a small random perturbation[2]. However this leads to, at least, a loss of efficiency in the estimation of the parameters of the model.

There is an additional, more important, reason why a count data stochastic frontier model is required. Approximating the distribution of a discrete random variable by that of a continuous random variable leads to a loss of efficiency but, more fundamentally, it can also constitute a source of model misspecification (see White, 1982, Cameron and Trivedi, 1986 and Winkelmann, 2008). Fé (2008) notes that discrete distributions often violate the third moment restrictions imposed by continuous data SFM. In particular, if output is discrete, the distribution of log(y) might exhibit too small a skewness or, more importantly, skewness of the wrong sign. In both cases, this results in SFM converging to OLS (the estimated *lambda* approaches zero), even when substantial inefficiency is present in the data[3].

To this date, the construction of count data stochastic frontiers (CDSF) has been the focus of only two papers. The first contribution, Fé (2007, 2008), studies the estimation of production functions and efficiency scores when output is an economic bad. That article proposes a whole class of CDSF models, but the Delaporte distribution (Delaporte 1962, Ruohonen, 1989) is explored at length and its use is illustrated by estimating a production curve for infant deaths in England[4]. The second contribution, Hofler and Scrogin (2008) focuses on the production

[1]For instance, let $y = 3$ and $f = 7$; then there is not an integer value for d solving the assumed identity

[2]A procedure illustrated by Denuit and Lambert (2005), albeit in a different context.

[3]Fé (2008) only considered the case of economic bads, commodities with a negative marginal utility. In this case, the ALS model relies on the distribution of $\varepsilon = v + u$ where v is, say, symmetric and u has non-negative domain, so that the distribution of log(y) should exhibit a long right tail; however simulations in Fé (2008) show that if y is discrete skewness can often be negative.

[4]In that context efficiency scores amount to an measure of relative inequality in the distribution of these deaths,

of economic goods. That paper emphasises that, since inefficiency reduces output, models for under-reported counts can be used to estimate the mean frontier as well as efficiency scores. Hofler and Scrogin (2008) explore the Beta-Binomial model in detail.

Despite their advantages, the models in Fé (2008) and Hofler and Scrogin (2008) are designed with either maximizing an economic good or minimizing an economic bad in mind, and they cannot be used to analyse both types of commodities. However, it is implicit in the literature that a proper stochastic frontier should be easily adaptable to economic goods and economic bads, so that a simple transformation within the model (normally a change of sign) is sufficient to capture the nature of the commodity under study. From this perspective, neither Fé (2008) nor Hofler and Scrogin (2008) present *proper* Stochastic Frontier models. Therefore, the first purpose of this paper is to provide a proper CDSF.

In section 2, we present a new CDSF model that can be used to study inefficiency in the production of discretely distributed economic goods or bads. The model is a conditional mixed Poisson distribution and is easy to implement using a modern computer (and is, therefore, accessible to practitioners). Cross-sectional estimates of inefficiency scores are also simple to calculate, following the same method described in Jondrow et al. (1982). Although our model is relatively adaptable, it might be unsuitable to describe certain data sets. Therefore, at the end of section 2, we propose a consistent kernel-based misspecification test to evaluate the suitability of our parametric model. The test is an straightforward variation of the statistic in Zheng (1996), and is simple to programme with standard software.

If the parametric model in this paper does not pass the test of adequacy, methods are available to fit the data nonparametrically. Following Kumbhakar et al. (2007) we suggest the use of Local Maximum Likelihood (Tibshirani and Hastie, 1987) to estimate our frontier model. Local Maximum Likelihood (LML) enables researchers to encompass some *anchorage* parametric model in a nonparametric fashion thus allowing the introduction of restrictions in the local conditional distribution of data. This attractive property has given the method some popularity (see Loader, 1996, Fan et al., 2002, Fan et al., 2006, Park et al., 2008 and related work by Gozalo and Linton, 2000) and makes it suitable for our application. Just as with the parametric model, its nonparametric counterpart is simple to program using standard software, however, as predicted by Cramer (1986), we found that standard optimisation algorithms experience occasional (but not frequent) problems while estimating the parameters in our models. We discuss this further in section 3.1 and suggest the use of a *global* optimisation algorithm. In particular, we advocate the use of simulated annealing (Goffe, 1996). This algorithm is designed to find global optima and its performance is insensitive to the choice of starting values.

Section 4 reports the results of a Monte Carlo experiment. The section explores the scope of the model when estimated either parametrically or nonparametrically. We find that simulated LML performs very well under a variety of scenarios, including misspecification of the mixing distribution assumed for the inefficiency term. This section includes evidence regarding

give socio-economic and environmental factors.

the performance of continuous output stochastic frontier models under the *wrong-skew* and the zero-output problem. Our results suggest that if output is discretely distributed and one uses a continuous data stochastic frontier models, the parameters of the model will be, at best, inefficiently estimated and, under certain scenarios, misleading. Further evidence regarding the role of model misspecification is provided in section 5, where we present an application of our methods to the study of the relationship between expenditure on research and development (R&D in what follows) and the number of patents at firm level. In our illustration we use the data set in Wang et al. (1998). This data set contains information about 70 pharmaceutical firms from the 1976 wave of the National Bureau of Economic Research R&D Masterfile (Hall et al., 1986), including the count of patents for each firm, the R&D expenditure for that year, and the value of sales in each firm. As emphasised in Wang et al. (1998), this data set contains very limited information, and is subject to substantial unobserved heterogeneity. The scope of the analysis is further restricted by the cross-sectional nature of the data (Hausman et al., 1984)[5]. However the data analysis in section 5 serves well the purpose of illustrating the role of model misspecification in the estimation of efficiency scores, which will be one of the points outlined in the conclusion, section 6.

2 Count Data Stochastic Frontier Model (CDSF).

Consider a collection of competing firms, institutions or administrative units, $(i = 1, \ldots, n)$ producing a commodity in discrete amounts, $y_i \in \{0, 1, 2, \ldots\}$. The commodity under study could be an economic good (for instance, the number of patents awarded to firms), but it could also be an economic bad (e.g. the number of infant deaths observed in a group of hospitals). In the first case, agents engaged in production try to maximize output, while in the second instance authorities will want to cap the incidence of the final product. We are interested in modelling output and efficiency in production, conditional on a number of inputs. As in the existing literature, we begin by defining a mean production frontier in logarithms,

$$\log \tilde{\lambda} = h(x; \beta).$$

where $\tilde{\lambda} \in \mathbb{R}^+$. Assume for now that inefficiency is fixed at a specific value $\varepsilon \in \mathbb{R}^+$, so that the efficiency-corrected, mean deterministic frontier is

$$\log \lambda = h(x; \beta) \pm \varepsilon$$

[5] Addressing unobserved heterogeneity and dynamics presents non-trivial technical challenges in terms of identification of parameters and tackling these issues are left for future work.

Up to here, we are replicating Aigner et al. (1977), except for the fact that ε is being kept fixed. Since we are modelling non-negative count data, we transform the last equation to

$$\lambda = \exp(h(x;\beta) \pm \varepsilon).$$

To progress towards a stochastic model, we borrow from existing literature and assume that actual output, y, has an Poisson distribution conditional on x and ε, with λ as the conditional mean of the distribution. As in Aigner et al. (1977), we have thus established the relationship between the average deterministic frontier and actual output[6]. Therefore, $\mathbb{P}(y|x,\varepsilon) = \text{Poisson}(\lambda)$. Up to this point we have kept inefficiency fixed at some level. However, as in Aigner et al. (1977) we also argue that cross-sectional inefficiency can vary randomly in accordance with an asymmetric probability law with domain on $\mathbb{R}^+$. As in Aigner et al. (1977), we endow ε with a half normal distribution, so that

$$f(\varepsilon) = f(\varepsilon;\sigma) = \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}} \mathbb{I}_{[0,\infty)} \text{ for } \sigma_\varepsilon > 0. \tag{2.1}$$

This density has the advantages of allowing some flexibility, thanks to its scale parameter and, more importantly, it leads to a model whose first order moments are well defined -a property that is not shared by some popular distributions[7].

If $f(\varepsilon)$ is half normal we can write $\varepsilon = |u|$, where u has a normal distribution. With this notation, and letting $h(x;\beta) = x'\beta$, the conditional distribution of y given x follows by averaging

[6] In Aigner et al. (1977) output has a normal distribution with mean $\log\lambda$, given x and ε.

[7] For example, assume that ε has gamma distribution with parameters $\alpha > 0$ and $\delta > 0$ such that $\delta = \alpha$. Then y is conditionally distributed as:

$$\mathbb{P}(y|x) = \frac{1}{\Gamma(\alpha)} \int_0^\infty \frac{(\exp(-\exp(x'\beta + \frac{\epsilon}{\delta}))) \exp(y(x'\beta + \frac{\epsilon}{\delta}))}{y!} \epsilon^{\alpha-1} e^{-\epsilon} d\epsilon. \tag{2.2}$$

Under the assumed gamma distribution, the density function of the transformation is

$$f(\nu) = f(\exp(\varepsilon)) = \frac{\delta^\alpha}{\Gamma(\alpha)} \frac{(\log(\nu))^{\alpha-1}}{\nu^{\delta+1}} \text{ for } \nu \in [1,\infty). \tag{2.3}$$

However, under the assumption $\delta = \alpha$, moments of order δ or above won't exist, since the integral $\int_1^\infty \log(x)/x dx$ is not convergent. This affects the distribution of y, whose moments depend on those of ν through the expression

$$E(y^r) = \sum_{s=1}^{r} S(r,j) E(\nu^j) \text{ for } j = 1, 2, \ldots$$

(see Karlis and Xekalaki, 2005) where $S(.,.)$ are Stirling's numbers of the second kind.

$\mathbb{P}(y|x,u)$ over the range of u,

$$
\begin{aligned}
&\mathbb{P}(y|x;\sigma,\beta) \\
= &\int_{-\infty}^{\infty} \frac{(\exp(-\exp(x'\beta \pm |u|)))\exp(y(x'\beta \pm |u|))}{y!} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\
= &E\left[\frac{(\exp(-\exp(x'\beta \pm \sigma|u|)))\exp(y(x'\beta \pm \sigma|u|))}{y!}\right] \\
= &E\left[\text{Poisson}(\exp(x'\beta \pm \sigma|u|))\right] \qquad (2.4)
\end{aligned}
$$

where expectations are taken with respect to the standard normal distribution. We refer to this specification as a Poisson log-half normal distribution or, for simplicity, Poisson half normal (PHN) distribution.

When $f(\varepsilon)$ is a half normal density function, it is not difficult to show (and this is done in Appendix A) that the transformation $\nu = \exp(+\varepsilon)$ has density and first two moments:

$$
\begin{aligned}
f(\nu) &= \frac{2}{\nu\sigma\sqrt{2\pi}} e^{-\frac{(\log(\nu))^2}{2\sigma^2}} \mathbb{I}_{[1,\infty)} && (2.5) \\
E(\nu) &= e^{\sigma^2/2}\left\{1 + \text{Erf}\left(\frac{\sigma}{\sqrt{2}}\right)\right\} \geq 1 && (2.6) \\
\text{var}(\nu) &= e^{2\sigma^2}\left\{1 + \text{Erf}\left(\sigma\sqrt{2}\right)\right\} - \left\{e^{\sigma^2/2}\left\{1 + \text{Erf}\left(\frac{\sigma}{\sqrt{2}}\right)\right\}\right\}^2 && (2.7)
\end{aligned}
$$

where Erf is the error function and $\mathbb{I}$ is an indicator function. When $\nu = \exp(-\varepsilon)$ we have,

$$
\begin{aligned}
f(\nu) &= \frac{2}{\nu\sigma\sqrt{2\pi}} e^{-\frac{(\log(\nu))^2}{2\sigma^2}} \mathbb{I}_{[0,1]} && (2.8) \\
E(\nu) &= e^{\sigma^2/2}\text{Erfc}\left(\frac{\sigma}{\sqrt{2}}\right) \in [0,1] && (2.9) \\
\text{var}(\nu) &= e^{2\sigma^2}\text{Erfc}\left(\sigma\sqrt{2}\right) - \left\{e^{\sigma^2/2}\text{Erfc}\left(\frac{\sigma}{\sqrt{2}}\right)\right\}^2 && (2.10)
\end{aligned}
$$

where now Erfc is the complementary error function. Note that the density functions are identical, and they only differ as to their domains. Both distributions have well defined means and variances, from which it follows that, for $\nu = \exp(\varepsilon)$,

$$
\begin{aligned}
E(y|x) &= \tilde{\lambda} e^{\sigma^2/2}\left\{1 + \text{Erf}\left(\frac{\sigma}{\sqrt{2}}\right)\right\} && (2.11) \\
\text{var}(y|x) &= E(y|x)\left[1 + \tilde{\lambda} W\right] && (2.12)
\end{aligned}
$$

where $\tilde{\lambda} = \exp(x'\beta)$ and

$$W = \frac{e^{2\sigma^2}\left\{1 + \mathrm{Erf}\left(\sigma\sqrt{2}\right)\right\} - e^{\sigma^2/2}\left\{1 + \mathrm{Erf}\left(\frac{\sigma}{\sqrt{2}}\right)\right\}}{e^{\sigma^2/2}\left\{1 + \mathrm{Erf}\left(\frac{\sigma}{\sqrt{2}}\right)\right\}}.$$

The mean and variance for the case when y is an economic good can be obtained similarly. In both cases, the first two moments of the conditional distribution of y are well defined: they exist and their existence does not hinge on the value of the parameters.

The PHN distribution does not have a closed form expression, however the integral in (2.4) can be approximated by simulation[8]. Gourieroux et al. (1984a,b) note that the unfeasible log-likelihood function derived from (2.4), say $L_n(\theta) = \sum_i \log \mathbb{P}(y_i|x_i;\theta)$ can be approximated by any of the following criteria:

$$L^I_{H,n}(\theta) = \sum_{i=1}^{n} \log\left[\frac{1}{H}\sum_{h=1}^{H}\mathbb{P}^*(y_i|x_i;\xi_h;\theta)\right] = \sum_{i=1}^{n}\log\hat{\mathbb{P}}^I(y_i|x_i,\xi_h;\theta) \quad (2.13)$$

$$L^D_{H,n}(\theta) = \sum_{i=1}^{n} \log\left[\frac{1}{H}\sum_{h=1}^{H}\mathbb{P}^*(y_i|x_i;\xi_{i,h};\theta)\right] = \sum_{i=1}^{n}\log\hat{\mathbb{P}}^D(y_i|x_i,\xi_{i,h};\theta) \quad (2.14)$$

where ξ are draws from the distribution of ε and the subscripts I(dentical) and D(ifferent) refer to whether the simulated values, ξ, are the same for each i or are drawn as i varies. The distinction is not trivial for the asymptotic properties of the method (Gourieroux et al. (1984b)), since

1. If $nH^{-1} \to 0$, then $\sqrt{n}(\hat{\theta}^I - \theta_0) \to N(0, \mathcal{I}^{-1}(\theta_0))$,
2. If $\sqrt{n}H^{-1} \to 0$, then $\sqrt{n}(\hat{\theta}^D - \theta_0) \to N(0, \mathcal{I}^{-1}(\theta_0))$

where $\mathcal{I}(.)$ is the information matrix. Since the number of Monte Carlo (MC) draws, H, is decided a priori by the researcher, arbitrary accuracy can be attained with each method by choosing H large enough.

In this paper we advocate combining Maximum Simulated Likelihood estimation of the PHN model with Halton sequences (Gentle, 2003, Greene, 2003). It is well known (Caflisch et al., 1997, Train, 2003) that standard Monte Carlo draws tend to form clusters and leave unexplored areas in the unit cube, thus reducing the accuracy of the Maximum Simulated Likelihood method. Low discrepancy methods such as Halton sequences give a better coverage and, as with antithetic draws, they generate negatively correlated nodes which results in a reduction of the error due to simulation[9].

[8] Gaussian Quadrature methods (see Press et al. 1992 or Judd 1998) are unlikely to work well with this model. The accuracy of quadrature methods depends on the smoothness of $f(x)$ in the sense of being well-approximated by a polynomial (see Press et al., 1992 and DeVuyst and Preckel, 2007).

[9] This occurs because the variance of the sum of any two draws is less than the variance of the sum of two independent draws (Gentle 2003). A clear drawback of Quasi-Monte Carlo methods is the deterministic nature of the sequence which results in negatively correlated random draws (even though the strength of the correlation might be small). This contradicts the assumptions of independent random draws on which the above asymptotic

Applying simulation, the conditional distribution of y_i (for $i = 1, \dots, n$) is approximated by the sum:

$$\mathbb{P}(y|x;\theta) \approx \hat{\mathbb{P}}(y|x;s_h,\theta) = \frac{1}{H}\sum_{h=1}^{H} \text{Poisson}(\exp(x'\beta \pm \sigma|s_h|)) \tag{2.15}$$

where s_h are the terms of a, possibly randomized, Halton sequence. The infeasible log-likelihood $L_n = \sum_{i=1}^{n} \log \mathbb{P}(y|x;\theta)$ can be approximated by $L_{n,h} = \sum_{i=1}^{n} \log \hat{\mathbb{P}}(y|x;s_h,\theta)$. The analytical derivatives of $L_{n,h}$ are given by

$$\frac{\partial L_{n,h}}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{\hat{\mathbb{P}}(y_i|x_i;s_h,\theta;)} \frac{1}{H} \sum_{h=1}^{H} \text{Poisson}(\tilde{\lambda}_{i,h})(y_i - \tilde{\lambda}_{i,h}) \left\{ \begin{array}{c} x_i \\ \pm|s_h| \end{array} \right\} \tag{2.16}$$

where $\tilde{\lambda}_{i,h} = \exp(x_i'\beta \pm \sigma|s_h|)$. The value $\hat{\theta}_{MSL}$ making the above system of equations equal to zero is the Maximum Simulated Likelihood estimator of θ. When the PHN is a correct representation of the underlying data generating process, $\hat{\theta}_{MSL}$ is a consistent, asymptotically normal and efficient estimator of the true parameter value, as follows from properties 1 and 2 above. Standard errors can be computed via the BHHH estimator or the minus inverse of the Hessian Matrix of $L_{n,h}$.

Tests of hypotheses can rely on the Wald-Score-LR trinity. Testing the null hypothesis of no inefficiency is of particular interest. The null hypothesis would be

$$H_0 : \sigma = 0$$

in which case PHN collapses to a standard Poisson model. Equation (2.11) already suggests that our model generates overdispersion in the data (in a vein identical to the Negative Binomial model[10]). Therefore this test is similar in nature to those commonly used for testing the Negative Binomial regression against the benchmark Poisson distribution. A formal test can be computed via a likelihood ratio comparing the resulting value with the quantiles of a χ_1^2 distribution. Note that this procedure does not yield a consistent test and, in general, rejection of the null hypothesis is not evidence in favour of the PHN or, indeed, any other Mixed Poisson Model.

Researchers might want to evaluate the overall validity of the PHN model. Then, a consistent misspecification test can be easily constructed following Zheng (1996). In particular, under the null hypothesis of correct specification, $E(Y|X)$ is given by equation (2.11), and therefore $E(e(\theta;Y,X)|X) = E(Y - E(Y|X)|X) = 0$, almost surely. A test procedure can be based on the

results hold. It is possible, however, to generate randomized Halton draws without compromising the properties of the original sequence. In particular Bhat (2003) advocates shifting the original terms of the Halton sequence, say s_h, by a quantity μ that has been drawn randomly from a uniform distribution on $[0, 1]$. In the resulting sequence those terms exceeding 1 are transformed so that $s_h^* = s_h + \mu - 1$.

[10]This is not a special feature of our model and is shared, in general, by Mixed Poisson models -see Karlis and Xekalaki, 2005.

following set of hypothesis,

$$\begin{aligned} H_o &: \mathbb{P}\left(E(e(\theta_o;Y_i,X_i)|X_i)=0\right)=1 \text{ for some } \theta_o\in\Theta \\ H_a &: \mathbb{P}\left(E(e(\theta;Y_i,X_i)|X_i)=0\right)<1 \text{ for all } \theta\in\Theta \end{aligned}$$

and the ensuing test (Zheng 1996) is given by

$$Z_n=\frac{\sum_{i=1}^{n}\sum_{i\neq j}^{n}K_{ij}\hat{e}_i\hat{e}_j}{\sqrt{2\sum_{i=1}^{n}\sum_{i\neq j}^{n}K_{ij}^2\hat{e}_i\hat{e}_j}} \tag{2.17}$$

where $K_{ij}=h^{-d}\prod_{l=1}^{d}k(X_{i,l}-X_{j,l}/h)$, $k(.)$ is a kernel function, h is a bandwidth parameter (such that $h\to 0$ as $n\to\infty$), $\hat{e}=e(\hat{\theta};Y_i,X_i)|X_i)$ and $\hat{\theta}=\hat{\theta}_{MSL}$ is, in this instance, the Maximum Simulated Likelihood estimator of θ. Zheng (1996) shows that the distribution of Z_n converges to an standard normal as $n\to\infty$ and, therefore, a *one-sided* test can be based on the quantiles of the standard normal distribution.

Although the parameters of the production frontier are of interest in themselves, the ultimate goal of our analysis is to obtain approximate efficiency scores for each individual in the sample. Following Jondrow et al. (1982) cross-sectional inefficiency scores, $v=\exp(\pm|u|)$, can be estimated via $E(v|y,x)$. Using Bayes' theorem

$$f(v|x,y)=\frac{\mathbb{P}(y|x,v)f(v)}{\mathbb{P}(y|x)} \tag{2.18}$$

so that

$$E(v|y,x)=\int vf(v|x,y)dv. \tag{2.19}$$

The latter expression does not have a closed form. However, we may still approximate the relevant integral via simulation. The simulated $E(v|y,x)$ for the parametric Mixed Poisson model is

$$\hat{v}_i \quad = \quad E(v_i|x_i,y_i)\approx\frac{\sum_{h=1}^{H}e^{\pm|s_h|\sigma}\text{Poisson}(\exp(x_i'\beta+\sigma|s_h|))}{\sum_{h=1}^{H}\text{Poisson}(\exp(x_i'\beta+\sigma|s_n|))} \tag{2.20}$$

It is convenient to reproduce here the observations made by Wang and Schmidt (2009). The distributions of v and $\hat{v}$ are not the same, as can be seen by noting that $\text{var}(v)=\text{var}(E(v|x,y))+E(\text{var}(v|x,y))$, (and hence, $\hat{v}$ has smaller variance). Furthermore, $\hat{v}$ is a shrinkage[11] of v toward its mean. As a result, the lower and upper tails of the distribution of v will be misreported, so that $\hat{v}$ penalizes outstanding firms and rewards the least efficient individuals -although the average efficiency in the sample is correctly approximated. Quoting Wang and Schmidt (2009) this "*...does not mean that there is anything wrong with the estimator* since ... *it is unbiased in the unconditional sense* $E(\hat{v}-v)=0$".

[11]See, for instance, Gourieroux and Monfort (1995))

3 Nonparametric Estimation via Local Likelihood.

One of the benefits of the parametric model is that they provide a large amount of (easily accessible) information about the frontier. Although the parameter σ of the mixing distribution allows some flexibility for data analysis, model misspecification would render the maximum likelihood estimator inconsistent and inefficient (White, 1982). Furthermore, PHN assumes a convex conditional mean production frontier. This is a sensible assumption if output is an economic bad, but it contradicts economic theory if output is an economic good.

Smoothing techniques can be used to mitigate the effects of model misspecification, but in the current context it is customary to accommodate the under-/over-production induced by inefficiency in the data. Standard nonparametric techniques (such as the local polynomial regression estimator) do not accommodate restrictions on the conditional distribution of the dependent variable and therefore, following Kumbhakar et al. (2007), we advocate the use of Local Maximum Likelihood (Tibshirani and Hastie (1987)). In Local Maximum Likelihood (LML) the parameters of the distribution, θ, but not necessarily the type of distribution itself, are allowed to vary along the domain of X, so that $\theta = \theta(x)$. Thus, in the normal regression case this would imply that $Y|X \sim N(\mu(x), \sigma^2(x))$ where $\mu(.)$ and $\sigma^2(.)$ are left unrestricted. Here x is the point or node in the domain of X at which estimation takes place

Let d denote the number of regressors in the model, and let p denote the number of parameters in θ. In its general formulation, the LML estimator of the parameters in a model is $\hat{\theta}_0 = \hat{\theta}_0(x)$ in

$$\left(\hat{\theta}_0(x), \hat{\theta}_1(x) \ldots \hat{\theta}_m(x)\right) = \underset{\theta_0, \theta_1 \ldots \theta_m}{\arg\max} \sum_{i=1}^{n} \log f(Y_i, \tilde{\theta}(x)) K_h(X_i - x) \tag{3.1}$$

where, in the univariate case[12]

$$\tilde{\theta}(x) = \theta_0 + \theta_1(X_i - x) + \theta_2(X_i - x)^2 + \ldots + \theta_m(X_i - x)^m$$

$K_H(X_i - x) = \prod_{j=1}^{d} h_d^{-1} k(x_{ij} - x_j/h_j)$, $k(.)$ is a univariate kernel function and $h_j = h_j(n)$ is a smoothing parameter such that $h_j \to 0$ as $n \to \infty$ for all $j = 1, \ldots, d$. Kumbhakar et al. (2007) show that $\hat{\theta}^{(0)}(x)$ is a consistent, asymptotic normal estimator of $\theta^{(0)}(x)$.

For the PHN distribution, the parameters of interest are $\theta(x) = (\mu(x), \sigma(x))'$ in the conditional distribution

$$\mathbb{P}(y|x; \theta(x)) = \int_{\varepsilon} \text{Poisson}(\exp(\mu(x) \pm \sigma(x)|u|) f(\varepsilon) d\varepsilon \tag{3.2}$$

The log-link in the conditional mean is present to guarantee the non-negativity of the mean parameter but, as we will show in the simulations, it turns out to be a rather innocuous assumption.

[12] In the multivariate case, for $m = 1$, $\tilde{\theta} = \theta_0 + \Theta_1(X_i - x)$, where Θ is a $p \times d$ matrix of parameters.

The infeasible local linear ($m = 1$) conditional maximum likelihood function is

$$\mathbb{P}(y_i|x_i;\theta(x)) = \sum_{i=1}^{n} \left[\log \int_u \text{Poisson}\left\{\lambda_i(x;u_i)\right\} f(u_i)du_i\right] K_H(X_i - x) \tag{3.3}$$

where

$$\lambda_i(x;u_i) = \exp\left(\mu^{(0)}(x) + \mu^{(1)}(x)(X_i - x)' \pm \left(\sigma^{(0)}(x) + \sigma^{(1)}(x)(X_i - x)'\right)|u_i|\right). \tag{3.4}$$

Local likelihood functions of higher order than one can be obtained similarly. As in the parametric case, the integral in (3.3) can be approximated by simulation, yielding the simulated local likelihood function, which can then be implemented with standard econometric software.

For the nonparametric model, the appropriate local estimator of inefficiency is

$$\hat{v}_i(X_i) \quad = \quad \approx \frac{\sum_{h=1}^{H} e^{\pm|s_h|\sigma^{(0)}}\text{Poisson}(\exp(\mu^{(0)}(X_i) + \sigma^{(0)}(X_i)|s_h|))}{\sum_{h=1}^{H}\text{Poisson}(\exp(\mu^{(0)}(X_i) + \sigma^{(0)}(X_i)|s_h|))} \tag{3.5}$$

Feasible estimators would follow by using the (local) maximum simulated likelihood estimates of β, σ, $\mu^{(0)}(x)$ and $\sigma^{(0)}(x)$. Non-parametric models may fit the data better, but it must be noted that they may provide limited information as to the shape of frontier. Thus it may be difficult to gauge whether the non-parametric model fits with economic expectations. As such there is still a role for the parametric model, and the tests for its adequacy presented in the previous sections are bound to be useful in this respect.

3.1 A note on maximisation algorithms.

In our experience, the BFGS[13] algorithm for maximisation tends to produce fairly reliable results when implementing the parametric and nonparametric PHN, and the use of the analytical score speeds up the implementation of the algorithms. However, when optimising the local likelihood, there were a small number of instances when this algorithm failed to converge to a global optimum (or to converge at all) at certain nodes. This could be expected. As pointed out by Cramer (1986), although the log-likelihood function is asymptotically concave, in small samples it might exhibit plateaus or several modes that will mislead conventional optimisation algorithms. Therefore, as in Cameron and Johansson (1997), we suggest the use of Simulated Annealing (SA) as the core optimisation algorithm. As described in Goffe et al. (1994) and Goffe (1996), SA enjoys some very desirable properties: it is insensitive to the starting values, it can escape from local optima and go on to find the global optimum, and it makes less stringent assumptions regarding the function than do conventional algorithms (for instance, it does not assume that the objective function is continuous). All these features come with an increase in computer time. However, this is a relatively low price to pay, given current computer hardware.

[13] Broyden-Fletcher-Goldfarb-Shanno (see, for instance, Press et al., 1992)

4 Evidence from a Monte Carlo Experiment.

To evaluate the small sample merits of the techniques discussed in the previous sections, we designed the following Monte Carlo experiments. All the simulations were run in Ox v.5 Doornik (2007). The optimisation algorithms used were `MaxSQP` a sequential quadratic programming technique to maximize a non-linear function subject to non-linear constraints and Charles Bos' `MaxSA`, an Ox implementation of the SIMAMM algorithm in Goffe (1996)[14].

4.1 Experiment 1.

In this experiment we explored the effect that different degrees of inefficiency have on the overall small sample performance of Maximum Likelihood when estimating the PHN using simulation methods. We generated data from the distribution

$$\mathbb{P}(y|x, \varepsilon) = \text{Poisson}(\exp(x'\beta - |u|)) \tag{4.1}$$

where $x' = (1, x_1, x_2)$, with $x_1 \sim \text{Uniform}(0, 5)$, $x_2 = \rho x_1 + \sqrt{1 - \rho^2}V$, $V \sim \text{Uniform}(0, 5)$, $\rho = 0.5$, and the vector of parameters β was set equal to $(0.3; 0.3; 0.3)'$. The distribution of u was $N(0, \sigma^2)$, and we considered $\sigma \in \{0.5, 1, 2, 4\}$, implying that $E(\exp(-|u|))$ took values 0.69924, 0.52316, 0.33620 and 0.18882. The sample sizes were 100, 400 and 1600 observations. In the simulations, the structural part of the conditional mean was correctly specified, so that $\tilde{\lambda} = \exp(x'\beta)$. The model was estimated 2000 times. For the Maximum Simulated Likelihood procedure, we used Halton sequences of 30 terms[15]. At the end of each cycle we calculated the Mean Squared Error of the estimated parameters and the mean estimated inefficiency. The results are summarised in Tables 1 and 2.

Table 1 shows that, for each model's parameters, the mean squared errors are small and, as expected, they decrease with the sample size because the (simulated) maximum likelihood estimator is consistent. The accuracy with which the intercept is estimated is sensitive to the value of σ: it decreases as σ increases, which is not surprising because as σ increases, so does the variation in the data. Similarly, the estimates of σ worsen as its true value increases. In Table 2 we report the average estimated inefficiency obtained with the estimator (2.20). The estimates are close to the true value although performance deteriorates slightly as σ increases, as anticipated by the results in Table 1.

In general, these results confirm that if the PHN is correctly specified, then reliable estimates can be obtained through the maximum simulated likelihood procedure outlined in the this article. However, in very small samples, large values of σ will imply a small loss of accuracy in the estimation of the intercept and efficiency scores.

[14] `MaxSa` is available for download at Bos' website: `http://www.tinbergen.nl/~cbos/`

[15] We used up to 200 Halton draws in our simulations, but did not observe fundamental variation in our results.

4.2 Experiment 2.

In this experiment we wanted to assess the ability of LML to capture alternative functional forms. Thus, we generated data from Poisson half normal models with conditional means

$$\begin{aligned}
\text{Model 1: } \tilde{\lambda} &= \theta_1^{\theta_2} - |u| & (4.2)\\
\text{Model 2: } \tilde{\lambda} &= \exp(\gamma_1 - \gamma_2^2 + \gamma_3 x^3 - |u|) & (4.3)\\
\text{Model 3: } \tilde{\lambda} &= \exp(\beta_1 + \beta_2 x - |v|) & (4.4)
\end{aligned}$$

where $(\theta_1, \theta_2) = (5, 0.5)$, $(\gamma_1, \gamma_2, \gamma_3) = (4, -1.65, 0.2)$, $\beta_1 = \beta_2 = 0.3$, $u \sim N(0, 1)$, $x \sim \text{Uniform}(0, 5)$ and $v \sim t_3$ (a Student's t distribution with three degrees of freedom). Model 1 assumes a concave Cobb-Douglas conditional mean, while model 2 assumes a highly non-linear conditional mean. Model 3 is devised to analyse the robustness of the method to distributional departures in the mixing parameter. The Student's distribution with 3 degrees of freedom has much longer tails than the standard normal, and therefore we would expect that the PHN will tend to under-estimate the level of the conditional mean.

We drew samples of 100 observations 1000 times, and we calculated the average estimated conditional mean and its fifth and ninety-fifth quantiles. We used Halton sequences of 20 terms[16]. The LML estimator was implemented over a grid of 20 equally spaced points in the interval $[0, 5]$. The result of the experiment are contained in Figures 1 to 3.

We observe in Figures 1 and 2 that the estimated conditional mean is almost indistinguishable from the true conditional mean. Only in Figure 2 do we observe a discrepancy between true and estimated curves about the interval [1.5,2.5]. However, the discrepancy is very small, and the overall performance of the nonparametric PHN model is very good. When data were generated from model 3, the estimator captured the essential features of the underlying conditional mean but, as expected, the mean estimated conditional mean tended to underestimate the true conditional mean. Therefore, misspecification of the mixing distribution has an effect, as least in the very small samples considered in this experiment. However, the LML procedure is still capable of unveiling the non-linearities in the underlying data generating process.

4.3 Experiment 3: The performance of continuous data methods.

How does the benchmark normal-halfnormal model in Aigner et al. (1977) behaves when output is discrete valued? To address this question, we simulated 100,000 observations from a simple PHN model with conditional mean $\tilde{\lambda}\theta$, where $\tilde{\lambda} = 1, 2, 4, 8, 16$ and $\theta = \exp(-\sigma|u|)$, where $\sigma^2 = 1, 2, 3, 4, 5$. We then calculated $W = \log(Y)$, the sample skewness coefficient of W, and the proportion of data dropped (because $Y = 0$). Finally, we fitted a basic normal halfnormal model,

$$\log(Y) = \alpha - u - v, \text{ where } v \sim N(0, \sigma_v^2), u \sim N^+(0, \sigma_u^2), \tag{4.5}$$

[16] Longer sequences could be used, at the cost of increased computer time.

where $N^+(.,.)$ denotes a half-normal distribution. The ratio $\eta = \sigma_u/\sigma_v$ measures the relative importance of inefficiency in the sample[17]. Values of η close to zero suggest that there is no inefficiency in the sample and that estimation of the model can be done via ordinary least squares. On this occasion the analysis was done in Stata (version 10). The results are given in table 6.

For the case of production frontiers, the normal halfnormal model requires that the distribution of W exhibits negative skew. Yet, table 6 shows that for $\tilde{\lambda} < 8$ the sample skewness coefficient of W is positive. In this case Y ranges between 0 and 13, but the prevalence of zero values is large (and increases as σ^2 increases). For instance, in the most favourable case ($\tilde{\lambda} = 4$ and $\sigma^2 = 1$) 20% of the observations had $Y = 0$ and had to be eliminated from the sample. Thus, not only does data violate one of the fundamental assumptions of the normal half-normal model, but estimation will be necessarily much less efficient, given the loss of data points. The effect of model misspecification can be evaluated by fitting a normal-halfnormal model to the data, and calculating the average efficiency score. Consider, for instance, the case $\tilde{\lambda} = 4$ and $\sigma^2 = 1$. We obtained the following estimates, $\sigma_u = 0.0017389$ (standard error equal to 0.0418485) and $\eta = \sigma_u/\sigma_v = 0.002783$ (0.0419141). In this case inefficiency had gone undetected.

When $\tilde{\lambda} \geq 8$, we begin to observe negative skewness in the distribution of W, and a decrease in the incidence of zeros. The most favourable setting in our simulation corresponds to $\tilde{\lambda} = 32$, $\sigma^2 = 1$. In this case $0 \leq Y \leq 51$, the estimated coefficients of skewness and kurtosis of Y are 0.3 and 2.4 respectively and the true mean efficiency equals 0.52022. The estimated coefficient of skewness of W was -1.07. When estimating the normal-halfnormal model we could make use of 99% of the data, obtaining estimates $\hat{\alpha} = 3.503$, (0.0019) $\hat{\eta} = 7.503$ (0.004), $\sigma_u = 1.11453$ (0.0030574). Thereforethe estimated mean and average efficiency were correctly approximated at values 33.11 and 0.4992015 respectively, even though not all data was used in the estimation. Together with results in White (1982), this implies that, even in the most favourable scenario, the estimates produced with the normal halfnormal model will be inefficient.

The results so far suggest that looking at the range of values of Y can be a useful first step to assess the validity of the normal-halfnormal model because, as we have seen, the normal halfnormal model performs well (albeit inefficiently) when the conditional mean takes large values. However this can be very misleading, as we now show. We simulated 10,000 observations from the PHN model for patent data that we estimate in the following section. In particular, based on tables 3 and 4, we defined covariates $X_1 \sim \text{Uniform}[-3, 5]$, $X_2 \sim \text{Uniform}(0, 8)$, $X_3 = \exp(X_1)/\exp(X_2)$, $\beta = (2.7, 1.2, -0.3, 4.6)$ and $\sigma = 1$. In this case, the maximum observed value of Y was 444. However only 53.23% of the observations could be used to estimate the parameters of the model -due to the zero problem- and the coefficient of skewness was small and positive. As a result, when the normal-halfnormal model was fitted to the data, we obtained a $\hat{\eta} = 0.0031583$ (0.2167242) so that no inefficiency was detected.

Overall, this simulation presents compelling evidence that the standard normal halfnormal

[17]In the literature, this ratio is usually denoted by the Greek letterλ. Here we had to use η to avoid clashes of notation.

model faces serious challenges when dealing with count data. The generating process considered often violate the assumptions underlying the continuous data models. Besides, the number of observations with zero value of Y is likely to be substantial, and this will lead to a significant loss of efficiency. Therefore, in practice, it is essential to obtain numerical measures of skewness for the empirical distribution of $\log(Y)$ prior to trying to fit a normal-halfnormal model. Only if the sample skewness coefficient has the right sign and is substantial can that model be informative. However, even in the most favourable situations, the PHN model will be more efficient, and so, given that the latter is parsimonious and easy to implement with modern software, we advocate the use of the PHN model whenever analysing efficiency with a count dependent variable.

5 Application.

In this section we use the PHN distribution to estimate (parametrically and nonparametrically) a production function for the number of patents awarded to a firm in a give year. The main goal of this section is, not only to shed further light regarding the applicability of the new model, but also to emphasise some of the issues outlined in the introduction regarding the estimation of a continuous data SFM on count data.

In the application we use the data set in Wang et al. (1998)[18]. This file contains information about 70 pharmaceutical firms from the 1976 wave of the National Bureau of Economic Research R&D Masterfile (Hall et al., 1986). The variables for each firm in the data set are the count of patents, the R&D expenditure and the value of sales. Restricting attention to a single industry mitigates the potential effect of neglected inter-industry heterogeneity. However the data set does not contain abundant information, and any unobserved heterogeneity will be allocated to the mixing variable of the conditional Poisson model (as would also happen in existing cross-sectional SFM), thus distorting the estimates of inefficiency. Omission of variables will also bias the estimated coefficients of the regressors unless the omitted and included variables are independent. Previous work in Hausman et al. (1984) also suggests that a full account of efficiency in the production of patents requires an analysis of the dynamics and cross-correlations between and within firms. Addressing dynamics and heterogeneity requires panel data versions of our models, and presents non-trivial challenges in terms of identification of parameters (Neyman and Scott, 1948, Lancaster, 2000). This is left for later work. However, these aspects of the application should not interfere with the goal of the section which is to illustrate the applicability of the new methods presented in this paper and to show some of the issues arising when estimating existing SFM on count data.

The relationship between patents and research and development has been studied at length by a number of authors. Key references are Pakes and Griliches (1980), Hausman et al. (1984), Hall et al. (1986), Pakes (1986), Griliches (1990) and Wang et al. (1998) among others. The general view, which we adopt here, is that annual expenditure in R&D is a type of investment,

[18]We thank the authors for kindly providing us with their data set.

and therefore, it enriches the stock of knowledge of a given firm. This knowledge materializes as new technologies, some of which are registered and, subsequently, a patent is awarded to the firm in order to protect its property rights on the invention. Therefore, patents represent the value of the underlying stock of knowledge. The authors mentioned above have acknowledged that patents are not the only output of R&D, and their economic value can be under question, as a large proportion of granted patents are of little economic value or depreciate too quickly (see Pakes (1986)). However, as has been pointed out, firms are awarded patents and this is the product of research and development undertaken within the firm, so that patents measure the levels of research activities -*...even though the information conveyed by an individual patent may be very small* (Hall et al. (1986)).

As in earlier contributions, we estimate a production function for patents, conditional on some function of R&D and other covariates. In addition, bearing in mind the restrictions presented by the limited data set, we use the PHN model to estimate inefficiency in the production of patents. The model we fit assumes that each observation is independently distributed, with a Poisson distribution, conditional on ε. The regressors in the conditional mean are $\log \text{R\&D}$, $\log \text{Sales}$ and the ratio R&D / Sales, which we use to approximate firms' research effort. We estimate the PHN model, parametrically and nonparametrically, as well as the standard continuous data normal-half normal stochastic frontier model in Aigner et al. (1977).

Summary statistics can be found in table 3. The first important feature of the data is the small and positive skewness coefficient of log(Patents). Because we are dealing with economic goods, implementation of the normal-half normal model requires negative skewness in the dependent variable. This disagreement between theory and data will have important consequences. Table 4 presents the results obtained with the normal half-normal model. The results are troubling because they are so implausible. Firstly, estimation of this model required that 13 observations with zero output were dropped from the sample (18% of the sample). More importantly, the normal-half normal model fails to capture any inefficiency in the sample. The mean efficiency is almost zero, meaning that the average firm is producing on the frontier of the patents that it can theoretically output in a year. On the contrary, the PHN model detects an average efficiency of 0.47 -that is average output is around 0.47λ. This is a high level of inefficiency, but the consistent model misspecification test in equation (2.17) takes a value of 0.561, so that the null hypothesis is not rejected at any significance level (the critical value is 1.645). In other words, this model misspecification test result supports the PHN model as probably being correct. Given the limitations imposed by the data set, it is conceivable that the estimated inefficiency has been distorted due to unobserved heterogeneity. Furthermore, the nonparametric test is consistent asymptotically, but in this occasion we have a very small sample of 70 observations. Therefore misspecification remains a concern. The results provided by the local PHN seem to confirm this (see figure 4). We implemented the LML method in section 3, with estimation at each X_i, 300 Halton numbers and bandwidth $h = 0.5\sigma_x n^{-1/5}$. The nonparametric method suggests an average efficiency of 0.56 which about 20 % higher than that reported by the parametric PHN model.

The empirical distribution of inefficiency is described in Table 5. Half of the firms in the sample attain efficiency levels above 56%, while a quarter of the firms attain efficiency levels above 75%. However, according to these results, about a quarter of the firms got efficiency scores below 32%, which still seems low. Bearing in mind the limitations of the data, what these results suggest is that, considering the consistent, distribution-free nature of LML estimator, the low mean efficiency reported by the parametric model was driven to a great extent by misspecification of the functional form. As we pointed out at the beginning of this section, our data set has limited information, and there is a concern that some of our results might be affected by unobserved heterogeneity that might affect our estimates of inefficiency. Furthermore, our model is static, but it has been emphasized in the literature that the dynamics of expenditure in R&D have been shown to be important, as well as the within and between correlations among firms. Therefore, these findings have limited external validity and should be viewed only as the results of an illustration of how our model can be applied to cases involving discretely-distributed economic goods or bads.

The parameter estimates reveal several facts about patents and R&D. First of all, the coefficient of the log (R&D) variable provides a direct estimate of the elasticity of innovative output with respect to R&D spending, and, consequently, the returns to scale to performing R&D. An elasticity less than (greater than) one implies decreasing (increasing) returns to scale. Our significant estimate of 1.2074, therefore, indicates increasing returns to scale. This compares to a range of elasticities from seven models in Wang et al. (1998) from .6716 up to 1.5884. That is, they report both decreasing and increasing returns to scale, depending on the model they estimated. Thus, our estimate, even though it is from a model completely different from those estimated by Wang et al. (1998), is near the mid-point of their range of elasticities. The measure of R&D intensity (R&D / Sales) has a negative coefficient in both of the two (out of seven) models where it is used by Wang et al. (1998), as does our estimate. However, our coefficient's standard error implies that this variable probably does not influence patent output. This agrees with one of those two models containing (R&D / Sales). We compare our results to those found in Wang et al because their models are most like our model in terms of the regressors used. However, our model is unlike any estimated in the literature on patents and R&D. Despite this fundamental difference, our elasticity and R&D intensity results are similar to those in Wang et al. This suggests that our PHN model can give sensible coefficient estimates when it is applied to count data.

6 Conclusion.

This article has introduced a new probability distribution function, the Poisson log-half normal (PHN), that can be used to estimate the production frontier -and efficiency scores- of a discretely distributed variable. This model solves a number of problems that arise when researchers try to fit data with existing continuous data Stochastic Frontier Models (SFM), among which the most important is the conflict between the restrictions that SFM impose on third moments (i.e. negative skewness when Y is being maximised) and the conflicting coefficient of skewness typically exhibited by the distribution of the logarithm of a count data variable. We have discussed two ways of estimating this model. Firstly, maximum simulated likelihood produces consistent estimates of the parameters of the model. However, it is often a mistake to assume that the same distribution applies throughout the entire data set. The adequacy of the model can be evaluated using the consistent misspecification test discussed in this article (which is a straightforward modification of the statistic in Zheng 1996), and if the model is inappropriate for the data being used, researchers can then resort to Local Maximum Simulated Likelihood, which is highly adaptable method and performs very well in a variety of settings.

The results of a Monte Carlo experiment and a empirical application have emphasised that model misspecification is, indeed, a concern, and that the cross-section estimates of efficiency can be severely misleading if the normal-halfnormal model is not a good representation of the model that generated the data. Therefore, is is important to test for misspecification and make use, when possible, of the local maximum simulated likelihood procedure discussed in this article.

Our model can be used outside the realm of efficiency analysis. By design, the PHN is a model for over- or under-reported counts (but not both at the same time). Thus, this distribution can complement the Beta-Binomial or Poisson-Logistic models often used in the social science when it is suspected that the variable under analysis is systematically inflated/deflated.

Further work must address two important questions which are heterogeneity and dynamics. Up to this date, only Greene (2004, 2005) addresses the former issue, but in the context of panel data for continuous output. The use of panel data seems to be crucial to address issues of identification of parameters, and we conjecture that the same will be true in our context. Panel data versions of the PHN would, of course, partially solve the issue of dynamics and within/between cross-correlations.

References

Aigner, D., C. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics 6*, 21–37.

Bhat, C. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B (Methodological) 37*, 837–855.

Caflisch, R., W. Morokoff, and A. Owen (1997). Valuation of mortgage backed securities using brownian bridges to reduce effective dimension. *Journal of Computational Finance 1*, 27–46.

Cameron, C. and P. Johansson (1997). Count data regression using series expansions: with applications. *Journal of Applied Econometrics 12*, 203–224.

Cameron, C. and P. Trivedi (1986). Econometric models based on count data: Comparison and applications of some estimators. *Journal of Applied Econometrics 1*, 29–53.

Cramer, J. (1986). *Econometric applications of Maximum Likelihood methods.* Cambridge University Press.

Delaporte, P. (1962). Sur l'efficacite des criteres de tarification de l'assurance contre les accidents d'automobiles. *ASTIN Bulletin 2*, 84–95.

Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis 93*, 40–57.

DeVuyst, E. and P. Preckel (2007). Gaussian cubature: A practitioner's guide. *Mathematical and Computer Modelling 45*, 787–794.

Doornik, J. A. (2007). *Ox: An Object-Oriented Matrix Programming Language.* Timberlake Consultants Ltd.

Fan, J., M. Farmen, and I. Gijbels (2002). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*, 591–608.

Fan, J., H. Lin, and Y. Zhou (2006). Local partial-likelihood estimation for lifetime data. *Annals of Statistics 34*, 290–325.

Fé, E. (2007). Exploring a stochastic frontier model when the dependent variable is a count. *The School of Economics Discussion Paper Series, University of Manchester 0725.*

Fé, E. (2008). On the production of economic bads and the estimation of production efficiency when the dependent variable is a count. *The School of Economics Discussion Paper Series, University of Manchester 0812.*

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods, 2nd Edition.* Springer.

Goffe, W. (1996). Simann: A global optimization algorithm using simulated annealing. *Studies in Nonlinear Dynamics & Econometrics. 1*, 169176.

Goffe, W., G. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*, 65–99.

Gourieroux, C. and A. Monfort (1995). *Statistics and Econometric Models.* Cambridge University Press.

Gourieroux, C., A. Monfort, and A. Trognon (1984a). Pseudo Maximum Likelihood Methods: Applications to Poisson models. *Econometrica 52*, 701–720.

Gourieroux, C., A. Monfort, and A. Trognon (1984b). Pseudo Maximum Likelihood methods: Theory. *Econometrica 52*, 681–700.

Gozalo, P. and O. Linton (2000). Local nonlinear least squares: Using parametric information in nonparametric regression. *Journal of Econometrics 99*, 63–106.

Greene, W. (2003). Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis 19*, 179–190.

Greene, W. (2004). Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics 13*, 959–980.

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics 126*, 269–303.

Griliches, Z. (1990). Patent statistics and economic indicators: A survey. *Journal of Economic Literature 28*, 1661–1707.

Hall, B., C. Cummins, E. Laderman, and J. Murphy (1986). The R&D master file documentation. *Technical Working Paper, NBER 72.*

Hall, B., Z. Griliches, and J. A. Hausman (1986). Patents and R&D: Is there a lag? *International Economic Review 27*, 265–283.

Hausman, J., B. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents R&D relationship. *Econometrica 52*, 909–938.

Hofler, R. and D. Scrogin (2008). A count data stochastic frontier. *Discussion Paper, University of Central Florida.*

Jondrow, J., I. Materov, K. Lovell, and P. Schmidt (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics 19*, 233–238.

Judd, K. L. (1998). *Numerical Methods in Economics*. The MIT Press.

Karlis, D. and E. Xekalaki (2005). Mixed Poisson distributions. *International Statistical Review 73*, 35–58.

Kumbhakar, S., B. Park, L. Simar, and E. Tsionas (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach. *Journal of Econometrics 137*, 1–27.

Kumbhakar, S. C. and K. Lovell (2003). *Stochastic Frontier Analysis*. Cambridge University Press.

Lancaster, A. (2000). The incidental parameter problem since 1948. *Journal of Econometrics 95*, 391–413.

Loader, C. (1996). Local likelihood density estimation. *Annals of Statistics 24*, 1602–1618.

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

Neyman, J. and E. Scott (1948). Consistent estimation from partially consistent observations. *Econometrica 16*, 1–32.

Pakes, A. (1986). Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica 54*, 755–784.

Pakes, A. and Z. Griliches (1980). Patents and R&D at the firm level: A first report. *Economic Letters 5*, 377–381.

Park, B., L. Simar, and V. Zelenyuk (2008). Local likelihood estimation of truncated regression and its partial derivatives: Theory and application. *Journal of Econometrics 146*, 185–198.

Press, W., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C (The art of Scientific Computing), Second Edition.* Cambridge University Press.

Ruohonen, M. (1989). On a model for the claim number process. *ASTIN Bulletin 18*, 57–68.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Economic Association 82*, 559–567.

Train, K. (2003). *Dicrete Choice Methods with Simulation*. Cambridge University Press.

Wang, P., I. M. Cockburn, and M. L. Puterman (1998). Analysis of patent data: A mixed-poisson-regression-model approach. *Journal of Business and Economic Statistics 16*, 27–41.

Wang, W. and P. Schmidt (2009). On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics 148*, 36–45.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 53*, 1–16.

Winkelmann, R. (2008). *Econometric Analysis of Count Data (Fifth Edition)*. Springer.

Zheng, J. (1996). A consistent test of functional form via non-parametric estimation techniques. *Journal of Econometrics 76*, 263–289.

Appendix A: Moments of the density function $f(\exp(\pm\varepsilon))$.

We now show how to calculate the first and second moments of the transformation $u = f(\exp(\pm\varepsilon))$, where ε follows a half normal distribution. Consider first the case $f(\exp(-\varepsilon))$. Then,

$$\begin{aligned}
\mathcal{E}(u) &= \frac{2}{\sigma\sqrt{2\pi}}\int_0^1 e^{-\log^2(u)/2\sigma^2}du = \frac{2}{\sigma\sqrt{2\pi}}\int_{-\infty}^0 e^{-(\frac{t^2}{2\sigma^2}-t)}dt = e^{\sigma^2/2}\frac{2}{\sqrt{\pi}}\int_{-\infty}^{-\frac{\sigma}{\sqrt{2}}} e^{-s^2}ds \\
&= e^{\sigma^2/2}\mathrm{Erfc}\left(\frac{\sigma}{\sqrt{2}}\right) \in [0,1] \qquad \text{(A-1)}
\end{aligned}$$

where Erfc(.) is the complementary error function[19] and we used the changes of variable $\log(x) = t$, $s = \frac{t}{\sqrt{2}\sigma} - \frac{\sigma}{\sqrt{2}}$ and the fact that $\frac{t^2}{2\sigma^2} - t = \left(\frac{t}{\sigma\sqrt{2}} - \frac{\sigma\sqrt{2}}{2}\right)^2 - \frac{\sigma^2}{2}$. Similar steps show that

$$\begin{aligned}
\mathcal{E}(u^2) &= e^{2\sigma^2}\mathrm{Erfc}\left(\sigma\sqrt{2}\right) \qquad \text{(A-2)} \\
\mathcal{V}(u) &= e^{2\sigma^2}\mathrm{Erfc}\left(\sigma\sqrt{2}\right) - \left\{e^{\sigma^2/2}\mathrm{Erfc}\left(\frac{\sigma}{\sqrt{2}}\right)\right\}^2 \qquad \text{(A-3)}
\end{aligned}$$

For the case $u = \exp(\varepsilon)$ the method is identical, but only the range of integration changes to $[1,\infty)$. Thus,

$$\begin{aligned}
\mathcal{E}(u) &= \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\int_1^\infty e^{-\log^2(u)/2\sigma^2}du = e^{\sigma^2/2}\frac{2}{\pi}\int_{-\frac{\sigma}{\sqrt{2}}}^\infty e^{-s^2}ds \\
&= e^{\sigma^2/2}\left\{1 + \mathrm{Erf}\left(\frac{\sigma}{\sqrt{2}}\right)\right\} \geq 1 \qquad \text{(A-4)} \\
\mathcal{E}(u^2) &= e^{2\sigma^2}\left\{1 + \mathrm{Erf}\left(\sigma\sqrt{2}\right)\right\} \qquad \text{(A-5)}
\end{aligned}$$

[19] Note that, among other relations,

$$\begin{aligned}
Erf(x) &= 2\Phi(x\sqrt{(2)}) - 1 \\
Erfc(x) &= 2\left[1 - \Phi(x\sqrt{(2)})\right]
\end{aligned}$$

where $\Phi(.)$ is the standard normal distribution function.

	Mean Square Errors		
		$\sigma = 0.5$	
	$n = 100$	$n = 400$	$n = 1600$
σ	0.0178	0.0057	0.0037
β_0	0.0349	0.0085	0.0030
β_1	0.0016	0.0004	0.0001
β_2	0.0016	0.0004	0.0001
		$\sigma = 1$	
	$n = 100$	$n = 400$	$n = 1600$
σ	0.0311	0.0133	0.0092
β_0	0.0572	0.0146	0.0058
β_1	0.0030	0.0007	0.0002
β_2	0.0030	0.0007	0.0002
		$\sigma = 2$	
	$n = 100$	$n = 400$	$n = 1600$
σ	0.0847	0.0314	0.0194
β_0	0.1122	0.0294	0.0116
β_1	0.0065	0.0016	0.0004
β_2	0.0066	0.0016	0.0003
		$\sigma = 4$	
	$n = 100$	$n = 400$	$n = 1600$
σ	0.4433	0.1165	0.0483
β_0	0.2410	0.0640	0.0237
β_1	0.0154	0.0033	0.0008
β_2	0.0164	0.0036	0.0008

Table 1: Parametric PHN Model. Mean Square Errors. 2000 Replications. β_i=0.3, for $i = 0, 1, 2$.

		Mean Estimated Inefficiency		
	True		Estimated	
		$n = 100$	$n = 400$	$n = 1600$
$\sigma = 0.5$	0.6992	0.6990	0.6848	0.6800
$\sigma = 1$	0.5232	0.5117	0.5028	0.5004
$\sigma = 2$	0.3362	0.3258	0.3168	0.3148
$\sigma = 4$	0.1888	0.1787	0.1708	0.1686

Table 2: Parametric PHN Model. Mean estimated inefficiency. 2000 Replications.

Patents Data Set (Wang et al. 1998); $N = 70$

	Patents	log R&D	log Sales
Mean	23.4570	1.3119	4.7089
S.D.	38.8230	2.0916	2.0361
Max	173.0000	4.9152	7.8330
Min.	0.0000	−2.9565	0.1527
Quantile 10	0.0000	−1.7880	1.5864
Quantile 25	1.0000	−0.1753	3.6753
Quantile 50	4.0000	1.3106	4.8651
Quantile 75	25.7500	2.9811	6.5212
Quantile 90	71.1000	4.0113	7.2076
Skewness Patents		2.1507	
Kurtosis Patents		7.2684	
Skewness log Patents		0.2708	
Kurtosis log Patents		1.7699	

Table 3: Descriptive Statistics. log Patents is based on 57 non-zero observations.

Parametric Models

	ALS ($N = 57$)	PHN ($N = 70$)
θ	0.72147	-
	(0.067577)	-
λ	3.22e-5	-
	(57.003)	-
σ	-	0.1596
	-	(0.17354)
Intercept	1.3144	2.7342
	(57.003)	(0.88869)
R&D / Sales	0.23034	-4.6948
	(0.29439)	(3.7552)
log(R&D)	0.82520	1.2074
	(0.11535)	(0.19456)
log(Sales)	-0.11252	-0.33579
	(0.13071)	(0.21197)
Mean Inefficiency	1.8559e-005	0.47453

Table 4: Results from the normal-half normal model (Aigner et al. 1977) and the Poisson log-half normal model. Standard errors in parentheses.

Distribution of Inefficiency						
5%	10%	25%	50%	75%	90 %	95%
0.10203	0.21137	0.32597	0.56231	0.76739	0.95002	0.99996

Table 5: Estimated distribution of inefficiency. PHN estimated via Local Maximum Likelihood.

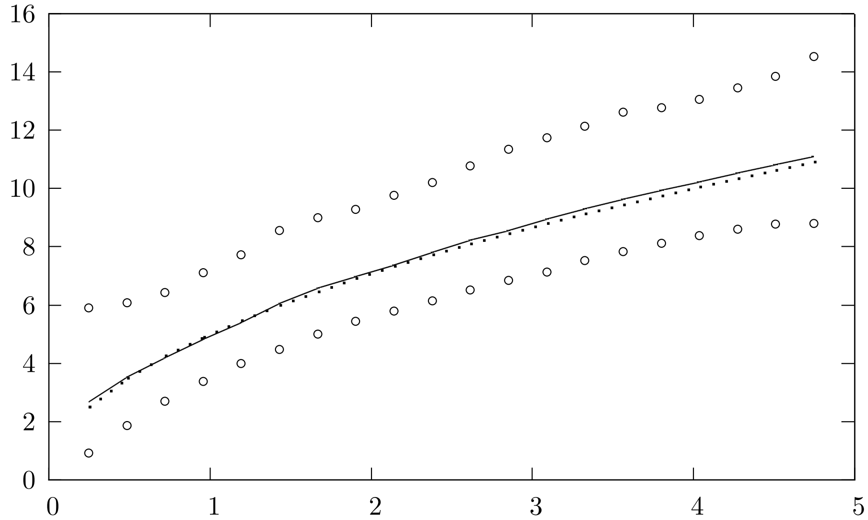

Figure 1: Monte Carlo Simulation. True $\lambda = 0.5x^{0.5}$. LML estimation over a grid of 20 nodes, using Gaussian Kernel and Simulated Annealing. True conditional mean (dotted line), LML estimator (continuous line) and 5 and 95 quantile lines. $N = 100$, $R = 1000$, $h = 0.5\sigma_x n^{-1/5}$.

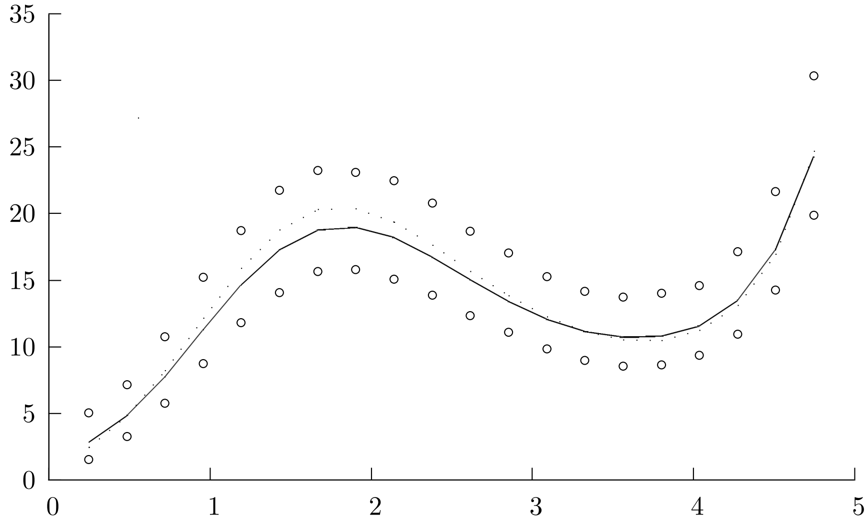

Figure 2: Monte Carlo Simulation. True $\lambda = \exp(4x - 1.65x^2 + 0.2x^3)$. LML estimation over a grid of 20 nodes, using Gaussian Kernel and Simulated Annealing. True conditional mean (dotted line), LML estimator (continuous line) and 5 and 95 quantile lines. $N = 100$, $R = 1000$, $h = 0.5\sigma_x n^{-1/5}$.

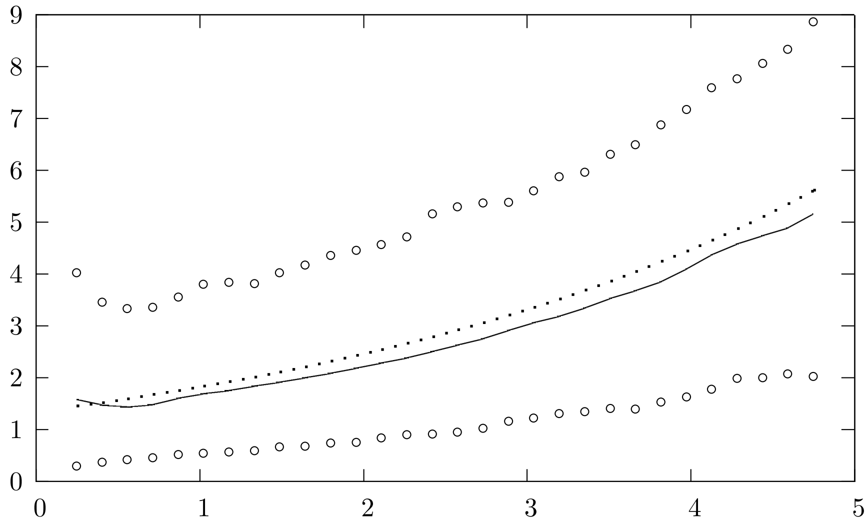

Figure 3: Monte Carlo Simulation. True $\lambda = \exp(0.3 + 0.3x + |t|)$, where $t \sim$ Student with 3 degrees of freedom. LML estimation over a grid of 20 nodes, using Gaussian Kernel and Simulated Annealing. True conditional mean (dotted line), LML estimator (continuous line) and 5 and 95 quantile lines. $N = 100$, $R = 1000$, $h = 0.5\sigma_x n^{-1/5}$.

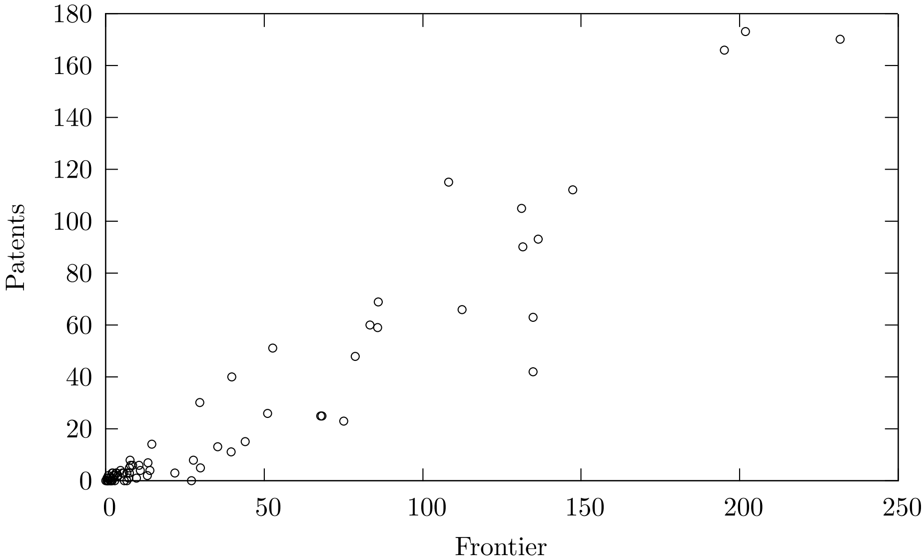

Figure 4: Patent data. Number of patents vs estimated frontier

Simulated Skewness.

		λ					
		1	2	4	8	16	32
	κ_3	1.369	0.66194	0.076942	-0.4487	-0.88223	-1.0769
σ^2 =1	$\min Y$	0	0	0	0	0	0
	$\max Y$	7	9	13	20	31	51
	% data	38.856	60.405	80.663	93.119	98.281	99.674
	κ_3	1.4648	0.79521	0.24615	-0.21696	-0.58242	-0.84267
σ^2 =2	$\min Y$	0	0	0	0	0	0
	$\max Y$	7	8	14	20	30	53
	% data	32.707	51.983	71.131	85.771	94.016	97.674
	κ_3	1.5393	0.86984	0.32644	-0.096093	-0.41775	-0.64045
σ^2 =3	$\min Y$	0	0	0	0	0	0
	$\max Y$	6	9	14	20	31	54
	% data	28.501	45.997	64.774	79.226	89.451	95.133
	κ_3	1.5834	0.9303	0.37888	-0.026182	-0.31567	-0.52977
σ^2 =4	$\min Y$	0	0	0	0	0	0
	$\max Y$	6	9	11	19	33	51
	% data	25.977	42.037	59.75	74.259	85.014	91.977
	κ_3	1.5731	0.93933	0.42126	0.042143	-0.23966	-0.44299
σ^2 =5	$\min Y$	0	0	0	0	0	0
	$\max Y$	6	10	12	21	31	49
	% data	24.118	38.929	55.495	70.045	81.339	89.113

Table 6: Simulated skewness of $\log(Y)$, where Y $\sim$Poisson$(\lambda\theta)$, with $\theta = \exp(-\sigma|u|)$. $N = 100,000$.