WPS6962

Policy Research Working Paper 6962

Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality

Chris Elbers
Roy van der Weide

WORLD BANK GROUP
Development Research Group
Poverty and Inequality Team
July 2014

Policy Research Working Paper 6962

Abstract

This paper proposes a method for estimating distribution functions that are associated with the nested errors in linear mixed models. The estimator incorporates Empirical Bayes prediction while making minimal assumptions about the shape of the error distributions. The application presented in this paper is the small area estimation of poverty and inequality, although this denotes by no means the only application. Monte-Carlo simulations show that estimates of poverty and inequality can be severely biased when the non-normality of the errors is ignored. The bias can be as high as 2 to 3 percent on a poverty rate of 20 to 30 percent. Most of this bias is resolved when using the proposed estimator. The approach is applicable to both survey-to-census and survey-to-survey prediction.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at rvanderweide@worldbank.org.

Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality

Chris Elbers and Roy van der Weide[1]

Keywords: Normal mixtures, linear mixed models, small area estimation, Empirical Bayes, poverty, inequality
JEL Classification: I32, C31, C42

[1] Chris Elbers (c.t.m.elbers@vu.nl) is at VU University Amsterdam and the Tinbergen Institute. Roy van der Weide (rvanderweide@worldbank.org) is at the World Bank. We gratefully acknowledge financial support from VU University Amsterdam and the World Bank's Knowledge for Change Program (KCP II). A thank you also goes to Peter Lanjouw for providing comments on earlier versions of this paper.

1 Introduction

We propose a method for estimating distribution functions that are associated with the nested errors in linear mixed models. The proposed estimator is accommodating Empirical Bayes prediction and makes minimal assumptions about the shape of the error distributions. Our objective is to accurately predict nonlinear functions of the dependent variable, where the shape of the error distribution functions potentially plays an important role. The application we have in mind is the small area estimation of poverty and inequality, although this denotes by no means the only application. This particular application has been made popular by the work of Elbers, Lanjouw and Lanjouw (2003; henceforward ELL). The approach they put forward has since been applied to obtain maps of poverty and inequality in over 60 countries worldwide. The highly disaggregated estimates of poverty and inequality have in turn inspired a range of other applications, for example: Demombynes and Ozler (2005) investigate whether inequality at the small area level has an impact on local crime rates, and find that it does using data from South Africa; Elbers et al. (2007) conduct an empirical experiment in order to estimate by how much one could potentially lower the costs of getting resources to the poor if one had access to a poverty map, and conclude that the gains can be substantial; Araujo et al. (2008) examine whether villages with higher levels of inequality are less likely to invest in public goods that would benefit the poor. Using data from Ecuador they find empirical support for this hypothesis which they attribute to elite capture. Recently, Fujii (2010) modified the approach to make it better suited for the small area estimation of child malnutrition outcomes with an application to Cambodia.

1.1 Problem statement

Consider the following standard linear mixed model for log income of household h residing in area a:

$$y_{ah} = x_{ah}\beta + u_a + \varepsilon_{ah}.$$

Let y_d denote a vector of log incomes for all households from domain d (where "domain" typically refers to a small geographic area). We are interested in estimating $W(y_d)$, where W is some (possibly nonlinear) function of y_d. It is assumed that we have data on x for the entire population, while data on both x and y is available for a sample S of households only.

ELL propose to estimate W by:

$$E[W(y_d)|x \in d; (x, y) \in S],$$

where the expectation is taken over the unobserved errors u_a and ε_{ah}. Because W is nonlinear, the shape of the error distributions will matter for the expected value of W. Estimation of β is obviously of primary importance when estimating $E[W]$, but in this paper we will concentrate on the error distributions and assume – except for the application in section 6 – that β is known. We argue that even if β is estimated perfectly, getting the error distribution wrong still has the

potential to introduce a significant bias.

ELL felt their approach would be most convincing if they make minimal assumptions about the errors u_a and ε_{ah}. They proceed by first obtaining estimates of the area errors u_a and ε_{ah}, which then allows them to sample from the empirical errors $\hat{u}_a$ and $\hat{\varepsilon}_{ah}$. u_a is estimated as the simple area average of the total residuals (appropriately re-scaled so that the sample variance of $\hat{u}_a$ equals the estimate of σ_u^2 that corrects for the contribution of the area average of ε_{ah}). The estimate for ε_{ah} is obtained by subtracting $\hat{u}_a$ from the total residual (re-scaled so that its sample variance matches σ_ε^2).[1]

This non-parametric approach to estimating the error distributions has two limitations. Firstly, this procedure does not adequately account for the fact that $\hat{u}_a$ equals a sum of u_a and $\bar{\varepsilon}_a$. This latter term is often large enough for it to affect the ability of the empirical distribution of $\hat{u}_a$ to reproduce the shape of the actual distribution of u_a. Simply ignoring this "contamination" may result in a significant bias. If the empirical distribution of u_a is biased, then this bias will also have implications for the empirical distribution of ε_{ah}. This is also referred to as a "convolution problem", where the objective is to estimate the distribution function of a random variable that is observed with error.

Secondly, the approach adopted by ELL does not easily lend itself for Empirical Bayes (EB) estimation where the distribution of u_a is tightened by conditioning on household data y and x available for domain d (i.e. in the event that some households in domain d are included in the sample). Working out the conditional distribution is not a trivial exercise without making further distributional assumptions. Consequently ELL decided to forego EB estimation altogether. In doing so, they have accepted a certain loss in efficiency by not fully utilizing all available information. Molina and Rao (2010) recently picked up on this and put forward an alternative approach that does implement EB estimation. They take ELL as a point of departure but then assume that both u_a and ε_{ah} are normal distributed, in which case the conditional distribution too will be normal distributed. Where ELL accept a loss in precision by not implementing EB estimation, Molina and Rao (2010) accept a loss in precision that might stem from a misspecification of the error distribution functions. The data at hand will ultimately determine which of the two will be accepting the larger loss. ELL are arguably most interested in estimating poverty and inequality in developing countries where the number of small areas (or domains) that are covered by the income surveys are often small, think of 5 to 25 percent of all domains in the population. In this case the benefits of EB estimation will be modest (as survey data are available for only few areas). However, in more developed countries, or countries where travel costs that are incurred when covering all small areas are manageable, income surveys often cover a much larger number of the domains. In fact, there are numerous examples where surveys cover between 50 and 100 percent of all domains in the country. In those instances, there may be clear benefits to adopting EB estimation.

The approach presented in this paper improves on both ELL and Molina and Rao (2010). Like ELL we make no restrictive assumptions about the error distributions. Our estimator

[1] Note that ELL allows for heteroskedasticity, so that σ_ε^2 can be household-specific.

for the distribution functions will generally be more accurate than the estimator adopted by ELL however, as we explicitly account for the nested error structure (that is responsible for the "convolution problem"). Unlike ELL we also accommodate EB estimation. We achieve this by fitting finite normal mixtures (NM) to the error distribution functions. Normal mixtures are extremely flexible; they are able to fit any well-behaved distribution function, and are ideally suited for accommodating EB estimation. If the marginal distributions of u_a and ε_{ah} can be described by normal mixtures, then the conditional distribution too can be described by a normal mixture with known parameters (that are functions of these parameters and of the data on which is being conditioned). Estimation of the normal mixtures for u_a and ε_{ah} is complicated by the fact that neither u_a nor ε_{ah} are observed. Our estimator for the NM parameters may be viewed as a modified version of the EM algorithm.

Monte Carlo simulations indicate that estimates of poverty and inequality can be severely biased when ignoring the non-normality of the errors. The bias can be as high as 2 to 3 percent on a poverty rate of 20 to 30 percent. Most of this bias is resolved when implementing our estimator. This is confirmed by an empirical application to US data.

1.2 Normal mixtures in nested-error models

There are a number of other studies that have explored different ways of relaxing the normality assumption in mixed linear models. Verbeke and Lesaffre (1996) is an early example that also considers normal mixtures as a "non-parametric" representation of the error distribution function. However, they impose a number of important restrictions. First, only the area random effects u_a are allowed to be non-normal; a normal-mixture is fitted to the distribution of u_a under the assumption that ε_{ah} is normally distributed. Second, it is assumed that the "component distributions" that make up the normal mixture share a common variance, which noticeably simplifies estimation but at the same time significantly limits the flexibility of the normal mixture to fit any given distribution function. This approach has also been followed by Cordy and Thomas (1997) who work with the same setup and adopt the same set of restrictions. There is another strand of the literature that permits both error terms to be non-normal distributed by imposing an alternative parametric family for the distribution functions. See for example Zhou and He (2008) who fit skewed t-distributions to the nested errors of the linear mixed model. Recently, there have also been efforts to explore the impact of misspecifications in the error distributions for Empirical Bayes predictions derived from linear mixed models, see for example Skrondal and Rabe-Hesketh (2009) and McCulloch and Neuhaus (2011). They conclude that the bias is reasonably small. It should be noted however that those studies focus on prediction of the dependent variable itself, in which case the Empirical Bayes estimates of the area random effects u_a denote the only source of bias. Misspecifications in the error distributions become considerably more important when predicting nonlinear functions of the dependent variable, such as measures of poverty and inequality, as we will show in this paper.

The remainder of this paper is organized as follows. In Section 2 we briefly discuss how the error distribution functions associated with the errors in a linear mixed model will matter for

prediction, with an application to poverty and inequality measurement. In this section we will also introduce normal mixtures as a flexible "non-parametric" representation of any given well-behaved distribution function, and demonstrate the implications for EB estimation. Estimation of the normal mixture distributions to both errors from the linear mixed model is presented in Section 3. A modest Monte Carlo simulation study followed by an equally modest empirical application is provided in Sections 4 and 5, respectively. Finally, Section 6 concludes.

2 Estimation of poverty and inequality: Distributions matter

2.1 Linear mixed model for income

Suppose that at the household level the data generating process (DGP) satisfies the equation already mentioned above:

$$y_{ah} = x_{ah}^T \beta + u_a + \varepsilon_{ah}, \tag{1}$$

where x_{ah} denotes a vector with independent variables, and where u_a and ε_{ah} denote zero expectation error terms that are independent of each other. The subscripts indicate target area (or domain) a and household h.[2] In this paper we assume that errors are homoskedastic, so that for each household h and area a we have: $var[y_{ah}|x_{ah}] = \sigma_u^2 + \sigma_\varepsilon^2$. Throughout the paper it is assumed that consistent estimators for the variance parameters are available, which we shall denote by $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$.[3] We will not make any assumptions about the shape of the error distributions.

Let A denote the total number of areas covered by the income survey and let n_a denote the number of households that have been sampled in area a, so that $n = \sum_{a=1}^{A} n_a$ denotes the total sample size of the survey. We shall denote the total household error by: $e_{ah} = y_{ah} - x_{ah}^T \beta$, and its area a average by $\bar{e}_a = \bar{y}_a - \bar{x}_a^T \beta$ where $\bar{e}_a = \sum_h e_{ah}/n_a$. We shall also use the notation $e_a = (e_{a,1}, \ldots, e_{a,n_a})$ which denotes the vector of length n_a with residuals for all households from area a. With a slight abuse of terminology we will at times refer to the errors e_{ah} and $\bar{e}_a$ as data (as if we know the parameter vector β).

Let the probability distribution functions for u_a and ε_{ah} be denoted by F_u and G_ε. We will propose a computationally attractive method for estimating these distribution functions, where we make no restrictive assumptions concerning their functional form, in particular allowing these functions to be other than normal distribution functions. Another appealing feature of our non-parametric estimator for the error distribution functions is that it can easily accommodate Empirical Bayes estimation.

[2] These areas may refer to geographic areas such as districts or municipalities, but also to non-geographic domains such as ethnic groups or age groups, say.

[3] A commonly used estimator for the variance parameters from a nested error model is Henderson's method III estimator (see Henderson, 1953; and Searle et al., 1992), which may be viewed as a method of moments estimator which does not require any assumptions about the shape of the error distributions. Alternative estimators are restricted maximum likelihood, minimum norm quadratic unbiased estimation (MINQUE; see e.g. Westfall, 1987; Searle et al., 1992), and so-called spectral decomposition estimation (see e.g. Wang and Yin, 2002; and Wu et al., 2009).

2.2 Empirical Bayes estimation

Empirical Bayes (EB) estimation, also known as Empirical Best estimation, tightens the error distributions by conditioning on all available data in the survey for the purpose of prediction. The 'observation' e_a (provided that area a is covered by the survey) clearly carries information about the area random effect u_a. In the extreme, for example, where n_a tends to infinity $\bar{e}_a$ will perfectly reveal u_a.

In effect EB estimates are obtained by integrating out area errors u_a using probability density $p(u_a|e_a)$, i.e. the probability density of u_a conditional on e_a observed from the survey. (Non-EB estimates are obtained by using the unconditional density $p(u_a)$.) The challenge is to work out $p(u_a|e_a)$ along with the marginals $p(u_a)$ and $p(\varepsilon_{ah})$ without imposing restrictions about their form so that the resulting conditional density $p(u_a|e_a)$ can also take on any form. Currently, the literature on EB estimation avoids this challenge by assuming normally distributed marginals, in which case the conditional distribution will be normal too.

For normally distributed errors it can be shown that $p(u_a|e_a)$ can be written as a function of $\bar{e}_a$, the mean of sample errors from domain a. This is not true for general error distributions. However, conditioning on the full vector e_a becomes computationally intractable. Therefore we will condition on $\bar{e}_a$ rather than the full vector e_a even in the case of non-normal errors.

A second simplification concerns the known regression residuals for sample households. When conditioning on e_a, we ignore that for these households the total error $e_{ah} = u_a + \varepsilon_{ah}$ is known (since it is one of the components of e_a). Instead we will assume that ε_{ah} is independent of e_a even for sample households. The reason is that in practice sample household h cannot be traced among households in the target domain a (households for which only data on x is available). In most practical settings, the error thus introduced will be negligible. In fact, this error tends to zero as the size of the sample relative to the population size tends to zero.

2.3 Distributions matter

Let $y_{(a)}$ and $e_{(a)}$ denote vectors of length N_a with elements y_{ah} and e_{ah} for all households from the population. Similarly, $x_{(a)}$ will denote a matrix with rows given by x_{ah}^T for all N_a households. y_a, e_a and x_a will denote the survey sample analogues. As mentioned above our objective is to estimate:

$$E[W(y_{(a)})|x_{(a)}, y_a] = \int W(x_{(a)}\beta + e)p(e|e_a)de \quad (2)$$

$$\simeq \int\int W(x_{(a)}\beta + u + \varepsilon)p(\varepsilon)p(u|e_a)d\varepsilon du, \quad (3)$$

where the function W will generally be non-linear.

Non-normality of the errors u and ε will affect the estimates of $E[W]$ via two different channels. Firstly, when the function W is indeed non-linear, the expected value $E[W]$ will be a function of the higher moments of the distributions of u and ε. Getting these moments wrong, in other words getting the distributions wrong, is then likely to introduce bias. Secondly, if

the distributions of u and ε are wrong then the conditional density $p(u|e_a)$, which concerns EB estimation, will also be wrong. This will affect all moments of the conditional distribution of u, including the first, and hence has the potential to introduce bias in the estimate of $E[W]$ even if W is linear. (Note that in the case of linear W, non-EB estimates of $E[W]$ are unbiased even if the distributions for u and ε are wrong.) Recently, Skrondal and Rabe-Hesketh (2009) and McCulloch and Neuhaus (2011) have explored the magnitude of the bias that is introduced by getting the first moment of $p(u|e_a)$ wrong, in the case of linear W, and found it to be modest. Our estimator would therefore be most relevant for the case of nonlinear W.

2.4 Empirical Bayes estimation with normal mixtures

Let us assume that F_u and G_ε can be represented by mixture distributions:

$$F_u = \sum_{i=1}^{i=m_u} \pi_i F_i \tag{4}$$

$$G_\varepsilon = \sum_{j=1}^{j=m_\varepsilon} \lambda_j G_j, \tag{5}$$

where the F_i's and G_j's denote a basis of distribution functions which we will also refer to as components or component distribution functions. The π_i's and λ_j's denote unknown non-negative probabilities that satisfy $\sum_i \pi_i = 1$ and $\sum_j \lambda_j = 1$, which we will also refer to as mixing probabilities. m_u and m_ε denote the number of components used to represent F_u and G_ε, respectively. We will denote the probability density functions associated with F_i and G_j by respectively f_i and g_j.

Mixture distributions are remarkably well equipped to fit any well-behaved distribution function. For example, kernel density estimators are closely related to mixture distributions. We will be working with normal component distributions, so that the mixture distributions are normal mixtures.

Assumption 1 *The components F_i are normal distribution functions with mean μ_i and variance σ_i^2. Similarly, components G_j are normal distribution functions with mean ν_j and variance ω_j^2.*

Note that the modeler is at liberty to work with a different basis of component distributions. This choice does not have real implications for the ability of the mixture to fit a given distribution function.

If $p(u_a)$ and $p(\varepsilon_{ah})$ are normal-mixtures, then $p(u_a|\bar{e}_a)$ is a normal mixture too. This is a powerful result as the integral in equation 3 will generally have to be computed by simulation, and sampling from normal mixtures is straightforward. Lemma 2 shows how the parameters that define $p(u_a|\bar{e}_a)$ can be obtained as a function of the parameters of the normal-mixtures $p(u_a)$ and $p(\varepsilon_{ah})$. Implementing EB estimation is thus as easy as sampling the area errors from the normal-mixture $p(u_a|\bar{e}_a)$ whenever $\bar{e}_a$ is observed in the survey sample.

Lemma 2 *The probability density function of u_a conditional on $\bar{e}_a$, which we denote by $p(u_a|\bar{e}_a)$, is a normal-mixture with known parameters:*

$$p(u_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \sum_{j=1}^{j=m_\varepsilon} \sum_{k=1}^{k=m_\varepsilon} w_{ijk} \varphi\left(u_a; m_{ijk}; s_{ijk}^2\right), \tag{6}$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\sigma_i^2}{\sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a}\right)(\bar{e}_a - (\nu_j + \nu_k)/n_a) + \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right)\mu_i \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2)\sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \tag{7}$$

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε, respectively.

If we apply the Central Limit Theorem to approximate the marginal distribution of $\bar{\varepsilon}_a$ to a normal distribution with mean zero and variance equal to σ_ε^2/n_a, the expression for $p(u_a|\bar{e}_a)$ simplifies considerably:

$$p(u_a|\bar{e}_a) \approx \sum_{i=1}^{i=m_u} \alpha_i \varphi\left(u_a; \gamma_{ai}\bar{e}_a + (1-\gamma_{ai})\mu_i; \left(1/\sigma_i^2 + n_a/\sigma_\varepsilon^2\right)^{-1}\right), \tag{8}$$

with $\gamma_{ai} = \sigma_i^2/(\sigma_i^2 + \sigma_\varepsilon^2/n_a)$, and where $\alpha_i = \tilde{\alpha}_i / \sum_i \tilde{\alpha}_i$ with:

$$\tilde{\alpha}_i = \pi_i \varphi(\bar{e}_a; \mu_i; \sigma_i^2 + \sigma_\varepsilon^2/n_a). \tag{9}$$

The expected value of u_a conditional on $\bar{e}_a$, given the density function $p(u_a|\bar{e}_a)$ from Lemma 2, is seen to solve:

$$E[u_a|\bar{e}_a] \approx \sum_i \alpha_i(\bar{e}_a)\left(\gamma_{ai}\bar{e}_a + (1-\gamma_{ai})\mu_i\right), \tag{10}$$

where $\gamma_{ai} = \sigma_i^2/(\sigma_i^2 + \sigma_\varepsilon^2/n_a)$, and where $\alpha_i(\bar{e}_a)$ denotes the mixing probabilities of $p(u_a|\bar{e}_a)$.

Note that the standard assumption of normal errors is nested as a special case, where there is just one component with $\mu_i = 0$ and $\sigma_i^2 = \sigma_u^2$. The first and second moment of the normal conditional density $p(u_a|\bar{e}_a)$ in this case are seen to solve:

$$E[u_a|\bar{e}_a] = \gamma_a \bar{e}_a \tag{11}$$

$$var[u_a|\bar{e}_a] = (1-\gamma_a)\sigma_u^2, \tag{12}$$

where $\gamma_a = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2/n_a)$ (see e.g. Molina and Rao, 2010).[4] For non-normal errors, we have that $E[u_a|\bar{e}_a]$ is generally a non-linear function of $\bar{e}_a$, and $var[u_a|\bar{e}_a]$ will generally be a function of $\bar{e}_a$.

2.5 Application to small area estimation of poverty and inequality

For our application let y_{ah} measure per capita income (or expenditure) for household h residing in area a, and let s_{ah} denote the number of household members for that same household. In vector notation, let $y_{(a)}$ and $s_{(a)}$ be vectors with elements y_{ah} and s_{ah} for all households from area a. The objective is to determine the level of welfare for area a which can be expressed as a function of $y_{(a)}$ and $s_{(a)}$: $W(y_{(a)}, s_{(a)})$. The welfare function W is typically non-linear. Popular examples are the share of individuals whose income falls below a pre-specified poverty line (also known as the head-count poverty rate), or the Gini index of income inequality.

Collecting data on income (or expenditure) y_{ah} for any given household is generally found to be expensive relative to collecting data on demographics, education, employment status, and housing. This is particularly true for developing countries where much of the income does not come from wage employment. Consequently, income data is often only available in the form of so-called income surveys. The sample size of these surveys is sufficient to estimate national and possibly sub-regional welfare, but too small to estimate welfare directly at the level of much smaller areas (i.e. the target areas a).

Elbers et al. (2003; henceforward ELL) advocate an approach that combines the income survey with unit record population census data. The census has data on the independent variables x_{ah} from equation (1), such as demographics, education, employment and housing, but not the household income variable y_{ah}. Crucially, the data on x_{ah} are also collected by the income survey. The idea is to use the income survey to estimate the parameters from equation (1), and then use the model to predict income for every household in the census. With these predicted incomes we can subsequently estimate welfare W for each target area a.

Standard errors can be obtained by means of simulation which is ideally suited for estimating quantities that are non-linear functions of the random variables at hand, as is the case with measures of poverty and inequality. Let R denote the number of simulations. The estimator then takes the form:

$$\hat{\mu} = \frac{1}{R}\sum_{r=1}^{R} W\left(\tilde{y}_{(a)}^{(r)}, s_{(a)}\right), \tag{13}$$

where $\tilde{y}_{(a)}^{(r)}$ denotes the r-th simulated (or predicted) income vector with elements $\tilde{y}_{ah}^{(r)} = x_{ah}^T\tilde{\beta}^{(r)} + \tilde{u}_a^{(r)} + \tilde{\varepsilon}_{ah}^{(r)}$. With each simulation, both the model parameters $\tilde{\beta}^{(r)}$ and the errors $\tilde{u}_a^{(r)}$ and $\tilde{\varepsilon}_{ah}^{(r)}$ are drawn from their estimated distributions.[5] In the end this gives R simulated poverty rates. The

[4] Note that the unconditional variance solves: $var[u_a] = var[\gamma_a\bar{e}_a] + (1-\gamma_a)\sigma_u^2 = \gamma_a^2(\sigma_u^2 + \sigma_\varepsilon^2/n_a) + (1-\gamma_a)\sigma_u^2$, which equals $var[u_a] = \gamma_a\sigma_u^2 + (1-\gamma_a)\sigma_u^2 = \sigma_u^2$, since $\sigma_u^2 + \sigma_\varepsilon^2/n_a = \sigma_u^2/\gamma_a$.

[5] Our preferred method is to draw $\tilde{\beta}^{(r)}$ by re-estimating the model parameters using the r-th bootstrap version of the survey sample. Alternatively, $\tilde{\beta}^{(r)}$ may be drawn from its estimated asymptotic distribution. The difference between these two alternatives is expected to be modest, unless the survey sample is particularly small so that finite sample effects may play a role.

point estimates and their corresponding standard errors are obtained by computing respectively the average and the standard deviation over these simulated values.

It should be noted that ELL draw the area error $\tilde{u}_a$ from the estimated unconditional distribution, which is estimated non-parametrically. (The distribution for $\tilde{\varepsilon}_{ah}$ too is estimated non-parametrically.) The advantage of this approach is that it is fully flexible in that it does not restrict the shape of the error distributions. A possible shortcoming is that it does not take full advantage of all the available data. Ideally one would want to draw $\tilde{u}_a$ from a distribution that is conditioned on all relevant data that has been sampled from area a.

Molina and Rao (2010; henceforward MR) do exactly that, they closely follow ELL but then draw the area error from the conditional distribution. This is referred to as Empirical Bayes (EB) estimation. Importantly, MR also differ from ELL in that they restrict the errors to be normally distributed. As we have seen, under this assumption the conditional distribution of $\tilde{u}_a$ is normal too and its parameter can be easily determined.[6] When errors are non-normal, it is not obvious what form the conditional distribution for $\tilde{u}_a$ will take; it will generally be of a different form than the unconditional distribution.

As is noted in section 2.3, getting the error distributions right is not merely a matter of efficiency. When the welfare function W is a non-linear function of the error terms, using wrong error distributions will also introduce a bias in the welfare estimates. Whether the magnitude of this bias due to misspecification is important in practice is an empirical question.

The choice between non-normal errors combined with non-EB estimation (which is more flexible, but does not fully utilize all available data) or normal errors combined with EB estimation (which is less flexible, but fully utilizes the available data) may be determined/motivated by: (a) the degree of non-normality found in the data, and (b) how much information one stands to ignore/lose. The latter depends largely on: (i) how many areas have been sampled by the income survey (as for areas not represented in the survey EB and non-EB estimation are equivalent), and (ii) the size of the area error relative to the total error.

The approach developed in this paper aims to combine the best of both worlds; we adopt EB estimation while permitting non-normal distribution functions. The non-parametric estimator for the distribution functions used by ELL does not lend itself for EB estimation. We propose a non-parametric estimator that does.

Remark 3 *Note that an added advantage of representing the non-normal error distributions by normal-mixtures is that* $E[w(y_{ah})]$ *can be evaluated as the sum of expectations over normal errors, i.e.* $E[w(y_{ah})] = E[w(x_{ah}^T\beta + u_a + \varepsilon_{ah})] = \sum_{ij} \kappa_{ij} E[w(x_{ah}^T\beta + u_a^{(i)} + \varepsilon_{ah}^{(j)})]$*, where* $u_a^{(i)}$ *and* $\varepsilon_{ah}^{(j)}$ *are normal distributed with means* μ_i *and* ν_j*, and variances* σ_i^2 *and* ω_j^2*. This means that evaluating* $E[W(y_{ah})]$ *analytically under normal-mixture errors is no more difficult than under normal errors.*

[6] This is arguably why errors are generally assumed normal in the literature on EB estimation.

3 Estimation of normal mixtures when errors are nested

3.1 Estimation of unrestricted normal mixtures

The objective is to estimate the parameters $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$, which determine the normal mixtures (NM) $F_u(u) = \sum_{i=1}^{i=m_u} \pi_i F_i(u; \mu_i, \sigma_i^2)$ and $G_\varepsilon(\varepsilon) = \sum_{j=1}^{i=m_\varepsilon} \lambda_j G_j(\varepsilon; \nu_j, \omega_j^2)$, respectively, where m_u and m_ε denote the number of components.

Estimation of NM to observed data is a routine task, mostly using the EM algorithm, see e.g. Dempster et al. (1977) and McLachlan and Peel, (2000). Suppose that some data y is drawn from a NM with m components and parameter vector θ. The algorithm introduces a latent random variable z_{ij} that equals 1 if observation y_j has been drawn from component i, and 0 otheriwse. Let the log-likelihood function that treats both y and z as data be denoted by $L(\theta; y, z)$, and its derivative with respect to θ by $L_\theta(\theta; y, z)$. The corresponding maximum-likelihood (ML) estimator solves the following moment condition:

$$E[L_\theta(\theta; y, z)] = E[E[L_\theta(\theta; y, z)|y]] = 0. \tag{14}$$

The EM algorithm essentially solves the same moment condition but iteratively:

$$E[E[L_\theta(\theta; y, z)|y, \hat{\theta}^{(k)}]] = 0, \tag{15}$$

where $\hat{\theta}^{(k)}$ denotes the iteration-k estimate for θ. Solving equation (15) yields $\hat{\theta}^{(k+1)}$. The procedure results in a non-decreasing sequence of likelihoods $L(y; \hat{\theta}^k)$. The advantage of the EM algorithm is that solving equation (15) is often considerably easier than solving equation (14). Note that the unobserved data z is integrated out conditional on the observed data y and the current estimate of θ.

The nested error structure however poses a challenge that prevents a straightforward application of the EM algorithm. Conventionally, mixture distributions are estimated to data that are observed directly. We wish to estimate the distributions for u_a and ε_{ah} but we do not observe either of them. In some sense u_a and ε_{ah} are both observed with error (even if the total error e_{ah} is known with certainty), which makes this a convolution problem.

We propose a modification of the EM algorithm that is able to deal with this convolution problem. In a nutshell, we treat the measurement error as a latent variable and integrate it out the same way as the latent variable z is integrated out by the EM algorithm. When estimating F_u (the NM for u_a), the observable data includes $\bar{e}_a = u_a + \bar{\varepsilon}_a$, where $\bar{\varepsilon}_a = \sum_h \varepsilon_{ah}/n_a$ will act as measurement error. The modified EM algorithm will then integrate out $\bar{\varepsilon}_a$ along with z_a. When estimating G_ε (the NM for ε_{ah}), the observable data includes $e_{ah} = u_a + \varepsilon_{ah}$. Now u_a takes on the role of measurement error and will hence be integrated out jointly with z_a. Note that we would need some initial estimate of G_ε when estimating F_u this way (as we have to integrate out $\bar{\varepsilon}$). Similarly, we need some initial estimate of F_u when estimating G_ε (to integrate out u). This by itself is very much in the spirit of the EM algorithm which requires an initial estimate of θ to determine to distribution of z conditional on the observed data.

Our modified EM algorithm solves the following moment condition when estimating F_u:

$$\sum_a E[L_{\theta_u}(\theta_u; \bar{e}_a, \bar{\varepsilon}_a, z_a)|\bar{e}_a, \hat{\theta}_u^{(k)}] = 0, \tag{16}$$

which is equivalent to:

$$\sum_a E[E[L_{\theta_u}(\theta_u; \bar{e}_a, \bar{\varepsilon}_a, z_a)|\bar{e}_a, \bar{\varepsilon}_a, \hat{\theta}_u^{(k)}]|\bar{e}_a] = 0. \tag{17}$$

Similarly, it solves the following moment condition when estimating G_ε:

$$\sum_a E[L_{\theta_\varepsilon}(\theta_\varepsilon; e_a, u_a, z_a)|e_a, \hat{\theta}_\varepsilon^{(k)}] = 0, \tag{18}$$

which is equivalent to:

$$\sum_a E[E[L_{\theta_\varepsilon}(\theta_\varepsilon; e_a, u_a, z_a)|e_a, u_a, \hat{\theta}_\varepsilon^{(k)}]|e_a] = 0, \tag{19}$$

where $e_a = (e_{a1}, \ldots, e_{a,n_a})$.

Specifically, in order to integrate out $\bar{\varepsilon}_a$ conditional on the observed $\bar{e}_a$ (see equation 17), we need an initial estimate of the probability distribution function for $\bar{\varepsilon}_a|\bar{e}_a$. It can be verified that $p(\bar{\varepsilon}_a|\bar{e}_a)$ is again a normal-mixture distribution.

Lemma 4 *The probability density function of $\bar{\varepsilon}_a$ conditional on $\bar{e}_a$, which we denote by $p(\bar{\varepsilon}_a|\bar{e}_a)$, is a normal-mixture with known parameters:*

$$p(\bar{\varepsilon}_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \sum_{j=1}^{j=m_\varepsilon} \sum_{k=1}^{k=m_\varepsilon} w_{ijk} \varphi\left(\bar{\varepsilon}_a; m_{ijk}; s_{ijk}^2\right), \tag{20}$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right)(\bar{e}_a - \mu_i) + \left(\frac{n_a\sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right)\left(\frac{\nu_j + \nu_k}{n_a}\right) \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2)\sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \tag{21}$$

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε, respectively.

Similarly, to integrate out u_a conditional on e_a (see equation 19), we need the probability distribution for $u_a|e_a$. As mentioned earlier however, we use the distribution of $u_a|\bar{e}_a$, assuming

that $p(u_a|e_a) \approx p(u_a|\bar{e}_a)$. It follows that $u_a|\bar{e}_a$ too is normal-mixture distributed.

Lemma 5 *The probability density function of u_a conditional on $\bar{e}_a$, which we denote by $p(u_a|\bar{e}_a)$, is a normal-mixture with known parameters:*

$$p(u_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \sum_{j=1}^{j=m_\varepsilon} \sum_{k=1}^{k=m_\varepsilon} w_{ijk} \varphi\left(u_a; m_{ijk}; s_{ijk}^2\right), \tag{22}$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\sigma_i^2}{\sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a}\right)(\bar{e}_a - (\nu_j + \nu_k)/n_a) + \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right)\mu_i \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2)\sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \tag{23}$$

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε, respectively.

It should be noted that by working with $p(u_a|\bar{e}_a)$ instead of $p(u_a|e_a)$ we will be solving a modified moment condition, namely:

$$\sum_a E[E[L_{\theta_\varepsilon}(\theta_\varepsilon; e_a, u_a, z_a)|\bar{e}_a, u_a, \hat{\theta}_\varepsilon^{(k)}]|\bar{e}_a] = 0. \tag{24}$$

This denotes a genuine departure from the original EM algorithm where one conditions on all the data that features in the log-likelihood function. We will refer to the resulting estimator as a pseudo-EM estimator. We expect the loss in precision to be minor, while the gain in practicality is substantial.

The resulting estimators, in the form of iterative equations, are presented below (see Annex 7.3 for a derivation). Given some initial estimate of $p(\bar{\varepsilon}_a|\bar{e}_a)$, we may implement the estimator for F_u. Subsequently, given this estimate for F_u, we may implement the estimator for G_ε. The newly obtained estimates can in turn be used to update our estimates for $p(\bar{\varepsilon}_a|\bar{e}_a)$ and $p(u_a|\bar{e}_a)$, after which we may obtain a new round of estimates for F_u and G_ε. This is continued until convergence. (In practice, one iteration is found to be sufficient to obtain accurate estimates.)

Estimator for F_u

The fixed-point solution to the following set of iterative equations yields the estimator $(\hat{\pi}_i, \hat{\mu}_i, \hat{\sigma}_i^2)$

for $(\pi_i, \mu_i, \sigma_i^2)$ for $i = 1, \ldots, m_u$:

$$\hat{\pi}_i^{(k+1)} = E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)\right]/A \tag{25}$$

$$\hat{\mu}_i^{(k+1)} = \frac{E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)(\bar{e}_a - \bar{\varepsilon}_a)|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)\right]}{E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)\right]} \tag{26}$$

$$\hat{\sigma}_i^{2(k+1)} = \frac{E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)\left(\bar{e}_a - \bar{\varepsilon}_a - \hat{\mu}_i^{(k+1)}\right)^2|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)\right]}{E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)\right]}, \tag{27}$$

with:

$$\hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a) = \frac{\hat{\pi}_i^{(k)}\varphi\left(\bar{e}_a - \bar{\varepsilon}_a; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}{\sum_i \hat{\pi}_i^{(k)}\varphi\left(\bar{e}_a - \bar{\varepsilon}_a; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}, \tag{28}$$

where φ denotes the normal probability density function, and where the expectations are taken over $\bar{\varepsilon}$ conditional on $\bar{e}_a$ using the iteration-k estimate of the conditional density function $\hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)$. Lemma 4 shows how $\hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)$ can be obtained as a function of $(\hat{\pi}_i^{(k)}, \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)})$.

Estimator for G_ε

The fixed-point solution to the following set of iterative equations yields the estimator $(\hat{\lambda}_j, \hat{\nu}_j, \hat{\omega}_j^2)$ for $(\lambda_j, \nu_j, \omega_j^2)$ for $j = 1, \ldots, m_\varepsilon$:

$$\hat{\lambda}_j^{(k+1)} = E\left[\sum_{ah} \hat{\tau}_{ahj}^{(k)}(u_a)|\bar{e}_a\right]/n \tag{29}$$

$$\hat{\nu}_j^{(k+1)} = \frac{E\left[\sum_{ah} \hat{\tau}_{ahj}^{(k)}(u_a)(e_{ah} - u_a)|\bar{e}_a\right]}{E\left[\sum_{ah} \hat{\tau}_{ahj}^{(k)}(u_a)|\bar{e}_a\right]} \tag{30}$$

$$\hat{\omega}_j^{2(k+1)} = \frac{E\left[\sum_{ah} \hat{\tau}_{ahj}^{(k)}(u_a)\left(e_{ah} - u_a - \hat{\nu}_j^{(k+1)}\right)^2|\bar{e}_a\right]}{E\left[\sum_{ah} \hat{\tau}_{ahj}^{(k)}(u_a)|\bar{e}_a\right]}, \tag{31}$$

with:

$$\hat{\tau}_{ahj}^{(k)}(u_a) = \frac{\hat{\lambda}_j^{(k)}\varphi\left(e_{ah} - u_a; \hat{\nu}_j^{(k)}, \hat{\omega}_j^{2(k)}\right)}{\sum_j \hat{\lambda}_j^{(k)}\varphi\left(e_{ah} - u_a; \hat{\nu}_j^{(k)}, \hat{\omega}_j^{2(k)}\right)}, \tag{32}$$

where φ denotes the normal probability density function, and where the expectations are taken over u_a conditional on $\bar{e}_a$ using the probability density function $p(u_a|\bar{e}_a)$. See Lemma 5 for the probability density function for $u_a|\bar{e}_a$.

In order to get the iterative scheme started, we would of course need suitable initial estimates for $p(\bar{\varepsilon}_a|\bar{e}_a)$ and $p(u_a|\bar{e}_a)$. Note that even without any prior knowledge about the distributions for u_a and ε_{ah}, we should be able to obtain a reasonable estimate of the distribution for $\bar{\varepsilon}_a$ by appealing to the Central Limit Theorem (CLT). Under the assumption that ε_{ah} are independend across households, we have that the distribution of $\bar{\varepsilon}_a$ tends to a normal distribution with mean zero and variance equal to σ_ε^2/n_a. In a typical household income survey, n_a will be in the range of 10 to 100, which is sufficiently large for the CLT to take effect. The corollaries below derive the initial estimates for $p(\bar{\varepsilon}_a|\bar{e}_a)$ and $p(u_a|\bar{e}_a)$ under the assumption of normal distributed $\bar{\varepsilon}_a$ (with known variance).

Corollary 6 *The probability density function of $\bar{\varepsilon}_a$ conditional on $\bar{e}_a$, which we denote by $p(\bar{\varepsilon}_a|\bar{e}_a)$, is a normal-mixture with known parameters:*

$$p(\bar{\varepsilon}_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} w_i \varphi\left(\bar{\varepsilon}_a; \left(\frac{\sigma_\varepsilon^2/n_a}{\sigma_\varepsilon^2/n_a + \sigma_i^2}\right)(\bar{e}_a - \mu_i); \left(\frac{\sigma_i^2\sigma_\varepsilon^2/n_a}{\sigma_i^2 + \sigma_\varepsilon^2/n_a}\right)\right), \tag{33}$$

where $w_i = \tilde{w}_i / \sum_i \tilde{w}_i$ with:

$$\tilde{w}_i = \pi_i \varphi(\bar{e}_a; \mu_i; \sigma_i^2 + \sigma_\varepsilon^2/n_a), \tag{34}$$

where φ denotes the normal probability density function, and where π_i, μ_i, and σ_i^2 denote the parameters associated with the normal-mixture distribution F_u.

Corollary 7 *The probability density function of u_a conditional on $\bar{e}_a$, which we denote by $p(u_a|\bar{e}_a)$, is a normal-mixture with known parameters:*

$$p(u_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \alpha_i \varphi\left(u_a; \gamma_{ai}\bar{e}_a + (1-\gamma_{ai})\mu_i; \left(1/\sigma_i^2 + n_a/\sigma_\varepsilon^2\right)^{-1}\right), \tag{35}$$

with $\gamma_{ai} = \sigma_i^2/(\sigma_i^2 + \sigma_\varepsilon^2/n_a)$, and where $\alpha_i = \tilde{\alpha}_i / \sum_i \tilde{\alpha}_i$ with:

$$\tilde{\alpha}_i = \pi_i \varphi(\bar{e}_a; \mu_i; \sigma_i^2 + \sigma_\varepsilon^2/n_a), \tag{36}$$

where φ denotes the normal probability density function, and where π_i, μ_i, and σ_i^2 denote the parameters associated with the normal-mixture distribution F_u.

The following propositions provide some properties of the estimators.

Proposition 8 *At every iteration-k, the mean and variance for the probability density function $\hat{g}_\varepsilon^{(k)}(\varepsilon) = \sum_j \hat{\lambda}_j^{(k)} \varphi(\varepsilon; \hat{\nu}_j^{(k)}, \hat{\omega}_j^{2(k)})$ solve:*

$$\begin{aligned} E[\varepsilon; \hat{g}_\varepsilon^{(k)}] &= \frac{1}{n}\sum_{ah} e_{ah} - E[u_a|\bar{e}_a; \hat{p}^{(k)}(u_a|\bar{e}_a)] \\ var[\varepsilon; \hat{g}_\varepsilon^{(k)}] &= \frac{1}{n}\sum_{ah} E[(e_{ah} - u_a)^2|\bar{e}_a; \hat{p}^{(k)}(u_a|\bar{e}_a)] - E^2[\varepsilon; \hat{g}_\varepsilon^{(k)}]. \end{aligned}$$

Proposition 9 *At every iteration-k, the mean and variance for the probability density function* $\hat{f}_u^{(k)}(u) = \sum_i \hat{\pi}_i^{(k)} \varphi(u; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)})$ *solve:*

$$\begin{aligned} E[u; \hat{f}_u^{(k)}] &= \frac{1}{A} \sum_a \bar{e}_a - E[\bar{\varepsilon}_a | \bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a | \bar{e}_a)] \\ var[u; \hat{f}_u^{(k)}] &= \frac{1}{A} \sum_a E[(\bar{e}_a - \bar{\varepsilon}_a)^2 | \bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a | \bar{e}_a)] - E^2[u; \hat{f}_u^{(k)}]. \end{aligned}$$

3.2 Some practical parameter restrictions

Let us also consider the case where the component distributions are assumed to be given, so that only the mixing probability vectors π and λ will have to be estimated. This is obviously a special case of the more general approach where the probabilities are jointly estimated with the parameters of the component distributions. Keeping the latter fixed markedly benefits the numerical convergence, indicating that the approach deserves to be considered a distinct option in and of itself. This is precisely the approach advocated by Cordy and Thomas (1997).

The estimators for π and λ are obtained as solutions to the following iterative equations:

$$\tilde{\pi}_i^{(k+1)} = \frac{1}{A} \sum_a \frac{\tilde{\pi}_i^{(k)} \varphi\left(\bar{e}_a; \mu_i, \sigma_i^2 + \sigma_{\bar{\varepsilon}}^2 / n_a\right)}{\sum_i \tilde{\pi}_i^{(k)} \varphi\left(\bar{e}_a; \mu_i, \sigma_i^2 + \sigma_{\bar{\varepsilon}}^2 / n_a\right)}, \tag{37}$$

and:

$$\tilde{\lambda}_j^{(k+1)} = \frac{1}{n} \sum_{ah} \frac{\tilde{\lambda}_j^{(k)} \left(\sum_i \tilde{\pi}_i \varphi\left(e_{ah}; \mu_i + \nu_j, \sigma_i^2 + \omega_j^2\right)\right)}{\sum_j \tilde{\lambda}_j^{(k)} \left(\sum_i \tilde{\pi}_i \varphi\left(e_{ah}; \mu_i + \nu_j, \sigma_i^2 + \omega_j^2\right)\right)}, \tag{38}$$

where $\tilde{\pi}_i$ denotes the estimator for π_i. Note that Cordy and Thomas (1997) only consider the estimator for F_u (i.e. the estimator for π_i for $i = 1, \ldots, m_u$), as they are not interested in the distribution function for ε_{ah}.

Since the parameters of the component distributions are not being estimated, the modeler will have to set the values for the means μ_i and ν_j, and variances σ_i^2 and ω_j^2 beforehand. Cordy and Thomas (1997) recommend to set a common variance (as is also done in kernel density estimators where the common variance is refered to as the bandwidth parameter). A natural choice is to set this common variance equal to: $\sigma_i^2 = \frac{\hat{\sigma}_u^2}{m_u}$ for all i (and hence $\omega_j^2 = \frac{\hat{\sigma}_\varepsilon^2}{m_\varepsilon}$). Note that for any given common variance $\bar{\sigma}^2$ we have: $var[u] = \bar{\sigma}^2 + \sum_i \pi_i \mu_i^2$, with $\sum_i p_i \mu_i^2 \geq 0$. This means that at a minimum the common variance must be chosen so that: $var[u] - \bar{\sigma}^2 \geq 0$. Obviously this is satisfied for $\bar{\sigma}^2 = \frac{\sigma_u^2}{m_u}$, which yields: $\sum_i \pi_i \mu_i^2 = (\frac{m_u - 1}{m_u}) \sigma_u^2$.

A convenient choice for the mean parameters is to take equally spaced μ_i (and ν_j) with a range that respects the values attained by the observed data. Note that the estimated mixing probabilities are expected to satisfy: $E[u] = \sum_i \hat{\pi}_i \mu_i = \sum_a \bar{e}_a / A$ (matching first moment), as well as: $\sum_i \hat{\pi}_i \mu_i^2 = (\frac{m_u - 1}{m_u}) \hat{\sigma}_u^2$ (matching second moment), although this is not guaranteed by the estimation approach. This is in contrast with the approach where the component distribution parameters are also being estimated, in which case the first and second moments are guaranteed.

Remark 10 *One possible extension to the set of component distributions proposed above (with different μ_i but common σ^2) can be obtained by adding components with common zero mean (i.e. $\mu_r = 0$) but different variances σ_r^2.*

4 A small simulation study

This section presents a modest Monte-Carlo simulation experiment. We will focus our attention to the following two questions: (1) How effective is our approach in fitting non-normal error distributions (that are not necessarily normal-mixtures)?, and (2) What are the implications for the estimation of poverty and inequality? Do distinct deviations from normality have the potential to introduce a meaningful bias in estimates of poverty and inequality if this non-normality is ignored?

We make the following assumptions. The simulated "country" consists of 500 domains (or target areas), which denotes the level at which measures of poverty and inequality will be estimated. Each domain is home to 3000 households. Household per capita incomes will be generated by means of the following model:

$$y_{ah} = \beta x_{ah} + u_a + \varepsilon_{ah}, \tag{39}$$

where y_{ah} denotes the log of household income, and where x_{ah} represents a single covariate. For the model parameters we consider the values: $var[u_a + \varepsilon_{ah}] = \sigma_u^2 + \sigma_\varepsilon^2 = 0.3$, $\beta = 1$, $E[x_{ah}] = 0$, and where $var[x_{ah}]$ is chosen so that $R^2 = 0.4$ (which represents a goodness-of-fit that is typical for empirical household income models).

It will be convenient to introduce a parameter that measures the size of the random area effect relative to the total error term. Let us denote this parameter by $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)$. We will be considering both the case of small "area effect" ($\rho = 0.05$) and medium/large "area effect" ($\rho = 0.25$).

The non-normal errors u_a and ε_{ah} will be drawn from a Log-Dagum distribution with probability density function:

$$f(x) = \frac{ape^{ap(x-log(b))}}{\left(1 + e^{a(x-log(b))}\right)^{p+1}}, \tag{40}$$

where $p > 0$ is the parameter that determines the shape (or skewness) of the distribution. Smaller values of p will result in larger deviations from normality. The remaining two parameters a and b will be fixed by imposing zero mean and setting the variance to σ_u^2 or σ_ε^2. Note that the Dagum distribution is not an uncommon choice when modeling income data (see e.g. Kleiber (2007), and the references therein). The covariate x_{ah} will be drawn from a normal distribution.

Finally, our artificial survey will sample 15 households from each domain.

4.1 Estimation of F_u and G_ε

The first test will be whether we can successfully uncover the probability density functions for u_a and ε_{ah}. We will keep the shape parameter for F_u fixed at $p_u = 0.5$, but will consider three different values of the shape parameter for G_ε: $p_\varepsilon = (0.10, 0.25, 0.50)$. The benchmark density functions are obtained as a kernel density estimate applied to the actual realizations of the errors in the census. Our estimators for F_u and G_ε are based on normal-mixtures with 3 and 2 component distributions, respectively.

Figure 1 presents our estimates of the probability density function for u_a. The estimates in the right panel show a remarkably good fit. These correspond to the case where the random area effect is large, with $\rho = 0.25$, in which case the contribution of u_a matters. The imperfect fit shown in the left panel suggests that it is harder to estimate F_u when the random area effect is small (in this case $\rho = 0.05$). This is arguably due to the poor signal-to-noise ratio in this case; u_a will only make up a small fraction of the total residual which is what is observed. However, this lack of precision will also have little to no implication for estimates of poverty and inequality, precisely because of the smallness of u_a. In sum, in this example estimates of F_u are precise when they matter, and less precise when they matter less.

Probability density function for u

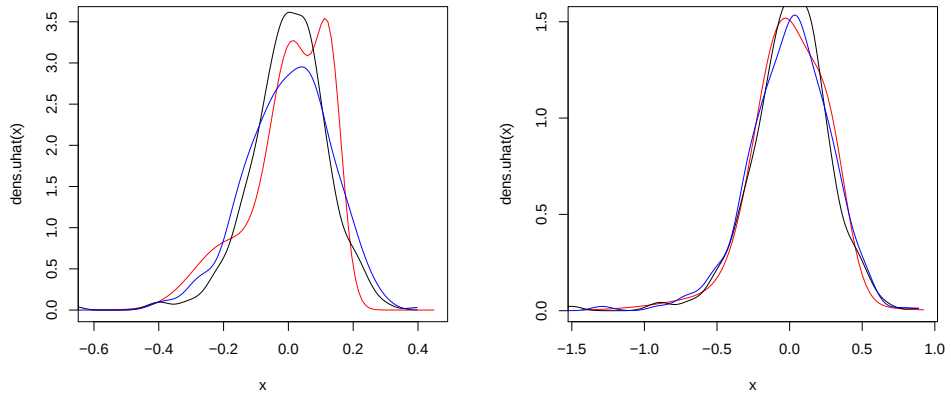

Figure 1: Normal-mixture density (Black) versus benchmark density (Red): (a) $\rho = 0.05$ (left panel); (b) $\rho = 0.25$ (right panel)

Our estimates for the probability density functions for ε_{ah} are presented in Figure 2. What is apparent is that, despite using no more than two components, our estimates provide remarkably accurate fits regardless of the size of the random area effect. When ρ is small, we have that ε_{ah} is large relative to u_a, and so the signal-to-noise ratio is in our favour when estimating G_ε. It matters less in this case that we do not have a precise estimate for F_u. On the other hand, when ρ is large and hence ε_{ah} is of a more modest magnitude relative to u_a, our estimation of G_ε may be helped by having a precise estimate of F_u. (Note that we first estimate F_u, and then use this estimate to subsequently estimate G_ε.)

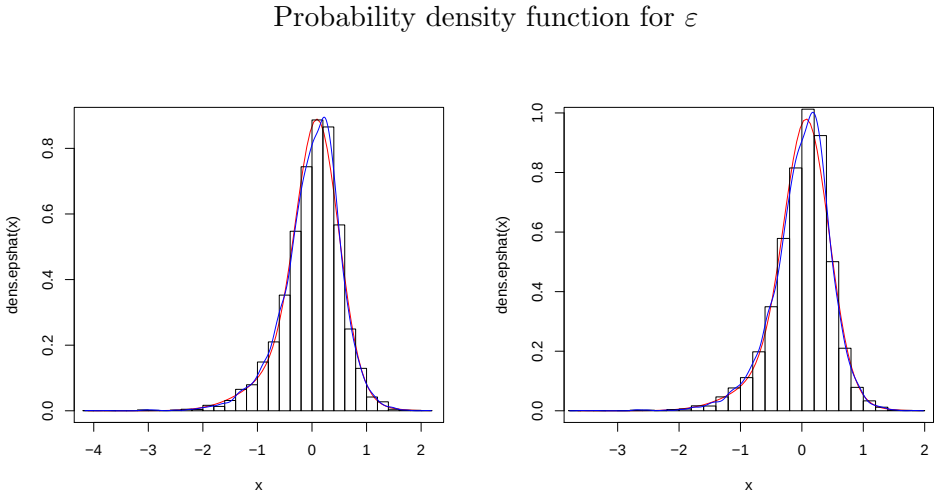

Figure 2: Normal-mixture density (Black) versus benchmark density (Red): (a) $\rho = 0.05$ (left panel); (b) $\rho = 0.25$ (right panel)

4.2 Implications for estimates of poverty

Next we investigate whether having more precise estimates of F_u and G_ε will also give us more precise estimates of poverty, and whether any gain in precision is economically meaningful. For ease of exposition we will focus on the percentage of households with incomes below the poverty line as the measure of poverty which we wish to estimate. Tables 1 and 2 provide estimates of the bias for different values of the shape parameter for G_ε (the shape parameter for F_u is fixed at $p_u = 0.5$), and for different values of the (log) poverty line. The poverty rates associated with the different (log) poverty lines are roughly 15, 20, 30, and 45 percent. It should be noted that the bias in this case is estimated as the difference between the estimated and the true poverty rate averaged over the 500 target areas for one given replication of the "census".

At least two observations stand out. First, estimates obtained under the (incorrect) assumption of normal errors can be severly biased, with a bias of 3 to 4 percent on a poverty rate that ranges between 20 and 30 percent depending on the shape parameter for G_ε. Our approach provides far superior estimates of poverty in these cases with a bias of less than a percent. Second, the benefit of accommodating non-normal errors as opposed to assuming normal errors changes with the value of the poverty line. At the far left tail of the (log) income distribution (i.e. for particularly low values of the poverty line), it will be hard to empirically separate the two distributions. The difference will be more pronounced where the distributions have more "mass", i.e. for intermediate to higher values of the poverty line.

4.3 Implications for estimates of inequality

Table 5 compares the bias for estimates of income inequality for different values of the shape parameter and different values of the area location effect. We focus on the Mean-Log-Deviation (MLD) as the measure of inequality which we wish to estimate. Note that by definition we have that the realizations of the area error u_a will drop out of the true measure of MLD for that area,

Poverty	$\rho = 0.05$			$\rho = 0.25$		
line/shape	0.50	0.25	0.10	0.50	0.25	0.10
-0.75	0.67 (1.61)	0.64 (1.69)	1.32 (2.06)	1.38 (2.51)	1.40 (2.70)	1.06 (2.55)
-0.50	0.76 (2.51)	0.82 (3.29)	1.35 (4.01)	1.39 (3.09)	1.55 (3.78)	1.32 (3.92)
-0.25	0.67 (2.73)	0.92 (4.16)	1.14 (5.15)	1.01 (2.82)	1.32 (3.92)	1.34 (4.37)
0	0.31 (1.96)	0.73 (3.60)	0.60 (4.51)	0.20 (1.53)	0.61 (2.73)	0.89 (3.32)

Table 1: Bias for EB estimates of head-count poverty: (a) Normal-Mixture distribution (left number); (b) Normal distribution (right number)

Poverty	$\rho = 0.05$			$\rho = 0.25$		
line/shape	0.50	0.25	0.10	0.50	0.25	0.10
-0.75	0.19 (0.98)	0.12 (1.07)	0.91 (1.45)	0.23 (0.79)	0.22 (0.97)	-0.08 (0.85)
-0.50	0.25 (1.95)	0.22 (2.73)	0.82 (3.47)	0.36 (1.63)	0.41 (2.31)	0.16 (2.48)
-0.25	0.30 (2.40)	0.44 (3.83)	0.62 (4.82)	0.41 (2.03)	0.57 (3.13)	0.53 (3.60)
0	0.23 (19.7)	0.58 (3.60)	0.35 (4.51)	0.28 (1.66)	0.57 (2.86)	0.79 (3.47)

Table 2: Bias for NON-EB estimates of head-count poverty: (a) Normal-Mixture distribution (left number); (b) Normal distribution (right number)

Poverty	$\rho = 0.05$			$\rho = 0.25$		
line/shape	0.50	0.25	0.10	0.50	0.25	0.10
-0.75	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)	0.04 (0.05)	0.04 (0.05)	0.04 (0.05)
-0.50	0.04 (0.05)	0.04 (0.05)	0.04 (0.06)	0.05 (0.06)	0.05 (0.06)	0.05 (0.06)
-0.25	0.05 (0.06)	0.05 (0.07)	0.05 (0.07)	0.06 (0.07)	0.06 (0.07)	0.06 (0.08)
0	0.06 (0.06)	0.06 (0.07)	0.06 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.08)

Table 3: RMSE for EB estimates of head-count poverty: (a) Normal-Mixture distribution (left number); (b) Normal distribution (right number)

Poverty	$\rho = 0.05$			$\rho = 0.25$		
line/shape	0.50	0.25	0.10	0.50	0.25	0.10
-0.75	0.04 (0.04)	0.04 (0.04)	0.04 (0.04)	0.10 (0.10)	0.09 (0.09)	0.09 (0.09)
-0.50	0.05 (0.06)	0.05 (0.06)	0.05 (0.06)	0.12 (0.12)	0.12 (0.12)	0.12 (0.12)
-0.25	0.07 (0.07)	0.07 (0.08)	0.07 (0.08)	0.14 (0.15)	0.14 (0.15)	0.14 (0.15)
0	0.07 (0.08)	0.07 (0.08)	0.08 (0.09)	0.15 (0.15)	0.16 (0.16)	0.16 (0.16)

Table 4: RMSE for NON-EB estimates of head-count poverty: (a) Normal-Mixture distribution (left number); (b) Normal distribution (right number)

which varies around 20 for our simulated data (we multiplied the numbers by a factor 100). Note however that the size of the "location effect" may still have a bearing on our estimates of inequality, to the extent that it impacts on our estimates of the distribution for ε_{ah}.

Similarly to what we observed for poverty, the potential gain of accommodating non-normal errors as opposed to incorrectly assuming normal errors can be quite large. We observe biases of 2 to 3 with true inequality around 20 when assuming normal errors, compared to biases of roughly 0.5 when using our approach.

	Shape parameter		
MLD	0.50	0.25	0.10
$\rho = 0.05$	0.36 (1.67)	0.52 (2.59)	1.12 (3.71)
$\rho = 0.25$	0.37 (1.26)	0.61 (2.18)	0.47 (2.25)

Table 5: Bias for estimates of inequality: (a) Normal-Mixture distribution (left number); (b) Normal distribution (right number)

5 A small empirical study

For this small empirical application we will treat the 2010 micro census from the United States, a 1 percent sample, as a "new" country (i.e. as if the data constitutes a full population census). The data is publically available at IPUMS USA. There are a total of 1.25 million households[7] residing in 422 counties, which is the level at which we will be estimating poverty and inequality (i.e. the "small area level"). The census includes individual income data as well as data on a wide range of individual characteristics. We use the income data to compute household income per capita. All other individual level variables are also aggregated at the household level as this denotes the level at which we will build the income model. We defined the head of the household as the member of the household with the highest level of individual income. (In many empirical applications much of the available data is at the household level.) Table 6 provides some descriptive statistics of the data.

Variable	Description	Mean	Std. Dev.
lnycap	Log of household income per capita	9.939	0.998
metro_suburb	In sub-urb of major city (dummy)	0.290	0.454
metro_none	Away from major city (dummy)	0.201	0.401
hh_size_1	Household of size one (dummy)	0.300	0.458
hh_size_2	Household of size two (dummy)	0.342	0.474
hh_size_3	Household of size three (dummy)	0.147	0.354
share_child15	Share of children aged younger than 15	0.108	0.199
hh_hedu2	At least one member with a masters degree (dummy)	0.366	0.482
hh_employed	Number of employed household members	1.104	0.936
hd_female	Head of household is female (dummy)	0.417	0.493
hd_logage	Log of age of household head	3.889	0.371
hd_ledu	Household head has lower education only (dummy)	0.048	0.214
hd_hedu1	Household head has college degree (dummy)	0.226	0.418
hd_hedu2	Household head has a masters degree (dummy)	0.312	0.463
hd_employed	Household head is employed (dummy)	0.660	0.474
hd_hisp	Household head is Hispanic (dummy)	0.091	0.288
hd_black	Household head is African-American (dummy)	0.106	0.308

Table 6: Descriptive statistics of variables used in empirical application (derived from 2010 IPUMS USA)

We draw a balanced sample of 6330 households (15 households per county). The log of

[7]The exact number of households in the census is $1,242,967$.

household per capita income will be our dependent variable. Our independent variables include: urbanity, household size, age composition, gender, ethnicity, education, and employment. There are a total of 16 regressors (excluding the constant). The regression model is shown in Table 7. We obtained an adjusted R-squared of 0.432, which denotes a typical level for this type of income (or consumption) regression model. The location effect is estimated at 0.027, which is not unusual for developed nations but rather small for developing countries where one might find location effects in the range of 0.05 to 0.25.

	lnycap	
metro_suburb	0.171***	(8.21)
metro_none	-0.0542*	(-1.68)
hh_size_1	0.733***	(16.32)
hh_size_2	0.620***	(16.00)
hh_size_3	0.344***	(9.73)
share_child15	-0.160**	(-2.20)
hh_hedu2	0.277***	(6.47)
hh_employed	0.257***	(14.24)
hd_female	-0.212***	(-10.80)
hd_logage	0.898***	(30.46)
hd_ledu	-0.331***	(-6.40)
hd_hedu1	0.175***	(7.06)
hd_hedu2	0.450***	(9.81)
hd_employed	0.484***	(15.25)
hd_hisp	-0.159***	(-4.36)
hd_black	-0.248***	(-7.78)
_cons	5.201***	(39.18)
N	6330	
adj. R^2	0.432	

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Regression model with log household income per capita as dependent variable (sample from 2010 IPUMS USA)

Figure 3 shows the estimates of the probability density function associated with the area error u_a and the household error ε_{ah}, respectively. The "red" line denotes the normal mixture density estimate, "blue" line denotes the normal density function that best fits the data. As we are dealing with empirical data, not simulated data, we do not know the distributions from which the errors are drawn, and hence are not able to include the "true" density function as a benchmark. The estimates suggest that the household errors are arguably drawn from a distribution that deviates visibly from a normal distribution. We do not see a similar deviation from the normal density for the area error. The smallness of the variance of the area error relative to the total error however, suggests that the signal to noise ratio is low making it difficult to identify the underlying density function. The small variance also means that the area error makes only a minor contribution to the total error and so its density function is less important. It is clearly more important that our estimate of the density associated with the

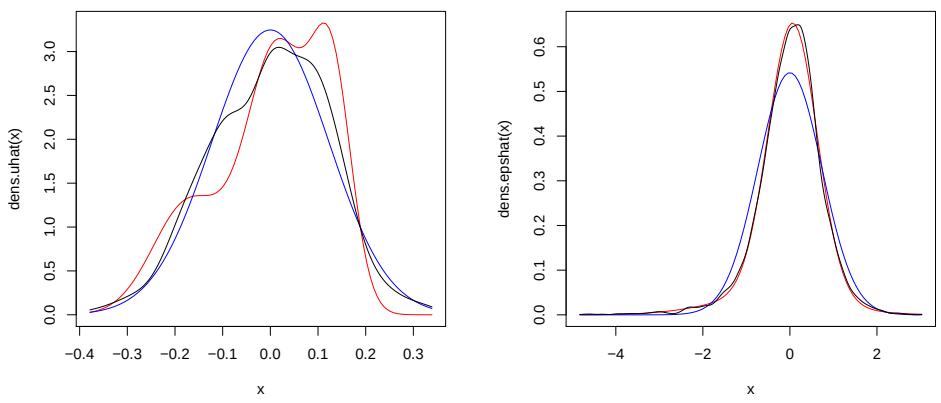

Figure 3: Normal-mixture density (Red) versus normal density (Blue) estimates for: (a) area error u_a (left panel); (b) household error ε_{ah} (right panel)

household error is accurate.

The log poverty line is set at 9.5 which yields an aggregate poverty rate of about 28.3 percent. County level estimates of poverty and inequality are obtained using both EB and non-EB estimates under the assumption of normal errors and under the less restrictive assumption of normal mixture errors. For each of the specifications we obtain 422 estimates of poverty and inequality which we then compare to the true estimates derived from the full census. We will judge the accuracy of the estimates on the basis of two summary statistics: the aggregate bias and the RMSE (both obtained by aggregating the county level differences between our estimates and the true rates over the 422 counties). The results are presented in Table 8.

	Bias	RMSE
EB	0.029 (0.048)	0.045 (0.059)
NON-EB	0.024 (0.044)	0.042 (0.056)

Table 8: Bias and RMSE for estimates of county-level poverty using 2010 data from the United States: (a) Normal-Mixture distribution (left number); (b) Normal distribution (right number)

We observe a rather large bias when estimating county level poverty under the assumption of normal distributed errors; almost 5 percentage points given a national poverty rate of about 28 percent. By relaxing the normality assumption, and assuming normal mixture distributions instead, we are able to reduce this bias to just below 3 percent which denotes a non-trivial improvement. We see similar reductions in the RMSE. Working with EB or non-EB estimates makes little to no difference in this empirical application. This was to be expected given the smallness of the location effect. Note that the fact that some bias still remains suggests that a degree of model misspecification still persists; these could include mis-specifications in the structural model but also forms of heteroskedasticity that are currently being ignored.

6 Concluding remarks

We have developed an estimator for the error distribution functions associated with the nested errors from a linear mixed model. An attractive feature of the estimator is that it accommodates Empirical Bayes prediction. This is achieved without making any restrictive assumptions about the shape of the distrbution functions. Monte-Carlo simulations presented in this paper show that estimates of poverty and inequality can be severely biased when non-normality of the errors is ignored. The bias can be as high as 2 to 3 percent on a poverty rate of 20 to 30 percent. Most of this bias is resolved when using our approach.

7 Appendix

7.1 A lemma and corollary that are used in the proofs of Lemmas 4 and 5

Lemma 11 *Let v_1 and v_2 be two independent normal random variates, with $Ev_i = \mu_i, \operatorname{var} v_i = \sigma_i^2$ for $i = 1, 2$. Let $v_3 = v_1 + v_2$. Denote the normal density function with expectation μ and variance σ^2 by $\varphi(\cdot; \mu, \sigma)$. Then for the joint density $p(v_1, v_3)$ and the conditional density $p(v_1|v_3)$ we have*

$$\begin{aligned} p(v_1, v_3) &= \varphi(v_1; \mu_1, \sigma_1)\varphi(v_3 - v_1; \mu_2, \sigma_2) \\ &= p(v_1|v_3)p(v_3) \\ &= \varphi\left(v_1; \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(v_3 - \mu_2) + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1, \frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \\ &\quad \times\varphi\left(v_3; \mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right). \end{aligned}$$

Proof This can be directly verified. Alternatively, note that both v_3 and $v_1|v_3$ are normally distributed. The distribution of $v_1|v_3$ has expectation

$$E(v_1|v_3) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(v_3 - \mu_2) + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1$$

and variance

$$\operatorname{var}(v_1|v_3) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

□

Corollary 12 *Let x and y be independent random variables distributed as normal mixtures, and let $q = x + y$. Then (i) q is also distributed as a normal mixture, and (ii) the conditional distributions $p(x|q)$ and $p(y|q)$ are also normal mixtures. Moreover, the mixing parameters of $p(q)$, $p(x|q)$ and $p(y|q)$ are closed expressions in terms of the distributional parameters of x and*

y. In particular, if the density function of x is

$$f(x) = \sum_{k=1}^{a} \pi_k \varphi(x; \mu_k, \sigma_k),$$

(with $\pi_k > 0$ and $\sum_k \pi_k = 1$) and the density function of y is

$$g(y) = \sum_{j=1}^{b} \lambda_j \varphi(y; \nu_j, \tau_j),$$

(with $\lambda_j > 0$ and $\sum_j \lambda_j = 1$) then the conditional distribution of x, given q is

$$h(x|q) = \sum_{jk} \kappa_{jk} \varphi(x; m_{jk}, s_{jk}),$$

where the mixing probabilities κ_{jk} are proportional to

$$\kappa_{jk} \propto \pi_k \lambda_j \varphi\left(q; \mu_k + \nu_j, \sqrt{\sigma_k^2 + \tau_j^2}\right)$$

and

$$\begin{aligned} m_{jk} &= \frac{\sigma_k^2}{\sigma_k^2 + \tau_j^2}(q - \nu_j) + \frac{\tau_j^2}{\sigma_k^2 + \tau_j^2}\mu_k \\ s_{jk}^2 &= \frac{\sigma_k^2 \tau_j^2}{\sigma_k^2 + \tau_j^2} \end{aligned}$$

Proof The random variable x can be expressed as

$$x = \sum_{k=1}^{a} z_k x_k,$$

where the x_k are independent (latent) random variables, and also independent of $(z_1, ..., z_a)$. x_k is normally distributed according to component distribution k, with density $\varphi(x_k; \mu_k, \sigma_k)$. The z_ks are indicator random variables taking values 0 and 1 with probability π_k. One and only one indicator takes the value 1, the others are 0.[8] Likewise y can be expressed as

$$x = \sum_{j=1}^{b} w_j y_j,$$

where the (w_j, y_j) independent of the (z_k, x_k). Note that $z_k = z_k \sum_j w_j$ and $w_j = w_j \sum_k z_k$, so that

$$q = x + y = \sum_k z_k x_k + \sum_j w_j y_j = \sum_{jk} z_k w_j (x_k + y_j).$$

[8] In other words, the $(z_1, ..., z_a)$ have a multinomial distribution with class probabilities π_k and a single draw.

Part (i) of the Corollary now immediately follows; the mixing probabilities for q are $P[z_k w_j = 1] = \pi_k \lambda_j$ and the component distribution for component (k, j) has expectation $\mu_k + \nu_j$ and variance $\sigma_k^2 + \tau_j^2$.

For the conditional distribution $p(x|q)$ we have

$$\begin{aligned} p(x|q) &\propto \\ p(x,q) &= \sum_{jk} \pi_k \lambda_j \varphi(x; \mu_k, \sigma_k) \varphi(q - x; \nu_j, \tau_j) \\ &= \sum_{jk} \pi_k \lambda_j \varphi\left(x; \frac{\sigma_k^2}{\sigma_k^2 + \tau_j^2}(q - \nu_j) + \frac{\tau_j^2}{\sigma_k^2 + \tau_j^2}\mu_k, \frac{\sigma_k \tau_j}{\sqrt{\sigma_k^2 + \tau_j^2}}\right) \\ &\quad \times \varphi\left(q; \mu_k + \nu_j, \sqrt{\sigma_k^2 + \tau_j^2}\right). \end{aligned}$$

The last equality follows from the Lemma.

The conditional distribution $p(x|q)$ is therefore again a normal mixture with mixing probabilities for the (k, j) component proportional to

$$\pi_k \lambda_j \varphi\left(q; \mu_k + \nu_j, \sqrt{\sigma_k^2 + \tau_j^2}\right)$$

and corresponding component distribution

$$\varphi\left(x; \frac{\sigma_k^2}{\sigma_k^2 + \tau_j^2}(q - \nu_j) + \frac{\tau_j^2}{\sigma_k^2 + \tau_j^2}\mu_k, \frac{\sigma_k \tau_j}{\sqrt{\sigma_k^2 + \tau_j^2}}\right).$$

The distribution of $p(y|q)$ follows in exactly the same fashion. □

7.2 Proofs

Lemma 4: The probability density function of $\bar{\varepsilon}_a$ conditional on $\bar{e}_a$, which we denote by $p(\bar{\varepsilon}_a|\bar{e}_a)$, is a normal-mixture with known parameters:

$$p(\bar{\varepsilon}_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \sum_{j=1}^{j=m_\varepsilon} \sum_{k=1}^{k=m_\varepsilon} w_{ijk} \varphi\left(\bar{\varepsilon}_a; m_{ijk}; s_{ijk}^2\right), \tag{41}$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2}\right)(\bar{e}_a - \mu_i) + \left(\frac{n_a \sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2}\right)\left(\frac{\nu_j + \nu_k}{n_a}\right) \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2)\sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2}\right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \tag{42}$$

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε, respectively.
Proof The result stated in this lemma follows from a direct application of Corollary 12. □

Lemma 5: The probability density function of u_a conditional on $\bar{e}_a$, which we denote by $p(u_a|\bar{e}_a)$, is a normal-mixture with known parameters:

$$p(u_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \sum_{j=1}^{j=m_\varepsilon} \sum_{k=1}^{k=m_\varepsilon} w_{ijk} \varphi\left(u_a; m_{ijk}; s_{ijk}^2\right), \tag{43}$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\sigma_i^2}{\sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a}\right)(\bar{e}_a - (\nu_j + \nu_k)/n_a) + \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right)\mu_i \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2)\sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a\sigma_i^2}\right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \tag{44}$$

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε, respectively.
Proof The result stated in this lemma follows from a direct application of Corollary 12. □

Proposition 9: At every iteration-k, the mean and variance for the probability density function $\hat{f}_u^{(k)}(u) = \sum_i \hat{\pi}_i^{(k)} \varphi(u; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)})$ solve:

$$\begin{aligned} E[u; \hat{f}_u^{(k)}] &= \frac{1}{A} \sum_a \bar{e}_a - E[\bar{\varepsilon}_a|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)] \\ var[u; \hat{f}_u^{(k)}] &= \frac{1}{A} \sum_a E[(\bar{e}_a - \bar{\varepsilon}_a)^2|\bar{e}_a; \hat{p}^{(k)}(\bar{\varepsilon}_a|\bar{e}_a)] - E^2[u; \hat{f}_u^{(k)}]. \end{aligned}$$

Proof Let us first evaluate the first moment:

$$E_k[u] = \sum_i \hat{\pi}_i^{(k)} \hat{\mu}_i^{(k)}, \tag{45}$$

where $E_k[\cdot]$ takes expectations using the iteration-k estimate of the pdf for u: $\hat{f}_u^{(k)}(u) = \sum_i \hat{\pi}_i^{(k)} \varphi(u; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)})$. Substituting the expressions for $\hat{\pi}_i^{(k)}$ and $\hat{\mu}_i^{(k)}$, we obtain:

$$\begin{aligned}
\sum_i \hat{\pi}_i^{(k)} \hat{\mu}_i^{(k)} &= \sum_i \left(\frac{1}{A} E_k[\sum_a \hat{\tau}_{ai}^{(k)} | \bar{e}_a] \right) \left(\frac{E_k[\sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon}) | \bar{e}_a]}{E_k[\sum_a \hat{\tau}_{ai}^{(k)} | \bar{e}_a]} \right) && (46) \\
&= \frac{1}{A} \sum_i \sum_a E_k[\hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon}) | \bar{e}_a] && (47) \\
E &= \frac{1}{A} \sum_a E_k[\sum_i \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon}) | \bar{e}_a]. && (48)
\end{aligned}$$

By definition we have that $\sum_i \hat{\tau}_{ai}^{(k)} = 1$, so that the latter simplifies to:

$$E_k[u] = \frac{1}{A} \sum_a \bar{e}_a - E_k[\bar{\varepsilon} | \bar{e}_a]. \tag{49}$$

Next is the expression for the variance of u:

$$var_k[u] = \sum_i \hat{\pi}_i^{(k)} \left(\hat{\sigma}_i^{2(k)} + \hat{\mu}_i^{2(k)} \right) - E_k^2[u]. \tag{50}$$

Substituting the expressions for $\hat{\pi}_i^{(k)}$, $\hat{\mu}_i^{(k)}$ and $\hat{\sigma}_i^{2(k)}$ yields:

$$\begin{aligned}
\sum_i \hat{\pi}_i^{(k)} \left(\hat{\sigma}_i^{2(k)} + \hat{\mu}_i^{2(k)} \right) &= \sum_i \left(\frac{1}{A} E_k[\sum_a \hat{\tau}_{ai}^{(k)} | \bar{e}_a] \right) \left(\frac{E_k[\sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon} - \hat{\mu}_i^{(k)})^2 | \bar{e}_a]}{E_k[\sum_a \hat{\tau}_{ai}^{(k)} | \bar{e}_a]} + \hat{\mu}_i^{2(k)} \right) \\
&= \frac{1}{A} \sum_i \left(E_k[\sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon} - \hat{\mu}_i^{(k)})^2 | \bar{e}_a] + E_k[\sum_a \hat{\tau}_{ai}^{(k)} | \bar{e}_a] \hat{\mu}_i^{2(k)} \right) \\
&= \frac{1}{A} \sum_i E_k \left[\sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon})^2 + 2 \sum_a \hat{\tau}_{ai}^{(k)} \hat{\mu}_i^{2(k)} - 2 \sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon}) \hat{\mu}_i^{(k)} | \bar{e}_a \right].
\end{aligned}$$

From the expression for $\hat{\tau}_{ai}^{(k)}$ we know that $E_k[\sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon}) | \bar{e}_a] = E_k[\sum_a \hat{\tau}_{ai}^{(k)} | \bar{e}_a] \hat{\mu}_i^{(k)}$. Plugging this into the last obtained equation, the last two terms are seen to cancel out, and we obtain:

$$\begin{aligned}
\sum_i \hat{\pi}_i^{(k)} \left(\hat{\sigma}_i^{2(k)} + \hat{\mu}_i^{2(k)} \right) &= \frac{1}{A} \sum_i E_k \left[\sum_a \hat{\tau}_{ai}^{(k)} (\bar{e}_a - \bar{\varepsilon})^2 | \bar{e}_a \right] && (51) \\
&= \frac{1}{A} \sum_a E_k \left[(\sum_i \hat{\tau}_{ai}^{(k)}) (\bar{e}_a - \bar{\varepsilon})^2 | \bar{e}_a \right] && (52) \\
&= \frac{1}{A} \sum_a E_k \left[(\bar{e}_a - \bar{\varepsilon})^2 | \bar{e}_a \right]. && (53)
\end{aligned}$$

This completes the proof. □

7.3 A derivation of the EM-type estimator for F_u

Below we will derive the EM-type estimator for the normal-mixture distribution F_u. The estimator for G_ε is obtained in a very similar fashion.

Let Ψ denote the vector that contains the parameters of this normal-mixture, which includes the mixing probabilities π_i as well as the parameters μ_i and σ_i^2 of the component distributions. Assume for the moment that u_a denotes observed data. The EM algorithm would then proceed as follows. It introduces the random variable z_{ai} that equals 1 if u_a comes from component i and 0 otherwise, and defines $L(\Psi; u, z)$ as the log likelihood function that would apply if both u and z were observed. It then integrates out z by taking expectations over z conditional on an estimate of Ψ that is available at iteration k, which we shall denote $\hat{\Psi}^{(k)}$. The solution $\hat{\Psi}^{(k+1)}$ that maximizes the resulting log likelihood function $Q(\Psi; \hat{\Psi}^{(k)}) = E[L(\Psi; u, z)|\hat{\Psi}^{(k)}]$ can be obtained analytically, unlike the solution to the original log likelihood function. The EM estimator for Ψ in this case is obtained as the fixed point of this iterative equation.

Unfortunately, we do not observe u_a. According to the nested error model, $u_a = \bar{e}_a - \bar{\varepsilon}_a$. We observe $\bar{e}_a$, but not $\bar{\varepsilon}_a$. The EM algortihm however can still be applied. Instead of integrating out z, we will be integrating out both z and $\bar{\varepsilon}_a$. With slight abuse of notation, the resulting log likelihood function is now given by: $Q(\Psi; \hat{\Psi}^{(k)}) = E[L(\Psi; \bar{e}, \bar{\varepsilon}, z)|\hat{\Psi}^{(k)}]$. Note that this function $Q(\Psi; \hat{\Psi}^{(k)})$ will be different from the function that is obtained when u is treated as observed data and only z is integrated out. (But it is obtained respecting the principles of the EM algorithm.)

It will be convenient to wait with taking the expectations over $\bar{\varepsilon}$ until after we have evaluated the partial derivatives of $Q(\Psi; \hat{\Psi}^{(k)})$ with respect to Ψ. It follows that:

$$Q(\Psi; \hat{\Psi}^{(k)}) = E[Q(\Psi; \hat{\Psi}^{(k)}, \bar{\varepsilon})|\bar{e}], \tag{54}$$

where expectations are taken over $\bar{\varepsilon}$ conditional on $\bar{e}$, and where:

$$Q(\Psi; \hat{\Psi}^{(k)}, \bar{\varepsilon}) = \sum_i \sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a) \log\left(\pi_i \varphi(\bar{e}_a - \bar{\varepsilon}_a; \mu_i, \sigma_i^2)\right), \tag{55}$$

with:

$$\hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a) = \frac{\hat{\pi}_i^{(k)} \varphi\left(\bar{e}_a - \bar{\varepsilon}_a; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}{\sum_i \hat{\pi}_i^{(k)} \varphi\left(\bar{e}_a - \bar{\varepsilon}_a; \hat{\mu}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}. \tag{56}$$

Evaluating the partial derivative of $Q(\Psi; \hat{\Psi}^{(k)})$ with respect to $\Psi = (\pi, \mu, \sigma^2)$ and bringing this inside the expectation operator yields:

$$\begin{aligned}
\frac{\partial Q}{\partial \pi_i} &= E\left[\frac{\partial Q(\Psi; \hat{\Psi}^{(k)}, \bar{\varepsilon})}{\partial \pi_i}|\bar{e}\right] = E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)\frac{1}{\pi_i}|\bar{e}_a\right] \\
\frac{\partial Q}{\partial \mu_i} &= E\left[\frac{\partial Q(\Psi; \hat{\Psi}^{(k)}, \bar{\varepsilon})}{\partial \mu_i}|\bar{e}\right] = E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)\frac{1}{\sigma_i^2}(\bar{e}_a - \bar{\varepsilon}_a - \mu_i)|\bar{e}_a\right]
\end{aligned}$$

$$\frac{\partial Q}{\partial \sigma_i^2} = E\left[\frac{\partial Q(\Psi;\hat{\Psi}^{(k)},\bar{\varepsilon})}{\partial \sigma_i^2}|\bar{e}\right] = E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)\frac{1}{2\sigma_i^2}\left(\frac{(\bar{e}_a-\bar{\varepsilon}_a-\mu_i)^2}{\sigma_i^2}-1\right)|\bar{e}_a\right].$$

The expressions for $\hat{\mu}_i^{(k+1)}$ and $\hat{\sigma}_i^{2(k+1)}$ presented in Estimation Method 1-A are readily obtained after setting the partial derivatives with respect to μ_i and σ_i^2 equal to zero and solving for μ_i and σ_i^2. Finding the value for π_i that maximizes $Q(\Psi;\hat{\Psi}^{(k)})$ is a Lagrange optimization problem that incorporates the constraint that $\sum_i \pi_i = 1$. The corresponding FOC solves:

$$E\left[\sum_a \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)\frac{1}{\pi_i}|\bar{e}_a\right] = \lambda, \tag{57}$$

where λ denotes the Lagrange multiplier. Rearranging terms and summing both sides over i yields:

$$E\left[\sum_a\sum_i \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a)|\bar{e}_a\right] = \sum_i \lambda\pi_i. \tag{58}$$

Since $\sum_i \hat{\tau}_{ai}^{(k)}(\bar{\varepsilon}_a) = \sum_i \pi_i = 1$ by definition, we are left with:

$$\lambda = \sum_a 1 = A. \tag{59}$$

By substituting this into eq. (57), and solving this for π_i, we obtain the desired expression for $\hat{\pi}_i^{(k+1)}$.

References

Araujo, M., Ferreira, F., Lanjouw, P. and Ozler, B. (2008). Local inequality and project choice: Theory and evidence from ecuador. *Journal of Public Economics*, **92,** 1022–1046.

Cordy, C. and Thomas, D. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association*, **92,** 1459–1465.

Demombynes, G. and Ozler, B. (2005). Crime and local inequality in south africa. *Journal of Development Economics*, **76,** 265–292.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Serie* B, **39,** 1–38.

Elbers, C., Fujii, T., Lanjouw, P., Ozler, B. and Yin, W. (2007). Poverty alleviation through geographic targeting: How much does disaggregation help? *Journal of Development Economics*, **83,** 198–213.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, **71,** 355–364.

Fujii, T. (2010). Micro-level estimation of child undernutrition indicators in cambodia. *The World Bank Economic Review*, **24,** 520–553.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, **9,** 226–253.

Kleiber, C. (2007). A guide to the dagum distributions. *WWZ Working Paper*, **23/07**.

McCulloch, C. and Neuhaus, J. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, **26,** 388–402.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Model*. New York: John Wiley & Sons.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, **38,** 369–385.

Searle, S., Casella, G. and McCulloch, C. (1992). *Variance components*. New York: Wiley.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society*, **172,** 659–687.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association*, **91,** 217–221.

Wang, S. and Yin, S. (2002). A new estimate of the parameters in linear mixed models. *Science in China Series A: Mathematics*, **45,** 1301–1311.

Westfall, P. (1987). A comparison of variance component estimates for arbitrary underlying distributions. *Journal of the American Statistical Association*, **82,** 866–874.

Wu, M., Yu, K. and Liu, A. (2009). Estimation of variance components in the mixed effects models: A comparison between analysis of variance and spectral decomposition. *Journal of Statistical Planning and Inference*, **139,** 3962–3973.

Zhou, T. and He, X. (2008). Three-step estimation in linear mixed models with skew-t distributions. *Journal of Statistical Planning and Inference*, **138,** 1542–1555.