Estimating King’s ecological inference
normal model via the EM Algorithm

Rogério Silva de Mattos
Universidade Federal de Juiz de Fora
Faculdade de Economia e Administração
Juiz de Fora, Brazil
rmattos@nutecnet.com.br

Álvaro Veiga
Universidade Católica do Rio de Janeiro
Departamento de Engenharia Elétrica
Rio de Janeiro, Brazil
alvf@ele.puc-rio.br

Recently, King (1997) introduced a new model for ecological inference (EI), based
on a truncated bivariate normal, which he estimates by maximum probability and
uses to simulate the predictive densities of the disaggregate data. This paper
reviews King’s model and its assumption of truncated normality, with the aim to
implement maximum probability estimation of his model and disaggregate data
prediction in an alternative fashion via the EM Algorithm. In addition, we
highlight and discuss important modeling issues related to the chance of non–
existence of maximum likelihood estimates, and to the degree that corrections for
this non–existence by means of suitably chosen priors are effective. At the end, a
Monte Carlo simulation study is run in order to compare the two approaches.

Keywords: ecological inference, disaggregate data, exponential families, truncated
normal; EM Algorithm.

1. Introduction

Ecological Inference (EI) research is concerned with developing techniques for the
recovery of information on individual behavior from aggregate data. As such, EI
techniques are suitable for a number of problems of disaggregate data estimation and
prediction that arise in a variety of areas from the social sciences. The typical problem
for which EI techniques were developed is associated to the estimation/prediction of
cells contents in a contingency or values tables, when it is known only the marginal
totals of the tables, say, the column and row sums.

In this paper, we present some findings of a study\(^2\) of ours on recent
contributions to EI that emerged from political methodology. We will be particularly

---

1 Prepared for presentation at the 2000 Midwest Political Science Association Meeting (April 27–30). The
paper is based on an ongoing research being developed under the Ph.D. thesis work from the first author
(http://www.ufjf.br/~rmattos), who acknowledges support from the PICDT–UFJF/CAPES Program and
from the Faculdade de Economia e Administração of the Universidade Federal de Juiz de Fora.

2 Our basic motivation to undertake this study are some important challenges, in large part associated to
the scarcity of disaggregate data, that some developing countries are facing nowadays to implement a
decentralized and more democratic system of public policy management. In the particular case of Brazil,
after the decline of a system of highly centralized decisions that marked the military dictatorship period,
it was launched a “municipalization” process, by means of which the responsibilities of policy decisions
regarding social development, formerly under control of the federal government, have been progressively
transferred to state and local authorities. A major challenge to efficient implementation of the new
decentralized system is the absence of sub–regional and local level data to guide public policy
management and planning at these levels.
interested here on the basic parametric model proposed by King (1997). Though innovative in various aspects, King’s model displays, as a major distinct feature from other EI models, the use of a truncated (over a rectangle in $R^2$) bivariate normal distribution to represent the disaggregate data generating process. For this reason, we will be calling it here as the EI normal model (EINM).

Even being a quite recent contribution, King’s method is now well known among political scientists and other social scientists who work with ecological data and models. Its launching had a sound impact on research by motivating a number of new EI studies, theoretical and applied (Cho, 1997; Rivers and Cho, 1997; Lewis, 1998; Penubarti and Schuessler, 1998, and King, Rosen and Tanner, 1999). However, it also provoked a good amount of controversy and debate (Cho, 1998; Freedman et al, 1998; and King, 1999). Among the various arguments set forth on the discussion, there is a complaint placed against King’s method regarding some lack of practical guidance, as long as the diagnostics and the methodological checklist proposed by this author are not taken to be effective. Disconsidering possible misundertstandings of King’s method, our view is that improvements in diagnostics and practical guidances for using King’s method, or of any new EI method, will develop from a deeper understanding of the intrinsic characteristics of the underlying statistical model.

The still preliminary results we present here, since our study is an ongoing research, runs in this direction. We studied a different approach to estimate King’s EINM based on the Expectation Maximization Algorithm (EMA). The EMA is an alternative optimization technique for maximizing likelihood/posterior functions in incomplete data problems of estimation. We are not proposing a different EI model, just an alternative way to implement the same maximum likelihood/posterior estimation of the EINM parameters as King (1997) does, though with a slightly different approach to predict the disaggregate data.

An important consequence of our efforts was that we had to work with the likelihood function for the unobservable disaggregate data (the complete data in the EMA formalism), which is an essential ingredient for the EMA to undertake the maximization of the aggregate (incomplete) data likelihood (or posterior if a prior is combined). For the new (complete data) likelihood function, we can state in precise terms its statistical properties (like the conditions for existence and non–existence of a stationary point and for a unique maximum), since, as will be shown here, it belongs to a regular exponential family of probability densities. We (and King) were not so lucky with regard to the likelihood function for the aggregate (incomplete) data, which does not belong to an exponential family.

The rest of this paper is organized as follows: In section 2, we state the EI problem and introduce some notation; in Section 3, the EINM is briefly described for reference purposes; in Section 4, the EM Algorithm and its advantages for regular

---

3 In addition, we shall refer to King’s EI method as the set formed by the EINM, its parameter estimation and its disaggregate data prediction approaches, the diagnostic procedures and the methodological checklist, which are all explained in King (1997).

4 King(1997, p. 311) says: “There is usually little uncertainty about convergence, which in my experience occurs almost every time. The main exceptions I find are artificially generated data sets that the model does not remotely fit”.
exponential family cases is reviewed; in Section 5, the representation of the EINM under the EM Algorithm formalism is introduced; in Section 6, properties of the TBNR and the possibility of non--existence of a solution to the complete data likelihood equations are discussed; in Section 7, corrections for non--existence are considered; in Section 8, a comparison of methods by Monte Carlo simulation is presented; and finally, in Section 9, concluding comments are made.

2 The EI problem
In technical terms, the EI problem represents a situation where the analyst/planner is interested in cell data for one or more contingency tables (or values tables), but he/she knows only the row and column totals of the table(s). These totals are called the aggregate (or ecological) data. Analyst's goal is to determine the contents of tables cells. Table 1 depicts this situation for the simplest case of a 2×2 tables problem.

Table 1
Representation of the EI problem for the 2×2 tables case

<table>
<thead>
<tr>
<th>Variable I</th>
<th>Variable II</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$B_i$</td>
<td>$1-B_i$</td>
</tr>
<tr>
<td>2</td>
<td>$W_i$</td>
<td>$1-W_i$</td>
</tr>
<tr>
<td>Totals</td>
<td>$T_i$</td>
<td>$1-T_i$</td>
</tr>
</tbody>
</table>

where:

$B_i$ = proportion of 1st category of variable II in 1st category of variable I;
$W_i$ = proportion of 1st category of variable II in 2nd category of variable I;
$X_i$ = proportion of 1st category of variable I in the total of its two categories;
$T_i$ = proportion of 1st category of variable II in the total of its two categories.

Variables in Table 1 are defined as proportions, instead of absolute values, because it allows for a direct interpretation of results. For instance, in voting behavior studies, variable I might be race, e.g. blacks and whites, and Variable II might be partisan candidate, e.g. Republican or Democrat; thus, $B_i$ would represent proportion of blacks, and $W_i$ proportion of whites, voting for the Republican candidate. By their turn, $X_i$ would be the proportion of blacks, and $T_i$ the proportion of people voting for the Republican candidate, both in total turnout of voting age population.

The notation in Table 1 is general and applicable to a variety of contexts: In economics, variable I might be levels of family income, and variable II number of goods
purchased; in sociology, variable I might be number of crimes by city regions, and variable II number of crimes by type; in transport planning, variable I might be number of residents by residential colonies, and variable II number of jobs by trade areas. Thus, EI techniques are suited for a wide range problems where the lack of disaggregate information is a major drawback. EI research is also concerned with developing techniques for $R \times C$ tables problems, though implementation of such techniques are still limited with this respect (King, 1999).

In Table 1, $T_i$ and $X_i$ represent the known aggregate data. Subscript $i$ indexes the tables, or sample units, ranging from 1 to $P$ – the number of tables used in the EI analysis. By their turn, $B_i$ and $W_i$ represent the unknown disaggregate data and the target of EI problem solving; for this reason, $B_i$ and $W_i$ are called the quantities of interest. Once they become known, the contents of all cells in all tables also become known. Figure 1 illustrates this distinct feature of EI techniques, as compared to others that allow for just one table at a time.

![Figure 1: The use of various tables. Adapted from Mattos and Veiga (1999).](image)

The EI problem is solved in such a way that the cells contents in all tables are determined simultaneously. The proportions appearing outside the $2 \times 2$ tables of Figure 1, say, the pairs $(T_i, X_i) : i = 1, \ldots, P$, are known aggregate data used to estimate each quantity of interest in each table. Various tables allows the use of more observations (each table is a sample unit), and for "borrowing strength" from the information in other tables, when cells values of a particular table are estimated. This latter aspect may bring efficiency to estimation, if it happens that all tables have "something in common" (King, 1997). In practice, not always exists such a communality, at least among all the $P$ tables, and King's model admits extensions that allows for different mean patterns of the quantities of interest in different tables.
3 King's EI normal model

In order to predict the quantities of interest, King (1997) proposed a new method that makes extensive use of the available information in the EI problem. Indeed, King’s EINM is a statistical model that combines deterministic features of the EI problem with mathematical and probabilistic assumptions. For purposes of further reference, we described it briefly in this section.

3.1 Model Features

The deterministic features of the EINM consist the following facts:

a) **Accounting identity**: \( T_i = B_i X_i + W_i (1 - X_i) \) is a (everywhere) true relationship between the two aggregate \( T_i \) and \( X_i \) variables, mediated by the disaggregate \( B_i \) and \( W_i \) ones, that is valid for any table like Table 1;

b) **Cells bounds**: given the aggregate data, the quantities of interest may belong to narrower intervals than \([0,1]\) in a way that is easy to be computed; it is important for it may promote substantial reduction in the uncertainty regarding the prediction of the cells values. Using the notations \( L \) and \( U \) to denote lower and upper bounds, respectively, for a quantity of interest, we have that \( B_i \in [L_i^b, U_i^b] \) and \( W_i \in [L_i^w, U_i^w] \), where the bounds of these intervals are computed as:\(^5\)

\[
L_i^b = \max \left( 0, \frac{T_i - (1 - X_i)}{X_i} \right) \geq 0 \quad L_i^w = \max \left( 0, \frac{T_i - X_i}{X_i} \right) \geq 0
\]

\[
U_i^b = \min \left( \frac{T_i}{X_i}, 1 \right) \leq 1 \quad U_i^w = \min \left( \frac{T_i}{1 - X_i}, 1 \right) \leq 1
\]

(3)

For more details and examples, see Duncan and Davis (1953), who introduced the method of bounds in the EI literature, Anchen and Schively (1995, pp. 190-193), King (1997) and Mattos and Veiga (1999).

By their turn, the probabilistic features of the EINM consist of the following assumptions:

1. **Non–stochastic regressors**: \( X_i; \ i = 1,\ldots,P \) are non–random deterministic variables;

2. **Truncated normality**: \((B_i, W_i)^T; i = 1,\ldots,P\), follows a bivariate normal distribution truncated on the unit square \([0,1]\times[0,1]\) \( \in \mathbb{R}^2 \), as:

\[
f_{bw}(b_i, w_i \mid \psi) = \frac{N_{bw}(b_i, w_i \mid \psi)}{R(\psi)} \quad i = 1,\ldots,P
\]

(4)

where \( \psi = [\bar{\mu}_b, \bar{\mu}_w, \sigma^2_b, \sigma^2_w, \rho]^T \) and:

---

\(^5\) The formulas for the accounting identity in a) and for the method of bounds in (3) formulas can be generalized to \( R \times C \) tables (King, 1997, Chapter 15).
\[
R(\psi) = \int_0^1 \int_0^1 N_{bw}(b_i, w_i | \psi) db_i dw_i
\]

is a normalizing factor that assures \( f_{bw} \) integrates to one (the symbol "∪" over the parameters in (5) is used to indicate they are from an untruncated distribution, which was truncated to produce (5));

3. **Constant means:** \((\mu_b, \mu_w)^T : i = 1, ..., P\) do not depend functionally on the regressors \(X_i\) and \(1 - X_i\); (here the notations \(\mu_b\) and \(\mu_w\) are not covered by a "∪" symbol to indicate they are means of the truncated variables);

4. **Spatial independence:** The conditional random variable \(T_i \mid X_i\) is independent across different tables or sample units.

### 3.2 Estimation and Prediction

From the facts and assumptions above, King (1997) developed an implementation of the EINM that works in two stages: First, model parameters are estimated; and second, the estimated model is used to generate point and interval predictions of the disaggregate data variables.

1. **Parameters estimation:** The likelihood function based on the observed aggregate data \(t = [t_1, ..., t_P]^T\) is given by:

\[
L(\bar{\psi} \mid t) = \prod_{i=1}^{P} \frac{S(\bar{\psi}, t_i, x_i)}{R(\bar{\psi})} N_{i} \left( \mu_{i}(\bar{\psi}, x_i), \sigma_{i}^2(\bar{\psi}, x_i) \right) \quad t_i \in [0,1]
\]

where \(N_{i}\) is the untruncated normal density of the aggregate data variable \(T_i\), \(\mu_{i}\) and \(\sigma_{i}^2\) are both functions of \(\bar{\psi}\) and \(x_i\); \(S(\bar{\psi}, t_i, x_i)\) is the normalizing factor for \(f_{bw}(b_i \mid t_i, \bar{\psi})\) (which is a doubly truncated normal density); and \(R(\bar{\psi})\) is the TBN factor defined in (5). However, a reparametIZATION \(\phi = c(\bar{\psi})\) is adopted, where \(\phi = [\phi_1, \phi_2, \phi_3, \phi_4, \phi_5]^T\) is such that:

\[
\begin{align*}
\phi_1 &= \frac{\mu_b - 0.5}{\sigma_b^2 + 0.25} \quad -\infty < \phi_1 < \infty \\
\phi_2 &= \frac{\mu_w - 0.5}{\sigma_w^2 + 0.25} \quad -\infty < \phi_2 < \infty \\
\phi_3 &= \ln \sigma_b \quad -\infty < \phi_3 < \infty \\
\phi_4 &= \ln \sigma_w \quad -\infty < \phi_4 < \infty \\
\phi_5 &= 0.5 \ln \left( \frac{1 + \bar{\rho}}{1 - \bar{\rho}} \right) \quad -\infty < \phi_5 < \infty
\end{align*}
\]

In addition, the likelihood is combined with a bayesian prior for \(\phi\), what in turn leads to a posterior function:

\[
p(\phi \mid t) \propto p(\phi) L(c^{-1}(\phi) \mid t)
\]
where \( p(\phi) \) is a prior density for \( \phi \). Then, parameter estimates \( \hat{\phi} \) are obtained by maximizing (12) for \( \phi \). This maximization is carried out through an iterative search algorithm, because the posterior (12) (and the likelihood) is implicitly non linear in \( \phi \). So, the produced estimates are:

\[
\hat{\phi} = \arg \max_{\phi} p(\phi \mid t)
\]

(13)

2. **Disaggregate data prediction**: Given the estimated parameter vectors \( \hat{\phi} \) and \( \hat{\theta} = c^{-1}(\hat{\phi}) \), the conditional (on \( t \) and \( \hat{\phi} \)) distributions \( p(b_i \mid t, \hat{\phi}) \) and a normal approximation to the posterior function \( p(\hat{\phi} \mid t) \) are used in composition to simulate the conditional (on \( t \) only) predictive densities \( p_b(b_i \mid t_i) \) and \( p_w(w_i \mid t_i) \) for the disaggregate data variables; then, point and interval predictions for \( B_i \) and \( W_i \) are produced by computing the mean and the standard deviations using simulated values for those predictive densities, as follows:

\[
\hat{B}_i = \frac{1}{K} \sum_{k=1}^{K} \tilde{B}_{i(k)} \\
S_{\hat{B}_i} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\tilde{B}_{i(k)} - \hat{B}_i)^2}
\]

(14)

\[
\hat{W}_i = \frac{1}{K} \sum_{k=1}^{K} \tilde{W}_{i(k)} \\
S_{\hat{W}_i} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\tilde{W}_{i(k)} - \hat{W}_i)^2}
\]

(15)

where \( \tilde{B}_{i(k)} \) and \( \tilde{W}_{i(k)} \) are the simulated values for \( B_i \) and \( W_i \).

An extended version of the EINM presented above was also developed by King (1997) that allows for the effects of explanatory variables. Since we will be discussing here an implementation only for the basic (with no explanatory variables) version of the EINM, we refer the reader to King (1997, Chapter 9) for more details on the extended version.

### 4 The EM Algorithm

The Expectation–Maximization Algorithm (EMA) is an optimization device for the maximization of likelihood/posterior functions in incomplete data problems. In principle, it is no more than an alternative to other optimization algorithms (like Newton Hapson and Quasi–Newton based algorithms) specifically designed for finding modes of incomplete data likelihoods/posteriors. The principles of the EMA have been applied in statistical analyses for a long time, but it was after the seminal paper from Dempster, Laird and Rubin (1977), in short DLR, that the EMA saw a widespread of its use. DLR introduced a structured formalism to apply the methodology, proved its mains useful properties, and gave it the name by which it is well known today.

Since DLR’s paper, the EMA theory has evolved in important aspects. For instance, its formalism has been exploited as a data augmentation technique (Tanner,
1996) and applied beyond specific incomplete/missing data environments. Also, a variety of extensions like the Monte Carlo EM and the Expected Conditional Maximization (ECM) algorithms, together with other practical improvements for speeding up the convergence of the EMA, and for computing the variance–covariance matrix of estimated parameters were developed. Comprehensive reviews of such developments can be found in a recent book from McLahlan and Krishnan (1997), and in Meng and van Dyk (1997), the last being a celebrating paper for the 20th anniversary of DLR’s paper.

The EMA can be useful to tackle the EI problem, since it can be suitably described as an incomplete data problem. In particular, the EMA is of value for the estimation of EI models because, in addition to providing estimates of models parameters, it also produces predictions of the quantities of interest. And also, it induces the exploitation of the modeling assumptions regarding the disaggregate data generating process, which can increase comprehension of intrinsic features of an EI model.

4.1 EMA concepts and functioning

EMA concepts are easy to be understood. In briefly explaining them here, we will take an incomplete/missing data perspective (and not the general data augmentation) point of view. Key concepts of the EM Algorithm formalism are those of complete and incomplete data. Sometimes, we are interested on a random variable for which observations are unavailable or taking measurements is impossible. It represents a variety of situations: For instance, a time series missing data for some periods; a survey research database with non-responded entries, or a set of variables for which we known only their sum. In a word, the complete data can be viewed as the sample information we should have to estimate the parameters of a model; actually, however, we cannot observe that. By its turn, the incomplete data is an associated sample we can observe, but whose information content for parameter estimation is lower than that of the complete data.

Implicit in the relation between the two forms of the data there is a many–to–one mapping linking them, in the sense that usually there are lots, even an infinity, of (unobservable) complete data samples associated to each single (observable) sample of incomplete data. In the EI problem, for instance, we observe a single sample of aggregate data (row and column totals for a number of tables); but, we know there may be many, indeed an infinity, of samples of unobserved dissaggregate data (cells contents) consistent with our (unique) sample of observed aggregate data.

In a more formal fashion, suppose we are interested in a phenomenon described by a continuous (it might be also a discrete) random variable $X$ that follows a known probabilistic model $f_X(x^* | \theta)$, with $x^*$ an observation from $X$ and $\theta$ a vector of

---

6 The data augmentation view of the EMA is a broader perspective taken on this technique. As stressed by Martin Tanner: “In the EM [algorithm], the data analyst augments the observed data with latent data to simplify the computations in the analysis. These latent data may be ‘missing’ data (…), parameter values (…) and values of sufficient statistics.” (Tanner, 1996: p.2). We believe that the particular incomplete/missing data perspective is the most sound way to look at the EI problem, since what we are searching, ultimately, is for the unobserved (thus, not available) disaggregate data.
unknown parameters. Suppose, in addition, that we cannot observe \( X \); however, we can write its likelihood function \( f(x | \theta) \), where \( x = [x_1^*, ..., x_p^*]^\top \) represents a non-observable random sample from \( X \). Vector \( x \) is called the complete data, and \( f(x | \theta) \) the likelihood of the complete data (LCD). Assume also that we can observe the sample vector \( y = [y_1, ..., y_Q]^\top \), which is related to \( x \) in a deterministic way such that: \( y = h(x) \) and \( Q < P \). Vector \( y \) is called the incomplete data. Now, let \( \Xi \) be the sample space for \( x \), and \( \Upsilon \) the sample space for \( y \). Since \( Q < P \), it is easy to have many \( x \in \Xi \) associated to each \( y \in \Upsilon \), so that \( h: \Xi \rightarrow \Upsilon \) is a many-to-one mapping. Say, a given single point \( y \in \Upsilon \) has associated to it a subset of \( \Xi \) — namely \( \Xi(y) \), the inverse image of \( y \) under the mapping \( h \) — containing many points \( x \in \Xi \). This situation is illustrated in Figure 3.

![Figure 3. Many–to–one mapping](image)

From the above, we are allowed to write the likelihood of the incomplete data (LID) as:

\[
g(y | \theta) = \int_{\Xi(y)} f(x | \theta) dx
\]  

(16)

What the EM Algorithm does is to find estimates of \( \theta \) by using the incomplete data \( y \) and the form of the LCD to indirectly maximize the LID. It is not undertaken at once, but in a sequence of iterations, where, in each iteration, two steps are performed: The Expectation Step (E-Step), and the Maximization Step (M-Step). The general EM scheme is as follows:

**S1.** Assume a guess for \( \theta \), say \( \theta_k \).

**S2.** E-Step: Use \( \theta_k \) and the observed complete data \( y \) to compute:

\[
Q(\theta, \theta_k) = E[\log f(x | \theta) | \theta_k, y] = \int_{\Xi(y)} \log f(x | \theta) \times f(x | \theta_k, y) dx
\]

(17)

**S3.** M-Step: Maximize \( Q(\theta, \theta_k) \) for \( \theta \), finding:

\[
\hat{\theta} = \arg \max_{\theta} Q(\theta, \theta_k)
\]

(18)

**S4.** Set \( \theta_{k+1} = \hat{\theta} \).

**S5.** Repeat steps S1–S4 a number of times until convergence, say, until:
\[ \left| \log g(y \mid \theta_{k+1}) - \log g(y \mid \theta_k) \right| = 0 \]  

(19)

and/or:

\[ \| \theta_{k+1} - \theta_k \| = 0 \]  

(20)

The EM Algorithm scheme S1-S5 presents good convergence properties: For instance, provided that \( Q(\theta_{k+1}, \theta_k) \geq Q(\theta_k, \theta_k) \), as is guaranteed in the M–Step, the LID never decreases in each iteration; and if, in addition, the LID is bounded above, the EMA will always converge to a stationary point. We can also use, in place of the LCD and the LID defined above, the posterior for \( \theta \) based on the complete data (PCD), written as \( p(\theta \mid x) \approx p(\theta) f(x \mid \theta) \) and the posterior for \( \theta \) based on the incomplete data (PID), written as \( p(\theta \mid y) \approx p(\theta) g(y \mid \theta) \). For a detailed treatment of the EM Algorithm in Bayesian analysis, see Gelman et al (1995, Chapters 7 and 9).

### 4.2 EMA for exponential families

The most general way to implement the EMA, say, for any probability model, involves the direct computation of the Q–function using the definition in (17). If the Q–function is easy to be computed in the E–Step and there exists a closed form solution for the maximization undertaken in the M–Step, then the EMA can be attractive (DLR, 1977). Otherwise, the opposite may be true. Particular instances where the EMA tends to be simpler and attractive occurs when the LCD is from a regular exponential family of probability densities. In such a case, we can write the LCD as:

\[ f(x \mid \theta) = b(x) \frac{\exp(\theta^T z(x))}{a(\theta)} \]  

(21)

where \( x \in X; \ \theta \in \Theta \subseteq \mathbb{R}^d; \ b(x) \geq 0 \) is some real valued function of \( x \); \( z(x) \) is a minimal vector of jointly sufficient statistics for the \( d \)–dimensional canonical parameter vector \( \theta \), and each element of \( z(x) \) is a real valued function of \( x \). Consider also the parameter space \( \Theta \subseteq \mathbb{R}^d \) is a convex open set (Barndorff–Nielsen, 1978). By taking the conditional expected log on both sides of (21), we find:

\[ Q(\theta, \theta_k) = E[\log b(X) \mid y, \theta_k] + \theta^T E[z(X) \mid y, \theta_k] - \log a(\theta) \]  

(22)

Now, by differentiating (22) with respect to \( \theta \), equating the result to 0 and making a few algebraic manipulations, we arrive at the following modified\(^7\) system of likelihood equations:

\[ E[z(X) \mid \theta] = E[z(X) \mid y, \theta_k] \]  

(23)

Expression (23) implies in a great simplification of the EMA. In order to maximize the Q–function at each iteration, we need only solve the system above by using, in the right hand side, the expected values of the (complete data based) sufficient statistics \( z(x)\)

\(^7\) We use the term “modified” because in a true system of likelihood equations from a distribution of an exponential family, the right hand side is given by the observed sufficient statistics, and not by their conditional expected values, as is the case here.
conditioned on the observed incomplete data \( y \) and on the instantaneous parameters guess \( \theta_k \). For a number of regular exponential families, \( E[z(X) \mid y, \theta_k] \) is easy to compute and the solution to the equation system (23) exists in closed form. The consequence is that the E–Step and the M–Step of the EMA scheme S1–S5 may be substituted by the simpler steps:

**S.2** E–Step (exponential family) – Use \( \theta_k \) and \( y \) to compute:

\[
z_k = E[z(X) \mid y, \theta_k] \tag{24}
\]

**S.3** M–Step (exponential family) – Maximize \( Q(\theta, \theta_k) \) for \( \theta \), by solving the equation system:

\[
E[z(X) \mid \theta] = z_k \tag{25}
\]

Even in some cases where no closed form exists for the equations system in (23) (or (25)), iterative search algorithms may run fast in the M–Step due to the well known property of strict convexity of minus the logarithm of a likelihood function belonging to a regular exponential family (Barndorff–Nielsen, 1978).

All we discussed above regarding likelihoods is valid also for posterior functions. But, in spite of the simplicity gained if the PCD or the LCD belongs to a regular exponential family, some difficulties remain. As pointed by DLR (1977; p.4), the equations system (23) does not always have a solution for \( \theta \) in \( \Theta \); in these cases, the maximand vector \( \hat{\theta} \) stays somewhere at the boundary of \( \Theta \) (or may not converge at all). However, provided that a solution exists for \( \theta \) in \( \Theta \), this must be unique due to the strict convexity property mentioned in the last paragraph.

## 5 The EINM under the EMA formalism

A major reason making the EMA attractive for EI is that it produces, jointly with the parameter estimates, predictions for the disaggregate data variables. According to the EMA scheme S1-S5, it is done in every E-Step performed in every iteration. Of interest are those complete data estimates (which correspond to the predictions of the disaggregate data variables) generated at the last EMA iteration, or after convergence has been achieved. By this property, a path is open to the use of the EMA with EI models. For instance, to write King’s EINM under the EM formalism, we may assume:

1. Complete (disaggregate) data vectors: \( b = [b_1, ..., b_p]^T \) and \( w = [w_1, ..., w_p]^T \);
2. Incomplete (aggregate) data vector: \( t = [t_1, ..., t_p]^T \);
3. Parameter vector: \( \psi = [\mu_b, \mu_w, \sigma_b^2, \sigma_w^2, \rho]^T \) (or \( \phi = [\phi_1, \phi_2, \phi_3, \phi_4, \phi_5]^T \);
4. Many to one mapping: \( t = h(b, w) = Xb + [I_p - X]w \); where \( I_p \) is the \( P \times P \) identity matrix, and \( X = \text{diag}([x_1, ..., x_p]^T) \);
5. LCD: \( f_{CD}(b, w \mid \psi) = \prod_{i=1}^{P} f_{bw}(b_i, w_i \mid \psi) \);

---

8 Note the many-to-one mapping is given by the accounting identity, presented in a generalized form in item 4.
6. LID: $f_{ID}(t \mid \psi) = \prod_{i=1}^{p} f_{t_i}(t_i \mid \psi)$;
7. PCD: $p_{CD}(\psi \mid b, w) \propto p(\psi) f_{CD}(b, w \mid \psi)$;
8. PID: $p_{ID}(\psi \mid t) \propto p(\psi) f_{ID}(t \mid \psi)$;
9. $Q$-Function: $Q(\psi, \psi_k) = E_{\beta} \left[ \log p_{CD}(\psi \mid b, w) \mid t, \psi_k \right]$

### 5.1 Using properties of the exponential family

A central feature of the EINM is that the disaggregate data variables follow a TBN with support over the unit square. Such a TBN is a particular case of a bivariate normal truncated over a rectangular region (TBNR) in Euclidean Space. Since a TBNR belongs to a regular exponential family, the joint density $f_{CD}(b, w \mid \psi)$, which is a product of TBNRs, also belongs to a regular exponential family. Indeed, it is easy to check that this last function can be written as:

$$f_{CD}(b, w \mid \psi) = q(b, w) \frac{\exp(c(b, w, \psi))}{a(\psi)}$$  \hspace{1cm} (26)

which is the exponential family representation for a multivariate density. Using the TBNR assumption, let us write it as:

$$f_{CD}(b, w \mid \psi) = \frac{\exp(g(b, w; \psi))}{(2\pi)^{p/2} |\Sigma(\psi)| |R(\psi)|}$$  \hspace{1cm} (27)

where:

$$g(b, w; \psi) = -\frac{\sum_{i=1}^{p} \left( \frac{(b_i - \mu_b)}{\sigma_b} \right)^2 + 2\rho \left( \frac{(b_i - \mu_b)}{\sigma_b} \right) \left( \frac{(w_i - \mu_w)}{\sigma_w} \right) + \left( \frac{(w_i - \mu_w)}{\sigma_w} \right)^2}{2(1 - \rho^2)}$$  \hspace{1cm} (28)

and note that $g(\cdot)$ can be written as:

$$g(b, w; \psi) = c(b, w; \psi) - k(\psi)$$  \hspace{1cm} (29)

where:

$$c(b, w; \psi) = \frac{\mu_b \sigma_w^2 - \mu_w \rho \sigma_b \sigma_w}{|\Sigma(\psi)|} \sum_{i=1}^{p} b_i + \frac{\mu_w \sigma_b^2 - \mu_b \rho \sigma_b \sigma_w}{|\Sigma(\psi)|} \sum_{i=1}^{p} w_i +$$
$$\frac{-\sigma_b^2}{|\Sigma(\psi)|} \sum_{i=1}^{p} b_i^2 + \frac{-\sigma_w^2}{|\Sigma(\psi)|} \sum_{i=1}^{p} w_i^2 + \frac{\rho \sigma_b \sigma_w}{|\Sigma(\psi)|} \sum_{i=1}^{p} b_i w_i$$

$$k(\psi) = \frac{\mu_b^2 \sigma_w^2 - 2\mu_b \mu_w \rho \sigma_b \sigma_w + \mu_w^2 \sigma_b^2}{2 |\Sigma(\psi)|}$$  \hspace{1cm} (30)

Substituting (29) in (27), we find:

$$f_{CD}(b, w \mid \psi) = \frac{\exp(c(b, w; \psi))}{\exp k(\psi) (2\pi)^{p/2} |\Sigma(\psi)| |R(\psi)|}$$  \hspace{1cm} (32)
Now, by defining $q(b, w; \psi) = 1$ and $a(\psi) = \exp k(\psi) \times [2\pi \sqrt{\Sigma(\psi) R(\psi)}]^p$, then (32) and (26) are the same. Take into account that $c(b, w; \psi) = \gamma(\psi)^T z(b, w)$ we can go a step further and write (32) in a canonical form:

$$f_{CD}(b, w | \psi) = \frac{\exp \left( \gamma(\psi)^T z(b, w) \right)}{\exp k(\psi) \left[ 2\pi \sqrt{\Sigma(\psi) R(\psi)} \right]^p}$$  \hspace{1cm} (32')

In (32'), $z(b, w)$ is a 5–dimensional vector of minimal sufficient statistics for $\gamma = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5]^T$, where $\gamma = m(\psi)$ is a vector of canonical parameters and $m: \Psi \rightarrow Y$ is a bijection from the parameter space $\Psi$ (of $\psi$) to the canonical parameter space $Y$ (of $\gamma$). The vector of sufficient statistics is given by:

$$z(b, w) = \left[ \sum b_i, \sum w_i, \sum b_i^2, \sum w_i^2, \sum b_i w_i \right]^T$$  \hspace{1cm} (33)

and the canonical parameter vector $\gamma$:

$$\gamma_1 = \frac{\bar{\mu}_w - \mu_w \rho \bar{\sigma}_b \bar{\sigma}_w}{|\Sigma(\psi)|}$$  \hspace{1cm} (34)

$$\gamma_2 = \frac{\bar{\mu}_b - \mu_b \rho \bar{\sigma}_b \bar{\sigma}_w}{|\Sigma(\psi)|}$$  \hspace{1cm} (35)

$$\gamma_3 = \frac{-\bar{\sigma}_w^2}{2|\Sigma(\psi)|}$$  \hspace{1cm} (36)

$$\gamma_4 = \frac{-\bar{\sigma}_b^2}{2|\Sigma(\psi)|}$$  \hspace{1cm} (37)

$$\gamma_5 = \frac{\bar{\rho} \bar{\sigma}_b \bar{\sigma}_w}{|\Sigma(\psi)|}$$  \hspace{1cm} (38)

where:

$$|\Sigma(\psi)| = \bar{\sigma}_b^2 \bar{\sigma}_w^2 (1 - \rho^2)$$  \hspace{1cm} (39)

By its turn, the inverse transformation $\psi = m^{-1}(\gamma)$ is given by:

$$\bar{\mu}_b = \frac{|\Sigma(\gamma)|}{2(\gamma_1 - \gamma_4 + \gamma_2 \gamma_5)}$$  \hspace{1cm} (40)

$$\bar{\mu}_w = \frac{|\Sigma(\gamma)|}{2(\gamma_2 - \gamma_3) + \gamma_1 \gamma_5}$$  \hspace{1cm} (41)

$$\bar{\sigma}_b = \frac{|\Sigma(\gamma)|}{2(-\gamma_4)}$$  \hspace{1cm} (42)

$$\bar{\sigma}_w = \frac{|\Sigma(\gamma)|}{2(-\gamma_3)}$$  \hspace{1cm} (43)

$$\bar{\rho} = \frac{\gamma_5}{2\sqrt{\gamma_3 \gamma_4}}$$  \hspace{1cm} (44)

---

9 The range of admissible values for $\gamma_5$ needs to be restricted as in equation (38) because the range of $\bar{\rho}$ is restricted to the open interval $(-1, 1)$. See ahead equation (45).
where $|\Sigma(\gamma)| = \frac{1}{4\gamma_5\gamma_4 - \gamma_5^2}$ \quad 0 < |\Sigma(\gamma)| < \infty \quad (45)

At this stage, it would be useful to write the LCD as:

$\ell_{CD}(m^{-1}(\gamma) \mid b, w) = \gamma_1 \sum_{i=1}^{p} b_i + \gamma_2 \sum_{i=1}^{p} w_i + \gamma_3 \sum_{i=1}^{p} b_i^2 + \gamma_4 \sum_{i=1}^{p} w_i^2 + \gamma_5 \sum_{i=1}^{p} b_i w_i \quad (46)$

$- \log K(m^{-1}(\gamma))$

where $K(m^{-1}(\gamma)) = K(\bar{\psi}) = \exp k(\bar{\psi})[2\pi \sqrt{\Sigma(\bar{\psi})} R(\bar{\psi})]^p$ is the denominator of (32).

Note in the present case that the $Q$–function – which is given by $Q(\gamma, \gamma_k)$ and which would result from us taking the expectation of both sides of (46) conditional on the incomplete (disaggregate) data $t$ and a guess $\gamma_k$ – is the $\ell_{CD}$ written as a function of the conditional expected sufficient statistics (see below equations (52)–(56)). Thus, we are able to state in more precise terms the application of the EMA for estimating the EINM. This would involve:

**E–Step:** Compute the predictions for the sufficient statistics:

$\sum \hat{b}_i(t_i; \bar{\psi}_k) = \sum E(B_i \mid t_i, \bar{\psi}_k) = \sum \int_{B_i}^B b_i f_{B \mid t_i}(b_i \mid t_i, \bar{\psi}_k)db_i \quad (47)$

$\sum \hat{w}_i(t_i; \bar{\psi}_k) = \sum E(W_i \mid t_i, \bar{\psi}_k) = \sum \left( a_i t_i - c_i \hat{b}_i(t_i; \bar{\psi}_k) \right) \quad (48)$

$\sum \hat{b}_i^2(t_i; \bar{\psi}_k) = \sum E(B_i^2 \mid t_i, \bar{\psi}_k) = \sum \int_{B_i}^B b_i^2 f_{B \mid t_i}(b_i \mid t_i, \bar{\psi}_k)db_i \quad (49)$

$\sum \hat{w}_i^2(t_i; \bar{\psi}_k) = \sum E(W_i^2 \mid t_i, \bar{\psi}_k) = \sum \left( a_i^2 t_i^2 - 2c_i \hat{b}_i(t_i; \bar{\psi}_k) + c_i^2 \hat{b}_i^2(t_i; \bar{\psi}_k) \right) \quad (50)$

$\sum \hat{b}_i \hat{w}_i(t_i; \bar{\psi}_k) = \sum E(B_i W_i \mid t_i, \bar{\psi}_k) = \sum \left( a_i t_i \hat{b}_i(t_i; \bar{\psi}_k) - c_i \hat{b}_i^2(t_i; \bar{\psi}_k) \right) \quad (51)$

**M–Step:** Maximize the $Q$–function by solving for $\gamma$ the following system of equations:

$E(B_i; \gamma) = \sum \hat{b}_i / P = \bar{b} \quad (52)$

$E(W_i; \gamma) = \sum \hat{w}_i / P = \bar{w} \quad (53)$

$Var(B_i; \gamma) = \sum (\hat{b}_i - \bar{b})^2 / P = \tilde{s}_b^2 \quad (54)$

$Var(W_i; \gamma) = \sum (\hat{w}_i - \bar{w})^2 / P = \tilde{s}_w^2 \quad (55)$

$Cov(B_i, W_i; \gamma) = \sum (\hat{b}_i - \bar{b})(\hat{w}_i - \bar{w}) / P = s_{bw} \quad (56)$

The last $\hat{\gamma}_K$, obtained at the last iteration of the EMA, can be easily transformed back to the original parameterization, thus providing $\hat{\psi} = m^{-1}(\hat{\gamma}_K)$. As stated, implementing the EMA to estimate parameters and predict quantities of interest of the EINM seems to be straightforward. However, two difficulties remain: First, the equation system above is not linear in the unknown parameters and it makes necessary the use of an iterative search algorithm to obtain a solution. This may not at all be a problem in the
present case, because the strict convexity property of \(-\log LCD\) in the case of regular exponential families ensures that, provided a solution to the equations system exist, this solution is unique. However, this leads to our second difficulty. As pointed by DLR (1977), the equations system above may not have a solution for \(\gamma\) in \(\Gamma\). The question then is: Under what conditions there does (or does not) exist such a solution?

6 TBNRs and maximum likelihood estimation

In this section we provide a partial (tough sufficient) answer to the question posed in the last section. Studying maximum likelihood estimation for the parameters of a doubly truncated (univariate) normal (DTN), Mittal and Dahyia (1987) showed that MLEs\(^{10}\) may not exist and presented sufficient conditions. Later, Hedge and Dahyia (1989) presented necessary and sufficient conditions for that to hold. In both papers, the authors also proposed a solution or correcting procedure, about which we will talk in the next section. We will make use here of the intuition presented in the paper of Mittal and Dahyia (1987) and generalize the authors’ arguments to show why the MLEs may not exist also in the bivariate case.

Let us consider initially the univariate case of a DTN with support over the interval \([a, b] \subset R\), where \(a\) and \(b\) are known. Mittal and Dahyia (1987) worked on a property that likelihood equations from exponential families can be obtained by equating distributional moments with their sampling counterparts. Thus, maximum likelihood estimation of the DTN parameters are obtained by solving a system of equations where DTN’s first and second orders moments (which are functions of the unknown parameters) are made equal to their first and second order sampling moments. The truncation imposed on a normal distribution to take values on a bounded interval restrain the moments (distributional and sampling) of the DTN random variable to be bounded. For the DTN under consideration, the distribution mean cannot be lesser than \(a\) or higger than \(b\), the same being true with regard to the sample mean. A similar situation occurs with the distribution and sample variances, as truncation made them both bounded also. However, the intervals where each take values are not the same.

Let \(X\) be a random variable following a DTN with known truncation limits \(a\) and \(b\), \(E(X)\) the distribution mean, \(Var(X)\) the distribution variance, \(\bar{x}\) the sample mean and \(s^2\) the sample variance. It can be shown that:

\[
0 < Var(X) < (b-a)^2 / 12
\]

and:

\[
0 \leq s^2 \leq (b-a)^2 / 4
\]

The proof of these statements relies on simple facts. Let \(\mu\) be the mean and \(\sigma^2\) the variance of the original untruncated normal, associated to the DTN in question. In limiting situations, say, when \(|\mu| \rightarrow \infty\) and \(\sigma^2 \rightarrow \infty\), two things may happen:

a) the DTN of \(X\) degenerate either on \(a\), or on \(b\) (when \(|\mu / \sigma^2| \rightarrow \infty\)); and

\(^{10}\) We will be using “MLEs” here to refer either to maximum likelihood estimators or to maximum likelihood estimates.
b) the DTN of $X$ converges to a uniform distribution, such that, in the limit, $E(X) = (b + a)/2$ and $Var(X) = (b - a)^2 / 12$ (when $\mu / \sigma^2 \to 0$).

These facts\(^{11}\) mean that the first likelihood equation $E(X) = \bar{x}$ will always have a solution, since both $E(X)$ and $\bar{x} \in [a, b]$; but, it is not true for the second likelihood equation, because $\text{sup}(Var(X)) = (b - a)^2 / 12$ while $\text{sup}(s^2) = (b - a)^2 / 4$. Thus, the supremum value for the latter is higher, and $s^2$ will attain it in the extreme case when half of the sample observations is equal to $a$ and the other half is equal to $b$ (no loss of generality in assuming the sample size is even). In summary, a MLE for the unknown parameters of a DTN will not exist whenever:

$$\frac{(b - a)^2}{12} \leq s^2 \leq \frac{(b - a)^2}{4}$$

(59)

We shall stress that (59) is just a sufficient condition for non-existence. For strictly necessary and sufficient conditions in a DTN case, see Hedge and Dahyia (1989).

Let us turn now to our particular bivariate case associated to the TBNR of the EINM. With analogous arguments, we can determine for this case similar sufficient conditions for non-existence of the MLEs for $\gamma$ (and $\psi$) in $\Gamma$ (and $\Psi$). Consider the LCD equations (52)–(56). They form a $5$–equations system we have to solve for $\gamma = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5]^T$ in $\Gamma$ at the M–Step. Both sides of those equations are bounded on finite intervals. The first two equations, namely (52) and (53), equates the distribution and sampling means of $B_i$ and $W_i$, respectively. For each of these random variables, their distribution and sampling means are constrained to take values on the $[0, 1]$ interval. As in the univariate case, there is no problem for existence of a solution in $\Gamma$ that satisfy these equations. However, with regard to each of the last three equations, namely (54), (55) and (56), their left–hand sides are differently bounded from their right–hand sides. It can be proved that\(^{12}\):

\begin{align}
1. & \begin{cases} 0 < \text{Var}(B_i; \gamma) < 1/12 \\ 0 \leq s^2_b \leq 1/4 
\end{cases} \\
2. & \begin{cases} 0 < \text{Var}(W_i; \gamma) < 1/12 \\ 0 \leq s^2_w \leq 1/4 
\end{cases} \\
3. & \begin{cases} -1/12 < \text{Cov}(B_i, W_i; \gamma) < 1/12 \\ -1/4 \leq s_{bw} \leq 1/14 
\end{cases}
\end{align}

(60) (61) (62)

Again, the proof of the above statements is based on the examination of limiting situations. When $|\bar{\mu}_b| \to \infty$, $|\bar{\mu}_w| \to \infty$, $\sigma^2_b \to \infty$, $\sigma^2_w \to \infty$, and $|\bar{\rho}| \to 1$, the following things may happen:

\(^{11}\) Mittal and Dahyia (1987) points to a third extreme situation, say, when $\mu / \sigma^2 \to c$, where $c$ is a positive or negative constant. In this case, the DTN converges to an exponential distribution, whose variance is even less than $(b - a)^2 / 12$.

\(^{12}\) From now on we will suppress the hat “$\hat{\phantom{a}}$” in the notation for the sampling moments of the TBNR, for the results discussed here are valid in general, say, when those moments are calculated from true samples and not from predicted ones.
a) the TBNR becomes *degenerated on a boundary point* of the unit square \([0,1] \times [0,1]\), so that \(\text{Var}(B_i) = \text{Var}(W_i) = 0\) and, as a consequence, \(\text{Cov}(B_i, W_i) = 0\) (this will happen if in the limiting situation we also have \(\mu_b^2 / \sigma_b^2 \to \infty\) and \(\mu_b^2 / \sigma_b^2 \to \infty\) or simply \(|\rho| \to 1\)) or;

b) the TBNR converges to a *bivariate uniform distribution*, with means \(E(B_i) = E(W_i) = 1/2\) and variances \(\text{Var}(B_i) = \text{Var}(W_i) = 1/12\), and covariance \(\text{Cov}(B_i, W_i) = 1/12\) or \(\text{Cov}(B_i, W_i) = -1/12\) (this will happen, for instance, when \(\mu_b^2 / \sigma_b^2 \to 0\) and \(\mu_b^2 / \sigma_b^2 \to 0\) sufficiently faster than \(|\rho| \to 1\))\(^{13}\).

A consequence of fact b) is that equations (54), (55) and (56) may not be solvable. The sampling variances will attain \(\sup s_b = \sup s_w = 1/4\) if and only if the sample covariance attains \(\sup |s_{bw}| = 1/4\). All of this will happen in the extreme case where half of the sample observations is equal to 0, the other half is equal to 1 (no loss of generality in assuming the sample size is even) and all observations stay on a same (positively or negatively) sloped line crossing the unit square, as illustrated in Figure 4.

**Figure 4.** Illustration of minimum and maximum sample correlations. The black circles represent clustering of half of sample observations. The positively sloped line indicates a situation with maximum attainable sample correlation, and the negatively sloped line a situation with minimum attainable sample correlation.

In summary, a MLE for the unknown parameters of a TBNR will not exist whenever at least one of the following conditions hold:

\(^{13}\) Note in equation (28) that the exponent of \(e\) is \(g(b_i, w_i; \psi) \in (-\infty, 0]\). In extreme cases (derived from limiting situations), \(g(b_i, w_i; \psi) \to -\infty\) and the TBNR may degenerate, or \(g(b_i, w_i; \psi) \to 0\) and the TBNR converges to a bivariate uniform distribution over the unit square.
Again, we stress these are just sufficient conditions for the non–existence of a MLE.

7 Alternative estimator

Mittal and Dahyia (1987) and Hedge and Dahyia (1989) proposed a solution to the problem discussed in the previous section for the univariate case. Their solution consists of maximizing a posterior function, instead of the likelihood, using a chi–square prior for the inverse of the variance parameters of the DTN. Such a posterior gives rise to a transformed system of likelihood equations which has a solution in \( \Gamma \) with probability one. Thus, they propose an alternative (maximum probability) estimator to the MLE.

We extend here their solution to the bivariate case. As we are discussing solutions to equations (52)–(56) in the natural parameter space \( Y \), we adopted chi–square priors with \( v \) degrees of freedom for the negative values of the canonical parameters \( \gamma_3 \) and \( \gamma_4 \):

\[
\begin{align*}
    p_{-\gamma_j}(-\gamma_j) &= c(v)(-\gamma_j)^{v-1} \exp\left(\frac{\gamma_j}{2}\right) -\gamma_j > 0 \quad j = 3,4 \\
    p_{\gamma_i}(\gamma_3) &= \left(\sqrt{2\pi h^2}\right)\exp\left(-\frac{\gamma_3^2}{2h^2}\right) -\infty < \gamma_3 < \infty
\end{align*}
\]

The new estimator results from maximizing the following log–posterior (with a proper exclusion of terms not involving \( \gamma \)):

\[
\ell^*(\gamma \mid x, y) = \log p(\gamma) - n \log a(\gamma) + \gamma^T \sum_{i=1}^{n} k_i(x_i, y_i)
\]

where \( p(\gamma) \propto p_{-\gamma_3}(-\gamma_3) p_{-\gamma_4}(-\gamma_4) p(\gamma_5) \). By differentiating partially both sides of (68) with respect to \( \gamma \), equating the result to zero and manipulating, we arrive at the following system of equations:

\[
\begin{align*}
    E(B_i; \gamma) &= \bar{b} \\
    E(W_i; \gamma) &= \bar{w} \\
    Var(B_i; \gamma) + \frac{(v-2)}{2n(-\gamma_3)} &= s_b^2 + \frac{1}{2n}
\end{align*}
\]

\[14\] In practice, this normal prior for \( \gamma_5 \) has a limited effect, because the optimization of the log–PCD is carried out under the restraint that this natural parameter stains inside the bounded interval determined as in equation (38).
\begin{align*}
\text{Var}(W; \gamma) + \frac{(v-2)}{2n(-\gamma_j)} &= s_w^2 + \frac{1}{2n} \\
\text{Cov}(B_j, W; \gamma) + \frac{\gamma_5}{nh^2} &= s_{bw}
\end{align*}
\tag{72}

\text{Cov}(B_j, W; \gamma) + \frac{\gamma_5}{nh^2} = s_{bw}
\tag{73}

The equation system (69)–(73) replaces (52)–(56). Note the first two equations of the new system remain the same as of the previous one, which displayed no problem to be satisfied. The last three equations have now different forms. The variance equations (71) and (72) were added by the term \(1/2n\) in their right-hand side, which tends to be negligible if \(n\) is large. Thus, both right sides are bounded above now by \(1/4 + 1/2n = 1/4\) so things did not change much. However, the left hand sides were modified so that the distribution variances were added by the term \((v-2)/(2n\times-\gamma_j)\), \(j = 3,4\). Provided that \(v > 2\), this term will always be positive and the equations (71)–(72) can always be satisfied.

A similar analysis of equation (73) for the covariance equation, that replaces now equation (56), indicates that a solution can always be found also. The right-hand side remains the same and equal to the sample covariance, whereas the left-hand side displays now the distribution covariance added by the term \(\gamma_5 / nh^2\). Since by definition \(h\) must be nonzero, this term is unbounded, taking values on the whole real line because \(-\infty < \gamma_5 < \infty\). So, a value for \(\gamma_5\) that makes this equation valid can always be found.

\section{8 A Monte Carlo experiment}

In this section, we present a Monte Carlo simulation exercise with the aims of testing the EM–EI implementation of King’s EINM, as explained in the previous sections, and to compare its behavior with King’s approach. The results are presented in Tables 2, 3 and 4, but before we discuss them some explanations regarding the development of the simulation experiment must be made. These are listed below:

\begin{enumerate}
\item \textbf{Computer codes:} for both approaches, specialized program routines were written using the Matlab (from Mathworks, Inc.) programming language, version 5.0. Though available from Gary King a software to implement his method, we had to write a specialized routine to read sequentially the simulated data sets and to compute specific evaluation statistics. We highlight that our routine \textit{does not use the final step of simulation of predictive densities for the disaggregate data variables}, as King’s software does. Our program just predict the disaggregate data variables from their conditional means (say, given the aggregate data and the estimated parameters);

\item \textbf{Data simulation:} it was created 9 datasets, divided in 3 groups, with each dataset containing 30 simulated sequences for each disaggregate and aggregate data variables. The data sets in each group differ in the number of observations per sequence: we simulated sequences with 20, 50 and 100 observations. Disaggregate data variables were simulated first, and by the accounting identity, the aggregate \(T_i\) variable was computed. Values of the \(X_i\) variable were made
constant across sequences (and simulated once according to a uniform distribution), for this variable is assumed deterministic.

3. Data groups: Each group had its data sequences simulated according to a particular value for the true parameters vector \( \psi = [\mu_b, \mu_w, \sigma_b^2, \sigma_w^2, \rho] \), as follows:

- **Group 1** (weak truncation): \( \psi = [0.05, 0.065, 0.065, 0.05, 0.05] \);
- **Group 2** (high truncation): \( \psi = [0.09, 0.065, 0.065, 0.01] \);
- **Group 3** (intermediate truncation): \( \psi = [0.09, 0.065, 0.065, 0.09] \).

As can be seen, the true parameters differ among the groups only in the means. The idea was to place the mode of a same bivariate normal over three distinct points inside the unit square, and then truncate it for simulation. So we have three types of simulated data: one subject to weak truncation effects (mode placed over the center of the unit square); one subject to high truncation effects (mode placed near the upper left corner) and one with intermediate truncation effects (mode placed near the center of the vertical right side of the square).

4. Parameterizations: we have talked here about three alternative parameterizations for the EINM: the original \( \psi \), the alternative \( \phi \) proposed by King, and the natural or canonical \( \gamma \) used in the exponential family representation. They are related, on a two by two basis, to each other under a one–to–one correspondence, so that we can easily transform any one of them in the others. The original and the natural parameterizations are not so useful for computational purposes as the alternative one used by King, since this last reduces the correlation between means and variances and makes unbounded the parameter space. Though we used the original parameterization to simulate the data, when maximizing the LPID (in King’s approach) or the LPCD (at each M–Step of the EM–EI approach) in our program routines we did that by searching iteratively in the parameter space for \( \phi \). At the end, the routines convert \( \phi \) back to \( \psi \).

5. Priors: In addition to choosing working with the alternative parameterization \( \phi \), we also chose working with the optional priors proposed by King (1997). Though it seems to be a step back from the developments we have presented herein, we decided for that with two purposes: first, the results can be more comparable to the original King’s method; and second, we could evaluate numerically (thus in a restricted fashion) the degree in which King’s priors are suitable as correcting devices for the possible non–existence of the MLEs (or to a solution to the LCD equations).

---

15 It is clear here also our purpose to study well–behaved cases (all EINMs are met), as they could provide a good starting–up to consider, for the future, more complex Monte Carlo studies. For this reason, we placed the mode of the TBNR only inside and not outside the unit square, and considered low values for the dispersion parameters, say, we assumed low variances and zero correlation in the 3 groups.
6. Termination criteria: For both approaches, the maximization subroutines terminates the iterative search when the two following criteria are jointly (not just one of them) met:

\[
\begin{align*}
\left| \log p_{ID}(\hat{\phi}_{k+1} \mid y) - \log p_{ID}(\hat{\phi}_k \mid y) \right| &< 10^{-4} \\
\left\| \hat{\phi}_{k+1} - \hat{\phi}_k \right\| &< 10^{-4}
\end{align*}
\]

or when a maximum number \( K \) of iterations are achieved. We set \( K_{\text{KING}} = p \times 10 \) (with \( p \) the number of observations) and \( K_{\text{EM-EI}} = 250 \) (we were more permissible here with King’s approach in order to reduce the total processing time of the experiment, as we will discuss ahead).

Tables 2, 3 and 4 are equally organized and display the results of the experiment for each of the three groups of data sets. They are horizontally divided according to the number of observations, 20, 50 or 100, used to simulate the sequences. Vertically, they are also divided in three parts: The first displays some computer performance indicators; the second the results of parameters estimation; and the third the results of disaggregate data prediction for each of the \( i_B \) and \( i_W \) variables.

The analysis of the results indicate that both methods have a paired performance regarding parameter estimation and disaggregate data prediction. King’s approach performs slightly better in the two cases, though not so much with regard to disaggregate data prediction. A great advantage of King’s method comes from its faster computations, as we can see by examining the columns labeled ‘Cpt’ (computer time in minutes.seconds) in the three tables.

Both methods estimates well the parameters in the weak truncation case (Table 2), and with a significant reduction in standard deviation of simulations when we pass from 20 to 50, as well as when we pass from 50 to 100, observations. These performance are not so good in the other two cases (Tables 3 and 4). More evident biased parameters estimates occur, with cases of estimated means for the \( i_B \) variable taking negative values (outside the unit square) for both methods (Table 3). There is also a less significant reduction in standard deviations of simulations when we augment the number of observations.

In an opposite direction, the methods perform a bit better in the strong and intermediate truncation cases. This was somewhat expected, as long as the simulated disaggregate observations tends to be more concentrated in a non-central area of the unit square, and may imply in somewhat more restricted inverse images on the space of the aggregate observations. As a consequence, larger coverages of predicted points closer to the true ones tends to occur (see columns labeled ‘C10j’ and ‘C15j’, \( j = b,w \)).

Finally, we introduced in each table a column labeled ‘Nc’ (see the Indicators division) which accounts for the number of sequences that may (do) not have converged, say, that did not stop before the maximum allowed number of iterations were achieved. Note that, among the 270 sequences analysed, it happens three times for each method:
a) For both of them, the 2nd sequence in the 1st data set (Table 1, 20 observations), and the 25th sequence of the 2nd data set (Table 1, 50 observations);

b) For King’s method only, the 10th sequence of the 8th data set (Table 3, 50 observations); and

c) For the EM–EI method only, the 5th sequence of the 3rd data set (Table 1, 100 observations).

We cannot say much about the two sequences mentioned in b) and c), for convergence was achieved by one of the methods. Since the two aims at maximizing the same function, the LPID, failure of one of them in obtain success performing its task may be due to some form of numerical problem, like propagation of errors.

However, the fact that a same sequence did not converge in both methods is more informative. It happened in the present experiment with two of the sequences analyzed. Anyway, it is indeed a too small portion of the whole set of 270 sequences. We shall remember, though, that the examples used here are quite well–behaved ones. Real data sets are in general not like the ones we tested here, even those which seem to display a behavior in more accordance with the EINM. Our point here is that we do not know enough yet about the intrinsic properties of the LID, as we do with respect to the LCD (which was the main subject of this paper). Thus, further research is needed with this regard.
## Table 2. Monte Carlo Analysis: performance in parameter estimation and disaggregate data prediction (weak truncation)

<table>
<thead>
<tr>
<th>Data Set / Method</th>
<th>Indicators</th>
<th>Parameters Estimation</th>
<th>Disaggregate Data Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>It</td>
<td>Cpt</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nobs = 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>108</td>
<td>0.05</td>
<td>(1(2))</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>25</td>
<td>0.01</td>
<td>3.77</td>
</tr>
<tr>
<td>EM–EI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>50</td>
<td>1.73</td>
<td>(1(2))</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>51</td>
<td>2.09</td>
<td>3.66</td>
</tr>
<tr>
<td>Nobs = 50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>113</td>
<td>0.1</td>
<td>(1(25))</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>30</td>
<td>0.0</td>
<td>3.86</td>
</tr>
<tr>
<td>EM–EI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>33</td>
<td>1.2</td>
<td>(1(25))</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>12</td>
<td>0.5</td>
<td>3.86</td>
</tr>
<tr>
<td>Nobs = 100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>120</td>
<td>0.1</td>
<td>34.44</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>17</td>
<td>0.0</td>
<td>5.61</td>
</tr>
<tr>
<td>EM–EI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>47</td>
<td>2.2</td>
<td>(1(5))</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>16</td>
<td>0.8</td>
<td>5.61</td>
</tr>
</tbody>
</table>

Notes: 30 sequences, with Nobs observations each, were simulated by data set; True parameters are **boldface typed**; Maximum no. of iterations = 5×Nobs for King and 250 for EM–EI; It = no. of iterations; Cpt = computer time in minutes.seconds; Nc = no. of non–convergent runs with sequence index inside parentheses; Lpid = log–posterior of the incomplete (aggregate) data. For J = B or W, Uj = mean of J; S2j = variance of J; Ro = correlation coefficient between B and W; Biasj = prediction bias for J; C10j: coverage of 10% deviation from true for J; C15j = idem, but considering 15%; Maej = mean absolute deviation for J; Rmseb = root mean squared error for J.
Table 3. Monte Carlo Analysis: performance in parameter estimation and disaggregate data prediction (strong truncation)

<table>
<thead>
<tr>
<th>Data Set / Method</th>
<th>Indicators</th>
<th>Parameters Estimation</th>
<th>Disaggregate Data Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>It</td>
<td>Cpt</td>
<td>Ne</td>
</tr>
<tr>
<td><strong>Nobs = 20</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>143</td>
<td>0.1</td>
<td>14.12</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>25</td>
<td>0.0</td>
<td>3.07</td>
</tr>
<tr>
<td><strong>EM–EI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>48</td>
<td>1.8</td>
<td>13.82</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>53</td>
<td>2.5</td>
<td>2.94</td>
</tr>
<tr>
<td><strong>Nobs = 50</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>166</td>
<td>0.1</td>
<td>33.10</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>37</td>
<td>0.0</td>
<td>4.75</td>
</tr>
<tr>
<td><strong>EM–EI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>38</td>
<td>1.4</td>
<td>32.39</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>15</td>
<td>0.7</td>
<td>4.70</td>
</tr>
<tr>
<td><strong>Nobs = 100</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>236</td>
<td>0.1</td>
<td>63.95</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>152</td>
<td>0.1</td>
<td>7.94</td>
</tr>
<tr>
<td><strong>EM–EI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>37</td>
<td>1.4</td>
<td>63.22</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>8</td>
<td>0.3</td>
<td>7.59</td>
</tr>
</tbody>
</table>

Notes: 30 sequences, with Nobs observations each, were simulated by data set; True parameters are **boldface typed**; Maximum no. of iterations = 5×Nobs for King and 250 for EM–EI; It. = no. of iterations; Cpt = computer time in minutes.seconds; Ne = no. of non-convergent runs with sequence index inside parentheses; Lpid = log-posterior of the incomplete (aggregate) data. For J = B or W, Uj = mean of J; S2j = variance of J; Ro = correlation coefficient between B and W; Biasj = prediction bias for J; C10j: coverage of 10% deviation from true for J; C15j = idem, but considering 15%; Maej = mean absolute deviation for J; Rmseb = root mean squared error for J.
Table 4. Monte Carlo Analysis: performance in parameter estimation and disaggregate data prediction (*intermediate* truncation)

<table>
<thead>
<tr>
<th>Data Set / Method</th>
<th>Indicators</th>
<th>Parameters Estimation</th>
<th>Disaggregate Data Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>It</td>
<td>Cpt</td>
<td>Ne</td>
</tr>
<tr>
<td>Nobs = 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>101</td>
<td>0.0</td>
<td>13.13</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>16</td>
<td>0.0</td>
<td>3.23</td>
</tr>
<tr>
<td>EM–EI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>43</td>
<td>1.4</td>
<td>11.98</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>44</td>
<td>1.7</td>
<td>2.67</td>
</tr>
<tr>
<td>Nobs = 50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td>114</td>
<td>0.1</td>
<td>1(10)</td>
</tr>
<tr>
<td>Mean</td>
<td>18</td>
<td>0.0</td>
<td>4.87</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>11</td>
<td>0.5</td>
<td>4.59</td>
</tr>
<tr>
<td>EM–EI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>34</td>
<td>1.1</td>
<td>24.71</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>11</td>
<td>0.5</td>
<td>4.59</td>
</tr>
<tr>
<td>Nobs = 100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KING</td>
<td>153</td>
<td>0.1</td>
<td>49.33</td>
</tr>
<tr>
<td>Mean</td>
<td>29</td>
<td>0.0</td>
<td>6.46</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>12</td>
<td>0.5</td>
<td>6.56</td>
</tr>
<tr>
<td>EM–EI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>43</td>
<td>1.5</td>
<td>47.66</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>12</td>
<td>0.5</td>
<td>6.56</td>
</tr>
</tbody>
</table>

Notes: 30 sequences, with **Nobs** observations each, were simulated by data set; *True* parameters are **boldface typed**; Maximum no. of iterations = 5×Nobs for King and 250 for EM–EI; *It* = no. of iterations; *Cpt* = computer time in minutes,seconds; *Ne* = no. of *non–convergent* runs with sequence index inside parentheses; *Lpid* = log–posterior of the incomplete (aggregate) data. For *J* = *B* or *W*, *Uj* = mean of *J*; *S2j* = variance of *J*; *Ro* = correlation coefficient between *B* and *W*; *Biasj* = prediction bias for *J*; *C10j* = coverage of 10% deviation from true for *J*; *C15j* = idem, but considering 15% ; *Maej* = mean absolute deviation for *J*; *Rmseb* = root mean squared error for *J*. 
9 Final Comments

In this paper, we presented some findings of a research on recent contributions to EI. Particularly, we presented and discussed a different way to implement and use the EINM introduced by King (1997), as part of this author’s method for making EI in contingency tables of order 2x2. We did not propose here a new EI model, but just an alternative way to implement an existing one, by means of the EM Algorithm.

In his EI method, King essentially suggests to estimate the parameters of the EINM via maximization of a log–PID, and to predict the disaggregate variables values (the complete data) via simulation of their predictive densities. The maximization of the log–PID undertaken by King uses standard optimization algorithms. By its turn, the EMA used by us is an alternative optimization device that works by predicting the complete data (E–step) and maximizing the expected conditional log–PCD (M–Step) in successive iterations until convergence. Provided that certain conditions are satisfied, the result is always a stationary point of the log–PID.

These efforts of ours were not particularly fruitful in providing a better way to making EI. But, as we pointed in the introduction of the paper, the demands placed by practitioners and applied researchers on improved methods and guidance for making EI will be fulfilled only from deeper understanding of intrinsic features of the proposed EI models and methods. The development of formal statistical tests and more effective practical guidance will depend on that. We believe that the study of the EINM we were led to, by means of the attempt to implement it via the EMA, has given some help to us in this direction.

We saw here that the operation of the EMA makes use of the PCD, which is proportional to the prior times the LCD. The latter is the product of TBNRs, which are densities belonging to a regular exponential family. Thus the LCD also belongs to a family of that type. We discussed here the properties of this last function, like the strict convexity of the negative of its log, and the possibilities that it does not have any maximum. We showed this is indeed a problem, by presenting sufficient conditions for non–existence, and proposed a correcting procedure based on the alternative maximization of a particular posterior function, instead of a likelihood one, built from a suitably chosen prior for the EINM parameters. Under the alternative estimation method, parameters estimates exist with certainty, but the prior and thus the posterior used are not the same ones used by King in his method. So, there is a chance that the log–PID maximized under this author’s method, and also by means of the EM–EI approach, does not always have a maximum and even neither a stationary point.

Finally, the numerical Monte Carlo experiment that we have undertaken was with the primary purpose to compare the performance of the two approaches, King’s and the EM–EI. We used for both methods the King’s priors, because of its more appealing properties to implement numerical search. We realized the two methods provide similar results, but King’s approach is substantially faster. But, in addition, the experiment showed that for the particular well–behaved sequences we simulated, the two methods converged for the vast majority of them (though not for all), which suggest in some degree that the log–PID of King’s method is fine. Anyway, we do not know enough about the LID properties as we know about the LCD, and we still do not know
also the detailed effects that King’s priors have for correcting the non–existence of MLEs, either for the LCD or for the LID.

References


