Inferential mixing distributions and applications

R.J. Biscay, T. Ozaki, E. Diaz-Frances
Inferential mixing distributions and applications

R. J. Biscay, T. Ozaki, E. Díaz-Frances

Abstract
Following Akaike's suggestions, inferential mixing distributions (IMDs) are defined as data-based distributions on the parameter space and their fitting is assessed according to the expected Kullback-Leibler divergence. Applications of IMDs to some statistical problems are explored. They include predictive inference, frequentist interpretation of fiducial probabilities under transformation models, and the elimination of nuisance parameters by integration.

Key words and phrases: mixing distributions; fiducial probability; predictive densities; nuisance parameters.

1 Introduction
Let \( f (\cdot; \theta) : \theta \in \Theta \) be a parametric statistical model specified by probability densities on a finite-dimensional euclidean sample space \( X \) with respect to some \( \sigma \)-finite measure \( \mu \) (for simplicity, we will use the notation \( \mu (dz) = dz \) in integrals with respect to \( \mu \)). Let \( x = (x_1, ..., x_n) \) be data obtained by i.i.d. sampling from \( f (\cdot; \theta_0) \) with unknown value of the parameter \( \theta_0 \in \Theta \).

A basic problem of predictive inference is to estimate, on the basis of the available data \( x \), the density \( f_Z (\cdot; \theta_0) \) of a vector \( Z = (Z_1, ..., Z_m) \) of i.i.d. future observations from \( f (\cdot; \theta_0) \). Notice that here \( Z \) are hypothetical, not actual observations.

For this problem, Akaike (1978) introduced a frequentist approach based on the following two main points:

1) The use of the expected Kullback-Leibler divergence (or relative entropy),
\[
E_{\theta_0} KL (g (\cdot; X), f (\cdot; \theta_0)) = -E_{\theta_0} \int f_Z (z; \theta_0) \ln (g (z; X)) dz + \int f_Z (z; \theta_0) \ln (f_Z (z; \theta_0)) dz,
\]
as criterion to assess the deviation of any estimator \( g (\cdot; X) = f_Z (\cdot; \theta_0) \) from \( f_Z (\cdot; \theta_0) \). Equivalently, preference is given to greater values of the expected log-likelihood (or Kullback-Leibler 'similarity')
\[
E_{\theta_0} S (g (\cdot; X), f (\cdot; \theta_0)) = E_{\theta_0} \int f_Z (z; \theta_0) \ln (g (z; X)) dz = E_{\theta_0} \ln (g (Z; X)).
\]

---

*Havana University, Faculty of Mathematics and Computer Sciences. San Lazaro y L. Vedado. Ciudad Habana. Cuba. (e-mail: rolando@matcom.uh.cu)

†Institute of Statistical Mathematics, Graduate University of Advanced Studies. 4-6-7 Minami Azabu, Minato-ku. Tokyo Japan (email: ozaki@isum.ac.jp)

II) The use of data-based distributions \( Q = Q (; x) \) on the parameter space \( \Theta \) as tools to define density estimators \( g (; x) = \tilde{f}_Z (; \theta_0) \) by means of predictive mixtures

\[
g (; x) = g (; Q) = \int f_Z (; \theta) Q (d\theta; x).
\]

Such random distributions for the parameter are called inferential distributions in Akaike (1978) but we will refer to them more specifically as inferential mixing distributions (IMDs). This role is motivated by the fact that any, possibly randomized selection \( f_Z (; \tilde{\theta}(x)) \) of a component of the model for \( Z \) (usually called an estimative probability density) is, with respect to the criterion (I), a worse estimator than the mixture density

\[
g (;) = \int f_Z (; \theta) dQ (\theta) = E_{\theta_0} f_Z (; \tilde{\theta}(X)) ,
\]

where \( Q \) is the distribution of \( \tilde{\theta}(X) \). Indeed, from Jensen's inequality it follows that

\[
E_{\theta_0} S \left( f_Z \left( ; \tilde{\theta}(X) \right), f \left( ; \theta_0 \right) \right) = \int f_Z (z; \theta_0) E_{\theta_0} \ln \left( f_Z \left( z; \tilde{\theta}(X) \right) \right) dz \leq \int f_Z (z; \theta_0) \ln \left( E_{\theta_0} f_Z \left( z; \tilde{\theta}(X) \right) \right) dz.
\]

Notice that (I) differs from the Bayesian utility function based on the Kullback-Leibler divergence (see, e.g., Aitchison, 1975, Section 2). In the former the expectation is taken with respect to the density corresponding to the fixed, unknown value of the parameter \( \theta = \theta_0 \), not with respect to a prior density \( f (\theta) \). This gives meaning to the criterion under repeated sampling. Specifically, IMDs \( Q (; x) \) are endowed with the following frequentist interpretation: \( Q (d\theta; x) \) is the weight given to the component \( f_Z (; \theta) \) of the model in defining the predictive mixture density \( g (; Q) \), and the optimality of such weighting is assessed by the closeness of \( g (; Q) \) to the true density \( f_Z (z; \theta_0) \) according to (I). This is in contrast with the other way that has been proposed in statistics to define data-based distributions on the parameter space—namely, fiducial arguments (Fisher, 1935) whose frequentist interpretation remains controversial (see e.g. comments in Sprott, 2000).

Akaike (1978) applied the approach (I) – (II) for the objective evaluation of prior distributions. Indeed, their associated posterior distributions \( Q (\theta; x) = Q (\theta/x) \) are inferential mixing distributions, and so can be evaluated by applying the criterion (I) to the corresponding predictive densities \( g (z; x) = f (z/x) \).

For some statistical models, it has been shown that Bayesian predictive densities (usually based on 'vague' prior distributions) perform better than common estimative densities (Aitchison, 1975). However, posterior distributions results to be a rather restricted subset of the class \( \mathcal{M} \) of all IMDs. A major reason for this is that prior distributions fail in general to express complete lack of information about the parameter. Hence any posterior distribution tends to act unfavorably towards some region of \( \Theta \). This causes deterioration of the criterion (I) when sampling from a density \( f (; \theta_0) \) with \( \theta_0 \) in such a region, as illustrated in Akaike (1978).

The limitation just mentioned points towards the need for searching inferential mixing distributions in the whole set \( \mathcal{M} \), not only in its Bayesian subset \( \mathcal{M}_B \). However, most non-Bayesian works on predictive inference have followed different approaches that lack an optimally criterion such as (I) (see, e.g., comments by Cox, 1986) and are not based on predictive mixtures (II). Typically, they lean on specific
structural assumptions that allow for conditioning on minimal sufficient statistics (as in Hinkley, 1979; Butler, 1986) and asymptotic arguments for exponential family models (see reviews in Bjørnstad, 1990; Geisser, 1993; and references therein). A single exception has been the use of IMFs defined by plug-in estimates of distributions of maximum likelihood estimators for exponential families (Harris, 1989; Vidoni, 1995).

The present work follows the original Akaīke's approach in looking further for good IMFs in the whole set $\mathcal{M}$ according to the expected Kullback-Leibler divergence. Furthermore, it explores applications of IMFs not only for predictive inference but also for other statistical problems: the frequentist interpretation of fiducial probabilities under transformation models and the elimination of nuisance parameters by integration.

Section 2 derives optimal IMFs for transformation models under specific invariances. For a number of examples they coincide with Bayesian posterior distributions based on 'vague' prior information and also with fiducial distributions. In Section 3 it is demonstrated that the distribution of the maximum likelihood estimator is an asymptotically optimal IMF. However, this is of limited relevance in practice because said distribution depends on the unknown value of the parameter. Plug-in estimation have been suggested (Harris, 1989) but this does not take explicitly into consideration the performance of the resulting IMD with respect to the target criterion (1). For this, Section 4 introduces a method that provides optimal nonparametric IMFs according to an empirical version of this criterion. This is applied to non-Bayesian integration of nuisance parameters in Section 5. All proofs are deferred to the Appendix.

2 Inferential mixing distributions for transformation models under invariances

The following assumptions are satisfied in a number of exponential families.

T1) For each $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ there are invertible affine transformations

$$\Gamma_x \theta = A_x \theta + a_x : \Theta \rightarrow \Theta$$

and

$$\Phi_x z = B_x z + b_x : Z \rightarrow Z,$$

such that

$$f_x (z; \Gamma_x \theta) = f_x (\Phi_x^{-1} z; \theta)$$

for all $\theta \in \Theta$, $z \in Z$. Here, the matrixes $A_x$, $B_x$ and the vectors $a_x$, $b_x$ are constant except for their dependency on $x$.

T2) The random vector $Y = \Phi_x^{-1} Z$ has a density $p$ which does not depend on the unknown parameter value $\theta_0$.

T3) There is a density $h^*$ on $\Theta$ that solves the equation

$$p (\cdot) = \int f_x (\cdot; \theta) h^* (\theta) d\theta.$$

We have found that under such conditions it is possible to derived the optimal IMD, as is shown in the following theorem.

Theorem 1. Assume (T1)-(T3). Then, in the class of inferential mixing densities $q (\theta; x)$ that are proportional to a function of $\Gamma_x^{-1} \theta$ the optimal one is

$$q (\theta; x) = h^* (\Gamma_x^{-1} \theta) |A_x|^{-1}.$$

The associated predictive density is

$$g (z; x) = \int f_x (z; \theta) q (\theta; x) d\theta = p (\Phi_x^{-1} z) |B_x|^{-1}.$$
Example 1. (Gaussian model with known variance). Let \( f(\cdot; \theta) = N_p(\theta, \Sigma) \), with known covariance matrix \( \Sigma \). Consider only one future observation \( Z \) (i.e., \( m = 1 \)). Then, assumptions (T1)-(T3) are satisfied with

\[
\Gamma_x \theta = \theta + \bar{x}, \quad \Phi_x z = z + \bar{x}, \quad p = N_p \left( \theta, \left( 1 + \frac{1}{n} \right) \Sigma \right) \quad \text{and} \quad h^* = N_p \left( 0, \frac{1}{n} \Sigma \right).
\]

Therefore, the optimal IMD that is function of \( \Gamma_x^{-1} \theta = \theta - \bar{x} \) has density

\[
g(\theta; x) = N_p \left( 0, \frac{1}{n} \sum \right)_{(\theta - \bar{x})}
\]

and the corresponding predictive mixture is

\[
g(z; x) = N_p \left( \bar{x}, \left( 1 + \frac{1}{n} \right) \sum \right)_{(z)}.
\]

Notice that such IMD coincides with the fiducial distribution for this example. Interestingly, also several other examples of fiducial densities on transformation models can be shown to be optimal in the Kullback-Leibler sense by applying Theorem 1.

3 Asymptotic optimality of the distribution of the maximum likelihood estimator as inferential mixing distribution

Let \( E_H(\theta) \) and \( V_H(\theta) \) denote, respectively, the expected value and covariance matrix of a probability distribution \( H \) on \( \Theta \), \( E_H(\theta) = \int \theta H(d\theta) \) and \( V_H(\theta) = E_H(\theta - E_H(\theta))(\theta - E_H(\theta))^T \). For each \( x \in X^n \), as measure of the precision of an IMD \( Q(\cdot; x) \) consider its mean squared error with respect to the true value \( \theta_0 \) of the parameter,

\[
MSE(Q(\cdot; x)) = E_{Q(\cdot; x)} \left( (\theta - \theta_0)(\theta - \theta_0)^T \right).
\]

Denote by \( \hat{\theta} = \hat{\theta}(X) \) the maximum likelihood estimator of \( \theta_0 \), and by \( F_{\hat{\theta}}(\cdot; \theta_0) \) its probability distribution.

Theorem 2. (Asymptotic optimality of the maximum likelihood IMD with respect to precision) Assume regularity conditions to guarantee the Rao-Cramer’s inequality and asymptotic Gaussian distribution of \( \hat{\theta} \). Suppose also that \( Q(\cdot; x) \) is an unbiased IMD in the sense that \( E_{\theta_0} E_{Q(\cdot; X)}(\theta) = \theta_0 \). Then,

\[
E_{\theta_0} MSE(Q(\cdot; X)) \geq V_{\theta_0} MSE(Q(\cdot; X)) \geq I^{-1}(\theta_0)/n.
\]

Here, \( A \geq B \) for matrices \( A, B \) means that \( A - B \) is a definite non-negative matrix. \( I^{-1}(\theta_0) \) denotes the Fisher’s information matrix corresponding to one observation of the model \( f(\cdot; \theta) \) evaluated at \( \theta_0 \).

The lower bound stated in the above inequality is asymptotically achieved by \( F_{\hat{\theta}}(\cdot; \theta_0) \). Thus, \( Q(\theta; x) = F_{\hat{\theta}}(\cdot; \theta_0) \) is an IMD with asymptotically optimal expected precision.

4 Nonparametric estimation of inferential mixing distributions

4.1 Method

Let \( \hat{\theta} = \hat{\theta}(X) \) be a consistent estimator of \( \theta_0 \), and define the simplex

\[
\Pi = \left\{ \pi = (\pi_1, \ldots, \pi_J) : \sum_{j=1}^{J} \pi_j = 1, \pi_j \geq 0 \text{ for } j = 1, \ldots, J \right\}.
\]
For each $\pi = (\pi_1, ..., \pi_J) \in \Pi$, define the following (discrete) IMD and its associated predictive mixture density as:

$$Q(\theta; \pi) = \sum_{j=1}^{J} \pi_j \delta_\theta_j(\theta) \quad \text{and} \quad g(z; Q(\pi)) = \sum_{j=1}^{J} \pi_j f(z; \theta_j^*).$$

Given data $x = (x_1, ..., x_n) \in \mathcal{X}^n$, generate $x^1, ..., x^J (J \geq n)$ by resampling from $x$. For instance, they can be bootstrap samples or leave-one-out samples $x^j = \{x_1, ..., x_n\} \setminus \{x_j\}, j = 1, ..., n$. Compute the estimates $\theta^j = \hat{\theta}(x^j), j = 1, ..., J$, corresponding to these samples.

Obtain the IMD $Q(\theta; \pi(x))$ as solution of the optimization problem

$$\max \left\{ \sum_{i=1}^{n} \ln g(x_i; Q(\pi)) : \pi \in \Pi \right\}.$$

Notice that $\sum_{i=1}^{n} \ln g(x_i; Q(\pi))$ is an empirical version of $KL(g(\cdot; Q(\pi)), f(\cdot; \theta_0))$ obtained by substituting the empirical distribution of the data by $f(\cdot; \theta_0)$.

This optimization problem has solution, and efficient algorithms are available (Lindsay, 1995).

5 Application to non-Bayesian integration of nuisance parameters

Consider in this Section the parameter of the model to be $(\theta, \lambda) \in \Theta \times \Lambda$, where $\lambda$ is a nuisance parameter while $\theta$ is of interest.

5.1 Profile likelihood in terms of Kullback-Leibler projections

As is well-known there is a close relation between the log-likelihood function and the Kullback-Leibler similarity $S$ for a parameter $\beta$. Specifically, given $n$ i.i.d. observations $x = (x_1, ..., x_n)$ from $f(\cdot; \beta_0)$ the log-likelihood (divided by $n$) is

$$\hat{l}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ln(f(x_i; \beta)) = \int \ln(f(z; \beta)) dF_n(z),$$

where $F_n$ is the empirical distribution of the data. The last expression can be thought of as an empirical version $S_n(f(\cdot; \beta), f(\cdot; \beta_0))$ of

$$l(\beta) = S(f(\cdot; \beta), f(\cdot; \beta_0)) = \int \ln(f(z; \beta)) f(z; \beta_0) dz,$$

that is obtained by substituting $dF_n(z)$ by $f(z; \beta_0) dz$ in $S$.

The profile log-likelihood is defined by

$$l_P(\theta) = \max \left\{ \hat{l}(\theta, \lambda) : \lambda \in \Lambda \right\} = \max \left\{ S_n(f(\cdot; \theta, \lambda), f(\cdot; \theta_0, \lambda_0)) : \lambda \in \Lambda \right\} = S_n(f(\cdot; \theta, \hat{\lambda}(\theta)), f(\cdot; \theta_0, \lambda_0)),$$

where $\hat{\lambda}(\theta)$ is the value of $\lambda$ at which this maximum is achieved. Its population version is then

$$\max \left\{ S(f(\cdot; \theta, \lambda), f(\cdot; \theta_0, \lambda_0)) : \lambda \in \Lambda \right\} = S(f(\cdot; \theta, \lambda(\theta)), f(\cdot; \theta_0, \lambda_0)),$$

where $\lambda(\theta)$ is the value of $\lambda$ at which this maximum is achieved.
This allows for the interpretation of the profile likelihood in terms of the following two steps:

1) For each $\theta$, the sampling density $f(\cdot; \theta_0, \lambda_0)$ is projected on the curve of densities $\{f(\cdot; \theta, \lambda) : \lambda \in \Lambda\}$ according to the empirical similarity $S_n$, giving the point $f(\cdot; \theta, \hat{\lambda}(\theta))$. In such way it is obtained a projected model $\{f(\cdot; \theta, \hat{\lambda}(\theta)) : \theta \in \Theta\}$ from the original one $\{f(\cdot; \theta, \lambda) : (\theta, \lambda) \in \Theta \times \Lambda\}$.

2) The distance of each member $f(\cdot; \theta, \hat{\lambda}(\theta))$ of the projected model from the true density $f(\cdot; \theta_0, \lambda_0)$ is assessed by the empirical similarity $l_P(\theta) = S_n \left( f(\cdot; \theta, \hat{\lambda}(\theta)), f(\cdot; \theta_0, \lambda_0) \right)$.

5.2 Integration of nuisance parameters by means of IMDs

It is known that in a number of cases the empirical projection $f(\cdot; \theta, \hat{\lambda}(\theta))$ does not provides a good approximation to $f(\cdot; \theta, \lambda(\theta))$. This has motivated the development of several modifications to the profile likelihood (see review in Severini, 1998, and references therein).

The interpretation of the profile likelihood as a projection suggests an alternative approach to approximate $f(\cdot; \theta, \lambda(\theta))$ that is based on IMDs. We introduce it through the following two steps:

1) For each $\theta$, obtain the IMD $Q(\cdot; \theta, x)$ on $\Lambda$ as solution of the problem

$$\max \left\{ \sum_{i=1}^{n} \ln g(x_i; \theta, Q(\pi)) : \pi \in \Pi \right\},$$

where

$$g(x_i; \theta, Q(\pi)) = \sum_{j=1}^{J} \pi_j f(x_i; \theta, \lambda^j(\theta)),$$

and $\lambda^1(\theta), \ldots, \lambda^J(\theta)$ ($J \geq n$) are the maximum likelihood estimates of $\lambda$ for fixed $\theta$ corresponding to given resampling data sets $x^1, \ldots, x^J$. Then, estimate the projection $f(\cdot; \theta, \lambda(\theta))$ by means of the predictive mixture $g(\cdot; \theta, Q(\cdot; \theta, x))$.

b) Take as approximation to the ideal profile log-likelihood the following ‘mixture profile log-likelihood’:

$$l_{MP}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln g(x_i; \theta, Q(\cdot; \theta, x)).$$

6 Examples

Example. (Pair of observations with common mean) Let $x_i, y_i$ be independent observations with distribution $f(\cdot; \theta, \lambda_i) = N(\lambda_i, \theta)$, $i = 1, \ldots, n$.

In this case all the variables are not i.i.d. but only the components within each pair. The approach introduced in the previous Section can be applied to each pair to obtain an IMD $Q_i(\cdot; \theta) = Q_i(\cdot; \theta, (x_i, y_i))$ and its associated mixture profile $g_i(\cdot; \theta, Q_i(\cdot; \theta, (x_i, y_i)))$. A global mixture likelihood is then obtained as

$$\tilde{l}_P(\theta) = \prod_{i=1}^{n} [g_i(x_i; \theta, Q_i(\cdot; \theta)) g_i(y_i; \theta, Q_i(\cdot; \theta))].$$

These computations may be carried out analytically. Indeed, each bootstrap sample from $(x_i, y_i)$ has the four possibilities $(x_i, x_i), (x_i, y_i), (y_i, x_i)$ and $(y_i, y_i)$ with equal probabilities. The corresponding maximum likelihood estimates of $\lambda_i$ are $\hat{\lambda}_i =$
\( x_i, (x_i + y_i)/2, (x_i + y_i)/2 \) and \( y_i \). For a number \( J \) of bootstrap samples sufficiently large the three possible estimates \( \lambda^1_i = x_i, \lambda^2_i = (x_i + y_i)/2 \) and \( \lambda^3_i = y_i \) will be in the set of estimates. Thus, the optimal IMD \( Q_i \) has to be found by solving the problem

\[
\max \{ \ln g_i (x_i; \theta, Q (\pi)) + \ln g_i (y_i; \theta, Q (\pi)): \pi \in \Pi \},
\]

where

\[
g_i (\cdot; \theta, Q (\pi)) = \sum_{j=1}^{3} \pi_j f (\cdot; \theta, \lambda^j).
\]

This is equivalent to solve the system of equations (Lindsay, 1995, pp. 60)

\[
\pi_j = \pi_j \frac{1}{2} \left[ f (x_i; \theta, \lambda^j) / g_i (x_i; \theta, Q (\pi)) + f (y_i; \theta, \lambda^j) / g_i (y_i; \theta, Q (\pi)) \right] \quad \text{for} \quad j = 1, 2, 3.
\]

It can be verified that these are satisfied by \( \pi_1 = \pi_3 = 1/2, \pi_2 = 0 \), and then

\[
Q_i (\cdot; \theta) = \frac{1}{2} \delta_{x_i} (\cdot) + \delta_{y_i} (\cdot),
\]

\[
g_i (\cdot; \theta, Q (\pi)) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2\theta} (\cdot - x_i)^2 \right) + \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2\theta} (\cdot - y_i)^2 \right).
\]

Thus, the following mixture profile likelihood is obtained:

\[
L_{MP} (\theta) = \frac{1}{\theta^{n/2}} \exp \left( -\frac{1}{2\theta} (y_i - x_i)^2 \right).
\]

Some relevant properties of this function are the following:

a) Maximization with respect to \( \theta \) leads to the consistent estimator

\[
\hat{\theta} = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i)^2.
\]

b) Relation with the modified profile-likelihood \( L_M \) (Barndorff-Nielsen, 1983):

\[
L_M (\theta) = (L_{MP} (\theta))^{1/2}.
\]

Acknowledgment

We thank Prof. D. A. Sprott for fruitful discussions on inferential issues relevant to the present work.

7 References


