Article

Exact Fit of Simple Finite Mixture Models [†]

Dirk Tasche [1,2,‡]

[1] Prudential Regulation Authority, Bank of England, 20 Moorgate, London EC2R 6DA, UK;
  E-Mail: dirk.tasche@gmx.net; Tel.: +44 20 3461 8174
[2] Department of Mathematics, Imperial College London, London SW7 2AZ, UK

[†] This paper is an extended version of our paper published in Fifth International Conference on Mathematics in Finance (MiF) 2014, organized by North-West University, University of Cape Town and University of Johannesburg, 24–29 September 2014, Skukuza, Kruger National Park, South Africa.

[‡] The opinions expressed in this note are those of the author and do not necessarily reflect views of the Bank of England.

External Editor: Michael McAleer

Abstract: How to forecast next year's portfolio-wide credit default rate based on last year's default observations and the current score distribution? A classical approach to this problem consists of fitting a mixture of the conditional score distributions observed last year to the current score distribution. This is a special (simple) case of a finite mixture model where the mixture components are fixed and only the weights of the components are estimated. The optimum weights provide a forecast of next year's portfolio-wide default rate. We point out that the maximum-likelihood (ML) approach to fitting the mixture distribution not only gives an optimum but even an exact fit if we allow the mixture components to vary but keep their density ratio fixed. From this observation we can conclude that the standard default rate forecast based on last year's conditional default rates will always be located between last year's portfolio-wide default rate and the ML forecast for next year. As an application example, cost quantification is then discussed. We also discuss how the mixture model based estimation methods can be used to forecast total loss. This involves the reinterpretation of an individual classification problem as a collective quantification problem.

1. Introduction

The study of finite mixture models was initiated in the 1890s by Karl Pearson when he wanted to model multimodal densities. Research on finite mixture models continued ever since but its focus changed over time as further areas of application were identified and available computational power increased. More recently the natural connection between finite mixture models and classification methods with their applications in fields like machine learning or credit scoring began to be investigated in more detail. In these applications, often it can be assumed that the mixture models are *simple* in the sense that the component densities are known (*i.e.*, there is no dependence on unknown parameters) but their weights are unknown.

In this note, we explore a specific property of simple finite mixture models, namely that their maximum likelihood (ML) estimates provide an exact fit of the observed densities if the estimates exist. Conceptually, exact fit is of interest because it inspires trust that also the estimates of the weights of the components are reliable. A practical consequence of exact fit is an inequality for the finite mixture model estimate of the class probabilities in a binary classification problem and the so-called covariate shift estimate (see Corollary 7 below for details). These observations extend observations made in the case "no independent estimate of unconditional default probability given" from [1] to the multi-class case and general probability spaces.

In Section 2, we present the result on the exact fit property in a general simple finite mixture model context. In Section 3, we discuss the consequences of this result for classification and quantification problems and compare the ML estimator with other estimators that were proposed in the literature. In Section 4, we revisit the cost quantification problem as introduced in [2] as an application. In Section 5, we illustrate by a stylised example from mortgage risk management how the estimators discussed before can be deployed for the forecast of expected loss rates. Section 6 concludes the note while an appendix presents the proofs of the more involved mathematical results.

2. The Exact Fit Property

We discuss the properties of the ML estimator of the weights in a simple finite mixture model in a general measure-theoretic setting because in this way it is easier to identify the important aspects of the problem. The setting may formally be described as follows:

Assumption 1. *μ is a measure on $(\Omega, \mathcal{H})$. $g > 0$ is a probability density with respect to μ. Suppose that the probability measure P is given by $P[H] = \int_H g \, d\mu$ for $H \in \mathcal{H}$. Write E for the expectation with regard to P.*

In applications, the measure μ will often be a multi-dimensional Lebesgue measure or a counting measure. The measure P might not be known exactly but might have to be approximated by the empirical measure associated with a sample of observations. This would give an ML estimation context.

The problem we study can be described as follows:

Problem 2. *Approximate g by a mixture of probability μ-densities $f_1, \ldots, f_k$, i.e., $g \approx \sum_{i=1}^k p_i f_i$ for suitable $p_1, \ldots, p_k \geq 0$ with $\sum_{i=1}^k p_i = 1$.*

In the literature, most of the time a sample version of the problem (*i.e.*, with P as an empirical measure) is discussed. Often the component densities f_i depend on parameters that have to be estimated in addition to the weights p_i (see [3] or [4], and for a more recent survey see [5]). In this note, we consider the simple case where the component densities f_i are assumed to be known and fixed. This is a standard assumption for classification (see [6]) and quantification (see [2]) problems.

Common approaches to the approximation problem 2 are

- **Least Squares.** Determine $\min_{(p_1,\ldots,p_k)} \int \left(g - \sum_{i=1}^k p_i f_i\right)^2 d\mu$ or its weighted version $\min_{(p_1,\ldots,p_k)} \int g \left(g - \sum_{i=1}^k p_i f_i\right)^2 d\mu$ (see [7] and [8] for recent applications to credit risk and text categorisation respectively). The main advantage of the least squares approach compared with other approaches comes from the fact that closed-form solutions are available. However, the achieved minimum will not be zero (*i.e.*, there is no exact fit) unless the density g happens to be a mixture of the densities f_i.

- **Kullback–Leibler (KL) distance.** Determine

$$\min_{(p_1,\ldots,p_k)} \int g \log\left(\frac{g}{\sum_{i=1}^k p_i f_i}\right) d\mu = \min_{(p_1,\ldots,p_k)} E\left[\log\left(\frac{g}{\sum_{i=1}^k p_i f_i}\right)\right] \tag{1}$$

(see [9] for a recent discussion). We show below (see Remark 1) that a minimum of zero (*i.e.*, exact fit) can be achieved if we allow the f_i to be replaced by densities g_i with the same density ratios as the f_i. In the case where P from Assumption 1 is the empirical measure of a sample, the r.h.s. of (1) becomes (up to an additive constant) a conventional log likelihood function and thus presents an ML problem.

In the following we examine in detail the conditions under which *exact fit* by the KL approximation can occur. First we note two alternative representations of the KL distance (assuming all integrals are well-defined and finite):

$$\int g \log\left(\frac{g}{\sum_{i=1}^k p_i f_i}\right) d\mu = \int g \log(g) \, d\mu - \int g \log\left(\sum_{i=1}^k p_i f_i\right) d\mu \tag{2}$$

$$= \int g \log(g) \, d\mu - \int g \log(f_k) \, d\mu$$

$$- \int g \log\left(1 + \sum_{i=1}^{k-1} p_i (R_i - 1)\right) d\mu \tag{3}$$

with $R_i = \frac{f_i}{f_k}$, $i = 1, \ldots, k-1$ if $f_k > 0$.

The problem in (2) to maximise $\int g \log \left(\sum_{i=1}^{k} p_i f_i \right) d\mu$ was studied in [10] (with $g = $ empirical measure, which gives ML estimators). The authors suggested the Expectation-Maximisation (EM) algorithm involving conditional class probabilities for determining the maximum. This works well in general but sometimes suffers from slow convergence.

The ML version of (2) had been studied before in [11]. There the authors analysed the same iteration procedure, which they stated, however, in terms of densities instead of conditional probabilities. In [11], the iteration was derived differently to [10], namely by studying the gradient of the likelihood function. We revisit the approach from [11] from a different angle by starting from (3).

There is, however, the complication that $g \log \left(1 + \sum_{i=1}^{k-1} p_i \left(R_i - 1 \right) \right)$ is not necessarily integrable. However, this observation does not apply to the gradient with respect to $(p_1, \ldots, p_{k-1})$. We therefore focus on investigating the gradient. With R_i as in (3), let

$$F(p_1, \ldots, p_{k-1}) \stackrel{\text{def}}{=} \int g \log \left(1 + \sum_{i=1}^{k-1} p_i \left(R_i - 1 \right) \right) d\mu \tag{4}$$

$$(p_1, \ldots, p_{k-1}) \in S_{k-1} \stackrel{\text{def}}{=} \left\{ (s_1, \ldots, s_{k-1}) : s_1 > 0, \ldots, s_{k-1} > 0, \sum_{i=1}^{k-1} s_i < 1 \right\}$$

From this we obtain for the gradient of F

$$\begin{aligned} G_j(p_1, \ldots, p_{k-1}) &\stackrel{\text{def}}{=} \frac{\partial F}{\partial p_j}(p_1, \ldots, p_{k-1}) \\ &= \int \frac{g \left(R_j - 1 \right)}{1 + \sum_{i=1}^{k-1} p_i \left(R_i - 1 \right)} d\mu, \quad j = 1, \ldots, k-1 \end{aligned} \tag{5}$$

$G_j(p_1, \ldots, p_{k-1})$ is well-defined and finite for $(p_1, \ldots, p_{k-1}) \in S_{k-1}$:

$$\begin{aligned} |G_j(p_1, \ldots, p_{k-1})| &\leq \int \frac{g \left(R_j + 1 \right)}{1 - \sum_{i=1}^{k-1} p_i + \sum_{i=1}^{k-1} p_i R_i} d\mu \\ &\leq \frac{1}{p_j} + \frac{1}{1 - \sum_{i=1}^{k-1} p_i} < \infty \end{aligned} \tag{6}$$

We are now in a position to state the main result of this note.

Theorem 3. *Let g and μ be as in Assumption 1. Assume that $R_1, \ldots, R_{k-1} \geq 0$ are $\mathcal{H}$-measurable functions on Ω. Suppose there is a vector $(p_1, \ldots, p_{k-1}) \in S_{k-1}$ such that*

$$0 = G_i(p_1, \ldots, p_{k-1}), \qquad i = 1, \ldots, k-1 \tag{7}$$

for G_i as defined in (5). Define $p_k = 1 - \sum_{j=1}^{k-1} p_j$. Then the following two statements hold:

(a) *$g_i = \frac{g R_i}{1 + \sum_{j=1}^{k-1} p_j \left(R_j - 1 \right)}$, $i = 1, \ldots, k-1$, and $g_k = \frac{g}{1 + \sum_{j=1}^{k-1} p_j \left(R_j - 1 \right)}$ are probability densities with respect to μ such that $g = \sum_{i=1}^{k} p_i g_i$ and $R_i = \frac{g_i}{g_k}$, $i = 1, \ldots, k-1$.*

(b) *Let $R_1 - 1, \ldots, R_{k-1} - 1$ additionally be linearly independent, i.e.,*

$$\text{P} \left[\sum_{i=1}^{k-1} a_i \left(R_i - 1 \right) = 0 \right] = 1 \quad \textit{implies} \quad 0 = a_1 = \ldots = a_{k-1}$$

Assume that $h_1, \ldots, h_{k-1} \geq 0$, $h_k > 0$ are μ-densities of probability measures on $(\Omega, \mathcal{H})$ and $(q_1, \ldots, q_{k-1}) \in S_{k-1}$ with $q_k = 1 - \sum_{i=1}^{k-1} q_i$ is such that $g = \sum_{i=1}^{k} q_i \, h_i$ and $R_i = \frac{h_i}{h_k}$, $i = 1, \ldots, k-1$. Then it follows that $q_i = p_i$ and $h_i = g_i$ for g_i as defined in (a), for $i = 1, \ldots, k$.

See appendix for a proof of Theorem 3.

Remark 1. *If the KL distance on the l.h.s. of (2) is well-defined and finite for all $(p_1, \ldots, p_{k-1}) \in S_{k-1}$ then under the assumptions of Theorem 3 (b) there is a unique $(p_1^*, \ldots, p_{k-1}^*) \in S_{k-1}$ such that the KL distance of g and $\sum_{i=1}^{k} p_i^* f_i$ is minimal. In addition, by Theorem 3 (a), there are densities g_i with $\frac{g_i}{g_k} = \frac{f_i}{f_k}$ for $i = 1, \ldots, k-1$ such that the KL distance of g and $\sum_{i=1}^{k} p_i^* g_i$ is zero—this is the exact fit property of simple finite mixture models alluded to in the title of this note.* $\square$

Remark 2. *(a) Theorem 3 (b) provides a simple condition for the uniqueness of the solution to (7). In the case $k = 2$ this condition simplifies to*

$$\mathrm{P}[R_1 = 1] \; < \; 1 \tag{8a}$$

For $k > 2$ there is no similarly simple condition for the existence of a solution in S_{k-1}. However, as noted in [12] (Example 4.3.1), there is a simple necessary and sufficient condition for the existence of a solution in $S_1 = (0, 1)$ to (7) in the case $k = 2$:

$$\mathrm{E}[R_1] \; > \; 1 \qquad and \qquad \mathrm{E}\big[R_1^{-1}\big] \; > \; 1 \tag{8b}$$

(b) Suppose we are in the setting of Theorem 3 (a) with $k > 2$ and all $R_i > 0$. Hence there are μ-densities $g_1, \ldots, g_k$, $(p_1, \ldots, p_{k-1}) \in S_{k-1}$, and $p_k = 1 - \sum_{i=1}^{k-1} p_i$ such that $g = \sum_{i=1}^{k} p_i \, g_i$ and $R_i = \frac{g_i}{g_k}$, $i = 1, \ldots, k$. Let $\overline{g} = \frac{\sum_{j=2}^{k} p_j \, g_j}{1 - p_1}$. Then we have another decomposition of g, namely $g = p_1 \, g_1 + (1 - p_1) \, \overline{g}$, with $p_1 \in (0, 1)$. The proof of Theorem 3 (b) shows that, with $\overline{R} = \frac{g_1}{g}$ this implies

$$0 \; = \; \int \frac{g \, (\overline{R} - 1)}{1 + p_1 \, (\overline{R} - 1)} \, d\mu \tag{9}$$

By (a) there is a solution $p_1 \in (0, 1)$ to (9) if and only if $\mathrm{E}[\overline{R}] > 1$ and $\mathrm{E}[(\overline{R})^{-1}] > 1$.

(c) As mentioned in (a), an interesting question for the application of Theorem 3 is how to find out whether or not there is a solution to (7) in S_{k-1} for $k > 2$. The iteration suggested in [11] and [10] will correctly converge to a point on the boundary of S_{k-1} if there is no solution in the interior (Theorem 2 of [11]). However, convergence may be so slow that it may remain unclear whether a component of the limit is zero (and therefore the solution is on the boundary) or genuinely very small but positive. The straightforward Newton–Raphson approach for determining the maximum of F defined by (4) may converge faster but may also become unstable for solutions close to or on the boundary of S_{k-1}.

However, when $k > 2$, the observation made in (b) suggests that the following Gauss–Seidel-type iteration works if the initial value $(p_1^{(0)}, \ldots, p_{k-1}^{(0)})$ with $p_k^{(0)} = 1 - \sum_{i=1}^{k-1} p_i^{(0)}$ is sufficiently close to the solution (if any) of (7):

> – *Assume that for some $n \geq 0$ an approximate solution $(q_1, \ldots, q_k) = (p_1^{(n)}, \ldots, p_k^{(n)})$ has been found.*

– *For $i = 1, \ldots, k$ try successively to update $(q_1, \ldots, q_k)$ by solving* (9) *with component i playing the role of component* 1 *in (b) and $p_1 = q_i$ as well as $\overline{g} = \frac{\sum_{j=1, j \neq i}^{k} q_j \, g_j}{1 - q_i}$. If for all $i = 1, \ldots, k$ the sufficient and necessary condition for the updated q_i to be in $(0, 1)$ is not satisfied then stop—it is then likely that there is no solution to* (7) *in S_{k-1}. Otherwise update q_i with the solution of* (9) *where possible, resulting in $q_{i,new}$, and set*

$$q_{j,new} = \frac{1 - q_{i,new}}{1 - q_i} \, q_j, \quad j \neq i$$

– *After step k set $(p_1^{(n+1)}, \ldots, p_k^{(n+1)}) = (q_{1,new}, \ldots, q_{k,new})$ if the algorithm has not been stopped by violation of the resolvability condition for* (7).

– *Terminate the calculation when a suitable distance measure between successive $(p_1^{(n)}, \ldots, p_k^{(n)})$ is sufficiently small.* $\square$

3. Application to Quantification Problems

Finite mixture models occur naturally in machine learning contexts. Specifically, in this note we consider the following context:

Assumption 4.

- $(\Omega, \mathcal{H})$ *is a measurable space. For some $k \geq 2$, $A_1, \ldots, A_k \subset \Omega$ is a partition of Ω. $\mathcal{A}$ is the σ-field generated by $\mathcal{H}$ and the A_i,* i.e.,

$$\mathcal{A} = \sigma(\mathcal{H} \cup \{A_1, \ldots, A_k\}) = \left\{ \bigcup_{i=1}^{k} (A_i \cap H_i) : H_1, \ldots, H_k \in \mathcal{H} \right\}$$

- P_0 *is a probability measure on $(\Omega, \mathcal{A})$ with $\mathrm{P}_0[A_i] > 0$ for $i = 1, \ldots, k$. P_1 is a probability measure on $(\Omega, \mathcal{H})$. Write E_i for the expectation with respect to P_i.*
- *There is a measure μ on $(\Omega, \mathcal{H})$ and μ-densities $f_1, \ldots, f_{k-1} \geq 0$, $f_k > 0$ such that*

$$\mathrm{P}_0[H \,|\, A_i] = \int_H f_i \, d\mu, \qquad i = 1, \ldots, k, H \in \mathcal{H}$$

The space $(\Omega, \mathcal{A}, \mathrm{P}_0)$ describes the training set of a classifier. On the training set, for each example both the features (expressed by $\mathcal{H}$) and the class (described by one of the A_i) are known. Note that $f_k > 0$ implies $A_k \notin \mathcal{H}$.

$(\Omega, \mathcal{H}, \mathrm{P}_1)$ describes the test set on which the classifier is deployed. On the test set only the features of the examples are known.

In mathematical terms, *quantification* might be described as the task to extend P_1 onto $\mathcal{A}$, based on properties observed from the training set, *i.e.*, of P_0. Basically, this means to estimate *prior class probabilities* (or prevalences) $\mathrm{P}_1[A_i]$ on the test dataset. In this note, the assumption is that $\mathrm{P}_1\big|\mathcal{A} \neq \mathrm{P}_0\big|\mathcal{A}$. In the machine learning literature, this situation is called *dataset shift* (see [6] and references therein).

Specifically, we consider the following two dataset shift types (according to [6]):

- **Covariate shift.** $P_1 \big| \mathcal{H} \neq P_0 \big| \mathcal{H}$ but $P_1[A_i \,|\, \mathcal{H}] = P_0[A_i \,|\, \mathcal{H}]$ for $i = 1, \ldots, k$. In practice, this implies $P_1[A_i] \neq P_0[A_i]$ for most if not all of the i. In a credit risk context, two portfolios of rated credit exposures differ by a covariate shift if the rating grade-level probabilities of default are equal but the rating grade distributions are different in the two portfolios.

- **Prior probability shift.** $P_1[A_i] \neq P_0[A_i]$ for at least one i but $P_1[H \,|\, A_i] = P_0[H \,|\, A_i]$ for $i = 1, \ldots, k$, $H \in \mathcal{H}$. This implies $P_1 \big| \mathcal{H} \neq P_0 \big| \mathcal{H}$ if $P_0[\cdot \,|\, A_1], \ldots, P_0[\cdot \,|\, A_k]$ are linearly independent. In a credit risk context, two portfolios of rated credit exposures differ by a prior probability shift if (1) the rating grade distributions of defaulting exposures in the two portfolios are equal and (2) the rating grade distributions of surviving exposures in the two portfolios are equal but (3) the portfolio-wide default rates of the two portfolios are different.

In practice, it is likely to have $P_1[A_i] \neq P_0[A_i]$ for some i in both covariate shift and prior probability shift. Hence, quantification in the sense of estimation of the $P_1[A_i]$ is important for both covariate shift and prior probability shift. Under covariate shift assumption, a natural estimator of $P_1[A_i]$ is given by

$$\widetilde{P}_1[A_i] \;=\; E_1 \big[P_0[A_i \,|\, \mathcal{H}] \big], \qquad i = 1, \ldots, k \tag{10}$$

Under prior probability shift, the choice of suitable estimators of $P_1[A_i]$ is less obvious.

The following result generalises the *Scaled Probability Average* method of [13] to the multi-class case. It allows to derive prior probability shift estimates of prior class probabilities from covariate shift estimates as given by (10).

Proposition 5. *Under Assumption 4, suppose that there are $q_1 \geq 0$, ..., $q_k \geq 0$ with $\sum_{i=1}^{k} q_i = 1$ such that P_1 can be represented as a simple finite mixture as follows:*

$$P_1[H] \;=\; \sum_{i=1}^{k} q_i \, P_0[H \,|\, A_i], \quad H \in \mathcal{H} \tag{11a}$$

Then it follows that

$$\begin{pmatrix} E_1 \big[P_0[A_1 \,|\, \mathcal{H}] \big] \\ \vdots \\ E_1 \big[P_0[A_k \,|\, \mathcal{H}] \big] \end{pmatrix} \;=\; M \begin{pmatrix} q_1 \\ \vdots \\ q_{k,} \end{pmatrix} \tag{11b}$$

where the matrix $M = (m_{ij})_{i,j=1,\ldots,k}$ is given by

$$\begin{aligned} m_{ij} \;&=\; E_0 \big[P_0[A_i \,|\, \mathcal{H}] \,|\, A_j \big] \\ &=\; E_0 \big[P_0[A_i \,|\, \mathcal{H}] \, P_0[A_j \,|\, \mathcal{H}] \big] \,/\, P_0[A_j] \end{aligned} \tag{11c}$$

Proof. Immediate from (11a) and the definition of conditional expectation. $\square$

For practical purposes, the representation of m_{ij} in the first row of (11c) is more useful because most of the time no exact estimate of $P_0[A_j \,|\, \mathcal{H}]$ will be available. As a consequence, there might be a non-zero difference between the values of the expectations in the first and second row of (11c) respectively. In contrast to the second row, for the derivation of the r.h.s. of the first row of (11c), however, no use of the specific properties of conditional expectations has been made.

Corollary 6. *In the setting of Proposition 5, suppose that $k = 2$. Define*

$$R_0^2 = \frac{\text{var}_0\big[P_0[A_1 \,|\, \mathcal{H}]\big]}{P_0[A_1]\,(1 - P_0[A_1])} \in [0, 1]$$

with var_0 *denoting the variance under* P_0. *Then we have*

$$E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] = P_0[A_1]\,(1 - R_0^2) + q_1\,R_0^2 \tag{12}$$

See appendix for a proof of Corollary 6. By Corollary 6, it holds that

$$
\begin{aligned}
P_0[A_1] \leq E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] \leq q_1, \quad \text{if } P_0[A_1] \leq q_1 \\
P_0[A_1] \geq E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] \geq q_1, \quad \text{if } P_0[A_1] \geq q_1
\end{aligned}
\tag{13}
$$

Hence, for binary classification problems the covariate shift estimator (10) underestimates the change in the class probabilities if the dataset shift is not a covariate shift but a prior probability shift. See Section 2.1 of [2] for a similar result for the *Classify & Count* estimator.

However, according to (12) the difference between the covariate shift estimator and the true prior probability is the smaller the greater the discriminatory power (as measured by the generalised R^2) of the classifier is. Moreover, both (12) and (11b) provide closed-form solutions for $q_1, \ldots, q_k$ that transform the covariate shift estimates into correct estimates under the prior probability shift assumption. In the following, the estimators defined in this way are called *Scaled Probability Average* estimators.

Corollary 6 on the relationship between covariate shift and Scaled Probability Average estimates in the binary classification case can be generalised to the relationship between covariate shift and KL distance estimates.

Corollary 7. *Under Assumption 4, consider the case $k = 2$. Let $R_1 = \frac{f_1}{f_2}$ and suppose that (8b) holds for $E = E_1$ such that a solution $p_1 \in (0, 1)$ of (7) exists. Then there is some $\alpha \in [0, 1]$ such that*

$$E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] = (1 - \alpha)\,P_0[A_1] + \alpha\,p_1$$

See appendix for a proof of Corollary 7. From the corollary, we obtain an inequality for the covariate shift and KL distances estimates, which is similar to (13):

$$
\begin{aligned}
P_0[A_1] \leq E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] \leq p_1, \quad \text{if } P_0[A_1] \leq p_1 \\
P_0[A_1] \geq E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] \geq p_1, \quad \text{if } P_0[A_1] \geq p_1
\end{aligned}
\tag{14}
$$

How is the KL distance estimator (or the ML estimator in the case of P being the empirical measure) of the prior class probabilities, defined by the solution of (7), in general, related to the covariate shift and Scaled Probability Average estimators?

Suppose the test dataset differs from the training dataset by a prior probability shift with positive class probabilities, *i.e.*, (11a) applies with $q_1, \ldots, q_k > 0$. Under Assumption 4 and a mild linear independence condition on the ratios of the densities f_i, then Theorem 3 implies that the KL distance and Scaled Probability Average estimators give the same results. Observe that in the context given by Assumption 4 the variables R_i from Theorem 3 can be directly defined as $R_i = \frac{f_i}{f_k}$, $i = 1, \ldots, k - 1$ or, equivalently, by

$$R_i = \frac{P_0[A_i \,|\, \mathcal{H}]}{P_0[A_k \,|\, \mathcal{H}]}\,\frac{P_0[A_k]}{P_0[A_i]}, \qquad i = 1, \ldots, k - 1 \tag{15}$$

Representation (15) of the density ratios might be preferable in particular if the classifier involved has been built by binary or multinomial logistic regression.

In general, by Theorem 3, the result of applying the KL distance estimator to the test feature distribution P_1, in the quantification problem context described by Assumption 4, is a representation of P_1 as a mixture of distributions whose density ratios are the same as the density ratios of the class feature distributions $P_0[\cdot \mid A_i]$, $i = 1, \ldots, k$.

Hence, the KL distance estimator makes sense under an assumption of identical density ratios in the training and test datasets. On the one hand, this assumption is similar to the assumption of identical conditional class probabilities in the covariate shift assumption but does not depend in any way on the training set prior class probabilities. This is in contrast to the covariate shift assumption where implicitly a "memory effect" with regard to the training set prior class probabilities is accepted.

On the other hand the "identical density ratios" assumption is weaker than the "identical densities" assumption (the former is implied by the latter), which is part of the prior probability assumption.

One possible description of "identical density ratios" and the related KL distance estimator is that "identical density ratios" generalises "identical densities" in such a way that exact fit of the test set feature distribution is achieved (which by Theorem 3 is not always possible). It is therefore fair to say that "identical density ratios" is closer to "identical densities" than to "identical conditional class probabilities".

Given training data with full information (indicated by the σ-field $\mathcal{A}$ in Assumption 4) and test data with information only on the features but not on the classes (σ-field $\mathcal{H}$ in Assumption 4), it is not possible to decide whether the covariate shift or the identical density ratios assumption is more appropriate for the data, because both assumptions result in the exact fit of the test set feature distribution $P_1 \big| \mathcal{H}$ but in general give quite different estimates of the test set prior class probabilities (see Corollary 7 and Section 5). Only if Equation (7) has no solution with positive components it can be said that "identical density ratios" does not properly describe the test data because then there is no exact fit of the test set feature distribution. In that case "covariate shift" might not be appropriate either but at least it delivers a mathematically consistent model of the data.

If both "covariate shift" and "identical density ratios" provide consistent models (*i.e.*, exact fit of the test set feature distribution) non-mathematical considerations of causality (are features caused by class or is class caused by features?) may help in choosing the more suitable assumption. See [14] for a detailed discussion of this issue.

4. Cost Quantification

"Cost quantification" is explained in [2] as follows:

> "The second form of the quantification task is for a common situation in business where a cost or value attribute is associated with each case. For example, a customer support log has a database field to record the amount of time spent to resolve each individual issue, or the total monetary cost of parts and labor used to fix the customer's problem. ... The cost quantification task for machine learning: given a limited training set with class labels, induce a cost quantifier that takes an unlabeled test set as input and returns its best estimate of the

total cost associated with each class. In other words, return the subtotal of cost values for each class."

Careful reading of Section 4.2 of [2] reveals that the favourite solutions for cost quantification presented by the author essentially apply only to the case where the cost attributes are constant on the classes. Only then the C^+ as used in Equations (4) and (5) of [2] stand for the same conditional expectations—the same observation applies to C^-.

Cost quantification can be more generally treated under Assumption 4 of this note. Denote by C the (random) cost associated with an example. According to the description of cost quantification quoted above then C is actually a feature of the example and, therefore, may be considered an $\mathcal{H}$-measurable random variable under Assumption 4.

In mathematical terms, the objective of cost quantification is the estimation of the total expected cost per class $\mathrm{E}_1[C\,\mathbf{1}_{A_i}]$, $i = 1, \ldots, k$. For a set M, its indicator function $\mathbf{1}_M$ is defined as $\mathbf{1}_M(m) = 1$ for $m \in M$ and $\mathbf{1}_M(m) = 0$ for $m \notin M$.

Covariate shift assumption. Under this assumption we obtain

$$\mathrm{E}_1[C\,\mathbf{1}_{A_i}] \;=\; \mathrm{E}_1\big[C\,\mathrm{P}_1[A_i\,|\,\mathcal{H}]\big] \;=\; \mathrm{E}_1\big[C\,\mathrm{P}_0[A_i\,|\,\mathcal{H}]\big] \tag{16}$$

This gives a probability-weighted version of the "Classify & Total" estimator of [2].

Constant density ratios assumption. Let $R_i = \frac{f_i}{f_k}$, $i = 1, \ldots, k-1$. If (7) (with $\mu = \mathrm{P}_1$ and $g = 1$) has a solution $p_1 > 0, \ldots, p_{k-1} > 0$, $p_k = 1 - \sum_{j=1}^{k-1} p_j < 1$, then we can estimate the conditional class probabilities $\mathrm{P}_1[A_i\,|\,\mathcal{H}]$ by

$$\mathrm{P}_1[A_i\,|\,\mathcal{H}] = \frac{p_i\,R_i}{1 + \sum_{j=1}^{k-1} p_j\,(R_j - 1)}, \quad i = 1, \ldots, k-1$$
$$\mathrm{P}_1[A_k\,|\,\mathcal{H}] = \frac{p_k}{1 + \sum_{j=1}^{k-1} p_j\,(R_j - 1)}$$

From this, it follows that

$$\mathrm{E}_1[C\,\mathbf{1}_{A_i}] \;=\; p_i\,\mathrm{E}_1\left[\frac{C\,R_i}{1 + \sum_{j=1}^{k-1} p_j\,(R_j - 1)}\right], \quad i = 1, \ldots, k-1$$
$$\mathrm{E}_1[C\,\mathbf{1}_{A_k}] \;=\; p_k\,\mathrm{E}_1\left[\frac{C}{1 + \sum_{j=1}^{k-1} p_j\,(R_j - 1)}\right] \tag{17}$$

Obviously, the accuracy of the estimates on the r.h.s. of both (16) and (17) strongly depends on the accuracy of the estimates of $\mathrm{P}_0[A_i\,|\,\mathcal{H}]$ and the density ratios on the training set. Accurate estimates of these quantities, in general, will make full use of the information in the σ-field $\mathcal{H}$ (*i.e.*, the information available at the time of estimation) and, because of the $\mathcal{H}$-measurability of C, of the cost feature C. In order to achieve this, C must be used as an explanatory variable when the relationship between the classes A_i and the features as reflected in $\mathcal{H}$ is estimated (e.g., by a regression approach). As one-dimensional densities are relatively easy to estimate, it might make sense to deploy (16) and (17) with the choice $\mathcal{H} = \sigma(C)$.

Note that this conclusion, at first glance, seems to contradict Section 5.3.1 of [2]. There it is recommended that "the cost attribute almost never be given as a predictive input feature to the classifier".

Actually, with regard to the cost quantifiers suggested in [2], this recommendation is reasonable because the main component of the quantifiers as stated in (6) of [2] is correctly specified only if there is no dependence of the cost attribute C and the classifier. Not using C as an explanatory variable, however, does not necessarily imply that the dependence between C and the classifier is weak. Indeed, if the classifier has got any predictive power and C is on average different on the different classes of examples, then there must be a non-zero correlation between the cost attribute C and the output of the classifier.

5. Loss Rates Estimation with Mixture Model Methods

Theorem 3 and the results of Section 3 have obvious applications to the problem of forecasting portfolio-wide default rates in portfolios of rated or scored borrowers. The forecast portfolio-wide default rate may be interpreted in an individual sense as a single borrower's unconditional probability of default. However, there is also an interpretation in a collective sense as the forecast total proportion of defaulting borrowers.

The statements of Theorem 3 and Assumption 4 are agnostic in the sense of not suggesting an individual or collective interpretation of the models under inspection. In contrast, by explaining Assumption 4 in terms of a classifier and the examples to which it is applied, we have suggested an individual interpretation of the assumption.

However, there is no need to adopt this perspective on Assumption 4 and the results of Section 3. Instead of interpreting $P_0[A_1]$ as an individual example's probability of belonging to class 1, we could as well describe $P_0[A_1]$ as the proportion of a mass or substance that has property 1. If we do so, we switch from an interpretation of probability spaces in terms of likelihoods associated with individuals to an interpretation in terms of proportions of parts of masses or substances.

Let us look at a retail mortgage portfolio as an illustrative example. Suppose that each mortgage has an associating loan-to-value (LTV) that indicates how well the mortgage loan is secured by the pledged property. Mortgage providers typically report their exposures and losses in tables that provide this information per LTV-band without specifying numbers or percentages of borrowers involved. Table 1 shows a stylised example of how such a report might look like.

Table 1. Fictitious report on mortgage portfolio exposure distribution and loss rates.

LTV Band	Last Year		This Year	
	% of exposure	**of this % lost**	**% of exposure**	**of this % lost**
More than 100%	10.3	15.0	13.3	?
Between 90% and 100%	28.2	2.2	24.2	?
Between 70% and 90%	12.9	1.1	12.8	?
Between 50% and 70%	24.9	0.5	25.4	?
Less than 50%	23.8	0.2	24.3	?
All	100.0	2.2	100.0	?

This portfolio description fits well into the framework described by Assumption 4. Choose events H_1 = "More than 100% LTV", H_2 = "Between 90% and 100% LTV", *etc.* Then the σ-field $\mathcal{H}$ is

generated by the finite partition $H_1, \ldots, H_5$. Similarly, choose $A_1 =$ "lost" and $A_2 =$ "not lost". The measure P_0 describes last year's observations, P_1 specifies the distribution of the exposure over the LTV bands as observed at the beginning of this year—which is the forecast period. We can then try and replace the question marks in Table 1 by deploying the estimators discussed in Section 3. Table 2 shows the results.

Table 2. This year's loss rate forecast for portfolio from Table 1. The R_0^2 used for calculating the Scaled Probability Average is 7.7%.

LTV Band	Covariate Shift % lost	Scaled Probability Average % lost	KL Distance % lost
More than 100%	15.0	34.6	35.4
Between 90% and 100%	2.2	6.3	6.5
Between 70% and 90%	1.1	3.2	3.3
Between 50% and 70%	0.5	1.5	1.5
Less than 50%	0.2	0.6	0.6
All	2.8	7.3	7.1

Clearly, the estimates under the prior probability shift assumptions are much more sensitive to changes in the distribution of the features (*i.e.*, LTV bands) than the estimates under the covariate shift assumption. Thus the theoretical results of Corollaries 6 and 7 are confirmed. However, recall that there is no right or wrong here as all the numbers in Table 1 are purely fictitious. Nonetheless, we could conclude that in applications with unclear causalities (like for credit risk measurement) it might make sense to compute both covariate shift estimates and KL estimates (more suitable under a prior probability shift assumption) in order to gauge the possible range of outcomes.

6. Conclusions

We have revisited the maximum likelihood estimator (or more generally Kullback–Leibler distance estimator) of the component weights in simple finite mixture models. We have found that (if all weights of the estimate are positive) it enjoys an exact fit property, which makes it even more attractive with regard to mathematical consistency. We have suggested a Gauss–Seidel-type approach to the calculation of the Kullback–Leibler distance estimator that triggers an alarm if there is no solution with all components positive (which would indicate that the number of modelled classes may be reduced).

In the context of two-class quantification problems, as a consequence of the exact fit property we have shown theoretically and by example that the straightforward "covariate shift" estimator of the prior class probabilities may seriously underestimate the change of the prior probabilities if the covariate shift assumption is wrong and instead a prior probability shift has occurred. This underestimation can be corrected by the Scaled Probability Average approach, which we have generalised to the multi-class case, or the Kullback–Leibler distance estimator.

As an application example, we then have discussed cost quantification, *i.e.*, the attribution of total cost to classes on the basis of characterising features when class membership is unknown. In addition,

we have illustrated by example that the mixture model approach to quantification is not restricted to the forecast of prior probabilities but can also be deployed for forecasting loss rates.

Both the theoretical (Corollary 7) and illustrative numerical results (Table 2) suggest that the difference between the Kullback–Leibler and the more familiar covariate shift estimates of prior class probabilities (default probabilities in the case of credit risk) can be material. This may actually be interpreted as an indication of model risk because the assumptions underlying the two estimators are quite different ("identical density ratios" *vs.* "identical conditional class probabilities"). It seems therefore reasonable to calculate both estimates and analyse the potential reasons for the difference if it turns out to be material.

Appendix: Proofs

Proof of Theorem 3. For (a) we have to show that $\int g_i \, d\mu = 1$, $i = 1, \ldots, k$. The other claims are obvious. By (5), Equation (7) implies that $\int g_i \, d\mu = \int g_k \, d\mu$, $i = 1, \ldots, k-1$. This in turn implies

$$1 = \int g \, d\mu = \sum_{i=1}^{k} p_i \int g_i \, d\mu = \int g_k \, d\mu \sum_{i=1}^{k} p_i = \int g_k \, d\mu$$

Hence $\int g_i \, d\mu = 1$ for all i.

With regard to (b), observe that

$$\frac{\partial G_i}{\partial q_j}(q_1, \ldots, q_{k-1}) = - \int \frac{g \, (R_i - 1) \, (R_j - 1)}{\left(1 + \sum_{\ell=1}^{k-1} q_\ell \, (R_\ell - 1)\right)^2} \, d\mu$$

Like in (6), for $(q_1, \ldots, q_{k-1}) \in S_{k-1}$ it can easily be shown that $\frac{\partial G_i}{\partial q_j}$ is well-defined and finite. Denote by $J = \left(\frac{\partial G_i}{\partial q_j}\right)_{i,j=1,\ldots,k-1}$ the Jacobi matrix of the G_i and let $\mathbf{R} = (R_1 - 1, \ldots, R_{k-1} - 1)$. Let $a \in \mathbb{R}^{k-1}$ and a^T be the transpose of a. Then it holds that

$$a \, J \, a^T = - \int g \, \frac{(\mathbf{R} \, a^T)^2}{\left(1 + \sum_{\ell=1}^{k-1} q_\ell \, (R_\ell - 1)\right)^2} \, d\mu \leq 0$$

In addition, by assumption on the linear independence of the $R_i - 1$, $0 = a \, J \, a^T$ implies $a = 0 \in \mathbb{R}^{k-1}$. Hence J is negative definite. From this it follows by the mean value theorem that the solution $(p_1, \ldots, p_{k-1})$ of (7) is unique in S_{k-1}.

By assumption on $(q_1, \ldots, q_k)$ and $h_1, \ldots, h_k$ we have for $i = 1, \ldots, k-1$

$$\begin{aligned}
1 &= \int h_i \, d\mu = \int \frac{g \, h_i}{\sum_{j=1}^{k} q_j \, h_j} \, d\mu = \int \frac{g \, R_i}{1 + \sum_{j=1}^{k-1} q_j \, (R_j - 1)} \, d\mu, \quad \text{and} \\
1 &= \int h_k \, d\mu = \int \frac{g \, h_k}{\sum_{j=1}^{k} q_j \, h_j} \, d\mu = \int \frac{g}{1 + \sum_{j=1}^{k-1} q_j \, (R_j - 1)} \, d\mu
\end{aligned} \tag{18}$$

Hence we obtain $0 = \int \frac{g \, (R_i - 1)}{1 + \sum_{j=1}^{k-1} q_j \, (R_j - 1)} \, d\mu$, $i = 1, \ldots, k-1$. By uniqueness of the solution of (7) it follows that $q_i = p_i$, $i = 1, \ldots, k$. With this, it can be shown similarly to (18) that

$$\int_H h_i \, d\mu = \int_H g_i \, d\mu, \quad H \in \mathcal{H}, \ i = 1, \ldots, k$$

This implies $h_i = g_i$, $i = 1, \ldots, k$. $\square$

Proof of Corollary 6. We start from the first element of the vector-Equation (11b) and apply some algebraic manipulations:

$$
\begin{aligned}
E_1\big[P_0[A_1 \,|\, \mathcal{H}]\big] &= \tfrac{q_1}{P_0[A_1]}\, E_0\big[P_0[A_1 \,|\, \mathcal{H}]^2\big] \\
&\quad + \tfrac{1-q_1}{1-P_0[A_1]}\, E_0\big[P_0[A_1 \,|\, \mathcal{H}](1 - P_0[A_1 \,|\, \mathcal{H}])\big] \\
&= \big(P_0[A_1]\,(1 - P_0[A_1])\big)^{-1}\Big(q_1\,(1 - P_0[A_1])\, E_0\big[P_0[A_1 \,|\, \mathcal{H}]^2\big] \\
&\quad + P_0[A_1]\,(1 - q_1)\,\big(P_0[A_1] - E_0\big[P_0[A_1 \,|\, \mathcal{H}]^2\big]\big)\Big) \\
&= \big(P_0[A_1]\,(1 - P_0[A_1])\big)^{-1}\Big(q_1\,\big(E_0\big[P_0[A_1 \,|\, \mathcal{H}]^2\big] - P_0[A_1]^2\big) \\
&\quad + P_0[A_1]\,\big(P_0[A_1] - E_0\big[P_0[A_1 \,|\, \mathcal{H}]^2\big]\big)\Big)
\end{aligned}
$$

From $P_0[A_1] - E_0\big[P_0[A_1 \,|\, \mathcal{H}]^2\big] = P_0[A_1]\,(1 - P_0[A_1]) - \mathrm{var}_0\big[P_0[A_1 \,|\, \mathcal{H}]\big]$ now the result follows. $\square$

Proof of Corollary 7. Suppose that $g > 0$ is a density of P_1 with respect to some measure ν on $(\Omega, \mathcal{H})$. The measure ν does not necessarily equal the measure μ mentioned in Assumption 4. We can choose $\nu = P_1$ and $g = 1$ if there is no other candidate. By Theorem 3 (a) then there are ν-densities $g_1 \geq 0$, $g_2 > 0$ such that $\frac{g_1}{g_2} = R_1$ and $g = p_1\, g_1 + (1 - p_1)\, g_2$.

We define a new probability measure $\widetilde{P}_0$ on $(\Omega, \mathcal{A})$ by setting

$$
\widetilde{P}_0[A_i] = P_0[A_i], \qquad i = 1, 2
$$

$$
\widetilde{P}_0[H \,|\, A_i] = \int_H g_i \, d\nu, \quad H \in \mathcal{H}, \ i = 1, 2
$$

By construction of $\widetilde{P}_0$ it holds that

$$
P_1[H] = p_1\, \widetilde{P}_0[H \,|\, A_1] + (1 - p_1)\, \widetilde{P}_0[H \,|\, A_2], \quad H \in \mathcal{H}
$$

Hence we may apply Corollary 6 to obtain

$$
E_1\big[\widetilde{P}_0[A_1 \,|\, \mathcal{H}]\big] \;=\; P_0[A_1]\,(1 - \widetilde{R}_0^2) + q_1\, \widetilde{R}_0^2
$$

where $\widetilde{R}_0^2 \in [0, 1]$ is defined like R_0^2 with P_0 replaced by $\widetilde{P}_0$. Observe that also by construction of $\widetilde{P}_0$ we have

$$
\widetilde{P}_0[A_1 \,|\, \mathcal{H}] = \frac{P_0[A_1]\, g_1}{P_0[A_1]\, g_1 + P_0[A_2]\, g_2} = \frac{P_0[A_1]\, R_1}{P_0[A_1]\, R_1 + P_0[A_2]} = P_0[A_1 \,|\, \mathcal{H}]
$$

With the choice $\alpha = \widetilde{R}_0^2$ this proves Corollary 7. $\square$

Conflicts of Interest

The author declares no conflict of interest.

References

1. Tasche, D. The art of probability-of-default curve calibration. *J. Credit Risk* **2013**, *9*, 63–103.

2. Forman, G. Quantifying counts and costs via classification. *Data Min. Knowl. Discov.* **2008**, *17*, 164–206.

3. Redner, R.; Walker, H. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **1984**, *26*, 195–239.

4. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*; Springer: New York, NY, USA, 2006.

5. Schlattmann, P. *Medical Applications of Finite Mixture Models*; Springer: Berlin, Germany, 2009.

6. Moreno-Torres, J.; Raeder, T.; Alaiz-Rodriguez, R.; Chawla, N.; Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **2012**, *45*, 521–530.

7. Hofer, V.; Krempl, G. Drift mining in data: A framework for addressing drift in classification. *Comput. Stat. Data Anal.* **2013**, *57*, 377–391.

8. Hopkins, D.; King, G. A Method of Automated Nonparametric Content Analysis for Social Science. *Am. J. Polit. Sci.* **2010**, *54*, 229–247.

9. Du Plessis, M.; Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Netw.* **2014**, *50*, 110–119.

10. Saerens, M.; Latinne, P.; Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.* **2002**, *14*, 21–41.

11. Peters, C.; Coberly, W. The numerical evaluation of the maximum-likelihood estimate of mixture proportions. *Commun. Stat. Theory Methods* **1976**, *5*, 1127–1135.

12. Titterington, D.; Smith, A.; Makov, U. *Statistical Analysis of Finite Mixture Distributions*; Wiley: New York, NY, USA, 1985.

13. Bella, A.; Ferri, C.; Hernandez-Orallo, J.; Ramírez-Quintana, M. Quantification via probability estimators. In Proceedings of the 2010 IEEE 10th International Conference on Data Mining (ICDM), Sydney, NSW, Australia, 13–17 December 2010; IEEE: Los Alamitos, California, 2010; pp. 737–742.

14. Fawcett, T.; Flach, P. A response to Webb and Ting's On the Application of ROC Analysis to Predict classification Performance under Varying Class Distributions. *Mach. Learn.* **2005**, *58*, 33–38.