On the Use of Prior and Posterior Information in the Sub-Pixel Proportion Problem

Eric D. Kolaczyk

Abstract—Although the problems of classification and sub-pixel proportion estimation in remote sensing land cover characterization generally are held to be distinct (though related), often elements of the former are adopted in addressing the latter that blur this distinction—particularly regarding the use of prior and posterior information. In this paper we examine this issue in more detail, using simple, canonical versions of the two problems and, in the course of this examination, provide analytical expressions upon which to build discussion of improvements to sub-pixel proportion estimation from a statistical viewpoint.

Index Terms—Bayes, land cover classification, and estimation.

I. INTRODUCTION

ACURATE characterization of land cover is an essential part of the process of ecosystem and climate modeling, monitoring, and prediction. A standard approach to producing such characterizations is through the classification of satellite images at the pixel level. However, land cover is a continuum and, with the exception of vast expanses of desert, grassland, or forests, typically is a heterogeneous mixture of different types. This fact has led recently to increased attention in characterizing land cover at a sub-pixel scale (e.g., [17], [3], [11], [13], [2]).

But the sub-pixel proportion problem (which we will call simply the sub-pixel problem\(^1\) here) differs from the perhaps better-known classification problem, in that the latter seeks to label regions (e.g., pixels) as being members of a single, pure land cover class, whereas the former aims to quantify the fraction of each land cover class contributing to the total land cover in the region. As such, they involve inherently different assumptions about the structure underlying a given set of remote sensing measurements. However, a number of authors (e.g., [4], [5], [14]) have had success in applying output from methods derived strictly speaking, for the classification problem to the sub-pixel problem. Specifically, it has been found that posterior class probabilities generated under a mixture model for the classification problem can do a reasonable job in estimating the unknown land cover fractions in the sub-pixel problem.

Nevertheless, inherent in this approach is a blurring of details between the classification and sub-pixel problems and, more fundamentally, of the notion and use of prior and posterior information in the two. In this paper we examine these issues in more detail, using simple, canonical versions of the two problems and, in the course of this examination, provide analytical expressions upon which to build discussion of improvements to sub-pixel proportion estimation from a statistical viewpoint.

Section II contains background on the classification and sub-pixel problems and definition of the basic notation, models, and procedures to be used throughout. Section III shows how the use of posterior class probabilities in the sub-pixel problem may be viewed as an extreme special case of a more general optimal solution to the sub-pixel problem. The implications of some of the formulas arising in this section are then examined in greater detail in Section IV, with the focus being on potential improvements to posterior-based methods for estimation of sub-pixel proportions. Section V contains closing comments.

II. BACKGROUND ON THE CLASSIFICATION AND SUB-PIXEL PROBLEMS

We begin by laying out separate frameworks and notation for canonical versions of the classification and sub-pixel problems, respectively. In Section III we will then examine the implications of using aspects of the former in solving the latter.

A. The Classification Problem.

In its simplest form, the remote sensing pixel classification problem asks, given observations \(x = (x_1, \ldots, x_n)\) for a pixel of interest, that the pixel be classified as one of \(K\) pure classes \(1, \ldots, K\). For example, based on measurements in the \(n = 6\) reflective bands of Landsat TM, one might wish to classify the land cover in a given pixel as being one of the \(K = 4\) classes conifer, hardwood, brush, or barren, as in [8]. The canonical classification model specifies something like the following set of conditions.

C1. To a given region of interest (ROI), associate a vector of proportions \(\pi = (\pi_1, \ldots, \pi_K)\).

C2. For each pixel in the ROI, generate the class membership of that pixel according to the distribution \(f(k) = \pi_k\).

C3. Given class membership \(k\) for a pixel, sample the measurement \(x\) according to a distribution \(f(x|k)\). For simplicity, the above is just a pixel-wise model, with sampling in steps C2 and C3 assumed statistically independent across pixels. The notation \(f(a|b)\) denotes the conditional
probability mass or density function (as the case may be) of \( \alpha \), given \( b \). For convenience we will often write \( f(x|k) \equiv f_k(x) \).

In selecting an optimal class, say \( \hat{k} \), for a pixel, based on the measurement \( x \), it is common to adopt misclassification error as the statistical measure of loss (i.e., where the loss is 0 if the pixel is classified correctly and 1 otherwise). Under the model specified by \( S_1 - S_3 \) and this choice of loss, the Bayes-optimal procedure is known to be to choose the class \( k \) most likely under the posterior distribution i.e.,

\[
\hat{k} = \arg \max_k f(k|x),
\]

where

\[
f(k|x) = \frac{f(x|k) \cdot P(k)}{f(x)} = \frac{f_k(x) \cdot \pi_k}{f(x)}
\]

is the posterior class probability for class \( k \) and

\[
f(x) = \sum_{k=1}^{K} f_k(x) \cdot \pi_k
\]

is the marginal distribution of \( x \) under this model.

### B. The Sub-pixel Problem.

In the remote sensing sub-pixel proportion problem, the belief in pure pixels no longer is felt tenable. Therefore, given a measurement \( x \) for a pixel, interest lies not in selecting a single class membership for that pixel but rather in estimating the proportion of each class \( 1, \ldots, K \) within the pixel. For example, in a pixel composed of what an expert would determine to be 70% conifer, 15% hardwood, 10% brush, and 5% barren land cover, one might wish to recover the numbers \((0.70, 0.15, 0.10, 0.05)\) from, say, Landsat TM measurements.

The distinction between this sub-pixel problem and the classification problem defined previously has implications not only on the specification and interpretation of the underlying model, but also on the choice of loss function and thus ultimately the optimal statistical procedure. A canonical model for the sub-pixel problem might be said to be formulated as follows.

- **S1.** To a given ROI, associate a distribution \( f(\pi^x) \) on proportions \( \pi^x = (\pi_1^x, \ldots, \pi_K^x) \).
- **S2.** For each pixel in the ROI, generate its sub-pixel proportions \( \pi^x \) according to \( f(\pi^x) \).
- **S3.** Given the proportions \( \pi^x \) for a pixel, sample the measurement \( x \) according to a distribution \( f(x|\pi^x) \).

Note, in particular, that there is no variable \( k \) for pixel class in this model, as there was in the classification model \( C_1 - C_3 \) under the assumption of pure pixels. Also, again for simplicity, we assume that the sampling in steps S2 and S3 is performed in a statistically independent manner between pixels.\(^2\)

When reporting numerical results in the literature to illustrate the performance of proposed procedures for the sub-pixel problem, it is common practice to use mean-squared error (e.g., \([2]\)). This usage implies the adoption of squared-error as the measure of statistical loss in estimating \( \pi^x \) by, say, \( \hat{\pi}^x \).

Under the model specified by \( S_1 - S_3 \) and this choice of loss, the Bayes-optimal procedure is known to be the mean of the posterior distribution i.e.,

\[
\hat{\pi}^x \equiv \mathbb{E}[\pi^x|x] = \int \pi^x f(\pi^x|x)d\pi^x,
\]

where

\[
f(\pi^x|x) = \frac{f(x|\pi^x) \cdot f(\pi^x)}{f(x)}
\]

is the posterior distribution of \( \pi^x \), given \( x \), and

\[
f(x) = \int f(x|\pi^x) \cdot f(\pi^x)d\pi^x
\]

is the marginal distribution of \( x \) under this model. Here and throughout the notation \( \mathbb{E}[\cdot] \) will denote statistical expectation, with \( \mathbb{E}[\cdot|\cdot] \) indicating expectation with respect to a conditional distribution (as in \((4)\)).

### III. ON ADAPTING ELEMENTS OF THE CLASSIFICATION MODEL TO THE SUB-PIXEL PROBLEM

As stated in Section I, it is not uncommon to see in the literature on the sub-pixel problem the use of posterior class probabilities to estimate sub-pixel proportions. In this section we show how this approach implicitly entails a curious “blending” of elements of the classification and sub-pixel models defined above.

#### A. Posterior Mean Under Mixture Model Assumptions

In specifying the sub-pixel model \( S_1 - S_3 \) above, no specific form was given to the (conditional) distribution governing the taking of measurements i.e., \( f(x|\pi^x) \). Suppose we now assume a finite mixture model (in the sense of \([10]\)), that is, that

\[
f(x|\pi^x) = \pi_1^x f_1(x) + \ldots + \pi_K^x f_K(x)
\]

is a linear mixture of densities \( f_k(x) \) in proportions \( \pi_k^x \). This form is in fact analogous to that induced for \( f(x) \) in the classification problem through assumptions \( C_1 - C_3 \), were one to integrate out (marginalize) the role of the random variable \( k \) (i.e., equation \((3)\)).

It is useful to examine the form taken by the posterior in \((4)\) now, under the addition of the assumption in \((7)\) to the model \( S_1 - S_3 \). A short sequence of probability calculations and some algebra yields that the \( k \)-th (scalar) component of the posterior mean has the form

\[
\pi_k^x \equiv \mathbb{E}[\pi_k^x|x] = \sum_{k'=1}^{K} \frac{f_k(x) \cdot \mathbb{E}[\pi_k^x|\pi_{k'}^x]}{f(x)}
\]

\[
= \sum_{k'=1}^{K} \frac{f_k(x) \cdot \mathbb{E}[\pi_k^x|\pi_{k'}^x]}{f(x)}.
\]

(Details may be found in the appendix.)

The expression in \((8)\) may at first glance appear to be relatively uninformative. However, under \( S_1 - S_3 \) and the assumption in \((7)\), note that \( \mathbb{E}[\pi_k^x] = \Pr(x \text{ sampled from } f_k) \).
Writing this probability as \( \hat{f}(k) \), one can replace the terms in the first expression inside the summation in (8) with ones of the form
\[
\hat{f}(k|x) \equiv \frac{f_k(x) \hat{f}(k)}{f(x)}.
\] (9)

If one then chooses to interpret \( \hat{f}(k) \equiv E[\pi_k^g] \) as a prior probability for class \( k \), in analogy to \( f(k) \equiv \pi_k^g \) in the classification problem, the expression for \( \hat{f}(k|x) \) resembles a posterior probability, in the spirit of \( f(k|x) \) in equation (2) in the classification problem. But since, as we have seen, the notion of requiring pure pixels (and hence the sampling of a class \( k \)) are not a part of (nor, indeed, even consistent with) the sub-pixel problem, we will call the values in (9) pseudo-posterior probabilities. Their presence is simply an interesting artifact of (i) the use of the finite mixture model, and (ii) our interpretation of \( \hat{f}(k) \).

Nevertheless, having made this association between posterior and pseudo-posterior probabilities, if we let \( \beta_{k,k'} = E[\pi_k^g \pi_{k'}^g] / E[\pi_k^g] \) we can then re-express (8) in the form
\[
\hat{\pi}_k^g = \sum_{k'=1}^K \beta_{k,k'} \hat{f}(k'|x).
\] (10)

In other words, the Bayes-optimal estimator in the sub-pixel problem now resembles a weighted combination of posterior probabilities.

### B. A Special Case

Both the pseudo-posterior probabilities \( \hat{f}(k'|x) \) and the weights \( \beta_{k,k'} \) depend upon the choice of prior distribution placed on the sub-pixel proportions \( \pi^g \) i.e., \( f(\pi^g) \). To better understand the role of this prior and, ultimately, its effect on the estimates \( \hat{\pi}^g \), it is useful to examine a special case.

The Dirichlet distribution is a standard choice of prior for a vector of positive weights constrained to sum to one \([7]\), and one that proves to be rather instructive in the current setting. The Dirichlet density function takes the form
\[
f(\pi^g) = \frac{\Gamma(\alpha_1 + \cdot + \cdot + \alpha_K)}{\Gamma(\alpha_1) \cdot \cdot \cdot \Gamma(\alpha_K)} (\pi_1^{\alpha_1-1} \cdot \cdot \cdot \cdot \pi_K^{\alpha_K-1}) ,
\] (11)

with parameter \( \alpha = (\alpha_1, \cdot, \alpha_K)^t \), where \( \alpha_k > 0 \) for each \( k \). From this form it follows that, setting \( \alpha^* = \sum_{k=1}^K \alpha_k \), we have
\[
E[\pi_k^g] = \frac{\alpha_k}{\alpha^*}.
\] (12)

and
\[
E[\pi_k^g \pi_{k'}^g] = \begin{cases} 
\frac{\alpha_k (\alpha_{k'} + 1)}{\alpha^*(\alpha^* + 1)}, & \text{if } k = k', \\
\frac{\alpha_k \alpha_{k'}}{\alpha^* (\alpha^* + 1)}, & \text{if } k \neq k'.
\end{cases}
\] (13)

Substituting these values appropriately in the definition of \( \beta_{k,k'} \), and some algebra, yields that for the Dirichlet prior distribution the expression in (10) becomes
\[
\hat{\pi}_k^g (\text{Dir}) = \frac{1}{1 + \alpha^*} \cdot \hat{f}(k|x) + \frac{\alpha^*}{1 + \alpha^*} \cdot E[\pi_k^g].
\] (14)

In other words, for this particular choice of distribution on \( \pi^g \) in S1 of the sub-pixel model, the Bayes optimal estimators for each component of \( \pi^g \) are given as a simple convex linear combination of (i) the expected value under the prior and (ii) the pseudo-posterior class probability.

Now consider the role of \( \alpha^* \) in (14). For large \( \alpha^* \) the sub-pixel estimator is governed primarily by the prior mean, while for small \( \alpha^* \) the pseudo-posterior is dominant. Hence this parameter dictates the relative weight given to the prior information (through \( E[\pi_k^g] \)) and the data (through \( \hat{f}(\pi_k^g|x) \)) in producing the final estimate of the proportion of sub-pixel land covered by class \( k \).

In particular, consider the special case in which some original choice of \( \alpha \) is scaled by a value \( \gamma > 0 \) i.e., \( \alpha \) is replaced by \( \gamma \alpha \), and we let \( \gamma \rightarrow 0 \). Then the components \( E[\pi_k^g] \) and \( \hat{f}(\pi_k^g|x) \) in (14) remain unchanged as \( \gamma \) decreases, but \( \alpha^* \rightarrow 0 \), and therefore \( \hat{\pi}_k^g \rightarrow \hat{f}(\pi_k^g|x) \). That is, the optimal procedure is just to use the individual pseudo-posterior class probabilities alone as estimators. However, it should be noted that for this extreme case, where \( \gamma \rightarrow 0 \), the corresponding sequence of Dirichlet distributions tend in their limit to a discrete distribution giving positive probability only to the set of all \( K \)-length vectors composed of \( K-1 \) zeros and one 1. And this limiting case for the prior distribution \( f(\pi^g) \) in fact dictates that only pure pixels are generated.

In other words, under the statistical model formed by S1–S3, the finite mixture distribution in (7), and this limiting case of the Dirichlet distribution, the resulting sub-pixel model effectively reduces to the classification model C1 – C3, in which case the posterior class probabilities are optimal.

### IV. ON IMPROVING POSTERIOR-BASED SUB-Pixel ESTIMATION

It is useful to consider the results of the previous section in more detail, particularly as they motivate discussion of how improvements can be made to the naive usage of individual posterior class probabilities in estimating sub-pixel proportions.

#### A. An Illustration

For the purpose of illustration, we will first restrict our attention to the specific form of (10) posed by (14), and frame our discussion within the context of the work of Ju et al. [8]. These authors address the problem of estimating the sub-pixel fractions of conifer, hardwood, and barren/brush land cover for regions of the Plumas National Forest of California.

Of primary relevance to our discussion here is that posterior class probabilities were used directly by Ju et al. to estimate the unknown sub-pixel fractions. (See [4], [5], [14] for similar usage.) That is, the fractions \( \pi^g \) in the sub-pixel model S1 – S3 were estimated using the probabilities \( f(k|x) \) that resulted from training the classification model C1 – C3. More specifically, class-specific densities \( f_k(x) \) of a certain form (motivated by the framework of mixture discriminant analysis (MDA) outlined in [6]) were fit for each of the \( K = 3 \) land cover classes, using the combination of field measurements of land cover and corresponding Landsat TM image data (in \( n = 6 \) spectral bands) first collected and processed by Carpenter et al. [2]. Prior probabilities \( \pi^c \) were chosen based on aggregate class characteristics over the training set.
Posterior probabilities $f(k|x)$ were then formed according to (2).

Recalling the discussion ending the previous section, note that this approach is equivalent to using the estimator (14) in the special case of $\alpha^* = 0$ (or $\eta \equiv (1/(1 + \alpha^*)) = 1$), with $\pi^c$ substituted for the quantity $E[\pi^c]$ in $f(k|x)$. Pursuing this observation further, and viewing $\eta$ as just an arbitrary tuning parameter (as opposed to being defined through the Dirichlet parameter $\alpha$), we are led to consider estimators of the form

$$\tilde{\pi}_k(\text{Illus}) \equiv \eta \cdot f(k|x) + (1 - \eta) \cdot \pi^c_k,$$  

(15) where $\eta \in [0,1]$, as ad hoc versions of the estimators in (14).

It is then illustrative to explore whether varying the parameter $\eta$ can yield estimators that improve upon use of the posterior class probabilities alone. Using the same test data as Ju et al., Figure 1 shows plots of the root mean-squared error (RMSE) for the collection of all estimators in (15) as a function of $\eta$. The performance of the estimator of Ju et al. corresponds to the values of the three curves at the far right hand side of the plot, yielding RMSEs of 21.9%, 11.4%, and 22.3%, for conifer, hardwood, and barren/brush, respectively. However, the total RMSE over the three categories is minimized at $\eta = 0.717$ (i.e., $\alpha^* = 0.394$), yielding RMSEs of 18.2%, 12.3%, and 19.5%. That is, the optimal choice of $\eta$ puts roughly $7/10$’s of the weight of the overall estimator on the pseudo-posterior (i.e., the quantity used by Ju et al.) and the remaining $3/10$’s on the prior probability $\pi^c_k$ (which, recall, is being used as a stand-in for $E[\pi^c_k]$ here).

The end result is, compared to the naive use of (pseudo) posterior probabilities, a notable decrease in error for the conifer and barren/brush classes, at the expense of only a slight increase in error for hardwood. In fact, these numbers are nearly equivalent to the error obtained by the ARTMAP neural network approach in the original analysis of this data by Carpenter et al. [2] – specifically, 18%, 12%, and 18%. Therefore, this illustration provides empirical evidence of the potential to improve upon the use of relatively simple posterior-based methods through the use of estimators like (14) or (10), and that the level of performance to be achieved may possibly rival that of comparatively complex neural network methods. We comment on these two estimators themselves more specifically in the remainder of this section.

B. Discussion of the Special Case: Equation (14)

From a theoretical perspective, the potential performance improvement just noted is not entirely a surprise. For example, the estimator (14), and its ad hoc analogue (15), as convex combinations of observed and prior information, have the form of so-called shrinkage estimators, representing one of the more powerful developments in statistical decision theory of the last 50 years (see [15], for example, for an intuitive introduction and references). Such estimators tend to take some standard estimator (like a sample mean, or here the pseudo-posterior) and “shrink” it towards an alternative estimator, such as one that might be used if only prior information were available (like a prior mean, or here the quantity $E[\pi^c]$). In a wide variety of settings such estimators have been found to uniformly offer improvement over the standard estimator they incorporate, generally by reducing the variance of that estimator without introducing excessive bias.

However, in order to implement the estimator in (14) explicitly, as opposed to the ad hoc version (15) used in the above illustration, there are both data and statistical hurdles to be overcome (which are beyond the aims and scope of this short paper). From the perspective of statistical modeling, it is necessary to elicit an appropriate member of the class of Dirichlet distributions for the prior, which can be reduced to choosing the parameter vector $\alpha$ in such a manner that the dynamics of the sub-pixel proportions $\pi^k$ across the region of interest are matched sufficiently well according to some criteria (e.g., maximum likelihood). But in order to pursue this task, the necessary data must be available. Specifically, there must be available representative measurements on land cover class proportions at the sub-pixel level; and these generally are lacking in standard land cover studies.

This lacking essentially reduces to a matter of scale: the scale at which measurements are taken, versus the scale at which statistical inferences are to be made. In fact, this issue also highlights an important point regarding the scale to which prior information pertains in the classification and sub-pixel problems, and is the reason for our having differentiated between $\pi^c$ and $\pi^k$, respectively, in our models. The former, as an image-level quantity, typically is inferred from ground truth data at scales consistent with that of the image or coarser, whereas the latter, as a pixel-level quantity, pertains to information aggregated over the sub-pixel level. Therefore, in using the posterior class probabilities (2) directly as estimators of sub-pixel proportions, not only are information from the prior and data not being combined in the fashion dictated optimal under mean squared error (e.g., equation (14)), but that prior information that is incorporated comes from a scale coarser than that at which inferences are to be made. The importance of this general issue has been the focus of a number of authors (e.g., [9], [12], [16]).
C. Discussion of the General Case: Equation (10)

Now consider equation (10) in its generality. The various issues raised in the discussion of (14) hold here as well. But while this general formula is perhaps less transparent than that in (14), it also is potentially more rich in its capacity to offer improvement. Specifically, whereas the formula in (14) can be viewed as augmenting each class-specific pseudo-posterior with first moment (i.e., mean) information about the prior \( f(\pi^*) \), that in (10) can be viewed as augmenting with both first and second moment (i.e., mean and covariance) information about the prior and, additionally, using that information to incorporate relevant information from the pseudo-posterior probabilities of all other classes.

That is, the Bayes-optimal estimate \( \hat{\pi}^*_k \) of the sub-pixel proportion \( \pi^*_k \) of class \( k \) is a linear combination involving not only the pseudo-posterior \( f(\pi^*_k|x) \) for class \( k \), but also those for all other classes i.e., \( f(\pi^*_k|x) \) for \( k' \neq k \). And the weights used in combining these values, the \( \beta_{kk'} \), are proportional to \( E[\pi^*_k|\pi^*] \), the (uncentered) covariance of the sub-pixel proportions \( \pi^*_k \) and \( \pi^*_k \) of classes \( k \) and \( k' \), respectively. Thus the sign and magnitude of correlations between \( \pi^*_k \) and the other \( \pi^*_k \) dictate whether incorporation of their pseudo-posteriors increases or decreases the estimate based just on the single pseudo-posterior class probability, and to what degree.

For example, generally speaking, if two sub-species of conifer are included in the classes \( 1, \ldots, K \) and known to be positively correlated in their prevalence, the expression in (10) would increase the overall estimated fraction of each by the other. Conversely, were either (or both) of these conifer sub-species to be negatively correlated with, say, a certain class of hardwood species, the estimator (10) differs from the use of a single pseudo-posterior class probability in allowing the weight of posterior evidence for the hardwood species to negatively influence the overall estimate for the fraction of the conifer species.

Interestingly, from a practical point of view, it is perhaps worth emphasizing that the form of (10) suggests that benefits can be expected from successfully modeling only the prior mean and covariance information for a given set of instruments and region of interest which, although challenging in and of itself, likely is much less daunting of a task than seeking to capture the form of the full distribution \( f(\pi^*) \).

Alternative choice of loss function (such as weighted squared-error loss) may be considered as well. Ultimately, however, it would seem that the most important challenge lies in how best to elicit useful prior information to be used in conjunction with the resulting estimation strategies.

APPENDIX

PROOF OF EQUATION 8

Begin with the expression in (4), and assume that equation (7) governs the form of the measurement distribution. The posterior \( f(\pi^*|x) \) is equal to the ratio of \( f(\pi^*,x) \) and \( f(x) \), and the former may be written as

\[
 f(\pi^*,x) = \sum_{k'=1}^{K} f(\pi^*,x,k') = \sum_{k'=1}^{K} f(x|k',\pi^*) f(k'|\pi^*) f(\pi^*) .
\]

Using this result to re-write (4), we get that

\[
 \pi^*_k = \sum_{k'=1}^{K} f(x|k',\pi^*) \int \pi^*_k \ f(k'|\pi^*) \ f(\pi^*) \ d\pi^* .
\]

The rest of the proof follows from noting that \( f(x|k',\pi^*) = f_k(c|x) \) and \( f(k'|\pi^*) = \pi^*_{k'} \), by definition, and that

\[
 E[\pi^*_k \pi^*_{k'}] = \int \pi^*_k \ f(\pi^*) \ d\pi^* .
\]

Finally, one multiplies top and bottom of each term inside the summation by \( E[\pi^*_k] \). Note that \( k \) here is used simply to denote the operable component of \( f(x|\pi^*) \), under sampling with respect to (7), a standard device when working with finite linear mixture models (e.g., [10, pg. 7]), as opposed to it being an assumption of pure pixels as in the classification model.

ACKNOWLEDGMENT

The author would like to thank Mark Friedl, Sucharita Gopal, and Alan Strahler for their feedback on earlier drafts of this manuscript. Thanks also to Junchang Ju for help in producing Figure 1.

REFERENCES


