Linear mixing and the estimation of ground cover proportions

J. J. SETTLE & N. A. DRAKE

a Department of Geography, NERC Unit for Thematic Information Systems, University of Reading, Whiteknights, Reading RG6 2AB, England

b Department of Geography, University of Reading, Whiteknights, Reading RG6 2AB, England

Available online: 24 Oct 2007
Linear mixing and the estimation of ground cover proportions

J. J. SETTLE
NERC Unit for Thematic Information Systems, Department of Geography,
University of Reading, Whiteknights, Reading RG6 2AB, England

and N. A. DRAKE†
Department of Geography, University of Reading, Whiteknights, Reading
RG6 2AB, England

(Received 15 March 1990; in final form 1 October 1991)

Abstract. In this paper we consider how we may determine the relative
proportions of ground cover components in a mixed pixel. We assume the usual
linear model for signal mixing and examine a number of methods, closely related,
for estimating the proportions. We also show how the precision of our estimates
can be defined. We introduce a new estimator which is based on regularisation
principles and which produces a smoother set of images than other methods, and
gives more accurate estimates. The methods are compared on a simulated data
set.

1. Introduction

The most widely used method for extracting information on surface cover from
remotely-sensed data is image classification. With this technique each image pixel is
allocated exclusively to one of a small number of known categories, producing an
image containing thematic information; one in which, for example, regions of water
or forest or arable crops would be depicted. The resulting thematic map can be an
appropriate interpretative aid for cases where the number of boundary pixels is small
and the scene is nicely partitioned into regions of homogeneous cover types. Many
agricultural scenes satisfy this requirement and counting the pixels allocated to each
cover type can give a fair estimate of the acreage occupied by each category.

However, classification is not always suitable and the classification map will
often be a poor representation of reality. Perhaps the most important example of this
is when the size of a ground pixel (i.e., the piece of ground that is viewed by an image
pixel) is comparable to, or much larger than the natural size of ground cover units
(as would be the case if we wanted to use AVHRR data for agricultural surveys).
Also, traditional classification is a far from ideal tool for the study of regions of
natural or semi-natural vegetation whenever there is a continuous variation in
ground surface cover, or for studying surface materials in desert regions. Geologists
are accustomed, in any case, to working with compositional rather than thematic
data.

†Present address: Department of Geography, King's College London, Strand, London
WC2 1LS.

0143-1161/93 $10.00 © 1993 Taylor & Francis Ltd
A more suitable way of extracting information from such scenes is to try to estimate how each ground pixel's area is divided up among different cover types. This approach is usually known as mixture modelling in the limited amount of work devoted to the subject in the remote sensing literature. What we are seeking is not a single map of symbols (the classification image) but a series of images, each the size of the original image, and each giving a map of the concentration of a different cover type across the scene. The result is then a mass of quantitative, not thematic, data. The early work in this area (e.g., Horwitz et al. 1971, Detchmendy and Pace 1972) was devoted to Landsat MSS data and agricultural contexts and much of the essential theory was laid down in those projects; however, the early promise of this approach was not then followed up to any great degree. In recent years interest in the unmixing of mixed signals has been revived, inspired largely by the efforts of a group at Washington State University (e.g., Smith et al. 1985, 1987) and by work carried out by CSIRO scientists in Australia (e.g., Pech et al. 1986).

This article has three aims. The first is to act as a general introduction to the problem of unmixing. The second is to show how the likely errors on standard estimates may be calculated; these are shown to have simple interpretations in terms of feature space geometry. The third aim is to introduce a new estimator. This is based upon regularisation techniques and generally produces smoother images of proportions than those obtained by unmixing signals on a pixel-by-pixel basis; we further show that for a range of values of the smoothing parameter the new approach gives more accurate answers than the classical approaches.

2. Estimation and calibration

Our task is to estimate for each pixel in the image the proportions of that pixel covered by each ground category given the multi-spectral observations. We shall denote these proportions by the $c \times 1$ vector \( \mathbf{f} = (f_1, f_2, \ldots, f_c)^T \), where \( c \) will stand throughout this paper for the number of ground cover components. The numbers \( f_i \) should strictly be non-negative and add to unity; however, there are certain advantages in not insisting on these constraints and so we shall generally be considering methods in which they are not built in \textit{ab initio}. The way the problem of obtaining proportions has generally been tackled to date has been by forming an estimator, \( \mathbf{\hat{f}} \), a function of the multi-variate point information for the pixel in question; in our case this is just the set of digital values for that pixel, which we shall denote by the \((n \times 1)\) vector \( \mathbf{x} = (x_1, x_2, \ldots, x_n)^T \). Here, and in the rest of this article, \( n \) will denote the number of spectral bands in our remotely-sensed image. We shall only be considering estimators such as this. More complicated approaches are possible; we might attempt for example to define an estimator which is a function not only of the signal for the pixel of interest, but also of the signals of neighbouring pixels. We might, alternatively, attempt to solve for more than one pixel simultaneously, modelling the way that the proportions change spatially. However, these possibilities lie beyond the scope of the present article.

A desirable estimator would be based on a sensible model of how a given mix of components gives rise to the pixel's signal, and the signal corresponding to the preferred mixture should be close to that observed. Ideally we would know exactly a function \( X(\mathbf{f}) \) which we might then invert to obtain our estimate \( \mathbf{\hat{f}}(\mathbf{x}) \). In practice we do not know \( X(\mathbf{f}) \) except for a simple case to be described below (the linear mixing model) and even then the function cannot always be inverted unambiguously; there may be more than one \( \mathbf{f} \) that can give rise to a given \( \mathbf{x} \). This happens when we have
more ground cover components than spectral bands, so one is in some sense asking for more information from the image (the proportions of the different cover types) than is present (the signals received by the different bands). The function $x(f)$, if it exists, will depend at the very least upon the reflectance spectra of the cover types of which the scene is composed. However, inadequate calibration of an image and atmospheric contamination often mean that data cannot be reduced to reflectance values and so laboratory reflectances are of limited value. A much more likely situation is that we have detailed field knowledge of proportional ground cover for a number of pixels. There are then a number of pixels for which both $f$ and $x$ are known, and a model of some sort can be used to deduce an appropriate expression for $f$, given $x$. Generalised at this level this is a statistical multivariate calibration problem (e.g., Brown 1982).

A review by Brown (1982) stressed three methods of multi-variate calibration, which have been examined in this context by Pech et al. (1986). One of these, known as the Lwin-Maritz estimator (Lwin and Maritz 1980), gives rise to a non-linear method of estimating proportions (the formula for $f$ given $x$ is a non-linear function of $x$); while this is a method favoured by Pech et al. (1986) it was the only one found to give bad results by Brown when applied to his test problem and it will not be considered further in this paper. The 'classical' method is equivalent to assuming a linear model for the generation of the composite signal, using training data to find the constants of this model, and then using that model to predict proportions for other pixels. It happens that the resulting formula gives the proportions as linear functions of the elements of the observation, and the third method of calibration considered by Brown fixes on this fact; it looks specifically for a model for the proportions that is just such a linear function of the signal and aims to find the best such model. Essentially, it is a matter of carrying out on the training data a linear regression of the $f$ values on the $x$ values. It is called 'inverse estimation' by Lwin and Maritz (1980). The classical estimator is perhaps the most preferable, since it is based on a sensible physical model of composite reflectance, namely the linear mixture model. It is the model which has been most commonly used or implicitly assumed in previous studies of mixing in remote sensing.

3. The linear mixture model

The basic physical assumption underlying the linear mixing model is that there is no significant amount of multiple scattering between the different cover types; each photon that reaches the sensor has interacted with just one cover type. Under these conditions the received energy can be considered a simple sum of the energies received from each cover component. Each field, or part field, in the scene will contribute an amount to the received signal which is characteristic of the cover type in that field and proportional to the area covered. These ground cover proportions are well-defined and add to unity. Let the vector $\mathbf{\mu}_i = \{\mu_{i1}, \mu_{i2}, \ldots, \mu_{in}\}^T$ denote the expected signal from pixels containing just the $i$th cover class. The expected value of the signal for a mixed pixel is, under the linear mixing assumption, given by the formula:

$$f_1\mathbf{\mu}_1 + f_2\mathbf{\mu}_2 + \cdots + f_n\mathbf{\mu}_n = \mathbf{M}f$$

(1)

The columns of the $(nxc)$ matrix $\mathbf{M}$ are the vectors $\mathbf{\mu}_i$. These quantities we shall henceforth refer to as the end-member spectra, and the categories themselves as the end-members. The term is a standard expression in geochemical studies; we adopt it...
here to mean the response that would be received in the absence of noise by a pixel containing nothing but the component of interest. Its close relationship to the concept of the 'class mean', that is such an important part of image classification, is obvious. These (largely hypothetical) pure pixels will not all have signal values corresponding to the appropriate end-member spectrum, but because of sensor noise will exhibit statistical fluctuations characterised by a variance-covariance matrix $N_i$. In consequence, pixels with the mixture $f$ will exhibit fluctuations around their mean value $MF$, characterised by a covariance matrix $N(f)$, given by:

$$N(f) = f_1N_1 + f_2N_2 + \ldots + f_cN_c$$  \hspace{1cm} (2)

provided there is no correlation between the signals emanating from the different fields within the instantaneous field of view of the sensor; this has been shown to be a good approximation by Horwitz et al. (1975). To keep the problem tractable we shall assume that each $N_i$ is equal to a common noise matrix $N$, so that $N(f)$ is independent of $f$ and the linear model is defined by:

$$x = MF + e$$  \hspace{1cm} (3)

where $e$ is a vector of errors satisfying

$$E(e) = 0 \text{ and } E(ee^T) = N,$$  \hspace{1cm} (4)

$E$ denoting expectation. We shall assume that the matrix $M$ is of full rank, consideration of all the possibilities that arise from assuming that this may not be so is not particularly useful, corresponding as it does to a poorly chosen set of end-members.

As with classification, it is instructive to consider the geometry of feature space, that is, the Euclidean space defined by the components of measurement vectors. Each of the end-member spectra are points in this space, as is each observation $x$. Pixels that are mixtures of just two components will, in the absence of noise, lie in feature space along the line joining the points corresponding to the two end-member spectra. Similarly, mixtures of three components will be found in the plane defined by the three end-member spectra: with non-negative proportions they should lie inside this triangle. Generalising, in the absence of noise mixture pixels lie in the hyperplane defined by the end-member spectra of the components contributing to the mixture. All possible legitimate mixtures, those with non-negative proportions, fill the volume enclosed by the generalised polyhedron whose vertices are points corresponding to the end-member spectra.

To carry out the calculations we need to know the quantities $\mu_i (i = 1, \ldots, c)$ and $N$. The situation is very similar to that in supervised classification, where we need to know precisely these quantities to carry out a maximum likelihood classification. Ideally we should like to be able to use laboratory based measurements to define the end-member spectra but there are substantial problems in correcting satellite data sufficiently well for atmospheric effects to allow direct comparison between laboratory data and the digital numbers of, for example, a Landsat image. Nevertheless, this is the novel approach adopted by Smith et al. (1987), who solve for the proportions by using laboratory spectra and adopting a linear mixing model together with a form of correction for the gains and offsets brought about by atmospheric scattering and absorption. For the everyday user of remotely-sensed data, however, it is more likely to be the case that, as in the case of supervised classification, we have a number of training pixels from which the end-member
spectra, and \( N \) must be inferred; appropriate formulas for these calculations are given in appendix A. We do not always have useful ground cover information when we set out to classify an image and so may be driven to use methods of unsupervised classification. Similarly, our ground reference knowledge may be imperfect when we wish to estimate ground cover proportions. In these circumstances the end-member spectra must somehow be inferred from the image data; this is an extremely difficult task. A number of methods may be attempted to determine the possible feature space locations of the end-member spectra; (Smith et al. 1985, Craig, 1990, Mackin et al. 1990).

3.1. The mixing equations

An exact match of the observation \( x \) to the quantity \( Mf \) for some suitable \( f \) can be made if we satisfy the following system of \((n+1)\) linear equations, one for each dimension of the feature space and a constraint equation. These are:

\[
\begin{align*}
  f_1 \mu_{11} + f_2 \mu_{21} + \ldots + f_n \mu_{n1} &= x_1 \\
  f_1 \mu_{12} + f_2 \mu_{22} + \ldots + f_n \mu_{n2} &= x_2 \\
  \vdots \\
  f_1 \mu_{1n} + f_2 \mu_{2n} + \ldots + f_n \mu_{nn} &= x_n \\
  f_1 + f_2 + \ldots + f_n &= 1
\end{align*}
\]

(5)

where \( \mu_{ij} = (M)_{ij} \) is the \( j \)th vector element of the end-member spectrum for the \( i \)th ground cover category. We shall refer to the first \( n \) of these as the mixing equations and the last as the sum-to-one constraint. In matrix-vector notation they may be written:

\[
\begin{align*}
  Mf &= x \quad \text{if} \quad f_1 + f_2 + \ldots + f_n = 1
\end{align*}
\]

(6)

From the theory of linear equations we know that in general a solution to (6), if one exists, will be unique if \( c = n + 1 \), and if \( c > n + 1 \) there will be an infinity of exact solutions. Finally, when \( c < n + 1 \) there may well be no exact solution. The important condition \( c \leq n + 1 \) that there should not be an infinity of solutions is evocatively named the 'condition of identifiability' by Kent and Mardia (1986). The same term is used in the study of unscrambling mixture populations (Titterington et al. 1985), where we have a number of observations from a given mixture. The condition must be modified to \( n < c \) if we are relaxing the sum-to-one constraint; for the rest of the paper we shall refer always to \((n+1)\) when discussing this problem although it is implicit that the slightly different condition may be relevant.

The superfluity of solutions when \( c > n + 1 \) is an embarrassment and there is no clear way in which we should attempt to define any sort of best solution from among the infinity of possibilities. Approaches to this problem will be considered elsewhere, but for the rest of this article we shall mainly restrict our attention to the case where we have fewer components than bands, and where as a consequence we know that we will not be faced with this infinity of legitimate solutions. Before we do this, though, we should point out that the relevant number is not really the number of spectral bands as such, but rather the intrinsic dimensionality of the spectra data, as might be revealed by principal components analysis. In other words, if we are dealing with the six reflective bands of Landsat Thematic Mapper data, and find that the fifth and sixth principal components of the data contained nothing but noise, then the true dimensionality of the data is four, and not six. While we might have
more components of interest than this, the linear model tells us that we cannot have fewer than that number in the scene.

When we have fewer unknowns than equations we cannot usually find \( \mathbf{f} \) such that the error vector \( \mathbf{e} \) is zero, and the problem we face is to find an approximate solution to the mixing equations, or to find an estimator \( \hat{\mathbf{f}} \) for which some appropriate measure of the error is suitably small. We might decide to try a simple least squares approach, where we would look for the vector of proportions which minimises the magnitude of \( \mathbf{e} \), i.e., for the set of \( \{ f_i \} \) that minimises the quantity

\[
(x-Mf)^T(x-Mf) = e^T e = e_1^2 + e_2^2 + \ldots + e_k^2
\]

(7)

More generally we might try to minimise the quadratic form:

\[
Q(x, f) = (x-Mf)^T \Sigma^{-1} (x-Mf)
\]

(8)

where \( \Sigma \) is some positive definite matrix; (7) is clearly of this form with \( \Sigma \) given by the identity matrix. This more general approach has been argued for on statistical grounds by Horwitz et al. (1971), with \( \Sigma \) the noise matrix \( \mathbf{N} \).

3.2. The classical estimator, and its variations

Without any constraints of \( \mathbf{f} \) we minimise the quadratic \( Q \) by setting to zero its partial derivatives with respect to each of the \( f_i \); this gives us the estimator:

\[
\hat{\mathbf{f}} = \mathbf{UM}^T \mathbf{N}^{-1} \mathbf{x} =: f_0, \text{ say}
\]

(9)

where the \( c \times c \) matrix \( \mathbf{U} \) is defined by \( \mathbf{U}^{-1} = \mathbf{M}^T \mathbf{N}^{-1} \mathbf{M} \). If we replace \( \mathbf{N} \) by \( \mathbf{I} \) in the definition of \( Q \), then the corresponding expression is:

\[
\hat{\mathbf{f}} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{x}
\]

(10)

which is the standard expression for the least squares approximation for the system of equations \( \mathbf{Mf} = \mathbf{x} \); it is thus the classical (least-squares) estimator for \( \mathbf{f} \) (Lwin and Martiz 1980). Expression (9) is just the natural extension of this for the case where the noise term, \( \mathbf{e} \) in (3), is correlated. If we impose the sum-to-one condition, then a standard Lagrangian analysis gives us the slightly different solution:

\[
\hat{\mathbf{f}} = \alpha (1 - \mathbf{j}^T \mathbf{f}_0) \mathbf{U} \mathbf{j} + \mathbf{f}_0
\]

\[
= \alpha \mathbf{U} \mathbf{j} + (1 - \alpha \mathbf{U} \mathbf{j}) \mathbf{f}_0
\]

\[
= \alpha \mathbf{U} \mathbf{j} + (\mathbf{U} - \alpha \mathbf{U} \mathbf{j}) \mathbf{M}^T \mathbf{N}^{-1} \mathbf{x}
\]

(11)

where \( \mathbf{I} \) is the \( c \times c \) identity matrix, \( \mathbf{J} (= \mathbf{j} \mathbf{j}^T) \) is a \( c \times c \) matrix consisting entirely of 1's and the constant \( \alpha \) is equal to \( \mathbf{j}^T \mathbf{U} \mathbf{j}^{-1} \); \( \alpha^{-1} \) is equal to the sum of the elements of \( \mathbf{U} \). Finally, we can obtain the solution when \( \mathbf{j}^T \mathbf{f} \) is constrained to lie within a range of values, say \( (k_1, k_2) \); it is:

\[
\hat{\mathbf{f}} = \mathbf{f}_0
\]

\[
= k_1 \alpha \mathbf{U} \mathbf{j} + (1 - \alpha \mathbf{U} \mathbf{j}) \mathbf{f}_0
\]

\[
= k_2 \alpha \mathbf{U} \mathbf{j} + (1 - \alpha \mathbf{U} \mathbf{j}) \mathbf{f}_0
\]

(12)

A reasonable pair of values for \( k_1 \) and \( k_2 \) might be 0 and 1, if we were happy that we had entirely removed atmospheric effects from our image. The variation in the sum of the proportions would correspond to variations in the amount of illumination received by each pixel; dividing each of the component proportions by their sum would then give a set of proportions in the correct ratios and that add to one.
Mixed pixels and ground cover

The simple linear relationship between the component based proportions and the elements of the signal has two consequences which may be of practical importance when implementing the algorithm. The first of these is that calculations of the kind (11) can be done for entire images in very little time (Settle and Briggs 1987). The great saving in computation times means that the generation of mixture maps can be carried out interactively and can readily become a standard tool of image analysis, and not just an end product as at the moment. The second feature of practical importance arises if it is the case that we are not interested in the patterns and distributions of the proportions over a scene, only in the average value over a large area (for forest inventory or crop yield estimates, for example). In this case the linear relationship means that the average value of any \( f_1 \) over a group of pixels is equal to the result obtained by estimating \( f \) using the average signal over those pixels. The required average of the signal over all pixels may well be already known, and if not can usually be readily calculated to help find a rapid estimate of global proportions.

3.3. The error matrix

If we take, for example, solution (11), and substitute for \( x \) from (3) we have:

\[
\hat{f} = \alpha (I - J^T UM^T N^{-1} U J) U^T M^T N^{-1} e
\]

\[
= f + \alpha (I - J^T U J) U M^T N^{-1} e
\]

\[
= f + (I - \alpha U J) U M^T N^{-1} e, \quad (13)
\]

Further:

\[
E((f - \hat{f})^T (f - \hat{f})) = (I - \alpha U J) U M^T N^{-1} E(\epsilon \epsilon^T) (I - \alpha U J) U M^T N^{-1}^T
\]

\[
= (I - \alpha U J) U M^T N^{-1} \cdot N \cdot N^{-1} M U (I - \alpha J U)
\]

\[
= (I - \alpha U J) U (I - \alpha J U)
\]

\[
= U - \alpha U J U = F', \text{ say.} \quad (14)
\]

which is independent of \( f \), a result which follows from our assumption that \( \epsilon \) is independent of \( f \). These two results define the statistical properties of the estimator \( \hat{f} \). (13) says that the estimator is unbiased; that is, if we have a pixel with true value of \( f, \hat{f} \), and for which linear mixing does take place, then the statistical expectation of \( \hat{f} \) is \( f \). The matrix \( F' \) is the variance-covariance matrix associated with the estimate \( \hat{f} \).

so that \( \sqrt{F'} \) is a measure of the error associated with the estimated value of \( f_1 \). If we adopt a specific distributional form for \( \epsilon \) we could perhaps place precise confidence limits on the estimates, although this is probably inadvisable given the number of assumptions we have already built into. It is probably better just to adopt the view that diagonal terms of \( F' \) give a rough estimate of the kind of error that our calculations are likely to suffer.

To illustrate the significance of the off-diagonal terms of \( F' \), let us consider the variance of the (unbiased) estimate of \( f_1 + f_2 \). This is easily shown to be given by \( (F_{11} + F_{22} + 2F_{12}) \). The non-diagonal terms are therefore important for determining the likely error on sub-compositions of the original set of components. We might, for instance, have cereals and grasslands as two components but want to know the combined acreage of the two types; this will be found from \( f_1 + f_2 \) and the error on this quantity is given by \( \sqrt{(F_{11} + F_{22} + 2F_{12})} \). Of incidental interest is the following geometrical representation of these terms. The error on \( f_1 \) is given by the square root of \( F_{11} \), is effectively a distance (suitably defined) in feature space between the
end-member spectrum for $\mu_1$ and the hyperplane defined by the remaining end-
members. Similarly, the error on the sum of a number of components is given by the
distance between the two hyperplanes defined by the two groups of components,
those contributing to the sum and those that are not. A proof is given in appendix B.

3.4. The positivity constraint

The unbiased nature of the estimates applies only to the estimators (9) and (11);
it is lost if we insist on the positivity of our estimate. This bias is occasioned only by
noise in the observations, however, and the matrix F is still likely to prove a useful
tool for defining the size of likely errors. If this positivity condition is always to
apply to our solution then in general the analytic solutions we have used will not be
valid and an alternative method of solution must be found; this may have to be some
sort of quadratic programming procedure, as outlined in, for example, Wolfe (1960).
Shimabukuro and Smith (1990) give a method suitable for a three class problem,
based on considering all possible outcomes from the classical estimator. Other
approaches are possible. An idea of Kent and Mardia (1986) is to work with the so-
called logit variables, $y_i = \log(f_i) = \ln\{(1 - f_i)/f_i\}$. As $f_i$ varies from 0 to 1, $y_i$ takes
on all real values, positive and negative. In the Kent and Mardia method linear
mixing is combined with a model for the spatial variability of the $y_i$. Equations for
these logit variables are derived, solved and legitimate proportions are then obtained
by inverting the logit transform. The authors report very good results using this
method. The simplest approach may simply be to use the classical estimator, or a
similar linear estimator, set any negative estimates to zero and then renormalise the
positive estimates so that they add to one; the resulting small loss of accuracy
compared to any of the more sophisticated approaches may be deemed acceptable
when balanced against the greatly reduced computational costs.

We have observed that the expression for $f$ which minimises $Q(f)$ is a linear
function of the components of $x$. Given that an acceptable solution is of this form,
we might consider simply looking for the linear estimator that gives the best such
representation. This approach leads to the use of the inverse estimator, which is
essentially a matter of carrying out a regression of the $f$ values on the $x$ values. For
identifiable cases the results tend to be very similar to those of the classical estimator
(e.g. Marsh et al. 1980). The method also produces solutions for the non-identifiable
case. For example, Pech et al. (1986) reported reasonable results from its use in the
estimation of vegetation cover over rangelands with a single Landsat band.
However, the linear nature of the estimator now means that, when $c > n + 1$, the
proportions will exhibit linear dependencies other than that they add to unity, and
these relationships are decided by the samples used to define the regression
constants. Care must be taken with the use of the inverse estimator in the non-
identifiable case.

4. The regularised estimator

The methods we have looked at so far find estimates of the proportions vector
for each pixel. To form an image representing the distribution of a particular cover
type we could simply form the aggregate of all these single pixel estimates. However,
while the application of a particular estimator may be a sensible strategy for an
individual pixel it is not true that a good strategy for whole scenes is to apply this to
each pixel in turn. Making the best possible fit to each pixel means we are inevitably
allowing the noise to influence the answers; we are fitting to both data and noise, in
Mixed pixels and ground cover

If we consider the quantity $Q(f)$, which we are normally trying to minimise, its expected value is given by:

\[
E(Q) = E(e^T N^{-1} e) = E(Tr(e^T N^{-1} e)) = E(Tr(e e^T N^{-1})) = Tr(E(e e^T N^{-1})) = Tr(N N^{-1}) = n
\]  

(15)

where $Tr(A)$ denotes the trace of the matrix $A$. This gives us a suitable indication of the average global error we should be prepared to tolerate in fitting the model to the data. By contrast, the average value of $(x - Mf)^T N^{-1} (x - Mf)$ over an image, with $\bar{f}$ the classical estimator, is readily shown to be $n - c + 1$. A better approach is to allow the computed mixture maps to be compatible, in the large, with the expected levels of noise. The proposed improved method of estimation will therefore be based upon the following strategy, viz. to choose $\bar{f}$ to minimise $\langle(f - g)^T (f - g)\rangle$, subject to:

\[
\langle(x - Mf)^T N^{-1} (x - Mf)\rangle = n,
\]  

(16)

where $\langle . \rangle$ denotes an average over all image pixels and $g$ is some favoured mixture. Without the sum-to-one constraint the required solution is found by determining the unconstrained minimum of:

\[
Q(f) + \lambda(f - g)^T (f - g)
\]  

(17)

where $\lambda$ is a Lagrangian multiplier. We find that the appropriate formula for the estimator $\bar{f}$ is:

\[
\bar{f} = \lambda V g + VM^T N^{-1} x
\]  

(18)

where

\[
V = (\lambda I + M^T N^{-1} M)^{-1}
\]  

(19)

If the sum of the proportions is forced to be unity then similar algebra gives us the estimator:

\[
\bar{f} = \lambda (V - \beta V J V) g + \beta V j + (V - \beta V J V) M^T N^{-1} x
\]  

\[
= \lambda R g + \beta V j + RM^T N^{-1} x
\]  

(20)

where $\beta$ and $R$ are defined by:

\[
R = V - \beta V J V
\]

\[
\beta = (j^T V j)^{-1}
\]  

(21)

These are further examples of linear estimators. They are defined for all values of $n$ and $c$, but there will once again be hidden linear dependencies between the elements of $f$ in the case of a non-identifiable problem. We have dubbed this approach the 'regularised estimator' because of the evident close link between it and the regularisation used in some areas of statistics. The similarity of (18) and (20) to (9) and (11) is obvious.

For the strategy stated, the parameter $\lambda$ would be chosen so that (16) be satisfied, although it is not immediately obvious how to solve the resulting equation; we show how to solve it in appendix C. On the other hand, we might take the view that (16) need only hold approximately for the noise in the image to have been adequately taken account of; in this way we are allowed a deal of latitude in our choice of $\lambda$. 


This leads us to considering the above formulae as two families of estimators, with \( \lambda \) acting as a truly free parameter.

5. Properties of the regularised estimator

For \( n > e + 1 \) the regularised estimator approaches the classical estimate for very small values of \( \lambda \). For larger values of the regularising parameter the estimate is drawn towards \( g \), and the very large values the spectral values make virtually no contribution and the estimator becomes essentially \( f = g + O(1/\lambda) \). Since one effect of the regularisation is to lessen the difference between \( \hat{f} \) and \( g \), another is to lessen the difference between values of \( \hat{f} \) for different pixels; thus the images derived from the use of the regularised estimator will generally be somewhat smoother than those obtained from the classical estimator.

The regularised estimator is generally biased. For a pixel with genuinely linear mixing, \( x \) will be equal to \( MF + e \) for some \( f \) and some zero-mean random \( e \), our estimate of the former being:

\[
\hat{f} = \lambda Rg + \beta Vj + RM^TN^{-1}(MF + e)
= \lambda Rg + \beta Vj + (V - \beta VJV)(V^{-1} - \lambda)f + RM^TN^{-1}e
= f + \lambda R(g - f) + RM^TN^{-1}e
\]

(22)

so that the bias is given by \( \lambda (V - \beta VJV)(g - f) \). If we would ensure that the average proportions vector is correct for the image as a whole, then we must set \( g \) to be equal to that average, \( \bar{x} \), say; it is a solution of the equation \( MF = \bar{x} \), where \( \bar{x} \) is the average signal for the image. The error on the estimate is found from:

\[
E((\hat{f} - f)(\hat{f} - f)^T) = \lambda^2 R(f - g)(f - g)^TR + RM^TN^{-1}E(ee^T)N^{-1}MR
= \lambda^2 R(f - g)(f - g)^TR + RM^TN^{-1}MR
= \lambda^2 R(f - g)(f - g)^TR + (V - \beta VJV)(V^{-1} - \lambda)(V - \beta VJV)
= \lambda^2 R(f - g)(f - g)^TR - \lambda R^2
\]

(23)

Expanding this to the first order in \( \lambda \) and taking the trace of both sides gives:

\[
E((\hat{f} - f)^T(\hat{f} - f)) = Tr(F) - 2\lambda Tr(F^2)
\]

(24)

where \( F \) is again defined to be equal to \( U - \alpha UJU \). Now since \( F \) is symmetric and non-zero, \( Tr(F^2) \) is positive, and so we have the remarkable result that whatever we choose for \( g \), the least squared error on \( \hat{f} \) is a decreasing function of \( \lambda \) for sufficiently small \( \lambda \). It is always possible, therefore, to find a value of \( \lambda \) that makes the regularised estimator more accurate, in a least squares sense, than the classical estimator. There is clearly a best value of \( \lambda \), one for which the error on \( \hat{f} \) is least; in appendix C we show how the optimal value of \( \lambda \) may be obtained from the solution of a polynomial equation.

The method can readily be adapted to include some sort of spatial context in the estimate, by making \( g \) depend on the values of the proportions for nearby pixels in an iterative scheme (Settle 1990); further significant improvements in accuracy can be obtained this way. Indeed, it is likely that the best use of this general technique will depend on the incorporation of a significant level of spatial information into our estimates.

6. Some examples

To illustrate some of these ideas and to test the different approaches, an artificial data set of 4096 pixels was generated. Five cover types were adopted, these being
green vegetation, dry vegetation and three minerals; muscovite, illite and calcite. A set of proportions was generated for each of the 4096 cases in the following way. A uniform random number in the range (0.1-0.3) was chosen to be the proportion of green vegetation. A second random number, uniform in the range (v-0.1, v+0.1) was chosen to give the proportion of dead vegetation, where v is the proportion of green vegetation; thus the proportions of green and dead vegetation are positively correlated. Next, one of the three minerals was chosen at random, and its proportion was chosen to fill between 80 and 100 per cent of the so-far unaccounted for pixel's area; one of the remaining two was chosen at random and assigned at random between 80 and 100 per cent of what was left, and the proportion of the third mineral was determined by the condition that the values should sum to one. This gives partial vegetation cover, and a mixture of soils, while any single pixel is dominated by just one soil type. These mixtures were intended to reflect the circumstances found in a number of semi-arid areas of geological interest. Spectra of 47 bands were generated for each pixel by forming the appropriate mixture of 'laboratory spectra', sampled at 10 nm resolution between 2.035 and 2.495 nm, a region in which the minerals can best be discriminated. The spectrum of green vegetation is very dark in this region, not rising much above 10 per cent reflectance (11 DN is the maximum in the spectrum for this component). A noise-free image was created, as well as two images with random noise added. The noise for these two images was taken to be zero mean Gaussian noise, with covariance matrix given by \( \sigma^2 \mathbf{I} \), where \( \sigma \) was taken to be 1 or 3DN. For the noise free case the image was generated as 32 bit real numbers in one instance, and as eight bit unsigned integers in another. This was intended to examine the effects of quantisation on our estimates. Simple reasoning suggests that if pixels are generally well mixed then the truncation of mixed signals into bytes will have an effect equivalent to a noise distribution given by a uniform random variable with mean zero and variance 1/12, as appropriate for a random variable distributed uniformly in \((-\frac{1}{4}, \frac{1}{4})\). This would lead for our example to our fitted values for \( \frac{1}{n} (x - \mathbf{M} \mathbf{I})^T (x - \mathbf{M} \mathbf{R}/n) \) being equal to \( \sqrt{(47-5+1)/(47 \times 12)} = 0.276 \), which is approximately the value we find after estimating proportions from the no-noise data which have been truncated to bytes. The effective variance of the noise for the image with 1DN of added noise is, similarly, 13/12 and is 109/12 for the image with 3DN added.

In table 1 we give the results of estimating proportions from each of five ways: (i) by unconstrained estimation, that is, applying the estimator (9) (ii) by renormalising the proportions vector obtained form (i), that is, setting to zero all negative estimates and rescaling the remainder so that they sum to one (iii) by partially constrained estimation, i.e., forcing the proportions to add to unity (the estimator given by (11))

<table>
<thead>
<tr>
<th></th>
<th>Unconstrained</th>
<th>Unconstrained/renormalised</th>
<th>Partially constrained</th>
<th>Part-constrained/renormalised</th>
<th>Fully constrained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\epsilon}_\text{u} )</td>
<td>( \hat{\epsilon}_\text{r} )</td>
<td>( \hat{\epsilon}_\text{p} )</td>
<td>( \hat{\epsilon}_\text{r} )</td>
<td>( \hat{\epsilon}_\text{f} )</td>
</tr>
<tr>
<td>No-noise</td>
<td>0.270</td>
<td>0.021</td>
<td>0.636</td>
<td>0.020</td>
<td>0.273</td>
</tr>
<tr>
<td>1DN noise</td>
<td>0.981</td>
<td>0.670</td>
<td>2.194</td>
<td>0.064</td>
<td>0.992</td>
</tr>
<tr>
<td>3DN noise</td>
<td>2.841</td>
<td>0.202</td>
<td>5.646</td>
<td>0.137</td>
<td>2.875</td>
</tr>
</tbody>
</table>

Table 1. Errors associated with different estimators.
(iv) by the application of (iii) followed by renormalising and (v) fully constrained estimation, where we use quadratic programming methods to find the best fit to the signal for non-negative proportions. For each method we give two measures of the error associated with the estimator; $e_x$ and $e_f$. The first of these is the quantity $\sqrt{\langle (x - Mf)^2 \rangle (x - Mf)/n}$ where $\langle \cdot \rangle$ denotes an average over the image; this is a measure of how well the signal can be reconstructed with the estimated proportions. The second, $e_f$, is defined to be $\sqrt{\langle (f - f_0)^2 \rangle (f - f_0)/c}$, where $f_0$ is the true proportions vector for a pixel; $e_f$ measures the root mean square error on the proportions vector. An upper limit to sensible values of this quantity may be obtained by considering the value it takes if $\hat{f} = \langle f_0 \rangle$ is used as the estimator; this upper limit we find to be equal to 0.19 for our image.

The errors behave much as expected. Unconstrained estimation would be expected to produce $e_x$ values of 0.273, 0.984 and 2.849 respectively (0.984 for example is the square root of $(n - c) \times (13/12 \div n$). We can see that the more the estimates are constrained, the poorer is the fit to the data but the better is the accuracy of the estimate. The renormalising estimators show clear improvements in the accuracy of the estimates but very sharp rises in $e_x$; the obvious improvements in accuracy that are to be attained by eliminating negative numbers are bought at the cost of a number of blunders with this simple procedure. The accuracy of some estimates for the image with 3DN of added noise are no better than the estimate which takes each pixel to be equally divided between the five cover types.

To illustrate the effect of regularisation, we carried out proportions estimates on these same images, for a number of different values of $\lambda$. The average proportions vector for the image was used for $g$, which happened to be very close to an equal mix of all cover types. In figure 1 we plot both $e_f$ and $e_x$ against $\lambda$ for the image with

![Figure 1](image.png)

**Figure 1.** Errors on the estimated proportions and on the fit to the image data for the 3DN noise example.
3DN added noise. $e_x$ is an increasing function of $\lambda$, as can readily be shown algebraically, while $e_f$ is seen to decrease, quite sharply at first, to a minimum then closely increase again. The smallest value of $e_f$ is less than half that corresponding to the classical estimator. The methods detailed in appendix C show that the optimal value of $\lambda$ can be predicted to be 26.7 in this case, and that $e_x$ takes the value 3 for $\lambda$ equal to about 85.

Regularisation can also be used with fully constrained estimation, one looks for the proportions vector with non-negative components that minimises (17). Here too, the effect is to decrease the error, though not by quite so much. In figure 2 we show how $e_f$ varies with $\lambda$ for three estimators. These are the partially constrained estimator (repeated from figure 1), the renormalised version of this, and the fully constrained estimator. The main observation to be made regarding this figure is that the best of the fully constrained estimators is not significantly better than the optimally regularised estimator. Thus the optimally regularised estimator has eliminated most of the spurious negative values that arise with the classical estimator, and gives virtually the same results as a fully constrained estimator but at a fraction of the computing time (typically an order of magnitude).

7. Conclusions

The method of information extraction described here, considered a superior alternative to image classification, is based on acknowledging the fact that most pixels are mixed. The linear mixing model, which we have concentrated on, is appropriate for a great many land-based scenes, and leads to a particularly simple mathematical description of the generation of the signal for a given mixture. The resulting algorithms for the estimation of cover information are in some cases much faster than maximum likelihood classification, although they generate a much larger quantity of data.

![Figure 2. Errors for 3 estimators—3DN noise image.](image-url)
The error matrix provides us with an idea of the errors on each of our proportions, and also on any linear combination of them. Quality estimation of this type is very much more difficult for classification. We have considered not only the standard methods of estimation for well conditioned problems but also an estimator based on regularisation. This we have presented in its simplest form, and shown that better results are always attainable with this method compared to results obtained from the standard estimator; in the case of our simulated data the root mean square error on the proportions estimate can be reduced by half with a suitable value of the smoothing parameter. Furthermore, the optimal value of this parameter can be found algebraically.

Overall, we believe that it has been shown that the benefits of proportions estimation are sufficient for this to become a serious alternative to classification for routine image processing of remotely sensed imagery, and that the use of non-standard estimators can lead to significant increases in the accuracy of estimation.

Acknowledgments

We would like to thank Dr. J. D. Wilson for modifying and making available to us her quadratic programming code when our own attempts to make proprietary software work had ended in despair. Thanks too to Professor J. R. G. Townshend of the University of Maryland, Dr John Louis of Wagga and to Professor Robert Gurney of NUTIS for reading early drafts of this paper and providing comments that have greatly improved its appearance. This work was started when the authors were working on the NERC special topic ‘Improved Algorithms for Classification and Image Processing of Remotely Sensed Data’. NAD was also supported by a Research Endowment Fellowship of the University of Reading.

Appendix A: Obtaining the parameters of the model

Suppose we have \( k \) training pixels and that the observation and known proportion vector for each of these are given by \( x_i \) and \( f_i \) respectively, \( i = 1, \ldots, k \). We suppose that linear mixing applies for each of these pixels; thus:

\[
x_k = Mf_k + e_k
\]

(A1)

where \( e_k \) is the error for the \( k \)th training pixel. If these errors are independent then the joint likelihood function of the signals on the proportions for the training pixels is given by:

\[
Pr(x_1, x_2, \ldots, x_k/f_1, f_2, \ldots, f_k) = Pr(x_1/f_1)Pr(x_2/f_2) \ldots Pr(x_k/f_k)
\]

(A2)

If we now assume that, for a given \( f \), signals have a multivariate normal distribution with mean \( Mf \) and covariance matrix \( N \), then the log-likelihood will be given by:

\[
2\ln(Pr(x_1, x_2, \ldots, x_k/f_1, f_2, \ldots, f_k)) = \text{constant} - k\ln|N| - \Sigma(x_i - Mf_i)^T N^{-1}(x_i - Mf_i).
\]

(A3)

If we now search for \( \hat{N} \) and \( \hat{M} \) to maximise this quantity, we find:

\[
\hat{M} = \langle xf^T \rangle \langle f^2 \rangle^{-1}
\]

(A4)

\[
\hat{N} = \langle (x - \hat{M}f)(x - \hat{M}f)^T \rangle
\]

(A5)

where \( \langle . \rangle \) denotes an average taken over all training pixels. These are the Maximum Likelihood estimates of \( M \) and \( N \). The equation for the common covariance matrix is identical in form to that encountered in supervised classification.
Mixed pixels and ground cover

1173

If we substitute from (A1) into (A4), we find

$$\tilde{M} = M + \langle ef^T \rangle \langle ff^T \rangle^{-1}$$  \hspace{1cm} (A6)

The estimate of $\tilde{M}$ is therefore unbiased. The sum of the squares of the differences of the elements of $M$ and $\tilde{M}$ is given by the trace of $(M - \tilde{M})(M - \tilde{M})^T$, and has expectation:

$$E(Tr(\langle ef^T \rangle \langle ff^T \rangle^{-2} (fe^T))) = \frac{1}{k} Tr(N) \{ \langle f^T \langle ff^T \rangle^{-2} f \rangle \}$$

$$= \frac{c}{k} Tr(N) Tr(\langle ff^T \rangle^{-1})$$  \hspace{1cm} (A7)

If $N$ should equal $\sigma^2 I$ for some $\sigma$ then the mean square estimation error per element of $M$ is:

$$\frac{1}{k} \sigma^2 Tr(\langle ff^T \rangle^{-1})$$  \hspace{1cm} (A8)

so to ensure that the training data provide good estimates of the end-member spectra we need the number of training pixels to be much larger than $Tr(\langle ff^T \rangle^{-1})$. This term is a measure of the dependence of the proportions vectors of the training data; it is very large when those vectors are nearly linearly dependent.

The estimate of $N$ is biased. For a given training pixel:

$$(x_k - \tilde{M} e_k)(x_k - \tilde{M} e_k)^T = (e_k - \langle ef^T \rangle \langle ff^T \rangle^{-1} f_k)(e_k - \langle ef^T \rangle \langle ff^T \rangle^{-1} f_k)^T$$  \hspace{1cm} (A9)

Averaging over the training data produces:

$$\langle (x_k - \tilde{M} e_k)(x_k - \tilde{M} e_k)^T \rangle = \langle ee^T \rangle - \langle ef^T \rangle \langle ff^T \rangle^{-1} \langle fe^T \rangle$$  \hspace{1cm} (A10)

The expectation of the first term on the right-hand side of (A10) is simply $N$, while that of the second is easily shown to be

$$-\frac{1}{k} N \langle f^T \langle ff^T \rangle^{-1} f \rangle = -\frac{c}{k} N$$  \hspace{1cm} (A11)

Thus we might prefer to adopt as our estimate of $N$ not (A5) but instead the unbiased estimate:

$$\frac{k}{k - c} \langle (x - \tilde{M} e)(x - \tilde{M} e)^T \rangle$$  \hspace{1cm} (A12)

The choice of either (A5) or (A12) has no effect on the analytical estimators of this paper, for they are unaffected by a scaling of the noise matrix $N$. However, the measures of precision of these estimates, based on the matrices $F$ and $U$, will be affected by the choice of estimate for $N$. The analysis of the unbiasedness or otherwise of these quantities, when estimates of $M$ and $N$ are used, is beyond the scope of this paper.

Appendix B

In this appendix the error on the estimate of the sum of a number of proportions is equated to a feature space distance between two hyperplanes, each spanned by the sub-compositions. The 'distances' we shall be dealing with are defined by the metric $d(u, v)$ where
\[(d(u, v))^2 = (u - v)^T N^{-1} (u - v) . \quad (B1)\]

This is really just a Euclidean distance modulated by the positive definite matrix \(N\), which we recall is the variance-covariance matrix of the error \(e\) on our measurement. We divide our set of components into subsets of \(m_1\) and \(m_2\) \((= c - m_1)\) members each; without loss of generality we may take these to be \(\{f_1, \ldots, f_{m_1}\}\) and \(\{f_{m_1 + 1}, \ldots, f_c\}\). We wish to know the error on \(f_1 + f_2 + \ldots + f_{m_2}\). We partition the matrix \(M\) thus, \(M = (M_1, M_2)\), where \(M_1\) is an \(n \times m_1\) matrix, the columns of which are the end-member spectra of our first group of components. The remaining components define the \(n \times m_2\) matrix \(M_2\). The distance, with the above-defined metric, between the hyperplanes defined by the two sets of end-members is found as follows. A point in the first hyperplane may be represented as \(M_1 \mathbf{u}\), where \(\mathbf{u}\) is an \(m_1\)-dimensional vector which is arbitrary apart from the fact that its elements must add to 1, and a point in the second hyperplane is given by \(M_2 \mathbf{v}\), where \(\mathbf{v}\) is a \((m_2 \times 1)\) vector, arbitrary apart from the fact that its elements must also sum to unity. The distance between the two hyperplanes is the minimum of the distances between pairs of points \((M_1 \mathbf{u}, M_2 \mathbf{v})\), \(\mathbf{u}\) from the first plane and \(\mathbf{v}\) from the second. Thus the distance squared is given by:

\[
\min(d(M_1 \mathbf{u}, M_2 \mathbf{v}))
\]

where the minimum is evaluated over all pairs of \(\mathbf{u}\) and \(\mathbf{v}\) satisfying

\[
j_1^T \mathbf{u} = 1, \quad j_2^T \mathbf{v} = 1
\]

(B3)

The solution is found by looking for the unconditional minimum of

\[
(M_1 \mathbf{u} - M_2 \mathbf{v})^T N^{-1} (M_1 \mathbf{u} - M_2 \mathbf{v}) - 2 \lambda_1 j_1^T \mathbf{u} - 2 \lambda_2 j_2^T \mathbf{v}
\]

regarded as a function of the components of \(\mathbf{u}\) and \(\mathbf{v}\); \(\lambda_1\) and \(\lambda_2\) are Lagrangian variables found by satisfying the constraints (B3).

Let us define the \((c \times 1)\) vectors \(\mathbf{w}\), \(\mathbf{h}_1\) and \(\mathbf{h}_2\) by:

\[
\mathbf{w}^T = (\mathbf{u}^T, -\mathbf{v}^T)
\]

\[
\mathbf{h}_1^T = (j_1^T, 0)
\]

\[
\mathbf{h}_2^T = (0, j_2^T)
\]

(B5)

where the zero vectors in the definitions of \(\mathbf{h}_1\) and \(\mathbf{h}_2\) have dimensions \(m_2\) and \(m_1\) respectively. Then \(M_1 \mathbf{u} - M_2 \mathbf{v}\) is equal to \(M \mathbf{w}\) and the problem may be redefined in terms of these variables as having to find the unconditional minimum of:

\[
\mathbf{w}^T M^TN^{-1}M \mathbf{w} - 2 \lambda_1 \mathbf{h}_1 \mathbf{w} + 2 \lambda_2 \mathbf{h}_2 \mathbf{w}
\]

(B6)

where the Lagrangian multipliers \(\lambda_1\) and \(\lambda_2\) are to be found from the conditions:

\[
\mathbf{h}_1 \mathbf{w} = -\mathbf{h}_2 \mathbf{w} = 1.
\]

(B7)

Differentiating (B6) with respect to each of the elements of \(\mathbf{w}\) and setting the resulting expressions to zero gives the equation to be satisfied:

\[
M^TN^{-1}M \mathbf{w} = \lambda_1 \mathbf{h}_1 - \lambda_2 \mathbf{h}_2
\]

with solution

\[
\mathbf{w} = \lambda_1 \mathbf{U} \mathbf{h}_1 - \lambda_2 \mathbf{U} \mathbf{h}_2.
\]

(B9)
Mixed pixels and ground cover

Taking the inner product of this equation with \( h_1 \) and \( h_2 \) gives the pair of equations:

\[
1 = \lambda_1 h_1^T U h_1 - \lambda_2 h_2^T U h_2,
\]
\[
-1 = \lambda_1 h_1^T U h_1 - \lambda_2 h_2^T U h_2 \quad (B10)
\]

with solution:

\[
\lambda_1 = (h_2^T U h_2 + h_1^T U h_1) / (h_1^T U h_1 h_2^T U h_2 - (h_1^T U h_2)^2)
\]
\[
\lambda_2 = (h_1^T U h_1 + h_2^T U h_2) / (h_1^T U h_1 h_2^T U h_2 - (h_1^T U h_2)^2) \quad (B11)
\]

Taking the inner product of (B9) with \( w \) gives on the left-hand side the required minimum of the distance function; this is:

\[
d_{mn}^2 = \lambda_1 h_1^T w - \lambda_2 h_2^T w = \lambda_1 - \lambda_2 \quad (B12)
\]

Combining this with the previous equations gives for the inverse of the minimum of the distance function:

\[
d_{mn}^{-2} = (h_1^T U h_1 h_2^T U h_2 - (h_1^T U h_2)^2) \cdot (h_1^T U h_1 + h_2^T U h_2)^{-1}
\]
\[
= (h_1^T U h_1 h_2^T U h_2 - (h_1^T U h_2)^2) \cdot (h_1^T U h_1 + h_2^T U h_2)^{-1}
\]
\[
= \alpha (h_1^T U h_1 h_2^T U h_2 - (h_1^T U h_2)^2) \quad (B13)
\]

Now \( (f_1 + f_2 + \ldots + f_m) \) is equal to \( h_1^T f \), our estimate of which is \( h_1^T \hat{f} \), and so the expected error on this quantity is given by:

\[
E \{ (h_1^T \hat{f} - h_1^T f) (h_1^T \hat{f} - h_1^T f)^T \} = h_1^T E \{ (f - \hat{f}) (f - \hat{f})^T \} h_1
\]
\[
= h_1^T (U - \alpha U J U) h_1 = h_1^T U h_1 - \alpha h_1^T U (h_1^T + h_2^T) U h_1
\]
\[
= \alpha (h_1^T U h_1 (h_1^T U h_1 + h_2^T U h_2) - (h_1^T U h_2)^2)
\]
\[
= \alpha (h_1^T U h_1 h_2^T U h_2 - (h_1^T U h_2)^2) \quad (B14)
\]

This is identical to (B13), proving the assertion that the standard deviation of the estimator is the inverse of the distance between the two groups of components.

Appendix C

In this appendix we consider the form of the error on our predicted proportions for the regularised estimator, and show how to find the value of the smoothing parameter that minimises this.

We recall (23) that the error is given by:

\[
\langle (f - \hat{f}) (f - \hat{f}) \rangle = Tr(R) - \lambda^2 Tr(R^2) + \lambda^3 \langle (f - \hat{f}) (f - \hat{f}) \rangle R (f - \hat{f}) \rangle \quad (C1)
\]

where \( Tr(A) \) denotes the trace of the matrix \( A \), \( \langle . \rangle \) denotes an average over the image and \( R \) is given by:

\[
R = V - \beta VJ V \quad (C2)
\]

To begin, we show that the eigenvectors of \( R \) are independent of \( \lambda \), so that they are the same as those of the matrix \( F \). If \( x \) is an eigenvector of \( R \) and \( \theta \) the corresponding eigenvalue, then:

\[
(V - \beta VJ V) x = \theta x \quad (C3)
\]
which we may rearrange to:

\[ x \propto (V - \theta I)^{-1} V j = (I - \theta V^{-1})^{-1} j = (I - \theta (U^{-1} + \lambda))^{-1} j = (U(1 - \theta \lambda) - \theta I)^{-1} j = (U - \tau I)^{-1} U j \]  

(C4)

where \( \tau \) is equal to \( \theta/(1 - \theta \lambda) \). It follows that \( \tau \) is an eigenvalue of \( F \), and that \( x \) is the corresponding eigenvector. The eigenvalues of \( R \) are therefore given by

\[ \theta_i(\lambda) = \tau_i/(1 + \lambda \tau_i) \]  

(C5)

Let \( P \) denote the eigenvector matrix for \( F \). Then:

\[ R = P \text{diag}(\theta_1(\lambda), \theta_2(\lambda), \ldots, \theta_s(\lambda)) P^T = P \Theta(\lambda) P^T \]

\[ \text{Tr}(R^2) = \text{Tr}(\Theta^2) = \Sigma \theta_i^2 \]  

(C6)

The third term on the right-hand side of (C1) is:

\[ \lambda^2 \langle f - \langle f \rangle \rangle^T R^2 \langle f - \langle f \rangle \rangle = \lambda^2 \text{Tr}(R^2 C_f) = \lambda^2 \text{Tr}(\Theta^2 P^T C_f P) = \lambda^2 \text{Tr}(\Theta^2 B) \]

\[ = \lambda^2 \sum \theta_i^2 B_{ii} \]  

(C7)

where \( C_f \) is the scatter matrix of the image proportions and the matrix \( B \) is defined by

\[ B = P^T C_f P \]  

(C8)

It is independent of \( \lambda \), since \( P \) is. Of course, \( C_f \) is not known but an estimate of it is:

\[ \hat{C}_f = U M^T N^{-1} C N^{-1} M U - U \]  

(C9)

where \( C_x \) is the variance covariance matrix of the image signals. This estimate is unbiased when exact values of \( M \) and \( N \) (and hence \( U \)) are used. Putting these together we find that the expression for the error of \( f \) is given by:

\[ \epsilon_f^2 = \sum_{i=1}^{c} \left( \frac{\tau_i}{(1 + \lambda \tau_i)} - \frac{\lambda \tau_i^2}{(1 + \lambda \tau_i)^2} + \frac{\lambda^2 \tau_i^2 B_{ii}}{(1 + \lambda \tau_i)^2} \right) = \sum_{i=1}^{c} \left( \frac{\tau_i + \lambda^2 \tau_i^2 B_{ii}}{(1 + \lambda \tau_i)^2} \right) \]  

(C10)

Differentiating with respect to \( \lambda \), and setting the resulting expression to zero gives the following equation that must be satisfied by the 'optimal' regularising parameter:

\[ 0 = \sum_{i=1}^{c} \frac{\tau_i B_{ii}}{(1 + \lambda \tau_i)^3} \]  

(C11)

which multiplies out to polynomial of degree \((3c-5)\) (one of the \( \tau_i \) is zero).

A similar expression for the rms error on the data can be obtained; it is:

\[ \epsilon_r^2 = n - c + 1 + \lambda^2 \sum_{i=1}^{c} \frac{B_{ii} \tau_i + \tau_i^2}{(1 + \lambda \tau_i)^2} \]  

(C12)

which is clearly a monotonic increasing function of \( \lambda \). Using this equation we can see to it that the estimate satisfies \( \epsilon_r^2 = n \), which is the condition we argued for at the outset of our discussion on regularisation.
Mixed pixels and ground cover

References


