Peer Reviewed Paper openaccess Special Issue on Chemometrics in Hyperspectral Imaging

Classification in hyperspectral images by independent component analysis, segmented cross-validation and uncertainty estimates

Beatriz Galindo-Prieto[a] and Frank Westad[b,*]

[a]Department of Engineering Cybernetics (ITK), Norwegian University of Science and Technology (NTNU), Norway.
ORCID: https://orcid.org/0000-0001-8776-8626
[b]Department of Engineering Cybernetics (ITK), Norwegian University of Science and Technology (NTNU), Norway and CAMO Software, Oslo, Norway. E-mail: frank.westad@ntnu.no

Independent component analysis combined with various strategies for cross-validation, uncertainty estimates by jack-knifing and critical Hotelling's T^2 limits estimation, proposed in this paper, is used for classification purposes in hyperspectral images. To the best of our knowledge, the combined approach of methods used in this paper has not been previously applied to hyperspectral imaging analysis for interpretation and classification in the literature. The data analysis performed here aims to distinguish between four different types of plastics, some of them containing brominated flame retardants, from their near infrared hyperspectral images. The results showed that the method approach used here can be successfully used for unsupervised classification. A comparison of validation approaches, especially leave-one-out cross-validation and regions of interest scheme validation is also evaluated.

Keywords: hyperspectral imaging, ROI selection, spectroscopy, independent component analysis, ICA, cross-validation, uncertainty test, jack-knifing, Hotelling's T^2, classification

Introduction

Hyperspectral images are useful for obtaining both qualitative and quantitative information, since they are high-dimensional data sets (hypercubes) that can be visually interpreted. As pointed by Vidal and Amigo,[1] dimensionality problems related to these large hypercubes of data, e.g. computer storage space or long transmission times, may make necessary the compression of the images and the acquisition of high performance computers. Some common ways of compressing hyperspectral images are data binning, variable selection and bytes encoding.[1,2]

In chemometrics, hyperspectral imaging (HSI) integrates spectroscopy [e.g., near infrared (NIR)] and imaging to create 3-D structures of data (high-dimensional hypercubes) that can be analysed using multivariate methods, such as principal component analysis (PCA),[3,4] independent component analysis (ICA)[5] and multivariate curve resolution (MCR),[6,7] *inter alia*. Prior to modelling of the hyperspectral data, preprocessing the images in the image domain, but foremost in the spectral domain, is usually needed (e.g., background removal, scatter correc-

Correspondence
Frank Westad (frank.westad@ntnu.no)

Received: 30 September 2017
Revised: 26 January 2018
Accepted: 31 January 2018
Publication: 25 February 2018
doi: 10.1255/jsi.2018.a4
ISSN: 2040-4565

Citation
B. Galindo-Prieto and F. Westad, "Classification in hyperspectral images by independent component analysis, segmented cross-validation and uncertainty estimates", *J. Spectral Imaging* **7,** a4 (2018). https://doi.org/10.1255/jsi.2018.a4

tion, de-noising, suppression of sample morphology effects or treatment of dead pixels).[1,8]

These big data, structured as hyperspectral image cubes, have relevance in many types of applications, for example agricultural and food sciences,[9,10] for data collection by drones[11] and in the pharmaceutical industry.[12] The applicability of multivariate data analysis for HSI is relevant for process analytical control (PAC) and quality by design (QbD) in a wide range of industrial sectors. Some applications of HSI multivariate data analysis relate, for instance, to the study of multi-component systems[12,13] or the understanding of the dynamics in time series analysis.[9]

As will be shown, ICA combined with various strategies for cross-validation and uncertainty estimates, proposed in this paper, is useful for an efficient and reliable multivariate analysis of hyperspectral images. To the best of our knowledge, the approach that we use in this paper for the combination of the abovementioned methodologies has not been applied to hyperspectral imaging analysis for interpretation and classification. More specifically, the data analysis performed here aims to distinguish, by unsupervised classification, between four different types of plastics from their hyperspectral images. The criterion followed is maximising independence of latent variables rather than orthogonality. Certainly, various discriminant methods may have been applied for a supervised classification, and other latent variable methods (such as MCR-ALS) could have been used; however, emphasis on the validation procedure and unsupervised classification ability inside the ICA framework has been pursued in this paper.

Independent component analysis (ICA)

Independent component analysis[5] is an alternative to principal component analysis (PCA)[3,4] for extracting pure and statistically independent pure profiles (components), such as pure spectra or original signals, from non-Gaussian distributed data.[14] The differences related to the extraction of components for ICA and PCA are further explained below. In ICA, the data matrix (**X**), which contains the mixture of pure profiles, is decomposed as shown in Equation 1, where **X** is the data matrix (also called *mixing system*), **A** is the "scores" matrix, **S** consists of statistical independent columns and **E** is the residuals matrix.

$$\mathbf{X} = \mathbf{A}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

The ICA "scores" can be calculated as shown in Equation 2. As pointed by Westad and Kermit,[15] unlike PCA scores, ICA "scores" (columns of **A**) are not restricted to be orthogonal among themselves.

$$\mathbf{A} = \mathbf{X}\mathbf{S}(\mathbf{S}^{\mathbf{T}}\mathbf{S})^{-1} \quad (2)$$

Independent component analysis decomposition is similar to other methodologies, e.g. multivariate curve resolution–alternating least squares (MCR-ALS)[6] that can be used for analysing spectral data that, for example, obey Lambert–Beer's law. These methods are able to provide quantitative information from hyperspectral images either at a global or at a pixel level.[16]

Validation and classification in multivariate models

The estimation of the optimal number of components in latent variable methods (e.g. PCA, MCR-ALS and ICA) is one of the cornerstones in model validation to avoid problems, such as, for example, a deficient profiles extraction or overfitting.[17,18] The most conservative approach is to set aside a number of objects, as an independent test set; this is the preferred procedure for multispectral imaging at the pixel level. Although visualisation of groups of pixels to elucidate differences (e.g., between origins or types of materials) provides a qualitative assessment if the groups are separated, a totally correct classification is not guaranteed. In this paper, one of the main objectives is to show how various validation strategies will affect the stability of the ICA models for hyperspectral image analysis.

As mentioned above, the model stability towards known and unknown sources of variation is very important. In all real applications, there will always be a reason to stratify the objects based on background information about their origin; such groups are a consequence of the experimental set-up of the study. Some typical stratifications are: (i) across treatment or origin (e.g., year or raw material), (ii) across instrumental replicates (repeatability), (iii) reproducibility (e.g., in relation to the analyst or the instrument), (iv) sampling (physical samples, time and location).[19]

Cross-validation (CV) performed at the various grouping levels will provide important information about the stability of the model and which are the sources of variation that need special attention. Thus, even if a test set is considered adequate for validating the model, the calibration set must be validated by cross-valida-

tion at the appropriate level.[19,20] Otherwise, the model dimensionality may not be conservative enough, leading to a classification/prediction of the test set where a suboptimal number of variables and/or components are used. In this paper, this concept is applied when pixels in various regions of interest (ROI) are selected from various physical objects of the same class.

Feature/variable selection for classification and regression

Many feature/variable selection methods have been presented and used in the literature.[21–28] Some of these methods aim to find a small subset of variables that gives the "best" possible model, others a selection of variables that are suitable for specific purposes; however, even if the latter perspectives are valid and useful for the applied cases, in this paper we will focus on how the model stability will influence estimates of the significance of individual variables. It is also worth noting that variable redundancy (i.e., when many variables contain the same information about the system under observation) provides valuable properties[21] in the model in order to (i) detect changes in the overall pattern due to unexpected events or (ii) avoid the unintentional elimination of variables that could help to understand causal effects. Therefore, some flexibility is needed when doing feature/variable selection.

Materials and methods

Materials

The codes of the algorithms were developed, tested and validated using MATLAB version R2016a (The MathWorks, Natick, MA, USA). The preprocessing of the hyperspectral images was performed using HYPER-Tools (downloaded from www.hypertools.org).

Plastics data set

The plastics NIR hyperspectral data set (downloaded from www.hypertools.org) consists of 142 wavelengths and 203 × 117 pixels.[2] The wavelength range goes from 1009 nm to 1694 nm with a spectral resolution of 4.85 nm. The samples represent four different plastics (PS, PA6, PP and ABS), see Figure 1, in the shape of small pellets (with a diameter of approximately 5 mm). PS and ABS plastic types contain brominated flame retardants (BFRs), whilst PA6 and PP types do not contain BFRs. After evaluation of the NIR spectra and hyperspectral images (which is a false RGB) of the four plastics, some pre-processing steps were required. The data were preprocessed by Savitzky–Golay smoothing[29] using a window of seven points (which is enough for the low levels of noise present in the data) and a second polynomial degree. Standard normal variate (SNV) transformation[30] was performed in order to remove scattering effects. Dead pixels detection was carried out by adding a standard deviation parameter equal to ±6 and allowing 25 % of zeros in the spectra; only one dead pixel was found and corrected. The background of the hyperspectral images was removed because it does not contain any information related to the plastics; this was achieved by visual inspection of the six clusters obtained by the *k*-means clustering[31] method, and removal of the two clusters that were only related to the background. Morphological masking was also carried out, including erosion of some remaining pixels that did not contain plastics information.

Multivariate analysis of the hyperspectral images

The methodology followed to perform the ICA included ROI selection in the image, cross-validation, uncertainty estimates and estimates of critical statistical limits (Hotelling's T^2). The results were visualised as plots and statistics.

The models were validated across ROIs selected for evaluating the robustness of the models related to the individual pellets and the so-called leave-one-out (LOO) cross-validation for comparison. Three ROIs were consecutively selected from each plastic type. For each ROI, a colour was assigned (see Figure 1). ROI 1 is represented in red, ROI 2 in green and ROI 3 in blue. The colour assignment of the ROIs matches the colours used for plotting the spectra (data) corresponding to those regions (see Figure 1).

The first step in the modelling was to perform PCA (often called whitening in the ICA literature) as the basis for the final ICA model; details about this procedure can be found in Westad.[32] It is worth mentioning that cross-validation for individual segments can sometimes give components that are mirrored or flipped when computing PCA; this flip in the order of the components can also occur in ICA. The difference lies in how ICA and PCA extract the components leading to a different ordering of them; whilst PCA extracts the components (PCs) according to the largest amount of variance, ICA does not

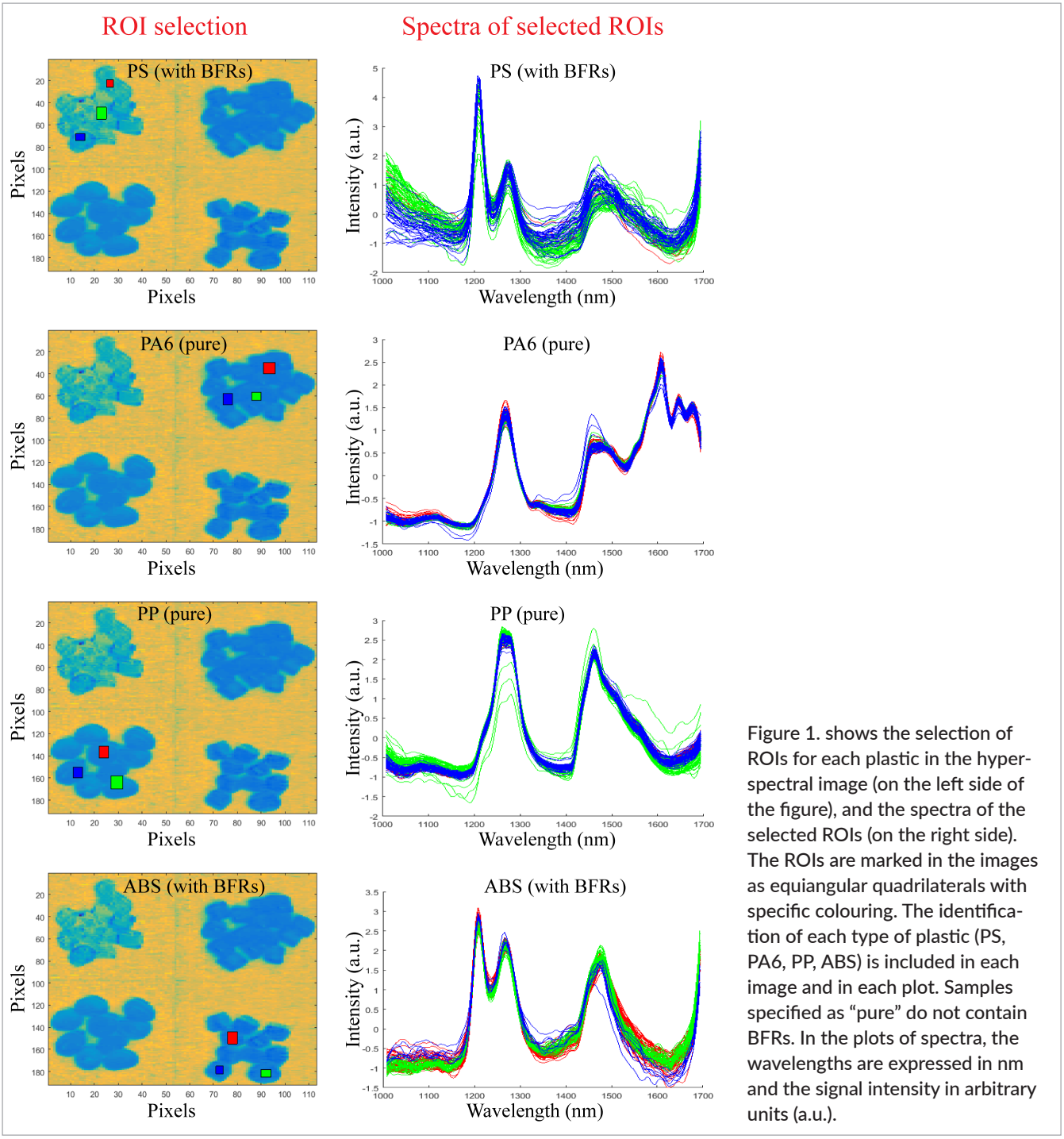

Figure 1. shows the selection of ROIs for each plastic in the hyperspectral image (on the left side of the figure), and the spectra of the selected ROIs (on the right side). The ROIs are marked in the images as equiangular quadrilaterals with specific colouring. The identification of each type of plastic (PS, PA6, PP, ABS) is included in each image and in each plot. Samples specified as "pure" do not contain BFRs. In the plots of spectra, the wavelengths are expressed in nm and the signal intensity in arbitrary units (a.u.).

take into consideration the amount of variance for the extraction of the components (ICs). Moreover, for some ICA algorithms, the order in the components extraction is also undetermined as consequence of random initialisation. Hence, the components obtained after cross-validation could be rotated towards the model based on all objects before the uncertainties can be estimated, if needed.[33] This is performed by flipping and ordering the loading vectors, followed by jack-knifing, as shown by Westad and Kermit in 2003.[15]

For the data analysis presented in this paper, CV was carried out at appropriate grouping levels in order to use the optimal number of components. In order to do this, pixels from different ROIs were selected from various physical objects (plastic pellets) of the same class. More specifically, three ROIs were selected as

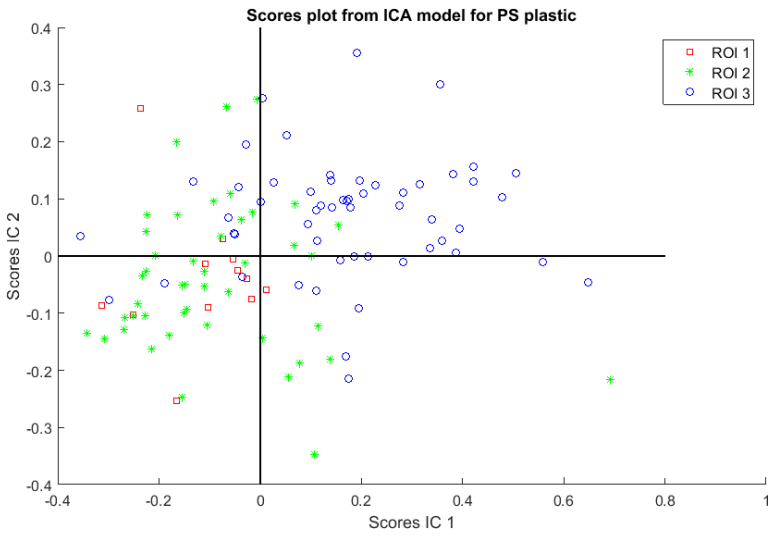

Figure 2. Distribution of score values along IC 1 and IC 2 for the PS individual ICA model. ROI 1 is marked in red (squares), ROI 2 in green (stars), and ROI 3 in blue (circles).

shown in Figure 1. The ROI selection was carried out by hand (using the computer mouse) deliberately with the purpose of choosing regions located in different areas of the groups of plastic pellets. As can be seen in the PS scores plot of Figure 2 (shown as example), the three regions selected for the PS plastic represent the variation between the ROIs (since they are positioned in different parts of the scores plot). After the ROI selection, the ICA modelling was performed as explained in the Results and discussion section. The ICA algorithm used was the so-called *joint approximate diagonalisation of eigenmatrices* (JADE).[34] The critical limits based on the Hotelling's T^2 statistics were applied for the classification of all pixels in the mosaic image of four types of plastics.[10]

Uncertainty estimates by jack-knifing

Resampling methods (such as bootstrap, jack-knifing and cross-validation) are used for assessing the importance of individual variables by means of their estimated uncertainty values. The jack-knifing method applied in this paper estimates the uncertainty of model parameters by calculating the difference between the model with all objects and the individual models (i.e., without some of the objects). These differences are squared and summed for all the cross-validation segments as the basis for the standard deviation of each model parameter.[19] Therefore, the uncertainty (σ) of the ICA loadings (s_a), can be estimated from Equation 3:[15,35]

$$\sigma^2(s_a) = \left[\sum_{m=1}^{M}\left(s_a - s_{a(-m)}\right)^2\right]\left[\frac{M-1}{M}\right] \quad (3)$$

where M is the number of cross-validation segments, $\sigma^2(s_a)$ is the estimated uncertainty variance of each variable (e.g., each wavelength) in the ICA loading for component a and s_a the ICA loading for component a using all the N objects (e.g., all pixels). The term $s_{a(-m)}$ indicates the ICA loading for component a using all objects except the object(s) left out in cross-validation segment m. The quantities s_a and $\sigma^2(s_a)$ may require a t-test to give the significance values (p-values) per individual variable and per component, and may also be used as an approximate confidence interval for each variable. We would like to emphasise that feature selection is not the goal of this paper, but the model stability expressed by uncertainties and their related p-values.

Results and discussion

The methods used for obtaining these results have been explained above. However, in this section we provide the necessary details for reproducing the results inside their description, as well as a discussion thereof.

Classification of four different plastics

The 142 channels (wavelengths) of the plastics hyperspectral data were used for the multivariate analysis (see above for details about the preprocessing applied). As basis for the ICA calculations, a number of principal components were selected based on the residual validation variance, which had average values of 14.4 % for the models of the pure plastics and 22.5 % for the models of the plastics that contained BFRs. Afterwards,

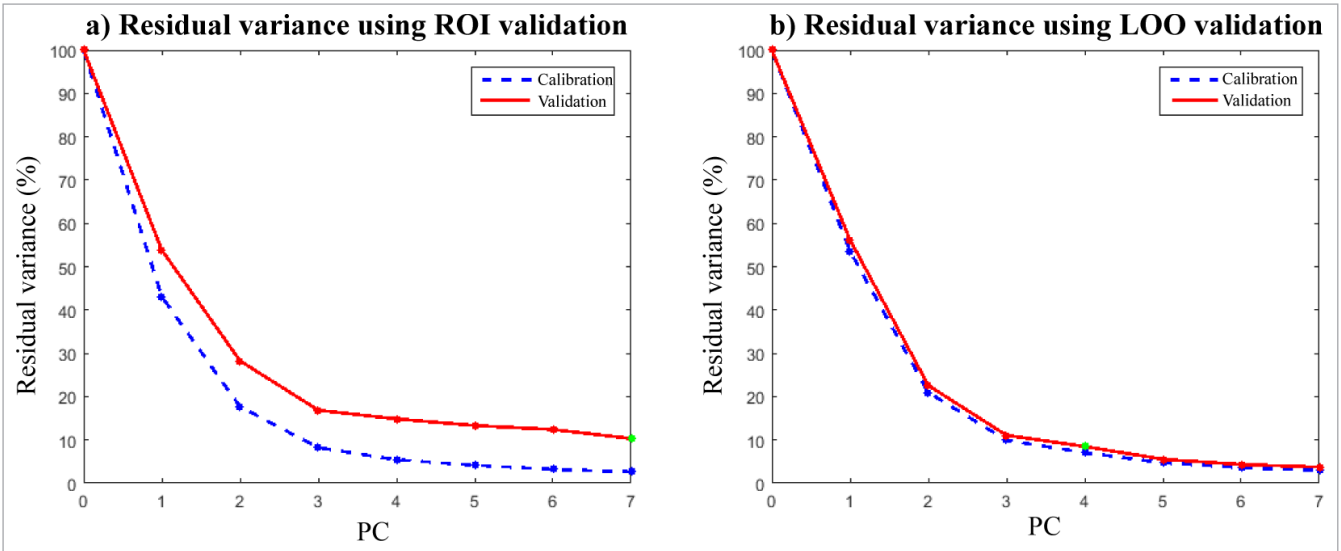

Figure 3. Residual variance curves for the two validation approaches. (a) shows the evolution of the residual variance when a ROI-based validation is used. (b) shows the analogous evolution for the same model when using the LOO validation method. For both representations, the calibration curve is marked with a blue dashed line, whist the validation curve is marked with a red solid line. PC stands for principal components.

the ICA algorithm extracted the corresponding number of components (ICs). It is worth mentioning that the result will be the same regardless if the ICA is performed on the loadings from PCA or the reconstructed data by multiplying scores and loadings. Table 1 summarises the number of ICs for all the models. The final ICA model (which contained all types of plastics) had a residual validation variance value of 2.9 %. A rotation of the model components was performed as explained above. The data was mean-centred prior to building the ICA models. In addition, the LOO cross-validation was carried out for the PS data (the top-left plastic sample in the images of Figure 1) to illustrate the conceptual difference between the two approaches.

Figures 3a and 3b show the residual calibration and validation variance for the two CV strategies. As expected, the LOO (full leave-one-out cross-validation) gives a residual validation curve that follows the calibration, whereas the ROI validation gives a more conservative validation curve because it reflects the differences between the individual pellets (since the ROIs are located at different pellets even in the same plastic sample), which is also the "operational" use of the model.

This difference in the validation results will also be reflected in the uncertainty estimates in Equation 3 and subsequently the estimated *p*-values. Figure 4 shows the *p*-values for the first independent component (IC 1) for the two validation schemes for the PS with BFRs case, the *p*-values obtained from the ROI validation are represented in blue (solid line) and the *p*-values from the LOO in red (dashed line). The LOO approach gives significantly more variables with *p*-values below 0.05 (see dashed lines in Figure 4). As the number of pixels in the ROI is rather small, the model is not stable towards the pellet-wise validation. One could have selected more pixels for more pellets and repeated this procedure for a more robust model, however, this was not the objective in this study.

Table 1. Summary of the number of independent components (ICs) extracted in the models.

ICA model for plastic sample	Number of ICs	
	Validation scheme ROI	LOO cross-validation
PS model	4	7
PA6 model	7	N/A
PP model	6	N/A
ABS model	4	N/A

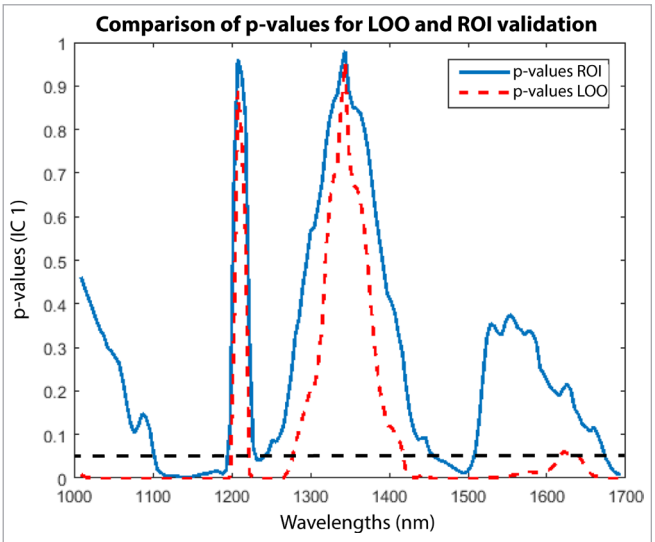

Figure 4. Comparison of the *p*-values from jack-knife estimates for LOO and ROI validation schemes. The *p*-values obtained from the ROI approach are represented by a blue solid line, and the *p*-values obtained from LOO approach by a red dashed line. A horizontal black dashed line has been added at p = 0.05 for easier visual assessment.

Hotelling's T^2 limits at the 0.005 level (almost equivalent to three standard deviations) were estimated for each individual model (Figure 5). These limits with the individual pixels for each model depicted as a line plot are shown in the right-hand plots of Figure 5. Although some pixels were outside for the individual models, the models were not recalculated without them as it was assumed that this variability was inherent in the pellets themselves. Then all pixels in the image were subject to projection onto the four models and individual pixels above the T^2 limits were filtered out before mapped back to the image. Thus, the resulting images on the left side of Figure 5 are based on classification of all pixels according to the class belonging (i.e., the plastic type used in the model). Pixels represented in white in Figure 5 correspond to the plastic type that has been modelled (i.e., they are the pixels that belong to that plastic class). The pixels of the plastic samples without BFRs (i.e., PA6 and PP) were classified almost perfectly, however, the pixels of the samples containing BFRs (i.e., PS and ABS) were confused when trying to classify them. Furthermore, pixels that belong to the PS class were classified much worse than pixels of the ABS class; the reason for this could be related to differences in the amount of BFR applied to the PS and ABS plastic samples (however, the detailed information about the amounts of BFR was not provided and the direct relationship between misclassification and increasing amount of BFR could not be verified). On the other hand, it is worth noting that the ICA algorithm extracted less independent components (ICs) in the cases of the plastics containing BFR than in the pure plastics (without BFR). The selection of the regions of interest (ROIs) could be also related to this.

The data matrices and the validation segments of the four individual models were used as inputs for a global ICA model. The scores plot for the third and the fourth independent components (IC 3 and IC 4 in Figure 6a) showed four clusters corresponding to the four plastics. However, as expected from Figure 5, the two plastics with BFRs (PS and ABS) had less compact clusters than the two pure plastics (PA6 and PP). In Figure 6a, each cluster has been marked off with an ellipse for better visualisation, and the names of the plastics have been added for easier cluster identification.

Therefore, the scores plot for the ICA global model containing the four plastics classes (Figure 6a) allowed the identification of each one of the plastics. Furthermore, the scores in Figure 6a showed that the two pure plastics (without BFRs) are similar among themselves, and the two plastics containing BFRs are similar among themselves. Therefore, IC3 separates the two pairs of plastics according to their content of BFRs (PP and PA6, the pure ones, have high scores for IC3; whilst PS and ABS, the ones containing BFRs, have low score values for IC3). In order to examine the influence of the wavelengths for IC 3, the loadings plot is shown in Figure 6b; as it can be seen, a large amount of variation, likely related to the BFR variance, is detected close to 1200 nm.

Conclusions

The plastics data set of this paper was previously used for supervised classification purposes using partial least squares discriminant analysis (PLS-DA) by Amigo *et al.* in 2015.[2] In this paper, we aimed to two main objectives: (i) an unsupervised classification of all pixels of the hyperspectral image by using local ICA models with segmented cross-validation and uncertainty estimates (including Hotelling's T^2 limit estimation) and (ii) a comparison of two different validation methods (leave-one-out and ROI selection based approaches) for ICA classification models. We challenged the unsupervised ICA-based classification method used in this paper in several ways, as, for

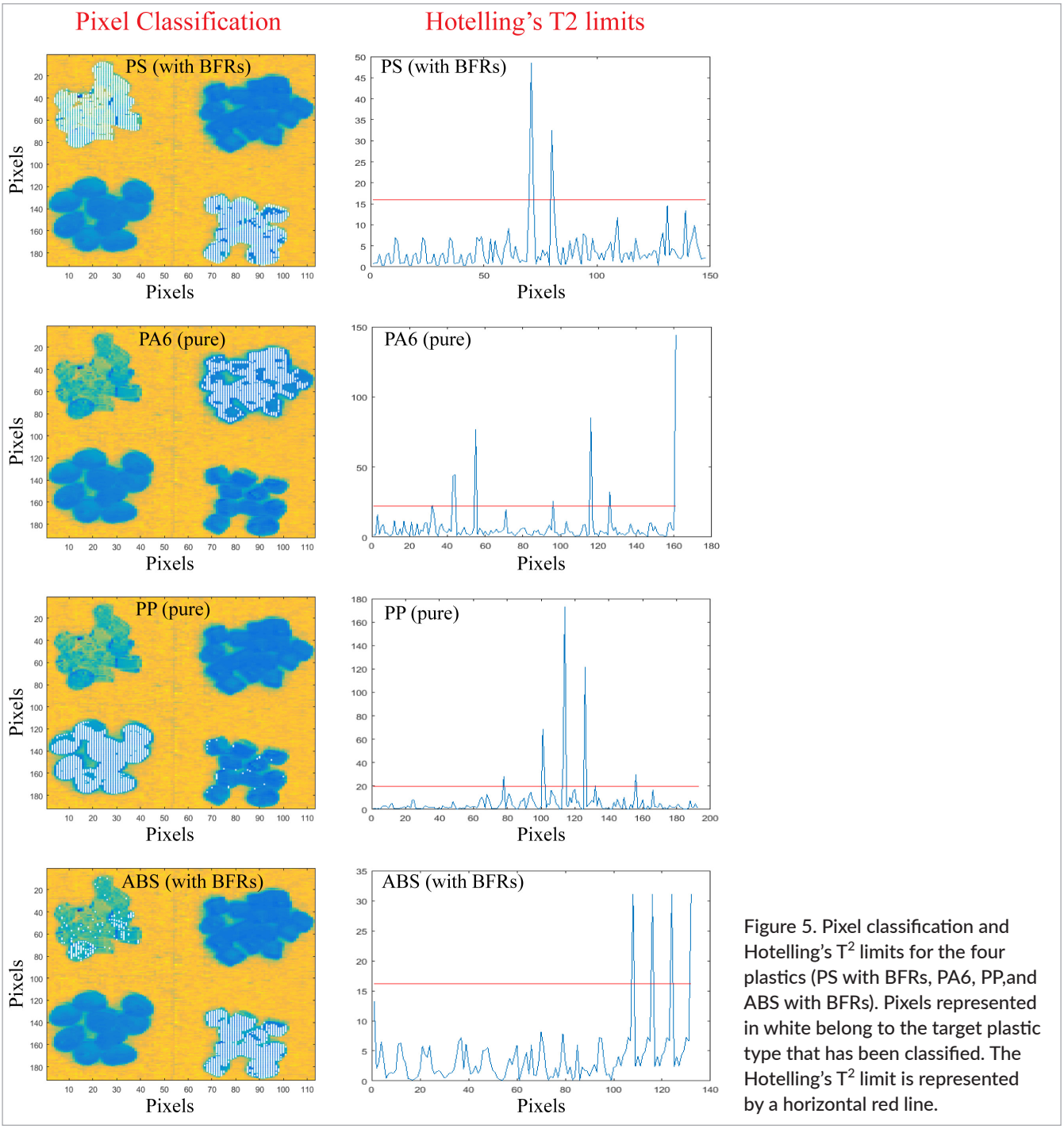

Figure 5. Pixel classification and Hotelling's T^2 limits for the four plastics (PS with BFRs, PA6, PP,and ABS with BFRs). Pixels represented in white belong to the target plastic type that has been classified. The Hotelling's T^2 limit is represented by a horizontal red line.

example, by either keeping some background pixels inside the holes of the pellets or mixing plastics with/without flame retardant content, to see if the classification was successful in those conditions. The data set was selected for these purposes.

The result of the ICA unsupervised classification showed the different composition of the four plastics as four clusters in the scores plot for the global ICA model; including the separation of the two plastics that contained brominated flame retardant (Figure 6a), although with more spread clusters. The classification of all the pixels (when mapped back to the image, see Figure 5) was perfect for the PA6 plastic, and almost perfect for the other pure plastic (PP). However, the mapping back of all the image pixels was not so good in the models for the BFR plastics (PS and ABS). This effect on the classification due to the presence of brominated flame retardant in the plastics was also reported by Amigo *et al.* in the final remarks

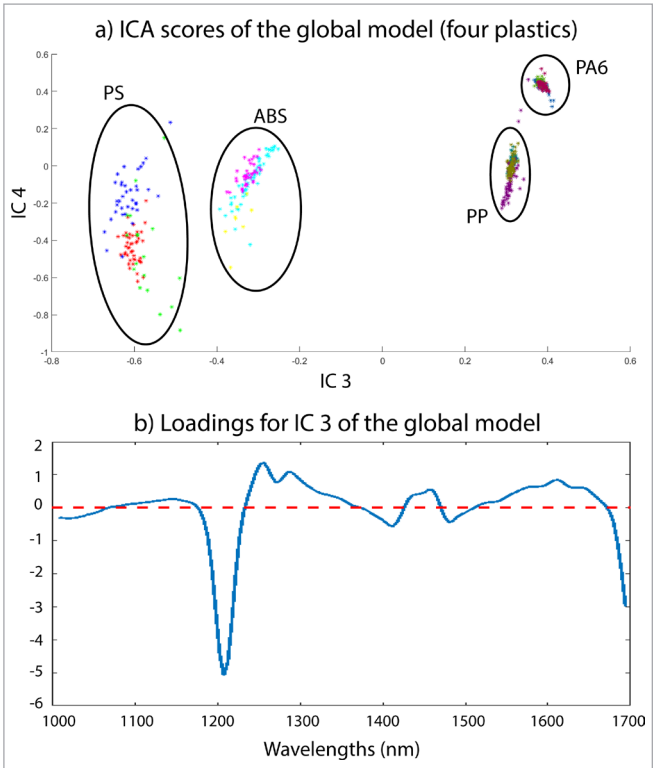

Figure 6. (a) shows the scores plot (for IC 3 vs IC 4) for the global ICA model. The third independent component (IC 3) is represented along the X-axis of the plot, and the fourth independent component (IC 4) along the Y-axis. The plastic classes are marked with ellipses and labelled according to the plastic name. (b) is a representation of the ICA loadings for IC 3, a red dashed line at the origin has been added for easier interpretation.

of Reference 2, therefore the choice of supervised vs unsupervised classification methods does not affect the outcome of the multivariate analysis.

The stability of the model was studied. The relation between the calibration and the validation curves (residual variance) is an indicator of the stability of the model across the ROIs. On the other hand, Hotelling's T^2 limits estimation helped to obtain an enhanced unsupervised ICA classification of all the pixels of the NIR hyperspectral images. The T^2 limits were based on pixels inside the ROIs for each class, and then applied to all pixels in the image.

Besides, the validation study reported some interesting conclusions. In any modelling similar to the one presented here, where different samples are used (e.g., pellets of plastic), the cross-validation must be done across samples. Otherwise, the number of components finally selected could be "not optimal". When using random validation (also tested during the data analysis) the optimal number of components was not evident from the results. And for the case of the LOO, as shown in Figure 4, the results showed too many variables (wavelengths) with *p*-values equal to zero when using the significance level; whilst the ROI-based validation provided a more informative set of *p*-values (which are related to the uncertainty assessment). Moreover, the uncertainty values (obtained by jack-knifing) were also checked to find the optimal number of significant model components; an assessment of the most relevant variables could be also performed by evaluating the jack-knife uncertainties, but this is out of the scope of this paper. Bootstrapping was not considered here because of the risk of obtaining many false positives.[19]

Acknowledgements

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The authors would like to thank Professor Rasmus Bro for help and advice, and Associate Professor José Manuel Amigo for providing the plastics data set, the HYPER-Tools toolbox (www.hypertools.org) and valuable help from the University of Copenhagen.

References

1. M. Vidal and J.M. Amigo, "Pre-processing of hyperspectral images. Essential steps before image analysis", *Chemometr. Intell. Lab. Sys.* **117,** 138–148 (2012). doi: https://doi.org/10.1016/j.chemolab.2012.05.009
2. J.M. Amigo, H. Babamoradi and S. Elcoroaristizabal, "Hyperspectral image analysis. A tutorial", *Anal. Chim. Acta* **896,** 34–51 (2015). doi: https://doi.org/10.1016/j.aca.2015.09.030
3. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space", *Philos. Mag. Ser. 6* **2,** 559–572 (1901). doi: https://doi.org/10.1080/14786440109462720
4. H. Hotelling, "Analysis of a complex of statistical variables into principal components", *J. Educ. Psych.* **24(6),** 417–441 (1933). doi: https://doi.org/10.1037/h0071325

5. P. Comon, "Independent component analysis, a new concept?", *Signal Process.* **36(3),** 287–314 (1994). doi: https://doi.org/10.1016/0165-1684(94)90029-9
6. R. Tauler, "Multivariate curve resolution applied to second order data", *Chemometr. Intell. Lab. Sys.* **30(1),** 133–146 (1995). doi: https://doi.org/10.1016/0169-7439(95)00047-X
7. A. de Juan and R. Tauler, "Multivariate curve resolution (MCR) from 2000: progress in concepts and applications", *Crit. Rev. Anal. Chem.* **36(3–4),** 163–176 (2006). doi: https://doi.org/10.1080/10408340600970005
8. C. Esquerre, A.A. Gowen, J. Burger, G. Downey and C.P. O'Donnell, "Suppressing sample morphology effects in near infrared spectral imaging using chemometric data pre-treatments", *Chemometr. Intell. Lab. Sys.* **117,** 129–137 (2012). doi: https://doi.org/10.1016/j.chemolab.2012.02.006
9. A.A. Gowen, F. Marini, C. Esquerre, C. O'Donnell, G. Downey and J. Burger, "Time series hyperspectral chemical imaging data: challenges, solutions and applications", *Anal. Chim. Acta* **705(1–2),** 272–282 (2011). doi: https://doi.org/10.1016/j.aca.2011.06.031
10. J.P. Wold, F. Westad and K. Heia, "Detection of parasites in cod fillets by using SIMCA classification in multispectral images in the visible and NIR region", *Appl. Spectrosc.* **55(8),** 1025–1034 (2001). doi: https://doi.org/10.1366/0003702011952929
11. J. Fortuna and H. Martens, "Multivariate data modelling for de-shadowing of airborne hyperspectral imaging", *J. Spectral Imaging* **6,** a2 (2017). doi: https://doi.org/10.1255/jsi.2017.a2
12. M. Dumarey, B. Galindo-Prieto, M. Fransson, M. Josefson and J. Trygg, "OPLS methods for the analysis of hyperspectral images—comparison with MCR-ALS", *J. Chemometr.* **28(8),** S687–S696 (2014). doi: https://doi.org/10.1002/cem.2628
13. A. de Juan, J. Jaumot and R. Tauler, "Multivariate curve resolution (MCR). Solving the mixture analysis problem", *Anal. Methods* **6(14),** 4964–4976 (2014). doi: https://doi.org/10.1039/C4AY00571F
14. F. Westad and M. Kermit, "Independent component analysis", in *Comprehensive Chemometrics*, Ed by S. Brown, R. Tauler and B. Walczak. Elsevier, pp. 227 (2009). doi: https://doi.org/10.1016/B978-044452701-1.00045-4
15. F. Westad and M. Kermit, "Cross validation and uncertainty estimates in independent component analysis", *Anal. Chim. Acta* **490(1–2),** 341–354 (2003). doi: https://doi.org/10.1016/S0003-2670(03)00090-4
16. S. Piqueras, J. Burger, R. Tauler and A. de Juan, "Relevant aspects of quantification and sample heterogeneity in hyperspectral image resolution", *Chemometr. Intell. Lab. Sys.* **117,** 169–182 (2012). doi: https://doi.org/10.1016/j.chemolab.2011.12.004
17. M. Stone, "Cross-validatory choice and assessment of statistical predictions", *J. Royal Stat. Soc. B* **36,** 111 (1974).
18. S. Wold, "Cross-validatory estimation of the number of components in factor and principal components models", *Technometrics* **20,** 397–405 (1978). doi: https://doi.org/10.2307/1267639
19. F. Westad and F. Marini, "Validation of chemometric models – a tutorial", *Anal. Chim. Acta* **893,** 14–24 (2015). doi: https://doi.org/10.1016/j.aca.2015.06.056
20. F. Westad, N.K. Afseth and R. Bro, "Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression", *Anal. Chim. Acta* **595(1–2),** 323–327 (2007). doi: https://doi.org/10.1016/j.aca.2007.02.015
21. C.M. Andersen and R. Bro, "Variable selection in regression—a tutorial", *J. Chemometr.* **24(11–12),** 728–737 (2010). doi: https://doi.org/10.1002/cem.1360
22. V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste and C. Sterna, "Elimination of uninformative variables for multivariate calibration", *Anal. Chem.* **68(21),** 3851–3858 (1996). doi: https://doi.org/10.1021/ac960321m
23. B. Galindo-Prieto, L. Eriksson and J. Trygg, "Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS)", *J. Chemometr.* **28(8),** 623–632 (2014). doi: https://doi.org/10.1002/cem.2627
24. B. Galindo-Prieto, J. Trygg and P. Geladi, "A new approach for variable influence on projection (VIP) in O2PLS Models", *Chemometr. Intell. Lab. Sys.* **160,** 110–124 (2017). doi: https://doi.org/10.1016/j.chemolab.2016.11.005
25. L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, "Interval partial least-

squares regression (IPLS): a comparative chemometric study with an example from near-infrared spectroscopy", *Appl. Spectrosc.* **54(3),** 413–419 (2000). doi: https://doi.org/10.1366/0003702001949500

26. C. Abrahamsson, J. Johansson, A. Sparén and F. Lindgren, "Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets", *Chemometr. Intell. Lab. Sys.* **69(1–2),** 3–12 (2003). doi: https://doi.org/10.1016/s0169-7439(03)00064-9

27. T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr and O.M. Kvalheim, "Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles", *Anal. Chem.* **81(7),** 2581–2590 (2009). doi: https://doi.org/10.1021/ac802514y

28. F. Marini, A. Roncaglioni and M. Novič, "Variable selection and interpretation in structure–affinity correlation modeling of estrogen receptor binders", *J. Chem. Inf. Model.* **45(6),** 1507–1519 (2005). doi: https://doi.org/10.1021/ci0501645

29. A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures", *Anal. Chem.* **36(8),** 1627–1639 (1964). doi: https://doi.org/10.1021/ac60214a047

30. R.J. Barnes, M.S. Dhanoa and S.J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra", *Appl. Spectrosc.* **43(5),** 772–777 (1989). doi: https://doi.org/10.1366/0003702894202201

31. J. MacQueen, "Some methods for classification and analysis of multivariate observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif., p. 281 (1967).

32. F. Westad, "Independent component analysis and regression applied on sensory data", *J. Chemometr.* **19(3),** 171–179 (2005). doi: https://doi.org/10.1002/cem.920

33. D. Jouan-Rimbaud Bouveresse, A. Moya-González, F. Ammari and D.N. Rutledge, "Two novel methods for the determination of the number of components in independent components analysis models", *Chemometr. Intell. Lab. Sys.* **112,** 24–32 (2012). doi: https://doi.org/10.1016/j.chemolab.2011.12.005

34. A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*. John Wiley & Sons, Inc., Ch. 7 (2002). doi: https://doi.org/10.1002/0471221317.ch7

35. B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation", *Am. Stat.* **37(1),** 36–48 (1983). doi: https://doi.org/10.2307/2685844