Supplementary Document: A Machine Learning Toolkit for CRISM Image Analysis

Emanuele Plebani[a], Bethany L. Ehlmann[b,e], Ellen K. Leask[c], Valerie K. Fox[d] and M. Murat Dundar[a,*]

[a]Computer and Information Sciences Department, Indiana University - Purdue University, Indianapolis, 46202, IN, USA
[b]Div. of Geological & Planetary Sciences, California Institute of Technology, Pasadena, 91125, CA, USA
[c]John Hopkins University Applied Physics Laboratory, Laurel, 20723, MD, USA
[d]Department of Earth and Environmental Sciences, University of Minnesota, Minneapolis, 55455, MN, USA

ARTICLE INFO

Keywords:

ABSTRACT

This document is related to the research article titled "A Machine Learning Toolkit for CRISM Image Analysis" and contains details of the derivation of the hierarchical Bayesian Model (HBM), and additional experiments by the developed toolbox.

1. Hierarchical Bayesian Model

We review here the generative model for the spectra in the dataset:

$$\text{Data model:} \quad \boldsymbol{x}_{ijk} \sim \mathcal{N}(\boldsymbol{\mu}_{jk}, \Sigma_k) \tag{1}$$

$$\text{Local prior:} \quad \boldsymbol{\mu}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k \kappa_1^{-1}) \tag{2}$$

$$\text{Global prior:} \quad \boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_k \kappa_0^{-1}) \tag{3}$$

$$\Sigma_k \sim \mathcal{IW}(\Sigma_0, m) \tag{4}$$

Different mineral classes are indexed by k, different images in which a mineral class k appear (observed instances) are indexed by j, and different pixels of a mineral class k in image j are indexed by i.

2. Model hyperparameters

The model assumes that pixels $\boldsymbol{x}_{ijk}$ are distributed according to a Gaussian distribution with mean $\boldsymbol{\mu}_{jk}$ and covariance matrix Σ_k. Each mineral class is characterized by the parameters $\boldsymbol{\mu}_k$ and Σ_k of their corresponding local prior, and the global prior is characterized by $\boldsymbol{\mu}_0$ and Σ_0. The parameter $\boldsymbol{\mu}_0$ is the mean of the Gaussian priors defined over the mean vectors of mineral classes and κ_0 is a scaling constant that adjusts the dispersion of the centers of mineral classes around $\boldsymbol{\mu}_0$. A smaller value for κ_0 suggests that means of mineral classes are expected to be farther apart from each other, whereas a larger value suggests they are expected to be closer. The parameters Σ_0 and m dictate the expected covariance under the inverse Wishart distribution assumption where the expected covariance is $\mathrm{E}(\Sigma|\Sigma_0, m) = \frac{\Sigma_0}{m-d-1}$, where d denotes the number of channels used. The minimum feasible value of m is equal to $d+2$, and the larger the m is, the less individual covariance matrices will deviate from the expected shape. The scaling constant κ_1 adjusts the dispersion of the means of detected mineral instances around the centers of their corresponding local prior. A larger κ_1 leads to smaller variations in instance means with respect to the means of their corresponding local prior, suggesting small variations among observed instances of the mineral class. On the other hand, a smaller κ_1 dictates larger variations among instances. μ_0 is set to the mean of the sample mean of mineral classes and Σ_0 is set to $\bar{\Sigma}/s$, where $\bar{\Sigma}$ is the average of the sample covariances estimated one for each mineral class and s is a scaling constant.

*Corresponding author
mdundar@iupui.edu (M.M. Dundar); (M.M. Dundar)
https://cs.iupui.edu/~mdundar (M.M. Dundar); (M.M. Dundar)
ORCID(s): 0000-0002-7809-9616 (E. Plebani); 0000-0001-5752-468X (M.M. Dundar)

The remaining hyperparameters (κ_0, κ_1, m, s) are tuned on a small subset of well-characterized images by visually inspecting classification maps and averaged spectra for each mineral detection. For all models trained in this study the same values are used for these parameters, i.e., $\kappa_0 = 1$, $\kappa_1 = 100$, $m = d + 2$, $s = 1$.

2.1. Derivation of the Posterior Predictive Distribution

The likelihood of a pixel $\boldsymbol{x}$ originating from a mineral class k is obtained by evaluating the posterior predictive distribution (PPD). For our two-layer Gaussian mixture architecture PPDs are derived in the form of *Student-t* distributions by integrating out unknown mean vectors and covariance matrices of local priors and their observed instances. In Bayesian inference "integrating out" is the process of eliminating the effect of one or more random variables during marginalization of the probability distribution so that the distribution only depends on a subset of the random variables. This is achieved by integrating the original probability distribution with respect to the random variables that need to be eliminated. This directly links observed spectral data with the hyperparameters of the model (κ_0,κ_1,m,μ_0, Σ_0). Optimizing hyperparameters with pixel data from the training dataset encodes information about observed spectral variations into the model. In what follows we sketch the derivations of PPDs.

Let $\boldsymbol{x}$ be the spectral representation of a pixel in an image. To classify $\boldsymbol{x}$ we need to evaluate the PPD for each mineral class, i.e., evaluate the distribution $P(\boldsymbol{x}|\bar{\boldsymbol{x}}_{1k}, \ldots, \bar{\boldsymbol{x}}_{n_k k}, S_{1k}, \ldots, S_{n_k k})$ where $\bar{\boldsymbol{x}}_{jk}$ and S_{jk} are the sample mean vector and sample covariance matrix of the observed instance j of local prior k. The derivation of $P(\boldsymbol{x}|\bar{\boldsymbol{x}}_{1k}, \ldots, \bar{\boldsymbol{x}}_{n_k k}, S_{1k}, \ldots, S_{n_k k})$ can be carried out in four steps.

In the first step we integrate over the unknown mean vector $\boldsymbol{\mu}_{jk}$ of instance j of mineral class k and obtain the conditional distribution of the sample mean $\bar{x}_{jk}$ with respect to the unknown mean vector $\boldsymbol{\mu}_k$ of the local prior:

$$P(\bar{\boldsymbol{x}}_{jk}|\boldsymbol{\mu}_k, \Sigma_k) = \mathcal{N}\left(\boldsymbol{\mu}_k, \Sigma_k(\frac{1}{n_{jk}} + \frac{1}{\kappa_1})\right) \tag{5}$$

where n_{jk} is the number of pixels available for instance j of mineral class k in the training set.

In the second step we derive the posterior distribution of the mean vector $\boldsymbol{\mu}_k$ from the Bayes rule and show that the posterior mean is the weighted average of the sample mean vectors of observed instances of mineral class k and their corresponding local prior mean:

$$\begin{aligned} P(\boldsymbol{\mu}_k|\bar{\boldsymbol{x}}_{1k}, \ldots, \bar{\boldsymbol{x}}_{n_k k}, \boldsymbol{\mu}_0, \Sigma_k, \kappa_0) &= \mathcal{N}(\bar{\boldsymbol{\mu}}_k, \bar{\Sigma}_k) \\ \bar{\boldsymbol{\mu}}_k &= \frac{\sum_{j=1}^{n_k} \frac{n_{jk}\kappa_1}{(n_{jk}+\kappa_1)} \bar{\boldsymbol{x}}_{jk} + \kappa_0 \boldsymbol{\mu}_0}{\sum_{j=1}^{n_k} \frac{n_{jk}\kappa_1}{(n_{jk}+\kappa_1)} + \kappa_0} \\ \bar{\Sigma}_k &= \bar{\kappa}_k^{-1} \Sigma_k \\ \bar{\kappa}_k &= \sum_{j=1}^{n_k} \frac{n_{jk}\kappa_1}{(n_{jk}+\kappa_1)} + \kappa_0 \end{aligned}$$

where n_k is the number of observed instances of mineral class k, i.e., the number of training images in which mineral class k is detected.

In the third step we derive the posterior distribution for Σ_k by combining Wishart terms corresponding to all observed instances of mineral class k.

$$P(\Sigma_k|S_{1k}, \ldots, S_{n_k k}) = \mathcal{W}^{-1}(\bar{S}_s, \bar{m}_s) \tag{6}$$

$$\bar{S}_s = \Sigma_0 + \sum_{j=1}^{n_k} S_{jk} \tag{7}$$

$$\bar{m}_s = m + \sum_{j=1}^{n_k} (n_{jk} - 1)) \tag{8}$$

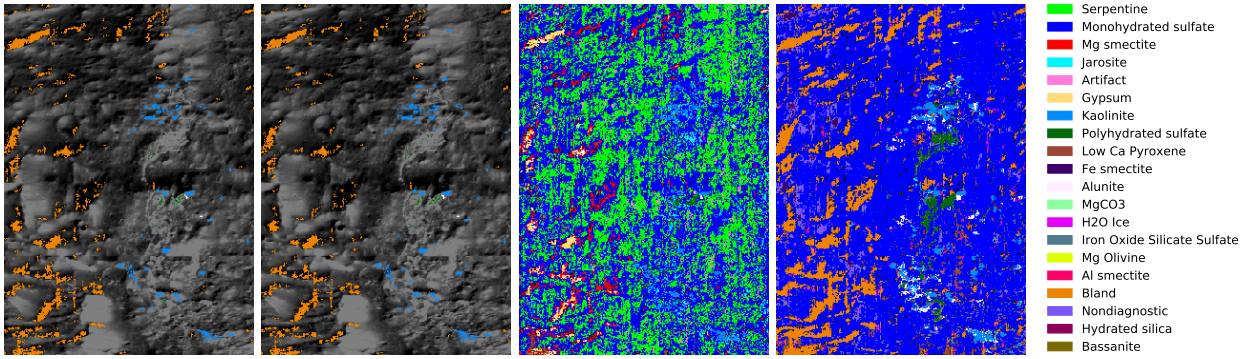
Finally, in the fourth step we derive the posterior predictive distribution for mineral class k as in (9) by integrating out parameters μ_k and Σ_k. Thanks to conjugacy, in our model this operation produces a closed form *Student-t*

distribution:

$$
\begin{aligned}
P(\boldsymbol{x}|\bar{\boldsymbol{x}}_{1k},\dots,\bar{\boldsymbol{x}}_{n_k k},S_{1k},\dots,S_{n_k k}) &= T(\boldsymbol{x}_{ji}|\bar{\boldsymbol{\mu}}_k,\bar{\Sigma}_s,\bar{\nu}_s) \\
\bar{\Sigma}_s &= \frac{\bar{S}_s}{\frac{\bar{\kappa}_s \nu_s}{\bar{\kappa}_s+1}} \\
\bar{\kappa}_s &= \frac{(\sum_{j=1}^{n_k}\frac{n_{jk}\kappa_1}{(n_{jk}+\kappa_1)}+\kappa_0)\kappa_1}{\sum_{j=1}^{n_k}\frac{n_{jk}\kappa_1}{(n_{jk}+\kappa_1)}+\kappa_0+\kappa_1} \\
\bar{\nu}_s &= m+\sum_{j=1}^{n_k}(n_{jk}-1)-d+1
\end{aligned}
\tag{9}
$$

where the three parameters $\bar{\boldsymbol{\mu}}_k$, $\bar{\Sigma}_s$, and $\bar{\nu}_s$ are the location vector, scale matrix, and the degrees of freedom, respectively.

3. Additional experiments



(a) ensemble, default m (b) subsampled, default m (c) single, default m (d) single $m = 10^5$

Figure 1: Predictions maps for different HBMs. The original ensemble and the ensemble on a subsampled dataset have very similar predictions, while the single model is overconfident and sensitive to hyperparameters, as shown by the completely different maps for different values of m.

We report experiments on different feature and dataset sampling strategies to show that the ensemble model we implemented for mineral classification generates more accurate predictions. We evaluated algorithms both quantitatively and qualitatively. For quantitative evaluation we selected a portion of the labeled images as validation and computed the F1 score on this split. For qualitative evaluation we trained models on the labeled dataset and manually verified detected minerals on the test image HRL00016CFE using the same post-processing steps and thresholds employed for the proposed ensemble.

3.1. Single HBM model

We trained a single model on all channels between 1.047 and 2.648 µm after excluding all spectral bands with known artifacts. In the first experiment we used the default value for m and in the second one we optimized m to maximize F1 score on the validation set. The model optimized for m generated the highest F1 score as shown in Table 1. We also tuned other hyperparameters (κ_0, κ_1, Σ_0, and μ_0) but none resulted in changes as drastic as m. Although the single model outperforms the ensemble model modestly in terms of F1 score for higher m, when we test models on a full CRISM image we observe that these results are

Model	F1 score
ensemble	0.58
ensemble, $m = 10^5$	0.69
single	0.49
single, $m = 10^5$	0.70
subsampled	0.60
subsampled, $m = 10^4$	0.68

Table 1: F1 scores of HBMs with different structure on the training-test split; the original *ensemble*, a *single* model on clean spectral bands and the ensemble on a *subsampled* training set.

quite misleading. As shown in Figure 1, the predictions by the single model (c and d) are highly overconfident; all pixels classified to a specific class are classified almost with certainty, making probability-based thresholding impractical. As a result, the segmentation maps for the single model are completely dominated by false positive detections and over-segmented true positives.

This comes as no suprise as the labeled dataset did not include any outlier pixels, validation data that was generated from this labeled set did not reflect the real-world characteristics of a typical CRISM image that includes a large number of mixed spectral pixels, pixels representing dust and soil, or other pixels that cannot be assigned to mineral classes used for training. As a result we see that models that perform well on the validation dataset may not necessarily perform well on the actual CRISM image. For the same image the proposed ensemble model generates weighted probability scores more uniformly distributed between 0 and 1, which makes thresholding-based post processing more practical and functional. As shown in Figures 1a and b the proposed model accurately detects several mineral phases in this image with no false positives. Using higher thresholds (focusing on the most confident detections) to reduce false positives may come with a trade-off; some true mineral phases might be undersegmented when some of the pixels from these minerals are detected with lower probability scores. Finally, the single model is also highly sensitive to the choice of hyperparameters as two very different segmentation maps were generated for the default m and for $m = 10^5$. On the other hand, the proposed ensemble is more robust to the value of m.

3.2. Ensemble HBM with subsampled dataset

We also trained an ensemble model with the same channels and weights as the original ensemble model but with each submodel trained on a different subset of the original labeled dataset. Each subset is obtained by sub-sampling with replacement the dataset by a factor of 0.3, stratified by class label. While this approach helps for methods prone to overfitting such as decision trees, for the proposed HBM it only increases the variance of the parameter estimates. The F1 score is comparable to the one obtained by the original ensemble without any subsampling. The best F1 score is achieved for $m = 10^4$ (see Table 1). The segmentation maps generated for the test image are comparable with only small differences in predicted pixels (see Figure 1).