# CLUSTER VALIDATION CRITERIA FOR IMAGE SEGMENTATION

*Jong-Kae Fwu and Petar M. Djurić*

Department of Electrical Engineering
State University of New York at Stony Brook
Stony Brook, New York 11794-2350
fwu@sbee.sunysb.edu djuric@sbee.sunysb.edu

## ABSTRACT

In this paper cluster validation criteria for piecewise constant image segmentation are proposed. All the criteria are based on the maximum a posteriori (MAP) principle and derived and implemented by four different, but related approaches. They are obtained by using Taylor expansions and three of them are derived by Bayesian predictive densities. The third and fourth criteria are implemented by the bootstrap technique, and their evaluations are, therefore, computationally more intensive than the evaluations of the first two. The proposed rules are compared by computer simulations with the widely used AIC and MDL criteria.

## 1. INTRODUCTION

One of the most important problems in image processing is the cluster validation. It arises usually with the task of image segmentation, where the image pixels are grouped into different classes according to some prespecified attributes. In general, the segmentation can be supervised or unsupervised, that is, done automatically without operator's assistance. In many practical applications, the automatic segmentation is the preferred or the only possible option. The majority of unsupervised segmentation methods known from the open literature assume that the number of classes is known [5], [9]. Since in practice however, this is usually not the case, there is an additional problem to be resolved, i.e., the determination of the number of various classes of pixels in the observed data [8], [11]. This problem is also known as cluster validation. It has been recognized that cluster validation is a very difficult task, so it is not surprising that there is still a significant ongoing research whose aim is getting improved solutions [10].

Some relatively new methods for cluster validation are based on the penalized maximum likelihood criteria. They are composed of data and penalty terms, where the data term is the negative of the loglikelihood function and the penalty term represents the cost for overparameterization. The data term quantifies the fitness of the segmented image to the original image and decreases as the number of classes increases, whereas the penalty term for larger number of classes increases. Among the most popular criteria of this form are the Akaike's information criterion (AIC) [11] and the Minimum Description Length (MDL) [10]. The AIC, however, is not consistent, that is, the probability of selecting the correct number of classes does not tend to one as the size of the image increases. This feature is a result of the incorrect penalty which is not a function of the data size. The MDL on the other hand, has a larger penalty that grows with the image size, and as a consequence, the MDL tends to choose models with fewer parameters than the AIC. In general, the straightforward application of the AIC and MDL can lead to incorrect results [3]. For example, in cluster validation it has been found that the AIC and MDL frequently overestimate the number of different classes.

Here we derive approximated MAP criteria whose forms are also penalized likelihoods. They all have identical data terms as the AIC and MDL, but different penalties. In our derivations, we have allowed for incorporation of spatial constraints, which can play a significant role in the criteria. The constraints and the inherent spatial ordering of the pixels are modeled by Markov Random Fields (MRF's). Since the exact MAP criterion for cluster validation is computationally infeasible, we have resorted to various approximations. In our work, one important deviation from the exact MAP solution for the number of classes is that our MAP criteria are obtained by using the joint probability of number of classes and the underlying image. Thus, the solution is the most probable number of classes and the underlying image considered together.

We applied our criteria with the tree-structure iterated conditional modes method (TS-ICM) for segmentation [5], [6]. Our criteria combined with the TS-ICM provide a complete automatic procedure for cluster validation and segmentation where no threshold settings are required. In the simulations, the procedure worked very well on all of the used images.

This paper is organized as follows: In Section 2, we define the problem we want to solve. In Section 3, four criteria for cluster validation are derived. In the derivation, the concepts of Bayesian predictive densities (BPD) and the bootstrap are exploited. The details of their implementation and simulation results are presented in Section 4 and 5, respectively.

## 2. PROBLEM STATEMENT

Let the observed image be given by $Y = \{y_{ij} : (i,j) \in \mathcal{S}\}$, where $\mathcal{S} = \{(i,j) : \ 1 \leq i < M_1, 1 \leq j < M_2\}$ denotes a

This work was supported by the National Science Foundation under Award No. MIP-9506743

rectangular lattice. The underlying image is represented by $X=\{x_{ij} : (i,j) \in \mathcal{S}\}$ and it is comprised of pixels that belong to one of $m$ classes, where the number of classes is *not known*. We model $X$ as an MRF whose probability distribution is denoted by $f(X)$. In our derivations we assumed that $f(X)$ is a Gibbs distribution. Each class of pixels is characterized by its own set of parameters which are also unknown.

The underlying image is corrupted by noise $W$ so that the observed image is given by

$$Y = X + W \tag{1}$$

where $W = \{w_{ij} : (i,j) \in \mathcal{S}\}$. The noise samples are considered to be independent and zero mean Gaussian. Their variance is unknown and in general may depend on the pixel class they are associated with.

Given the observed data and the assumptions of the data and noise models, our objective is to estimate the number of different classes $m$.

## 3. PROPOSED CRITERIA

The number of classes can be determined by using the MAP criterion given by

$$\begin{aligned} \hat{m}_{\text{MAP}} &= \arg\max_k \{f(k|Y)\} && (2)\\ &= \arg\max_k \{\sum_X f(k|X,Y)\} && (3) \end{aligned}$$

where $k \in \{1, 2, \cdots, m_{\max}\}$ with $m_{\max}$ being the maximum number of possible classes, $f(k|Y)$ is the posterior probability mass function of $k$ classes given the data $Y$, and $f(k|X,Y)$ is the probability mass function of $k$ given the underlying image $X$ and the observed image $Y$. This criterion is very difficult to implement because of its extremely large number of terms in the summation in (3).

An alternative is to seek for the MAP solution that jointly yields the number of classes and the most probable underlying image. For an image with $k$ classes, we can write

$$f(X,k|Y) = \frac{f(Y|X,k)f(X|k)f(k)}{f(Y)} \tag{4}$$

where $f(Y|X,k)$ is the density function of the data conditioned on the number of classes and the underlying image, and $f(X|k)$ is the Gibbs distribution of the underlying image with $k$ classes. Finally, $f(k)$ is the a priori probability of $k$ classes. The maximization of $f(X,k|Y)$ does not depend on $f(Y)$, and therefore $f(Y)$ can be ignored. If we assume uniform $f(k)$, the MAP solution of (4) becomes

$$(\hat{X}, \hat{m})_{\text{MAP}} = \arg\max_{(X,k)} \{f(Y|X,k)f(X|k)\}. \tag{5}$$

Given the number of classes $k$, we can find the underlying image $X$ for $k = 1, 2, \cdots, m_{\max}$ by an unsupervised image segmentation technique [5], [6], where the segmented image is obtained from

$$\hat{X} = \arg\max_X \{f(Y|X,k)f(X|k)\}. \tag{6}$$

After we find the underlying images $\hat{X}$ for $k = 1, 2, \cdots, m_{\max}$, we select the number of classes by

$$\begin{aligned} \hat{m}_{\text{MAP}_1} &= \arg\max_k \{f(Y|\hat{X},k)f(\hat{X}|k)\} && (7)\\ &= \arg\min_k \{-\ln f(Y|\hat{X},k) - \ln f(\hat{X}|k)\}. && (8) \end{aligned}$$

Now, to determine $f(Y|\hat{X},k)$ we use

$$f(Y|\hat{X},k) = \int_{\Theta_k} f(Y|\hat{X},\theta,k)f(\theta|\hat{X},k)d\theta \tag{9}$$

where $\theta$ is the parameter vector, and $\Theta_k$ is the parameter space when there are $k$ classes. If we Taylor expand $f(Y|\hat{X},\theta,k)$ around the maximum likelihood estimate of $\theta$, $\hat{\theta}$, we obtain

$$\begin{aligned} &\int_{\Theta_k} f(Y|\hat{X},\theta,k)f(\theta|\hat{X},k)d\theta \\ &\simeq (2\pi)^{\frac{k}{2}} |\mathcal{H}_k|^{-\frac{1}{2}} f(Y|\hat{X},\hat{\theta},k)f(\hat{\theta}|\hat{X},k) \end{aligned} \tag{10}$$

where $\mathcal{H}_k$ is the Hessian of $-\ln f(Y|\hat{X},\theta,k)$ evaluated at $\hat{\theta}$.

From (7) and (10), and using the approximation $\ln|\mathcal{H}_k| \simeq \sum_{i=1}^{k} 2\ln n_i$, where $n_i$ is the number of pixels in class $i$, we write our first MAP criterion as

$$\begin{aligned} \hat{m}_{\text{MAP}_1} = \arg\min_k \{&-\ln f(Y|\hat{X},\hat{\Theta},k) + \sum_{i=1}^{k} \ln n_i \\ &- \ln PL(\hat{X}|k)\}. \end{aligned} \tag{11}$$

In (11) we have dropped irrelevant terms and used the concept of pseudo-likelihood, to avoid the intractable partition function in $f(X)$ [1]. Thus, $PL(\hat{X}|k)$ is the pseudo-likelihood function evaluated for $X = \hat{X}$.

We continue with the description of the second criterion which is based on BPD's. These densities have already been successfully applied to model selections in other signal processing problems [2]. To employ them, we partition the available data into a training set $Y_T$ and a validation set $Y_V$. The MAP criterion can then be written as

$$\hat{m}_{\text{MAP}_2} = \arg\min_k \{-\ln f(Y_V|Y_T,\hat{X},k) - \ln f(\hat{X}|k)\} \tag{12}$$

where $Y_T$ are the training data and $Y_V$ are all the remaining data. The first term in (12) can be manipulated similarly as $f(Y|\hat{X},\theta,k)$ in (9). As a result, we obtain

$$\begin{aligned} &\int_{\Theta_k} f(Y_V|Y_T,\hat{X},\theta,k)f(\theta|Y_T,\hat{X},k)d\theta \\ &\simeq (2\pi)^{\frac{k}{2}} |\mathcal{H}'_k|^{-\frac{1}{2}} f(Y_V|Y_T,\hat{X},\hat{\theta},k)f(\hat{\theta}|Y_T,\hat{X},k) \end{aligned}$$

where $\mathcal{H}'_k$ is the Hessian of $-\ln f(Y_V|Y_T,\hat{X},\theta,k)$ evaluated at $\hat{\theta}$. We then approximate (12) by

$$\begin{aligned} \hat{m}_{\text{MAP}_2} = \arg\min_k \{&-\ln f(Y_V|Y_T,\hat{X}_k,\hat{\theta},k) + \frac{1}{2}\ln|\mathcal{H}'_k| \\ &- \ln PL(\hat{X}|k)\}. \end{aligned} \tag{13}$$

The criteria $\text{MAP}_1$, and $\text{MAP}_2$ are both penalized maximum likelihood functions with three terms, a data term

and two penalty terms. The data term corresponds to the fitting error of the image model based on $k$ classes. The first penalty term penalizes for using additional model parameters and the second for spatial discontinuities.

Next we outline the derivation of the third criterion. Instead of using the Hessians in the Taylor expansion, we use the covariance matrices of the estimated parameters. Then, the estimation of these matrices is carried out by the well known bootstrap method [4], [7]. In this criterion, the penalty is built in the covariance matrix of the estimated parameters, and thereby, we avoid the derivation of the penalty due to additional model parameters. By using the Bayes' theorem, we can write for the BPD

$$\begin{aligned} & f(Y_V|Y_T,\hat{X},k) \\ = & \frac{\int_{\Theta_k} f(Y|\hat{X},\theta,k)f(\theta|\hat{X},k)d\theta}{\int_{\Theta'_m} f(Y_T|\hat{X},\theta',k)f(\theta'|\hat{X},k)d\theta'}. \end{aligned} \tag{14}$$

Again, by Taylor expansion, the $-\ln(\cdot)$ of (14) can be approximated by

$$-\ln f(Y_V|Y_T,\hat{X},k) = L(Y,Y_T,k) + P(\mathcal{C}_k,\mathcal{C}'_k) \tag{15}$$

where

$$L(Y,Y_T,k) = -\ln f(Y|\hat{X},\hat{\theta},k) + \ln f(Y_T|\hat{X},\hat{\theta}',k)$$

and

$$P(\mathcal{C}_k,\mathcal{C}'_k) = -\frac{1}{2}\ln|\mathcal{C}_k| + \frac{1}{2}\ln|\mathcal{C}'_k|.$$

Here $\mathcal{C}_k$ and $\mathcal{C}'_k$ are the covariance matrices of the estimated parameters from $Y$ and $Y_T$, respectively. Note that, $L(Y,Y_Y,m)$ represents the difference of log-likelihood functions obtained from the whole data and the training data. From (12) and (15), the final estimator becomes

$$\hat{m}_{\mathrm{MAP}_3} = \arg\min_k\{L(Y,Y_T,k) + P(\mathcal{C}_k,\mathcal{C}'_k) - \ln PL(\hat{X}|k)\}. \tag{16}$$

Finally, the fourth criterion is also based on (12), however, it is implemented in a different way. Namely, if we express the Bayesian predictive density by

$$\begin{aligned} & f(Y_V|Y_T,\hat{X},k) = \\ & \int_{\Theta_k} f(Y_V|Y_T,\hat{X},\theta,k)f(\theta|Y_T,\hat{X},k)d\theta \end{aligned} \tag{17}$$

and by Taylor expanding $\ln f(Y_V|Y_T,\hat{X},\theta,k)$, we have

$$\begin{aligned} & f(Y_V|Y_T,\hat{X},k) = \\ \int_{\Theta_k} & f(Y_V|Y_T,\hat{X},\hat{\theta},k)e^{D(\theta,\mathcal{C}_k)}f(\theta|Y_T,\hat{X},k)d\theta \end{aligned}$$

where

$$D(\theta,\mathcal{C}_k) = -\frac{1}{2}(\theta-\hat{\theta})^T\hat{C}_k^{-1}(\theta-\hat{\theta}). \tag{18}$$

By moving the $f(Y_V|Y_T,\hat{X},\hat{\theta},k)$ in front of the integration, the equation becomes

$$\begin{aligned} & f(Y_V|Y_T,\hat{X},k) \\ & = f(Y_V|Y_T,\hat{X},\hat{\theta},k)G(\theta,\hat{C}_k,k) \end{aligned} \tag{19}$$

where $\hat{C}_k$ is obtained from the bootstrap data and

$$\begin{aligned} G(\theta,\hat{C}_k,k) = & \int_{\Theta_k}[e^{-\frac{1}{2}(\theta-\hat{\theta})^T\hat{C}_k^{-1}(\theta-\hat{\theta})} \\ & \times f(\theta|Y_T,\hat{X},k)]d\theta. \end{aligned} \tag{20}$$

The MAP criterion becomes

$$\begin{aligned} \hat{m}_{\mathrm{MAP}_4} = & \arg\min_k\{-\ln f(Y_V|\hat{X},\hat{\theta},k) - \ln PL(\hat{X}|k) \\ & - \ln G(\theta,\hat{C}_k,k)\}. \end{aligned} \tag{21}$$

The implementations of (16), and (21) are discussed in more detail in the next section.

## 4. EVALUATION OF THE MAP CRITERION BY THE BOOTSTRAP TECHNIQUE

To compute the MAP criteria (16) and (21), we first use a method for unsupervised image segmentation. Once the image is segmented, we apply the bootstrap scheme to generate sets of bootstrap data. Using these data, we evaluate the covariance matrices needed in our criteria. The bootstrap procedure was applied as follows. More specifically, given the observed image data $Y$ and the number of classes equal to $k$, we segment the image into $k$ classes, i.e. we find $\hat{X}$. Knowing $\hat{X}$, we then generate the bootstrap data $Y_b^* = \{y_s^* : s \in S\}$, for $b = 1,2,\cdots B$, by randomly selecting $y_s^*$ from $\{y_{s'} : s' \in \mathcal{S}, x_{s'} = x_s\}$. With $Y_b^*$ for $b = 1,2,\cdots B$, we calculate straightforwardly the parameters $\theta^{*b}$, and subsequently obtain the covariance matrix of the estimated parameters $\hat{C}_k$.

By partitioning the data into training and testing subsets and using the above procedure, we easily evaluate the criterion (16). For the fourth rule we also compute the value of the integral in (21), that is (20). The evaluation of the criterion is carried out according to

$$\begin{aligned} \hat{m}_{\mathrm{MAP}_4} = & \arg\min_k\{-\ln f(Y_V|\hat{X},\hat{\theta},k) \\ & - \ln\left[\frac{1}{B}\sum_{b=1}^{B} e^{-\frac{1}{2}(\theta^{*b}-\hat{\theta})^T\hat{C}_k^{-1}(\theta^{*b}-\hat{\theta})}\right] - \ln PL(\hat{X}|k)\} \end{aligned}$$

where the parameters $\theta^{*b}$ and the covariance matrix $\hat{C}_k$ are obtained from the bootstrap data.

## 5. SIMULATION RESULTS

To verify the performance of the MAP criteria, we tested them on synthesized MR brain images. We also compared them with the widely used AIC, and MDL criteria. The size of these images was $256 \times 256$. In Figure 1, from top to bottom we present the true image, noisy image, and segmented image. Table 1 displays the simulation results from 100 independent trials of the $\mathrm{MAP}_1$, $\mathrm{MAP}_2$, $\mathrm{MAP}_3$, $\mathrm{MAP}_4$, AIC, and MDL rules for various contrast-to-noise (CNR) ratios. The results show that the $\mathrm{MAP}_2$, $\mathrm{MAP}_3$, and $\mathrm{MAP}_4$ perform best among the six. The $\mathrm{MAP}_1$ criterion chooses the correct number of classes for high CNR's, whereas the

| rule\segm. (CNR = 1) | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| MAP1 | 0 | 0 | 0 | 40 | 60 | 0 |
| MAP2 | 0 | 0 | 0 | 98 | 2 | 0 |
| MAP3 | 0 | 0 | 0 | 98 | 2 | 0 |
| MAP4 | 0 | 0 | 0 | 99 | 1 | 0 |
| AIC | 0 | 0 | 0 | 0 | 18 | 82 |
| MDL | 0 | 0 | 0 | 0 | 21 | 79 |
| rule\segm. (CNR = $\frac{3}{4}$) | 2 | 3 | 4 | 5 | 6 | 7 |
| MAP1 | 0 | 0 | 0 | 45 | 55 | 0 |
| MAP2 | 0 | 0 | 0 | 100 | 0 | 0 |
| MAP3 | 0 | 0 | 0 | 99 | 1 | 0 |
| MAP4 | 0 | 0 | 0 | 99 | 1 | 0 |
| AIC | 0 | 0 | 0 | 0 | 23 | 77 |
| MDL | 0 | 0 | 0 | 0 | 42 | 58 |
| rule\segm. (CNR = 2) | 2 | 3 | 4 | 5 | 6 | 7 |
| MAP1 | 0 | 0 | 0 | 73 | 27 | 0 |
| MAP2 | 0 | 0 | 0 | 99 | 1 | 0 |
| MAP3 | 0 | 0 | 0 | 100 | 0 | 0 |
| MAP4 | 0 | 0 | 0 | 100 | 0 | 0 |
| AIC | 0 | 0 | 0 | 0 | 31 | 69 |
| MDL | 0 | 0 | 0 | 0 | 66 | 34 |

Table 1: Comparison of the MAP, AIC, and MDL rules for image segmentation. The true number of classes is 5.

AIC and the MDL usually tend to overestimate the class number.

## 6. REFERENCES


[1] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. R. Stat. Soc,* ser. B 48, pp. 259-302, 1986.

[2] P. M. Djurić, *Selection of Signal and System Models by Bayesian Predictive Densities,* Ph. D. dissertation, Univ. of Rhode Island, 1990.

[3] P. M. Djurić, "Model Selection Based on Asymptotic Bayes Theory," *Proceeding of 7-th SP Workshop on Statistical Signal & Array Processing* , pp. 7-10, 1994.

[4] B. Efron and R. J. Tibshurani, *An Introduction to the Bootstrap,* New York: Chapman and Hall, 1993.

[5] J. K. Fwu and P. Djurić, "Unsupervised MR Image Segmentation by ICM," *ICASSP,* Atlanta, Georgia, pp. 2235-2238, 1996.

[6] J. K. Fwu and P. Djurić, "Unsupervised Vector Image Segmentation by a Tree Structure – ICM Algorithm," *IEEE Transactions on Medical Imaging,* vol. 15, pp. 871-880, 1996.

[7] J. S. Hjorth, *Computer Intensive Statistical Methods,* Chapman & Hall, NY, 1994.

[8] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data,* Prentice Hall, NJ, 1988.

[9] Z. Kato, J. Zerubia, and M. Berthod, "Unsupervised Adaptive Image Segmentation," *ICASSP,* pp. 2399-2402, 1995.

[10] C. Won and H. Derin, "Unsupervised Segmentation of Noisy and Textured Images Using Markov Random Fields," *CVGIP: Graphical Models and Image Processing,* pp. 308-328, 1992.

[11] J. Zhang and J. W. Modestino, "A Model-Fitting Approach to Cluster Validation with Application to Stochastic Model-Based Image Segmentation," *IEEE PAMI,* vol. 12, pp. 1009-1017, 1990.


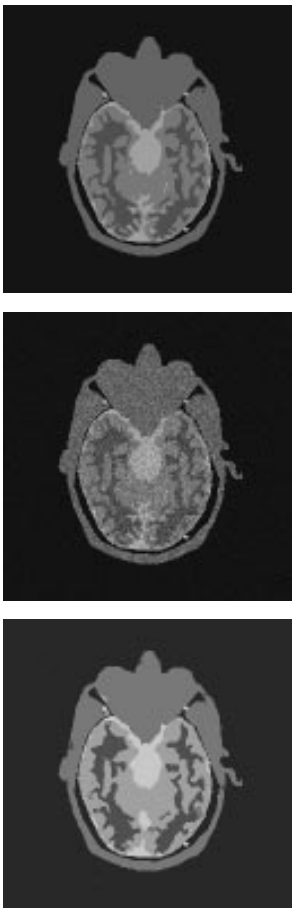

Figure 1: From top to bottom: Noiseless, noisy and segmented MR image.