# Quantization from Bayes Factors with Application to Multilevel Thresholding

F. Murtagh (1) and J.L. Starck (2)
(1) School of Computer Science, Queen's University Belfast, Belfast BT7 1NN
(2) DAPNIA/SEI-SAP, CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France
f.murtagh@qub.ac.uk, jstarck@cea.fr

## ABSTRACT

We are concerned with the optimal selection of multiple thresholds in image analysis. We propose the use of the Bayes information criterion, a minimal information measure, for this and illustrate its use in practical cases. Applications of multiple threshold selection of interest to us include the closely related problems of (i) quantization for lossy encoding, and (ii) segmentation. Our examples relate to segmentation as a post-processing phase in edge detection.

**Keywords:** image thresholding, model selection, Bayes factor, Bayes information criterion, edge detection, wavelet transform.

## 1. INTRODUCTION

Optimal selection of multiple thresholds is a difficult problem for a number of reasons. Firstly, we are often concerned with non-fixed intervals, unlike e.g. Yin (2002) who considers fixed interval thresholding. Secondly, the distribution of the signal and/or noise in our image data is rarely a distribution which is amenable to fixed a priori setting of thresholds. We will now look in somewhat greater depth at these two reasons for considering optimal multiple selection. First we consider optimal quantization with non-fixed thresholds. This will allow us to introduce notation to be used in the section to follow where we will discuss how to choose optimally the number of classes.

Optimal non-uniform quantization is reviewed in Gray and Neuhoff (1998) and Gray (2002). In the univariate case, one-dimensional (scalar) observations $x_i$ are taken, with $n$ pixels or observations in total: $x = \{x_1, \ldots, x_n\}$. A lossy encoder is a classification function $\gamma$ which maps the observations onto a label or index set of class labels or sequence numbers: $\gamma : x \longrightarrow K$. We will write $K = \{1, 2, \ldots, g, \ldots, G\}$. The classification function $\gamma$ is the (initially unknown) $n \times G$ assignment matrix, where $\gamma_{ig} = 1$ if $x_i$ is assigned to the $g$th group, and $\gamma_{ig} = 0$ otherwise.

In non-uniform quantization, with each label we associate a codebook entry, or associated cluster mean. The function defining this, the "reproduction decoder", is $\mu : K \longrightarrow \text{approx}(x)$, where $\text{approx}(x)$ is an approximation or distortion function. For class label $g$, $g \in K$, we write $\mu_g$ as the mean of the $g$th class. Therefore $\text{approx}(x_i) = \mu_g$ when observation $i$ is assigned to class $g$. The approximation or distortion function is usually defined such that the minimal average or expected distortion, $E[d(x, \mu(\gamma(x))]$, is minimized. $\mu(\gamma(x))$ is to be read as: first assign each observation in $x$ to a class which gives us the assignment matrix, $\gamma$; and then determine the mean of the class. I.e. $E[d(x, \mu \mid \gamma)]$, where each term is written $d(x_i, \mu_g \mid \gamma_{ig} = 1)$. The measure $d$ will be taken as the probability of the random variable or observation, $x_i$, given that the class is Gaussian of mean $\mu_g$ and variance $\sigma_g^2$. If the class variance is constant then $d$ is the Euclidean distance.

The Lloyd (1957) quantizer, originally developed by Lukaszewicz and Steinhaus (1955), has the following among its properties: The optimal reproduction decoder $\mu$ is given by $\mu_K = \text{argmin}_\gamma E[d(x, \mu) \mid \gamma]$ I.e., we

minimize the conditional expectation of the distortion between the codebook entries and the input, given that the encoder produced index label set $K$. In the section to follow, this property will reappear as one step, the M step, of the expectation-maximization algorithm.

Our discussion of optimal lossy encoding has not dealt with the optimal value of $G$, the number of classes. In the next section we will return to this.

The second reason why we are studying optimal multiple threshold selection is due to the fact that signal and/or noise is often not distributed as, e.g., a Gaussian.

Let us take the particular case of the wavelet transform, which can be used as a preliminary to segmentation or quantization. Wavelet coefficients have been shown to be long-tailed or of generalized Gaussian distribution (Tsakalides et al., 2000; Buccigrossi and Simoncelli, 1999; Belge et al., 2000; and elsewhere), for a wide class of input data signals. The generalized Gaussian distribution includes the Gaussian and Cauchy as special cases. The generalized Gaussian distribution is difficult to quantize in an analytic way (Tsakalides et al., 2000).

A fortiori, products of wavelet scales are found to be long-tailed. A product of wavelet scales, based on use of a redundant wavelet transform, is the pixelwise product of wavelet coefficients. The wavelet transform, as is well-known, highlights local transitions in data signals. The persistence of large wavelet coefficients across scales gives further evidence of the presence of edges (Xu et al., 1994; Lu et al., 1994; Lee and Kozaitis, 2000). A straightforward way to study wavelet coefficient persistence is using (pixelwise) wavelet products. The probability density of wavelet products for a wide range of data has been shown to be long-tailed, and distributed as a generalized Gaussian or alpha-stable distribution (Sadler and Swami, 1999). See also Murtagh and Starck (2002).

In raising the problem of adaptive quantization of wavelet coefficient product distributions, we are thereby raising the issue of adaptively quantizing long-tailed distributed data.

## 2. OPTIMALLY CHOOSING THE NUMBER OF CLASSES

### 2.1. The Prior: Gaussian Mixture Model

In the previous section, we have defined our observations or pixel values, $x$, the $G$ classes which we seek, and the matrix $\gamma$ of ones and zeros which represents the assignment of observations to classes. This is the univariate normal finite mixture model, or Gaussian mixture model. Mixture model fitting is by now a very common way to cluster data. Our goal is to determine the number of classes, to determine the class assignment of each observation, and to estimate the parameters $\mu_g$ and $\sigma_g$ of each class. The probability density for this model is

$$f(x_i|\theta,\lambda) = \sum_{g=1}^{G} \lambda_g f_g(x_i|\theta_g), \tag{1}$$

where the class parameters are denoted $\theta = (\theta_1, \ldots, \theta_G)$, with $\theta_g = (\mu_g, \sigma_g^2)^T$; $f_g(\cdot|\theta_g)$ is a Gaussian density with mean $\mu_g$ and variance $\sigma_g^2$; and $\lambda = (\lambda_1, \ldots, \lambda_G)$ is a vector of mixture probabilities such that $\lambda_g \geq 0$ $(g = 1, \ldots, G)$ and $\sum_{g=1}^{G} \lambda_g = 1$.

We estimate the parameters by maximum likelihood using the EM (expectation-maximization) algorithm (Dempster et al., 1997; Mclachlan and Krishnan, 1997). The EM algorithm iterates between the E step and the M step. In the E step, the conditional expectation, $\hat{\gamma}$, of $\gamma$ given the data and the current estimates of $\theta$ and $\lambda$ are computed, so that $\hat{\gamma}_{ig}$ is the conditional probability that $x_i$ belongs to the $g$th class. In the M step, conditional maximum likelihood estimators of $\theta$ and $\lambda$ given the current $\hat{\gamma}$ are computed. Although the EM algorithm has some limitations (e.g. it is not guaranteed to converge to a global rather than a local maximum of the likelihood), it is generally efficient and effective for Gaussian clustering problems.

## 2.2. BIC as a Minimum Description Length

We have characterized the mixture model fitting, with a fixed number of classes $G$, as our prior. We denote this overall model, which we have fit to our data, as $M_G$. As seen in the previous subsection, $M_G$ is defined as firstly the class parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_G)$, and secondly the mixture probabilities $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_G)$.

We now wish to investigate one such model versus another, i.e. $M_G$ versus $M_{G'}$ for two choices of numbers of classes, $G$ and $G'$.

The posterior probability of model $M_G$ is

$$f(M_G \mid x) = \frac{f(x \mid M_G) f(M_G)}{\sum_{L=1}^{G\max} f(x \mid M_L) f(M_L)} \tag{2}$$

We can ignore $f(M_G)$ and the influence of $M_L$ if each model is equilikely a priori.

The Bayes factor is the posterior odds of one hypothesis when the prior probabilities of the two hypotheses are equal: $f(x \mid M_G)/f(x \mid M_{G'})$. The term $f(x \mid M_G)$ is the integrated likelihood rather than the maximized likelihood.

The integrated likelihood, $f(x \mid M_G)$, is given by

$$f(x \mid M_G) = \int f(x \mid \theta_G, M_G) f(\theta_G) d\theta_G \tag{3}$$

where $\theta_G$ has now been redefined to be the set of all parameters for model $M_G$ (i.e. including both $\theta$ and $\lambda$ terms for all classes). We have that $f(x \mid \theta_G, M_G)$ is the usual likelihood. Finally $f(\theta_G)$ is the prior, which we will assume as equilikely for all $M_G$.

A good approximation to the integrated likelihood is given in terms of BIC, Bayes information criterion (Schwarz, 1978; Kass and Raftery, 1995):

$$\mathrm{BIC} = 2 \log f(x \mid \hat{\theta}_G, M_G) - N \log(\dim(\theta_G)) \tag{4}$$

where $\hat{\theta}_K$ is the maximum likelihood estimator of $\theta_K$, i.e. the result of the Gaussian mixture fitting. $N$ is the dimensionality of the observation vectors, and $\dim(\theta_G)$ is the cardinality of the parameter set.

Finally the Bayes factor is approximated by the difference of BIC terms, which in turn are the maximized likelihood results of Gaussian model fits for different numbers of classes, $G$ and $G'$:

$$2 \log \frac{f(x \mid M_G)}{f(x \mid M_{G'})} \approx \mathrm{BIC}(G) - \mathrm{BIC}(G') \tag{5}$$

In operation, a plot of BIC for increasing numbers of classes, $G$, generally shows increase to an approximate plateau. We can usually increase the model fit indefinitely by increasing $G$. It is usual to consider the first peak in this plot, or the effective reaching of the plateau, to provide the optimal value of $G$.

We can also derive the BIC term as a parsimony or minimum information measure (Hansen and Yu, 2001; Rissanen, 1986).

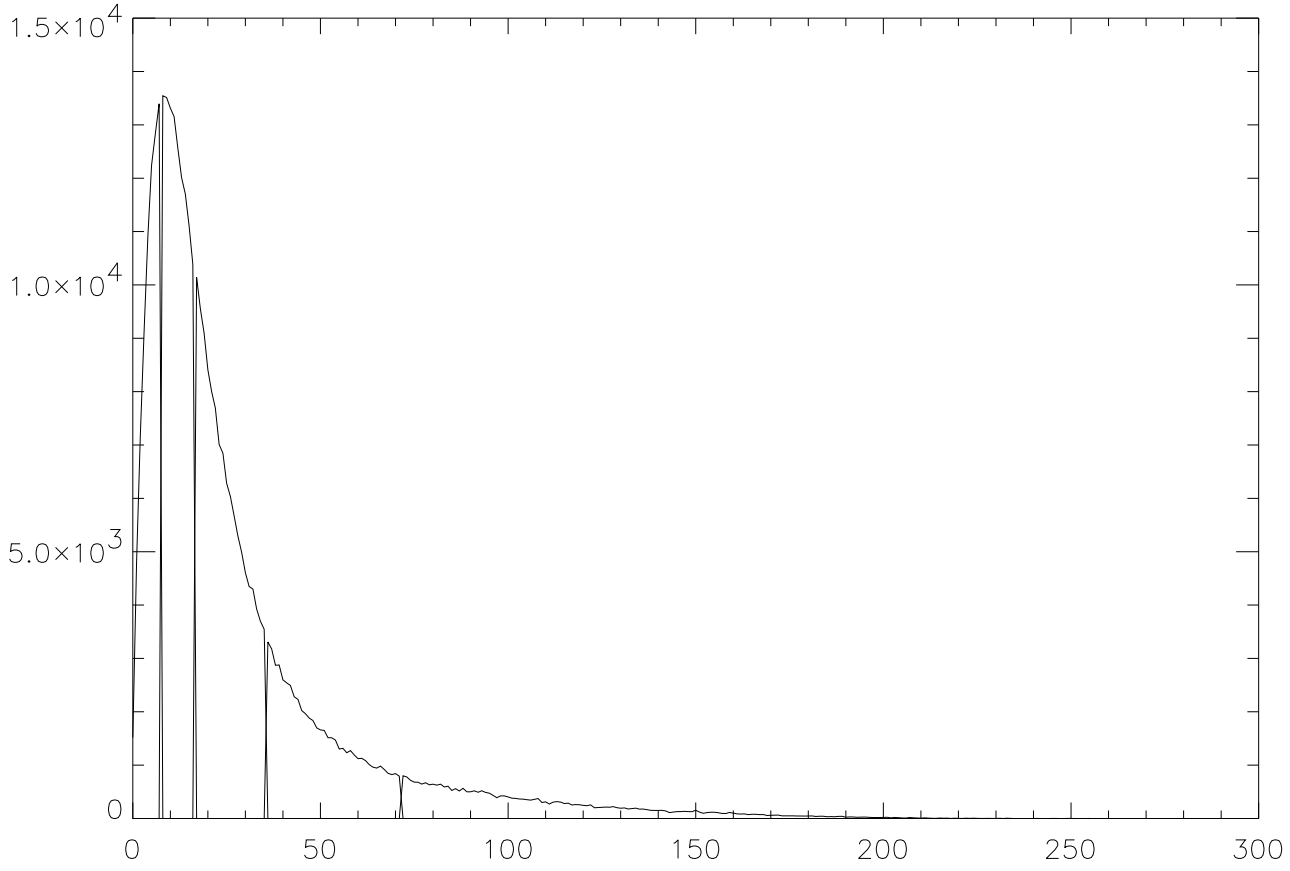


**Figure 1.** Histogram of Canny filtered data, showing clearly long-tailed behavior. Shown also are the classification regions mapped out by the 5 Gaussian mixture components which were fit to this distribution.

## 3. APPLICATION

A test data set from Kominek (2000) was used, the red component of a color $768 \times 512$ image.

Our first set of results is based on use of the Canny (1986) derivative of Gaussian edge detector, $\frac{\partial(g*f)}{\partial x} + \frac{\partial(g*f)}{\partial y}$ with $g = \exp -\frac{x^2+y^2}{2\sigma^2}$, with scale parameter $\sigma = 1/\sqrt{3}$, and where $*$ is convolution of image $f$. Visually, long-tailed behavior in the marginal density can be noted. In a second set of results, also with the objective of edge finding, we use wavelet correlation.

Fig. 1 shows the histogram, and the classification boundaries or thresholds resulting from a 5-component Gaussian fit. Fig. 2 explains why we selected a 5-component fit: a plateau is reached for this number of components. For reference, Fig. 3 shows the 2-component result. Fig. 4 shows the 5-component result, with all components represented. Fig. 5 shows the 5-component result with only the component of highest mean shown.

With reference to Fig. 1, Fig. 5 shows only the tail of the distribution. The threshold used in Fig. 5 is defined by the separation line between the fourth and fifth components. This threshold resulted from the use of the BIC minimum information principle.

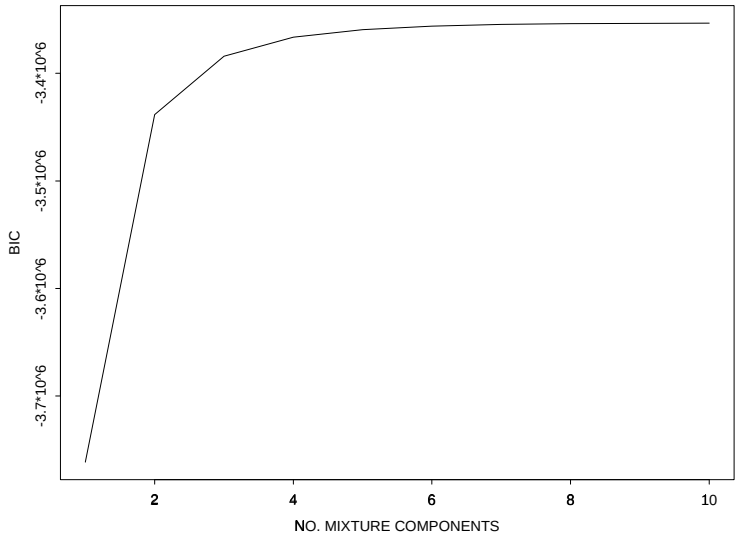


**Figure 2.** Plot of BIC values.

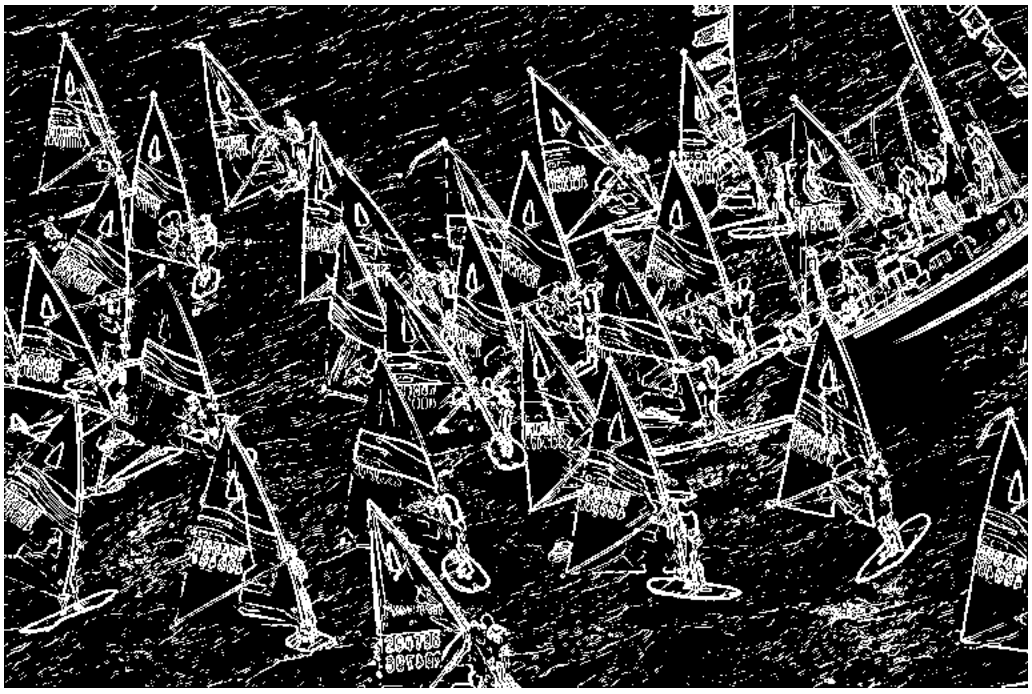

**Figure 3.** Mixture model fit, with 2 components, for result of Canny filtering.

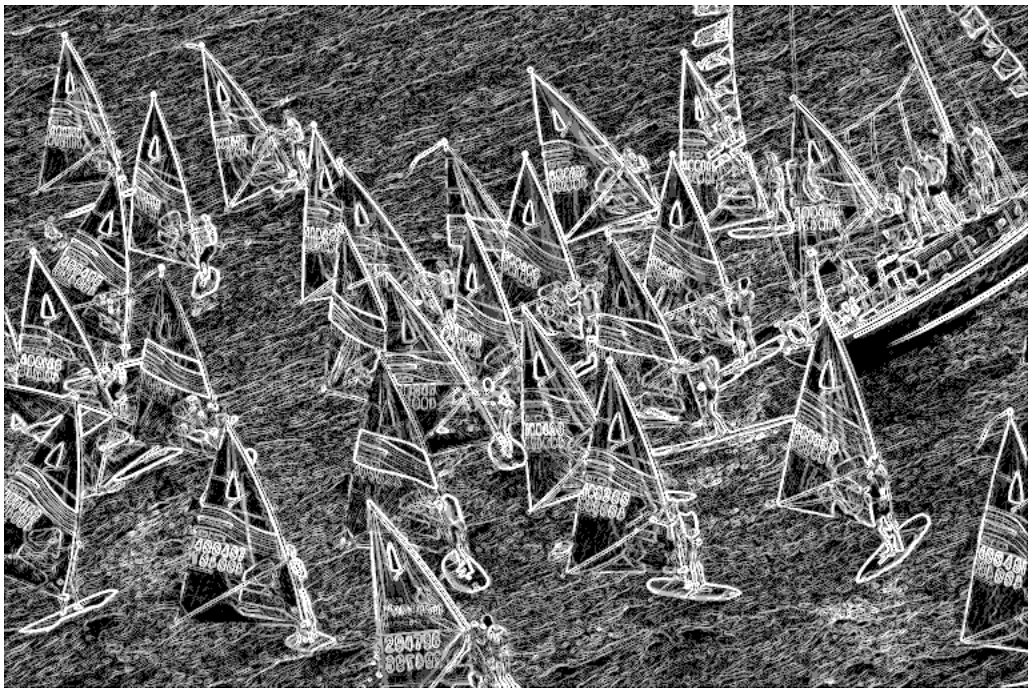

**Figure 4.** Mixture model fit, with 5 components, for result of Canny filtering.

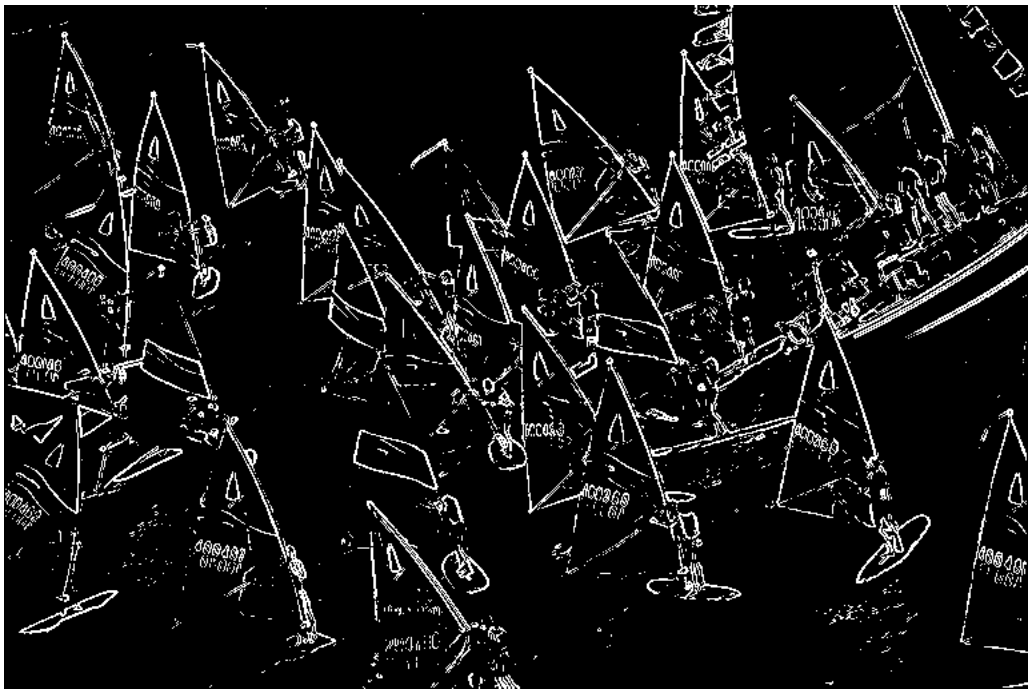

**Figure 5.** Mixture model fit, with one component of largest mean, for result of Canny filtering.

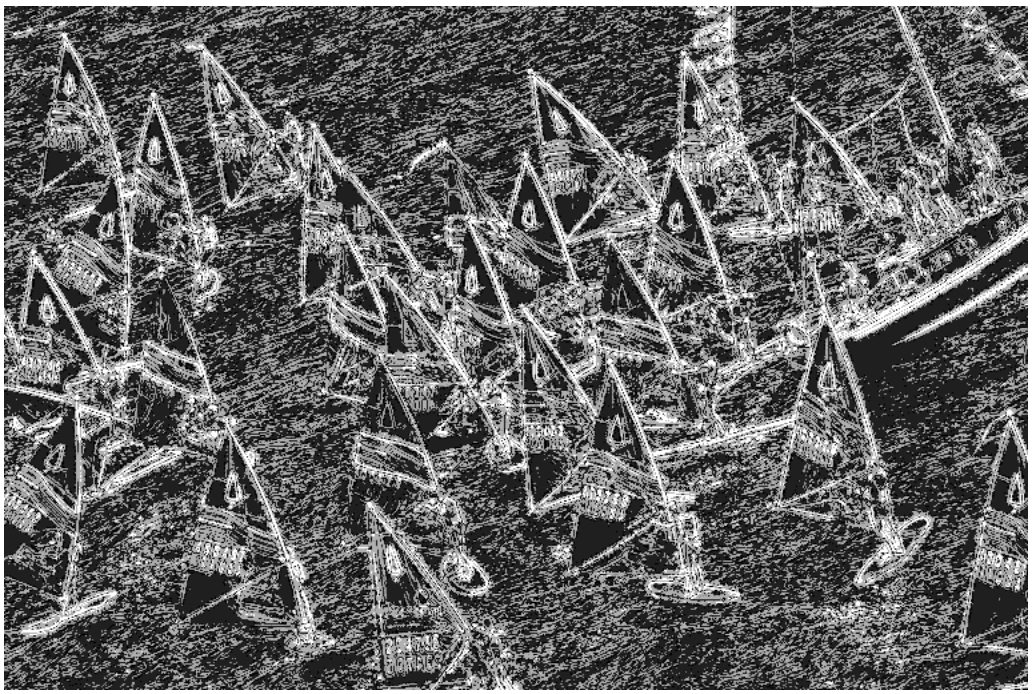

**Figure 6.** Mixture model fit, with 5 components, for product of first and second wavelet scales.

The product of wavelet scales has been proposed as an edge detector, in that persistence of large wavelet coefficients across scales increases our confidence of the presence of edges (Murtagh and Starck, 2002). Here we use a redundant transform, the à trous $B_3$ spline wavelet transform (Starck et al., 1998) which is shift invariant, and also allows a pixel's information at varying resolution scales to be read off very straightforwardly. Consider an image $x$, which in the à trous $B_3$ spline wavelet transform can be decomposed additively as $x = w_S + \sum_j w_j$ where each set of wavelet coefficients $w_j$ constitutes an image of the same dimensions as $x$, where $j$ is the number of resolution scales in use, and finally where $w_S$ is the last smooth version of the data (or DC, "direct current", component). The product of wavelet scales 1 and 2, which we used, is then the pixelwise product $w_1.w_2$.

Fig. 6 shows the result following a 5-component mixture model fit, which gives a fully adequate indication of the wavelet product itself. Wavelet scales 1 (i.e., the highest frequency scale) and 2 were used in this pixelwise product. Why we selected five components is explained in Fig. 7: essentially a plateau is reached at this number. Fig. 8 shows the somewhat poorer 2-component fit. Fig. 9 shows the class assignment corresponding to the tail of the wavelet product density. It corresponds to the class with the highest mean.

In conclusion, the procedure for selecting the optimal number of threshold components has been applied to the gradient maps provided by Canny and by the wavelet product operations.

Table 1 summarizes what we expect from these results: the top class from the 5-class fit has a high mean value. In addition, as is visually clear from the figures, it is more concentrated in regard to class variance. The issue to be stressed here is that both approaches provide a sound basis for a necessary post-processing phase when determining edges.

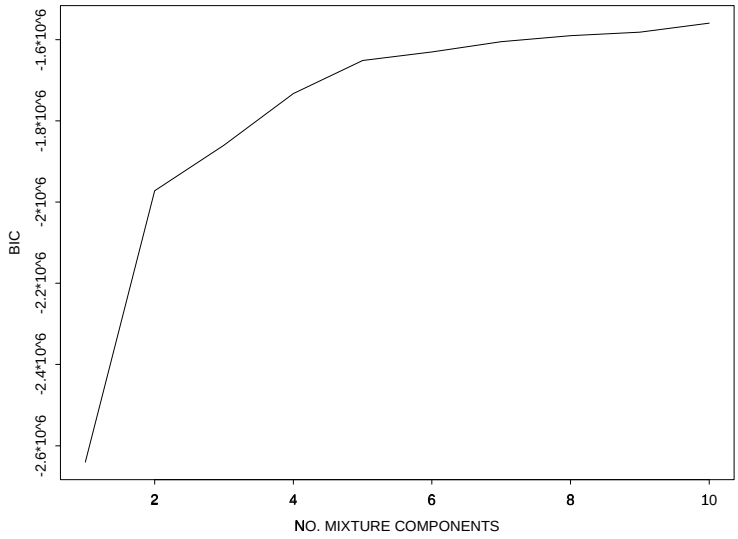


**Figure 7.** Plot of BIC values for mixture model fits to product of first and second wavelet scales.

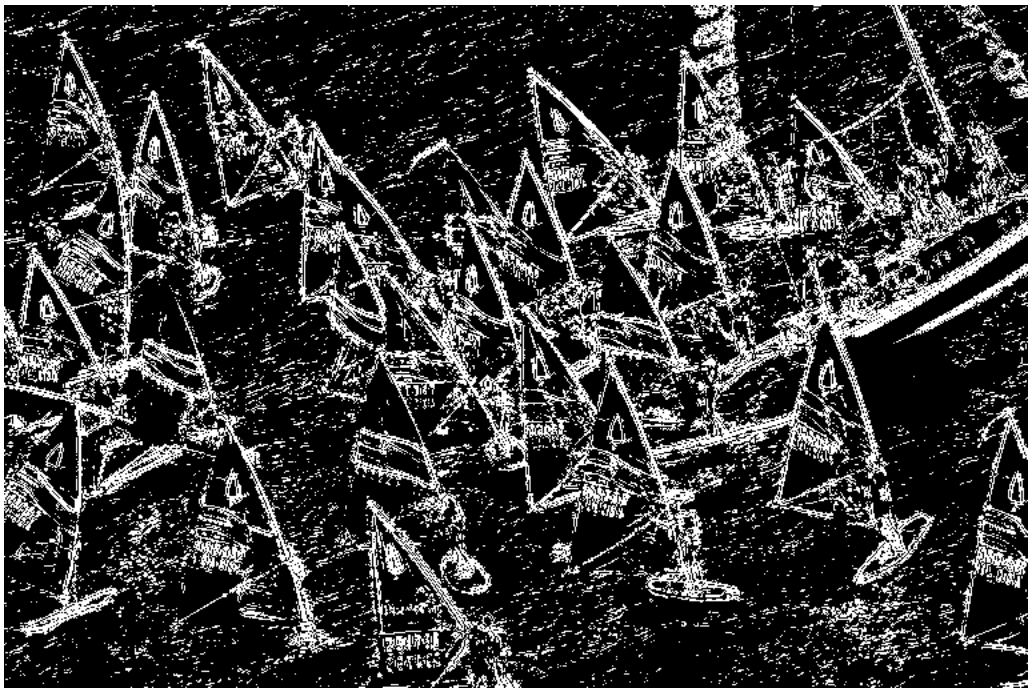

**Figure 8.** Mixture model fit, with 2 components, on product of first and second wavelet scales.

**Figure 9.** Mixture model fit, with one component of largest mean from a 5-component fit, on product of first and second wavelet scales.

| | mean | standard deviation |
|---|---|---|
| Canny result | | |
| 2-class fit | 117.29 | 54.36 |
| Top class of 5-class fit | 126.64 | 36.39 |
| Wavelet product result | | |
| 2-class fit | 117.72 | 54.22 |
| Top class of 5-class fit | 131.10 | 36.40 |

**Table 1.** Comparison of, respectively from top to bottom, Figs. 3, 5, and 8, 9. The comparison is based on class label and, for the two edge finding approaches, the same input image.

## 4. DISCUSSION AND CONCLUSION

In Chipman et al. (1997), Choi and Baraniuk (2001), and elsewhere, a prior model is defined for wavelet coefficients which takes them as variates from an additive mixture of two Gaussians. One of these mixture components (let us call it L) has large wavelet coefficients and large variance, and the other component (we will call it S) has small wavelet coefficients and small variance. As an optimal model fit, the {L, S} solution defines simultaneously an optimal bilevel threshold. It is not unduly surprising that we have found a multilevel set of thresholds, defined from a mixture model with more than two components, to provide a better approximation. What is a lot more interesting is that a Bayes factor criterion gives us a performance metric for choosing the optimal mixture, and thereby the optimal set of thresholds.

Two differences between our work and hierarchical models as pursued by Choi and Baraniuk (2001) deserve to be noted. We do not explicitly consider correlations between neighboring coefficiences, but the redundant wavelet transform does implicitly take correlation into account. Secondly, we do not consider persistence of wavelet coefficient values across more than two scales, although we could do so in a way similar to Xu et al. (1994).

Coates and Kuruoğlu (2002) consider the different problem of signal source detection in long-tailed noise, but similar to this article they fit a Gaussian mixture model to the data. In the problem of long-tailed distribution parameter estimation, Swami and Sadler (2002) focus on the limits of Gaussian mixture model fitting.

The edge detectors used for illustrative purposes in this article, viz. the Canny detector and the product of wavelet planes resulting from a redundant wavelet transform, are powerful approaches. Nonetheless it is clear that post-processing is needed to define edge presence and edge absence. This was the application area which we selected in order to exemplify optimal specification of multiple thresholds.

We took a Canny edge detector and a wavelet product, both to provide localization information on edges in the image. In the first case we showed long-tailed behavior in the marginal density, and in the case of wavelet products we cited where such a distribution has been investigated in the literature. Setting ourselves the task of variable-width multiple thresholding, we showed how a Gaussian mixture fit to the marginal density permitted such a result to be derived.

Next came the question of the optimality of such a result. Model fitting gives a maximum likelihood solution. We used the ratio of integrated likelihoods of one model against another, or in other words the differences of log likelihoods, for model selection. In this work, our objective in using this procedure was to choose objectively the number of mixture components.

From different perspectives, this procedure can be viewed in terms of Bayes classification, minimum information, maximum likelihood, and minimum description length. The approach described here is a powerful one, both in terms of supporting theory and in terms of practicality and ease of deployment.


## Acknowledgements

We acknowledge the helpful comments of referees on an earlier version of this paper.


## References


1. Belge, M., Miller, E. and Kilmer, M., 2000. Wavelet domain image restoration with adaptive edge-preserving regularization, IEEE Trans. Image Proc., 9, 598–608.

2. Buccigrossi, R.W. and Simoncelli, E.P., 1999. Image compression via joint statistical characterization in the wavelet domain, IEEE Trans. Image Proc. 8, 1688–1701.

3. Canny, J., 1986. Computational approach to edge detection, IEEE Trans. Pattern Analysis and Machine Intelligence 8, 679–698.

4. Yan Chang, Fu, A.M.N., Hong Yan and Mansuo Zhao, 2002. Efficient two-level image thresholding method based on Bayesian formulation and the maximum entropy principle, Optical Engineering 41, 2487–2498.

5. Chipman, H., Kolaczyk, E. and McCulloch, R., 1997. Adaptive Bayesian wavelet shrinkage, J. Amer. Statist. Assoc. 92, 1413–1421.

6. Choi, H. and Baraniuk, R.G., 2001. Multiscale image segmentation using wavelet-domain hidden Markov models, IEEE Trans. Image Proc. 10, 1309–1320.

7. Coates, M.J. and Kuruoğlu, 2002. Time-frequency-based detection in impulsive noise environments using $\alpha$-stable noise models, Signal Proc. 82, 1917–1925.

8. Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm, J. Royal Statist. Soc., Series B 39, 1–22.

9. Gray, R.M., and Neuhoff, D.L., 1998. Quantization, IEEE Trans. Information Theory 44, 2325–2384.

10. Gray, R.M., 2002. Gauss mixtures quantization: clustering Gauss mixtures, in Nonlinear Methods in Estimation and Classification, Springer-Verlag, forthcoming.

11. Hansen, M.H. and Bin Yu, 2001. Model selection and the principle of minimum description length, J. American Statist. Assoc. 96, 746–774.

12. Kass, R.E. and Raftery, A.E., 1995. Bayes factors. J. American Statist. Assoc. 90, 773–795.

13. Kominek, J., Waterloo BragZone, 2000,
http://links.uwaterloo.ca/bragzone.base.html
Images: ftp://links.uwaterloo.ca/pub/BragZone/ColorSet

14. Lee, Y. and Kozaitis, S.P., 2000. Multiresolution gradient-based edge detection in noisy images using wavelet domain filters, Optical Engineering 39, 2405–2412.

15. Lloyd, S.P., 1957. Least squares quantization in PCM, Bell Labs technical note, partly presented at Institute of Mathematical Statistics Meeting, Atlantic City, NJ, Sept. 1957. Published in IEEE Trans. Information Theory 28, 129–137, 1982.

16. Lu, J., Healy, D.M. and Weaver, J.B., 1994. Contrast enhancement of medical images using multiscale edge representation, Optical Engineering 33, 2151–2161.

17. Lukaszewicz, J. and Steinhaus, H., 1955. On measuring by comparison, Zastos. Mat. 2, 225–231 (in Polish).

18. McLachlan, G. and Krishnan, T., 1997. The EM Algorithm and Extensions, Wiley.

19. Murtagh, F. and Starck, J.L., 2002. Bayes factors for edge detection from wavelet product spaces, Optical Engineering, in press.

20. Rissanen, J., 1986. Stochastic complexity and modeling, Annals of Stastics 14, 1080–1100.

21. Sadler, B.M. and Swami, A., 1999. Analysis of multiscale products for step detection and estimation, IEEE Trans. Information Theory 45, 1043-1051.

22. Schwarz, G., 1978. Estimating the dimension of a model, Annals of Statistics 6, 461–464.

23. Starck, J.L., Murtagh, F. and Bijaoui, A., 1998. Image and Data Analysis: The Multiscale Approach, Cambridge University Press.

24. Swami, A. and Sadler, B.M., 2002. On some detection and estimation problems in heavy-tailed noise, Sigal Proc. 82, 1829–1846.

25. Tsakalides, P., Reveliotis, P. and Nikias, C.L., 2000. Scalar quantisation of heavy-tailed signals, IEE Vision, Image and Signal Proc. 147, 475–484.

26. Xu, Y., Weaver, J.B., Healy, D.M., and Liu, J., 1994. Wavelet transform domain filters: a spatially selective noise filtration technique, IEEE Trans. Image Proc. 3, 747–758.

27. Peng-Yeng Yin, 2002. Maximum entropy-based optimal threshold selection using deterministic reinforcement learning with controlled randomization, Signal Proc., 82, 993–1006.