# Histogram Thresholding using Kernel Density Estimates

Joakim Lindblad
Centre for Image Analysis, Uppsala University
Lägerhyddv. 17, 752 37 Uppsala
E-mail: joakim@cb.uu.se

## Abstract

A non-parametric method for thresholding of data which makes very weak assumptions about the underlying distribution is presented. The output consists of one or more recommended threshold points, along with a scalar measurement of significance which enables for easy comparison of different thresholds. The suggested thresholds are given as numbers on the real line, and are not restricted to integer values or anything similar to bin-size of an ordinary histogram. The method has been shown to work well in different situations, including cases where it is hard to find a good threshold by visual inspection.

## Key words

kernel density estimation, non-parametric histogram thresholding, bandwidth selection, second derivative

## 1 Introduction

Thresholding of histograms is one of the first things a student in image analysis gets acquainted with and it remains one of the most common operations performed in image analysis, also on higher levels of research. Thresholding may be used in many areas to separate different parts of a data set. It can be done in the form of a direct thresholding of image data, or used for object-wise classification on measured features, or perhaps hidden inside other methods, as for example is done in the Canny edge detector [1].

If we have spatial relations or shading in an image, it is rarely a good choice to directly apply a global threshold. But with the right pre-processing and shading correction, histogram based thresholding may in fact show up to be one of the best methods available. Also there is nothing keeping us from using local histogram thresholding, or thresholding on secondary features which take in account the spatial relations that may exist. Add to this the appealing simplicity of histogram thresholding and the good understanding of the situation that follows with that, and one easily realises that histogram based thresholding is certainly something which is worth having a look at, and so has also been done for very many years, not only in the field of image analysis.

## 2 Taking a histogram

Histogram thresholding can be, as is indicated by the name, separated into two parts, taking a histogram and then use that to select a suitable threshold. Let us start with the first part.

Taking a histogram of a data set is essentially a method to estimate the distribution of data by plotting the number of samples that exist in a set of intervals. Commonly the set of intervals are selected as an equal sized partitioning of $\mathbb{R}$ (or $\mathbb{Z}$). An obvious problem at this point is, how to select the partitioning size, i.e. the bin-size of the histogram. This is really a problem of great importance, as the choice of bin-size greatly affects the way the histogram looks. If we choose the bin-size too small, the histogram becomes wiggly and noisy. If we on the other hand select a too large bin-size, it will hide features of the underlying distribution, it may also give bad accuracy of the threshold, as this usually can only be decided down to the precision of a single bin. In addition, using large bins hides the variation that really takes place inside each bin. A data point near the left edge of a bin gives exactly the same contribution to the histogram as one near the right edge.

If the input data is quantised, e.g. integer valued, it may feel natural to select the bin-size to be equal to the quantisation step, but this is really just a lower limit of the bin-size, and may be very far from optimal.

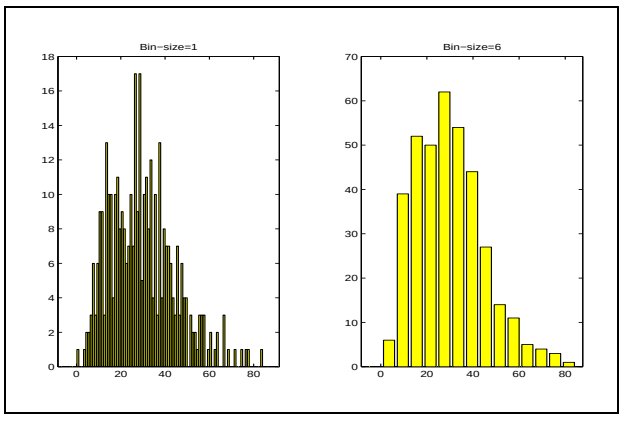


Figure 1: Too small bin-size makes the histogram noisy and hard to interpret (left). Too large bin-size hides features of the underlying distribution (right).

## 3 Thresholding of histograms

Probably the most common method to threshold a histogram is by visual inspection. This is, of course, not to be considered as a very objective method, and it varies a lot between individuals. It is also non-automatic and therefore both slow and expensive. In addition, visual inspection is sensitive to the aspect ratio of the histogram. Re-scaling the axes of the histogram should of course not change the outcome of a thresholding algorithm, and care should be taken when designing automatic methods to avoid such an undesired dependency.

A multitude of methods has been invented to automatically find thresholds in a histogram. Some methods rely on taking percentiles of modelled distributions, others on finding local minima in the histogram or smoothed versions of it, and some use direct geometrical relations, like e.g. the triangle algorithm [7]. Usually, these methods incorporate a priori knowledge about the shape and modality of the distribution to be analysed. While the triangle algorithm relies on unimodality, the minimum finding methods must have multimodality, and if we cut at percentiles we certainly make strong assumptions about the underlying distribution.

To make such assumptions about the underlying distribution is always dangerous. What happens e.g. if we make the wrong assumptions about symmetries, smoothness, modality, etc? It would be nice if we could find some general all-purpose method that works well regardless of the modality of the histogram, and which only makes weak assumptions about the underlying distribution.

As a good starting point, one could ask the somewhat philosophical question: What is a good threshold point really? It makes sense to say that a good threshold point is a point where there is a change between different distributions. If we do not want to model the distribution directly (to avoid making possibly false assumptions about their shape), we can see this change between different distributions as a point where we have a change of "trends". Here, the word trend should be interpreted as "the way things are heading", as it e.g. is used on the stock market. "The way things are heading" is in the one-dimensional case represented by a derivative. The rate, at which the trends change, is correspondingly expressed by the second derivative. Concluding this reasoning, we come to the suggestion that a good threshold point should be located at an extreme value in the second derivative of our histogram.

Here we trip right into another problem regarding the use of histograms to find suitable threshold values. It is not at all trivial how to calculate derivatives of a discontinuous histogram, preferably in such a way that it gives stable and reliable results and without introducing "magic numbers" into the theory.

## 4 Kernel Density Estimates

Having run into all this trouble regarding histograms, one naturally ends up asking the question: Does there not exist some more elaborate theory about how to make an estimate of a probability density function?

The answer to that question is certainly yes. To estimate the probability density functions from a random sample is a basic problem in several domains of applied science such as pattern recognition, function approximation, machine learning, econometrics, etc. It is also one of the more fundamental problems of the statistical sciences, and consequently there has been a lot of work done in this field.

The basic question is, how one can estimate a probability density function $f(x)$ given a sequence of independent identically distributed random variables $X_1, ..., X_n$ from this density $f$.

There exist a lot of parametric approaches to this problem, e.g. mixture modelling, where one tries to fit a mixture of standard distributions to sum up to the unknown distribution $f$. As already touched upon, parametric modelling has some unappealing properties. What happens e.g. if the unknown distribution is not at all a mixture of our model distributions?

Not willing to take the risk of making wrong assumptions, the non-parametric approaches are often more popular, and by now a rich basket of non-parametric density estimators (kernel, spline, ML and orthogonal series) exists. As the length of this paper is limited, we will simply focus on one of the most popular methods, namely the well studied class of kernel density estimators (KDE) (which also includes the standard histogram) as introduced by Rosenblatt [4] and Parzen [3]

These class of estimators $\hat{f}$ are defined by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \qquad (1)$$

where $h$ is called the bandwidth and $K$ is a kernel, $K_h(u) = K(u/h)/h$.

Common choices of the kernel include the uniform, triangular and the Gaussian kernel functions, as well as more advanced ones. If we choose the uniform kernel, and $h$ equal to the bin-size, we are back at our well known histogram case.

It is easily verified that the optimal bandwidth should decrease with increasing number of samples $n$, and the class of kernel density estimates has been shown to converge asymptotically to the true distribution $f(x)$ in the $L^2$ metric, if the bandwidth $h$ is chosen such that $h \to 0, nh \to \infty$ as $n \to \infty$ [3].

The choice of the kernel $K$ is not critical for the asymptotic convergence. However, it does influence the smoothness of the estimate, and, as we may wish to analyse extreme values and inflection points of the estimated distribution, it is clearly a good thing to choose a smooth kernel. The Gaussian is one of the more popular choices, due to its nice properties and

statistically sane foundation, and it will also be the choice of this study.

It turns out that the choice of the bandwidth $h$, is much more important for the behaviour of $\hat{f}$ than the choice of $K$. Small values of $h$ make the estimate look "wiggly" and show spurious features due to the sampling, whereas too big values of $h$ will lead to an estimate which is too smooth in the sense that it may hide structural features, like e.g. bimodality, of the underlying density $f$ (see. figure 2).

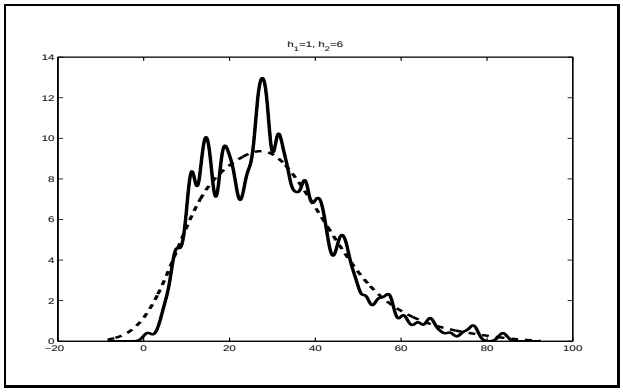


Figure 2: Too small bandwidth (solid line) makes the KDE noisy and hard to interpret. Too large bandwidth (dashed) hides underlying features.

So in some sense, by moving from a standard histogram to a KDE, we have just moved from a bin-size problem to a bandwidth problem. Compared to the histogram case though, by using a better suited kernel than the uniform, we do neither loose in threshold precision, nor miss the information of inside bin positioning, when we choose a larger bandwidth.

# 5 Bandwidth selection

Nevertheless, the bandwidth selection is a crucial step. Unfortunately, a satisfactory theory for kernel width selection is not available for fixed sample sizes. This is not surprising, since the kernel width must be smaller than the sharpest "detail" in the unknown target density, in order to prevent undesirable "blurring" in the approximation. At the same time, it should be as large as possible to filter out the random variability of the sample.

A lot of research has been done in the last years to find objective data-driven bandwidth selection methods and plenty of different techniques have been developed [6].

Without going too much into detail, we will settle for one of the more popular choices, which is to derive an optimal global bandwidth by minimising the asymptotic mean of the integrated squared error (AMISE). It does not really make sense to include a complete mathematical derivation here, and I will jump directly to the optimal choice of bandwidth $h$ for a Gaussian kernel.

$$h = \left( \frac{\int_{\mathbb{R}} K^2(x)\,dx}{(\int_{\mathbb{R}} x^2 K(x)\,dx)^2 \int_{\mathbb{R}} f''^2(x)\,dx} \right)^{1/5} n^{-\frac{1}{5}} \quad (2)$$

Here the term $\int_{\mathbb{R}} f''^2(x)\,dx$ has to be estimated. The rule of thumb in this case is to replaces the unknown density function $f$ in this functional by a reference distribution function which is re-scaled to have variance equal to the sample variance. Taking $K$ as the Gaussian kernel and the standard normal distribution as reference function, this rule of thumb yields the estimate

$$\hat{h}_{rot} = 1.06 \hat{\sigma} n^{-\frac{1}{5}} \quad (3)$$

where $\hat{\sigma}^2$ is the sample variance.

A version more robust against outliers can be constructed if the interquartile range $R$ is used as a measure of spread instead of the variance. This gives us the following modified estimator

$$\hat{h}_{rot} = 1.06\, min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-\frac{1}{5}} \quad (4)$$

which is a commonly used bandwidth selection [5] and will also be the choice of this paper.

Note however that there exist many other bandwidth selection methods, e.g. cross-validation and plug-in methods. See e.g. Turlach [6] for a good overview.

In our computerised world, data is not always continuous, but rather a digital sample of a continuous world. For histograms we noticed that we cannot use a bin-size which is smaller than the quantisation step. This holds also for the KDE, and we are not allowed to select a bandwidth in the range of or smaller than the sampling resolution. E.g. if we have a data resolution of 8 bits, all values will be integers between 0 and 255. If we model this situation with a bandwidth smaller than one, we will experience dips in the KDE at the real values in-between these integers. This is, of course, not resembling the real situation, and is an artifact due to the sampling. We can reduce the effect of the quantisation either by selecting a larger bandwidth, or by performing some de-convolution or smoothing before making the KDE.

# 6 Conclusions

The KDE is obviously a better estimator of a sampled distribution than an ordinary histogram, and we have some theory on how to select a suitable bandwidth. What about the problem of derivation then? Yes, the KDE solves that problem too. A nice property of the way a KDE is expressed, is that it is very easy to calculate derivatives. As it consists of a sum of translated versions of the kernel function, calculating derivatives becomes just to sum the derivatives of the

kernel function instead.

$$\hat{f}_h^{(p)}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h^{(p)}(x - X_i) \qquad (5)$$

Note: The fact that we will use the second derivative of the estimate $\hat{f}$ should really affect the choice of the bandwidth. This has not been further investigated in this study.

Now that all problems are solved, it seems time to conclude: It makes good sense to use the maxima of the second derivative of a KDE of a data set as possible good threshold values for this data set. The magnitude of the second derivative suggest which maximum is most likely to represent a significant and useful threshold.

However, in most cases, a priori knowledge must be applied to select the most appropriate of a small set of prominent maxima. E.g. we usually do not want to put a threshold to the left of a dark background peak in a pixel distribution, despite the fact that this is an obvious change of trend point, where we are going from the case of no data at all, to the case of a large background signal. A priori knowledge also rules out the minima of the second derivative, as they, in most cases represent the place were we have a peak in the distribution. (This is, however, the place where the stock broker usually wants to put a threshold, commonly called "Time to sell!")

# 7 Performance

The suggested method has shown to be fairly robust, and performs well on both noisy and clean data, performing (due to the choice of a global bandwidth) somewhat better on fairly normal distributed data. There is not at all place in the assigned four pages to include a representative test suit, so you will have to be satisfied with just a few examples.

The reason this paper came to be in the first place was the need of segmenting cells according to their cyclein content using fluorescence microscopy [2]. When a first approach using fixed percentiles of a modelled Gaussian failed, the above described method was applied, with very promising results. In a total of 45 different data sets it worked very well, and in only one case visual inspection suggested that the second largest peak of the second derivative should be used instead of the largest one. Below are two fairly representative cases shown.


# Acknowledgements

Thanks to Fredrik Erlandsson, Carolina Linnman, Ewert Bengtsson and Felix for data, help and ideas.


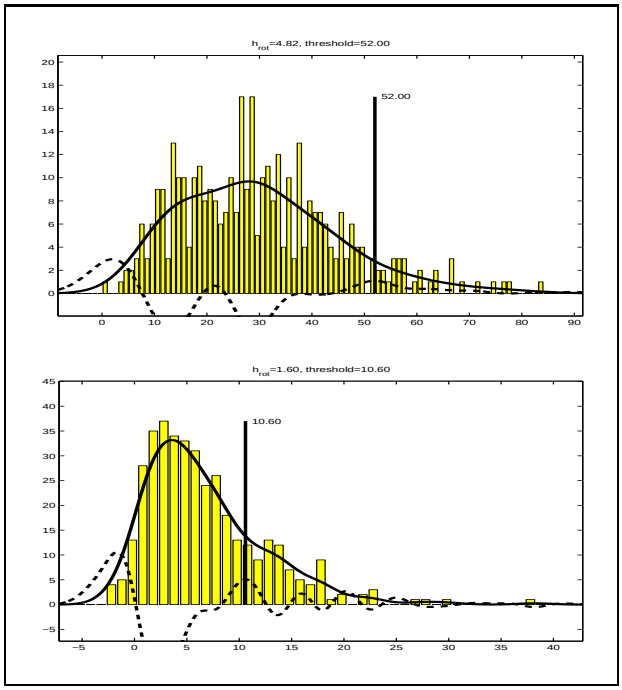


Figure 3: Background histogram is number of cells plotted against average internal fluorescence intensity. Solid line is the KDE $\hat{f}$ and the dashed line is the second derivative $\hat{f}''$ (scaled). Suggested threshold is taken to be the largest peak in the second derivative, that is to the right of the main peak in the KDE.

# References


[1] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. and Machine Intelligance*, 8:679–698, 1986.

[2] F. Erlandsson, C. Linnman, S. Ekholm, E. Bengtsson, and A. Zetterberg. A detailed analysis of cyclin A accumulation at the G1/S border in normal and transformed cells. *submitted to Experimental Cell Research*, 2000.

[3] E. Parzen. On estimation of probability and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[4] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:642–669, 1956.

[5] B. Silverman. *Density estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[6] B. A. Turlach. Bandwidth selection in kernel density estimation: a review. Discussion papers series, Institute of Statistics, Louvain-la-Neuve, Belgium, 1993.

[7] G. Zack, W. Rogers, and S. Latt. Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry and Cytochemistry*, 25(7):741–753, 1977.