# H-MEANS IMAGE SEGMENTATION TO IDENTIFY SOLAR THERMAL FEATURES

*Nathan Stein,*[1] *Vinay Kashyap,*[2] *Xiao-Li Meng,*[1] *David van Dyk*[3]

[1] Department of Statistics, Harvard University, Cambridge MA 02138 USA
[2] Harvard-Smithsonian Center for Astrophysics
[3] Statistics Section, Imperial College London

## ABSTRACT

Properly segmenting multiband images of the Sun by their thermal properties will help determine the thermal structure of the solar corona. However, off-the-shelf segmentation algorithms are typically inappropriate because temperature information is captured by the relative intensities in different passbands, while the absolute levels are not relevant. Input features are therefore pixel-wise proportions of photons observed in each band. To segment solar images based on these proportions, we use a modification of $k$-means clustering that we call the $H$-means algorithm because it uses the Hellinger distance to compare probability vectors. $H$-means has a closed-form expression for cluster centroids, so computation is as fast as $k$-means. Tempering the input probability vectors reveals a broader class of $H$-means algorithms which include spherical $k$-means clustering. More generally, $H$-means can be used anytime the input feature is a probabilistic distribution, and hence is useful beyond image segmentation applications.

***Index Terms***— Astronomy, Astrophysics, Sun, Clustering algorithms, Image segmentation

## 1. INTRODUCTION

The thermal structure of the solar corona is a crucial factor in a proper understanding of the Sun-Earth connection. It places constraints on coronal heating mechanisms, controls the characteristics of the solar wind, and has the potential to predict flare onsets and mass ejections. Determining the thermal structure is however difficult since the data are obtained as images at high cadence in multiple filters. Hence, fast image segmentation methods that segregate regions of similar thermal properties are needed.

Image segmentation is a well-studied problem, with applications in fields such as medical imaging and computer vision; see [1], [2], [3], [4], [5] for examples of recent work. However, for studying the thermal structure of the solar corona, standard segmentation approaches suffer two disadvantages. First, off-the-shelf segmentation algorithms often assume that the absolute intensities in each band of multiband images are the natural input features to be compared, which is not the case in the thermal structure problem. Second, model-based approaches that incorporate more complicated observation models or input features often suffer from computation requirements that prevent their use on streams of high-resolution images. Fast computation is necessary when analyzing data from the Atmospheric Imaging Assembly (AIA) of the Solar Dynamics Observatory, which captures high resolution multiband images with a cadence of approximately ten seconds.

In the corona, absolute brightness roughly corresponds to the amount of plasma in a particular region, which is not relevant for studying thermal properties. Thermal properties such as average temperature are instead captured by the relative intensities in the different passbands. Motivated by a probabilistic model of the data generating process, we cluster pixels based on the vectors of proportions of photons observed in each passband. To cluster these probability vectors, we use a modified $k$-means algorithm that replaces Euclidean distance with Hellinger distance, inspiring the name $H$-means.

In Section 2, we describe the $H$-means algorithm and discuss tempering the input distributions. In Section 3, we describe our application and the statistical model that inspires our approach. In Section 4, we illustrate the performance of our algorithm on AIA images. Section 5 concludes.

## 2. $H$-MEANS CLUSTERING

Clustering algorithms based on distances other than Euclidean distance have been proposed and studied in, for instance, [6], [7], [8], [9], [10], and [11]. The Hellinger distance has been suggested as an alternative to Euclidean distance in clustering and dimension reduction problems in [12], [13], [14], and [11], but there has been relatively little study of substituting the Hellinger distance for Euclidean distance in $k$-means clustering. We call this modification to $k$-means the *$H$-means algorithm.*

Professor van Dyk was supported in part by NSF grant DMS-09- 07522 and by a British Royal Society Wolfson Research Merit Award. N. Stein and Professor Meng were supported in part by NSF grant DMS-09-07185. V. Kashyap was partially supported by CXC NASA contract NAS8-39073. The authors are grateful to three reviewers for their very helpful and encouraging comments.

A standard generalization of the $k$-means algorithm replaces squared Euclidean distance with another dissimilarity measure $d(y_1, y_2)$. The so-called *$k$-medoids algorithm* alternates between updating cluster assignments and updating cluster medoids. To initialize the $k$-medoids algorithm, we can randomly chose the starting medoids $m_1, \ldots, m_k$ from the data $\{y_1, \ldots, y_n\}$. We assign each unit to the cluster with the closest medoid as measured by the dissimilarity $d$. Formally, cluster assignments $\{c_1, \ldots, c_n\}$ are determined by setting $c_i$ equal to the $j$ that minimizes $d(m_j, y_i)$. Then, the following two steps are repeated until a convergence criterion is met:

- Update cluster assignments by setting $c_i$ to the $j$ that minimizes $d(m_j, y_i)$ for each unit $i$.
- Update the cluster medoids by setting $m_j$ equal to the minimizer
$$\arg\min_m \sum_{i:c_i=j} d(m, y_i). \tag{1}$$

In cases where the distance is an arbitrary dissimilarity measure only defined pairwise between units in the data set, the cluster medoid is found as the point from the original data set in the given cluster that minimizes the total distance (1). The usual problem with $k$-medoids is that, if there is not a closed-form expression for (1), then finding the minimizer in (1) requires an expensive search, and the computation consequently scales quadratically in the number of observations.

When there is a closed-form solution for the minimizer (1), then it is possible to replace Euclidean distance with another metric without the usual computational disadvantages of the $k$-medoids algorithm. Fortunately, the Hellinger distance (2) has such a closed-form solution. The squared Hellinger distance between two discrete[1] probability distributions $p_1$ and $p_2$ on a state space $\mathcal{X}$ is

$$d_H^2(p_1, p_2) = \frac{1}{2} \sum_{x \in \mathcal{X}} \left( \sqrt{p_1(x)} - \sqrt{p_2(x)} \right)^2. \tag{2}$$

For a collection $p_1, \ldots, p_n$ of discrete probability distributions, the minimizer

$$p^* = \arg\min_p \sum_{i=1}^n d_H^2(p, p_i) \tag{3}$$

is

$$p^*(x) = \frac{\left( \sum_{i=1}^n \sqrt{p_i(x)} \right)^2}{\sum_{x' \in \mathcal{X}} \left( \sum_{i=1}^n \sqrt{p_i(x')} \right)^2}, \tag{4}$$

which can be easily proved using the Cauchy-Schwarz inequality.

[1] In this paper, we focus on the discrete case, but $H$-means can in principle also be applied to continuous probability distributions, although the closed-form solution then requires evaluating a possibly difficult integral. The method is therefore quite general and can be applied whenever features of interest can be expressed as probability distributions.

### 2.1. Spherical $k$-means and $H$-means with tempering

The spherical $k$-means algorithm [6] is closely related to $H$-means. The spherical $k$-means algorithm takes input vectors $\{y_1, \ldots, y_n\}$ with nonnegative entries and normalizes them to have unit length. Unit vectors are then compared according to their cosine similarity. That is, the distance between vectors $y_i$ and $y_j$ is given by

$$d_{\cos}(y_i, y_j) = 1 - \frac{y_i^\top y_j}{||y_i||\,||y_j||}. \tag{5}$$

Like Euclidean $k$-means and $H$-means, the spherical $k$-means algorithm has a closed-form expression for the cluster centroids, enabling fast computation [6].

In fact, the spherical $k$-means algorithm can be viewed as a member of a class of generalized $H$-means algorithms. The Hellinger distance (2) can be expressed as

$$d_H^2(p_1, p_2) = 1 - \sum_{x \in \mathcal{X}} \sqrt{p_1(x) p_2(x)}.$$

We can also "temper" these distributions with a parameter $\alpha \geq 0$, defining

$$p_i^\alpha(x) \equiv \frac{(p_i(x))^\alpha}{\sum_{x' \in \mathcal{X}} (p_i(x'))^\alpha}, \quad i = 1, 2.$$

Then, Hellinger distance and cosine distance are related by

$$d_H^2\left(p_1^2, p_2^2\right) = d_{\cos}\left(p_1, p_2\right). \tag{6}$$

Thus, spherical $k$-means can be viewed as $H$-means with tempered inputs.

## 3. IDENTIFYING SOLAR THERMAL FEATURES

Much is still unknown about the processes that govern thermal features in the corona. The Atmospheric Imaging Assembly is a four-telescope array on the Solar Dynamics Observatory satellite that captures near-simultaneous images through seven extreme ultraviolet wavelength passband filters. Because these different filters have differing responses to temperature, they should be useful in reconstructing the temperature distribution of the emitting plasma in each image pixel. If it were possible to accurately infer these underlying temperature distributions, then astronomers could study the thermal characteristics of interesting features on the Sun, such as loops of hot plasma, and trace the evolution of these thermal properties over time. However, with only at most seven filters, each responding to a fairly wide range of temperatures, reconstructing the temperature distributions is an extremely underconstrained problem. Instead of adding constraints from prior information about the likely shapes of temperature distributions, we choose to bypass the reconstruction step altogether.

Our approach is motivated by a statistical model for the data generating process. For simplicity of illustration, we

assume no background contamination and no spatial dependence. The $J \times 1$ observed vector $y_i$ in pixel $i$ is modeled as a Poisson random variable with mean

$$\lambda_i = PA\mu_i, \tag{7}$$

where $P$ is a diagonal $J \times J$ matrix of exposure times, $A = (a_{jk})$ is a $J \times K$ matrix encoding the response of the $j$th filter to temperature bin $k$, and $\mu_i$ is a $K \times 1$ vector of true intensities in each temperature bin. We parameterize the true intensities as

$$\mu_i = \gamma_i \theta_i, \tag{8}$$

where $\gamma_i$ is a scalar roughly corresponding to the amount of plasma in pixel $i$ and $\theta_i$ is a discretized probability distribution for temperature. For the scientific applications we consider, we focus on properties of $\theta_i$ such as average temperature, and $\gamma_i$ is a nuisance parameter. It is difficult to reliably estimate the temperature distributions because $K$ is larger than $J$.

To avoid directly modeling the niusance parameters $\gamma_i$, we use the conditional likelihood of $y_i$ given the pixel-wise total $\sum_j y_{ij}$, a multinomial distribution free of $\gamma_i$. The multinomial probabilities

$$\pi_i = \frac{PA\theta_i}{\mathbf{1}^\top PA\theta_i}$$

capture all of the temperature information available after the degradation by the response matrix $A$. Thus, in an attempt to retain the available information with respect to this model, we identify solar thermal features by clustering pixels with similar values of the maximum likelihood estimates $\hat{\pi}_i = y_i / \sum_j y_{ij}$.

## 4. RESULTS

### 4.1. Simulations

To investigate the performance of $H$-means clustering, we applied it to six $128 \times 128$ images simulated under the model (7) and (8) using the same response matrix $A$ that we use on AIA data in Section 4.2. Figure 1a plots the simulated temperature map. We used two different underlying temperature distributions arranged in a vertical stripe pattern. The temperature distributions on the $\log_{10}$ of the temperature in Kelvin were Gaussian with standard deviation 0.25 and means 6.00 and 6.05. Values for $\gamma_i$ in (8) were set proportional to the observed total intensity summed across all filters in a slice of an observation of the Sun, shown in Figure 1b.

Figure 1c shows the results of $H$-means clustering with ten clusters based on the data simulated in six bands under (7) and (8). To make the graphical display more meaningful, gray scale values of each cluster were set to the pooled proportions of counts in the 171 Angstrom band. The clustering successfully reveals the vertical stripe pattern of the true underlying temperature distributions, without being obscured by the patterns in the $\gamma_i$ map.

(a) (b) (c)

**Fig. 1**. Results for a simulated image under the model (7) and (8). (a) The simulated map of the true temperature distributions. (b) The simulated map of the nuisance parameters $\gamma_i$, corresponding to the total intensity. (c) The results of $H$-means clustering with 10 clusters.

### 4.2. Application to AIA Data

We have applied $H$-means clustering to six full-resolution ($4096 \times 4096$) images of the Sun, using the 94, 131, 171, 193, 211, and 335 Angstrom filters, observed on October 2, 2010, at 05:57am UT. Figure 2 shows the original, untransformed data in all six bands. Figure 3a shows the estimated proportions $\hat{\pi}_{i,171}$ in the 171 Angstrom filter. That is, Figure 3a plots the pixel-wise ratios of counts in the 171 Angstrom band to the sums of counts across all bands. Figure 3b displays the results of applying $H$-means clustering with 64 clusters. As in Section 4.1, gray scale values for each cluster were set to the pooled proportions of counts in the 171 Angstrom band.

The $H$-means clustering results capture features in the estimated probability images that do not appear in the original images, indicating that clustering is operating on features of the temperature distributions that are of scientific interest. This is reflected in the similarity between Figures 3a and 3b. Moreover, the 'S'-shaped region evident in Figure 3 is a much larger scale feature than the typical solar features that astronomers study, suggesting a direction for further investigation.

## 5. CONCLUSION

$H$-means clustering is a general method for clustering probability distributions based on the Hellinger distance. It is based on the $k$-means algorithm, but by tempering the input probability distributions, $H$-means encompasses other methods designed to reduce the influence of total counts or magnitudes across categories, such as spherical $k$-means. We applied $H$-means in a model-inspired image segmentation algorithm to reveal thermal features in multiband images of the Sun. A key advantage of this method is that is does not require the reconstruction of the underlying temperature distributions in each pixel. A remaining question is the effect of different choices of the tempering parameter $\alpha$, given the connection between $H$-means and spherical $k$-means. Future work will aim to

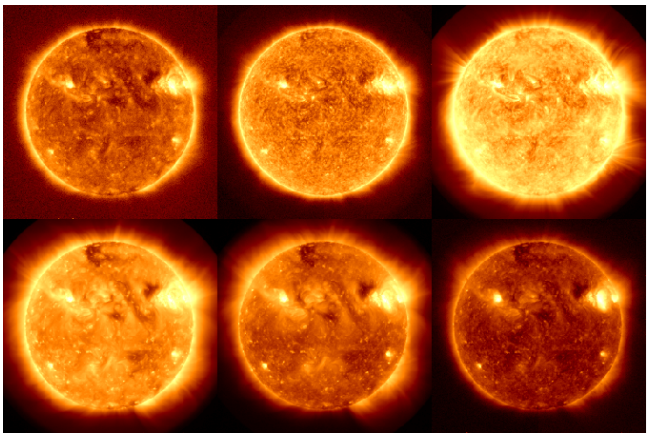
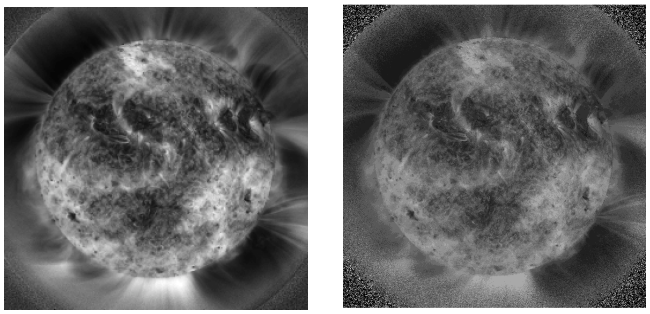


**Fig. 2**. AIA images of the Sun from 05:57am UT on October 2, 2010. The images in the top row, left to right, are the 94, 131, and 171 Angstrom filters. The images in the bottom row are the 193, 211, and 335 Angstrom filters.



(a) (b)

**Fig. 3**. (a) Estimated probabilities $\hat{\pi}_{i,171}$ in the 171 Angstrom filter. (b) $H$-means clustering results using 64 clusters.

develop a statistically principled way to choose $\alpha$ in practice.

## 6. REFERENCES


[1] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.

[2] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 530–549, 2004.

[3] F. Murtagh, A. E. Raftery, and J.-L. Starck, "Bayesian inference for multiband image segmentation via model-based cluster trees," *Image and Vision Computing*, vol. 23, pp. 587–596, 2005.

[4] M. W. Woolrich and T. E. Behrens, "Variational Bayes inference of spatial mixture models for segmentation," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 1380–1391, 2006.

[5] P. Orbanz and J. Buhmann, "Smooth image segmentation by nonparametric Bayesian inference," in *Computer Vision ECCV 2006*, Ale Leonardis, Horst Bischof, and Axel Pinz, Eds., vol. 3951 of *Lecture Notes in Computer Science*, pp. 444–457. Springer Berlin / Heidelberg, 2006.

[6] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, pp. 143–175, 2001.

[7] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, 2003.

[8] J. Kogan, M. Teboulle, and C. Nicholas, "The entropic geometric means algorithm: An approach to building small clusters for large text datasets," in *Proceedings of the Workshop on Clustering Large Data Sets (held in conjunction with the Third IEEE International Conference on Data Mining)*. IEEE, 2003, pp. 63–71.

[9] J. Kogan, M. Teboulle, and C. Nicholas, "Data driven similarity measures for $k$-means like clustering algorithms," *Information Retrieval*, vol. 8, pp. 331–349, 2005.

[10] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[11] M. Teboulle, P. Berkhin, I. Dhillon, Y. Guan, and J. Kogan, "Clustering with entropy-like $k$-means algorithms," in *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds., chapter 5, pp. 127–160. Springer, New York, 2006.

[12] C. Radhakrishna Rao, "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance," *Qüestiió*, vol. 19, pp. 23–63, 1995.

[13] P. Legendre and E. Gallagher, "Ecologically meaningful transformations for ordination of species data," *Oecologia*, vol. 129, pp. 271–280, 2001.

[14] S. A. Gagné and R. Proulx, "Accurate delineation of biogeographical regions depends on the use of an appropriate distance measure," *Journal of Biogeography*, vol. 36, pp. 561–567, 2009.