# The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics

Elizaveta Levina, Peter Bickel
Department of Statistics
University of California, Berkeley
Berkeley CA 94720
[levina, bickel]@stat.berkeley.edu

## Abstract

*The Earth Mover's distance was first introduced as a purely empirical way to measure texture and color similarities. We show that it has a rigorous probabilistic interpretation and is conceptually equivalent to the Mallows distance on probability distributions. The two distances are exactly the same when applied to probability distributions, but behave differently when applied to unnormalized distributions with different masses, called signatures. We discuss the advantages and disadvantages of both distances, and statistical issues involved in computing them from data. We also report some texture classification results for the Mallows distance applied to texture features and compare several ways of estimating feature distributions. In addition, we list some known probabilistic properties of this distance.*

## 1. Introduction

The Earth Mover's distance (EMD) was first introduced by Rubner *et al*. for color and texture images [11, 12]. This distance can be applied to distributions of points (e.g., colors or texture features) as long as the space of points is equipped with some similarity measure. Rubner *et al*. demonstrated that it works well for image retrieval [11]. In addition, it was shown that the EMD outperforms many other texture similarity measures when used for texture classification and segmentation [9]. The EMD has many attractive properties – to some extent it mimics the human perception of texture similarities, it allows for partial matches, and there exist efficient algorithms for computing it. However, so far there has been almost no theoretical justification for the EMD.

The concept of EMD is not new, although implementations and applications vary. The match distance for histograms of pixel intensities introduced in [13] in 1983 and its multidimensional extension [14] are based on the same idea of matching the closest values. And there is an equivalent metric on probability distributions known as Mallows, or Wasserstein, distance, which has a clear probabilistic interpretation. It was introduced in the statistical literature in 1972 by [7], but it had also independently appeared a little earlier in the physics and probability literatures, and some date it all the way back to the 1940s [10]. For the case of two distributions with equal masses, the EMD is exactly the same as the Mallows distance. The case of unequal masses is not formally covered by the Mallows distance as all probability distributions are normalized to have total mass 1. In this case, the EMD and Mallows behave differently, and one may have an advantage over the other depending on the context; this issue will be discussed in detail in section 2.2.

This paper is organized as follows: in section 2, we define the EMD and Mallows distances, demonstrate their equivalence for the case of equal masses and discuss the differences for the case of unequal masses. In section 3, we discuss how the Mallows distance can be computed from data, including the special case of one-dimensional data which does not require solving the optimization problem. Section 4 presents some empirical results for texture classification, comparing several ways of applying the Mallows distance to textures. Section 5 concludes with a summary, and the Appendix lists some mathematical properties of the Mallows distance.

## 2. Comparing the Earth Mover's and Mallows distances

### 2.1. Definitions and equivalence

Let us start with the formal definitions of the two distances. The Earth Mover's distance is defined for "signatures" of the form $\{(x_1, p_1) \ldots, (x_m, p_m)\}$, where $x_i$

is the center of data cluster $i$ and $p_i$ is the number of points in the cluster. The signatures are not normalized, so the total masses of two signatures may not be equal. Given two signatures $P = \{(x_1, p_1), \ldots, (x_m, p_m)\}$ and $Q = \{(y_1, q_1), \ldots, (y_n, q_n)\}$, the EMD is defined in terms of an optimal flow $F = (f_{ij})$, which minimizes

$$W(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$$

where $d_{ij} = d(x_i, y_j)$ is some measure of dissimilarity between $x_i$ and $y_j$, e.g., the Euclidean distance in $R^d$. In the EMD terminology, $W(P, Q, F)$ is the work required to move earth from one signature to another. The flow $(f_{ij})$ must satisfy the following constraints:

$$\begin{aligned} f_{ij} &\geq 0, \quad 1 \leq i \leq m, 1 \leq j \leq n && (1) \\ \sum_{j=1}^{n} f_{ij} &\leq p_i, \quad 1 \leq i \leq m && (2) \\ \sum_{i=1}^{m} f_{ij} &\leq q_j, \quad 1 \leq j \leq n && (3) \\ \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} &= \min(\sum_{i=1}^{m} p_i, \sum_{j=1}^{n} q_j) && (4) \end{aligned}$$

Once the optimal flow $f^*_{ij}$ is found, the Earth Mover's distance between $P$ and $Q$ is defined as

$$\text{EMD(P,Q)} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f^*_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f^*_{ij}} \qquad (5)$$

Now let us switch to statistical terminology and introduce the Mallows distance. $X$ and $Y$ are now random variables with distributions $P$ and $Q$ in $R^d$, respectively. The Mallows distance between $P$ and $Q$ can in general be defined by a minimum of the expected difference between $X$ and $Y$, taken over all joint probability distributions $F$ for $(X, Y)$ such that the marginal distribution of $X$ is $P$ and the marginal of $Y$ is $Q$:

$$\begin{aligned} M_p(P, Q) &= \min_F \{(E_F \|X - Y\|^p)^{1/p} : (X, Y) \sim F \\ & \quad X \sim P, \ Y \sim Q\}. \end{aligned}$$

Here $p$ can be any number greater or equal to 1, but the most interesting cases are $p = 1$ and $p = 2$. $\| \cdot \|$ is usually taken to be the Euclidean or $L^1$ vector norm. For the definition to make sense, the distributions $P$ and $Q$ must have finite $p$-th moments ($E[\|X\|^p] < \infty$ and $E[\|Y\|^p] < \infty$).

Now let us write out this definition for the case of two discrete distributions $P = \{(x_1, p_1), \ldots, (x_m, p_m)\}$ and $Q = \{(y_1, q_1), \ldots, (y_n, q_n)\}$. Note that signatures can always be converted to proper probability distributions by normalizing the weights to add up to 1. We need to minimize the expectation under $F = (f_{ij})$, the joint distribution of $X$ and $Y$:

$$E_F \|X - Y\|^p = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \|x_i - y_j\|^p = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}.$$

The distribution $F$ is subject to the following constraints:

$$\begin{aligned} f_{ij} &\geq 0, \quad 1 \leq i \leq m, 1 \leq j \leq n && (6) \\ \sum_{j=1}^{n} f_{ij} &= p_i, \quad 1 \leq i \leq m && (7) \\ \sum_{i=1}^{m} f_{ij} &= q_j, \quad 1 \leq j \leq n && (8) \\ \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} &= \sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j = 1. && (9) \end{aligned}$$

Constraints 6 and 9 ensure that $F$ is indeed a distribution and are the same for the EMD flow (since both $P$ and $Q$ are proper probability distributions with total mass 1, 4 is the same as 9). Moreover, as long as $P$ and $Q$ have the same total mass, the EMD constraints 2 and 3 are forced to become equalities in order to satisfy 4. As noted in [11], for signatures with the same total mass the EMD is a true metric on distributions, and it is *exactly* the same as the Mallows distance (note that normalizing signatures with the same mass to have total mass 1 does not affect their EMD).

## 2.2. The case of unequal total masses

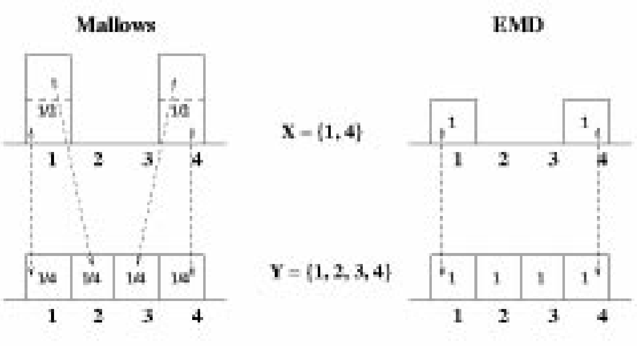

**Figure 1. The difference between distributions and signatures: same data, different normalization**

When the two signatures have different masses, the EMD does something truly different from Mallows. Let us start with a toy example: suppose we have two sets of data, $X = \{1, 4\}$, and $Y = \{1, 2, 3, 4\}$. If we normalize these to have total mass 1, then each point in $X$ has weight $1/2$, each point in $Y$ has weight $1/4$, and it is easy to check that the joint distribution of $X$ and $Y$ that gives mass $1/4$ to pairs $(1,1), (1,2), (4,3), (4,4)$ and 0 to all others satisfies the constraints and solves the optimization problem. If we use the $L^1$ norm to measure the ground distance and set $p = 1$, then $M_1(X,Y) = \text{EMD}(X,Y) = 1/2$. However, if we use signatures and give every point weight 1 (so that the total mass of $X$ is 2 and the total mass of $Y$ is 4), it is easy to see that $\text{EMD}(X,Y) = 0$ (one can either compute it directly or note that $X$ is a subset of $Y$ and the EMD allows for partial matching). Arrows on Figure 1 show how the points are matched in both cases.

From the statistical point of view, this property of the EMD is probably a disadvantage. In the toy example above, even if $Y$ contained a thousand other points with very different values, the distance between $X$ and $Y$ would still be 0, so just two points from a sample of a thousand would determine the distance. However, there exist other non-statistical contexts where partial matches may be appropriate, such as image retrieval. Nevertheless, since the EMD allows for matching any part of the distribution, no matter how small, partial matches may be spurious, especially if the sizes of the two signatures being compared are very different. It is quite possible that textures would produce spurious matches. (Note that the excellent EMD texture classification results reported in [9] were obtained by comparing signatures of the same size, so partial matching was never a problem.) Also, the EMD on signatures is not invariant to weight scaling, unless both signatures are scaled by the same factor. So if, for example, one of the two texture patches is duplicated to produce a larger image, the distance between the two textures will change. Partial matching is a computationally efficient and convenient way to search a large image for a small match, but it should be used with caution, especially outside the image retrieval context.

## 3. Computing the Mallows distance from data

In practice, the distributions which we want to compare are unknown, so the distance between them cannot be computed exactly. In the texture framework, for example, if we believe that two textures have "true" feature distributions $P$ and $Q$, then our goal is to estimate the distance $d(P,Q)$. However, we do not know $P$ and $Q$, so one way to estimate the distance would be to construct some distribution estimates $\hat{P}$ and $\hat{Q}$ from data and estimate $d(P,Q)$ by $d(\hat{P},\hat{Q})$. The triangle inequality implies that

$$|d(P,Q) - d(\hat{P},\hat{Q})| \leq d(\hat{P},P) + d(\hat{Q},Q),$$

so if the distribution estimates $\hat{P}$ and $\hat{Q}$ are good, then the distance will also be estimated accurately. This is not the only possible way to estimate the distance, but it is rather natural.

It is important to distinguish between the issues of picking the right distance for the problem (e.g., Mallows or $\chi^2$ or $L^1$) and estimating the distributions well from the available data (by a fixed-bin histogram, adaptive-bin histogram, signature, or some other method). There is an abundance of statistical literature on how to estimate distributions; the choice depends on the amount of available data, the dimensionality of the data, and the questions about the distribution one needs to answer. Once the distributions are estimated, one must make another choice on how to measure the distance between them, which again depends on the problem. There are no a priori reasons for these two issues to be connected, other than perhaps computational complexity.

In theory, the Mallows distance can be computed for any probability distribution, discrete or continuous; in practice, it is convenient to use optimization algorithms for the transportation problem [5], so the distributions need to be discrete. The optimization problem can be stated especially compactly if we have two samples of the same size $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ and use the empirical distribution function as our estimate (i.e., give every point weight $1/n$ and do not bin). Then the Mallows distance between empirical distributions is

$$M_p(X,Y) = \left( \frac{1}{n} \min_{(j_1,\ldots,j_n)} \sum_{i=1}^{n} \|x_i - y_{j_i}\|^p \right)^{1/p} \qquad (10)$$

where the minimum is taken over all possible permutations of $\{1,\ldots,n\}$. In this case it is convenient to use the Hungarian algorithm for the optimal assignment problem [4], a special case of the transportation problem. If the observations are one-dimensional, the optimization problem can be solved explicitly: let $x_{(1)} \leq \ldots \leq x_{(n)}$ and $y_{(1)} \leq \ldots \leq y_{(n)}$ be the sorted vectors $X$ and $Y$; then the Mallows distance is just the $L^p$ vector distance between the sorted vectors:

$$M_p(X,Y) = \left( \frac{1}{n} \sum_{i=1}^{n} |x_{(i)} - y_{(i)}|^p \right)^{1/p}.$$

It is interesting to note that the match distance on one-dimensional "unfolded histograms" for texture intensities [13] and its multidimensional extension [14] can be written in the form of equation 10. Both of them are nothing but the Mallows distance applied to the empirical distributions.

If we have two samples of unequal sizes $m$ and $n$, it is still possible to apply the Hungarian algorithm by replicating each observation so that both samples have the size of the least common multiple of $m$ and $n$. Of course it only makes sense to do so if the least common multiple of $m$ and $n$ is not too large; otherwise in practice one bins the distributions or applies the general algorithm to the rectangular matrix.

Using an adaptive binning technique like signatures has some attractive properties. The way the signatures are constructed for textures – clustering filter responses into a few clusters and then computing the frequency of each cluster – corresponds to the concept of textons in the sense of [6]. There textons were defined as frequently occurring filter responses, or texture prototypes, and texton distributions were estimated by the same technique as the one used for texture signatures in [11], i.e., clustering filter responses and computing cluster frequencies. Both [6] and [11] demonstrate that the distributions of textons provide an accurate and compact way to describe textures. However, one must be aware that even though this approach does not depend on a fixed bin size, there are still artifacts from the choice of the clustering algorithm, the number of clusters, etc. Using the empirical distributions, on the other hand, does not involve any additional algorithms or parameters, and it may lead to a more accurate estimate of the distance between distributions (if it is computationally feasible to use the whole sample). However, the dimension of the data should also be taken into account: one-dimensional data is special, since it only requires sorting, but in dimensions 2 and higher it may be necessary to coarsen the distribution estimates if the transportation problem algorithm becomes slow. There is some empirical evidence that for the EMD-based texture classification estimating high-dimensional distributions may be avoided altogether [9]. We discuss this in more detail in the next section.

## 4. Experimental results

Extensive empirical evidence on the usefulness of the EMD for texture analysis and image retrieval has already been published [11, 12, 9]; therefore we chose not to conduct large-scale experiments. Instead, we tested the use of the empirical distribution as an estimate on the relatively small MeasTex texture database [8], which consists of 16 Brodatz textures [2]. Following the benchmarking strategy of [9], we extracted sets of 16 random non-overlapping blocks from each texture, with sizes $16 \times 16$, $32 \times 32$, $64 \times 64$, and $128 \times 128$. Then for each sample size we computed the average classification rate by the "leave-one-out" method (cross-validation in statistical terminology). This means leaving out each image in turn, computing its distance to all the other images, and assigning it to the class of its nearest neighbor. The classification error rate is then estimated by the percentage of incorrectly assigned images.

First we describe how the Mallows distance can be applied to pixel intensities. The texture synthesis method of [3] suggests that the appearance of texture is largely determined by the joint distribution of pixel intensities in a window of a suitable size. It is also consistent with the idea that filter responses determine texture appearance, since the joint distribution of pixel intensities in the filter support window determines the distribution of filter responses. The sample windows were created by sampling at random a fixed number (300) of overlapping square texture patches from each image. The ground distance between patches was measured as the sum of squared differences of pixel intensities. This method produces reasonable classification errors (12% on the $128 \times 128$ size), although it seems so simple-minded that one would not expect it to work at all. However, the methods based on filter responses are a lot better, and this example is intended only as an illustration.

Estimating the Mallows distance between distributions of filter responses can be done in two ways: using the one-dimensional marginals (distributions of individual filter responses) or the joint distribution of all filter responses. If one uses the empirical distributions rather than binned histograms, for the one-dimensional marginals all one needs to do is sort the vectors, which can be done very fast. On the other hand, the Hungarian algorithm needed for the joint distributions becomes slow for large images, and using empirical distributions is no longer feasible. The "curse of dimensionality" is also an issue, because the high-dimensional joint distribution is harder to estimate that the one-dimensional marginals. (The number of filters in our experiments is 40, all of them first or second derivatives of a 2-D Gaussian at different scales and orientations.) We have done a small number of experiments with joint distributions and found that using the marginals of filter responses gives better classification results and is faster to compute. This agrees with results in [9], where using marginals of filter responses rather than the joint distribution also produced somewhat more accurate classification. For these reasons, we only report detailed Mallows distance results for the marginals of filter responses, comparing four methods of estimating the distribution: empirical distribution (no binning), coarse fixed-bin histogram (16 bins), fine fixed-bin histogram (256 bins), and adaptive-bin histogram where responses are clustered into 16 bins by a $k$-means type algorithm. In all cases, the Mallows distance between two textures is the sum of the distances between individual filter marginals, the vector norm is $L^1$, and $p = 2$. These results are presented in Table 1.

The results confirm what one might expect – the empirical distribution function contains the most information and consistently does better than other estimates. The adaptive-bin histogram is nearly as good and requires less memory, but takes longer to compute, so the choice should depend on the particular application. The fixed-bin histograms perform substantially worse. Finally, the larger the image size, the easier the classification problem, and for $128 \times 128$ textures all methods perform reasonably well. One should keep

| Distribution | Image size | | | |
|---|---|---|---|---|
| estimate | 16 | 32 | 64 | 128 |
| Empirical | 35.94 | 5.86 | 1.56 | 0 |
| Adaptive hist. | 36.33 | 8.20 | 1.56 | 0 |
| Coarse hist. | 45.31 | 12.50 | 4.69 | 1.17 |
| Fine hist. | 51.95 | 28.91 | 14.06 | 7.03 |

**Table 1. Texture classification results: percent misclassified**

in mind, however, that on a larger database the differences we see on small images may show on large image sizes as well.

## 5. Summary and conclusions

In this paper we demonstrated the connection between the Earth Mover's distance and Mallows distance on distributions, which has a clear probabilistic interpretation. The solid theoretical foundation may be helpful for further understanding of why Earth Mover's distance performs so well for various vision tasks and for establishing its properties. We also discussed different methods of estimating the distributions and advantages and disadvantages of using unnormalized signatures. A few experimental results were presented as an illustration of the methods. It is our hope that the computer vision community will find it useful to be aware of the statistical theory and issues behind this successful but so far mostly empirical technique.

## Acknowledgments

We would like to thank Jitendra Malik and other current and former members of the Berkeley vision group, in particular Serge Belongie, Alex Berg, Alexei Efros, Thomas Leung, Jan Puzicha, and Jianbo Shi for helpful discussions and shared code.

## References

[1] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9:1196–1217, 1981.

[2] P. Brodatz. *Textures*. Dover, New York, 1966.

[3] A. Efros and T. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1033–1038. Corfu, Greece, Sept. 1999.

[4] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[5] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, 1984.

[6] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours, and regions: cue combination in image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 918–925. Corfu, Greece, Sept. 1999.

[7] C. L. Mallows. A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43(2):508–515, 1972.

[8] Meastex image texture database and test suite. Website: `http://www.cssip.elec.uq.edu.au/˜guy/meastex_v1.1/meastex.html`.

[9] J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1165–1173. Corfu, Greece, Sept. 1999.

[10] S. T. Rachev. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications*, 29:647–676, 1984.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's distance as a metric for image retrieval. Technical Report STAN-CS-TN-98-86, Department of Computer Science, Stanford University, Sept. 1998.

[12] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 59–66. Bombay, India, Jan. 1998.

[13] H. C. Shen and A. K. C. Wong. Generalized texture representation and metric. *Computer Vision, Graphics, and Image Processing*, 23:187–206, 1983.

[14] M. Werman, S. Peleg, and A. Rozenfeld. A distance metric for multidimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32:328–336, 1985.

## A Mathematical properties of the Mallows distance

This appendix lists some of the Mallows distance properties that have been studied in the statistical literature. A more detailed and mathematical treatment of these properties can be found in [1]. Unless stated otherwise, $p \in [1, \infty)$ and all distributions have finite $p$-th moments.

1. $M_p$ is a metric, i.e.,

   (i) $M_p(F, G) = 0$ if and only if $F = G$,

   (ii) $M_p(F, G) = M_p(G, F)$,

   (iii) $M_p(F, G) \leq M_p(F, H) + M_p(H, G)$.

2. Convolution property for $p = 2$: if $\int x \, dF_i(x) = \int x \, dG_i(x)$ for $i = 1 \ldots n$, then

   $$M_2^2(F_1 * \ldots * F_n, G_1 * \ldots * G_n) \leq \sum_{i=1}^{n} M_2^2(F_i, G_i).$$

   The $*$ stands for convolution of cumulative distribution functions, so $F_1 * \ldots * F_n$ is the distribution of the sum of independent random variables with distributions $F_1, \ldots, F_n$. This property is stronger than the triangle inequality.

3. $M_p(F_n, F) \to 0$ if and only if

   (i) $F_n \to F$ weakly, that is, $F_n(x) \to F(x)$ for every $x$ at which $F$ is continuous, and

   (ii) $\int \|x\|^p dF_n(x) \to \int \|x\|^p dF(x)$.

4. If $X_1, \ldots, X_n$ are independent observations from a distribution $F$, and $F_n$ is their empirical distribution, i.e., $F_n(t) = 1/n \sum_{i=1}^{n} \mathbf{1}(X_i \leq t)$, then $M_p(F_n, F) \to 0$.

5. If $F$ and $G$ are distributions on the real line, then

   $$M_p(F, G) = \left( \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt \right)^{1/p}.$$

   The case $p = 1$ is especially simple because

   $$\int_0^1 |F^{-1}(t) - G^{-1}(t)| dt = \int_{-\infty}^{\infty} |F(t) - G(t)| dt.$$