On the Generalization of the Mahalanobis Distance

Gabriel Martos[1], Alberto Muñoz[1], and Javier González[2]

[1] University Carlos III, Department of Statistics, Madrid, Spain
[2] J. Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands
{gabrielalejandro.martos,alberto.munoz}@uc3m.es, j.gonzalez.hernandez@rug.nl

Abstract. The Mahalanobis distance (MD) is a widely used measure in Statistics and Pattern Recognition. Interestingly, assuming that the data are generated from a Gaussian distribution, it considers the covariance matrix to evaluate the distance between a data point and the distribution mean. In this work, we generalize MD for distributions in the exponential family, providing both, a definition in terms of the data density function and a computable version. We show its performance on several artificial and real data scenarios.

1 Introduction

The Mahalanobis distance (MD) [5], widely used in Statistics and Machine Learning for classification and outlier detection tasks, is a scale-invariant metric that provides a measure of distance between two points taking into account the correlation between the variables. It can be seen as the composition of the linear transformation $T_M : \mathbf{x} \xrightarrow{T_M} \mathbf{x}' = \Sigma^{-\frac{1}{2}}\mathbf{x}$, where Σ is the covariance matrix of a vector of random variables $\mathbf{x}$, plus the computation of the ordinary Euclidean distance (ED) between the transformed data. This is illustrated in Fig. 1 for two data points from a bivariate Gaussian distribution. The distance in probability (d_M) from B to the mean μ is larger than the distance from A to μ, which is correctly detected by the MD, but not by the ED (d_E).

The Mahalanobis distance is a particular case of the Bregman Divergence (see Def. 1), a generalization of the concept of distance. We will show that this connection allows us to generalize the concept of distance from a point to the center of a distribution (the densest point) for density functions in the exponential family, a quite general case. The rest of this paper is organized as follows. In Section 2 we introduce the new distance, in terms of the data density function and then we provide a computable version of the distance. In Section 3 we show the performance of the generalized MD for outlier detection and classification problems.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 125–132, 2013.
© Springer-Verlag Berlin Heidelberg 2013

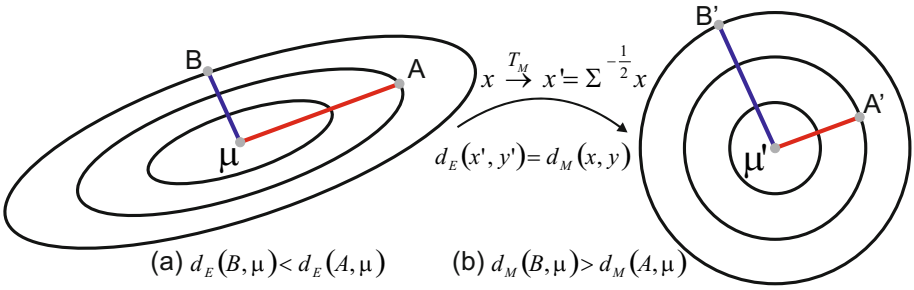

Fig. 1. The effect of the Mahalanobis transformation T_M

2 A Generalized Mahalanobis Bregman Divergence

Our goal in this section is to define a Generalized Mahalanobis distance to the center of a general Probability Measure (distribution), that is, a distance for distributions non necessarily Gaussian.

Consider a measure space $(\mathcal{X}, \mathcal{F}, \mu)$, where $\mathcal{X}$ is a sample space (here a compact set of a real vector space), $\mathcal{F}$ a σ-algebra of measurable subsets of $\mathcal{X}$ and $\mu : \mathcal{F} \to \mathbb{R}^+$ the ambient σ-additive measure, the Lebesgue measure. A probability measure $\mathbb{P}$ is a σ-additive finite measure absolutely continuous w.r.t. μ that satisfies the three Kolmogorov axioms. By Radon-Nikodym theorem, there exists a measurable function $f : \mathcal{X} \to \mathbb{R}^+$ (the density function) such that $P(A) = \int_A f d\mu$, and $f = \frac{d\mathbb{P}}{d\mu}$ is the Radon-Nikodym derivative.

In the Multivariate Gaussian case, say $f = \mathbf{N}_d(\mu, \mathbf{\Sigma})$ where $\mu \in \mathbb{R}^d$ and $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ are respectively the mean vector and the covariance matrix, it holds that for $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x}|\mu, \mathbf{\Sigma}) \propto e^{-\frac{1}{2} d_M^2(\mathbf{x}, \mu)}$ and MD is defined by:

$$d_M(\mathbf{x}, \mu) = \sqrt{(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)}.$$

Next we show that MD is as a particular case of the Bregman Divergence:

Definition 1. *(**Bregman Divergence**): Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact domain and ξ a strictly convex and differentiable function $\xi : \mathcal{X} \to \mathbb{R}$. Define the* Bregman Divergence *(BD) for a pair of points $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}$ as follows*

$$BD_\xi(\mathbf{x}, \mathbf{y}) = \xi(\mathbf{x}) - \xi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla \xi(\mathbf{y}) \rangle, \tag{1}$$

where $\nabla \xi(\mathbf{y})$ is the gradient vector evaluated at the point $\mathbf{y}$. Taking $\xi(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mathbf{x}$, it is immediate to verify that BD is the square of MD.

In general, there exists a bijective correspondence between Bregman divergences and the class of (regular) exponential distributions [1,3]. An example is the mentioned Normal distribution whose corresponding BD is the square of the MD. However, the square of the MD can be expressed in an alternative and interesting way as follows:

$$f(\mathbf{x}) \propto e^{-\frac{1}{2} d_M^2(\mathbf{x}, \mu)} \Longrightarrow d_M^2(\mathbf{x}, \mu) \propto \log\left(\frac{1}{f(\mathbf{x})}\right), \tag{2}$$

Now, if f belongs to the regular exponential family, f can be expressed by $f(\mathbf{x}) \propto e^{-\frac{1}{2}BD_\xi(\mathbf{x},\mu)}$ for appropriate ξ [1,3] and, thus:

$$f(\mathbf{x}) \propto e^{-\frac{1}{2}BD_\xi(\mathbf{x},\mu)} \Longrightarrow BD_\xi(\mathbf{x},\mu) \propto \log\left(\frac{1}{f(\mathbf{x})}\right), \tag{3}$$

which gives us the hint to generalize the MD to any distribution in the exponential family.

Definition 2. *(**Generalized Mahalanobis Distance**): Given a (d-dimensional) distribution f in the exponential family and denote by $\boldsymbol{m}_o$ the mode of f, that is, $f(\boldsymbol{m}_o) = \max_{\mathbf{x}} f(\mathbf{x})$, we define the Generalized Mahalanobis distance (GM) between $\mathbf{x} \in \mathcal{X}$ and the mode ($\boldsymbol{m}_o$) of f by*

$$d^2_{GM}(\mathbf{x}, \boldsymbol{m}_o) = \log\left(\frac{f(\boldsymbol{m}_o)}{f(\mathbf{x})}\right). \tag{4}$$

When $\mathbf{x} = \mathbf{m}_o$, $d^2_{GM}(\mathbf{x}, \mathbf{m}_o) = \log(1) = 0$, and $d^2_{GM}(\mathbf{x}, \mathbf{m}_o)$ increases when $\mathbf{x}$ moves off from the mode $\mathbf{m}_o$. What is the connection between BD and GM? As already told, BD is only defined for distributions on the exponential family. In the important case of the normal distribution[1], $BD_\xi(\mathbf{x}, \mathbf{m}_o) = 2d^2_{GM}(\mathbf{x}, \mathbf{m}_o)$. In the case of the gamma distribution[1] with shape parameter α, $BD_\xi(\mathbf{x}, \mathbf{m}_o) = \frac{\alpha}{\alpha-1} d^2_{GM}(\mathbf{x}, \mathbf{m}_o)$ (provided that there exist a mode: $\alpha > 1$). Thus, BD and GM are "formally" equivalent for distributions in the exponential family. The advantage of the GM are two: First, it is always defined for any continuous regular distribution, but BD is not out of the exponential family. Second, it is possible to derive a sample version of the GM by just providing an estimator of $f(\mathbf{x})$.

From a practical point of view, we are interested in the GM to solve classification and outlier detection problems. Thus the relevant information here is not the exact value of the distance, but the relative order among the distances from data points to the center of the distribution (the densest point). Therefore, we do not need to know $f(\mathbf{x})$, but given $\mathbf{x}$ and $\mathbf{y}$, it is enough to know if $f(\mathbf{x}) < f(\mathbf{y})$ or $f(\mathbf{x}) > f(\mathbf{y})$. To this aim, we just need to estimate the α-level sets of f: Given a probability measure $\mathbb{P}$ with density function $f_\mathbb{P}$, the minimum volume sets (or α-level sets) are defined by $S_\alpha(f_\mathbb{P}) = \{\mathbf{x} \in \mathcal{X} | \, f_\mathbb{P}(\mathbf{x}) \geq \alpha\}$, such that $P(S_\alpha(f_\mathbb{P})) = 1 - \nu$, where $0 < \nu < 1$. If we consider an ordered sequence $\alpha_1 < \ldots < \alpha_m$, then $S_{\alpha_{i+1}}(f_\mathbb{P}) \subseteq S_{\alpha_i}(f_\mathbb{P})$. Let us define $A_i(\mathbb{P}) = S_{\alpha_i}(f_\mathbb{P}) - S_{\alpha_{i+1}}(f_\mathbb{P})$, $i \in \{1, \ldots, m-1\}$. We can choose $\alpha_1 \simeq 0$ and $\alpha_m \geq \max_{\mathbf{x}\in\mathcal{X}} f_\mathbb{P}(\mathbf{x})$ (which exists, given that $\mathcal{X}$ is compact and $f_\mathbb{P}$ continuous). If the $\{\alpha_i\}_{i=1}^m$ sequence is long enough, we can assume constant density for the points contained in $A_i(\mathbb{P})$, that is, they have the same value $f(\mathbf{x})$.

If $\mathbf{x} \in A_i(\mathbb{P})$, and because of the definition of $A_i(\mathbb{P})$, then $f(\mathbf{x}) \approx \alpha_i$ and thus:

$$d^2_{GM}(\mathbf{x}, \mathbf{m}_o) = \log\left(\frac{f(\mathbf{m}_o)}{f(\mathbf{x})}\right) \approx \log\left(\frac{f(\mathbf{m}_o)}{\alpha_i}\right). \tag{5}$$

Next we introduce the algorithm to estimate the $A_i(\mathbb{P})$ sets.

[1] Proof is omitted for lack of space.

Table 1. Algorithmic formulation of Theorem 1

Obtention of $\mathbf{R_n} = \hat{\mathbf{S}}_\alpha(\mathbf{f})$:
1 Choose a constant $\nu \in [0, 1]$.
2 Consider the order induced in the sample s_n by the sparsity measure $g_n(\mathbf{x})$, that is, $g_n(\mathbf{x}_{(1)}) \leq \cdots \leq g_n(\mathbf{x}_{(n)})$, where $\mathbf{x}_{(i)}$ denotes the i^{th} sample, ordered after g.
3 Consider the value $\rho_n^* = g(\mathbf{x}_{(\nu n)})$ if $\nu n \in \mathbb{N}$, $\rho_n^* = g_n(\mathbf{x}_{([\nu n]+1)})$ otherwise, where $[\mathbf{x}]$ stands for the largest integer not greater than $\mathbf{x}$.
4 Define $h_n(\mathbf{x}) = \text{sign}(\rho_n^* - g_n(\mathbf{x}))$.

2.1 Level Set Estimation

Usually the available data are given as a finite sample. We will consider an *iid* sample $s_n(\mathbb{P}) = \{\mathbf{x}_i\}_{i=1}^n$ drawn from the density function $f_{\mathbb{P}}$. To estimate level sets from a data sample (useful to obtain $\hat{S}_\alpha(f_{\mathbb{P}})$) we present the following definitions and theorems, concerning the One-Class Neighbor Machine [7,8].

Definition 3 (Neighbourhood Measures). *Consider a random variable X with density function $f(\mathbf{x})$ defined on $\mathbb{R}^d$. Let S_n denote the set of random independent identically distributed (iid) samples of size n (drawn from f). The elements of S_n take the form $s_n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$. Let $M : \mathbb{R}^d \times S_n \longrightarrow \mathbb{R}$ be a real-valued function defined for all $n \in \mathbb{N}$. (a) If $f(\mathbf{x}) < f(\mathbf{y})$ implies $\lim_{n\to\infty} P(M(\mathbf{x}, s_n) > M(\mathbf{y}, s_n)) = 1$, then M is a **sparsity measure**. (b) If $f(\mathbf{x}) < f(\mathbf{y})$ implies $\lim_{n\to\infty} P(M(\mathbf{x}, s_n) < M(\mathbf{y}, s_n)) = 1$, then M is a **concentration measure**.*

The Support Neighbour Machine [7,8] solves the following optimization problem:

$$\begin{aligned} \max_{\rho,\xi} \; & \nu n\rho - \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \; & g(\mathbf{x}_i) \geq \rho - \xi_i \,, \\ & \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{6}$$

where $g(\mathbf{x}) = M(\mathbf{x}, s_n)$ is a sparsity measure, $\nu \in [0, 1]$, ξ_i with $i = 1, \ldots, n$ are slack variables and ρ is a threshold induced by the sparsity measure.

Theorem 1. *The set $R_n = \{\mathbf{x} : h_n(\mathbf{x}) = \text{sign}(\rho_n^* - g_n(\mathbf{x})) \geq 0\}$ converges to a region of the form $S_\alpha(f) = \{\mathbf{x} | f(\mathbf{x}) \geq \alpha\}$, such that $P(S_\alpha(f)) = 1 - \nu$.*

Therefore, the Support Neighbour Machine estimates a density contour cluster $S_\alpha(f)$ (around the mode). Theorem 1 [7,8] can be expressed in algorithmic form as in Table 1: Hence, we take $\hat{A}_i(\mathbb{P}) = \hat{S}_{\alpha_i}(f_{\mathbb{P}}) - \hat{S}_{\alpha_{i+1}}(f_{\mathbb{P}})$, where $\hat{S}_{\alpha_i}(f_{\mathbb{P}})$ is estimated by R_n defined above (for further details on the estimation refers to [7,8]). Whit the estimation of level sets and the relation presented in Equation 2, we will test with some experiment the performance of the proposed distance.

3 Experimental Section

To demonstrate the capability of the proposed distance, we test it in one artificial and two real data experiments.

Artificial Experiments

The goal of the first experiment is to demonstrate that the GM adequately captures a significant amount of outliers in non-Gaussian scenarios. We keep the distribution simple and visually tractable in this example. We simulate 1000 points from a bimodal and asymmetric bi-logistic distribution [9], with parameters $BL(\alpha = 0.5, \beta = 0.9)$. The values of the parameters $\alpha = 0.5$ and $\beta = 0.9$ where selected in order to obtain a bi-modal distribution in the sampled data. We replace some of these observations with contaminated observations (noise) normally distributed with parameters $\mathbf{N}_d(\mu = (3,3), \mathbf{\Sigma} = 5I_{2\times 2})$. The simulation process was: first we generate a vector u of size 1000, uniformly distributed in $[0,1]$. Then for each value of $u \leq .95$ we generate a data point from the $BL(\alpha = .5, \beta = .9)$ distribution, in the other case we generate a data from a $\mathbf{N}_d(\mu = (3,3), \mathbf{\Sigma} = 5I_{2\times 2})$.

Table 2. Outlier detection performance

Metric/Technique	% of: Outliers captured	False-positives (Type I error)	False-negatives (Type II error)
pc-Outlier[2]	36.5%	23.2%	65.8%
sign-Outlier[2]	23.1%	7.4%	76.9%
locoutPercent[2]	13.4%	**7.3**%	86.4%
Percentile 5% Euclidean Distance	3.8%	10.7%	96.1%
Percentile 5% Mah. Distance	23.1%	10.4%	76.9%
Percentile 5% Gen. Mah. Distance	**38.5**%	10.3%	**65.4**%

We use a battery of different algorithms [2,10] to identify contaminated points (outliers) for the simulated data. The results are summarized in Table 2. Our metric outperforms the other metrics in the detection of the contaminated points. We also get the lowest rate of unidentified outliers (false-negatives rate) and a very competitive rate of false identification of outliers (false-positives rate) compared to other more sophisticated techniques. In Figure 2, we present the points revealed as contaminated points in all the considered cases. The GM adequately capture those points that are far apart from the "center" of the bimodal and asymmetric sampled distribution.

Real Data Experiments

For the first real example, we consider a collection of 1774 documents (corresponding to 13 topics) extracted from three bibliographic data bases (LISA,

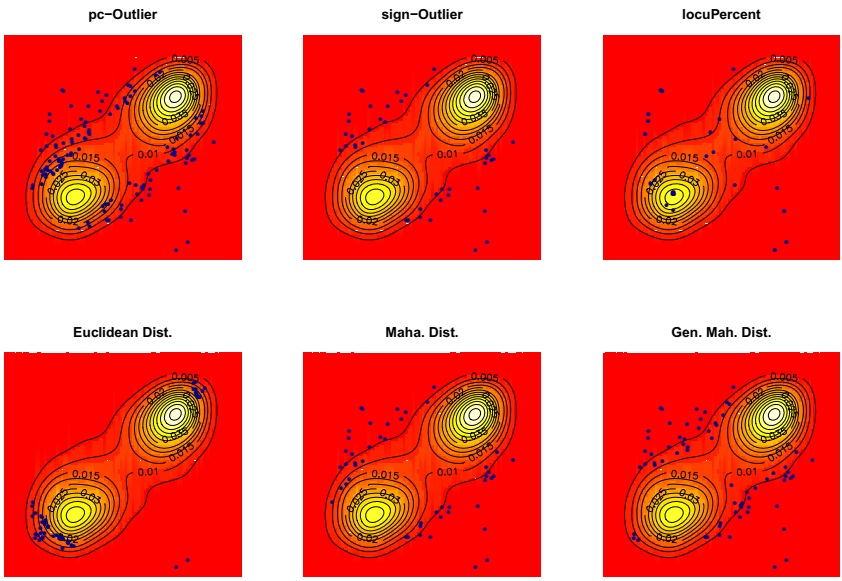

Fig. 2. Contaminated points detected for each of the method/metric

Table 3. Classification percentage errors for a three-class text database and four classification procedures. In parenthesis the St. Error on test samples are shown.

Method	% of: **Train Error**	**Test Error**
SVM	0.000%	0.005% (0.000)
LDA	6.100%	7.035% (0.007)
QDA (Mahalanobis)	6.426%	6.960% (0.001)
Generalized Mahalanobis	2.553%	2.761% (0.002)

INSPEC and Sociological Abstracts). Each document is converted into a vector into the Latent Semantic Space using the Singular Value Decomposition. We considers 3 classes of similar topics: "dimensionality reduction" and "feature selection" (311 documents), "optical cables" and "power semiconductor devices" (384 documents) and "rural areas" and "retirement communities" (165 documents). In order to implement the classification we divide the 860 documents into a training sample (516 documents, 60% of the data) and a test sample (the remaining 344 documents). In order to give a robust classification result we repeat the experiment 100 times. We report in Table 3 the average error rate on the test sample and the standard error for each classifier. We can see that our metric clearly outperforms Mahalanobis distance. This is explained because we are dealing with highly dimensional data and few observations, therefore it is difficult to estimate an appropriate covariance matrix in order to adequately compute the Mahalanobis distance to the centers. Our distance does not suffer this inconvenience and is capable to approximate the classification performance of a variety of very sophisticated classification methods.

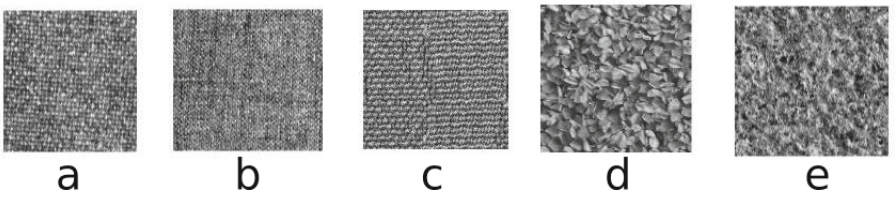

Fig. 3. Textures images: a) blanket, b) canvas, c) seat, d) linseeds and e) stone

Table 4. Outlier detection performance

Metric/Technique	% of: Outliers captured	False-positives (Type I error)	False-negatives (Type II error)
pc-Outlier[2]	60%	13.23%	28.65%
sign-Outlier[2]	40%	5.13%	37.75%
locoutPercent[2]	35%	2.80%	39.39%
Percentile 5% Euclidean Distance	25%	4.00%	42.85%
Percentile 5% Mah. Distance	35%	3.60%	39.39%
Percentile 5% Gen. Mah. Distance	**100**%	5.10%	**0.00**%

The second real data example considers the detection of outliers in sample of texture images. We consider the texture images from the Kylberg texture database [4]. We use 500 texture images with a resolution of 576×576 pixels. The first 480 texture images are very similar textures (Fig. 3 a) to c)). We also consider 20 "outliers" images with different textures (Fig. 3 d) and e)). We represent each image using the 32 parameters of the wavelet coefficient histogram proposed in [6]. We report the results in Table 4. Only the proposed distance is able to capture all the outliers in the sample. We also get an acceptable performance regarding the Type I Error rate (with 5.1%).

Future Work: The list of tasks for next future include an exhaustive simulation study of the performance of the proposed metric (some of this work is not included because the lack of space), the generalization of the proposed metric to define a Generalized "inter-point" Mahalanobis distance, and the study of properties of the proposed metric and its relations with the strictly convex and differentiable function ξ that originates the definition of the Bregman Divergences.

Acknowledgments. This work was partially supported by projects **MIC 2012/00084/00**, **ECO2012-38442**, **DGUCM 2008/00058/002** and **MEC 2007/04438/001**.

References

1. Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J.: Clustering with Bregman Divergences. Journal of Machine Learning Research, 1705–1749 (2005)
2. Filzmoser, P., Maronna, R.A., Werner, M.: Outlier identification in high dimensions. Computational Statistics & Data Analysis 52(3), 1694–1711 (2008)

3. Forster, J., Warmuth, M.K.: Relative Expected Instantaneous Loss Bounds. In: Annual Conference on Computational Learning Theory, pp. 90–99 (2000)
4. Kylberg, G.: The Kylberg Texture Dataset v. 1.0. In: Centre for Image Analysis. Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, http://www.cb.uu.se/
5. Mahalanobis, P.C.: On the generalised distance in statistics. In: Proceedings of the National Institute of Sciences of India, pp. 49–55 (1936)
6. Mallat, S.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans. on Pattern Analysis and Machine Intelligence 11(7), 674–693
7. Muñoz, A., Moguerza, J.M.: Estimation of High-Density Regions using One-Class Neighbor Machines. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(3), 476–480
8. Muñoz, A., Moguerza, J.M.: A Naive Solution to the One-Class Problem and Its Extension to Kernel Methods. In: Sanfeliu, A., Cortés, M.L. (eds.) CIARP 2005. LNCS, vol. 3773, pp. 193–204. Springer, Heidelberg (2005)
9. Smith, R.L.: Extreme value theory. In: Ledermann, W. (ed.) Handbook of Applied Mathematics, vol. 7, pp. 437–471 (1990)
10. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining 5(5), 363–387 (2012)