35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data

Carlos Sáez[1], Montserrat Robles[1], and Juan Miguel García-Gómez[1]

Abstract— **Research biobanks are often composed by data from multiple sources. In some cases, these different subsets of data may present dissimilarities among their probability density functions (PDF) due to spatial shifts. This, may lead to wrong hypothesis when treating the data as a whole. Also, the overall quality of the data is diminished. With the purpose of developing a generic and comparable metric to assess the stability of multi-source datasets, we have studied the applicability and behaviour of several PDF distances over shifts on different conditions (such as uni- and multivariate, different types of variable, and multi-modality) which may appear in real biomedical data. From the studied distances, we found information-theoretic based and Earth Mover's Distance to be the most practical distances for most conditions. We discuss the properties and usefulness of each distance according to the possible requirements of a general stability metric.**

I. Introduction

Research biobanks are often composed by data from multiple sources: different hospitals, health services, physicians, etc. A common research task consists in developing a hypothesis or model based in the whole set of multi-source data. However, dissimilarities in the probability density function (PDF) among the different subsets of data may complicate such research, lead to wrong hypothesis, or harm the further use of results on new data. In addition, detecting such dissimilarities may be difficult due to the heterogeneous conditions present in biomedical research data: (1) variables of different types (categorical, ordinal or not; and numerical, continuous or discrete), (2) data coming from uni-modal or multi-modal distributions, and (3) univariate or multivariate data. We classify the presence of such dissimilarities in PDFs as a problem in the stability of multi-source data, categorized as a data quality (DQ) problem in [15].

Providing accurate information about the data stability may help data managers and researchers to take decisions during the definition and development of research studies, as well as to feedback data providers about their acquisition procedures. In addition, a generic metric comparable among different studies, may provide a measurement of the degree of stability of multi-source biomedical data as a DQ metric.

In this work, we have studied the applicability and behaviour of several pairwise PDF distances on a set of simulations of data shifts based on the aforementioned biomedical data conditions. These pairwise distances provide stability information between pairs of sources. Hence, this study is the first stage towards the development of a global stability metric for any arbitrary number of sources, where pairwise PDF distances will serve as baseline measurements. We present the results of such comparative study as well as a discussion aimed to the next research steps.

II. Background

In [18], several studies on biomedical DQ are reviewed. Most focus on measuring DQ dimensions of a data repository as a whole, however, the concept of data source agreement is introduced aligned with the problem we are focusing. Besides, dataset shifts have also been related to DQ problems [6][15]. Dataset shifts are dissimilarities in the underlying distributions of data which can be originated through the course of time or across spatial factors. Our aim is to assign a distance to spatial dataset shifts among several sources of data, as a measurement of the overall data source agreement[1].

Most studies aim to detect dataset shifts in data streams, e.g. based on specific statistical tests [11] or distributional divergences [8]. Some of these approaches can be suited to obtain dissimilarity measures among the PDF of different data sources. Some works have also been published comparing PDF dissimilarity measures [12][4], although aimed to image retrieval. To the best of our knowledge, no similar comparisons have been carried out to assess the stability among biomedical data distributions, envisaging the multi-source, multivariate, multimodal and multi-type conditions, as well as the adequateness to a global stability metric.

III. Methods

A. Simulation

We evaluated the distances on a set of simulations to cover: (1) variable types, (2) multi-modality, and (3) dimensionality. We focused on numerical and categorical data, the most common post-processed research data, which facilitate the statistical analysis. In each simulation, two random datasets, (a) and (b), were defined following the same statistical distribution, where a null dissimilarity is expected. Then, we sequentially increased their dissimilarity until a predefined maximum state, where a maximum dissimilarity is expected. Distances were measured at each dissimilarity level.

We started evaluating the effect of shifts in different univariate variable types, covering (1) and (2). Simulation U1 consisted in a Normal $N(\mu, 1)$ continuous variable (cont.v.)

[1] C. Sáez, M. Robles and J.M. García-Gómez are with Grupo de Informática Biomédica (IBIME), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain carsaesi@ibime.upv.es

[1] The semantic compatibility among sources is out of the scope of this work, where multi-source biobanks are uniformly represented.

where dataset means $\mu^{(a)}$ and $\mu^{(b)}$ separated each other —e.g. due to an acquisition device that becomes biased. U2: $N(\mu,1)$ cont.v. where dataset b becomes bi-modal as a mixture of two Normal PDFs defined as $\frac{1}{2}\sum_{c=1,2}N(\mu_c^{(b)},1)$, which component means $\mu_1^{(b)}$ and $\mu_2^{(b)}$ symmetrically separate from the original —e.g. due to the appearance of a new pathological pattern. U3: Chi-squared $\chi^2(k)$ cont.v. where degrees of freedom $k^{(b)}$ separated from $k^{(a)}=0$ —e.g. due to an increase in the occurrence of a biomarker. U4: Binomial $B(1,p)$ ordinal categorical variable (cat.v.) where $p^{(a)}=p^{(b)}=0.5$ shifted to 0 and 1 respectively —e.g. due to variation in gender percentages in a diagnostic group. U5: Multinomial $Mult_3(1,p)$ non-ordinal cat.v. which priors shifted from an equal to a maximum difference state —e.g. due to a variation in the number of uses of treatments.

The multivariate simulations consisted in a combination of the previous variables, completing then (1), (2), and (3). M1: bivariate $N(\mu,1)$ cont.v. which means separated respectively. M2: bivariate $N(\mu 1)$ cont.v. where dataset b becomes multi-modal as a mixture which component means symmetrically separated from the original. M3: two $B(1,p)$ cat.v. where $p_1^{(a)}=p_2^{(a)}=p_1^{(b)}=p_2^{(b)}=0.5$ shifted to $p_1^{(a)}=p_2^{(a)}=1, p_1^{(b)}=p_2^{(b)}=0$. M4: a combination of a $N(\mu,1)$ cont.v. with a $B(1,p)$ cat.v. combining the shifts of U1 and U4.

B. Estimation of probability densities

To ensure the applicability to any non-parametric continuous PDF, we estimated empirical PDF histograms of the compared datasets using a Kernel-density smoothing method (or Parzen-window)[3], with Gaussian kernels and establishing the optimum bandwith based in [16]. Additionally, to homogenize the support, we estimated the common PDF from both datasets, and then, its bin centers were used as reference to interpolate the PDF of the independent datasets.

C. Studied distances

PDF distances measure how far two statistical distributions are in a metric space. A distance metric must be (I) non-negative, (II) zero only if the two compared distributions are the same (identity of indiscernibles), (III) symmetric, and (IV) must satisfy the triangle inequality. Divergences also provide a measure of dissimilarity, however, do not require to be symmetric nor satisfy the triangle inequality. Distances are then consistent with our purpose of a generic and comparable stability metric.

One type of studied distances included the statistics obtained in classical two-sample statistical hypothesis tests, including parametric: Student's t from t-test and Kolmogorov-Smirnov test statistic, and non-parametric: Kruskal-Wallis difference in mean ranks and the obtained χ^2 statistic from the Kruskal-Wallis test. We discarded the χ^2 test statistic for categorical data because it does not accomplish the identity of indiscernibles condition of a metric. Despite these type of distances are conceived for univariate[2] numerical data, we kept these tests for two reasons. First, we want to compare their behaviour in univariate multi-modal data. And second, dimensionality reduction of multivariate datasets may lead to a univariate sample making these methods feasible. Other advantage is that these statistics can be directly associated to p-values which permit significance tests on the differences.

We also studied information-theoretic based distances, which derive from Shanon's entropy theory [5], including the Jeffrey divergence and the square root of the Jensen-Shannon divergence, both symmetrized versions of the Kullback-Leibler divergence, the second also smoothed. We also studied the Hellinger distance, which can be defined as a metric version of the Bhattacharyya distance, commonly used in Pattern Recognition. These distances belong to the family of f-divergences[1][7], which measure the difference between PDFs. The main advantage of these metrics to the aforementioned statistics is that they apply to any type of binned PDF. Jeffrey and Jensen-Shannon, nevertheless, can not be measured when any of the PDFs has 0-probability bins —e.g. a categorical value not present in a source—, hence, in such cases an absolute discounting method was used to smooth the estimated PDFs.

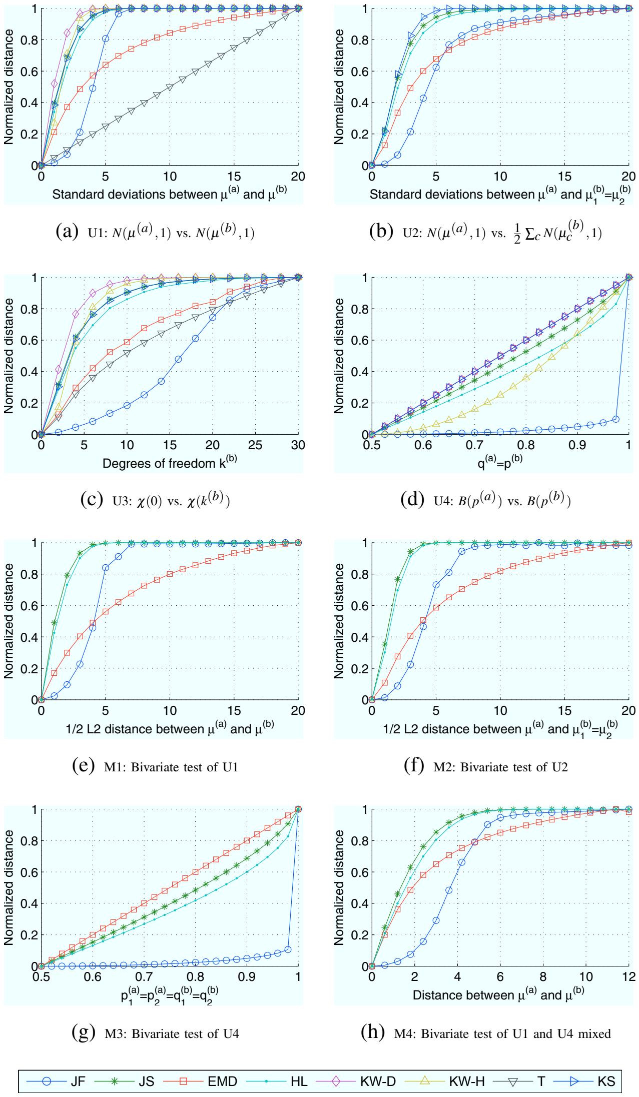
Finally, we studied the Earth Mover's Distance metric (EMD, a.k.a. Mallows or Wasserstein distance) [14]. EMD calculates the minimum cost required to transform one PDF into the other, using a predefined cost matrix of the probability mass flow between the bins in the support (ground distances). Originally conceived for image retrieval, in recent studies [2][9] EMD has been used to measure dissimilarities in multidimensional distributions. EMD envisages inter-bin information, in contrast to information-theoretic distances which make bin-by-bin comparisons, however, involves a higher computational cost. Additionally, EMD relaxes possible losses of information caused by binning, and permits defining custom cost matrices. To adapt the multivariate experiments to EMD algorithm, we embedded the two dimensions into one histogram using a normalized L1 ground distance matrix.

IV. Results

Figure 1 shows the results of the experiments. Each distance was normalized between zero and one to facilitate the comparison. As expected, in all simulations the evaluated distances behave monotonically increasing. In addition, all distances begin in 0, and we can observe that while most converge in continuous tests, these are approximately linear in discrete.

In experiment U1 non-parametric statistics behave similarly, converging around a distance between means of 4σ. Information-theoretic distances behave similarly, except Jeffrey divergence, which begins convex and converges when the tails of the PDFs leave each other. The t-test statistic

[2] Bivariate Kolmogorov-Smirnov test approaches [13] require further study since in d-dimensions imply 2^d-1 possible orderings. MANOVA tests entail a linear combination of the two or more normally-distributed dependent variables.



(a) U1: $N(\mu^{(a)},1)$ vs. $N(\mu^{(b)},1)$

(b) U2: $N(\mu^{(a)},1)$ vs. $\frac{1}{2}\sum_c N(\mu_c^{(b)},1)$

(c) U3: $\chi(0)$ vs. $\chi(k^{(b)})$

(d) U4: $B(p^{(a)})$ vs. $B(p^{(b)})$

(e) M1: Bivariate test of U1

(f) M2: Bivariate test of U2

(g) M3: Bivariate test of U4

(h) M4: Bivariate test of U1 and U4 mixed

Fig. 1. Results of univariate, (a), (b), (c) and (d), and multivariate, (e), (f), (g) and (h) experiments. JF: Jeffrey, JS: Jensen-Shannon, EMD: Earth Mover's Distance, HL: Hellinger, KW-D: Kruskal-Wallis mean rank difference, KW-H: Kruskal-Wallis statistic, T: t-test statistic, KS: Kolmogorov-Smirnov statistic.

behaves linearly, since we are separating two Normal PDFs with equal variance. The EMD resultant function also converges, but later.

In U2, t-test and Kruskal-wallis statistics were not able to capture the bi-modal shifting —despite the dissimilarity, sample means were the same— resulting in zero. The rest of distances behave equivalently to U1, capturing the bi-modality.

In U3, distances behave similarly to U1, but as it can be appreciated in t-test series, PDF means did not vary linearly with the shift in degrees of freedom.

Categorical simulations U4 and U5 resulted in equivalent results in information-theoretic and EMD distances. However, statistics distances were not applicable to U5, since non-ordinal categorical. Thus, we only show the results of U4, where the first are equivalent. Due to the linear shift in probability masses in these categorical experiments — in contrast to when separating Normal PDFs— none of the distances resulted convex. In addition, some captured this linear density shift resulting in linear functions. Despite the smoothing, we can observe in the Jeffrey distance series the tendency to infinite with smoothed 0-probability elements in the last iteration.

Results of multivariate experiments M1 and M2 are equivalent to their univariate relative U1 and U2, with the exception that statistic tests were not applicable. Thus, all distances converge, although EMD does later. Analogously, the results in Binomial experiment M3 are equivalent to those in U4. We can appreciate, however, slight differences in the results of mixed variable types experiment M4: while Jensen-Shannon and Hellinger distances seem to average the results of its independent continuous and categorical shifts, the EMD transformation cost seem to be slightly higher across the central iterations due to the abrupt density flow through the categorical dimension —we remember that EMD envisages inter-bin information.

V. Discussion

We come back to the studied conditions: (1) variable types, (2) multi-modality, and (3) dimensionality. Biomedical data can be considered heterogeneous and multi-modal by nature. Even univariate data may be formed by different 'natural' components, such as a mixture of healthy and different components of unhealthy parameters, or 'artificial' components, such as differences in the quality of data among their generating sources. Thus, an effective distance must be able to capture the dissimilarity in any of these conditions.

Regarding to the evaluation of (1), only information-theoretic and EMD are suited to any type of variable — statistics are only to numerical. Additionally, EMD is the only distance which permits setting specific costs to the difference between categories in unordered categorical data. Regarding to (2), t-test and Kruskal-Wallis had problems detecting multi-modality (U2, M2), however, Kolmogorov-Smirnov, information-theoretic and EMD were successful. Thus, despite the advantage of information-theoretic and EMD in (1) (and as we will see next, in (3)), Kolmogorov-Smirnov might still be used for obtaining a p-value on the difference in a continuous univariate variable resultant from a possible dimensionality reduction on multi-type data. At this point, information-theoretic and EMD distances seem the most practical for most situations. From these, we may consider the issue with null-probability elements of Jeffrey divergence a reason for discarding it. We can also observe that Jensen-Shannon and Hellinger are within a small constant each other [10]. Additionally, as we already mentioned, EMD is able to capture inter-bin information, and it is possible to define any cost between them, what may be useful in categorical data or when grouping PDF signatures [14].

Finally, regarding to the evaluation of (3), we already mentioned that statistics distances were not suited to multivariate data. In contrast, information-theoretic distances and EMD are theoretically suited to any number of dimensions.

However, direct estimation of PDFs in high-dimensional biobanks may be impractical due both to computational requirements and sparsity in the probabilistic space. Hence, dimensionality reduction methods may be applied to make feasible low-dimensional distances. For instance, we could reduce the dataset into a lower-dimensional statistical manifold. Additionally, in massive-data environments, we could represent groups of similar cases based in PDF signatures to facilitate the distance calculus.

On the other hand, results show that, in general, most distances have a convergence limit. They converge when the volume of the joint density between the two PDFs is minimized converging as well. However, EMD does later, what may suppose two advantages. First, it behaves approximately linear until the saturation level of those that converge first. And second, it can still express dissimilarity farther from this level. Furthermore, a bounded PDF support, e.g. in categorical data or bounded continuous, obviously entails a maximum limit in all the distances. Under these assumptions, we may choose between using the Jensen-Shannon, Hellinger or EMD, depending on the dissimilarity level at which we need the distance to converge.

To be generic, pairwise measurements should provide a dissimilarity level comparable across different datasets, or even different domains —imagine we wish to provide a stability mark in a DQ consulting. Jensen-Shannon, Hellinger, and Kolmogorov-Smirnov distances are bounded by definition between zero and one, what applies here (Kolmogorov-Smirnov, however, did not achieve its maximum value in the bimodal experiment (U2)). On the other hand, we noticed that the normalization applied to the EMD ground distance matrix, where a maximum cost of 1 is given when moving density between extreme bins, makes comparable the resultant transformation cost. This solution, however, requires predefining the possible support of all variables in order to identify the maximum inter-bin costs —equivalent to establishing the bounds of the probabilistic space.

We have not focused on other common types of biomedical data such as free text, signals or images. In some research tasks, a specific preprocessing may be used to obtain quantitative or qualitative measurements which will permit the use of the methods presented in this work. For instance, the Quantitative Magnetic Resonance (MR) methodology [17] is based on different quantitative parameters from brain MR images or MR spectroscopy signals, which may be used to assess the stability across radiology data sources.

VI. Conclusions

Providing information about the stability of biomedical research data among its sources may be of crucial importance. We have studied the behaviour and application of pairwise PDF distances on simulations of multi-type, multi-modal and multivariate conditions of biomedical data. Distances based in hypothesis contrast statistics are only suited to numerical univariate data, and have difficulties in multi-modality. Information-theoretic distances and EMD can handle multivariate, both continuous and discrete, and mixed types data. In general, all distances converge when the joint probability mass between the compared PDFs converges to the minimum, however EMD does later, what may provide more versatility in bounded supports. Additionally, EMD permits setting custom inter-bin costs. These results establish the basis for further studies of a general stability metric. Additionally, in further work we will generalise this study to real biomedical data studying the effect of dimensionality reduction methods on the PDF distances.

References

[1] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

[2] D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek. Unsupervised clustering of multidimensional distributions using earth mover distance. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 636–644, New York, NY, USA, 2011. ACM.

[3] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations (Oxford Statistical Science Series)*. Oxford University Press, USA, Nov. 1997.

[4] M. Budka, B. Gabrys, and K. Musial. On accuracy of pdf divergence estimators and their applicability to representative data sampling. *Entropy*, 13(7):1229–1266, 2011.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 99th edition, Aug. 1991.

[6] R. J. Cruz-Correia, P. Pereira Rodrigues, A. Freitas, F. Canario Almeida, R. Chen, and A. Costa-Pereira. Data quality and integration issues in electronic health records. *Information Discovery On Electronic Health Records*, pages 55–96, 2010.

[7] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[8] T. Dasu, S. Krishnan, D. Lin, S. Venkatasubramanian, and K. Yi. Change (detection) you can believe in: Finding distributional shifts in data streams. In *Proc. of the 8th Intl. Symp. on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII*, IDA '09, pages 21–34, Berlin, Heidelberg, 2009. Springer-Verlag.

[9] T. Dasu and J. M. Loh. Statistical distortion: consequences of data cleaning. *Proc. VLDB Endow.*, 5(11):1674–1683, July 2012.

[10] T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *Proc. of the 12th Intl. Workshop and 13th Intl. Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, APPROX '09 / RANDOM '09, pages 562–573, Berlin, Heidelberg, 2009. Springer-Verlag.

[11] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proc. of the 13th intl. conf. on Very large data bases - Volume 30*, VLDB '04, pages 180–191. VLDB Endowment, 2004.

[12] H. Liu, D. Song, S. Rüger, R. Hu, and V. Uren. Comparing dissimilarity measures for content-based image retrieval. In *Proc. of the 4th Asia information retrieval conf. on Information retrieval technology*, AIRS'08, pages 44–50, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] R. Lopes, I. Reid, and P. Hobson. The two-dimensional kolmogorov-smirnov test. In *XI Int. Workshop on Advanced Computing and Analysis Techniques in Physics Research, Nikhef, Amsterdam, the Netherlands, April 23-27, 2007*, 2007.

[14] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, Nov. 2000.

[15] C. Saez, J. Martinez-Miranda, M. Robles, and J. M. Garcia-Gomez. Organizing data quality assessment of shifting biomedical data. *Stud Health Technol Inform*, 180:721–725, 2012.

[16] H. Shimazaki and S. Shinomoto. Kernel bandwidth optimization in spike rate estimation. *J Comput Neurosci*, 29(1-2):171–182, Aug 2010.

[17] D. Wagnerova, V. Herynek, A. Malucelli, M. Dezortova, J. Vymazal, D. Urgosik, M. Syrucek, F. Jiru, A. Skoch, R. Bartos, M. Sames, and M. Hajek. Quantitative MR imaging and spectroscopy of brain tumours: a step forward? *Eur Radiol*, 22(11):2307–2318, Nov 2012.

[18] N. G. Weiskopf and C. Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*, 20(1):144–151, Jan 2013.