Mach Learn (2012) 86:369–389
DOI 10.1007/s10994-011-5267-2

Local equivalences of distances between clusterings—a geometric perspective

Marina Meilă

Received: 20 February 2006 / Accepted: 10 October 2011 / Published online: 17 December 2011

Abstract In comparing clusterings, several different distances and indices are in use. We prove that the Misclassification Error distance, the Hamming distance (equivalent to the unadjusted Rand index), and the χ^2 distance between partitions are equivalent in the neighborhood of 0. In other words, if two partitions are very similar, then one distance defines upper and lower bounds on the other and viceversa. The proofs are geometric and rely on the concavity of the distances. The geometric intuitions themselves advance the understanding of the space of all clusterings. To our knowledge, this is the first result of its kind.

Practically, distances are frequently used to compare two clusterings of a set of observations. But the motivation for this work is in the theoretical study of data clustering. Distances between partitions are involved in constructing new methods for cluster validation, determining the number of clusters, and analyzing clustering algorithms. From a probability theory point of view, the present results apply to any pair of finite valued random variables, and provide simple yet tight upper and lower bounds on the χ^2 measure of (in)dependence valid when the two variables are strongly dependent.

Keywords Clustering · Comparing partitions · χ^2 divergence · Misclassification error · Rand index · Convexity

1 Introduction and motivation

1.1 Why study distances?

In modern machine learning, there is a tendency to move from the perspective of the space where the data points lie, to that of the space of "learned functions." This change in paradigm accompanied a number of significant and lasting advances. Two such examples are kernel machines, whose development is tightly related to reproducing kernel

Editor: Carla Brodley.

M. Meilă (✉)
Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA
e-mail: mmp@stat.washington.edu

Hilbert spaces (RKHS) (Cortes and Vapnik 1995; Schölkopf and Smola 2002), and compressed sensing (Candès and Tao 2005; Donoho 2006), an algorithm for sparse regression, grounded in high dimensional vector spaces. In both cases the geometric intuitions about these spaces were instrumental in the discovery of the new techniques (Donoho 2006; Cortes and Vapnik 1995) and continue to be productive to this day.

For clustering, the natural space of "learned functions" is the space of all partitions of a set with n elements. Unlike the two examples above, which have intuitive, Euclidean or almost Euclidean geometries, well understood for a long time, the space of all partitions is much more challenging from the geometric point of view.

For instance, the space of partitions is not a vector space, which means that operations like "shift" and "rescaling" do not make sense for clusterings. Consequently, it does not admit a norm (while common vector spaces in machine learning admit the popular l_1, l_2, l_∞ norms) and so a clustering does not have a "magnitude". But it does admit several metrics, or distances, and in the absence of a norm, distances are the best means to analyze the geometric properties of clusterings.

Evaluating, comparing, predicting, and averaging distances are basic, ubiquitous mental operations when one reasons about clustering. For instance, one may want to know how accurate an algorithm can be, or how fast it converges. These are measured by the distance between the algorithm's output and an optimal clustering. Or, one may want to know how much variation in the result will be induced by randomness in the clustering algorithm. Again, since the result is a clustering, natural measure of variation is a distance between clusterings. Regarding clustering quality criteria (for instance, the quadratic distortion optimized by the k-means algorithm), one may want to know how fast they vary with the change in the partition, i.e., how "smooth" they are. This also requires a way to express the change in the partition, i.e., a distance. Finally, if one takes a statistical point of view and regards the data set itself as a sample from some distribution, one deals with averages and limits of such distances.[1]

Unfortunately, distances between partitions are both little studied and significantly less intuitive than the familiar l_p norms for other spaces of functions. This paper, which is a quantitative analysis of the relationships between several distances, represents a most basic result. Such results must exist before the more advanced results pertaining to algorithms, learning theory or statistics can be formulated.

1.2 The distances

Thus, we will be interested in distances $d(X, Y)$ between two clusterings X, Y of the same data set. A variety of different distances and indices[2] are in use today. While some work in understanding the properties of these distances and their relative merits exists, very little is known about how the values of various distances translate into each other. For instance, if we know the Rand index (Rand 1971) $r(X, Y)$ between two clusterings of a data set, can we evaluate from it the value of another index or distance, say the Misclassification Error distance $d_{ME}(X, Y)$?

With few exceptions, there is no one-to-one transformation between two different distances d, d' between clusterings. In other words, from the Rand index alone, we cannot compute the d_{ME} value exactly. But we can provide bounds on the range of values that $d_{ME}(X, Y)$ can take. This is what the present paper sets out to do.

[1] Usually over *different* spaces of partitions.

[2] An index $i(X, Y)$ is typically between 0 and 1, with 1 indicating identity of X with Y.

We will consider three distances between clusterings, defined in the next section: the Misclassification Error distance d_{ME}, the Hamming distance d_H (equivalent to the unadjusted Rand index), and the χ^2 distance d_{χ^2} and we will show that they are equivalent in the neighborhood of 0. In other words, as two clusterings X, Y become more similar to each other, all three distances will tend to 0, but at different rates. We establish these rates, by obtaining upper bounds on one distance, given another distance.

The Misclassification Error is widely used in the computer science literature on clustering. The Hamming distance is equivalent to the well known Rand index, and is popular in machine learning. The χ^2 distance originated in statistics. It is less used in practice but is a convenient vehicle for proofs.

Various properties of the Misclassification Error and of the Hamming distance, that are relevant to the task of comparing clusterings have been established and discussed in Meilă (2005). The three distances are defined in Sect. 2.

1.3 Equivalence, local equivalence, and a summary of the results

Two distances d and d' are called *equivalent* iff there exist constants $\underline{\beta}, \bar{\beta} > 0$ such that for any two clusterings X, Y, $\underline{\beta} d(X,Y) \leq d'(X,Y) \leq \bar{\beta} d(X,Y)$. If two distances are equivalent, then they behave essentially in the same way; for instance, d' can be approximated by d and viceversa, and if one distance is small, the other one cannot be too large. For finite-dimensional vectors, it is well known that all the norms are equivalent, and so are the distances derived from them.

As we shall see in Sect. 2, d_{ME}, d_H and d_{χ^2} are bounded respectively by 1, $\frac{1}{2}$ and $\sqrt{(K+K')/2}$, where K, K' are the number of clusters of the two clusterings. Thus, for fixed or bounded K, K', global equivalence is trivial.

In this paper we are concerned with the property of *local equivalence* which is weaker than equivalence in two respects: (a) it holds only locally, when the distances are small, or (b) the constants $\underline{\beta}, \bar{\beta}$ depend on certain properties of the clusterings X and Y, and thus they vary over the space of all partitions. As we shall see, by choosing this framework we will obtain finer grained relationships between the distances.

The next table summarizes the results obtained, and indicates in which section they are presented; the quantities p_{min}, p_{max} and β, β' are defined in the respective sections. The results are followed by a discussion and conclusions contained in Sect. 8.

Section	Relation	Global?	Proof approach
3 (Theorem 9)	$d^2_{\chi^2} \geq \frac{1}{p_{max}} d_{ME}$	no	convexity, extreme points
4 (Theorem 19)	$d^2_{\chi^2} \leq \frac{2}{p_{min}} d_{ME}$	yes[a]	convexity, first order definition
5 (Theorem 26)	$d_H \leq 4 p_{max} d_{ME}$	yes[a]	convexity, first order definition
6 (Theorem 27)	$d_{ME} \leq \frac{1}{2p_{min}} d_H$	no	convexity, extreme points
7 (Theorem 28)	$d_H \leq 4 p^2_{max} d^2_{\chi^2}$	no	Theorems 19 and 27
	$d_H \geq p^2_{min} d^2_{\chi^2}$	no	Theorems 9 and 26
7 (Theorem 29)	$d^2_{\chi^2} \leq \beta d_H + \beta'$	yes	matrix calculus
	$d^2_{\chi^2} \geq \beta d_H - \beta'$	yes	matrix calculus

[a]Restricted to $K = K'$ or $K \leq K'$

2 Definitions and representation

We consider a finite set $\mathcal{D}_n$ with n elements. A *clustering* is a *partition* of $\mathcal{D}_n$ is into sets $C_1, C_2, \ldots, C_K$ called *clusters* such that

$$C_k \cap C_l = \emptyset \quad \text{and} \quad \bigcup_{k=1}^{K} C_k = \mathcal{D}_n.$$

Let the cardinality of cluster C_k be n_k. We have, of course, that $n = \sum_{k=1}^{K} n_k$. We assume that $n_k > 0$; in other words, that K represents the number of non-empty clusters.

Representing clusterings as matrices Without loss of generality the set $\mathcal{D}_n$ can be taken to be $\{1, 2, \ldots, n\} \stackrel{def}{=} [n]$. Denote by X a clustering $\{C_1, C_2, \ldots, C_K\}$; X can be represented by the $n \times K$ matrix A_X with $A_{ik} = 1$ if $i \in C_k$ and 0 otherwise. In this representation, the columns of A_X are indicator vectors of the clusters and are orthogonal.

Representing clusterings as random variables The clustering X can also be represented as the random variable (denoted abusively by) $X : [n] \longrightarrow [K]$ taking value $x \in [K]$ with probability $\frac{n_x}{n}$. One typically requires distances between two partitions to be invariant to the permutations of the labels $1, \ldots, K$. By this representation, any distance between two clusterings can be seen as a particular type of distance between random variables that is invariant to permutations.

Let a second clustering of $\mathcal{D}_n$ be $Y = \{C'_1, C'_2, \ldots, C'_{K'}\}$, with cluster sizes n'_y. Note that the two clusterings may have different numbers of clusters.

Lemma 1 *The joint distribution of variables X, Y is given by*

$$p_{XY} = \frac{1}{n} A_X^T A_Y \tag{2.1}$$

In other words, $p_{XY}(x, y)$ is the x, y-th element of the $K \times K'$ matrix in (2.1).

In the above, the superscript $()^T$ denotes matrix transposition. The proof is immediate and is left to the reader. We now define the three distances between two clusterings in terms of the joint probability matrix defined above.

Definition 2 The misclassification error distance d_{ME} between clusterings X, Y (with $K \leq K'$) is

$$d_{ME}(X, Y) = 1 - \max_{\pi \in \Pi_{K'}} \sum_{x \in [K]} p_{XY}(x, \pi(x))$$

where $\Pi_{K'}$ is the set of all permutations of K' objects represented as mappings $\pi : [K'] \to [K']$.

Although the maximization above is over a set of $(K')!$ permutations, d_{ME} can be computed in polynomial time by a maximum bipartite matching algorithm (Papadimitriou and Steiglitz 1998). It can be shown that d_{ME} is a metric (see e.g., Meilă 2005). This distance is widely used in the computer science literature on clustering, due to its direct relationship

with the misclassification error cost of classification. It has indeed very appealing properties as long as $d_{ME}(X, Y)$ takes small values (i.e., the clusterings are "close") (Meilă 2007). Otherwise, its poor resolution (Meilă 2007) represents a major hindrance.

It can be seen that d_{ME} is always smaller than 1. The bound 1 is never attained, but is approached arbitrarily closely. For example, between the clustering with a single cluster and the clustering with n singleton clusters the Misclassification Error distance is $1 - \frac{1}{n}$.

Definition 3 The χ^2 distance d_{χ^2} is defined as

$$d^2_{\chi^2}(X, Y) = \frac{K + K'}{2} - \chi^2(p_{XY})$$

with

$$\chi^2(p_{XY}) = \sum_{x,y} \frac{p_{XY}(x, y)^2}{p_X(x)p_Y(y)} \tag{2.2}$$

The above definition and notation are motivated as follows.

Lemma 4 *Let $p_X = (p_x)_{x\in[K]}$, $p'_Y = (p'_y)_{y\in[K']}$ be the marginals of p_{XY}. Then, the function $\chi^2(p_{XY})$ defined in (2.2) represents the functional $\chi^2(f, g) + 1$ applied to $f = p_{XY}, g = p_X p'_Y$.*

Proof Denote $p_{xy} = p_{XY}(x, y)$. By the definition of Lancaster (1969),

$$\chi^2(f, g) = \sum_{xy} \frac{(p_{xy} - p_x p'_y)^2}{p_x p'_y} = \sum_{xy} \left[\frac{p^2_{xy}}{p_x p'_y} - 2p_{xy} + p_x p'_y \right] = \sum_{xy} \frac{p^2_{xy}}{p_x p'_y} - 2 + 1 \qquad \square$$

Hence, $d^2_{\chi^2}$ is a measure of independence. It is equal to 0 when the random variables X, Y are identical up to a label permutation, and to $(K + K')/2$ when they are independent. Once can also show that $d^2_{\chi^2}$ is a squared metric (Bach and Jordan 2006) and for completeness this result will be included in a lemma to follow shortly.

The d_{χ^2} distance with slight variants has been used as a distance between partitions by Hubert and Arabie (1985), Bach and Jordan (2006) with the obvious motivation of being related to the familiar χ^2 functional. The following definition and lemma give another, technical motivation for paying attention to d_{χ^2}.

Definition 5 The normalized matrix representations for X is defined by $\tilde{A}_X(i, k) = \frac{1}{\sqrt{n_k}}$ if $i \in C_k$ and 0 otherwise.

The columns of $\tilde{A}_X$ have thus unit length, and this representation has orthonormal columns, being an *orthogonal* matrix.

Lemma 6 (Bach and Jordan 2006) *Let $\|\|_F$ represent the Frobenius norm. Then*

$$\chi^2(p_{XY}) = \|\tilde{A}^T_X \tilde{A}_Y\|^2_F \tag{2.3}$$

and

$$d^2_{\chi^2}(p_{XY}) = \|\tilde{A}_X \tilde{A}^T_X - \tilde{A}_Y \tilde{A}^T_Y\|^2_F \tag{2.4}$$

Proof To prove (2.3) note that $(\tilde{A}_X^T \tilde{A}_Y)_{xy} = \frac{p_{xy}}{\sqrt{p_x p'_y}}$. To prove the second equality, note that $\|\tilde{A}_X^T \tilde{A}_X\|_F^2 = K$, $\|\tilde{A}_Y^T \tilde{A}_Y\|_F^2 = K'$. Then, we use the identity $\|A\|_F^2 = \text{trace} A^T A$ and basic properties of the matrix trace. □

The above lemma shows that $d^2_{\chi^2}$ is a quadratic function, making it a convenient instrument in proofs. Contrast this with the apparently simple d_{ME} distance, which is not everywhere differentiable and is theoretically much harder to analyze.

A third distance between partitions, which has a long history, is the distance known under the names of *Hamming distance* (Ben-David et al. 2006), *Rand index* (Rand 1971), or *Mirkin metric* (Mirkin 1996). The three names refer to slightly different forms of the same criterion for comparing partitions.

Definition 7 The Hamming distance d_H between clustering X, Y is defined as

$$d_H(X, Y) = \frac{1}{2n^2} \|A_X A_X^T - A_Y A_Y^T\|_F^2 \tag{2.5}$$

Because A_X, A_Y are $\{0, 1\}$ matrices representing clusterings, $A_X A_X^T$, $A_Y A_Y^T$ are also $\{0, 1\}$ matrices, and the Frobenius norm on the r.h.s of (2.5) counts the positions in which they differ. Hence, d_H represents the Hamming distance between the matrices $A_X A_X^T$, $A_Y A_Y^T$. Note the strong similarity with the expression of $d^2_{\chi^2}$ in (2.4), which shows that $\sqrt{d_H}$ is also a metric.

Other interpretations and variants of this distance are given by the following lemma.

Lemma 8

1. *The Hamming distance is the probability of the event "i, j are in the same cluster under X but in different clusters under X' or viceversa" when the two points $i, j \in [n]$ are picked uniformly and independently.*
2. *The Mirkin metric* (Mirkin 1996) *is defined as*

$$d_{Mirkin}(X, Y) = \sum_{x \in [K]} n_x^2 + \sum_{y \in [K']} {n'_y}^2 - 2 \sum_{x \in [K]} \sum_{y \in [K']} n_{xy}^2 \tag{2.6}$$

$$d_H(X, Y) = \frac{1}{2n^2} d_{Mirkin}(X, Y) \tag{2.7}$$

3. *The Rand index defined in* Rand (1971) $r(X, Y)$ *is given by*

$$r(X, Y) = 1 - \frac{d_{Mirkin}(X, Y)}{n(n-1)} \tag{2.8}$$

Proof (1) This probabilistic interpretation of the Hamming distance was put forward in Rand (1971) and later in Ben-David et al. (2006).

(2) One can easily verify that $\|A_X^T A_X\|_F^2 = \sum_{x \in [K]} n_x^2$, $\|A_Y^T A_Y\|_F^2 = \sum_{y \in [K']} {n'_y}^2$, $\|A_X^T A_Y\|_F^2 = \sum_{x \in [K]} \sum_{y \in [K']} n_{xy}^2$ which shows that

$$d_H(X, Y) = \frac{1}{2} \sum_{x \in [K]} p_x^2 + \frac{1}{2} \sum_{y \in [K']} {p'_y}^2 - \sum_{x \in [K]} \sum_{y \in [K']} p_{xy}^2 \tag{2.9}$$

(3) This was proved in Meilă (2007). □

Moreover, the Hamming distance is bounded above by $\frac{1}{2}$, which can be seen from (2.5) because the number of elements of the matrices involved is n^2, and all their entries are 0 or 1. The value $\frac{1}{2}$ is approached asymptotically, as one can verify by calculating the distance between the clustering with a single cluster and the clustering with n singletons.

We close this section by noting that the functions d_{ME}, $d^2_{\chi^2}$, and d_H are concave in p_{XY}. For $d^2_{\chi^2}$, this follows from the convexity of the χ^2 functional (Vajda 1989). The d_{ME} can be expressed as the minimum of a set of linear functions[3]; therefore it is concave. The concavity of d_H is proved in Sect. 5.

The remaining sections prove the local equivalences between the three distances, in the following sequence: $d^2_{\chi^2}$ upper bounds d_{ME} in Sect. 3, d_{ME} upper bounds $d^2_{\chi^2}$ in Sect. 4, d_{ME} upper bounds d_H in Sect. 5, d_H upper bounds d_{ME} in Sect. 6. A slightly different kind of relation between d_H and $d^2_{\chi^2}$ is proved in Sect. 7. The paper concludes with a discussion of the results (Sect. 8).

3 Small d_{χ^2} implies small d_{ME}

This is the first of the bounds in the paper. We first state the result precisely, then describe the geometric intuition underlying it. We also establish a framework for the proof approach. This framework is shared by the proofs in Sects. 4, 5 and 6.

Theorem 9 *For two clusterings with the same number of clusters K represented by the joint distribution p_{XY}, denote $p_{min} = \min_{[K]} p_x$, $p_{max} = \max_{[K]} p_x$. Then, for any $\varepsilon \le p_{min}$, if $d^2_{\chi^2}(p_{XY}) \le \frac{\varepsilon}{p_{max}}$ then $d_{ME}(p_{XY}) \le \varepsilon$.*

Before we embark on the proof, we give an example where $d^2_{\chi^2}/d_{ME}$ is arbitrarily close to this bound.

Example 10 Consider the following p_{XY}, with $K = K'$.

$\frac{1}{K} - \frac{1}{n}$	$\frac{1}{n}$		$\cdots$
	$\frac{1}{K}$		$\cdots$
		$\frac{1}{K}$	$\cdots$
			$\cdots$

$$d_{ME} = \frac{1}{n} \quad \text{and} \quad d^2_{\chi^2} = \frac{K}{n} - \frac{2}{n^2/K^2 - 1}$$

Hence, $d^2_{\chi^2}/d_{ME}$ approaches $K = 1/p_{max}$.

Geometric ideas All the bounds in this paper with the exception of Theorem 29 are based on the concavity of the respective distances. The proofs make use of two geometric facts about concave functions. The first is that a concave function attains its minimum at an extreme point of its domain. In our case the domain is a (convex) set of joint probability distributions p_{XY} (that will be defined below) and the minimum value of 0 is attained at multiple "corners" of this domain. Therefore, we expect all distances (i.e., $d_{ME}, d^2_{\chi^2}, d_H$) to be small near these corners and large far away from them. This is the crucial idea of the proof of Theorem 9, reiterated in Theorem 27. A second fact is that a concave function is always below any tangent to its graph. This will be the main approach in the proofs of Theorems 19 and 26.

[3] d_{ME} = minimum of the off-diagonal mass of p_{XY} over all permutations.

Proof outline First we introduce some basic notation that will be used for the rest of the paper. For any distribution p_{XY}, we denote by $\bar{\bar{p}}$ the table of values of this distribution, by p_{xy} the probability of pair $(x, y) \in [K] \times [K']$ under p_{XY} (i.e., an entry in $\bar{\bar{p}}$), and by $p_X = (p_1 \dots p_K)$, $p_Y = (p'_1 \dots p'_K)$ respectively the X and Y marginals of p_{XY} (or equivalently of $\bar{\bar{p}}$). As a matter of usage, the p_{XY} notation will be used in the statements of the main theorems, while the $\bar{\bar{p}}$ notation will be used in the proofs and minor results. This dual notation corresponds to viewing a pair of clusterings as a distributions p_{XY} in the statements of the theorems, but viewing the same pair as a point $\bar{\bar{p}}$ in the $K \times K'$ space while proving the theorems.

We adopt the following framework, which will also be common to all proofs. We will assume without loss of generality that partition X is fixed, while Y is allowed to vary. In terms of random variables, the assumption describes the set of distributions over $[K] \times [K']$ that have a fixed marginal $p_X = (p_1, \dots, p_K)$. We denote this domain by $\mathcal{P}$. Thus, $\mathcal{P} = \{\bar{\bar{p}} = [p_{xy}]_{x \in [K], y \in [K']}, p_{xy} \geq 0, \sum_y p_{xy} = p_x \text{ for } x \in [K]\}$, a convex and bounded set. Note also that since X is fixed, each $\bar{\bar{p}}$ corresponds to one or more clusterings Y; thus we will sometimes speak of clusterings (Y) when we refer to points in $\mathcal{P}$. Note also that our setting is slightly more general than needed by Theorem 9. Indeed, some of the intermediate results we prove do not require that $K' = K$.

For this particular theorem, the intermediate results are simpler in terms of the (convex) χ^2 function; the reader will keep in mind that $d^2_{\chi^2} = K - \chi^2$ by (2.2).

We will show that the maxima of χ^2 over $\mathcal{P}$ have value K and are attained when the second random variable is a one-to-one function of the first (note that these correspond to the minima of $d^2_{\chi^2}$ which are 0). We call such a point *optimal*; the set of optimal points of $\mathcal{P}$ is denoted by E^*. Any element $\bar{\bar{p}}^*$ in E^* is defined as:

$$p^*_{xy} = \begin{cases} p_x & \text{if } y = \pi(x) \\ 0 & \text{otherwise} \end{cases}$$

where π represents a permutation of the indices $1, 2, \dots, K$.

We prove that if a joint distribution $\bar{\bar{p}}$ in $\mathcal{P}$ is more than ε away from any optimal point, then $\chi^2(\bar{\bar{p}})$ will be bounded away from K. A schematic description of the proof outline and underlying geometry is given in Fig. 1.

For a fixed π, we denote the corresponding optimal point by $\bar{\bar{p}}^*_\pi$ and the points which differ from $\bar{\bar{p}}^*_\pi$ by ε in p_{aa}, p_{ab} by $\bar{\bar{p}}_{\varepsilon,\pi}(a, b)$. We shall see that the regions where d_{ME} is small/large are defined by these points. Below is the definition of $\bar{\bar{p}}_{\varepsilon,\pi}(a, b)$ in the case of the identical permutation. In what follows, whenever we consider one optimal point only, we shall assume without loss of generality that π is the identical permutation, and omit it from the notation.

$$[\bar{\bar{p}}_\varepsilon(a, b)]_{xy} = \begin{cases} \varepsilon & x = a, y = b \\ p_a - \varepsilon, & x = y = a \\ p_x, & x = y \neq a \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

and thus

$$[\bar{\bar{p}}^* - \bar{\bar{p}}_\varepsilon(a, b)]_{xy} = \begin{cases} \varepsilon, & x = y = a \\ -\varepsilon & x = a, y = b \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

- the square represents $\mathcal{P}$
- its corners are $E^* = \{\bar{\bar{p}}^*_\pi\}$ (permutations of X)
- the • dots are $\bar{\bar{p}}^\pi_\varepsilon$, special clusterings at exactly ε from a $\bar{\bar{p}}^*$
- the crosshatched regions near a $\bar{\bar{p}}^*$ are clusterings for which $d_{ME}(\bar{\bar{p}}) \le \varepsilon$ (Lemma 13); if $\varepsilon \le p_{min}$ these regions are disjoint
- the white central region is A, the convex hull of the • points (Lemma 15)
- at the square corners $\chi^2 = K$, the maximum value (Lemma 11)
- at • $\chi^2 \le K - \varepsilon/p_{max}$ (3.5)
- on A, by convexity, $\chi^2(\bar{\bar{p}}) \le K - \varepsilon/p_{max}$ (Proof of Theorem 17)
- therefore, if $d^2_{\chi^2}(\bar{\bar{p}}) \le \varepsilon/p_{max}$, then $\bar{\bar{p}}$ must be in the complement of A, the crosshatched regions where $d_{ME}(\bar{\bar{p}}) \le \varepsilon$

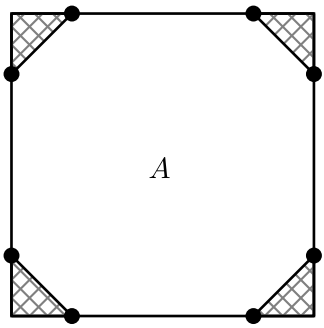

Fig. 1 Simplified geometric view of the proof of Theorem 9. Recall that because clustering X is fixed, each $\bar{\bar{p}} \in \mathcal{P}$ represents one or more clusterings Y

For $\varepsilon \le p_{min} = \min_x p_x$ let $E^\pi_\varepsilon = \{\bar{\bar{p}}_{\varepsilon,\pi}(a,b), a, b \in [K] \times [K'], a \ne b\}$. We lower bound the value of χ^2 at all points in $E_\varepsilon = E^{identity}_\varepsilon$. We then show that if d_{ME} is greater than ε, the value of χ^2 cannot be lower than the aforementioned lower bound.

These results will be proved as a series of lemmas, after which the formal proof of the theorem will close this section. Figure 1 shows a schematic walk-through the lemmas that follow.

The first result says that the extreme points of $\mathcal{P}$ are the clusterings Y that do not break up the clusters in X.

Lemma 11

1. *The set of extreme points of* $\mathcal{P}$ *is*

$$E = \{\bar{\bar{p}} \mid \exists \phi : [K] \longrightarrow [K'], p_{xy} = p_x \text{ if } y = \phi(x), 0 \text{ otherwise}\}$$

2. *For* $\bar{\bar{p}} \in E$, $\chi^2(\bar{\bar{p}}) = |\text{Range } \phi|$.

Proof The proof of part 1 is immediate and left to the reader. To prove part 2, let $\bar{\bar{p}} \in E$. We can write successively

$$\chi^2(\bar{\bar{p}}) = \sum_y \sum_{x \in \phi^{-1}(y)} \frac{p_x^2}{p_x \sum_{z \in \phi^{-1}(y)} p_z}$$

$$= \sum_y \frac{\sum_{x \in \phi^{-1}(y)} p_x}{\sum_{z \in \phi^{-1}(y)} p_z} = \sum_y 1 = |\text{Range } \phi|$$

□

If $\text{Range}(\phi) = K$, then ϕ is a permutation and we denote it by π; $E^* = \{\bar{\bar{p}}^*_\pi\}$ is the set of extreme points for which $\chi^2 = K$ and $E^- = E \setminus E^*$ the set of the extreme points for which $\chi^2 = K' \le K - 1$. Hence E^- contains the clusterings Y that join several clusters of X and

E^* the clusterings identical to X (up to a relabeling of the clusters). Note also that E^- is non-empty only when $K' < K$ and that for $K' > K$ no additional extreme points are created.

The second step, in two lemmas, describes the regions near the optimal clusterings, where d_{ME} is small. These are the crosshatched corners in Fig. 1.

Lemma 12 *Let $B_1(r)$ be the 1-norm ball of radius r centered at $\bar{\bar{p}}^* \in E^*$. For all $\bar{\bar{p}} \in B_1(2\varepsilon) \cap \mathcal{P}$*

$$d_{ME}(\bar{\bar{p}}) \leq \varepsilon$$

Proof For a point $\bar{\bar{p}} \in B_1(2\varepsilon) \cap \mathcal{P}$ let e be defined as

$$e = \sum_x \sum_{y \neq x} p_{xy} \tag{3.3}$$

Note that $\|\bar{\bar{p}}^* - \bar{\bar{p}}\|_1 = 2e$. Now it is obvious that $d_{ME}(\bar{\bar{p}}) \leq \sum_x \sum_{y \neq x} p_{xy} = e \leq \varepsilon$. □

Lemma 13

$$B_1(2\varepsilon) \cap \mathcal{P} = \text{convex}(\{\bar{\bar{p}}^*\} \cup E_\varepsilon)$$

Proof First we show that $\|\bar{\bar{p}}^* - \bar{\bar{p}}_\varepsilon(a,b)\|_1 = 2\varepsilon$.

$$\begin{aligned}\|\bar{\bar{p}}^* - \bar{\bar{p}}_\varepsilon(a,b)\|_1 &= \sum_{x,y} |p^*_{xy} - p_\varepsilon(a,b)_{xy}| = |p^*_{aa} - p_\varepsilon(a,b)_{aa}| + |p^*_{ab} - p_\varepsilon(a,b)_{ab}| \\ &= \varepsilon + \varepsilon = 2\varepsilon\end{aligned}$$

Then, it is easy to check (with e defined in (3.3)) that

$$\bar{\bar{p}} = \left(1 - \frac{e}{\varepsilon}\right)\bar{\bar{p}}^* + \sum_a \sum_{b \neq a} \frac{p_{ab}}{\varepsilon} \bar{\bar{p}}_\varepsilon(a,b)$$

and

$$\left(1 - \frac{e}{\varepsilon}\right) + \sum_a \sum_{b \neq a} \frac{p_{ab}}{\varepsilon} = 1$$

□

Next, we focus on the region (denoted by A) where d_{ME} is large. In the following two results we characterize it and show that it is convex.

Lemma 14 *Let $x = \sum_i \alpha_i x_i$ with $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$ and, for all i, let y_i be a point of the segment $[x, x_i)$. Then x is a convex combination of $\{y_i\}$.*

Proof Let $y_i = \beta_i x + (1 - \beta_i)x_i$, $\beta \in [0, 1)$. Then $x_i = \frac{y_i - \beta_i x}{1 - \beta_i}$ and replacing the above in the expression of x we get successively

$$x = \sum_i \left[\frac{\alpha_i}{1 - \beta_i} y_i - \frac{\alpha_i \beta_i}{1 - \beta_i} x\right] = \sum_i \frac{\alpha_i}{1 - \beta_i} y_i - x \sum_i \frac{\alpha_i \beta_i}{1 - \beta_i}$$

and

$$x = \sum_i \underbrace{\frac{\frac{\alpha_i}{1-\beta_i}}{1+\sum_j \frac{\alpha_j\beta_j}{1-\beta_j}}}_{\gamma_i} y_i \quad \text{with } \gamma_i \geq 0 \text{ and } \sum_i \gamma_i = \frac{\sum_i \frac{\alpha_i}{1-\beta_i}}{1+\sum_j \frac{\alpha_j\beta_j}{1-\beta_j}} = \frac{1+\sum_i \frac{\alpha_i\beta_i}{1-\beta_i}}{1+\sum_j \frac{\alpha_j\beta_j}{1-\beta_j}} = 1.$$

□

Lemma 15 *The set* $\{\bar{\bar{p}} | d_{ME}(\bar{\bar{p}}) \geq \varepsilon\}$ *with* $\varepsilon \leq p_{min}$ *is included in the convex hull of* $\{E^\pi_\varepsilon\}_{\Pi_K} \cup E^-$.

Proof Let $A = \{d_{ME}(\bar{\bar{p}}) \geq \varepsilon\}$ and $\bar{\bar{p}} \in A$. Because $\bar{\bar{p}} \in \mathcal{P}$ is a convex combination of the extreme points of $\mathcal{P}$, it can be written as

$$\bar{\bar{p}} = \sum_{i=1}^{K!} \alpha_i \bar{\bar{p}}^*_{\pi_i} + \sum_{i=1}^{|E^-|} \alpha_{i+K!} \bar{\bar{p}}^-_i, \quad \alpha_i \geq 0, \sum_i \alpha_i = 1$$

where by $\bar{\bar{p}}^-_i$ we have denoted the points in E^-. Let us look at the segment $[\bar{\bar{p}}, \bar{\bar{p}}^*_{\pi_i}]$; its first end, $\bar{\bar{p}}$ is in A, while its other end is outside A and inside the ball $B_1^{\pi_i}(\varepsilon)$. As the ball is convex, there is a (unique) point $\bar{\bar{p}}_i = [\bar{\bar{p}}, \bar{\bar{p}}^*_{\pi_i}] \cup \partial B_1^{\pi_i}(\varepsilon)$. This point being on the boundary of the ball, it can be written as a convex combination of points in $E^{\pi_i}_\varepsilon$ by Lemma 13. We now apply Lemma 14, with $x_i = \bar{\bar{p}}^*_{\pi_i}$ and $y_i = \bar{\bar{p}}_i$ for $i = 1, \ldots, K!$ and $x_i = y_i = \bar{\bar{p}}^-_{i-K!}$ for $i > K!$. It follows that $\bar{\bar{p}}$ is a convex combination of $\bar{\bar{p}}_i, i = 1, \ldots, |E|$, which completes the proof.[4] □

The last step is to look at the extreme points of A from the point of view of χ^2 and show that its values are bounded away from the optimal value K.

Lemma 16 *For* $\varepsilon \leq p_{min}$

$$\chi^2(\bar{\bar{p}}^*) - \chi^2(\bar{\bar{p}}_\varepsilon(a,b)) \geq \frac{\varepsilon}{p_{max}}$$

Proof Compute $\chi^2(\bar{\bar{p}}_\varepsilon(a,b))$:

$$\begin{aligned}
\chi^2(\bar{\bar{p}}_\varepsilon(a,b)) &= K - 2 + \frac{(p_a-\varepsilon)^2}{p_a(p_a-\varepsilon)} + \frac{\varepsilon^2}{p_a(p_b+\varepsilon)} + \frac{p_b^2}{p_b(p_b+\varepsilon)} \\
&= K - 2 + 1 - \frac{\varepsilon}{p_a} + \frac{\varepsilon^2}{p_a(p_b+\varepsilon)} + 1 + \frac{\varepsilon}{p_b+\varepsilon} \\
&= K - \frac{\varepsilon(p_a+p_b)}{p_a(p_b+\varepsilon)} \qquad (3.4) \\
&\leq K - \frac{\varepsilon}{p_a} \qquad (3.5)
\end{aligned}$$

Therefore

$$\chi^2(\bar{\bar{p}}^*) - \chi^2(\bar{\bar{p}}_\varepsilon(a,b)) \geq \frac{\varepsilon}{p_a} \geq \frac{\varepsilon}{p_{max}}$$

□

We are now ready to prove a result relating χ^2 and d_{ME} that holds for any K, K'.

[4] In fact, it can be easily shown (left to the reader) that A *equals* the convex hull of $\{E^\pi_\varepsilon\}_{\Pi_K} \cup E^-$.

Theorem 17 *For two clusterings X, Y with number of clusters K, respectively K', represented by the joint distribution p_{XY}, denote $p_{min} = \min_{[K]} p_x$, $p_{max} = \max_{[K]} p_x$. Then, for any $\varepsilon \le p_{min}$, if $\chi^2(p_{XY}) \ge K - \frac{\varepsilon}{p_{max}}$ then $d_{ME}(p_{XY}) \le \varepsilon$.*

Proof of Theorem 17 By contradiction. Assume $d_{ME}(\bar{\bar{p}}) \ge \varepsilon$. Then, $\bar{\bar{p}} \in A$ by Lemma 15. Because χ^2 is convex on A, $\chi^2(\bar{\bar{p}})$ cannot be larger than the maximum value at the extreme points of A, which are contained in $E^- \cup (\bigcup_\pi E_\varepsilon^\pi)$. But we know by Lemma 16 that the value of χ^2 is bounded above by $K - \varepsilon/p_{max}$ at any point in E_ε^π and equals K' at any point in E^-. Also E^- is not empty only when $K' \le K - 1$.

Note also that a tight, non-linear bound can be obtained by maximizing (3.4) over all a, b. □

Proof of Theorem 9 The theorem now follows from Theorem 17 when we set $K = K'$, since in this case $d_{\chi^2}^2(\bar{\bar{p}}) = K - \chi^2(\bar{\bar{p}})$. □

It is interesting to see what happens if $K \ne K'$. Assume first that $K' > K$. We can write

$$d_{\chi^2}^2(p_{XY}) = \frac{K' - K}{2} + \big(K - \chi^2(p_{XY})\big).$$

Therefore, if $d_{\chi^2}^2(p_{XY}) \le \frac{K-K'}{2} + \frac{\varepsilon}{p_{max}}$ then $d_{ME}(p_{XY}) \le \varepsilon$.

Now assume $K' < K$. In this case $\chi^2(p_{XY}) \le K'$, but the values of $K' - \chi^2(p_{XY})$ do not bound d_{ME}, as can be seen from the following example.

Example 18 $K > 2$, $K' = 2$ and $p_1 = p_2 = \cdots = \frac{1}{K}$.

$\frac{1}{K}$	0
0	$\frac{1}{K}$
0	$\frac{1}{K}$
...	...

$$\chi^2 = K' = 2 \quad \text{and} \quad d_{ME} = \frac{K-2}{K}$$

Hence, χ^2 is equal to its maximum while d_{ME} can become arbitrarily close to 1.

4 Small d_{ME} implies small d_{χ^2}

This is the converse bound to the bound in the previous section. Together, the two results prove the local equivalence between d_{ME} and d_{χ^2}.

Theorem 19 *Let p_{XY} represent a pair of clusterings with the same number of clusters. Then*

$$d_{\chi^2}^2(p_{XY}) \le \frac{2 d_{ME}(p_{XY})}{p_{min}}$$

Example 20 Consider the following p_{XY}, with $K = K' = 2$

$1 - \frac{2}{n}$	$\frac{1}{n}$
0	$\frac{1}{n}$

$$d_{ME} = \frac{1}{n} \quad \text{and} \quad d_{\chi^2}^2 = \frac{1}{2} + \frac{1}{2(n-1)}$$

Hence, $d_{\chi^2}^2/d_{ME}$ is of order $n = 1/p_{min}$.

Proof outline The proof is based on the fact that a convex function is always above any tangent to its graph. We pick a point $\bar{\bar{p}}$ that has $d_{ME}(\bar{\bar{p}}) = \varepsilon$ and lower bound $\chi^2(\bar{\bar{p}})$ by the tangent to χ^2 in the nearest $\bar{\bar{p}}^*$ (which always exists). We first prove three intermediate results then follow with the formal proof of the theorem.

First, we calculate the tangent slope at $\bar{\bar{p}}^*$.

Lemma 21 *The unconstrained partial derivatives of χ^2 in $\bar{\bar{p}}^*$ are*

$$\left.\frac{\partial \chi^2}{\partial p_{xy}}\right|_{\bar{\bar{p}}^*} = \begin{cases} -\frac{1}{p_y}, & x \neq y \\ \frac{1}{p_x}, & x = y \end{cases} \quad \text{for } x, y \in [K]$$

Proof

$$\begin{aligned} \frac{\partial \chi^2}{\partial p_{ab}} &= \frac{\partial}{\partial p_{ab}} \left[\sum_x \frac{1}{p_x} \sum_y \frac{p_{xy}^2}{\sum_{x'} p_{x'y}} \right] \\ &= \frac{1}{p_a} \frac{\partial}{\partial p_{ab}} \left(\frac{p_{ab}^2}{\sum_{x'} p_{x'b}} \right) + \sum_{x \neq a} \frac{1}{p_x} \frac{\partial}{\partial p_{ab}} \left(\frac{p_{xb}^2}{\sum_{x'} p_{x'b}} \right) \\ &= \frac{1}{p_a} \frac{2 p_{ab} p'_b - p_{ab}^2 . 1}{p_b'^2} + \sum_{x \neq a} \frac{-p_{xb}^2}{p_x p_b'^2} \\ &= \frac{2 p_{ab}}{p_a p'_b} - \sum_x \frac{p_{xb}^2}{p_x p_b'^2} \end{aligned} \tag{4.1}$$

The result follows now by setting $p_{xb} = p_x \delta_{xb}$, $p'_b = p_b$. □

Then, we calculate a first order approximation of $\chi^2(\bar{\bar{p}})$ by projecting on the tangent direction.

Lemma 22 *For any $\bar{\bar{p}} \in \mathcal{P}$ with $K = K'$*

$$\chi^2(\bar{\bar{p}}^*) - \chi^2(\bar{\bar{p}}) \leq \sum_x \sum_{y \neq x} \left(\frac{p_{xy}}{p_x} + \frac{p_{xy}}{p_y} \right)$$

Proof χ^2 is convex, therefore $\chi^2(\bar{\bar{p}})$ is above the tangent at $\bar{\bar{p}}^*$, i.e.,

$$\chi^2(\bar{\bar{p}}) \geq \chi^2(\bar{\bar{p}}^*) + \operatorname{vec}(\nabla \chi^2(\bar{\bar{p}}^*)) \cdot \operatorname{vec}(\bar{\bar{p}} - \bar{\bar{p}}^*)$$

$$\operatorname{vec}(\nabla \chi^2(\bar{\bar{p}}^*)) \cdot \operatorname{vec}(\bar{\bar{p}} - \bar{\bar{p}}^*) = \sum_x \frac{1}{p_x} \left(-\sum_{y \neq x} p_{xy} \right) + \sum_x \left(-\frac{1}{p_y} \sum_{y \neq x} p_{xy} \right) \tag{4.2}$$

$$= -\sum_x \sum_{y \neq x} \left(\frac{p_{xy}}{p_x} + \frac{p_{xy}}{p_y} \right) \tag{4.3}$$

□

Denote

$$\varepsilon_x = \frac{1}{p_x} \sum_{y \neq x} p_{xy}, \quad x \in [K] \tag{4.4}$$

$$\varepsilon'_y = \frac{1}{p_y} \sum_{x \neq y} p_{xy}, \quad y \in [K] \tag{4.5}$$

These quantities represent the relative leak of probability mass from the diagonal to the off-diagonal cells in row x, respectively in column y of the matrix $\bar{\bar{p}}$ w.r.t. $\bar{\bar{p}}^*$.

The bound in Lemma 22 depends on all the p_{xy} entries in $\bar{\bar{p}}$. Therefore, the next step is to upper bound it by something that depends only on ε and p_{min}.

Lemma 23 *Let $\varepsilon_x, x \in [K]$ be as defined above, and assume that the marginals p_x are sorted so that $p_{min} = p_1 \leq p_2 \leq p_3 \leq \cdots \leq p_K = p_{max}$ with $\sum_x p_x \varepsilon_x = \varepsilon$. Then,*

$$\max_{\{\varepsilon_x\}} \sum_x \varepsilon_x = \begin{cases} \frac{\varepsilon}{p_1}, & \text{if } \varepsilon \in [0, p_1] \\ 1 + \frac{\varepsilon - p_1}{p_2}, & \text{if } \varepsilon \in (p_1, p_1 + p_2] \\ \cdots & \\ k + \frac{\varepsilon - \sum_{x \leq k} p_x}{p_{k+1}}, & \text{if } \varepsilon \in (p_1 + \cdots + p_k, p_1 + \cdots + p_{k+1}] \\ \cdots & \end{cases}$$

Proof It is easy to verify the solution for $\varepsilon \leq p_1$. For the other intervals, one verifies the solution by induction over $k \in [K]$. □

Proof of Theorem 19 Assume that $d_{ME}(\bar{\bar{p}}) = \varepsilon$. Then, without loss of generality one can assume that the off-diagonal elements of $\bar{\bar{p}}$ sum to ε. It is easy to see from Lemma 23 that

$$\sum_x \varepsilon_x \leq \frac{\varepsilon}{p_{min}}$$

By symmetry, this bound also holds for $\sum_y \varepsilon'_y$. Therefore, by Lemma 22

$$K - \chi^2(\bar{\bar{p}}) = \chi^2(\bar{\bar{p}}^*) - \chi^2(\bar{\bar{p}}) \leq \frac{2\varepsilon}{p_{min}} \tag{4.6}$$

from which the desired result follows. The case $K' < K$ was proved in the previous section. □

This theorem holds for every value of d_{ME}. Because of the linear approximation in Lemma 22, the bound is not tight. However, the proof of Lemma 23 indicates that the bound will be tighter when $d_{ME} \leq p_{min}$, (when Lemma 23 gives a tight bound); that is, for smaller differences between the two partitions.

5 Small d_{ME} implies small d_H

This section and the next show the local equivalence of d_H and d_{ME}. We start by presenting a few useful facts about the Hamming distance d_H, including the fact that it is concave.

The first set of helpful facts can be obtained by direct calculations, and the proofs are omitted. They prepare the ground for the more interesting concavity theorem.

Lemma 24

1. *The Hamming distance d_H can be expressed as*

$$d_H = 2\sum_x \sum_{y \neq y'} p_{xy} p_{xy'} + 2\sum_y \sum_{x \neq x'} p_{xy} p_{x'y} \tag{5.1}$$

where the sums are taken over the unordered pairs (x, x') and respectively (y, y').

2. *Its partial derivatives are given by*

$$\frac{\partial d_H}{\partial p_{ab}} = 2\sum_{y \neq b} p_{ay} + 2\sum_{x \neq a} p_{xb} \tag{5.2}$$

3. *Its second order partial derivatives are given by*

$$\frac{\partial^2 d_H}{\partial p_{ab}^2} = 0 \quad \textit{for all } a, b \tag{5.3}$$

$$\frac{\partial^2 d_H}{\partial p_{ab} \partial p_{a'b}} = \frac{\partial^2 d_H}{\partial p_{ab} \partial p_{ab'}} = 1 \quad \textit{for all } a, b, a', b', a \neq a', b \neq b' \tag{5.4}$$

$$\frac{\partial^2 d_H}{\partial p_{ab} \partial p_{a'b'}} = 0 \quad \textit{otherwise} \tag{5.5}$$

Theorem 25 *The Hamming distance d_H is concave in p_{XY}.*

Proof From (5.3), (5.4) and (5.5) we derive that the Hessian H of d_H can be written as a square matrix with $K \times K$ blocks of size $K' \times K'$. The off-diagonal blocks are of the form $I_{K'}$, the unit matrix of dimension K', and the diagonal blocks are of the form $\bar{\bar{1}}_{K'} - I_{K'}$, with $\bar{\bar{1}}_{K'}$ being the matrix of all ones.

It is immediate to verify that any v of dimension $K \times K'$ satisfying $\sum_x v_{xy} = \sum_y v_{xy} = 0$ is an eigenvector of H with eigenvalue -2 (for compatibility with p_{XY} we index the "vector" in the same way as we index probability tables). Now note that for any two probabilities $p_{XY}^{(1)}$, $p_{XY}^{(2)}$ the difference $v = p_{XY}^{(1)} - p_{XY}^{(2)}$ is exactly such a v. Therefore, the Hessian projected on the probability simplex is always negative definite, hence d_H is strictly concave. □

Now we are ready to prove this section's main result.

Theorem 26 *Let p_{XY} represent a pair of clusterings with $K \leq K'$. Then*

$$d_H(p_{XY}) \leq 4 p_{max} d_{ME}(p_{XY})$$

Proof The proof is similar to that of Theorem 19, using the fact that a concave function is always below any tangent to its graph. We pick a point $\bar{\bar{p}}$ that has $d_{ME}(\bar{\bar{p}}) = \varepsilon$ and upper bound $d_H(\bar{\bar{p}})$ by the tangent to d_H in the "nearest" extreme point of $\mathcal{P}$. We define this to be the point $\bar{\bar{p}}^*_{\pi^{ME}}$, with π^{ME} the permutation of cluster assignments that realizes the d_{ME} distance according to Definition 2. Assume without loss of generality that π^{ME} is the

identity, so the extreme point in question is $\bar{\bar{p}}^*$. We consider the tangent through $\bar{\bar{p}}^*$ and obtain

$$d_H(\bar{\bar{p}}) \leq 0 + \mathrm{vec}(\nabla d_H(\bar{\bar{p}}^*)) \cdot \mathrm{vec}(\bar{\bar{p}} - \bar{\bar{p}}^*) \tag{5.6}$$

From (5.2) we get

$$\left.\frac{\partial d_H}{\partial p_{aa}}\right|_{\bar{\bar{p}}^*} = 0 \quad \text{and} \quad \left.\frac{\partial d_H}{\partial p_{ab}}\right|_{\bar{\bar{p}}^*} = 2(p_a + p_b) \quad \text{for all } a \neq b$$

The expression of $\bar{\bar{p}} - \bar{\bar{p}}^*$ is given in (3.2). Hence, (5.6) becomes

$$\begin{aligned} d_H(\bar{\bar{p}}) &\leq \sum_x \sum_{y \neq x} 2(p_x + p_y) p_{xy} \\ &= 4 \sum_x \sum_{y \neq x} p_x p_{xy} \\ &\leq 4 p_{max} \sum_{xy, y \neq x} p_{xy} = 4 p_{max} d_{ME}(\bar{\bar{p}}) \end{aligned}$$

□

Note that this is a global bound, holding for any values of d_{ME} and d_H. Moreover, it can be used to upper bound by d_{ME} any other concave distance between clusterings.

6 Small d_H implies small d_{ME}

This result is formulated and proved similarly to the result of Sect. 3. Thus, we prove that if a joint distribution $\bar{\bar{p}}$ in $\mathcal{P}$ is more than ε away w.r.t. d_{ME} from any optimal point $\bar{\bar{p}}^*$ then $d_H(\bar{\bar{p}})$ will be bounded away from 0.

Theorem 27 *For two clusterings represented by the joint distribution p_{XY}, denote $p_{min} = \min_{[K]} p_x$. Then, for any $\varepsilon \leq p_{min}$, if $d_H(p_{XY}) \leq 2\varepsilon p_{min}$ then $d_{ME}(p_{XY}) \leq \varepsilon$.*

Proof The reasoning follows that of Theorem 9. We assume that $d_{ME} \geq \varepsilon$, and we already know that the subset of $\mathcal{P}$ where this is true is included in the convex hull of $\{E_\varepsilon^\pi\}_{\Pi_K} \cup E^-$. Because d_H is concave, its minimum over this convex set is attained in an extreme point. We will find the minimum of d_H over $\{E_\varepsilon^\pi\}_{\Pi_K} \cup E^-$; this is a lower bound for d_H when $d_{ME} \geq \varepsilon$. By contradiction, we get that d_H upper bounds d_{ME}.

We now need to find the minimum of d_H over the points $\bar{\bar{p}}_\varepsilon(a,b) \in \{E_\varepsilon^\pi\}_{\Pi_K} \cup E^-$, as all the rest is taken care of as part of Theorem 9. For the points in E_ε^π we have

$$\begin{aligned} d_H(\bar{\bar{p}}_\varepsilon(a,b)) &= 2\left[\sum_x \sum_{y \neq y'} p_{xy} p_{xy'} + \sum_y \sum_{x \neq x'} p_{xy} p_{x'y}\right] \\ &= 2 p_{ab}(p_{aa} + p_{bb}) \\ &= 2\varepsilon(p_a - \varepsilon + p_b) \\ &\geq 2\varepsilon(2p_{min} - \varepsilon) \geq 2\varepsilon p_{min} \end{aligned}$$

Let $\bar{\bar{p}}^-$ be a point in E^-. This means that the corresponding clustering Y, with $K' < K$ clusters, merges some clusters in X. For simplicity we will write $x \in y$ to denote that cluster C_x is one of the clusters included in C'_y. Note also that for $\bar{\bar{p}}^-$, we have that $p_{xy}p_{xy'} = 0$ always, and that $p_{xy}p_{x'y} > 0$ only if $x, x' \in y$. We will also write K_y for the number of clusters of X that were merged to form cluster C'_y. Then

$$d_H(\bar{\bar{p}}^-) = 2\sum_y \sum_{x,x'\in y, x\neq x'} p_x p_{x'} \geq 2\underbrace{\sum_y \frac{K_y(K_y-1)}{2}}_{\kappa} p_{min}^2$$

It is easy to verify that $\kappa = 1$ if $K = 2, K' = 1$ and $\kappa \geq 2$ otherwise. Hence, in general $d_H(\bar{\bar{p}}^-) \geq 2p_{min}^2 \geq 2\varepsilon p_{min}$ and if $K \neq 2$ or $K' \neq 1$ $d_H(\bar{\bar{p}}^-) \geq 4p_{min}^2 \geq 2\varepsilon(2p_{min} - \varepsilon)$. □

This result holds for all possible numbers of clusters. It is easy to see from the proof that the slightly stronger bound $2\varepsilon(2p_{min} - \varepsilon)$ can be used in all cases except $K = 2, K' = 1$.

7 The relation between $d^2_{\chi^2}$ and d_H

We now turn to the last pair of distances. Based on the inequalities that we already have, one can derive local equivalence relations between $d^2_{\chi^2}$ and d_H.

Theorem 28 *Let p_{XY} represent a pair of clusterings X, Y with the same number of clusters, and let $\varepsilon \leq p_{min}$.*

1. *If $d_H \leq 2\varepsilon p_{min}$, then $d^2_{\chi^2} \leq \frac{2\varepsilon}{p_{min}}$.*
2. *If $d^2_{\chi^2} \leq \frac{\varepsilon}{p_{max}}$, then $d_H \leq 4\varepsilon p_{max}$.*

Proof The proof of part 1 follows immediately from Theorems 27 and 19; part 2 follows from Theorems 9 and 26. □

However, for this pair of distances we can also prove another relationship.

Theorem 29 *For any two clusterings X, Y we have*

$$d^2_{\chi^2}(X,Y) \leq \frac{d_H(X,Y)}{p_{max}p'_{max}} + \left(\frac{K+K'}{2} - \frac{\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{max}p'_{max}}\right) \tag{7.1}$$

$$d^2_{\chi^2}(X,Y) \geq \frac{d_H(X,Y)}{p_{min}p'_{min}} + \left(\frac{K+K'}{2} - \frac{\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{min}p'_{min}}\right) \tag{7.2}$$

where $p_{max}, p_{min}, p'_{max}, p'_{min}$ represent the probabilities of the largest and smallest clusters in X, respectively in Y.

The additive terms in (7.1) and (7.2) cannot be removed, as it is shown in Lemma 30 below. Therefore, the result above is not strictly speaking a local equivalence.

There are several other differences between Theorem 29 and the previous theorems. First, the additive terms depend on both marginals p_X, p_Y, thus require more detailed knowledge

of both clusterings. The second difference pertains to the proof; the proof is not geometric and does not have a simple geometric interpretation due to the extra terms. Third, even the simple coefficients $\frac{1}{p_{max}p'_{max}}, \frac{1}{p_{min}p'_{min}}$ depend on both clusterings.

It is however worth noticing that the additive terms become 0 when all the clusters have equal sizes, i.e., when p_X, p_Y are uniform distributions over $[K]$. Hence, we find again that the bounds become looser and the distances differ more when the clusterings are less balanced.

Proof of Theorem 29 By definition

$$A_X = \tilde{A_X}\text{diag}(\sqrt{n_1}\sqrt{n_2}\dots\sqrt{n_K})$$
$$A_Y = \tilde{A_Y}\text{diag}(\sqrt{n'_1}\sqrt{n'_2}\dots\sqrt{n'_{K'}}).$$

We introduce these expressions in the definition of $d^2_{\chi^2}$.

$$\begin{aligned} &d^2_{\chi^2}(X,Y) \\ &\quad= \frac{K+K'}{2} - \text{trace}\,\tilde{A}_X^T\tilde{A}_Y\tilde{A}_Y^T\tilde{A}_X \\ &\quad= \frac{K+K'}{2} - \text{trace}\,[\text{diag}(n_1^{-1}, n_2^{-1}, \dots, n_K^{-1})A_X^TA_Y\text{diag}(n_1'^{-1}, n_2'^{-1}, \dots, n_{K'}'^{-1})(A_X^TA_Y)^T] \end{aligned} \tag{7.3}$$

The matrix $A_X^TA_Y$ has non-negative elements, and the diagonal matrices have positive diagonals, with $np_{min} \le n_x \le np_{max}$, and $np'_{min} \le n'_y \le np'_{max}$. Hence, if we replace n_x, n'_y with their lower (upper) bounds in (7.3) we obtain upper (lower) bounds for this expression. It follows that

$$\begin{aligned} &d^2_{\chi^2}(X,Y) \\ &\quad\ge \frac{K+K'}{2} - \text{trace}\left[\frac{1}{np_{min}}A_X^TA_Y\frac{1}{np'_{min}}(A_X^TA_Y)^T\right] \\ &\quad= \frac{K+K'}{2} - \frac{1}{n^2p_{min}p'_{min}}\text{trace}\,A_X^TA_Y(A_X^TA_Y)^T \\ &\quad= \frac{K+K'}{2} - \frac{1}{p_{min}p'_{min}}\left[\frac{1}{2}\left(\sum_{x\in[K]}p_x^2 + \sum_{y\in[K']}p_y^2\right) - d_H(X,Y)\right] \\ &\quad= \frac{1}{p_{min}p'_{min}}d_H(X,Y) + \left(\frac{K+K'}{2} - \frac{\sum_{x\in[K]}p_x^2 + \sum_{y\in[K']}p_y^2}{2p_{min}p'_{min}}\right) \end{aligned}$$

The lower bound is proved in a similar way. □

We now show that the rightmost terms of the inequalities (7.1) and (7.2) are negative, respectively positive, and hence that the equations cannot be simplified by removing them.

Lemma 30

$$\frac{\sum_{x\in[K]}p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{max}p'_{max}} \le \frac{K+K'}{2} \le \frac{\sum_{x\in[K]}p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{min}p'_{min}}$$

Proof We only prove the first inequality, as the second one is proved similarly.

One first notes that as $p_{max}, p'_{max} \geq 1/K$ it follows that

$$K \geq \frac{1}{p_{max}}, \frac{1}{p'_{max}}$$

Then we write successively

$$\begin{aligned}\frac{\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{max}p'_{max}} &= \frac{\sum_{x\in[K]} p_x \frac{p_x}{p_{max}}}{2p'_{max}} + \frac{\sum_{y\in[K']} p'_y \frac{p'_y}{p'_{max}}}{2p_{max}} \\ &\leq \frac{\sum_{x\in[K]} p_x}{2p'_{max}} + \frac{\sum_{y\in[K']} p'_y}{2p'_{max}} \\ &= \frac{1}{2p_{max}} + \frac{1}{2p'_{max}} \leq \frac{K+K'}{2}\end{aligned}$$

□

Theorem 29 would be strictly stronger than Theorem 28, if the additive terms in the former could be removed. However, these term can't be ignored as shown above, but are small if the two clusterings have balanced clusterings, with all cluster sized close to $1/K$. This makes it probable that for imbalanced clusterings, Theorem 28 provides the tighter bound, while for well balanced clusterings Theorem 29 is the tighter one. But although at times looser than the algebraic bounds, the geometric ones have the advantage of simplicity.

8 Concluding remarks

With few exceptions, there is no formula to transform one distance between clusterings into another distance in the absence of additional information. Here we have proved computable bounds on the range of one distance, given another distance, for the case of three specific distances in use. The bounds show that the three distances are in an approximate linear relation (if one considers $d^2_{\chi_2}$ instead of d_{χ^2}) to each other for small distances, provided quantities like p_{min}, p_{max} are kept constant. However, the distances can become arbitrarily different when p_{min} becomes small.

Another characteristic of all the bounds is that they depend on additional features of the clusterings. For Theorems 9, 19, 26 and 27, this information consists only of p_{min} or p_{max} of one of the clusterings. This matters for two reasons: first, it highlights what are the primary factors that govern the variability of a distance given another distance. These are the cluster sizes, and most importantly, the size of the smallest/largest cluster.

Third, it can be seen that all bounds become tighter and hold for a larger range of ε when the clusterings have approximately equal sized clusters, that is when p_{min}, p_{max} approach $1/K$.[5] This confirms the general intuition that clusterings with equal sized clusters are "easier" (and its counterpart, that clusterings containing clusters of very small size are "hard"). From this perspective, here it was shown that clustering with equal sized clusters are "easy to compare."

[5]It is worth noticing that if either p_{min} or p_{max} are near $1/K$ this is sufficient to imply a balanced clustering. This follows from the easy to prove fact that $p_{max} = 1 + \delta$ implies $p_{min} \geq \frac{1}{K} - (K-1)\delta$. The symmetric relation is also true.

The aforementioned Theorems 9, 19, 26 and 27 are more useful than, say, Theorem 28, which depends on all p_x, p'_y, because they depend on p_{min} or p_{max} of one clustering only. Hence, they can be applied in cases when only one clustering is known. For example, Meilă et al. (2005) used this result in the context of spectral clustering, to prove that any clustering with low enough normalized cut is close to the (unknown) optimal clustering of that data set.

Some of the bounds involving d_{ME} found here are only correct when $d_{ME} < p_{min}$ the minimum cluster size of one clustering. This value can be considered the boundary within which two clusterings can be considered "close". Indeed if a proportion of points in X smaller than p_{min} changes labels, none of the clusters in X will lose all its points. Thus, the "identities" of the clusters in X are preserved in Y.

The proof techniques based on convexity/concavity that were developed in Sects. 3 and 4 can be extended to compare d_{ME} with any other concave distance, the way we did for the d_H. One can find bounds between arbitrary pairs of concave distances by using d_{ME} as intermediary the way we did in Sect. 7. The Lemmas and proofs can also be immediately applied to lower bound a distance by another distance, which is how we obtained the table at the end of Sect. 1.

A natural question that was hinted at in the introduction is: can we hope to prove global equivalence relationships between these distances instead of local ones? Our results provide some answers. Example 20 in Sect. 4 shows that $d^2_{\chi^2}/d_{ME}$ can be as large as n (or $1/p_{min}$) which is essentially unbounded, *even when the number of clusters K, K' are bounded*. For the reverse relationship the situation looks better, because $d_{ME}/d^2_{\chi^2}$ can be bounded by $1/p_{max} \leq K$ for small d_{ME}. Because $\frac{\max d_{ME}}{\max d^2_{\chi^2}} = \frac{1}{K}$ we expect that Theorem 9 can be extended to all X, Y with a constant equal or close to K.

From Sect. 5 we know already that $d_H \leq 4d_{ME}$ (by setting $p_{max} = 1$); since the converse result is fundamentally similar to Theorem 9, there is hope that this can also be extended to a global bound. For the relationship between d_H and $d^2_{\chi^2}$, we have the one-sided *local relationship* $d_H \leq 4/K^2 d^2_{\chi^2}$; this can be extended to a global relationship if Theorem 9 can be. However, in Example 20, $d_H = 2\frac{1}{n}(1 - \frac{1}{n})$, and consequently $d^2_{\chi^2}/d_H$ is of order n (or $1/p_{min}$), hence it is unbounded.

Although the motivation for this work is in clustering, we have proved results which hold for any two finite-valued random variables. The non-linear bound (3.4) in Theorem 17 is tight. The proof of this theorem holds even when $K' \to \infty$.

Of interests to statisticians, the two theorems give lower and upper bounds on the χ^2 measure of independence between two random variables, holding locally when the two variables are strongly dependent. The present approximation complements an older approximation of χ^2 by the mutual information $I_{XY} = \sum_{xy} p_{xy} \ln \frac{p_{xy}}{p_x p'_y}$. It is known (Cover and Thomas 1991) that the second order Taylor approximation of I_{XY} is $\frac{1}{2}(\chi^2(p_{XY}) - 1)$ with χ^2 defined as in (2.2). This approximation is good when $p_{XY} \approx p_X p'_Y$, hence in the weak dependence region, while the bounds we introduce here work for the strong dependence region.

This result is, to our knowledge, the first ever to give a detailed local comparison of two distances between partitions. The case of small distances is of utmost importance, as it is in this regime that one desires the behaviour of any clustering algorithm to lie. The paper concludes with a few examples where the present results can be used.

The first and simplest example is the empirical evaluation of clustering algorithms, where understanding the small distances regime is necessary in order to make fine distinctions among different algorithms. The present equivalence theorems represent a step toward removing the dependence of the distance from the evaluation outcome. One notes that w.r.t.

the d_{ME} distance, the fact that the equivalence holds only for small values ($d_{ME} \leq p_{min}$) is not a hindrance, because this distance becomes too coarse to be useful when its values are large.

Second, any statistical analysis of clustering deals with small perturbations and with the asymptotic limit $n \to \infty$, and our results apply to both situations.

The third example relates to the recent and on-going efforts to relate clustering stability with other "good" properties of a clustering. Various distances between clusterings were used to quantify stability (Ben-David et al. 2006; Bach and Jordan 2006). A relationship between a low distortion and clustering stability has been established (Meilă 2006), and questions of the informational limits of clustering have been investigated (Srebro et al. 2006). While the area of clustering stability is outside the scope of this paper, all work in this area is intimately tied with distances between partitions and their small fluctuations.

References

Bach, F., & Jordan, M. I. (2006). Learning spectral clustering with applications to speech separation. *Journal of Machine Learning Research*, *7*, 1963–2001.

Ben-David, S., von Luxburg, U., & Pal, D. (2006). A sober look at clustering stability. In *19th annual conference on learning theory, COLT 2006*. Berlin: Springer.

Candès, E. J., & Tao, T. (2005). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, *35*, 2313–2351.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*, 273–297.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*(4), 1289–1306.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Lancaster, H. (1969). *The Chi-squared distribution*. New York: Wiley.

Meilă, M. (2005). Comparing clusterings—an axiomatic view. In S. Wrobel & L. De Raedt (Eds.), *Proceedings of the international machine learning conference (ICML)*. New York: ACM Press.

Meilă, M. (2006). The uniqueness of a good optimum for K-means. In A. Moore & W. Cohen (Eds.), *Proceedings of the international machine learning conference (ICML)* (pp. 625–632). Princeton: International Machine Learning Society.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873–895.

Meilă, M., Shortreed, S., & Xu, L. (2005). Regularized spectral learning. In R. Cowell & Z. Ghahramani (Eds.), *Proceedings of the artificial intelligence and statistics workshop (AISTATS 05)*.

Mirkin, B. G. (1996). *Mathematical classification and clustering*. Dordrecht: Kluwer Academic.

Papadimitriou, C., & Steiglitz, K. (1998). *Combinatorial optimization. Algorithms and complexity*. Minneola: Dover.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge: MIT Press.

Srebro, N., Shakhnarovich, G., & Roweis, S. (2006). An investigation of computational and informational limits in Gaussian mixture clustering. In *Proceedings of the 23rd international conference on machine learning (ICML)*.

Vajda, I. (1989). *Theory of statistical inference and information. Theory and decision library. Series B: Mathematical and statistical methods*. Norwell: Kluwer Academic Publishers.