Local equivalences of distances between clusterings

Marina Meilă
University of Washington
Department of Statistics
Box 354322
Seattle, WA 98195-4322
phone:(206)543-8484
e-mail:mmp@stat.washington.edu

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 d_{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 proof is geometric and relies on the convexity of a certain set of probability measures. To my knowledge, this is the first result of its kind.

The motivation for this work is in the area of data clustering. Practically, these distances are frequently used to compare two clusterings of a set of observations. Theoretically, such distances are involved in formulating and proving properties of clusterig algorithms. Besides, our results apply to any pair of finite valued random variables, and provides simple yet tight upper and lower bounds on the χ^2 measure of (in)dependence valid when the two variables are strongly dependent.

1 Motivation

Clustering, or finding partitions in data, has become an increasingly popular part of data analysis. In order to theoretically study clustering, or in order to assess its behaviour empirically, one often needs to compute a distance $d(X, Y)$ between two clusterings X, Y of the same data set. A variety of different distances and indices[1] 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. We will consider three distances between clusterings: the Misclassification Error distance, the Hamming distance (equivalent to the unadjusted Rand index), and the d_{χ^2} distance and we 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 bound 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.

This result is, to my 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. Therefore, this proof provides a theoretical tool for the analysis of algorithms behavior and for the analysis of clustering criteria. For instance, various distances between clusterings were used to quantify stability in [Ben-David et al., 2006], [Bach and Jordan, 2006], a relationship between a low distortion and clustering stability was established in [Meilă, 2006], and questions of the informational limits of

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

clustering were investigated in [Srebro et al., 2006].

In the empirical evaluation of clustering algorithms, understanding the small distances regime is necessary in order to make fine distinctions between different algorithms. The present equivalence theorems represent a step towards removing the dependence on the distance from the evaluation outcome.

2 Definitions and representation

We consider a finite set $\mathcal{D}_n$. A *clustering* is a *partition* of $\mathcal{D}_n$ is into sets C_1, C_2, $\dots$ 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 also assume that $n_k > 0$; in other words, that K represents the number of non-empty clusters.

Representing clusterings as matrices. W.l.o.g. the set $\mathcal{D}_n$ can be taken to be $\{1, 2, \dots n\} \stackrel{def}{\equiv} [n]$. Denote by X a clustering $\{C_1, C_2, \dots C_K\}$; 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 $x \in [K]$ w.p. $\frac{n_k}{n}$. One typically requires distances between partitions to be invariant to the permutations of the labels $1, \dots K$. By this representation, any distance between two clusterings can be seen as a particular type of distance between random variables which is invariant to permutations.

Let a second clustering of $\mathcal{D}_n$ be $Y = \{C'_1, C'_2, \dots C'_{K'}\}$, 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 two distances between 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 $\pi : [K'] \rightarrow [K']$.

Although the maximization above is over a set of size $(K')!$, 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 X, Y are close[Meilă, 2007]. Otherwise, its poor resolution represents a major hindrance.

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 $p_X = (p_x)_{x \in [K]}$, $p'_Y = (p'_y)_{y \in [K']}$ 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],

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

□

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 it equals 1 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 lemma gives 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 colums 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*

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

and

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

Proof To prove (2.3) note that $(\tilde{A}_X^T \tilde{A}_Y)_{xy} = \frac{p_{xy}}{\sqrt{p_x p'_y}}$. Then, to prove the second equality, note that $||\tilde{A}_X^T \tilde{A}_X||_F^2 = K$, $||\tilde{A}_Y^T \tilde{A}_Y||_F^2 = K'$. □

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*

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

Because A_X, A_Y are $\{0,1\}$ matrices representing clusterings, $A_X^T A_X$, $A_Y^T A_Y$ are also $\{0,1\}$ matrices, and the Frobenius norm on the r.h.s of equation (2.5) counts in how many positions they differ. Hence, d_H is the Hamming distance between the matrices $A_X^T A_X$, $A_Y^T A_Y$. Note also the strong similarity with the expression of $d^2_{\chi^2}$, which shows that the Hamming distance is also a squared 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] $Rand(X,Y)$ is given by

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

Proof 1) This probabilitistic interpretation of the Hamming distance was put forward 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] □

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[2]; therefore it is concave. The concavity of d_H is proved in section 5.

The following sections prove the local equivalences between the three distances, in the following sequence: $d^2_{\chi^2}$ upper bounds d_{ME} in section 3, d_{ME} upper bounds $d^2_{\chi^2}$ in section 4, d_{ME}

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

upper bounds d_H in section 5, d_H upper bounds d_{ME} in section 6. A local equivalence under slightly different conditions between d_H and $d^2_{\chi^2}$ is proved in section 7. The paper concludes with a discussion.

3 Small d_{χ^2} implies small d_{ME}

In the rest of the paper we will adopt the following notation: p_{XY} and $\bar{\bar{p}}$ denote a distribution from $\mathcal{P}$, p_{xy} is the probability of pair $(x, y) \in [K] \times [K]$ under $\bar{\bar{p}}$, $p_X = (p_1 \ \ldots \ p_K)$, $p_Y = (p'_1 \ \ldots p'_K)$ are respectively the X and Y marginals of $\bar{\bar{p}}$.

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

An example Before we embark on the proof, we give an example where $d^2_{\chi^2}/d_{ME}$ is arbitrarily close to this bound. Consider the following p_{XY}, with $K = K'$.

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

$$d_{ME} = \frac{1}{n} \quad \text{and} \quad d^2_{\chi^2} = \frac{K}{n} - \frac{2}{n^2/K^2 - 1}$$ Hence, $d^2_{\chi^2}/d_{ME}$ approaches $K = 1/p_{max}$.

Outline of proof To prove this statement, we adopt the following framework. First, for simplicity, we assume that $K' = K$; the generalization to $K' \neq K$ is straightforward. Second, we will assume w.l.o.g that partition X is fixed, while Y is allowed to vary. In terms of random variables, the two assumptions describe the set of distributions over $[K] \times [K]$ that have a fixed marginal $p_X = (p_1, \ \ldots \ p_K)$. We denote this domain by $\mathcal{P}$. Thus, $\mathcal{P} = \{\bar{\bar{p}} = [p_{xy}]_{x,y\in[K]}, \ p_{xy} \geq 0, \ \sum_y p_{xy} = p_x \text{ for } y \in [K]\}$, a convex and bounded set.

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. 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 $1, 2, \ldots K$.

In the following it will be proved 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.

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 $\bar{\bar{p}}_{\varepsilon,\pi}(a,b)$. Below is the definition of $\bar{\bar{p}}_{\varepsilon,\pi}$ in the case of the identical permutation. In what follows, whenever we consider one optimal point only, we shall assume w.l.o.g 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

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

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

These results will be proved as a series of lemmas, after which the formal proof of the theorem will close this section.

Lemma 10 *(i) The set of extreme points of $\mathcal{P}$ is*

$$E \;=\; \{\bar{\bar{p}} \,|\, \exists \phi : [K] \longrightarrow [K'],\ p_{xy} = p_x \text{ if } y=\phi(x),\ 0 \text{ otherwise}\} \tag{3.3}$$

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

Proof The proof of (i) is immediate and left to the reader. To prove (ii) let $\bar{\bar{p}} \in E$. We can write successively

$$\begin{aligned} \chi^2(\bar{\bar{p}}) \;&=\; \sum_{y|p'_y>0} \sum_{x\in\phi^{-1}(y)} \frac{p_x^2}{p_x \sum_{z\in\phi^{-1}(y)} p_z} \\ &=\; \sum_{y|p'_y>0} \frac{\sum_{x\in\phi^{-1}(y)} p_x}{\sum_{z\in\phi^{-1}(y)} p_z} \;=\; \sum_{y|p'_y>0} 1 \;=\; |\text{Range}\,\phi| \end{aligned}$$

□

If Range(ϕ) = K, then ϕ is a permutation and we denote it by π. Let $E^* = \{\bar{\overline{p}}^*_\pi\}$ the set of extreme points for which $\chi^2 = K$ and $E^- = E \setminus E^*$ the set of the extreme points for which $\chi^2 \leq K - 1$.

Lemma 11 *Let $B_1(\varepsilon)$ be the 1-norm ball of radius ε centered at $\bar{\overline{p}}^* \in E^*$. Then,*

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

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

$$||\bar{\overline{p}}^* - \bar{\overline{p}}_\varepsilon(a,b)||_1 \;=\; \sum_{x,y} |p^*_{xy} - p_\varepsilon(a,b)_{xy}| \;=\; |p^*_{aa} - p_\varepsilon(a,b)_{aa}| + |p^*_{ab} - p_\varepsilon(a,b)_{ab}| = \varepsilon + \varepsilon \;=\; 2\varepsilon \tag{3.4}$$

For any point $\bar{\overline{p}} \in B_1(2\varepsilon) \cap \mathcal{P}$ denote by

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

Then, it is easy to check that

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

and

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

□

Lemma 12 *For all $\bar{\overline{p}} \in B_1(2\varepsilon) \cap \mathcal{P}$*

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

Proof Obvious, since $d_{ME}(\bar{\overline{p}}) \leq \sum_x \sum_{y \neq x} p_{xy} \;=\; \varepsilon$. □

Lemma 13 *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:

$$x = \sum_i \left[\frac{\alpha_i}{1-\beta_i} y_i - \frac{\alpha_i \beta_i}{1-\beta_i} x \right] \tag{3.5}$$

$$= \sum_i \frac{\alpha_i}{1-\beta_i} y_i - x \sum_i \frac{\alpha_i \beta_i}{1-\beta_i} \tag{3.6}$$

Hence

$$x = \sum_i \underbrace{\frac{\frac{\alpha_i}{1-\beta_i}}{1+\sum_j \frac{\alpha_j \beta_j}{1-\beta_j}}}_{\gamma_i} y_i \tag{3.7}$$

with $\gamma_i \geq 0$ and

$$\sum_i \gamma_i = \frac{\sum_i \frac{\alpha_i}{1-\beta_i}}{1+\sum_j \frac{\alpha_j \beta_j}{1-\beta_j}} = \frac{1+\sum_i \frac{\alpha_i \beta_i}{1-\beta_i}}{1+\sum_j \frac{\alpha_j \beta_j}{1-\beta_j}} = 1 \tag{3.8}$$

□

Lemma 14 *The set* $\{\bar{\overline{p}} \,|\, d_{ME}(\bar{\overline{p}}) \geq \varepsilon\}$ *with* $\varepsilon \leq p_{min}$ *is included in the convex hull of* $\{E_\varepsilon^\pi\}_{\Pi_K} \cup E^-$.

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

$$\bar{\overline{p}} = \sum_{i=1}^{|E|} \alpha_i \bar{\overline{p}}^*_{\pi_i}, \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1$$

$$= \sum_{i=1}^{K!} \alpha_i \bar{\overline{p}}^*_{\pi_i} + \sum_{i=1}^{|E^-|} \alpha_{i+K!} \bar{\overline{p}}_i$$

Let us look at the segment $[\bar{\overline{p}}, \bar{\overline{p}}^*_{\pi_i}]$; its first end, $\bar{\overline{p}}$ 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 $\bar{\overline{p}}_i = [\bar{\overline{p}}, \bar{\overline{p}}^*_{\pi_i}] \cup \partial B_1^{\pi_i}(\varepsilon)$. This point being on the boundary of the ball, it can be written as a convex combination of points in $E_\varepsilon^{\pi_i}$ by Lemma 11. We now apply Lemma 13, with $y_i = \bar{\overline{p}}^*_{\pi_i}$ for $i = 1, \ldots K!$ and $y_i = \bar{\overline{p}}_{i-K!}$ for $i > K!$. It follows that $\bar{\overline{p}}$ is a convex combination of $\bar{\overline{p}}_i$, $i = 1, \ldots m$, which completes the proof. □

Lemma 15 *For* $\varepsilon \leq p_{min}$

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

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

$$\begin{aligned}
\chi^2(\overline{\overline{p}}_\varepsilon(a,b)) &= K-2+\frac{(p_a-\varepsilon)^2}{p_a(p_a-\varepsilon)}+\frac{\varepsilon^2}{p_a(p_b+\varepsilon)}+\frac{p_b^2}{p_b(p_b+\varepsilon)} && (3.9)\\
&= K-2+1-\frac{\varepsilon}{p_a}+\frac{\varepsilon^2}{p_a(p_b+\varepsilon)}+1+\frac{\varepsilon}{p_b+\varepsilon} && (3.10)\\
&= K-\frac{\varepsilon(p_a+p_b)}{p_a(p_b+\varepsilon)} && (3.11)\\
&\leq K-\frac{\varepsilon}{p_a} && (3.12)
\end{aligned}$$

Therefore

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

□

Proof of Theorem 9 By contradiction. Assume $d_{ME}(\overline{\overline{p}}) \geq \varepsilon$. Then, $\overline{\overline{p}} \in A$ by lemma 14. Since χ^2 is convex on A, $\chi^2(\overline{\overline{p}})$ cannot be larger than the maximum value at the extreme points of A, which are contained in $E^- \cup (\cup_\pi E^\pi_\varepsilon)$. But we know by lemma 15 that the value of χ^2 is bounded above by $K-\varepsilon/p_{max}$ at any point in E^π_ε and by $K-1$ at any point in E^-.

Note also that a tight, non-linear bound can be obtained by maximizing (3.11) over all a, b.
□

4 Small d_{ME} implies small d_{χ^2}

Theorem 16 *Let p_{XY} represent a pair of clusterings with $d_{ME}(p_{XY}) \leq \varepsilon$. Then*

$$d^2_{\chi^2}(p_{XY}) \;\leq\; \frac{2\varepsilon}{p_{min}}$$

An example Consider the following p_{XY}, with $K = K' = 2$

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

$$d_{ME} \;=\; \frac{1}{n} \quad \text{and} \quad d^2_{\chi^2} \;=\; \frac{1}{2}+\frac{n-3}{2(n-1)(n-2)}$$

Hence, $d^2_{\chi^2}/d_{ME}$ is of order n.

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

Lemma 17 *The unconstrained partial derivatives of* χ^2 *in* $\overline{\overline{p}}^*$ *are*

$$\left.\frac{\partial \chi^2}{\partial p_{xy}}\right|_{\overline{\overline{p}}^*} = \begin{cases} -\frac{1}{p_y}, & x \neq y \\ \frac{1}{p_x}, & x = y \end{cases}$$

Proof

$$\frac{\partial \chi^2}{\partial p_{ab}} = \frac{\partial}{\partial p_{ab}} \left[\sum_x \frac{1}{p_x} \sum_y \frac{p_{xy}^2}{\sum_{x'} p_{x'y}} \right] \tag{4.1}$$

$$= \frac{1}{p_a} \frac{\partial}{\partial p_{ab}} \left(\frac{p_{ab}^2}{\sum_{x'} p_{x'y}} \right) + \sum_{x \neq a} \frac{1}{p_x} \frac{\partial}{\partial p_{ab}} \left(\frac{p_{xb}^2}{\sum_{x'} p_{x'b}} \right) \tag{4.2}$$

$$= \frac{1}{p_a} \frac{2p_{ab}p'_b - p_{ab}^2.1}{p_b'^2} + \sum_{x \neq a} \frac{-p_{xb}^2}{p_x p_b'^2} \tag{4.3}$$

$$= \frac{2p_{ab}}{p_a p'_b} - \sum_x \frac{p_{xb}^2}{p_x p_b'^2} \tag{4.4}$$

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

Lemma 18 *For any* $\overline{\overline{p}} \in \mathcal{P}$

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

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

$$\chi^2(\overline{\overline{p}}) \geq \chi^2(\overline{\overline{p}}^*) + \mathrm{vec}(\nabla \chi^2(\overline{\overline{p}}^*)) \cdot \mathrm{vec}(\overline{\overline{p}} - \overline{\overline{p}}^*) \tag{4.5}$$

$$\mathrm{vec}(\nabla \chi^2(\overline{\overline{p}}^*)) \cdot \mathrm{vec}(\overline{\overline{p}} - \overline{\overline{p}}^*) = \sum_x \frac{1}{p_x} \left(\sum_{y \neq x} p_{xy} \right) + \sum_x \left(-\frac{1}{p_y} \sum_{y \neq x} p_{xy} \right) \tag{4.6}$$

$$= -\sum_x \sum_{y \neq x} \left(\frac{p_{xy}}{p_x} + \frac{p_{xy}}{p_y} \right) \tag{4.7}$$

□

Denote

$$\varepsilon_x = \frac{1}{p_x} \sum_{y \neq x} p_{xy}, \quad x \in [K] \tag{4.8}$$

$$\varepsilon'_y = \frac{1}{p_y} \sum_{x \neq y} p_{xy} \quad y \in [K] \tag{4.9}$$

These quantitites 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 $\overline{\overline{p}}$ w.r.t $\overline{\overline{p}}^*$.

Lemma 19 *Let $\varepsilon_x, x \in [K]$ be as defined above, and assume that the marginals p_x are sorted so that $p_{min} = p_1 \leq p_2 \leq p_3 \leq \ldots \leq p_K = p_{max}$ with $\sum_x p_x \varepsilon_x = \varepsilon$. Then,*

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

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

Proof [Theorem 16] Assume that $d_{ME}(\overline{\overline{p}}) = \varepsilon \leq \varepsilon_0$. Then, w.l.o.g. one can assume that the off-diagonal elements of $\overline{\overline{p}}$ sum to ε. It is easy to see that under the conditions of lemma 19

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

By symmetry, this bound also holds for $\sum_y \varepsilon'_y$. Therefore, by lemma 18

$$\chi^2(\overline{\overline{p}}^*) - \chi^2(\overline{\overline{p}}) \;\leq\; \frac{2\varepsilon}{p_{min}} \tag{4.10}$$

or

$$d^2_{\chi^2}(\overline{\overline{p}}) \;\leq\; \frac{2\varepsilon}{p_{min}} \;\leq\; \frac{2\varepsilon_0}{p_{min}}$$

□

While the theorem holds for every ε_0, it is only interesting for $\varepsilon \in [0, 1]$. Because of the linear approximation in Lemma 18, the bound is not tight. However, the proof of Lemma 19 indicates that the bound will be tighter when $\varepsilon \leq p_{min}$, that is, for smaller differences between the two partitions.

5 Small d_{ME} implies small d_H

We start by proving a few useful facts about the Hamming distance d_H, including the fact that it is concave.

Lemma 20 *1) The Hamming distance d_H can be expressed as*

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

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

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

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

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

Proof By direct calculation. □

Lemma 21 *The Hamming distance d_H is concave in p_{XY}.*

Proof From (5.3) and (5.4) 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 $I_{K'}$ where $I_{K'}$ represents the unit matrix, and the diagonal blocks are of the form $\bar{\bar{\mathbf{1}}}_{K'} - I_{K'}$, where $\bar{\bar{\mathbf{1}}}_{K'}$ is the vector 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 was as we index probability tables). Now note that for any two probabilities $p^{(1)}_{XY}, p^{(2)}_{XY}$ the difference $v = p^{(1)}_{XY} - p^{(2)}_{XY}$ 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 22 *Let p_{XY} represent a pair of clusterings with $d_{ME}(p_{XY}) = \varepsilon \leq p_{min}$. Then*

$$d_H(p_{XY}) \;\leq\; 4\varepsilon p_{max}$$

Proof The proof is similar to that of Theorem 16, 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 $d_H(\overline{\overline{p}})$ by the tangent to d_H in the nearest $\overline{\overline{p}}^*$, i.e

$$d_H(\overline{\overline{p}}) \;\leq\; \nabla d_H(\overline{\overline{p}})^* \mathrm{vec}(\overline{\overline{p}} - \overline{\overline{p}}^*) \tag{5.6}$$

From (5.2) we get

$$\left.\frac{\partial d_H}{\partial p_{aa}}\right|_{\overline{\overline{p}}^*} \;=\; 0 \quad \text{and} \quad \left.\frac{\partial d_H}{\partial p_{ab}}\right|_{\overline{\overline{p}}^*} \;=\; 2(p_a + p_b) \quad \text{for all } a \neq b \tag{5.7}$$

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

$$\begin{aligned} d_H(\overline{\overline{p}}) &\leq \sum_x \sum_{y \neq x} 2(p_x + p_y) p_{xy} && (5.8)\\ &= 4 \sum_x p_x \sum_{y \neq x} p_{xy} && (5.9)\\ &\leq 4 p_{max} \varepsilon && (5.10) \end{aligned}$$

□

6 Small d_H implies small d_{ME}

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

Theorem 23 *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 $d_H(p_{XY}) \leq 4\varepsilon p_{min} - 2\varepsilon^2$ 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 $\{E^{\pi}_{\varepsilon}\}_{\Pi K} \cup E^{-}$. Since 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 $\{E^{\pi}_{\varepsilon}\}_{\Pi K} \cup E^{-}$; 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 $\bar{\overline{p}}_\varepsilon(a,b) \in \{E^\pi_\varepsilon\}_{\Pi K} \cup E^-$, as all the rest is taken care of as part of Theorem 9.

$$\begin{aligned} d_H(\bar{\overline{p}}_\varepsilon(a,b)) &= 2\left[\sum_x \sum_{y\neq y'} p_{xy}p_{xy'} + \sum_y \sum_{x\neq x'} p_{xy}p_{x'y}\right] && (6.1)\\ &= 2p_{ab}(p_{aa}+p_{bb}) && (6.2)\\ &= 2\varepsilon(p_a - \varepsilon + p_b) && (6.3)\\ &\geq 2\varepsilon(2p_{min}-\varepsilon) \;\geq\; 2\varepsilon p_{min} && (6.4) \end{aligned}$$

□

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

For d_H and the distance $d^2_{\chi^2}$ we can prove the following

Theorem 24 *For any two clusterings X, Y we have*

$$\begin{aligned} d^2_{\chi^2}(X,Y) &\leq \frac{d_H}{p_{max}p'_{max}} + \left(K - \frac{\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{max}p'_{max}}\right) && (7.1)\\ d^2_{\chi^2}(X,Y) &\geq \frac{d_H}{p_{min}p'_{min}} + \left(K - \frac{\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']}(p'_y)^2}{2p_{min}p'_{min}}\right) && (7.2) \end{aligned}$$

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

Proof By definition, $\tilde{A_X}\text{diag}(\sqrt{n_1}\sqrt{n_2}\ldots\sqrt{n_K})$, $\tilde{A_Y}\text{diag}(\sqrt{n'_1}\sqrt{n'_2}\ldots\sqrt{n'_{K'}})$.

$$\begin{aligned} & d^2_{\chi^2}(X,Y)\\ &= K - \text{trace}\tilde{A}^T_X\tilde{A}_Y\tilde{A}^T_Y\tilde{A}_X\\ &= K - \text{trace}[\text{diag}(n_1^{-1/2}\,n_2^{-1/2}\ldots n_K^{-1/2})A^T_XA_Y\text{diag}(n_1'^{-1}\,n_2'^{-1}\ldots n_{K'}'^{-1})A^T_YA_X\text{diag}(n_1^{-1/2}\,n_2^{-1/2}\ldots n_K^{-1/2})]\\ &= K - \text{trace}[\text{diag}(n_1^{-1}\,n_2^{-1}\ldots n_K^{-1})A^T_XA_Y\text{diag}(n_1'^{-1}\,n_2'^{-1}\ldots n_{K'}'^{-1})(A^T_XA_Y)^T] \qquad (7.3) \end{aligned}$$

The matrix $A^T_XA_Y$ has non-negative elements, and the diagonal matrices D, D' have positive diagonals, with $np_{min} \leq n_x \leq np_{max}$, and $np'_{min} \leq n'_y \leq np'_{max}$. Hence, if we replace the

diagonal elements of D, D' with their lower (upper) bounds in (7.3) we obtain upper (lower) bounds for this expression. It follows that

$$\begin{aligned}
d^2_{\chi^2}(X,Y) &= K - \text{trace}[\text{diag}(\, n_1^{-1}\, n_2^{-1}\, \ldots\, n_K^{-1}\,) A_X^T A_Y \text{diag}(\, {n'_1}^{-1}\, {n'_2}^{-1}\, \ldots\, {n'_{K'}}^{-1}\,)(A_X^T A_Y)^T] \\
&\leq K - \text{trace}[\frac{1}{np_{min}} A_X^T A_Y \frac{1}{np'_{min}} (A_X^T A_Y)^T] \qquad (7.4) \\
&= K - \frac{1}{n^2 p_{min} p'_{min}} \text{trace} A_X^T A_Y (A_X^T A_Y)^T \qquad (7.5) \\
&= K - \frac{1}{p_{min} p'_{min}} [\frac{1}{2}(\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']} p_y^2) - d_H(X,Y)] \qquad (7.6) \\
&= \frac{1}{p_{min} p'_{min}} d_H(X,Y) + \left(K - \frac{\sum_{x\in[K]} p_x^2 + \sum_{y\in[K']} p_y^2}{2 p_{min} p'_{min}} \right) \qquad (7.7)
\end{aligned}$$

The lower bound is proved in a similar way. □

8 Remarks

With a few exceptions, there is no formula to transform one distance between clusterings into another distance between clustering 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 between clusterings 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, but 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, 16, 22, 23 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.

It can be seen that all bounds become tighter and hold for a larger range of ε when the clusterings have approximately equal clusters, that is when p_{min}, p_{max} approach $1/K$. This confirms the general intuition that clusterings with equal sized clusters are "easier" (and its counterpart, that clusterings containing very small clusters are "hard").

Another reason why the above results are more useful than, say, Theorem 24, which depends on all p_x, p'_y, is that 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.

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 withing 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.

It is also not hard to see that the proof techniques based on convexity/concavity that we developed in sections 3 and 4 can be extended to compare d_{ME} with any other concave distance. We already did so for the d_H. The Lemmas and proofs can also be immediately applied to lower bound a distance by another distance. For instance, the Lemmas in section 3 imply (proof omitted) the result

Lemma 25 *For any $d_{ME} < p_{min}$, $d^2_{\chi^2} \geq \frac{d_{ME}}{p_{max}}$*

This result is given for illustration only as is not tight, but the lemmas also imply a tight bound (where the factor $1/p_{max}$ would be replaced with a more complicated expression involving $1/p_x$, $x \in [K]$).

Although the original motivation for this work stems from comparing partitions, we have proved results which hold for any two finite-valued random variables. The non-linear bound (3.11) in Theorem 9 is tight. The proofs hold when the condition $K' = K$ is replaced by $K' \geq K$ or even by $K' \to \infty$.

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 around $p_{XY} = p_X p'_Y$, hence in the weak dependence region, while the bounds we introduce here work for the strong dependence region.

References

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

[Ben-David et al., 2006] Ben-David, S., von Luxburg, U., and Pal, D. (2006). A sober look at clustering stability. In *19th Annual Conference on Learning Theory, COLT 2006*. Springer.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

[Lancaster, 1969] Lancaster, H. (1969). *The Chi-Squared Distribution*. Wiley.

[Meilă, 2005] Meilă, M. (2005). Comparing clusterings – an axiomatic view. In Wrobel, S. and De Raedt, L., editors, *Proceedings of the International Machine Learning Conference (ICML)*. ACM Press.

[Meilă, 2006] Meilă, M. (2006). The uniqueness of a good optimum for K-means. In Moore, A. and Cohen, W., editors, *Proceedings of the International Machine Learning Conference (ICML)*, pages 625–632. International Machine Learning Society.

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

[Meilă et al., 2005] Meilă, M., Shortreed, S., and Xu, L. (2005). Regularized spectral learning. In Cowell, R. and Ghahramani, Z., editors, *Proceedings of the Artificial Intelligence and Statistics Workshop(AISTATS 05)*.

[Mirkin, 1996] Mirkin, B. G. (1996). *Mathematical classification and clustering*. Kluwer Academic Press.

[Papadimitriou and Steiglitz, 1998] Papadimitriou, C. and Steiglitz, K. (1998). *Combinatorial optimization. Algorithms and complexity*. Dover Publication, Inc., Minneola, NY.

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

[Srebro et al., 2006] Srebro, N., Shakhnarovich, G., and 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, 1989] Vajda, I. (1989). *Theory of statistical inference and information*. Theory and Decision Library. Series B: Mathematical and Statistical methods. Kluwer Academic Publishers, Norwell, MA.