Towards an Information-Theoretic Approach to Population Structure

Omri Tal[1]
[1] School of Philosophy and The Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University, Tel Aviv 69978, Israel.
talomri@post.tau.ac.il

Abstract

This paper uses an information-theoretic perspective to propose multi-locus *informativeness* measures for ancestry inference. These measures describe the potential for correct classification of unknown individuals to their source populations, given genetic data on population structure. Motivated by Shannon's axiomatic approach in deriving a unique information measure for communication (Shannon 1948), we first identify a set of intuitively justifiable criteria that any such quantitative information measure should satisfy, and then select measures that comply with these criteria. It is shown that standard information-theoretic measures such as multidimensional *mutual information* cannot completely account for informativeness when source populations differ in size, necessitating a decision-theoretic approach.

1 Introduction

Information is the resolution of uncertainty.
-- Claude Shannon

The research on the genetic structure of human populations takes diverse paths and involves complex statistical learning and analysis methods. Two of the most powerful approaches are *clustering*, which attempts to infer the underlying population structure, and *classification,* which assigns data of unknown origin to a most probable group. A third category of analysis is high dimensional *structural analysis*, such as principal component analysis, used for constructing low dimensional qualitative representations. Another common approach is to use *diversity*, *differentiation* and *distance* measures to quantify population relatedness and individual variability. The various links between genetic structure and principles from classical information theory have been pointed out from a variety of perspectives: examining the causal, semantic and transmission sense of information embedded in DNA (Maynard Smith 2000; Godfrey-Smith 2006; Bergstrom and Rosvall 2011), using information-theoretic terms to quantitatively model processes such as drift, mutation, selection, gene flow (Smith 2011), or modeling the evolution of complex traits (Plotkin and Nowak 2000).

The use of an information-theoretical approach to derive measures for the information content of genetic markers has been utilized for assignment of genotypes to their source populations (Rosenberg, Li, et al. 2003) and measuring the informativeness of a marker for relationship or relatedness inference (Wang 2006). Previous work on classification performance and *informativeness for assignment* in the context of genetic data had identified three basic properties of related measures: [a] higher informativeness with additional loci, [b] higher informativeness with wider population

divergence, and [c] higher informativeness with larger population sample size (Estoup and Angers 1998; Cornuet, et al. 1999; Edwards 2003; Rosenberg, Li, et al. 2003; Rosenberg 2005; Witherspoon, et al. 2007; Tal 2012). A central goal of this paper is to extend this set of properties, revealing important intrinsic aspects of such information measures. For simplicity, we consider a haploid population model with known allele frequencies from biallelic loci from two subpopulations with given class priors. The approach is motivated by Shannon's axiomatic program in deriving a measure for of the rate of information produced by a discrete information source, formally resembling Entropy in statistical mechanics (Shannon 1948). We first propose a set of necessary and sufficient justifiable properties that any *quantitative* information measure should satisfy. The next step is to examine a host of candidate measures based on popular divergence measures, distance metrics and classification schemes against these criteria, to finally arrive at a measure that strictly complies with the complete set.

1.1 The Motivation from Shannon

In Shannon's groundbreaking 1948 paper, each message received represents gain in information and decrease in uncertainty. Information and uncertainty are thus two sides of the same coin: the more uncertainty there is, the more information we gain by removing the uncertainty (Frigg and Werndl 2011). Shannon posits a measure defined on discrete probabilities of events from an information source:

> Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process [a discrete information source], or better, at what rate information is produced? Suppose we have a set of possible events whose probabilities of occurrence are p_1, p_2,…, p_n. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

Shannon proceeds to propose a set of three "reasonable" properties any such measure H should comply with: [1] H should be continuous in the p_i; [2] If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n; and [3] H should be the weighted sum of the values of H resulting from splitting the probabilities into successive choices. It is then proved that the only H satisfying the three assumptions is of the form,

$$H(X) = -k\sum_{i=1}^{n} p_i \log p_i$$

The discussion follows with additional three "interesting properties which further substantiate it as a reasonable measure of choice or information", namely [a] $H >= 0$ and is zero only if all the p_i are 0, apart from one p_i which is 1; [b] maximum value of H is $log(n)$, attained when all $p_i=1/n$; and [c] any change toward equalization of the probabilities p_i increases H. Shannon found that only three properties [1-3] were required for arriving at a unique formulation for H, whereas these latter properties [a-c] only substantiate H as a reasonable measure. Similarly, we first impose a set of reasonable *criteria* on any informativeness measure, and justify them from considerations of aspects of classification. Several of these criteria will resemble in structure and motivation Shannon's properties, particularly when expressed symbolically.[*] Subsequently, we specify a small set of *properties* that do not have strong prior intuitive justification, but nevertheless provide more insight into the nature of an *informativeness* measure. In the spirit of Shannon, we specify no optimality

[*] Symbolic forms for the three *necessary* properties of H were described subsequent to Shannon's paper. Expressed symbolically, they bear a strong resemblance to the formal definition for the criteria in the next section. For instance, $H_{n+1}(p_1,\ldots,p_n,0) = H_n(p_1,\ldots,p_n,0,)$, or $H_n(p_1,\ldots,p_n) \leq H_n(1/n,\ldots 1/n)$ or $H_n(1/n,\ldots 1/n) \leq H_{n+1}(1/(n+1),\ldots 1/(n+1))$.

criterion: the numerical value of H is only meaningful in *relation* to other distributions to which H is applied (H is unique only up to a constant).[†]

2 The Criteria

We denote by C_n any information measure across a set of n loci from two haploid populations that complies with a given set of criteria.[‡] More specifically, $C_n(P,Q)$ is a measure of *informativeness for classification* given a set of n biallelic markers, where P and Q are *vectors* of known allele frequencies $(p_1,\ldots, p_n)$ and $(q_1,\ldots, q_n)$ in $\boldsymbol{Q^n}$ from populations 1 and 2 respectively, and where $0<p_i<1$ and $0<q_i<1$. We assume p_i and q_i are true population parameters of *polymorphic* loci, i.e., each locus in each population is properly biallelic. Let α be the prior of population 1 (such that $1-\alpha$ is the prior of population 2), the probability that an individual belongs to population 1 when its genotype is unknown (reflecting possible discrepancy in population sizes). In effect, if we denote by N_X the size of population X, then $\alpha = N_1/(N_1+N_2)$. The full notation then becomes $C_n(\alpha,P,Q)$ with the shorter notation appearing where contextually sufficient.

[1] *Zero*: For equal priors ($\alpha=½$), $C_n=0$ *if* $n=0$ or $p_i=q_i$ for all loci i. Less formally, for populations of equal size, if there are no loci to examine, or if allele frequencies are exactly equal across loci between populations, there is *zero* information available for classification. This criterion resembles Shannon's first descriptive property for a condition under which $H=0$.

[2] *Non-negativity*: $C_n \geq 0$. An information measure is expected to be non-negative. This criterion also resembles Shannon's first descriptive property for H, namely, $H\geq 0$.

[3] *Bound*: $C_n < 1$. The information measure should have an upper bound that signifies the possibility of *definite* classification. This bound along with non-negativity and the minimum of zero allows C_n to be interpreted as a *probability* or subjective Bayesian certainty about accurate classification. This criterion resembles Shannon's second descriptive property for H.[§]

[4] *Performance*: C_n should be a monotonic non-decreasing function of n. Informally stated, each additional locus may only add information for classification. This criterion resembles Shannon's second requirement for H.

[5] *Convergence*: *if* allele frequencies at each locus differ between populations by *at least* ε, where ε is any *predefined* value as small as we wish, $0<\varepsilon<1$, then C_n asymptotically equals 1 as $n\rightarrow\infty$. In other words, complete information exists with an infinite number of loci given any minimal predefined frequency difference.[**] This criterion implies that in *practical* sequencing situations, where frequencies of polymorphisms differ by some infinitesimal amount between populations, the inclusion of additional markers eventually results in $C_n\rightarrow 1$. Note that criteria #3 together with #4 are not equivalent to this criterion.

[6] *Neutrality*: The inclusion of *uninformative* loci (for which $q_i = p_i$) should not affect C_n. Note that this does *not* imply that an informative locus should necessarily modify C_n.

[7] *Continuity*: C_n should be continuous in p_i, q_i and α. This criterion resembles Shannon's requirement for continuity in H.

[†] Subsequent generalizations of entropy, such as *Tsallis* entropy or *Renyi* entropy are similarly relative, merely highlighting different aspects of the distribution.

[‡] The use of the term *measure* with respect to C_n should not be interpreted in the strict sense as a mathematical notion of *measure over sets.*

[§] If we define $C'_n=-log(1-C_n)$ so that it ranges from 0 to infinity like entropy H, the measure would lack the benefit of representing maximal informativeness (zero classification error rate). Indeed, H is sometimes normalized to express a "deficiency in entropy" from optimal distribution: $Efficiency = H/\log(n)$.

[**] Note that criteria #3 with #4 are not equivalent to criterion #5. Without the latter criterion, C_n could be less than 1 at the limit.

[8] *Dominance*: $C_n \to 1$ if $\delta_i = |q_i - p_i| \to 1$. A single locus with maximal allele frequency difference should enable *definite* classification of any individual.[††] The asymptotic limit ($\to 1$) follows from the *Continuity* criterion. We note that human SNP data may reveal very high differentiation at many sites, especially at X-linked loci. For e.g., Casto et al. on report 159 X-linked SNPs with frequency delta greater than 0.9 in the Yoruba-Han dataset (Casto, et al. 2010).[‡‡]

[9] *Loci invariance*: C_n should be invariant to different ordering of sequenced loci, i.e., the components of the frequency vectors P and Q may be specified in any order, as long as they remain in sync.

[10] *Symmetry invariance*: $C_n(P,Q) = C_n(Q,P)$.[§§]

[11] *Allocation invariance*: C_n should be invariant to the arbitrary choice of the alleles to which we assign the frequency parameters. Thus the concurrent substitution of p_i with $(1- p_i)$ in P and q_i with $(1- q_i)$ in Q should not affect C_n.

[12] *Priors*: $C_n \to 1$ *if* $\alpha \to 0$ or $\alpha \to 1$. At a limit, extremely unequal priors induce complete information for classification; i.e., if one population is infinitely larger than the other, the probability for correct classification is 1, irrespective of allele frequencies.

In more formal terms:

$Let\ P = (p_1,.., p_n), Q = (q_1,..., q_n),$

$allele\ \ frequencies\ 0 < p_i, q_i < 1,\ population\ \ prior\ 0 < \alpha < 1, and\ d_i = | q_i - p_i |.$

$C_n(\alpha, P, Q)$, *abbreviated* $C_n(P,Q)$ *or* C_n, *should satisfy*:

[1] $if\ \forall i\ p_i = q_i\ and\ \alpha = \frac{1}{2} \Rightarrow C_n = 0$

[2] $C_n \geq 0$

[3] $C_n < 1$

[4] $d_{n+1} > 0 \Rightarrow C_{n+1}((p_1,..., p_n, p_{n+1}),(p_1,..., q_n, q_{n+1})) \geq C_n(P,Q)$

[5] $\forall \varepsilon > 0, if\ \forall i\ d_i > \varepsilon \Rightarrow lim_{n\to\infty} C_n = 1$

[6] $p_{n+1} = q_{n+1} \Rightarrow C_{n+1} = C_n$

[7] $C_n\ is\ coninuous\ in\ p_i, q_i\ and\ \alpha$

[8] $lim_{d_i \to 1} C_n = 1$

[9] $C_n(P,Q) = C_n((p_1,..., p_{i+k},..., p_i,... p_n),(q_1,..., q_{i+k},..., q_i,... q_n))$

[10] $C_n(\alpha, P, Q) = C_n(1-\alpha, Q, P)$

[11] $C_n(P,Q) = C_n((p_1,...,1- p_i,..., p_n),(q_1,...,1- q_i,..., q_n))$

[12] $lim_{\alpha\to 0} C_n = 1\ and\ lim_{\alpha\to 1} C_n = 1$

Following Shannon, we suggest a few properties which further substantiate C_n as a reasonable measure of information for assignment.

[a] *Triviality*: At *equal priors*, $C_1 = |q-p|$. This formulation for C_1 reflects the simplest measure for divergence at a single locus: *absolute allele frequency difference* ($\delta=|q-p|$).

[b] *Population subadditivity*: C_n complies with a triangle inequality when interpreted as a distance measure between populations. Formally,

$$C_n(\alpha, P, Q) + C_n(\beta, Q, R) \geq C_n(\gamma, P, R)$$

[††] With k alleles and c populations, we can have Dominance if $k \geq c$.

[‡‡] However, distinctive alleles present in all individuals of one region but absent from individuals outside the region do not exist for human *microsatellite* data (Rosenberg 2005).

[§§] Nevertheless, exchanging only *some* of the p_i in P with the corresponding q_i in Q may change C_n.

where the three priors are naturally defined in terms of relative population sizes, $\alpha=N_1/(N_1+N_2)$, $\beta=N_2/(N_2+N_3)$, $\gamma=N_1/(N_1+N_3)$. The first two priors define the third: solving for N_1 and N_3 in terms of α and β,

$$\gamma = \frac{\alpha\beta}{\alpha\beta + (1-\alpha)(1-\beta)}.$$

The following section examines several proposals based on standard differentiation, divergence, distance and information measures, illuminating the drawbacks of each and preparing the ground for introducing a proposal satisfying the complete set of criteria.

3 Proposals for C_n

3.1 Differentiation Measures

Wright's fixation index F_{ST} is a classic measure of population differentiation. It is commonly defined in terms of expected heterozygosity (Boca and Rosenberg 2011; Tal 2012). In terms of the underlying allele frequencies at a single locus *i* from two populations we have,

$$F_{ST}(i) = \frac{H_T - H_S}{H_T} = \frac{(q_i - p_i)^2}{(p_i + q_i)(2 - p_i - q_i)}$$

A multilocus F_{ST} requires separately averaging the numerator and denominator across loci (Weir 1996) such that our first proposal becomes,

$$C_n = F_{ST} = \frac{\sum_{i=1}^{n} (p_i - q_i)^2}{\sum_{i=1}^{n} (p_i + q_i)(2 - p_i - q_i)}$$

However, this proposal fails both the *Dominance* and the *Performance* criteria (e.g., it is evident that *Performance* fails when alleles across loci have the same frequency in each population: as *n* increases C_n remains constant).

3.2 Information-Theoretic Measures

A powerful measure of statistical dependency is the *mutual information*. The use of mutual information at a single-locus has been explored in the context of *feature selection* (Peng, Long and Ding 2005) and informativeness (Rosenberg, Li, et al. 2003). Let $X=\{0,1\}$ represent source populations 1 and 2 respectively, and let $Y_i=\{0,1\}$ represent the allele at our biallelic haploid locus *i*, with $p_i = Pr(Y_i=1 \mid X=0)$ and $q_i = Pr(Y_i=1 \mid X=1)$. Formally, assuming equal class priors,

$$(Y_i \mid X = 0) \sim Bernoulli(p_i), \quad (Y_i \mid X = 1) \sim Bernoulli(q_i), \quad X \sim Bernoulli(\tfrac{1}{2})$$

From basic definitions,

$$I(X;Y_i) = \sum_{x=0}^{1} \sum_{y_i=0}^{1} p(y_i, x) log \frac{p(y_i, x)}{p(y_i)\,p(x)} =$$

$$\frac{1-p_i}{2} log \frac{2(1-p_i)}{2-p_i-q_i} + \frac{p_i}{2} log \frac{2p_i}{p_i+q_i} + \frac{1-q_i}{2} log \frac{2(1-q_i)}{2-p_i-q_i} + \frac{q_i}{2} log \frac{2q_i}{p_i+q_i}$$

Note that this is identical to I_n (Rosenberg, Li, et al. 2003, Eq. 4). Averaging across loci and normalizing (assuming all logarithms are to base *e*) to comply with the required bounds for C_n,

$$C_n = \frac{1}{log(2)\cdot n}\sum_{i=1}^{n} I(X;Y_i)$$

However, this proposal fails the *Performance* criterion. Since mutual information can incorporate multivariate random variables, we may quantify the dependency of the source population (X) and the *joint* distribution alleles across n loci (Y_i).[***] From basic definitions of mutual information and conditional probability,

$$I(X;[Y_1,...,Y_n]) = \sum_{x=0}^{1}\sum_{y_i=0}^{1}\cdots\sum_{y_n=0}^{1} p(y_1,...,y_n,x)log\frac{p(y_1,...,y_n,x)}{p(y_1,...,y_n)\,p(x)} =$$

$$= \sum_{y_i=0}^{1}\cdots\sum_{y_n=0}^{1}(p(y_1,...,y_n \mid x=0)\,p(x=0)log\frac{p(y_1,...,y_n \mid x=0)}{p(y_1,...,y_n)} + \quad (1)$$

$$+ p(y_1,...,y_n \mid x=1)\,p(x=1)log\frac{p(y_1,...,y_n \mid x=1)}{p(y_1,...,y_n)})$$

Assuming linkage equilibrium, the joint multivariate distributions $[Y_1,\ldots,Y_n \mid X]$ and $[Y_1,\ldots,Y_n]$ may be expressed in terms of the allele frequencies $p(y_i \mid X)$,

$$p(y_1,...,y_n \mid X) = \prod_{i=1}^{n} p(y_i \mid X)$$

and from the law of total probability,

$$p(y_1,...,y_n) = p(y_1,...,y_n \mid X=0)\,p(X=0) + p(y_1,...,y_n \mid X=1)\,p(X=1)$$

Finally, we introduce class priors, $P(X=0)=\alpha, \quad P(X=1)=1-\alpha$, such that,

$$I(X;[Y_1,...,Y_n]) =$$

$$\sum_{k=0}^{2^n-1}\left[\alpha h_k log\frac{h_k}{\alpha h_k + (1-\alpha)g_k} + (1-\alpha)g_k log\frac{g_k}{\alpha h_k + (1-\alpha)g_k}\right] \quad (2)$$

Where (noting that by definition h_k=g_k=1 for n=0),

$$h_k = \prod_{i=1}^{n}|1-f_n(k,i)-p_i|, \quad g_k = \prod_{i=1}^{n}|1-f_n(k,i)-q_i| \quad (3)$$

and where f_n is an indicator function for traversing the 2^n genotypes of each population, transforming the n-multiple sum (1) to a single sum,

$$f_n(k,i) = \left\lfloor \frac{k}{2^i} \right\rfloor mod \;\; 2 \quad (the\; i^{th}\; bit\; of\; k)$$

If all logarithms are taken to base e then $I(X;[Y_1,...,Y_n])$ is bounded above by $log(2)$, which we use for normalization,

$$C_n = \frac{I(X;[Y_1,...,Y_n])}{log(2)}$$

This proposal fully satisfies the *Performance* criterion, even for unequal priors.[†††] However, this measure only satisfies the *Dominance* criteria for equal priors. It is sufficient to show by counter-example that *Dominance* fails at some scenario (illustrated by Fig. 1A). Furthermore, this measure

[***] A similar use of mutual information by (Peng, Long and Ding 2005) is targeted for *feature selection*, where the object is to find a feature set S with m features that *jointly* have the largest dependency on the target class – a theoretical scheme called *Max-Dependency*.

[†††] Interestingly, this C_n satisfies a strong version of the *Performance* criterion: it is monotonic *increasing* at each extra locus, rather than merely monotonic non-decreasing (also for unequal priors).

also fails the *Priors* criterion since C_n approaches 0 (instead of 1) as the prior approaches the extremes of 0 or 1 (Fig 1B).

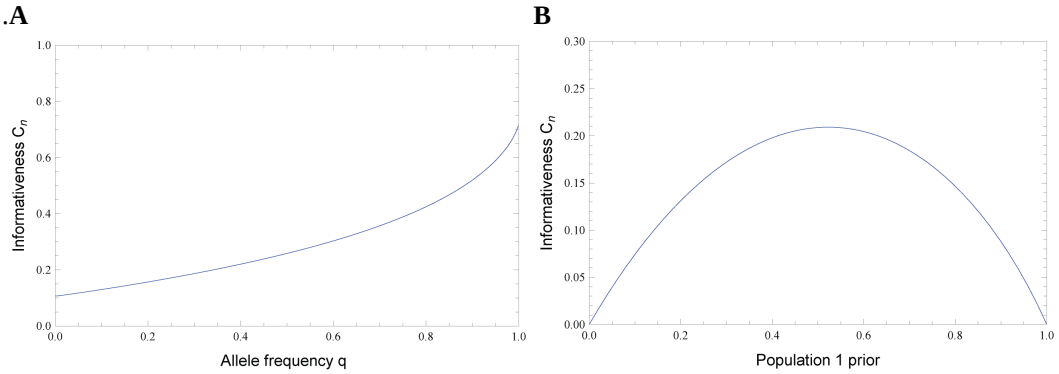

Fig 1. C_n based on multilocus mutual information fails two criteria. **A:** C_n does not satisfy the *Dominance* criterion for unequal population priors. Shown for population 1 prior α=0.2, and 4 loci with frequencies p_i=0.1/q_i=0.3 with the 5th locus reaching full differentiation, p_5=0/q_5→1.| **B:** C_n does not satisfy the *Priors* criterion. Shown for 5 loci with frequencies p_i=0.1/q_i=0.3 and prior α ranging from 0 to 1.

3.3 *f-divergence* Measures

An *f-divergence* is a function $D_f(P\|Q)$ that measures the difference between two probability distributions *P* and *Q*. The divergence is intuitively an average of the odds ratio given by *P* and *Q*, weighted by the function *f* that is convex over (0,∞) and satisfies $f(1) = 0$.

$$D_f(P \parallel Q) = \sum_{x\in\Omega} p(x)\, f(\log \frac{p(x)}{q(x)})$$

The use of *f*-divergences here is motivated by the properties they hold, which resemble our set of criteria. Such properties are non-negativity (equal to zero if and only if probability densities coincide), convexity, boundedness and various features of invariance (Cichocki and Amari 2010). Moreover, these functions are classically interpreted as modeling *discrimination information* between hypotheses (Toussaint 1975).

Kullback-Leibler Divergence

The *Kullback-Leibler* divergence can be intuitively considered as a distance measure between the two probability densities. For probability distributions *P* and *Q* of a discrete random variable their *KL* divergence (non-negative and unbounded) is defined to be,

$$D_{KL}(P \parallel Q) = \sum_{x\in\Omega} p(x) \log \frac{p(x)}{q(x)}$$

Kullback and Leibler defined a symmetric version of their divergence (Lin 1991),

$$D_{KL-S} = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)$$

The symmetric KL divergence seems to be a good candidate for C_n. In terms of the genotype probabilities *h* and *g* in (3),

$$D_{KL-S} = \sum_{k=0}^{2^n-1} \left[(h_k - g_k) \cdot \log(\frac{h_k}{g_k}) \right]$$

Normalization is required to transform its unbounded range of [0, ∞) (Nielsen and Boltz 2011) to [0, 1),

$$C_n = 1 - \frac{1}{1 + D_{KL-S}}$$

However, this formulation can be shown to fail the *Dominance* criterion. We also note that it does not have the *Population subadditivity* property, since the symmetric *KL* divergence fails the triangle inequality and is therefore not a metric (Khosravifard, Fooladivanda and Gulliver 2007).

Jensen-Shannon Divergence

A popular symmetrized version of *KL* divergence is the Jensen-Shannon (*JS*) divergence, which has some attractive features: it is symmetric, bounded and always defined (even for zero probabilities). A generalized form of the *JS* Divergence incorporates distribution weights (corresponding to our prior α),

$$D_{JS} = \alpha D_{KL}(P \parallel M) + (1-\alpha) D_{KL}(Q \parallel M)$$

where,

$$M = \alpha P + (1-\alpha) Q$$

In terms of the genotype probabilities *h* and *g* (2.3.0) and setting $m_k = \alpha h_k + (1-\alpha) g_k$,

$$D_{JS} = \sum_{k=0}^{2^n-1} \left[\alpha h_k \cdot \log\left(\frac{h_k}{\alpha h_k + (1-\alpha) g_k}\right) + (1-\alpha) g_k \cdot \log\left(\frac{g_k}{\alpha h_k + (1-\alpha) g_k}\right) \right]$$

Noting that the square root of the *JS* Divergence is a metric (Endres and Schindelin 2003) and using normalization, we produce the following expression,

$$C_n = \sqrt{D_{JS} / ln(2)}$$

It is immediately evident that D_{JS} is *precisely* in the form of the multilocus mutual information in (2). Therefore a formulation of C_n based on D_{JS} would also fail the *Dominance* criterion for *unequal* priors. An interesting corollary is that if $X \sim Bernoulli(1-\pi)$,[‡‡‡]

$$I(X;[Y_1,...,Y_n]) = D_{JS}(P(Y_1,...,Y_n \mid X=0) \parallel P(Y_1,...,Y_n \mid X=1)),$$

The Hellinger Distance

The *Hellinger* distance (H_D) is also a type of *f*-divergence and is related to the *Bhattacharyya* coefficient (B_C),[§§§]

$$H_D = \sqrt{1 - B_C}$$

$$B_C(P,Q) = \sum_{x \in \Omega} \sqrt{p(x)q(x)} = \sum_{k=0}^{2^n-1} \sqrt{h_k \cdot g_k}$$

Given that $0 \le H_D \le 1$ (Nielsen and Boltz 2011), we simply let,

$$C_n = H_D = \sqrt{1 - \sum_{k=0}^{2^n-1} \sqrt{h_k \cdot g_k}}$$

However, there is no way to introduce the class priors in this formulation: simply replacing the likelihoods h_k and g_k with the posterior probabilities αh_k and $(1-\alpha) g_k$ would render the priors arbitrarily interchangeable in the product.

---

‡‡‡ This result was implied in relation to *KL* Divergence and multiple classes (Vasconcelos and Vasconcelos 2004).

§§§ The *Bhattacharyya* coefficient is closely related to the Bayes error for the special case of equal priors and two classes, and can be used to provide upper and lower bounds for this error (Djouadi, Snorrason and Garber 1990; Aherne, Thacker and Rockett 1998).

The Mahalanobis Distance

The *Mahalanobis* squared distance (Δ^2) is not strictly an *f-divergence* but a part of a broader class of (*Bregman*) divergences, and may be seen as an extension of the Euclidean metric,

$$\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \quad \Sigma = \pi \cdot \Sigma_1 + (1-\pi) \cdot \Sigma_2$$

The population means, or *centroids*, are simply the vectors of our allele frequencies (irrespective of possible linkage disequilibrium or the choice of distance metric),

$$M_1 = (p_1,...,p_n) \quad M_2 = (q_1,...,q_n)$$

Assuming linkage equilibrium, the diagonal covariance matrix is the mean of the two covariance matrices,

$$\Sigma = \begin{pmatrix} \alpha p_1(1-p_1) + (1-\alpha) q_1(1-q_1) & & \\ & \alpha p_2(1-p_2) + (1-\alpha) q_2(1-q_2) & \\ & ..... & \\ & & \alpha p_n(1-p_n) + (1-\alpha) q_n(1-q_n) \end{pmatrix}$$

The distance between the two centroids is therefore,

$$\Delta(M_1, M_2) = \sqrt{\sum_{i=1}^{n} \frac{(p_i - q_i)^2}{\alpha p_i(1-p_i) + (1-\alpha) q_i(1-q_i)}}$$

Alternatively, using an L_1 norm $\| x - y \|_1 = \sum_{i=1}^{n} | x_i - y_i |$,

$$\Delta(M_1, M_2) = \sum_{i=1}^{n} \frac{| p_i - q_i |}{\sqrt{\alpha p_i(1-p_i) + (1-\alpha) q_i(1-q_i)}}$$

Finally, with proper normalization (since the distances are unbounded),

$$C_n = 1 - \frac{1}{1+\Delta}$$

This formulation (using either *Mahalanobis* or the *L1-Mahalanobis*) satisfies the *Dominance* criterion also for *unequal* priors, but fails the *Priors* criterion, as Fig. 2 demonstrates.

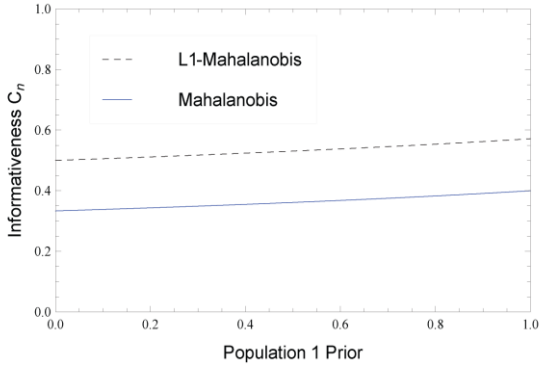

Fig 2. C_n based on *Mahalanobis distances* does not satisfy the *Priors* criterion (we require $C_n \to 1$ *if* $\alpha \to 0$ or $\alpha \to 1$); shown for 4 loci, p_i=0.1/q_i=0.2 and α ranging from 0 to 1.

3.4 Classifiers

The Naïve Bayes Classifier

Quantifying multi-locus information could also be approached from a *decision-theoretic* perspective, using insights from supervised learning classification schemes. The optimal classifier under known class-conditional densities is the *Bayes* or *maximum-likelihood* (ML) classifier, where data are classified according to the most probable class (due to the *MAP Rule*). The expected error or misclassification rate of the Bayes classifier is called the *Bayes error* (Hastie, Tibshirani and Friedman 2009). The standard assumption of linkage equilibrium within populations (absence of within-class dependencies) motivates the use of a *naïve Bayes* classifier, where class-conditional likelihoods are expressed as the product of allele frequencies across the independent loci (Cornuet, et al. 1999; Phillips, et al. 2007). The Bayes error can be formulated as a prior-weighted sum of probabilities over the 2^n possible genotypes indexed by k,

$$E_n = \sum_{k=0}^{2^n-1} min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) \tag{4}$$

where the genotype frequencies g_k and h_k are defined in (3). To gain some intuition consider the expressions for E_1 and E_2,

$$\begin{aligned} E_1 &= min(\alpha p_1, (1-\alpha) q_1) + min(\alpha(1-p_1), (1-\alpha)(1-q_1)) \\ E_2 &= min(\alpha p_1 p_2, (1-\alpha) q_1 q_2) + min(\alpha p_1 (1-p_2), (1-\alpha) q_1 (1-q_2)) + \\ &\quad + min(\alpha(1-p_1) p_2, (1-\alpha)(1-q_1) q_2) + \\ &\quad + min(\alpha(1-p_1)(1-p_2), (1-\alpha)(1-q_1)(1-q_2)) \end{aligned} \tag{5}$$

Since for two classes the error rate E of any classifier ranges 0 to ½, a straightforward transformation conforming with the *Range* criterion of [0,1) is C_n=1-2E,

$$C_n = 1 - 2\sum_{k=0}^{2^n-1} min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) \tag{6}$$

This is equivalent to the formulation of *ORCA* (Rosenberg, Li, et al. 2003). An alternative formulation derives from the equivalence of the Bayes error for two classes and the *variational distance* (a form of *f-divergence*), even for unequal priors (Nguyen, Wainwright and Jordan 2009, section 2.1.1); see Appendix A here for a simple proof given *equal* priors.[****] This equivalence results in an alternative expression to (6),

$$C_n = \sum_{k=0}^{2^n-1} \left| \alpha \cdot h_k - (1-\alpha) \cdot g_k \right| \tag{7}$$

This formulation of C_n obeys all 12 criteria and attains the extra properties (Appendix B). Finally, we note that if linkage disequilibrium is nevertheless a feature of the data, the naïve Bayes classifier will have reduced performance, but C_n would remain compliant with all criteria and properties. Fig. 3 is a numerical simulation of two criteria.

[****] For equal priors it is the only *f*-divergence that is a metric (Khosravifard, Fooladivanda and Gulliver 2007).

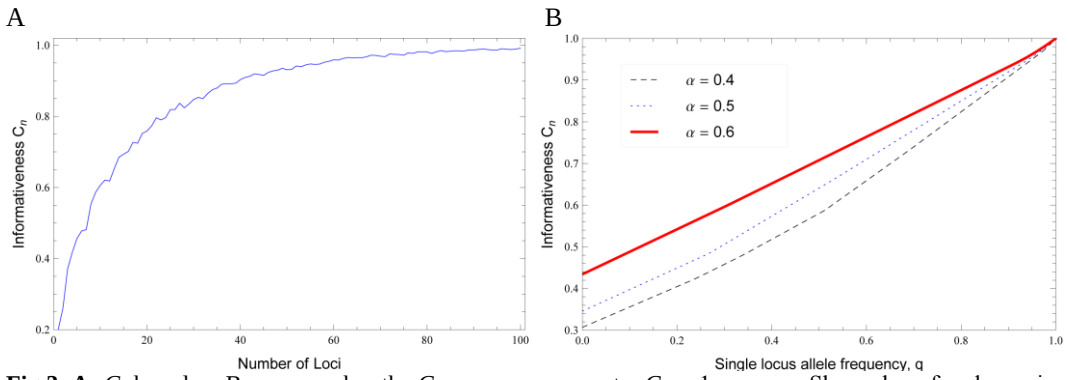

Fig 3. A: C_n based on Bayes error has the *Convergence* property: $C_n \to 1$ as $n \to \infty$. Shown here for class prior α=0.4 and allele frequencies p_i=0.1/q_i=0.3 using Monte Carlo simulation of (6). | B: C_n based on Bayes error has the *Dominance* property: $C_n \to 1$ as $|q_i - p_i| \to 1$. Shown here for three cases of class priors, under a five loci scenario, where q_5 changes from ≈0 through ≈1 while $p_5 \approx 0$; the other 4 loci have frequencies: p_1=0.01/q_1=0.3, p_2=0.2/q_2=0.35, p_3=0.4/q_3=0.3, p_4=0.1/q_4=0.15.

4 Discussion and Conclusion

In this paper, we have formulated a measure complying with a predefined set of criteria to capture the informativeness of a collection of markers for population assignment. The informativeness may be seen as the reduction in uncertainty regarding the ancestry of the genotype given these markers. This reduction in uncertainty originates from the fact that each ancestral population has a distinct distribution of over this set of markers (Bercovici and Geiger 2009). We have taken an approach analogous to Shannon's axiomatic program for deriving a measure of the rate of information produced by a discrete information source. Criteria for an information measure are first justified from intuitive considerations, and subsequently employed to arrive at viable formulations. A similar approach was adopted by Lewontin in a widely cited paper on the apportioning of genetic variation in human populations (Lewontin 1972). Lewontin had adopted Shannon's information measure *H* as a diversity measure for the purpose of ascertaining average population differentiation with respect to a limited set of genetic loci. Four characteristics reminiscent of Shannon's properties for information were specified as a requirement for any (single-locus) diversity measure: [a] minimum when a single allele is present, [b] maximum when all alleles have equal frequencies, [c] generally increase with the number of alleles present, and finally, [d] the pooling of two populations should result in a higher diversity, compared to the average of their separate diversities. In particular, while the first three characteristics are equivalent to three of the properties of *H* proposed by Shannon, the fourth is original, deemed pertinent specifically to a measure of genetic diversity.

We have shown that a measure of informativeness based on an optimal Bayes classifier complies with the full set of the proposed criteria. However, it is important to emphasize that it cannot be assumed a-priori that any classifier operating on genetic data could similarly be utilized. For e.g., the simple 'population-trait' classifier (Witherspoon, et al. 2007), with an error rate modeled by *generalized binomial* distributions for haploid populations (Tal 2012), can be shown to fail the *Dominance* criterion, and in some instances, the *Performance* criterion. There does not seem to be a simple general heuristic for identifying classifiers that may be utilized for the present purpose; instead, each candidate must be checked against all target criteria.

Interestingly, compliance with a subset of criteria and one property would deem C_n a *metric* above a set of populations, for equal priors ($\alpha = ½$ is required from criterion #1). A metric on a set X is a *distance function* d: $X \times X \rightarrow \mathbf{R}$, where $\mathbf{R}$ is the set of real numbers. If x, y, z are in X, this function is required to satisfy four conditions, which express intuitive notions about the concept of distance:[††††]

1. $d(x, y) \geq 0$ (our *non-negativity* criterion)[‡‡‡‡]
2. $d(x, y) = 0$ if and only if $x = y$ (our *zero* criterion)
3. $d(x, y) = d(y, x)$ (our *symmetry* criterion)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (our *Population subadditivity* property)

Indeed, if we represent populations by their allele frequency vectors, then C_n may be seen as a *population distance metric*. In contrast, standard differentiation measures such as F_{ST} are often perceived as population distances but are not properly metrics (see (Jardine 1971) for early proposals of *dissimilarity coefficients* as metrics over a set of populations).[§§§§]

It was also shown that a previously proposed *multi-locus* information-theoretic measure based on mutual information I_n (Rosenberg, Li, et al. 2003) fails both the *Dominance* and *Priors* criteria under unequal population priors (Fig. 1). Therefore, I_n could not function as a measure of *absolute* informativeness; i.e., it could not be used to compare the information content of a collection of loci from one set of populations with a collection of loci from a *different set of populations*. For instance, consider two populations of *unequal size* where a different allele is fixed in each at a single locus (i.e., maximum frequency differentiation, $|q\text{-}p|=1$). If we compute a multi- or single-locus I_n that includes this locus, we would have $I_n<1$ (since it was shown that *Dominance* fails for unequal priors), although it is obvious that any proper classification method of unknown genotypes should achieve a 100% success rate. In contrast, a different pair of populations without complete differentiation at any single locus may have higher I_n although it is obvious that there is less information for correct assignment. We note that I_n can still function perfectly well for its intended purpose of *panel selection* (also known as 'feature selection' in statistical learning applications), where different sets of loci (or single markers) from the *same* meta population are compared for highest informativeness.

Future work will be required to extend the set of criteria and resulting formulations to capture various aspects of practical applications – the use of population samples for estimating allele frequencies and the possible presence of linkage disequilibrium. In addition, it would be useful to generalize the model to comply with diploid genotypes, multiple population scenarios and multi-allelic loci.

5 Appendix

Appendix A – *The equivalence of Bayes error and the variational distance*

Here we develop a proof for *equal priors* only. For the purposes of this proof, p_i and q_i do not denote allele frequencies but rather genotype frequencies. It is required to prove that,

$$\sum_i min(p_i, q_i) = 1 - \frac{1}{2}\sum_i |\, p_i - q_i \,|,$$

where,

[††††] While $C_n(P,Q)$ is a metric under equal priors, it is not a *norm* since the sum $P+Q$ is meaningless.

[‡‡‡‡] Note that the upper bound of 1 from criterion #3 does not detract from the metric quality of C_n. Take for example the distance metric $d(x,y)=min\{|x\text{-}y|,1\}$.

[§§§§] Another option is to standardize by the number of loci, using C_n/n, to remove the dependency on n.

$$\sum_i p_i = 1, \sum_i q_i = 1, \quad p_i, q_i > 0, \quad i = 2^k, k = 1,2,3....$$

Proof: Let S denote the subset of all indices i, for which $p_i<q_i$ and T denote all other indices i. Then for all p_i, q_i whose indices are in S we have $|p_i-q_i|=q_i-p_i$; similarly, for all p_i and q_i whose indices are in T we have $|p_i-q_i|=p_i-q_i$. Therefore,

$$2 - \sum_i |\, p_i - q_i \,| = \sum_i p_i + \sum_i q_i + \sum_{i\in S}(p_i - q_i) + \sum_{i\in T}(q_i - p_i) =$$
$$= (\sum_{i\in S} p_i + \sum_{i\in T} p_i + \sum_{i\in S} q_i + \sum_{i\in T} q_i) + \sum_{i\in S}(p_i - q_i) + \sum_{i\in T}(q_i - p_i) =$$
$$= (\sum_{i\in S} p_i + \sum_{i\in S} q_i + \sum_{i\in S} p_i - \sum_{i\in S} q_i) + (\sum_{i\in T} p_i + \sum_{i\in T} q_i + \sum_{i\in T} q_i - \sum_{i\in T} p_i) =$$
$$= 2\sum_{i\in S} p_i + 2\sum_{i\in T} q_i = 2\sum_{i\in S} min(p_i, q_i) + 2\sum_{i\in T} min(p_i, q_i) = 2\sum_i min(p_i, q_i)$$

Thus, $1 - \frac{1}{2}\sum_i |\, p_i - q_i \,| = \sum_i min(p_i, q_i)$.

Appendix B – *Proof of compliance of Bayes C_n with all criteria*

[1] *Zero*: If $p_i = q_i$ for all i then $h_k = g_k$ for all k, and therefore from (7),

$$C_n = |\alpha - (1-\alpha)| \cdot \sum_{k=0}^{2^n-1} h_k = |\alpha - (1-\alpha)| \cdot 1 = |2\alpha - 1|$$

And if $\alpha = ½$ (equal priors) we have $C_n = 0$, and also $C_0 = 0$. Note that since $C_n(P,P)=|2\alpha -1|$ it is follows that with *unequal* priors C_n cannot strictly be a metric over vectors of frequencies.

[2] *Non-negativity*: $C_n \geq 0$ follows from the sum of absolute values in (7) being ≥ 0. A corollary of #1 above provides the stronger result of $C_n \geq |2\alpha -1|$.

[3] *Bound*: $C_n < 1$ follows from the lower bound of zero of the Bayes error and the relation $C_n=1-2E_n$.

[4] *Performance*: $C_{n+1}\geq C_n$ follows from $E_{n+1}\leq E_n$, where E_n is the Bayes error given the same distributions (Appendix B.1). An interesting corollary is incorporating an extra locus is not always informative and will not improve classification performance. Formally, without loss of generality,

$$\forall p_{n+1}, \exists\, \varepsilon > 0, C_{n+1}((p_1,...,p_n,p_{n+1}),(q_1,...,q_n,p_{n+1}+\varepsilon)) = C_n(P,Q).$$

It can be shown that the range of ε for which this invariance is satisfied is wider with more unequal priors.

[5] *Convergence*: First, note that a corollary of criteria #2 and #3 is only that $lim_{n\to\infty} C_n \in [|\, 2\alpha - 1\,|, 1]$, a weaker result than is demanded by this criterion.[*****] For the proof that $C_n \to 1$ as $n\to\infty$ see Appendix B.2.

[6] *Neutrality*: To prove that $C_n=C_{n+1}$ if $p_{n+1}=q_{n+1}$ first note from (5) and (6) that C_{n+1} has twice the number of terms than C_n. After arranging of terms we have,

$$C_{n+1} = p_{n+1}C_n + (1 - p_{n+1})C_n = C_n$$

[7] *Continuity*: Since there are no singularities in C_n and since $p_i=q_i$ are real-valued parameters, C_n is continuous with respect to its parameters.

[*****] The compliance of C_n with this criterion suggests that Theorem 4 in (Rosenberg 2005), which proves for the *informativeness* measure I_n that convergence at infinity is to a number possibly smaller than 1 (for two populations and prior α, $lim_{n\to\infty} I_n \in [max\{\alpha, 1-\alpha\}, 1]$) is a *weaker* version of this criterion. Our stronger version applies under practical implementations, where frequencies of polymorphisms analyzed differ by any minimal amount between populations.

[8] *Dominance*: To see that $C_n \to 1$ *if* $|q_i - p_i| \to 1$ consider the expression for E_n in (4) and without loss of generality examine the effect of $|q_n - p_n| \to 1$. Each of the 2^n summands is one of either four forms,

$$[a] \quad min\{\alpha p_1 \cdots p_n,\ (1-\alpha)q_1 \cdots q_n\}$$
$$[b] \quad min\{\alpha(1-p_1) \cdots p_n,\ (1-\alpha)(1-q_1) \cdots q_n\}$$
$$[c] \quad min\{\alpha p_1 \cdots (1-p_n),\ (1-\alpha)q_1 \cdots (1-q_n)\}$$
$$[d] \quad min\{\alpha(1-p_1) \cdots (1-p_n),\ (1-\alpha)(1-q_1) \cdots (1-q_n)\}$$

Now, from $|q_i - p_i| \to 1$, if $p_n \to 0$ then $q_n \to 1$ and all summands in the form of *a* and *b* reduce to zero due to the *first* term, and all summands in the form of *c* or *d* reduce to zero due to the *second* term. If $p_n \to 1$ then $q_n \to 0$ and all summands in the form of *a* and *b* reduce to zero due to the *second* term, and all summands in the form of *c* and *d* reduce to zero due to the *first* term. This results in $E_n \to 0$ and consequently $C_n \to 1$.

[9] *Loci invariance*: Since the genotype probabilities h_k and g_k in (6) are each a commutative product of allele frequencies from all loci, C_n is invariant to different ordering of loci.

[10] *Symmetry invariance*: The truth of $C_n(P,Q) = C_n(Q,P)$ simply follows from the presence of an *absolute value* in the formulation of (7).

[11] *Allocation invariance*: The simultaneous substitution of p_i with $(1\text{-}p_i)$ and q_i with $(1\text{-}q_i)$ simply changes the order of the summation terms in (7) and thus does not affect C_n.

[12] *Priors*: If $\alpha \to 0$ or $\alpha \to 1$ then one of the two terms within the sum in (7) diminishes to zero and what remains in the limit is the sum over all genotype probabilities in one population, which equals 1, and therefore $C_n \to 1$.

The formulation of C_n in terms of the Bayes error also conforms to the two extra properties:

[a] *Triviality*: In the general case from (7),

$$C_1 = \sum_{k=0}^{1} |\, h_k\alpha - g_k(1-\alpha)\,| = |\, p_1\alpha - q_1(1-\alpha)\,| + |\,(1-p_1)\alpha - (1-q_1)(1-\alpha)\,|$$

and for equal priors,

$$C_1 = \frac{1}{2}|\, p_1 - q_1\,| + \frac{1}{2}|\,(1-p_1) - (1-q_1)\,| = |\, p_1 - q_1\,|$$

[b] *Population subadditivity*: The compliance of C_n with the triangle inequality follows from the formulation in (7) which describes the *variational distance* - a measure that satisfies the triangle inequality, even for unequal priors (Khosravifard, Fooladivanda and Gulliver 2007).

Appendix B.1 – *Proof of the Performance criterion for Bayes-based* C_n

We are required to prove that $C_{n+1} \geq C_n \ \ \forall n \geq 0$ where, $0 < p_i, q_i < 1 \quad i = 1 \ldots n$.

For simplicity, instead of C_n we examine on E_n, where $C_n = 1\text{-}2E_n$, and would like to prove that $E_{n+1} \leq E_n$. From (4),

$$E_n = \sum_{k=0}^{2^n-1} min(\alpha \cdot h_k, (1-\alpha) \cdot g_k)$$

In general, the *k*-th summand X_k of E_n (there are 2^n) is split to $X_{2k} + X_{2k+1}$ in E_{n+1},

$$E_n = \ldots + \tfrac{1}{2} min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) + \ldots$$

$$E_{n+1} = \ldots + \tfrac{1}{2} min(\alpha \cdot h_k p, (1-\alpha) \cdot g_k q) + \frac{1}{2} min(\alpha \cdot h_k (1-p), (1-\alpha) \cdot g_k (1-q)) + \ldots$$

where *p*, *q* are the frequencies at the (*n*+1)-th locus. There are two cases:

If $min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) = \alpha \cdot h_k$, then by simple algebra,

$$min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) = \alpha \cdot h_k = \alpha \cdot h_k p + \alpha \cdot h_k (1-p) \geq$$
$$min(\alpha \cdot h_k p, (1-\alpha) \cdot g_k q) + min(\alpha \cdot h_k (1-p), (1-\alpha) \cdot g_k (1-q))$$

else, $min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) = (1-\alpha) \cdot g_k$, and then similarly,

$$min(\alpha \cdot h_k, (1-\alpha) \cdot g_k) = (1-\alpha) \cdot g_k = (1-\alpha) \cdot g_k q + (1-\alpha) \cdot g_k (1-q) \geq$$
$$min(\alpha \cdot h_k p, (1-\alpha) \cdot g_k q) + min(\alpha \cdot h_k (1-p), (1-\alpha) \cdot g_k (1-q))$$

Finally, summing these inequalities over k:1 to n, results in $E_{n+1} \leq E_n$. [†††††]

Appendix B.2 - *Proof of the Convergence criterion for Bayes-based C_n*

We need to prove that for any *predefined* value $0<\varepsilon<1$, if $|q_i-p_i|>\varepsilon$ for all i then $C_n \rightarrow 1$. We prove here the equivalent result, $E_n \rightarrow 0$, where $C_n = 1-2E_n$. First, we relate E_n to the error rate of the 'population-trait' classifier (Witherspoon, et al. 2007), modeled by *generalized* binomial distributions (Tal 2012), call it G_n, by making a small correction of some allele frequencies: the generalized binomial model requires that $p_i<q_i$, so whenever this is not the case replace p_i with $1-p_i$ and q_i with $1-q_i$, so that $q_i-p_i>\varepsilon$. This does not change the Bayes error, as can be seen from,

$$E_{n+1} = ... + \tfrac{1}{2} min(\alpha \cdot h_k p, (1-\alpha) \cdot g_k q) + \tfrac{1}{2} min(\alpha \cdot h_k (1-p), (1-\alpha) \cdot g_k (1-q)) + ...$$

Now, since the Bayes error is a lower bound on the error achievable from any classifier we have $E_n \leq G_n$. Next, we relate G_n to the misclassification rate from a simple binomial model (Tal, 2012), call it B_n. For a population with allele frequencies p_i the *generalized* binomial distribution has mean and variance,

$$E(S_n) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p_i = n\pi$$

$$Var(S_n) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} p_i (1-p_i) = n\pi - n\pi^2 - (\sum_{i=1}^{n} p_i^2 + n\pi^2 - 2\pi \sum_{i=1}^{n} p_i) =$$
$$= n\pi(1-\pi) - nV_p$$

where π and V_p are the mean and variance of p_i, respectively. Since $Var(X)$ is always lesser or equal to a binomial with $p=\pi$, and since $q_i - p_i > \varepsilon$ for all i implies that $\pi_q - \pi_p > \varepsilon$, we have that $G_n \leq B_n$. Next, we show that $B_n \rightarrow 0$ as $n \rightarrow \infty$, given $q-p > \varepsilon$.

$$B_n = \sum_{i=0}^{n} min\left\{ \alpha \cdot \binom{n}{i} p^i (1-p)^{n-i}, (1-\alpha) \cdot \binom{n}{i} q^i (1-q)^{n-i} \right\}$$

Since the minimum of two nonnegative quantities is bounded by their geometric mean, we have,

$$B_n \leq \alpha(1-\alpha) \sum_{i=0}^{n} \binom{n}{i} p^{i/2} q^{i/2} (1-p)^{(n-i)/2} (1-q)^{(n-i)/2}$$

We now note that the geometric mean is bounded by the arithmetic mean, with equality *iff* the two quantities are equal. In particular, $\sqrt{p}\sqrt{q} = r(p+q)/2$ for some $0 < r < 1$, and $\sqrt{(1-p)}\sqrt{(1-q)} = s(1-(p+q)/2)$ for some $0 < s < 1$. Note we assume $p \neq q$ here. So,

[†††††] A different proof can be found in (Rosenberg 2005).

$$B_n \le \alpha(1-\alpha)\sum_{i=0}^{n}\binom{n}{i} r^i s^{n-i} (\frac{p+q}{2})^i (1-\frac{p+q}{2})^{n-i}$$

Next, observe that $r^i s^{n-i} \le t^n$ where $0 < t = max(r,s) < 1$. So we have,

$$B_n \le \alpha(1-\alpha)t^n \sum_{i=0}^{n}\binom{n}{i} (\frac{p+q}{2})^i (1-\frac{p+q}{2})^{n-i}$$

By the binomial theorem the sum is just 1. So we get,

$$B_n \le \alpha(1-\alpha)t^n$$

Since $0 < t < 1$, we finally get $lim_{n\to\infty} B_n = 0$ as needed.

Finally, from $E_n \le G_n \le B_n$ we have $E_n \to 0$ as $n \to \infty$, and from $C_n = 1\text{-}2E_n$ we get $C_n \to 1$.

Acknowledgements

The author would like to thank Jonathan Rosenblatt and Tamir Tassa for helpful suggestions on mathematical and computational issues.

References

Aherne, F., N. Thacker, and P. Rockett. "The Bhattacharyya metric as an absolute similarity measure for frequency coded data." *Kybernetika, vol. 34, no. 4*, 1998: 363–368.

Bercovici, S., and D. Geiger. "Inferring Ancestries Efficiently in Admixed Populations with Linkage Disequilibrium." *Journal of Computational Biology 16:8*, 2009.

Bergstrom, C., and M. Rosvall. "The transmission sense of information." *Biology and Philosophy 26(2)*, 2011: 159-176.

Boca, S.M., and N.A. Rosenberg. "Mathematical properties of Fst between admixed populations and their parental source populations." *Theoretical Population Biology 80*, 2011: 208-216.

Casto, AM, JZ Li, D Absher, R Myers, S Ramachandran, and MW. Feldman. "Characterization of X-Linked SNP genotypic variation in globally distributed human populations." *Genome Biology 11:R10*, 2010.

Cichocki, A., and S-I. Amari. "Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities." *Entropy, 12*, 2010: 1532-1568.

Cornuet, J.M., S. Piry, G. Luikart, A. Estoup, and M. Solignac. "New methods employing multilocus genotypes to select or exclude populations as origins of individuals." *Genetics 153*, 1999.

Djouadi, A., O. Snorrason, and F. Garber. " The quality of Training-Sample estimates of the Bhattacharyya coefficient." *IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1)*, 1990: 92–97.

Edwards, A.W.F. "Human genetic diversity: Lewontin's fallacy." *BioEssays 25*, 2003: 798-801.

Endres, D.M., and J. E. Schindelin. "A new metric for probability distributions." *IEEE Trans. Inf. Theory, 49*, 2003: 1858–60.

Estoup, A., and B. Angers. "Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations." In *Advances in Molecular Ecology*, by G. Carvalho, 55-86. NATO press, 1998.

Frigg, R., and C. Werndl. "Entropy - A Guide for the Perplexed." In *Probabilities in Physics*, by Beisbart & S. Hartmann, 115–42. Oxford University Press, 2011.

Godfrey-Smith, P. "Information in Biology." In *The Cambridge Companion to the Philosophy* of *Biology*, by M Ruse and Hull D. Cambridge University Press, 2006.

Hastie, T., R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning, 2nd Edition.* Springer Verlag, 2009.

Jardine, N. "Patterns of Differentiation Between Human Local Populations." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. Vol. 263, No. 846*, 1971: 1-33.

Khosravifard, M., D. Fooladivanda, and T.A. Gulliver. "Confliction of the convexity and metric properties in f-divergences." *IEICE Trans. on Fundamentals E90-A 9*, 2007: 1848–1853.

Lewontin, R. "The apportionment of human diversity." *Evolutionary Biology, 6*, 1972: 381-398.

Lin, J. "Divergence measures based on the shannon entropy." *IEEE Transactions on Information Theory 37 (1)*, 1991: 145–151.

Maynard Smith, J. "The Concept of Information in Biology." *Philosophy of Science 67*, 2000: 177-194.

Nguyen, X., M.J. Wainwright, and M.I. Jordan. "On surrogate loss functions and f-divergences." *Annals of Statistics, 37*, 2009: 876-904.

Nielsen, F., and S. Boltz. "The Burbea-Rao and Bhattacharyya centroids." *IEEE Transactions on Information Theory, arXiv:1004.5049v2*, 2011.

Peng, H., F. Long, and C. Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8)*, 2005: 1226–1238.

Phillips, C., et al. "Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs." *Forensic Sci Int Genetics 1*, 2007: 233–235.

Plotkin, J.B., and M.A. Nowak. "Language evolution and information theory." *Journal* of *Theoretical Biology, 205(1)*, 2000: 147--159.

Rannala, B., and J. L. Mountain. "Detecting immigration by using multilocus genotypes." *Proc. Natl. Acad. Sci. USA 94*, 1997: 9197-9221.

Reza, F.M. *An Introduction to Information Theory.* New York: McGraw-Hill. New York: Dover 2010 Ed., 1961.

Rosenberg. "Algorithms for selecting informative marker panels for population assignment." *Journal of Computational Biology 12*, 2005: 1183-1201.

Rosenberg, N.A. "A Population-Genetic Perspective on the Similarities and Differences among Worldwide Human Populations." 2011: Human Biology: Vol. 83: Iss. 6.

Rosenberg, N.A., L.M. Li, R. Ward, and J.K. Pritchard. "Informativeness of Genetic Markers for Inference of Ancestry." *Am. J. Hum. Genet., 73*, 2003: 1402-1422.

Shannon, C.E. "A mathematical theory of communication." *The Bell System Technical Journal, 27*, 1948: 379–423.

Smith, R.D. "Information Theory and Population Genetics." *submitted to PLoS One*, 2011.

Tal, O. "The Cumulative Effect of Genetic Markers on Classification Performance: Insights from Simple Models." *Journal of Theoretical Biology. Volume 293, 21 January*, 2012: 206-218.

Toussaint, G.T. "Sharper lower bounds for discrimination information in terms of variation." *IEEE Trans. inform. Theory, vol. IT-21. no. 1*, 1975: 99-100.

Vasconcelos, N., and M. Vasconcelos. "Scalable Discriminant Feature Selection for Image Retrieval and Recognition." *CVPR (2)*, 2004: 770-775.

Wang, J. "Informativeness of genetic markers for pairwise relationship and relatedness inference." *Theoretical Population Biology 70*, 2006: 300–321.

Weir, B.S. *Genetic Data Analysis II: Methods for discrete population genetic data (2nd. ed.).* Sinauer Assoc., Sunderland, MA. , 1996.

Witherspoon, D.J., et al. "Genetic similarities within and between human populations." *Genetics 176*, 2007: 351–359.