Journal of Machine Learning Research 25 (2024) 1-27 Submitted 4/19; Revised 2/24; Published 12/24

Approximate Information Tests on Statistical Submanifolds

Michael W. Trosset MTROSSET@IU.EDU
Department of Statistics
Indiana University
Bloomington, IN 47408, USA

Carey E. Priebe CEP@JHU.EDU
Department of Applied Mathematics & Statistics
Johns Hopkins University
Baltimore, MD 21218-2682, USA

Editor: Edo Airoldi

Abstract

Parametric inference posits a statistical model that is a specified family of probability distributions. Restricted inference, for example, restricted likelihood ratio testing, attempts to exploit the structure of a statistical submodel that is a subset of the specified family. We consider the problem of testing a simple hypothesis against alternatives from such a submodel. In the case of an unknown submodel, it is not clear how to realize the benefits of restricted inference. To do so, we first construct information tests that are locally asymptotically equivalent to likelihood ratio tests. Information tests are conceptually appealing but (in general) computationally intractable. However, unlike restricted likelihood ratio tests, restricted information tests can be approximated even when the statistical submodel is unknown. We construct approximate information tests using manifold learning procedures to extract information from samples of an unknown (or intractable) submodel, thereby providing a roadmap for computational solutions to a class of previously impenetrable problems in statistical inference. Examples illustrate the efficacy of the proposed methodology.

Keywords: Restricted Inference, Dimension Reduction, Information Geometry, Minimum Distance Test

1. Introduction

An engrossing challenge arises when an appropriate statistical model is a subset of a familiar family of probability distributions: how to exploit the structure of the restricted model for the purpose of subsequent inference? This challenge encompasses theoretical, methodological, computational, and practical concerns. The reasons to address these concerns are especially compelling when the restricted model is of lower dimension than the unrestricted model, as parsimony principles encourage the selection of less complicated models.

The following example illustrates the concerns of the present manuscript.

Motivating Example Consider a multinomial experiment with 7 possible outcomes and probability vector $\theta \in \Re^7$. To test the simple null hypothesis

$$H_0 : \theta = \bar{\theta} = (0.09, 0.09, 0.09, 0.25, 0.16, 0.16, 0.16)$$

at significance level $\alpha = 0.05$, we perform $n = 30$ trials and observe

$$o = (3, 5, 4, 6, 9, 2, 1).$$

Should we reject H_0?

The likelihood ratio test statistic of

$$G^2 = 2 \sum_{j=1}^{7} o_j \log \left(o_j / n\bar{\theta}_j \right) = 11.93649$$

results in an (approximate) significance probability of $\mathbf{p} = 0.0634$. Pearson's $X^2 = 11.23519$ results in $\mathbf{p} = 0.0814$. Neither test provides compelling evidence against H_0.

Suppose, however, that it is possible to perform an auxiliary experiment that randomly generates possible values of θ for the primary experiment. The auxiliary experiment is performed $m = 100$ times and it is found that 96% of the variation in the $m = 100$ values of θ is explained by 2 principal components. This finding suggests the possibility that θ is restricted to a (slightly curved) 2-dimensional submanifold of the 6-dimensional simplex Δ^6. Can this revelation be exploited to construct a more powerful test?

If the submanifold was known, then one could perform a restricted likelihood ratio test. But *the submanifold is not known.* ■

In fact, the family of multinomial distributions provides numerous examples of dimension-restricted submodels. In statistical genetics, the phenomenon of Hardy-Weinberg equilibrium corresponds to a much-studied 1-parameter subfamily of trinomial distributions. Spherical subfamilies of multinomial distributions (Gous, 1999) are potentially valuable in a variety of applications, for example, text mining (Hall and Hoffman, 2000). In a recent effort to discover brainwide neural-behavioral maps from optogenetic experiments on *Drosophila* larvae (Vogelstein et al., 2014), each neuron line was modeled by a 29-dimensional vector of multinomial probabilities but the available evidence suggested that these vectors resided on an unknown 4-dimensional submanifold. These examples suggest a natural progression, from a submodel that is known and tractable, to a submodel that is known but possibly intractable, to an unknown submodel that can be sampled, to an unknown submodel that must be estimated. The particular challenge of how to exploit low-dimensional structure that is apparent but unknown motivated our investigation. The present manuscript addresses the case of known submodels and unknown submodels that can be sampled; a sequel will address the case of unknown submodels that must be estimated.

For unknown submodels that can be sampled, we propose the computationally intensive *approximate information test* summarized in Figure 1. The theory that underlies and motivates this procedure originates in information geometry, specifically in the well-known fact that Fisher information induces Riemannian structure on a statistical manifold. It leads to *information tests* that are conceptually appealing but (in general) computationally intractable. Approximate information tests circumvent the intractability of information tests.

Sections 2–5 develop and illustrate the theory of information tests. Section 2 establishes the mathematical framework that informs our investigation. We review the fundamental concepts of a statistical manifold and the Riemannian structure induced on it by Fisher

Suppose that known distributions $\bar{p}, p_1, \ldots, p_m$ lie on an unknown statistical submanifold. To test $H_0 : p = \bar{p}$ against alternatives that lie on the submanifold, we propose the following procedure.

1. Compute h_{ij}, the pairwise Hellinger distances between $\bar{p}, p_1, \ldots, p_m$.
2. Construct $\mathcal{G}$, a graph whose vertices correspond to the known distributions. Connect vertices i and j when h_{ij} is sufficiently small.
3. Compute the pairwise shortest path distances in $\mathcal{G}$.
4. Construct $\bar{z}, z_1, \ldots, z_m \in \Re^r$, an embedding of $\mathcal{G}$ whose pairwise Euclidean distances approximate the pairwise shortest path distances.
5. From $x_1, \ldots, x_n \sim p$, construct a nonparametric density estimate $\hat{p}_n$. Compute the Hellinger distances of $\hat{p}_n$ from $p_1, \ldots, p_m$ and embed $\hat{p}_n$ as $y(\vec{x}) \in \Re^r$ in the previously constructed Euclidean representation. The proposed test rejects $H_0 : \theta = \bar{\theta}$ if and only if the test statistic $\|y(\vec{x}) - \bar{z}\|$ is sufficiently large.
6. Estimate a significance probability by generating simulated random samples from the hypothesized distribution $\bar{p}$.

Figure 1: An approximate information test for the case of an unknown submodel that can be sampled. Steps 2–4 are essentially isomap (Tenenbaum et al., 2000), used here to represent the Riemannian structure of a statistical manifold rather than a data manifold. Details are provided in Section 6.

information. We demonstrate that information distance, that is, geodesic distance on this Riemannian manifold, is more practically derived from Hellinger distance, and we briefly review minimum Hellinger distance estimation. Sections 3–5 develop tests of simple null hypotheses using the concept of information distance. Section 3 demonstrates that information tests are locally asymptotically equivalent to various classical tests (Hellinger distance, Wald, likelihood ratio, and Hellinger disparity distance). Section 4 derives information tests for submodels of the multinomial model. Section 5 provides examples using the Hardy-Weinberg submodel of the trinomial model.

Despite their conceptual appeal, the information tests developed in Sections 3–5 are of limited practical application. Hence, our primary contribution lies in Section 6, which proposes a discrete approximation of an information test and illustrates its effectiveness in two cases for which an unknown submodel can be sampled. Section 7 reports a small simulation study designed to explore the effect of sampling density on performance. Section 8 discusses implications and possible extensions.

2. Preliminaries

2.1 Statistical Manifolds

We begin by recalling some basic properties of differentiable manifolds. See Matsushima (1972) for a more detailed explication of these concepts. Let M denote a completely separable Hausdorff space. Let $U \subseteq M$ and $V \subseteq \Re^k$ denote open sets. If $\varphi : U \to V$ is a homeomorphism, then $\varphi(u) = (x_1(u), \ldots, x_k(u))$ defines a coordinate system on U. The x_i are the coordinate functions and φ^{-1} is a parametrization of U. The pair (U, φ) is a chart. An atlas on M is a collection of charts $\{(U_a, \varphi_a)\}$ such that the U_a cover M.

The set M is a k-dimensional topological manifold if and only if it admits an atlas for which each $\varphi_a(U_a)$ is open in $\Re^k$. It is a differentiable manifold if and only if the transition maps $\varphi_b \varphi_a^{-1}$ are diffeomorphisms. A subset $S \subset M$ is a d-dimensional embedded submanifold if and only if, for every $p \in S$, there is a chart (U, φ) such that $p \in U$ and

$$\varphi(U \cap S) = \varphi(U) \cap \left(\Re^d \times \{\vec{0} \in \Re^{k-d}\} \right) = \{y \in \varphi(U) : y_{d+1} = \cdots = y_k = 0\}.$$

Our explication of statistical manifolds follows Murray and Rice (1993), from whom much of our notation is borrowed. Let $(\Omega, \mathcal{B}, \mu)$ denote a measure space. Let $\mathcal{M}$ denote the nonnegative measures on $(\Omega, \mathcal{B})$ that are absolutely continuous with respect to μ. We write an element of $\mathcal{M}$ as $p\,d\mu$, where p is a density function with respect to μ. We write $p\,d\mu \sim q\,d\mu$ and say that $p\,d\mu$ and $q\,d\mu$ are equivalent up to scale if and only if

$$\frac{\int_B p(x)\,d\mu(x)}{\int_\Omega p(x)\,d\mu(x)} = \frac{\int_B q(x)\,d\mu(x)}{\int_\Omega q(x)\,d\mu(x)}$$

for every $B \in \mathcal{B}$. Murray and Rice (1993) regard a probability measure as an equivalence class of finite measures. Let $\mathcal{P}$ denote the space of probability measures in $\mathcal{M}$, that is, the set of finite measures up to scale.

Let $\Re_\Omega$ denote the vector space of measurable real-valued functions on Ω and define the log-likelihood map $\ell : \mathcal{M} \to \Re_\Omega$ by $\ell(p\,d\mu) = \log(p)$. We say that the log-likelihood map is smooth if and only if, for each $x \in \Omega$, the corresponding real-valued component map defined by $p\,d\mu \mapsto [\log(p)](x)$ is sufficiently differentiable.

Definition 1 *Let $P = \left\{ p(\cdot, \theta)\,d\mu \;:\; \theta \in \Theta \subseteq \Re^k \right\}$ denote a parametric family of probability distributions in $\mathcal{P}$. We say that P is a statistical manifold if and only if P is a differentiable manifold, the log-likelihood map is smooth, and, for any $p\,d\mu \in P$, the random variables*

$$\frac{\partial \ell}{\partial \theta^1}(p\,d\mu), \ldots, \frac{\partial \ell}{\partial \theta^k}(p\,d\mu)$$

are linearly independent.

We might dispense with the parametric structure of P, but many of the familiar concepts and results of classical statistics are stated in terms of index sets rather than families of distributions. For example, fix $p\,d\mu \in P$. Then the random vector

$$d_p \ell = \left(\frac{\partial \ell}{\partial \theta^1}(p\,d\mu), \ldots, \frac{\partial \ell}{\partial \theta^r}(p\,d\mu) \right)$$

is the score vector at $p\,d\mu$, and the set of vectors obtained by observing the score vector at each $x \in \Omega$ is the tangent space of P at $p\,d\mu$, denoted T_pP. Our exposition will emphasize the manifold structure of P itself, but one can just as easily regard P as indexed by a k-dimensional manifold Θ—and it is often convenient to do so.

2.2 Riemannian Geometry and Fisher Information

A metric tensor on the statistical manifold P is a collection of inner products on the tangent spaces of P. If P admits a metric tensor, then P is a Riemannian manifold. See Milnor (1963, Part II) and Hicks (1971) for concise introductions to Riemannian geometry. Note that many authors refer to the metric tensor as a Riemannian metric. In neither case is the word "metric" used in the sense of a distance function.

Let E_p denote expectation with respect to $p\,d\mu$, that is, $E_pf = \int_\Omega f(x)p(x)\,d\mu(x)$. Define an inner product on the space of square-integrable f by $\langle f, g\rangle_p = E_pfg$. If the log-likelihood map is smooth, then the Fisher information matrix $I(p) = [g_{ij}(p)]$ has entries

$$g_{ij}(p) = E_p \frac{\partial \ell}{\partial \theta^i}\frac{\partial \ell}{\partial \theta^j} = \left\langle \frac{\partial \ell}{\partial \theta^i}, \frac{\partial \ell}{\partial \theta^j} \right\rangle_p .$$

Because the scores are linearly independent, $I(p)$ is the matrix of the inner product $\langle\cdot,\cdot\rangle_p$ with respect to the basis defined by the scores.

Rao (1945) observed that Fisher information induces a natural metric tensor on P. To obtain a coordinate-free representation of this tensor, that is, a representation that does not involve Fisher information matrices, suppose that $v \in T_pP$ and let $\gamma : (-\epsilon, \epsilon) \to P$ be any variation with tangent vector v at $p = \gamma(0)$. The differential of the log-likelihood map at p is the function $d_p\ell : T_pP \to \Re_\Omega$ defined by

$$d_p\ell(v) = \frac{d}{dt}\ell(\gamma(t))|_{t=0} = \lim_{t\to 0}\frac{\ell(\gamma(t)) - \ell(\gamma(0))}{t}$$

and the Fisher information tensor is the collection of inner products

$$g_p(v, w) = E_p d_p\ell(v) d_p\ell(w).$$

Henceforth we regard P as a Riemannian manifold and assume that P is connected. Given $p\,d\mu, q\,d\mu \in P$, let $\gamma : [0, 1] \to P$ be a smooth variation such that $\gamma(0) = p$ and $\gamma(1) = q$. Let $\dot\gamma(t) \in T_{\gamma(t)}P$ denote the tangent vector to γ at $\gamma(t)$. The distance traversed by γ is

$$\text{length}(\gamma) = \int_0^1 \left[g_{\gamma(t)}\left(\dot\gamma(t), \dot\gamma(t)\right)\right]^{1/2} dt = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, dt$$

and the infimum of these lengths over all such variations defines $i(p, q)$, the *information distance* between $p\,d\mu$ and $q\,d\mu$ in P.

2.3 Hellinger Distance

Murray and Rice (1993, Section 6.8) remarked that the fact that the inner products g_p vary with p makes it difficult to discern the global structure of the statistical manifold P directly

from Fisher information. To remedy this difficulty they defined the square root likelihood, here denoted $s : P \to \Re_\Omega$, by $s(p) = s(p\,d\mu) = 2\sqrt{p}$. Defining the inner product

$$\langle f, g \rangle_\mu = \int_\Omega f(x) g(x)\, d\mu(x)$$

and noting that $2 d_p s = s d_p \ell$, we discover that

$$\begin{aligned} g_p(v,w) &= E_p d_p\ell(v) d_p\ell(w) \\ &= \int_\Omega \left[d_p\ell(v)\right](x) \left[d_p\ell(w)\right](x) p(x)\, d\mu(x) \\ &= \int_\Omega \left[\frac{s(p)}{2} d_p\ell(v)\right](x) \left[\frac{s(p)}{2} d_p\ell(w)\right](x)\, d\mu(x) \\ &= \int_\Omega \left[d_p s(v)\right](x) \left[d_p s(w)\right](x)\, d\mu(x) \\ &= \langle d_p s(v), d_p s(w) \rangle_\mu . \end{aligned}$$

Hence, if γ is a variation in P and $\sigma = s(\gamma)$ is the corresponding variation in $s(P)$, then

$$\text{length}(\gamma) = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, dt = \int_0^1 \|\dot\sigma(t)\|_\mu\, dt = \text{length}(\sigma).$$

The quantity

$$h(p,q) = \|s(p) - s(q)\|_\mu = \|2\sqrt{p} - 2\sqrt{q}\|_\mu \tag{1}$$

is the Hellinger distance between the densities p and q. Thus, information distances can be computed by working with Hellinger distance rather than Fisher information, and information distance on P behaves locally like Hellinger distance. Proof of the following approximation is relegated to Appendix A.

Theorem 1 *Under standard regularity conditions,*

$$h^2\left(p\left(\cdot,\theta\right), p\left(\cdot,\theta_0\right)\right) = \left(\theta - \theta_0\right)^\top I\left(\theta_0\right)\left(\theta - \theta_0\right) + o\left(\left\|\theta - \theta_0\right\|^2\right).$$

2.4 Minimum Hellinger Distance Estimation

Following Basu et al. (2011), with minor changes in notation, suppose that $x_1, \ldots, x_n \sim p\,d\mu = p(\cdot,\theta)\,d\mu$ and let $\bar\theta$ denote the true value of θ. Let $u(x_i,\theta) = \nabla_\theta \log p(x_i,\theta)$ denote the score function for P and let

$$Z_n(\theta) = \sqrt{n}\frac{1}{n}\sum_{i=1}^n u\left(x_i,\theta\right).$$

Under standard regularity conditions, the maximum likelihood estimator $\tilde\theta_n$ of θ is first-order efficient; in particular,

$$\sqrt{n}\left(\tilde\theta_n - \bar\theta\right) = I^{-1}\left(\bar\theta\right) Z_n\left(\bar\theta\right) + o_p(1). \tag{2}$$

Let $\hat{p}_n$ denote a nonparametric density estimate of p and define the minimum Hellinger distance estimate (MHDE) of θ by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} h\left(p(\cdot,\theta),\hat{p}_n\right) = \arg\min_{\theta\in\Theta}\int_\Omega \left[\sqrt{p(x,\theta)} - \sqrt{\hat{p}_n(x)}\right]^2 d\mu(x).$$

Under suitable regularity conditions (see Beran, 1977; Basu et al., 2011, Section 3.2.2),

$$\sqrt{n}\left(\hat{\theta}_n - \bar{\theta}\right) = I^{-1}\left(\bar{\theta}\right) Z_n\left(\bar{\theta}\right) + o_p(1). \tag{3}$$

Thus, both $\tilde{\theta}_n$ and $\hat{\theta}_n$ are first-order efficient estimators. Typically, $\tilde{\theta}_n$ is more readily computed and $\hat{\theta}_n$ has better robustness properties.

3. Information Tests

Suppose that $x_1,\ldots,x_n \sim p\,d\mu$, where $p\,d\mu$ lies in the connected k-dimensional statistical manifold P. We write $p = p(\cdot,\theta)$, $\bar{p} = p(\cdot,\bar{\theta})$ and test the simple null hypothesis $H_0 : p = \bar{p}$ against the composite alternative hypothesis $H_1 : p \neq \bar{p}$. Equivalently, we test $H_0 : \theta = \bar{\theta}$ against $H_1 : \theta \neq \bar{\theta}$.

Let $\hat{\theta}_n$ denote the MHDE of θ and consider the test statistic

$$\mathrm{ID}_n = i^2\left(p\left(\cdot,\hat{\theta}_n\right), p\left(\cdot,\bar{\theta}\right)\right) = i^2\left(p\left(\cdot,\hat{\theta}_n\right),\bar{p}\right),$$

the squared information distance between $p(\cdot,\hat{\theta}_n)$ and $\bar{p}$ on the statistical manifold P. Because information distance on P behaves locally like Hellinger distance (see Section 2.3), we begin by studying the local behavior of the related test statistic

$$\mathrm{HD}_n = h^2\left(p\left(\cdot,\hat{\theta}_n\right),\bar{p}\right).$$

Notice that $n\mathrm{HD}_n$ differs from the standard Hellinger disparity difference statistic described in Basu et al. (2011, Section 5.1), although it turns out that they are locally asymptotically equivalent. More precisely, the relation of tests based on HD to Wald tests is analogous to the relation of disparity difference tests to likelihood ratio tests. Proof of the following approximation is relegated to Appendix A.

Theorem 2 *Let*

$$W_n = n\left(\tilde{\theta}_n - \bar{\theta}\right)^\top I\left(\bar{\theta}\right)\left(\tilde{\theta}_n - \bar{\theta}\right)$$

denote the Wald statistic for testing $H_0 : \theta = \bar{\theta}$ versus $H_1 : \theta \neq \bar{\theta}$. If (2) and (3) hold, then

$$nHD_n - W_n = o_p(1).$$

Our Theorem 2 is analogous to Theorem 1 in Simpson (1989), which relates a Hellinger deviance test statistic to the likelihood ratio test statistic

$$G_n^2 = 2\sum_{i=1}^n \log p\left(x_i,\tilde{\theta}_n\right)/p\left(x_i,\bar{\theta}\right).$$

The asymptotic null distribution of G_n^2 and W_n is $\chi^2(k)$; it follows that the asymptotic null distribution of Simpson's test statistic and our $n\mathrm{HD}_n$ is also $\chi^2(k)$. Furthermore, a contiguity argument (for details see Simpson, 1989) establishes that these tests have the same asymptotic power at local alternatives of the form $\bar{\theta} + \eta/\sqrt{n}$. (Local alternatives are defined by fixing η and assuming that $\bar{\theta} + \eta/\sqrt{n} \in \Theta$ for n sufficiently large.) In this sense, our HD test, the Wald test, the likelihood ratio test, and Simpson's Hellinger deviance test are all locally equivalent.

To extend the equivalence to our ID test, we demonstrate that $i^2(p_t, p_0)$ behaves locally like (6). Recall that a geodesic arc is a variation with zero curvature, hence with constant velocity. Given $p_0 \in P$, use Lemma 10.3 in Milnor (1963) to choose a neighborhood W of p_0 and $\bar{\epsilon} > 0$ such that $q \in W$ implies the existence of a unique geodesic variation γ connecting p_0 and q with $\epsilon = \mathrm{length}(\gamma) < \bar{\epsilon}$. It then follows from Theorem 10.4 in (Milnor, 1963) that $i(q, p_0) = \epsilon$, that is, that γ is the unique path of shortest distance from p_0 to q. Parametrizing γ by arc length and letting $q = p_\epsilon$, we obtain

$$\epsilon = i\left(p_\epsilon, p_0\right) = \mathrm{length}(\gamma) = \int_0^\epsilon \left\|\dot{\gamma}(t)\right\|_{\gamma(t)} \, dt$$

with constant unit velocity

$$1 = \left\|\dot{\gamma}(t)\right\|_{\gamma(t)} = I_\gamma\left(p_t\right).$$

It follows from (6) that

$$h^2\left(p_\epsilon, p_0\right) = I_\gamma\left(p_0\right)\epsilon^2 + o\left(\epsilon^2\right) = \epsilon^2 + o\left(\epsilon^2\right) = i^2\left(p_\epsilon, p_0\right) + o\left(\epsilon^2\right). \tag{4}$$

Set $\theta_0 = \bar{\theta}$. By arguments analogous to those used to establish Theorem 2, we then obtain the following relation.

Theorem 3 *If (3) holds, then* $nHD_n - nID_n = o_p(1)$.

Thus, ID_n and HD_n are locally asymptotically equivalent for testing $H_0 : \theta = \bar{\theta}$ versus $H_1 : \theta \neq \bar{\theta}$.

Although the information distance, Hellinger distance, Wald, likelihood ratio, and Hellinger disparity distance tests are all locally asymptotically equivalent, only the information distance test attempts to exploit the Riemannian geometry of P when testing nonlocal alternatives.

4. Restricted Information Tests

Let $Q = \{p(\cdot, \theta)\, d\mu \ : \ \theta \in \Psi \subset \Theta\}$ denote a parametric subfamily of probability distributions in P. Suppose that Q is a d-dimensional embedded submanifold of P; equivalently, suppose that Ψ is a d-dimensional embedded submanifold of Θ. Suppose that $\bar{\theta} \in \Psi$ and that we want to test $H_0 : \theta = \bar{\theta}$, restricting attention to alternatives that lie in Ψ. We emphasize that we are restricting inference to the submanifold, *not* testing the null hypothesis that θ lies in the submanifold. Two information tests are then available: the unrestricted information test computes information distance on the statistical manifold P, whereas the restricted information test computes information distance on the statistical submanifold Q. It is

tempting to speculate that restricted information tests are more powerful than unrestricted information tests.

An analogous investigation of restricted likelihood ratio tests was undertaken by Trosset et al. (2016), who indeed established that, if $d = \dim(\Psi) < \dim(\Theta) = k$, then the restricted likelihood ratio test is asymptotically more powerful than the unrestricted likelihood ratio test at local alternatives. As information tests are locally asymptotically equivalent to likelihood ratio tests, they must enjoy the same property. However, Trosset et al. (2016) also constructed examples in which the restricted likelihood ratio test is less powerful than the unrestricted likelihood ratio test for certain nonlocal alternatives. Unlike restricted likelihood ratio tests, restricted information tests potentially exploit the *global* structure of the statistical submanifold. This observation motivates investigating the behavior of information tests at nonlocal alternatives.

In what follows we specialize to the case of multinomial distributions, which are widely used (as in Kass, 1989) to illustrate the ideas of information geometry. Accordingly, consider an experiment with $k+1$ possible outcomes. The probability model $P = \text{Multinomial}(\theta)$ specifies that the outcomes occur with probabilities $\theta = (\theta_1, \ldots, \theta_{k+1})$. It is parametrized by the k-dimensional unit simplex in $\Re^{k+1}$,

$$\Theta = \Delta^k = \{\theta \in [0,1]^{k+1} : \theta_1 + \cdots + \theta_{k+1} = 1\},$$

or (upon setting $\sigma = \sqrt{\theta}$, defined by setting each $\sigma_i = \sqrt{\theta_i}$) by that portion of the k-dimensional unit sphere that lies in the nonnegative orthant of $\Re^{k+1}$,

$$\Sigma = \{\sigma \in [0,1]^{k+1} : \sigma_1^2 + \cdots + \sigma_{k+1}^2 = 1\}.$$

One advantage of studying multinomial distributions is the availability of explicit formulas. If $p = p(\cdot, \theta = \sigma^2)$ and $q = p(\cdot, \pi = \rho^2)$, then

$$h^2(p,q) = \sum_{i=1}^{k+1} \left(2\sqrt{\theta_i} - 2\sqrt{\pi_i}\right)^2 = 4\sum_{i=1}^{k+1} (\sigma_i - \rho_i)^2 = 4\,\|\sigma - \rho\|^2$$

and we see that Hellinger distance between multinomial distributions corresponds to chordal (Euclidean) distance on Σ. Hence, by the law of cosines,

$$h^2(p,q) = 4(2 - 2\cos\delta) = 8 - 8\langle\sigma, \rho\rangle,$$

where δ is the angle between σ and ρ. But δ is also the great circle (geodesic) distance between σ and ρ; hence,

$$i(p,q) = 2\delta = 2\arccos\langle\sigma, \tau\rangle,$$

where the factor of 2 accrues from (1). It follows that

$$h^2(p,q) = 8 - 8\cos\left(i(p,q)/2\right),$$

establishing that the information and Hellinger distances between multinomial distributions are monotonically related.

A second advantage of studying multinomial distributions is that empirical distributions from multinomial experiments are themselves multinomial distributions. Suppose

that one draws n independent and identically distributed observations from Multinomial(θ) and counts $\vec{x} = (x_1, \ldots, x_{k+1})$, where x_i records the number of occurrences of outcome i. The empirical distribution of $\vec{x}$ is $\hat{p}_n(\vec{x}) = \vec{x}/n$ and furthermore, because $\vec{x}/n \in \Theta$, the unrestricted MHDE of $\theta \in \Theta$ is $\hat{\theta}_n(\vec{x}) = \vec{x}/n$. The restricted MHDE of $\theta \in \Psi$ is

$$\check{\theta}_n(\vec{x}) = \arg\min_{\theta \in \Psi} h^2(\theta, \vec{x}/n) = \check{\sigma}_n^2(\vec{x}),$$

where

$$\check{\sigma}_n(\vec{x}) = \arg\max_{\sigma^2 \in \Psi} \left\langle \sigma, \sqrt{\vec{x}/n} \right\rangle.$$

Depending on the submanifold Ψ, the calculation of $\check{\theta}_n(\vec{x})$ may require numerical optimization.

Let $i(\cdot, \cdot; \Theta)$ denote information distance on the unrestricted model and let $i(\cdot, \cdot; \Psi)$ denote information distance on the restricted model. The nonrandomized unrestricted information test with critical value c_2 rejects $H_0 : \theta = \bar{\theta}$ if and only if

$$i_n(\vec{x}; \Theta) = i\left(p\left(\cdot, \hat{\theta}_n\right), p\left(\cdot, \bar{\theta}\right); \Theta\right) = 2\arccos\left\langle \sqrt{\vec{x}/n}, \sqrt{\bar{\theta}} \right\rangle > c_2.$$

The nonrandomized restricted information test with critical value c_1 rejects $H_0 : \theta = \bar{\theta}$ if and only if

$$i_n(\vec{x}; \Psi) = i\left(p\left(\cdot, \check{\theta}_n\right), p\left(\cdot, \bar{\theta}\right); \Psi\right) > c_1.$$

Because $\vec{x}$ is discrete, randomization may be needed to attain a specified size. For n sufficiently large, we can use the $1-\alpha$ quantiles $q_{1-\alpha}(k)$ and $q_{1-\alpha}(d)$ of chi-squared distributions with k and d degrees of freedom to select the critical values:

$$c_2 = (q_{1-\alpha}(k)/n)^{1/2} \quad \text{and} \quad c_1 = (q_{1-\alpha}(d)/n)^{1/2}$$

The power functions of the above tests are

$$\beta_2(\theta) = P_{\theta \in \Psi}(i_n(\vec{x}; \Theta) > c_2)$$

for the unrestricted information test and

$$\beta_1(\theta) = P_{\theta \in \Psi}(i_n(\vec{x}; \Psi) > c_1)$$

for the restricted information test.

5. Two Trinomial Examples

The probability model Trinomial(θ) specifies that $k + 1 = 3$ outcomes occur with probabilities $\theta = (\theta_1, \theta_2, \theta_3)$. Define $\psi : [0, 1] \to \Theta = \Delta^2$ by $\psi(\tau) = \left(\tau^2, 2\tau(1-\tau), (1-\tau)^2\right)$. The Hardy-Weinberg subfamily of trinomial distributions is parametrized by the embedded submanifold $\Psi = \{\psi(\tau) : \tau \in [0, 1]\}$. Notice that $\dim \Psi = 1 < 2 = \dim \Theta$. We write $\mathrm{HW}(\tau) = \mathrm{Trinomial}(\psi(\tau))$.

Fix $\bar{\tau} \in (0, 1)$ and set $\bar{\theta} = \psi(\bar{\tau})$. We test the simple null hypothesis $H_0 : \theta = \bar{\theta}$ against alternatives of the form $\theta = \psi(\tau)$. The unrestricted information test statistic is

$$i_n(\vec{x}; \Theta) = 2\arccos\left(\bar{\tau}(x_1/n)^{1/2} + [2\bar{\tau}(1-\bar{\tau})x_2/n]^{1/2} + (1-\bar{\tau})(x_3/n)^{1/2}\right).$$

x_1, x_2, x_3	$p(\vec{x}, \psi(0.3))$	$i_3(\vec{x}; \Theta)$	$\check{\tau}(\vec{x})$	$i_3(\vec{x}; \Psi)$
3, 0, 0	0.000729	2.532207	1	2.803414
2, 1, 0	0.010206	1.806363	0.8535517	1.692687
2, 0, 1	0.011907	1.728807	1	2.803414
1, 2, 0	0.047628	1.584191	0.7236016	1.237653
1, 1, 1	0.111132	0.625338	0.5	0.581973
1, 0, 2	0.064827	1.461264	0	1.639469
0, 3, 0	0.074088	1.731487	0.5	0.581973
0, 2, 1	0.259308	0.734627	0.2763984	0.073708
0, 1, 2	0.302526	0.662028	0.1464483	0.528741
0, 0, 3	0.117649	1.590798	0	1.639469

Table 1: Unrestricted (Trinomial) and restricted (Hardy-Weinberg) information tests of $H_0 : \theta = \psi(0.3)$ with $n = 3$ observations. Columns 1–2 list the possible outcomes and their exact probabilities under H_0; Column 3 lists the unrestricted information distance of the empirical distributions from the null distribution; Columns 4–5 list the minimum Hellinger distance estimates of the Hardy-Weinberg parameter, τ, and the restricted information distance of the corresponding distributions from the null distribution.

The restricted MHDE of $\theta \in \Psi$ is

$$\check{\theta}_n(\vec{x}) = \psi\left(\check{\tau}(\vec{x})\right),$$

where

$$\check{\tau}(\vec{x}) = \arg\max_{\tau \in [0.1]} \left(\tau \left(x_1/n\right)^{1/2} + \left[2\tau(1-\tau)x_2/n\right]^{1/2} + (1-\tau)\left(x_3/n\right)^{1/2} \right).$$

Letting $\sigma(\tau) = 2\psi(\tau)^{1/2}$, the restricted information test statistic, $i_n(\vec{x}; \Psi)$, is computed by integrating

$$\|\dot{\sigma}(\tau)\| = 2\left[1^2 + \frac{(1-2\tau)^2}{2\tau(1-\tau)} + 1^2\right]^{1/2}$$

as τ varies between $\bar{\tau}$ and $\check{\tau}(\vec{x})$.

Example 1 The trinomial experiment with $n = 3$ has 10 possible outcomes, enumerated in the first column of Table 1. Consider the unrestricted and restricted information tests of $H_0 : \theta = \psi(0.3)$ with size $\alpha = 0.1$. The exact unrestricted test rejects H_0 with certainty if

$$C_{2a} = \{(3,0,0), (2,1,0), (0,3,0), (2,0,1)\}$$

is observed, and with probability $(0.1 - 0.09693)/0.117649 \doteq 0.02609457$ if $C_{2b} = (0,0,3)$ is observed. The exact restricted test rejects H_0 with certainty if

$$C_{1a} = \{(3,0,0), (2,0,1), (2,1,0)\}$$

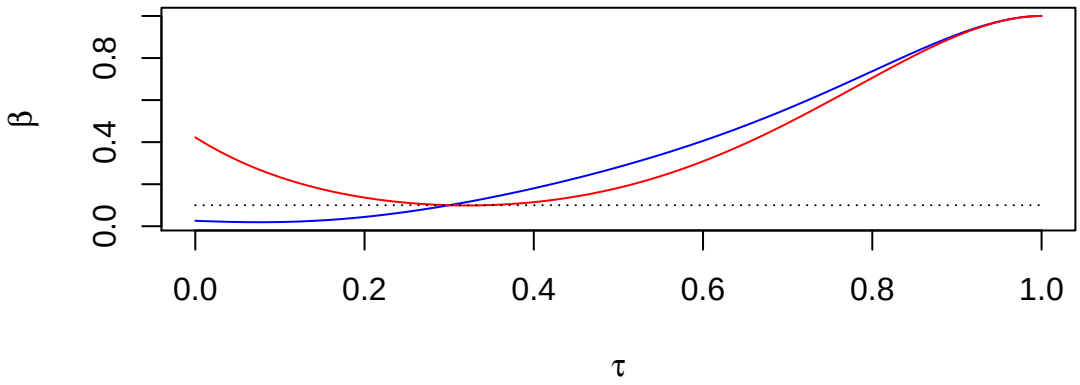

Figure 2: Power of the exact unrestricted (β_2, plotted in blue) and restricted (β_1, plotted in red) information tests for testing $H_0 : \theta = \psi(0.3)$ with $\alpha = 0.1$ (dotted line) and $n = 3$. The alternatives $\{\theta = \psi(\tau) : \tau \in [0, 1]\}$ are displayed on the horizontal axis. The restricted test is greatly superior for $\tau < 0.3$, slightly inferior for $\tau > 0.3$.

is observed, and with probability $(0.1 - 0.022842)/0.182476 \doteq 0.4228392$ if

$$C_{1b} = \{(1, 0, 2), (0, 0, 3)\}$$

is observed. The respective power functions are plotted in Figure 2. The restricted test is dramatically more powerful for $\tau < 0.3$, slightly less powerful for $\tau > 0.3$. ■

The small sample size in Example 1 allows us to illustrate the construction of the unrestricted and restricted information tests, but understates the superiority of the restricted test. It is curious that the restricted test is *less* powerful than the unrestricted test for alternatives $\tau > 0.3$, but Trosset et al. (2016) demonstrated the same anomaly for likelihood ratio tests. For larger sample sizes, the superiority of the restricted test is unambiguous.

Example 2 The trinomial experiment with $n = 20$ has 231 possible outcomes. Consider the unrestricted and restricted information tests of $H_0 : \theta = \psi(0.3)$ with size $\alpha = 0.05$. The exact unrestricted test has a critical region of 169 possible outcomes, with a boundary of one outcome that requires randomization. The exact restricted test has a critical region of 152 possible outcomes, with a boundary of one outcome that requires randomization. The difference in power functions, $\beta_1(\psi(\tau)) - \beta_2(\psi(\tau))$, is plotted in Figure 3. The restricted test is clearly superior, although careful examination reveals that it is slightly inferior for alternatives slightly greater than 0.3. For example,

$$\beta_1(\psi(0.305)) - \beta_2(\psi(0.305)) \doteq -0.0002842388.$$

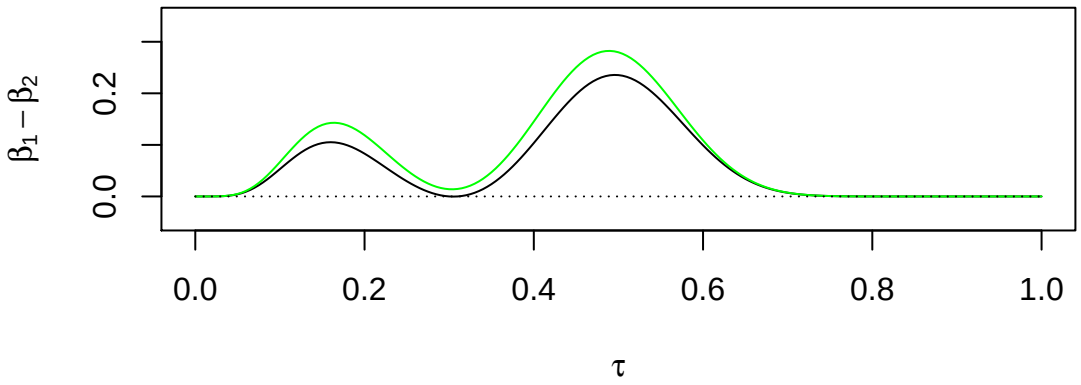

Figure 3: Powers of two restricted information tests (β_1) minus power of the exact unrestricted information test (β_2) for testing $H_0 : \theta = \psi(0.3)$ with $\alpha = 0.05$ and $n = 20$. The black curve corresponds to the exact restricted information test, which has size α. The green curve corresponds to the restricted information test with critical value determined by $\chi^2(1)$, which has size 0.064. The alternatives $\{\theta = \psi(\tau) : \tau \in [0, 1]\}$ are displayed on the horizontal axis.

For comparison, a $\chi^2(1)$ approximation yields a critical value of $c_1 = 0.4382613$. The corresponding critical region is slightly larger than the exact critical region, containing an additional 5 outcomes. Using a larger critical region increases the probability of rejection, in particular to a size of 0.06402558. This power function, minus $\beta_2(\psi(\tau))$, is also plotted in Figure 3. ■

6. Approximate Information Tests

So far, our exposition has glossed the computational challenges posed by information tests. For multinomial manifolds, the empirical distributions lie on the manifold and information distance can be computed by a simple formula. For the 1-dimensional Hardy-Weinberg submanifold, minimum Hellinger distance estimates require numerical optimization, geodesic variations are apparent by inspection, and computing an information distance requires numerical integration. In general, however, the information tests described in Sections 3 and 4 necessitate overcoming the following challenges:

1. Numerical optimization on the submanifold to determine the minimum Hellinger distance estimate, $\breve{\theta}_n$.

2. Determining the geodesic variation between $\check{\theta}_n$ and the hypothesized $\bar{\theta}$. If the submanifold is 1-dimensional, then this is easily accomplished by inspection; if $d > 1$, then the geodesic variation must be determined by solving a potentially intractable problem in the calculus of variations.

3. Numerical integration along the geodesic variation to determine the information distance between $\check{\theta}_n$ and $\bar{\theta}$.

We now propose procedures that circumvent these challenges. The key idea that underlies these procedures is that information distance is locally approximated by Hellinger distance.

In what follows, we assume that the problems described above are difficult or intractable, but that we can identify a finite set of distributions in the submanifold $Q = \{p(\cdot, \theta)\, d\mu \ : \ \theta \in \Psi \subset \Theta\}$. For example, in the case of the Hardy-Weinberg submanifold, we might identify m trinomial distributions by drawing $\tau_1, \ldots, \tau_m \sim \text{Uniform}(0, 1)$. Combined with the hypothesized distribution, we thus have $m + 1$ distributions in Q from which we hope to learn enough about the Riemannian structure of Q to approximate the methods of Section 4.

Elaborating on Figure 1, we propose the following procedure for testing $H_0 : \theta = \bar{\theta}$.

1. Identify $\theta_1, \ldots, \theta_m \in \Psi \subset \Re^k$ and compute the $(m+1)m/2$ pairwise Hellinger distances h_{ij} between the $\bar{p}, p_1, \ldots, p_m$ that correspond to $\bar{\theta}, \theta_1, \ldots, \theta_m$.

   Remark. We include this step for clarity of exposition; however, the localization graph constructed in the next step may not require pre-computing all $(m + 1)m/2$ pairwise distances.

2. Use the pairwise Hellinger distances to form $\mathcal{G}$, a graph with $m + 1$ vertices corresponding to the $m + 1$ distributions. Connect vertices i and j when h_{ij} is sufficiently small, so that $\mathcal{G}$ localizes the structure of the submanifold Q. Weight edge $i \leftrightarrow j$ by h_{ij}.

   This is a standard construction in manifold learning, for example, Tenenbaum et al. (2000); Roweis and Saul (2000), although our application of manifold learning techniques to statistical rather than data manifolds appears to be novel. The most popular constructions are either (a) connect i and j if and only if $h_{ij} \leq \epsilon$, or (b) connect i and j if and only if i is a K-nearest neighbor (KNN) of j or j is a KNN of i. The choice of the localization parameter (ϵ or K) is a model selection problem. It is imperative that the localization parameter be chosen so that $\mathcal{G}$ is connected.

   The computational complexity of the first two steps depends on the type of localization graph. Traditional algorithms for computing exact KNN graphs require $O(km)$ time. However, there exist faster algorithms that compute approximate KNN graphs. See Giles et al. (2008, Section 3.3) and the references therein.

3. Compute $\Delta = [\delta_{ij}]$, the $(m + 1) \times (m + 1)$ dissimilarity matrix of pairwise shortest path distances in $\mathcal{G}$.

   Here we appropriate the key idea of the popular manifold learning procedure isomap (Tenenbaum et al., 2000). A path in $\mathcal{G}$ is a discrete approximation of a variation in Q. The length of a path is the sum of its Hellinger distance edge weights, hence a discrete

approximation of the integral that defines the length of the approximated variation. The shortest path between vertices i and j approximates the geodesic variation between distributions i and j, hence the shortest path distance δ_{ij} approximates the information distance between distributions i and j.

A number of algorithms are available for computing the entire set of pairwise shortest path distances on a graph. The famous Floyd-Warshall algorithm has computational complexity $O(m^3)$.

4. For a suitable choice of r, embed Δ in $\Re^r$ by minimizing a suitably weighted raw stress criterion,

$$\sigma(Z) = \sum_{i<j} w_{ij} \left[\|z_i - z_j\| - \delta_{ij} \right]^2 ,$$

where the coordinates of $z_i \in \Re^r$ appear in row i of the $(m+1) \times r$ configuration matrix Z.

Isomap (Tenenbaum et al., 2000) embeds shortest path distances by classical multidimensional scaling (Torgerson, 1952; Gower, 1966), which minimizes a squared error criterion for pairwise inner products. The widely used raw stress criterion is more directly related to our objective of modeling shortest path distance with Euclidean distance; it also provides greater flexibility through its ability to accommodate different weighting schemes. The raw stress criterion can be numerically optimized by majorization (de Leeuw, 1988), several iterations of which usually provides a useful embedding, or by Newton's method (Kearsley et al., 1998), which has better local convergence properties.

An initial embedding can be constructed in $O(rm)$ time, although more expensive efforts may be less prone to finding nonglobal minimizers and may result in better overall performance. For general weights, the computational complexity of the Guttman majorization algorithm is dominated by an initial Cholesky factorization that requires $O(m^3)$ time. If $w_{ij} = 1$, then the initial Cholesky factorization is not needed and each application of the Guttman transform requires $O(rm)$ time.

Remark. The choice of r is a model selection problem. While $r = d$ is nearly universal in conventional manifold learning, $r > d$ may provide a more faithful Euclidean representation of the geodesic structure of Q.

5. From $x_1, \ldots, x_n \sim p$, construct a nonparametric density estimate $\hat{p}_n$. Compute the Hellinger distances of $\hat{p}_n$ from $p_1, \ldots, p_m$ and let $j_1, \ldots, j_\ell$ index the nearest $\ell \geq r$ distributions. Embed $\hat{p}_n$ in the previously constructed representation by a suitable out-of-sample embedding technique. Let $y(\vec{x}) \in \Re^r$ denote the resulting representation of $\hat{p}_n$. The proposed approximate information test rejects $H_0 : \theta = \bar{\theta}$ if and only if the test statistic

$$\hat{i}_n(\vec{x}; \Psi) = \|y(\vec{x}) - \bar{z}\| ,$$

where $\bar{z}$ corresponds to $\bar{\theta}$, is sufficiently large.

A comprehensive discussion of how to embed $\hat{p}_n$ using only its ℓ nearest neighbors is beyond the scope of this manuscript. For $r = 1$ and $\ell = 2$, one can use the law of

cosines to project $\hat{p}_n$ into the line that contains z_{j_1} and z_{j_2}. This construction is a special case of out-of-sample embedding into a principal components representation. See Gower (1968) for a general formula that uses pairwise squared distances; see Williams and Seeger (2001) for a general formula that uses pairwise inner products. For Example 4 and Section 7, we simply set $y(\vec{x})$ equal to the centroid of $z_{j_1}, z_{j_2}, z_{j_3}$.

The computational complexity of this step depends on the out-of-sample embedding technique. The technique used in Example 4 and Section 7 is $O(m)$.

6. Estimate a significance probability by generating simulated random samples $\vec{x}_i$ of size n from the hypothesized distribution $\bar{p}$. Perform the previous step for each $\vec{x}_i$ and compute the fraction of $\vec{x}_i$ for which

$$\hat{i}_n(\vec{x}_i; \Psi) \geq \hat{i}_n(\vec{x}; \Psi).$$

Remark. While the computational complexity of the previous step is just $O(m)$, good estimates of the correct significance probability will usually require generating a large number of simulated samples.

Example 3 As in Section 5, we consider the Hardy-Weinberg submanifold of Trinomial(θ), defined by $\psi(\tau) = \left(\tau^2, 2\tau(1-\tau), (1-\tau)^2\right)$ for $\tau \in [0,1]$. Using $n = 30$ trials, we test $H_0 : \theta = \psi(0.3)$ by two methods:

a The information test on the unrestricted manifold of trinomial distributions, for which information distance can be computed by explicit calculation.

b Ten approximate information tests on estimated 1-dimensional submanifolds, each constructed using $\bar{\tau} = 0.3$ and $\tau_1, \ldots, \tau_9 \sim$ Uniform$[0,1]$. Shortest path distances on 5NN graphs weighted by pairwise Hellinger distances were embedded in $\Re$ using the unweighted raw stress criterion. Empirical distributions were then embedded by applying the law of cosines to the $\ell = 2$ nearest neighbors.

In each case, a randomized test was constructed to have size $\alpha = 0.05$. Note that we use the adjectives *exact* and *approximate* to indicate whether the information distance was computed exactly or approximated by random sampling and manifold learning, not to describe the size of the test.

The power function of the exact unrestricted test was subtracted from the power functions of the ten approximate restricted tests, resulting in the ten difference functions displayed in Figure 4. Except occasionally for values of τ slightly less than 0.3, the approximate restricted tests are consistently more powerful than the exact unrestricted test—often dramatically so. ■

We now return to the Motivating Example in Section 1 and illustrate the proposed methodology.

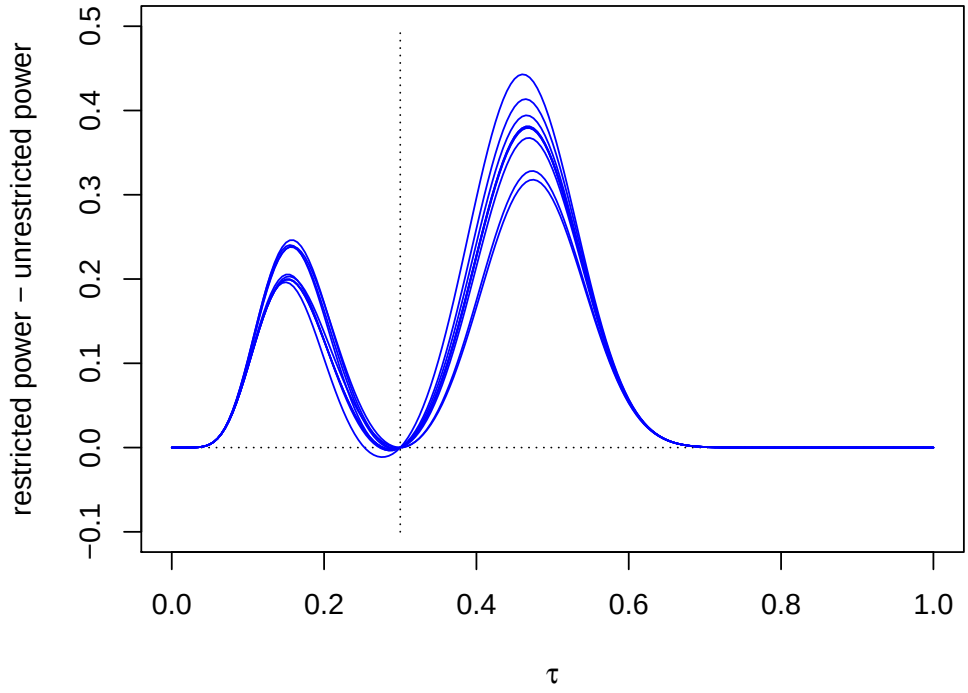

Figure 4: Powers of ten approximate restricted information tests minus power of the exact unrestricted information test for testing $H_0 : \theta = \psi(0.3)$ versus $H_1 : \theta \in \{\psi(\tau) : \tau \in [0, 1]\}$ with $\alpha = 0.05$ and $n = 30$. Each test was randomized to have size α. Each restricted test was constructed using only a random sample of $m = 9$ points from the Hardy-Weinberg submanifold.

Example 4 We parametrize the family of multinomial distributions with 7 possible outcomes by Σ, the portion of the 6-dimensional unit sphere in $\Re^7$ that lies in the nonnegative orthant. The null hypothesis to be tested is

$$H_0 : \sigma = \bar{\sigma} = (0.3, 0.3, 0.3, 0.5, 0.4, 0.4, 0.4).$$

Define $\psi : [0, \pi/2]^2 \to \Sigma$ by

$$\psi(\tau) = (0.3, 0.3, 0.3, 0.5, \rho \cos \tau_1 \sin \tau_2, \rho \sin \tau_1 \sin \tau_2, \rho \cos \tau_2)\,,$$

where $\rho^2 = 0.48$. The 2-dimensional subfamily of multinomial distributions defined by the embedded submanifold $\Psi = \{\psi(\tau) : \tau \in [0, \pi/2]^2\}$ is a spherical subfamily in the sense of Gous (1999). Notice that setting $\tau_1 = \pi/4$ and $\tau_2 = \arctan \sqrt{2}$ results in $\psi(\tau) = \bar{\sigma}$.

We want to test H_0 against alternatives that lie in Ψ. If Ψ was known, then we could perform a restricted likelihood ratio test. The likelihood of $o = (3, 5, 4, 6, 9, 2, 1)$ under $\sigma = \psi(\tau)$ is

$$L_o(\psi(\tau)) = C \cdot 0.09^{3+5+4} \cdot 0.25^6 \cdot (\rho \cos\tau_1 \sin\tau_2)^{2 \cdot 9} \cdot (\rho \sin\tau_1 \sin\tau_2)^{2 \cdot 2} \cdot (\rho \cos\tau_2)^{2 \cdot 1} .$$

To find the restricted maximum likelihood estimate of τ, it suffices to minimize

$$f(\tau) = (-18 \log\cos\tau_1 - 4 \log\sin\tau_1) + (-2 \log\cos\tau_2 - 22 \log\sin\tau_2) = f_1(\tau_1) + f_2(\tau_2)$$

subject to simple bound constraints $\tau \in [0, \pi/2]^2$. The objective function f is separable: it suffices to choose τ_1 to minimize f_1 and τ_2 to minimize f_2. Furthermore, f_1 and f_2 are each strictly convex on $[0, \pi/2]$ (each has a strictly positive second derivative on $(0, \pi/2)$), with unique global minimizers at

$$\breve{\tau}_1 = \arcsin\sqrt{2/11} \doteq 0.4405107 \quad \text{and} \quad \breve{\tau}_2 = \arcsin\sqrt{11/12} \doteq 1.277954.$$

The restricted likelihood ratio test statistic is then

$$\begin{aligned} -2 \log L_o(\bar{\sigma}) / L_o(\psi(\breve{\tau})) &= -2 \log 0.16^{12} / \left(0.36^9 \cdot 0.08^2 \cdot 0.04\right) \\ &= 36 \log 3 - 44 \log 2 \\ &\doteq 9.051566. \end{aligned}$$

The standard asymptotic approximation of the null distribution of the test statistic is a chi-squared distribution with 2 degrees of freedom, resulting in an approximate significance probability of $\mathbf{p} = 0.01082623$. This significance probability is considerably smaller than the significance probabilities that resulted from the unrestricted Pearson and likelihood ratio tests performed in the Motivating Example. Unlike them, it causes rejection of H_0 at significance level $\alpha = 0.05$.

Of course, it is only possible to perform a likelihood ratio test of $H_0 : \sigma = \bar{\sigma}$ versus $H_1 : \sigma \in \Psi$ if Ψ is known. We are concerned with the case that Ψ is unknown, but elements of Ψ can be obtained by sampling. To simulate that scenario, we drew $\tau_1, \ldots, \tau_{100} \sim \text{Uniform}[0, \pi/2]^2$ and computed $\sigma_i = \psi(\tau_i)$. As reported in Section 1, the first two principal components of the corresponding θ_i account for 96% of the variation in the $m = 100$ multinomial parameter values. The vectors $\bar{\sigma}, \sigma_1, \ldots, \sigma_m \in \Re^7$ were then embedded in $\Re^2$ by the manifold learning procedure described above. Shortest path distances on 10NN graphs weighted by pairwise Hellinger distances were embedded in $\Re^2$ using the unweighted raw stress criterion. Empirical distributions were then embedded by averaging the embedded points corresponding to the $\ell = 3$ nearest neighbors.

The resulting representation of the estimated submanifold, $\hat{\Psi}$, is displayed in Figure 5, in which $\sigma_1, \ldots, \sigma_m$ are indicated by $\bullet$, $\bar{\sigma}$ is indicated by $\bullet$, and $y(\vec{x})$ is indicated by $\bullet$. Repeating this procedure on 10000 simulated samples of size $n = 30$ drawn from the null distribution resulted in just 259 larger values of the test statistic, that is, the estimated significance probability is $259/10000 = 0.0259$. The evidence against H_0 produced by the restricted approximate information test is slightly less compelling than the evidence produced by the restricted likelihood ratio test (for which Ψ is known), but is more compelling than the unrestricted Pearson or likelihood ratio tests. ■

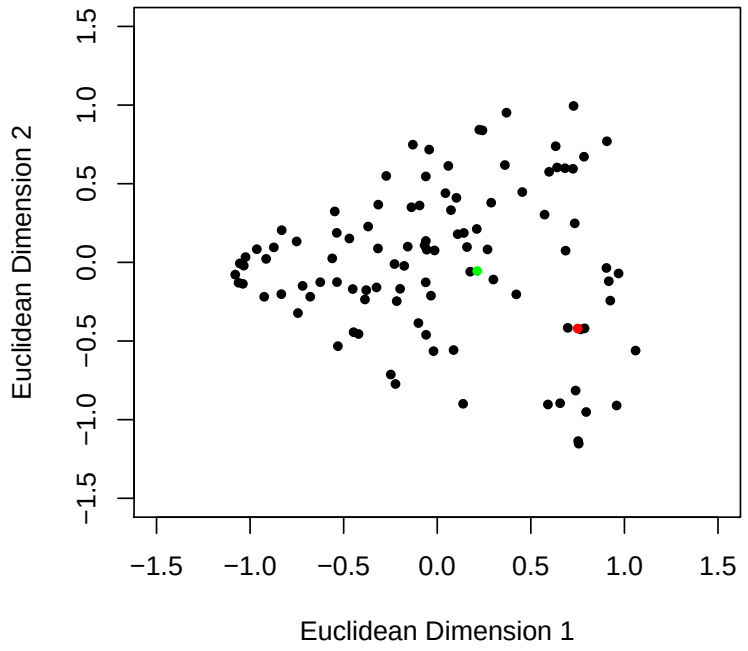

Figure 5: The estimated submanifold in Example 4. The $m = 100$ possible distributions generated by sampling are indicated by •; the null hypothesis is indicated by •; and the minimum distance estimate based on the empirical distribution is indicated by •. The proposed test statistic is $\|\bullet - \bullet\|$, which leads to an estimated significance probability of 0.0259.

7. Effect of Sampling the Submanifold

The proposed approximate information test depends on how densely the unknown statistical submanifold is sampled. We explore this dependence via a small simulation study.

Consider the 2-parameter spherical subfamily Ψ of multinomial distributions defined in Example 4. For $\bar{\sigma} = (\pi/4, \arctan\sqrt{2})$, we test the null hypothesis $H_0 : \psi(\tau) = \psi(\bar{\sigma})$ against alternatives in Ψ at significance level $\alpha = 0.05$. We assume that Ψ is unknown, but that we can generate $\tau_1, \ldots, \tau_m \sim \text{Uniform}[0, \pi/2]^2$. For $m = 25, 100, 400$, we investigate the power of the approximate information test described in Example 4.

For simplicity, we report power functions along two arcs of alternatives,

$$\gamma_1(t) = \psi((t, \arctan\sqrt{2})) \quad \text{and} \quad \gamma_2(t) = \psi((\pi/4, t)),$$

for $t \in (0, \pi/2)$. For each choice of m, we generate 5 random samples of points in $[0, \pi/2]^2$. For each random sample, we estimate the power of the approximate information test as t varies in $(0, \pi/2)$. Details are reported in Figure 6. The resulting power functions are plotted in Figures 7 and 8.

Fix $\bar{\sigma} = \psi((\pi/4, \arctan\sqrt{2}))$, $n = 30$, and $\alpha = 0.05$. For $m = 25, 100, 400$ and $a = 1, \ldots, 5$, generate $\tau_1, \ldots, \tau_m \sim \text{Uniform}[0, \pi/2]^2$. Compute $\sigma_i = \psi(\tau_i)$. Set $B = 1000$.

1. Construct a representation of the submanifold in $\Re^2$.
   (a) Compute the pairwise Hellinger distances between $\bar{\sigma}, \sigma_1, \ldots, \sigma_m$. Construct $\mathcal{G}$ by connecting vertices i and j if either vertex i is one of vertex j's $K = 10$ nearest neighbors or vice versa.
   (b) Compute the pairwise shortest path distances in $\mathcal{G}$. Embed the shortest path distances in $\Re^2$ by minimizing the raw stress criterion, obtaining $\bar{z}, z_1, \ldots, z_m$.
2. Estimate the critical value. For $b = 1, \ldots, B$, draw o from a multinomial distribution with n trials and probability vector $\bar{\sigma}$.
   (a) Compute the Hellinger distances between o/n and $\bar{\sigma}, \sigma_1, \ldots, \sigma_m$ and determine the $\ell = 3$ nearest neighbors of o/n.
   (b) Compute $y(\vec{x})$ and $\|y(\vec{x}) - \bar{z}\|$. The estimated critical value $\hat{C}$ is the $1 - \alpha$ quantile of the B values of $\|y(\vec{x}) - \bar{z}\|$.
3. Estimate power at each alternative. For $k = 1, 2$, $t \in \{1, \ldots, 99\} \cdot \pi/200$, and $b = 1, \ldots, B$, draw o from a multinomial distribution with n trials and probability vector $\psi(\gamma_i(t))$.
   (a) Compute the Hellinger distances between o/n and $\bar{\sigma}, \sigma_1, \ldots, \sigma_m$ and determine the $\ell = 3$ nearest neighbors of o/n.
   (b) Compute $y(\vec{x})$ and $\|y(\vec{x}) - \bar{z}\|$. The estimated power is the proportion of the B values of $\|y(\vec{x}) - \bar{z}\|$ that exceed $\hat{C}$.

Figure 6: Design of the simulation experiments that produced the estimated power functions plotted in Figures 7 and 8.

For comparison, Figures 7 and 8 also display the estimated power function of the unrestricted likelihood ratio test. Its performance, which hardly exceeds $\alpha = 0.05$ at any alternative on either curve, is profoundly unsatisfying. (To confirm preliminary impressions, we used $B = 20000$ replications for each alternative. We also estimated the power function of Pearson's chi-squared test, obtaining similar results.) Such low power reminds us of the limitations of likelihood ratio tests.

For composite hypotheses, likelihood ratio tests are heuristic procedures inspired by the Neyman-Pearson Lemma. The theory that supports their use is asymptotic, and $n = 30$ is a rather small sample size for a multinomial distribution with 7 possible outcomes. Furthermore, a summand in G^2 will be unstable if either o_j or $n\bar{\theta}_j$ is small. It may well be that one of the goodness-of-fit statistics proposed by Read and Cressie (1988) is better suited to the demands of Example 4.

In any event, our interest lies in the ability of approximate information tests to learn an unknown statistical submanifold well-enough to derive some benefit from exploiting its structure. The estimated power functions displayed in Figures 7 and 8 clearly demonstrate the viability of our approach. Substantial variation in the cyan power functions suggests that sampling $m = 25$ points on the submanifold in question may be insufficient to construct a reliable test; in contrast, the blue ($m = 100$) and magenta ($m = 400$) power functions are quite similar. In general, the extent of sampling needed to construct a reliable test will surely depend on the dimension of the submanifold, the extent of its curvature, and the probability distribution that generates points on the submanifold.

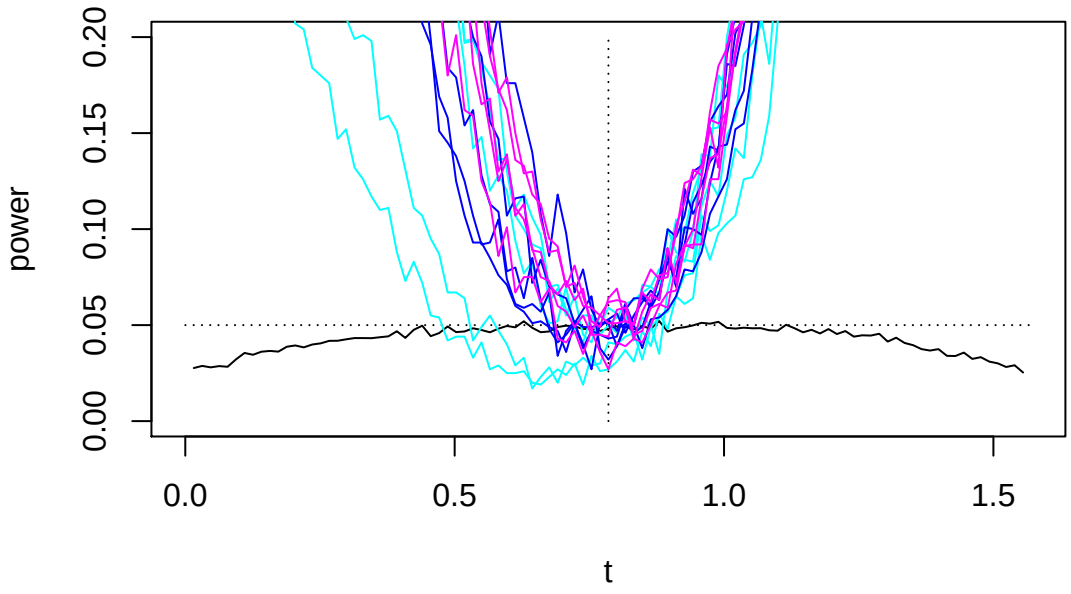

Figure 7: Estimated power functions produced by the simulation experiment described in Figure 6 for alternatives lying on the curve $\gamma_1(t)$. Five replications of the approximate information test with $m = 25, 100, 400$ points drawn from the submanifold are plotted in cyan, blue, and magenta. The estimated power function of the unrestricted likelihood ratio test is plotted in black.

It is worth pausing to reflect on what we have learned. We have posited a small-sample situation in which we want to test a hypothesis about the probability vector of a multinomial distribution with 7 possible outcomes. The parameter space of all such multinomial distributions is the 6-dimensional simplex Δ^6, but the parameters of interest lie on an unknown 2-dimensional submanifold of that simplex. The unrestricted likelihood ratio test has extremely low power against alternatives on the submanifold; nevertheless, the ability to sample points from the submanifold allows us to construct a test with excellent power on the submanifold.

8. Discussion

It is widely believed throughout the statistics community that restricted tests are more powerful than unrestricted tests. Indeed, although restricted tests may not be uniformly more powerful than unrestricted tests, our experience has been that the former generally outperform the latter. In consequence, we prefer restricted likelihood ratio tests to unrestricted likelihood ratio tests. But restricted likelihood ratio tests can only be constructed when the restriction to a parametric family

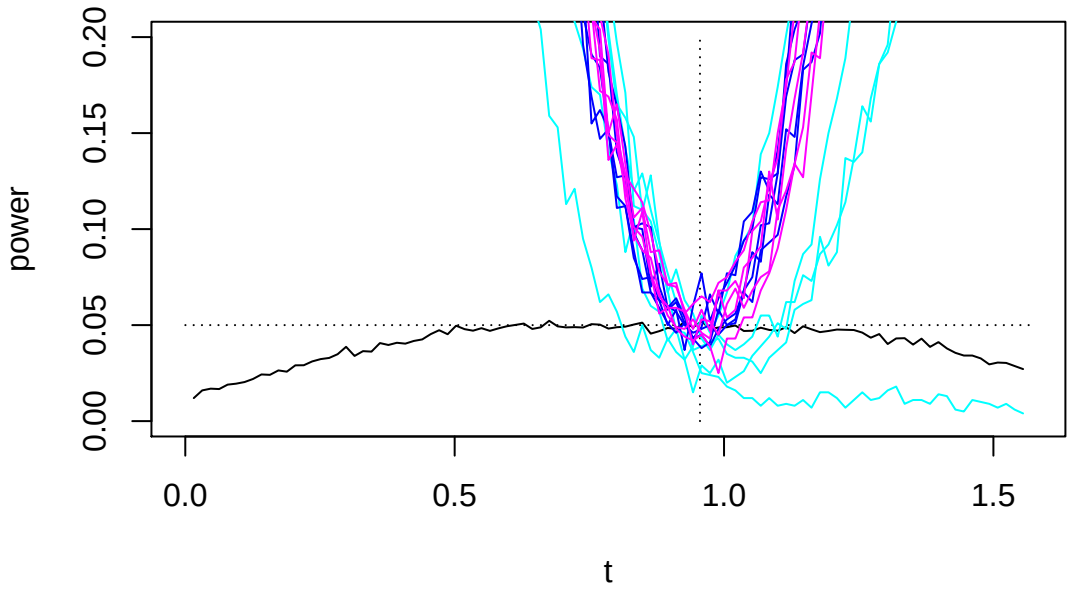

Figure 8: Estimated power functions produced by the simulation experiment described in Figure 6 for alternatives lying on the curve $\gamma_2(t)$. Five replications of the approximate information test with $m = 25, 100, 400$ points drawn from the submanifold are plotted in cyan, blue, and magenta. The estimated power function of the unrestricted likelihood ratio test is plotted in black.

of probability distributions is known and tractable. It is not clear that the low-dimensional structure of a restricted submanifold of distributions can be exploited when the submanifold is unknown.

For 1-sample problems with simple null hypotheses, we have proposed information tests that are locally asymptotically equivalent to likelihood ratio tests. Except in the special case of 1-dimensional submanifolds, these tests are computationally less tractable than likelihood ratio tests—typically intractable. Unlike likelihood ratio tests, however, information tests can be approximated when the relevant submanifold of distributions is unknown.

Implementing an approximate information test requires the user to make numerous decisions. These include: (1) how to construct the localization graph $\mathcal{G}$ and choose a suitable localization parameter, e.g., ϵ or K, an important problem in manifold learning; (2) how to choose a suitable dimension r in which to embed $\mathcal{G}$; (3) what criterion to optimize when embedding $\mathcal{G}$ in $\Re^r$, as well what algorithm to compute the embedding; (4) how to construct the nonparametric density estimate $\hat{p}_n$ (not an issue for the multinomial examples considered herein); (5) what out-of-sample embedding technique to use for embedding $\hat{p}_n$; and (6) how many simulated samples to draw from the hypothesized distribution. Clearly, a comprehensive investigation of how these decisions affect power is beyond the scope of the present manuscript.

The proposed approximate informations tests rely on isomap to learn the unknown statistical submanifold well enough that the advantages of restricted inference can be realized by working on the learnt submanifold. Doing so exemplifies manifold learning for subsequent inference, as opposed to the more conventional use of manifold learning for generic nonlinear dimension reduction, as an end in itself. Without specification of a subsequent exploitation task, it is difficult to provide guidance on how to select the inevitable tuning parameters, for example, the localization parameter ϵ or K in step 1 of isomap. Indeed, one often hears the quip that such choices are "features, not bugs." The framework of approximate information testing makes it possible to ask specific questions, such as "how must one choose the tuning parameters so as to guarantee that the power of the approximate information test converges to the power of the information test as one samples more extensively from the unknown submanifold?" Although our current implementation has not advanced beyond the feature-not-bug approach to manifold learning, we submit that the possibility of doing so is an advance in and of itself.

While local asymptotic theory commends the use of restricted tests, it does not guarantee that finite approximations of restricted tests will outperform unrestricted tests using finite sample sizes. Nevertheless, we report examples in which the unknown submanifold of distributions can be estimated well enough to realize gains in power. A preliminary version of our methodology has already been used to infer brainwide neural-behavioral maps from optogenetic experiments on *Drosophila* larvae (Vogelstein et al., 2014). However, in contrast to the methods discussed herein, that application involved a 2-sample problem with a composite null hypothesis. Moreover, the data observed in that application are estimates of parameter values that only lie near (not necessarily on) the statistical submanifold. These insights suggest two natural extensions of our methods.

First, consider the 2-sample problem of testing the composite null hypothesis $H_0 : \theta_a = \theta_b$ against the alternative $H_1 : \theta_a \neq \theta_b$. A corresponding test statistic for this problem is the information distance between $\hat{\theta}_a$ and $\hat{\theta}_b$, but corresponding theory remains to be developed. Second, consider the problem that results from replacing the randomly generated $\theta_1, \ldots, \theta_m \in \Psi$ with randomly generated $\vec{\theta}_1, \ldots, \vec{\theta}_m$ near Ψ. The same methods can be used, but replacing known θ_i with approximated $\vec{\theta}_i$ introduces another layer of uncertainty. We are currently exploring both extensions in related work.

Acknowledgments

This work was partially supported by the Naval Engineering Education Consortium (NEEC), Office of Naval Research (ONR) Award Number N00174-19-1-0011, as well as by DARPA XDATA contract FA8750-12-2-0303, SIMPLEX contract N66001-15-C-4041, GRAPHS contract N66001-14-1-4028, and D3M contract FA8750-17-2-0112.

Appendix A.

This appendix contains the proofs of Theorems 1 and 2.

A.1 Proof of Theorem 1

Let $\gamma = \{p_t \, d\mu : t \in (-\epsilon, \epsilon)\}$ denote a smooth variation in the statistical manifold P and consider the Taylor expansion

$$h^2\left(p_t, p_0\right) = h^2\left(p_0, p_0\right) + \left.\frac{d}{dt} h^2\left(p_t, p_0\right)\right|_{t=0} t + \frac{1}{2}\left.\frac{d^2}{dt^2} h^2\left(p_t, p_0\right)\right|_{t=0} t^2 + o\left(t^2\right). \tag{5}$$

Of course $h^2(p_0, p_0) = 0$. Writing

$$h^2\left(p_t, p_0\right) \quad = \quad \int_\Omega \left[2\sqrt{p_t(x)} - 2\sqrt{p_0(x)}\right]^2 d\mu(x)$$

$$
\begin{aligned}
&= 4 \int_\Omega \left[p_t(x) - 2\sqrt{p_t(x) p_0(x)} + p_0(x) \right] d\mu(x) \\
&= 8 - 8 \int_\Omega \left[p_t(x) p_0(x) \right]^{1/2} d\mu(x)
\end{aligned}
$$

and assuming standard regularity conditions that permit differentiation under the integral sign, we obtain

$$
\begin{aligned}
\left. \frac{d}{dt} h^2 \left(p_t, p_0 \right) \right|_{t=0} &= -8 \int_\Omega \left. \frac{d}{dt} \left[p_t(x) p_0(x) \right]^{1/2} \right|_{t=0} d\mu(x) \\
&= -4 \int_\Omega \left[p_0(x) p_0(x) \right]^{-1/2} p_0(x) \left. \frac{d}{dt} p_t(x) \right|_{t=0} d\mu(x) \\
&= -4 \left. \frac{d}{dt} \int_\Omega p_t(x) \, d\mu(x) \right|_{t=0} \\
&= -4 \left. \frac{d}{dt} 1 \right|_{t=0} = 0.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\left. \frac{d^2}{dt^2} h^2 \left(p_t, p_0 \right) \right|_{t=0} &= -8 \int_\Omega \left. \frac{d^2}{dt^2} \left[p_t(x) p_0(x) \right]^{1/2} \right|_{t=0} d\mu(x) \\
&= -8 \int_\Omega \left. \frac{d}{dt} \left\{ \frac{1}{2} \left[p_t(x) p_0(x) \right]^{-1/2} p_0(x) \frac{d}{dt} p_t(x) \right\} \right|_{t=0} d\mu(x) \\
&= -8 \int_\Omega \left\{ -\frac{1}{4} \left[p_t(x) p_0(x) \right]^{-3/2} p_0(x) \frac{d}{dt} p_t(x) p_0(x) \frac{d}{dt} p_t(x) + \right. \\
&\qquad \left. \frac{1}{2} \left[p_t(x) p_0(x) \right]^{-1/2} p_0(x) \frac{d^2}{dt^2} p_t(x) \right\} \Bigg|_{t=0} d\mu(x) \\
&= 2 \int_\Omega \left[\frac{\left. \frac{d}{dt} p_t(x) \right|_{t=0}}{p_0(x)} \right]^2 p_0(x) \, d\mu(x) - 4 \int_\Omega \left. \frac{d^2}{dt^2} p_t(x) \right|_{t=0} d\mu(x) \\
&= 2 \int_\Omega \left[\left. \frac{d}{dt} \log p_t(x) \right|_{t=0} \right]^2 p_0(x) \, d\mu(x) - 4 \left. \frac{d^2}{dt^2} \int_\Omega p_t(x) \, d\mu(x) \right|_{t=0} \\
&= 2 I_\gamma \left(p_0 \right),
\end{aligned}
$$

where I_γ denotes Fisher information with respect to the 1-dimensional submanifold γ. Substituting the preceding expressions into (5) yields

$$
h^2 \left(p_t, p_0 \right) = I_\gamma \left(p_0 \right) t^2 + o\left(t^2 \right). \tag{6}
$$

Passing from variations to the (parametrized) manifold P, we write $p_t = p(\cdot, \theta_t)$ and obtain

$$
h^2 \left(p\left(\cdot, \theta_t \right), p\left(\cdot, \theta_0 \right) \right) = \left(\theta_t - \theta_0 \right)^\top I\left(\theta_0 \right) \left(\theta_t - \theta_0 \right) + o\left(\left\| \theta_t - \theta_0 \right\|^2 \right). \tag{7}
$$

Having derived this expression, the variation γ is vestigial and we replace θ_t in (7) with θ. ■

A.2 Proof of Theorem 2

The proof of Theorem 2 requires a technical result about the remainder term in (7).

Lemma 4 *For any well-defined events A and B, $P(A) \geq 1 - \alpha/2$ and $P(B) \geq 1 - \alpha/2$ entails $P(A \cap B) \geq 1 - \alpha$.*

Proof Notice that $(A \cap B)^c = A^c \cup B^c$. Applying Boole's Inequality,

$$P\left(A^c \cup B^c\right) \leq P\left(A^c\right) + P\left(B^c\right) \leq \alpha/2 + \alpha/2 = \alpha.$$

■

Lemma 5 *Let*

$$r(\theta) = h^2\left(p(\cdot,\theta), \bar{p}\right) - \left(\theta - \bar{\theta}\right)^\top I\left(\bar{\theta}\right)\left(\theta - \bar{\theta}\right).$$

If (3) holds with $\theta_t = \theta$ and $\theta_0 = \bar{\theta}$, then

$$n\left|r\left(\hat{\theta}_n\right)\right| = o_p(1).$$

Proof Given $c, \alpha > 0$, we seek to demonstrate the existence of N such that $n \geq N$ entails

$$P\left(n\left|r\left(\hat{\theta}_n\right)\right| \geq c\right) < \alpha.$$

Let T denote the random variable to which $\sqrt{n}(\hat{\theta}_n - \bar{\theta})$ converges in distribution and choose $\epsilon > 0$ such that

$$P\left(\epsilon\|T\|^2 \geq c\right) < \frac{\alpha}{4}.$$

Choose N_1 such that $n \geq N_1$ entails

$$\left|P\left(\epsilon\left\|\sqrt{n}\left(\hat{\theta}_n - \bar{\theta}\right)\right\|^2 \geq c\right) - P\left(\epsilon\|T\|^2 \geq c\right)\right| < \frac{\alpha}{4},$$

and hence that

$$P\left(B_n^c\right) = P\left(\epsilon\left\|\sqrt{n}\left(\hat{\theta}_n - \bar{\theta}\right)\right\|^2 \geq c\right) < \frac{\alpha}{4} + \frac{\alpha}{4} = \frac{\alpha}{2}.$$

Because $r(\theta) = o(\|\theta - \bar{\theta}\|^2)$, there exists $\delta > 0$ such that $\|\hat{\theta}_n - \bar{\theta}\| < \delta$ entails

$$\frac{\left|r\left(\hat{\theta}_n\right)\right|}{\left\|\hat{\theta}_n - \bar{\theta}\right\|^2} < \epsilon, \quad \text{hence} \quad n\left|r\left(\hat{\theta}_n\right)\right| < \epsilon\left\|\sqrt{n}\left(\hat{\theta}_n - \bar{\theta}\right)\right\|^2.$$

Choose N_2 such that $n \geq N_2$ entails

$$P\left(\left\|\hat{\theta}_n - \bar{\theta}\right\| < \delta\right) \geq 1 - \frac{\alpha}{2},$$

hence

$$P\left(A_n\right) = P\left(n\left|r\left(\hat{\theta}_n\right)\right| < \epsilon\left\|\sqrt{n}\left(\hat{\theta}_n - \bar{\theta}\right)\right\|^2\right) \geq 1 - \frac{\alpha}{2}.$$

Let $N = \max(N_1, N_2)$. Then $n \geq N$ entails

$$P\left(n\left|r\left(\hat{\theta}_n\right)\right| < c\right) \geq P\left(A_n \cap B_n\right) \geq 1 - \alpha$$

by Lemma 4. ■

The relation between the HD and Wald statistics is now straightforward. Applying (2), (3), and Lemma 5,

$$\begin{aligned}
n\mathrm{HD}_n - W_n &= n\left(\hat{\theta}_n - \bar{\theta}\right)^\top I\left(\bar{\theta}\right)\left(\hat{\theta}_n - \bar{\theta}\right) + o_p(1) - n\left(\tilde{\theta}_n - \bar{\theta}\right)^\top I\left(\bar{\theta}\right)\left(\tilde{\theta}_n - \bar{\theta}\right) \\
&= \left[I^{-1}\left(\bar{\theta}\right) Z_n\left(\bar{\theta}\right) + o_p(1)\right]^\top I\left(\bar{\theta}\right)\left[I^{-1}\left(\bar{\theta}\right) Z_n\left(\bar{\theta}\right) + o_p(1)\right] + o_p(1) \\
&\quad - \left[I^{-1}\left(\bar{\theta}\right) Z_n\left(\bar{\theta}\right) + o_p(1)\right]^\top I\left(\bar{\theta}\right)\left[I^{-1}\left(\bar{\theta}\right) Z_n\left(\bar{\theta}\right) + o_p(1)\right] \\
&= o_p(1).
\end{aligned}$$

■

References

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Press, Boca Raton, FL, 2011.

R. J. Beran. Minimum Hellinger distance estimation for parametric models. *Annals of Statistics*, 5: 445–463, 1977.

J. de Leeuw. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, 5:163–180, 1988.

K. E. Giles, M. W. Trosset, D. J. Marchette, and C. E. Priebe. Iterative denoising. *Computational Statistics*, 23(4):497–517, 2008.

A. Gous. Spherical subfamily models. Available at `http://yaroslavvb.com/papers/gous-spherical.pdf`, November 10, 1999.

J. C. Gower. Some distance properties of latent root and vector methods in multivariate analysis. *Biometrika*, 53:325–338, 1966.

J. C. Gower. Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55(3):582–585, 1968.

K. Hall and T. Hoffman. Learning curved multinomial subfamilies for natural language processing and information retrieval. In *ICML 2000: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 351–358. Morgan Kaufmann, 2000.

N. J. Hicks. *Notes on Differential Geometry*. Van Nostrand Reinhold Company, London, 1971.

R. E. Kass. The geometry of asymptotic inference. *Statistical Science*, 4(3):188–219, 1989.

A. J. Kearsley, R. A. Tapia, and M. W. Trosset. The solution of the metric STRESS and SSTRESS problems in multidimensional scaling using Newton's method. *Computational Statistics*, 13(3): 369–396, 1998.

Y. Matsushima. *Differentiable Manifolds*. Marcel Dekker, New York, 1972.

J. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963. Annals of Mathematical Studies, Study 51.

M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*. Chapman & Hall, London, 1993.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

T. R. C. Read and N. A. C. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York, 1988.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

D. G. Simpson. Hellinger deviance test: Efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, 84:107–113, 1989.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

W. S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17:401–419, 1952.

M. W. Trosset, M. Gao, and C. E. Priebe. On the power of likelihood ratio tests in dimension-restricted submodels. arXiv:1608.00032, 2016.

J. T. Vogelstein, Y. Park, T. Ohyama, R. Kerr, J.W. Truman, C. E. Priebe, and M. Zlatic. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science*, 344 (6182):386–392, 25 April 2014.

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.