Machine Learning in Mental Health Data (Design Project MATH F376)

Moitrish Majumdar

BITS Pilani, K.K Birla Goa Campus

Semester II, 2019-20

Contents

1 Introduction **1**

2 Overview of differential geometric concepts in econometrics **2**

3 Elements and areas of Econometrics **4**

4 Geodesic distance and hypothesis testing **6**

5 References **8**

1 Introduction

In the course of this study, I intend to review the tools required to sufficiently understand differential geometric applications in Econometrics. Broadly speaking, econometrics refers to empirical and quantitative analysis done in order to test economic theories, forecast outcomes and evaluate policy outcomes. This study does not seek a comprehensive exploration of differential geometric theory and nor does it claim to be an extensive treatise on the relevant theorems and their respective proofs involved in such an exploration. It is merely a basic introduction to some special mathematical structures and their properties which are exploited for developing econometric models.

At the outset, it is established that econometric models can take the form of objects known as *manifolds* , the study of which is the central idea in a differential geometric treatment. The development of this idea leads to a discussion on other properties of these structures such as *tangent spaces* and *metrics*.

In the subsequent discussions, we focus our interest on a specific metric, known as the Riemannian metric. Further, we see its' application in the Fisher information setup, which is indispensable for any differential geometric study of econometrics. The development of these concepts finally leads to a number of interesting applications as wide ranging as hypothesis testing, analysing regression models and maximum likelihood estimators. For the purpose of this report,

I focus on a specific application which falls within the ambit of hypothesis testing, usng an approach known as a *distance test*. This method of hypothesis testing uses the idea of the *Rao distance* to test a null of the form

$$H_0 : g(\theta) = 0$$

versus the alternative

$$H_1 : g(\theta) \neq 0$$

For a more exhaustive and detailed review of current differential geometric statistical theory see Amari (2015) or from a more purely mathematical background, see Kumaresan (2001).

2 Overview of differential geometric concepts in econometrics

The theory of manifolds is fundamental to the development of differential geometry, although we do not need the full abstract theory. A few preliminary concepts are important to develop this theory.

Consider any set X and let

$$\tau \subset P(X)$$

where $P(X)$ is the power set of X such that

a) The empty set $\phi \in \tau$ and the set itself $X \in \tau$

b) τ is closed under arbitrary union

c) τ is closed under finite intersection,

Then, we say that τ defines a **topology** on X or (X, τ) is a **topologial space**, and elements of τ are called open sets with respect to the topology τ . Some examples of topological spaces include the indiscrete topology, where $\tau = \{\phi, X\}$, discrete topology ($\tau = P(X)$) and the Zariski topology, where $X = \mathbf{R}$ and $U \in \tau$ iff U^c is a finite set.

Moreover, a topological space (X, τ) is called **Hausdorff** if for every $x, y \in X$ with $x \neq y \; \exists \; U_x, U_y \in \tau$ such that $x \in U_x$ and $y \in U_y$ and $U_x \cap U_y = \phi$.

Let X be a topological space, then we say that X is a **differentiable manifold** of dimension n if-

a) $\forall a \in X$, $\exists$ an open set $U_a \subset X$ such that $a \in U_a$ and there is a homeomorphism ϕ of U onto some open subset $\phi(U)$ where the image $\phi(U)$ is an open subset of $\mathbf{R}^n$ (A homeomorphism is a continuous, bijective map whose inverse is continuous).

b) If there is an open covering U_α where $\alpha \in X$ and homeomorphisms ϕ_α from U_α onto an open subset $\phi_\alpha(U_\alpha)$ of $\mathbf{R}^m(\alpha)$ where $m(\alpha)$ is a non-negative integer, moreover for $\alpha, \beta \in X$, $U_\alpha \cap U_\beta$ is non-empty, then the map

$$\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta)$$

is a diffeomorphism (it is a bijective, differentiable map whose inverse is also differentiable). Here, $\phi_\beta(U_\alpha \cap U_\beta) \subset \mathbf{R}^m(\alpha)$ and $\phi_\alpha(U_\alpha \cap U_\beta) \in \mathbf{R}^{m_1(\alpha)}$. Hence, all $m(\alpha)$'s are the same, and equal to the dimension n .

This characterisation tells us that a differentiable manifold is a space which locally looks like some $\mathbf{R}^n$ on which we can speak of differentiable functions.

However, to deal with the concept of distance on differentiable manifolds, we should have a well-defined norm for the tangent vectors of the curves in a manifold. By a **curve** in a manifold X , we mean a smooth map from $(\alpha, \beta) \to X$. Now, if X is a differentiable manifold of dimension n, some $V \in \mathbf{R}^n$ is called a **tangent vector** to X at point $p \in X$ if there is a differentiable function

$$\gamma : (-\epsilon, \epsilon) \to X$$

such that $\gamma(0) = p$ and $\gamma^{'}(0) = V$.

Intuitively, the concept of distances on manifolds can be understood as follows $(\forall x, y \in X)$

$$d(x, y) = \inf\{l(\gamma([a, b])) : \gamma : [a, b] \to X\}$$

Here, $\gamma([a, b]) \subset \mathbf{R}^n$ and $l(\gamma)$ is defined as

$$l(\gamma) := \int_a^b \|\gamma(t)\| dt$$

Using the idea of **directional derivatives**, we can understand the method of computing tangent vectors in an n-dimensional manifold.

If we consider an n-dimensional manifold X having a local co-ordinate system $\xi = (\xi^1, \xi^2, ...\xi^n)$, the **tangent space** T_ξ at a point ξ is a vector space spanned by n tangent vectors along the co-ordinate curves of ξ^i , i.e the n tangent vectors (each denoted by e_i) form the basis of the tangent space $T_\xi X$. Here,

$$e_i = \frac{\partial}{\partial \xi_i}$$

The tangent vector e_i operates on the differentiable function $f(\xi)$ and gives its partial derivative in the direction of co-ordinate curve ξ^i .

Amari (2015) identifies the tangent vector with the "score function" incase of a manifold of probability distributions. He identifies the expression for e_i as

$$e_i = \partial_i \log\Big(p(x, \xi)\Big)$$

Broadly speaking, a **Riemannian metric** g on a manifold X is a map $p \mapsto g_p$ where g_p is a positive definite inner product on $T_p X$. With this definition of tangent vectors, consider the special inner product:

$$\langle e_i, e_j \rangle = \langle \partial_i \log(p(x, \xi)) \partial_j \log(p(x, \xi)) \rangle = E[\partial_i \log(p(x, \xi)) \partial_j \log(p(x, \xi))] = g_{ij}$$

where $E[X]$ flows from the notion of expectation in probability, i.e :

$$E[X] = \int_{-\infty}^{+\infty} x p(x) dx$$

The matrix defined by these g_{ij}'s forms the **Fisher information matrix** (Fisher information is formally defined as the covariance of the score function, or the amount of information a variable carries about the unknown parameter

that models it) . Thus, we see that the Fisher information induces a Riemannian metric as the g_{ij}'s form a positive definite matrix.

We can explore how this matrix looks for different cases and distributions. It is important to note that the classical notation for a Riemannian metric is given by

$$ds^2 = \sum_{i,j} g_{ij} dx_i dx_j$$

Now, let us consider a general setup where we look at a 2-dimensional submanifold of $\mathbf{R}^3$ where the co-ordinates on are given by (u, v) . Thus, it can be seen that the are:

$$g_{11} = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2$$

$$g_{12} = g_{21} = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v}$$

$$g_{22} = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2$$

Furthermore, it should be noted that in classical notation, $g_{11} = E$, $g_{12} = g_{21} = F$ and $g_{22} = G$. Thus, in co-ordinate representation,

$$g = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

From the classical notation of the Riemannian metric, we see in this case,

$$ds^2 = Edu^2 + 2Fdu.dv + Gdv^2$$

This is called the first fundamental form of the surface. The first fundamental forms in a normal distribution and further metrics shall be described in section 5.

In this exercise, I define **Christoffel symbols** instead of the theory of connections purely for convenience of use. Using the theory of Fisher information, the Christoffel symbols are simply defined as:

$$\Gamma^l_{ki} = 1/2 \sum_j \left(\frac{\partial g_{li}}{\partial x_j} + \frac{\partial g_{lj}}{\partial x_i} - \frac{\partial g_{ij}}{\partial x_l}\right) g^{lk}$$

The Christoffel symbols are used to derive geodesics on surfaces in section 5. Note that here, g^{lj} are the elements of the *inverse* of the Fisher information matrix.

3 Elements and areas of Econometrics

As a discipline, econometrics is relatively new and has seen rapid expansion and development in the past century. By emphasizing the quantitative aspects of economic relationships, econometrics calls for a 'unification' of measurement and theory in economics. Theory without measurement can have only limiled relevance for the analysis of actual economic problems; while measurement without theory, being devoid of necessary explanations for the interpretation

of the statistial observation, is unlikely to result in a satisfactory understanding of the way economic forces interact with each other. Neither 'theory' nor 'measurement' on its own is sufficient to further our understanding of economic phenomena.

Heterogeneity of economic relations across individuals, firms and industries is increasingly acknowledged, and attempts have been made to take them into account either by integrating out their effects or by remodeling the sources of heterogeneity when suitable panel data exists. Multiple elements of statistical inference such as hypothesis testing also find applications in econometrics. Thus, it becomes important to analyse the geometry of various distributions and their properties, which form the fundamental basis of models.

The major focus will be on *exponential families*. Let $\theta \in \Theta \subset \mathbf{R}^r$ be an r-dimensional parameter vector, then we write it in component terms as $\theta = (\theta^1, \theta^2, ...\theta^r)'$. Let X be a random variable and $s(X) = (s_1(x), ..., s_r(x))'$ be an r-dimensional statistic. Then, we can represent the probability distribution of an exponential family as

$$p(x|\theta) = exp\Big\{ \sum_{i=1}^{r} \theta^i s_i - \psi(\theta) \Big\} m(x).$$

The densities are defined with respect to some fixed dominating measure ν . The function $m(x)$ is non-negative and independent of the parameter vector θ . We further assume that the components of $s(X)$ are not linearly dependent. We call Θ the natural parameter space and assume it contains all θ such that

$$\int exp\{\theta^i s_i\} m(x)\, d\nu < \infty$$

A parametric set of densities of this form is called a *full exponential family* . If Θ is open in $\mathbf{R}^r$ then the family is said to be regular, and the statistics $(s_1, ...s_r)'$ are called the *canonical statistics* .

The function $\psi(\theta)$ plays an important role and is represented as

$$\log \Big(\int exp\{\theta^i s_i\} m(x)\, d\nu \Big)$$

It is easy to check that the expression for $\psi(\theta)$ follows from the property that the integral of the density $p(x|\theta)$ is one. It can also be interpreted in terms of the moment generating function of the canonical statistic S. This is given by $M(S; t; \theta)$ where

$$M(S; t; \theta) = exp\{\psi(\theta + t) - \psi(\theta)\}$$

Full exponential families have a natural geometrical characterisation similar to the affine sub-spaces in the space of all density functions. They therefore play the role that lines and planes do in three-dimensional Euclidean geometry. Common distributions such as normal and exponential distributions (which are utiised in several applications later) fall within this family. The simplest examples of full exponential families in econometrics are the standard regression model and the linear simultaneous equation model. The other standard building blocks of univariate statistical theory including the Poisson, gamma, Bernoulli, binomial and multinomial families are also full exponential families.

In a seminal paper Rao (1945) introduced the concept of **Geodesic Distance** (or **Rao Distance**) into statistics. This concept has important theoretical properties and is based on the demanding differential- geometrical approach to statistics.

If we let $S = \{p(x|\theta)|\theta \in \Theta\}$ be a statistical model, we therefore see that such an n-parametric family of distributions behaves only locally like $\mathbf{R}^n$, so that its potential curvature can be analysed. Coordinates in this statistical model can be changed by admissible (smooth) transformations being continuously differentiable and having a non-singular functional determinant. In this way, S is equipped with a differentiable structure and S is a differentiable n-manifold. Then, local coordinates are transferred from $\Theta \in \mathbf{R}^n$ to open neighbourhoods of the points on the n-manifold.

Now let $C_\lambda = \{c_\lambda|\lambda \in \Lambda\}$ be the set of all curves lying completetely in S and connecting two distributions F_1 and F_2 repreented by parameter values θ_1 and θ_2 . With some $t_1 < t_2, c_\lambda(t_1) = \theta_1$ and $c_\lambda(t_2) = \theta_2$, the Rao distance (also called the geodesic distance or the Riemannian distance) is the minimum arc-length of all the curves:

$$d(F_1, F_2) = \min_{c_\lambda \in C_\Lambda} \int_{t_1}^{t_2} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}(\theta(t)) \frac{d\theta_i(t)}{dt} \frac{d\theta_j(t)}{dt}} dt$$

$d(F_1, F_2)$ is a mathematical distance, and the idea behind a distance test is as follows: the null hypothesis is rejected if the distance (weighted with the sample size) between the estimated distribution and the distribution under H_0 is too big. In some standard test problems, Rao distance tests are equivalent to classical t, χ^2 or F tests.

4 Geodesic distance and hypothesis testing

We look at Rao distances for two-parameter statistical families. Picking up from section 3, if we find the Fisher information matrix for a normal distribution $N(\mu, \sigma^2)$ in a setup where the co-ordinates are given by (μ, σ^2), it is:

$$g_{ij}(\theta) = 1/\sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

From section 3, it was seen that

$$ds^2 = \sum_{i,j} g_{i,j}(\theta) dx_i dx_j$$

was a positive definite differential form based on the information matrix. This can be used as a measure of distance between two distributions, whose parameter values are points in the parameter space Ξ (Atkinson and Mitchell, 1981).

If we let

$$\theta_i = \theta_i(t)$$

(where t is a parameter) denote a curve in Ξ joining two distributions with parameters θ_1 and θ_2 , Rao(1945) considered the variance of $\sum_{i=1}^{n} \frac{\partial log p(x,\xi)}{\partial \theta_i} d\theta_i$

The calculus approach can be used in order to find the Rao geodesic between two distributions. In this case, the curve joining two distributions F_1 and F_2 for which $d(F_1, F_2)$ is the shortest is of interest. This curve can be obtained as a solution of the differential equations (called the Euler-Lagrange equations in the calculus of variations approach):

$$\frac{d^2\theta_i}{dt^2} + \sum_{i,j} \Gamma^k_{ij} \frac{d\theta_i}{dt} \frac{d\theta_j}{dt} = 0$$

where Γ^k_{ij} are the Christoffel symbols as defined in section 3. The term $d(F_1, F_2)$ obtained as an integral expression in the last section is a solution of this equation.

An alternative approach using calculus was given by Garabedian (1964) and Courant and Hilbert (1961). In this approach, for the metric

$$ds^2 = \sum_{i,j} g_{i,j}(\theta) dx_i dx_j$$

consider s_0 as the distance *along* the geodesic. Define

$$\psi_i = 2 \sum_{j=1}^{n} g_{ij} \frac{d\theta_j}{ds_0}$$

and

$$\frac{d\psi_i}{ds_0} = (-1/4) \sum_{l,m=1}^{n} \frac{\partial g^{lm}}{\partial \theta_i} \psi_l \psi_m$$

where g^{lm} are the inverse elements of the infomation matrix. These two equations are called Hamiltons' equations and are satisfied by the geodesics in terms of the parameter s_0. It is assumed that

$$\psi_i \frac{d\theta_i}{ds_0} = c$$

where c is any constant, this is done to specify a length scale.

As an extension, if we specify $s^*(\theta_1, \theta_2)$ as the geodesic between the distribution described by θ_2 and the point of the outer end of θ_1, it can be shown that $s^*(\theta_1, \theta_2)$ is a solution of the non-linear partial differential equation

$$\sum_{i,j=1}^{n} \frac{\partial s^*(\theta_1, \theta_2)}{\partial \theta_{i1}} \frac{\partial s^*(\theta_1, \theta_2)}{\partial \theta_{j1}} = 1$$

This equation is known as the Hamilton-Jacobi equation and can also be expressed as

$$\sum_{i,j=1}^{n} g_{ij} \frac{\partial [s^*(\theta_1, \theta_2)]^2}{\partial \theta_{i1}} \frac{\partial [s^*(\theta_1, \theta_2)]^2}{\partial \theta_{j1}} = 4[s^*(\theta_1, \theta_2)]^2$$

It is important to note that the geodesic equations are mostly non-linear or partial differential equations, and are usually difficult to solve explicitly.

Another possible approach to find geodesic curves is the idea of isometry. A map $f : S_1 \to S_2$ is an isometry iff the following conditions hold-

a) f is a diffeomorphism and

b) f preserves distances in the two surfaces.

Moreover, the first fundamental forms in both setups must be equal. Thus, an instance would be the Poincaré metric whose first fundamental form is:

$$(d\mu^{*2} + d\sigma^{*2})/\sigma^{*2}$$

Focussing specifically on normal distributions, it is important to write

$$ds^2 = \sum_{i,j} g_{ij}(\theta) dx_i dx_j$$

explicitly for the Riemannian metric. Considering the local coordinate system to be (μ, σ) , the expression for ds^2 becomes:

$$ds^2 = (d\mu^2 + 2d\sigma^2)/\sigma^2$$

Now, if we look at two normal distributions whose means are same, i.e $\mu_1 = \mu_2$ the expression for the first fundamental form becomes $ds^2 = 2d\sigma^2/\sigma^2$. On integrating, we obtain the geodesic equation as

$$2^{1/2} \log |\sigma_1/\sigma_2|$$

Similarly, considering two distributions where the variations are the same, we obtain

$$|\mu_1 - \mu_2|/\sigma$$

However, when the two distributions differ in both mean and variance the Rao distance between them is obtained as :

$$2tanh^{-1}\left[\left((\mu_1^{*2} - \mu_2^{*2}) + (\sigma_1^{*2} - \sigma_2^{*2})\right)/\left((\mu_1^{*2} - \mu_2^{*2}) + (\sigma_1^{*2} + \sigma_2^{*2})\right)\right]^{1/2}$$

See Burbea and Rao (1982)

Burbea and Oller (1989) also derived the critical region a Rao distance test as

$$C = \left\{ U = (m_1 m_2/m_1 + m_2) d^2((\mu_1^*, \sigma_1^*), (\mu_2^*, \sigma_2^*)) \middle| U > u^* \right\}$$

where U follows $\chi^2(2)$

Several other interesting applications relating to the idea of statistical inferences exist, which include studying the effects of model curvature on building the critical region.

5 References

Shun-ichi Amari. *Information Geometry and its' Applications*, Springer, 2015

S. Kumaresan. *A Course in Differential Geometry and Lie Groups*, Hindustan Book Agency , 2001

Paul Marriott and Mark Salmon. Applications of Differential Geometry to Econometrics, Cambridge University Press, 2000

Colin Atkinson and Ann F. S. Mitchell. *Rao's Distance Measure*, Sankhya: The Indian Journal of Statistics , Vol. 43 , 1981