Geometry of Parametric Binary Choice Models

Hisatoshi Tanaka

School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku, Tokyo 169-8050, Japan hstnk@waseda.jp

Abstract. Parametric binary choice models are studied from the viewpoint of information geometry. The model set is a dually flat manifold with dual connections, which are naturally derived from the Fisher information metric. Under the dual connections, the canonical divergence and the Kullback-Leibler (KL) divergence of the binary choice model coincide if and only if the model is a logit. The results are applied to a logit estimation with linear constraints.

Keywords: Binary Choice Models · Discrete Choice Models · Logit · Multinomial Logit · Single-Index Models.

1 Introduction

Information geometry has been applied to econometric models such as the standard linear model, a Poisson regression, Wald tests, the ARMA model, and many other examples [3, 4, 8, 10]. In this work, we apply the method to a standard binary choice model. Let x be an $\mathbb{R}^d$-valued random vector. Let $y \in \{0, 1\}$ be a binary outcome such that

$$y = \begin{cases} 1 & \text{if} \quad y^* \geq 0 \\ 0 & \text{if} \quad y^* < 0 \end{cases}, \tag{1}$$

where $\theta \in \mathbb{R}^d$, $y^* = x \cdot \theta - \epsilon$, $\epsilon \perp\!\!\!\perp x$, and $E[\epsilon] = 0$. The choice probability conditioned on x is given by

$$\mathbf{P}\{y = 1 \mid x\} = \mathbf{P}\{\epsilon \leq x \cdot \theta \mid x\} = F(x \cdot \theta), \tag{2}$$

where the distribution F of ϵ is known to a statistician. Let p_θ be the density of the binary response model given by

$$p_\theta(y, x) = F(x \cdot \theta)^y (1 - F(x \cdot \theta))^{1-y} p_X(x), \quad (y, x) \in \{0, 1\} \times \mathbb{R}^d, \tag{3}$$

where p_X denotes the marginal density of x.

The model is widely used in social sciences to describe decision-makers' choices between two alternatives. These alternatives may represent school, labor supply, marital status, or transportation choices. See [9, 11] for a list of empirical applications in social sciences.

The model is called *probit* when F is the standard normal distribution, that is,

$$F(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds,$$

and *logit* when F is the standard logistic distribution, that is,

$$F(u) = \frac{\exp u}{1 + \exp u}. \tag{4}$$

In particular, the logit model is widely used because the choice probability $F(x \cdot \theta)$ has a closed form and is readily interpretable [11]. A goal of the paper is to show that, among parametric binary response models, the logit model exhibits good geometric properties because of its 'conditional' exponentiality.

The remainder of this paper is organized as follows. In Section 2, the geometry of the binary choice model is formulated. In Section 3, two divergences, the canonical divergence and the Kullback-Leibler (KL) divergence, are introduced. In particular, the logit is shown to be a unique model whose canonical divergence is equal to the KL divergence. In Section 4, the logit model with linear constraints is studied. In Section 5, the conclusions of this study are presented.

2 Geometry of the Binary Choice Models

Assume that $F : \mathbb{R} \to [0, 1]$ is a smooth distribution function of ϵ. Let $\Theta \in \mathcal{O}(\mathbb{R}^d)$ be a set of parameters θ. Given p_X, the model set

$$\mathcal{P} = \{p_\theta \mid \theta \in \Theta\} \tag{5}$$

is considered as a d-dimensional C^∞ manifold with a canonical coordinate system $\theta \mapsto p_\theta$.

The tangent space of $\mathcal{P}$ at p_θ is simply denoted as $T_\theta\mathcal{P}$ and is given by

$$T_\theta\mathcal{P} = \mathrm{Span}\left\{(\partial_1)_\theta, \cdots, (\partial_d)_\theta\right\},$$

where $\partial_i = \frac{\partial}{\partial\theta_i}$ for $i = 1, \cdots, d$. The score of the model is

$$\frac{\partial}{\partial\theta} \log p_\theta(y, x) = \frac{y - F(x \cdot \theta)}{F(x \cdot \theta)(1 - F(x \cdot \theta))} f(x \cdot \theta) x \tag{6}$$

and the Fisher information matrix is

$$G(\theta) = E\left(\frac{\partial}{\partial\theta} \log p_\theta\right)\left(\frac{\partial}{\partial\theta} \log p_\theta\right)^\top = E\left[\frac{f(x \cdot \theta)^2}{F(x \cdot \theta)(1 - F(x \cdot \theta))} x x^\top\right]$$

because $E\left[(y - F(x \cdot \theta))^2 \middle| x\right] = F(x \cdot \theta)(1 - F(x \cdot \theta))$.

The Fisher information metric g is introduced on $T_\theta\mathcal{P}$ by

$$g_\theta(X, Y) = E\left[X\left(\int_0^{x\cdot\theta} \frac{f(u)}{\sqrt{F(u)(1 - F(u))}} du\right) Y\left(\int_0^{x\cdot\theta} \frac{f(u)}{\sqrt{F(u)(1 - F(u))}} du\right)\right]$$

for X, $Y \in T_\theta \mathcal{P}$. In particular, the (i,j) component of g at θ is

$$g_{ij}(\theta) = g_\theta(\partial_i, \partial_j) = E\left[\frac{f(x\cdot\theta)^2}{F(x\cdot\theta)(1-F(x\cdot\theta))} x_i x_j\right].$$

Hence, the Levi-Civita connection ∇ of $(\mathcal{P}, g)$ is given by the connection coefficients

$$\begin{aligned}\Gamma_{ij,k}(\theta) &= \frac{1}{2}\Big[(\partial_i)_\theta g_{kj}(\theta) + (\partial_j)_\theta g_{ik}(\theta) - (\partial_k)_\theta g_{ij}(\theta)\Big] \\ &= E\left[\left(\frac{f(x\cdot\theta)f'(x\cdot\theta)}{F(x\cdot\theta)(1-F(x\cdot\theta))} + \frac{f(x\cdot\theta)^3(F(x\cdot\theta)-1/2)}{F(x\cdot\theta)^2(1-F(x\cdot\theta))^2}\right) x_i x_j x_j\right]\end{aligned}$$

for $1 \le i,j,k \le n$. The coefficients show symmetry on (i,j,k). In particular, $\Gamma_{ij,k}(\theta) = \Gamma_{ji,k}(\theta)$, which implies that $(\mathcal{P}, g, \nabla)$ is a tortion-free manifold.

The symmetry of the connection is caused by the single-index structure of the model: p_θ depends on θ only through the linear index $x\cdot\theta$. To see this, consider a general linear-index model $p(y, x\cdot\theta)$ of the joint density of (y,x). The score is

$$\frac{\partial}{\partial\theta}\log p(y, x\cdot\theta) = \frac{p_2(y, x\cdot\theta)}{p(y, x\cdot\theta)} x,$$

where $p_2(y, x\cdot\theta) := \frac{\partial}{\partial u} p(y,u)\big|_{u=x\cdot\theta}$. The Fisher information matrix is $G(\theta) = E[\gamma(y, x\cdot\theta) x x^\top]$ with

$$\gamma(y, x\cdot\theta) = \left(\frac{p_2(y, x\cdot\theta)}{p(y, x\cdot\theta)}\right)^2.$$

Then,

$$(\partial_i)_\theta g_{kj}(\theta) = (\partial_j)_\theta g_{ik}(\theta) = (\partial_k)_\theta g_{ij}(\theta) = E\left[\gamma_2(y, x\cdot\theta) x_i x_j x_k\right]$$

with $\gamma_2(y, x\cdot\theta) := \frac{\partial}{\partial u}\gamma(y,u)\big|_{u=x\cdot\theta}$, and symmetric connection coefficients

$$\Gamma_{ij,k}(\theta) = \frac{1}{2} E\left[\gamma_2(y, x\cdot\theta) x_i x_j x_k\right]$$

are obtained.

By the symmetry of the Levi-Civita connection, the α-connection $\nabla^{(\alpha)}$ is naturally defined by

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \frac{1-\alpha}{2}\Gamma_{ij,k}(\theta) \tag{7}$$

for each $\alpha \in \mathbb{R}$. A pair $(\nabla^{(\alpha)}, \nabla^{(-\alpha)})$ provides the dual connections of $(\mathcal{P}, g)$, such that

$$X g_\theta(Y,Z) = g_\theta(\nabla^{(\alpha)}_X Y, Z) + g_\theta(Y, \nabla^{(-\alpha)}_X Z)$$

for every $X, Y, Z \in \mathcal{X}(\mathcal{P})$, where $\mathcal{X}(\mathcal{P})$ is the family of smooth vector fields on $\mathcal{P}$.

Theorem 1. *$(\mathcal{P}, g, \nabla^{(+1)}, \nabla^{(-1)})$ is a dually flat space with dual affine cordinates (θ, η); θ is the $\nabla^{(+1)}$-affine coodinate and $\eta = (\eta_1, \cdots, \eta_d)$ given by*

$$\eta_j = E\left[\left(\int_0^{x\cdot\theta} \frac{f(u)^2}{F(u)(1-F(u))}du\right) x_j\right] \tag{8}$$

for $1 \leq j \leq d$ is the $\nabla^{(-1)}$-affine coodinate.

Proof. For $\alpha = 1$, $\Gamma_{ij,k}^{(+1)} \equiv 0$ holds for all i, j, and k. Moreover, since

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta)$$

holds with potential $\psi : \Theta \to \mathbb{R}$ defined by

$$\psi(\theta) = E\left[\int_0^{x\cdot\theta} \left(\int_0^v \frac{f(u)^2}{F(u)(1-F(u))}\, du\right) dv\right],$$

the dual-affine coordinates are obtained as follows:

$$\eta_j = \partial_j \psi(\theta) = E\left[\left(\int_0^{x\cdot\theta} \frac{f(u)^2}{F(u)(1-F(u))}\, du\right) x_j\right]$$

for $1 \leq j \leq d$. □

For later convenience, we denote the inverse function of

$$\partial\psi : \Theta \to \mathbb{R}^d, \theta \mapsto \eta = (\partial_1\psi(\theta), \cdots, \partial_d\psi(\theta))$$

by $(\partial\psi)^{-1}$. Because the Hesse matrix $\partial^2\psi(\theta)$ is equal to the Fisher information matrix $G(\theta)$ and is therefore positive definite, $\partial\psi : \Theta \to \partial\psi(\Theta)$ is invertible at any $\eta \in \partial\psi(\Theta)$.

The dual potential $\varphi(\eta)$ is given by

$$\varphi(\eta) = \max_\theta\ \eta \cdot \theta - \psi(\theta) = \eta \cdot (\partial\psi)^{-1}(\eta) - \psi((\partial\psi)^{-1}(\eta)), \tag{9}$$

which is the Legendre transformation of $\psi(\theta)$. Let $\partial^i = \frac{\partial}{\partial\eta_i}$ for $1 \leq i \leq d$, then $\theta^i = \partial^i\varphi(\eta)$ holds.

Corollary 1. *The $\nabla^{(\pm 1)}$-geodesic path connecting $p, q \in \mathcal{P}$ is given by $t \in [0,1] \to p_{\theta_t^{(\pm 1)}} \in \mathcal{P}$, where*

$$\theta_t^{(+1)} = (1-t)\theta_p + t\theta_q \tag{10}$$

and

$$\theta_t^{(-1)} = (\partial\psi)^{-1}((1-t)\eta_p + t\eta_q) \tag{11}$$

for $0 \leq t \leq 1$.

The $\nabla^{(-1)}$-geodesic is a solution to the ordinary differential equation,

$$\dot{\theta}_t = G(\theta_t)^{-1}(\eta_q - \eta_p),\ \theta_0 = \theta_p.$$

To see this, let $\eta_t^{(-1)} = \partial\psi(\theta_t^{(-1)}) = (1-t)\eta_p + t\eta_q$. Then,

$$\frac{d}{dt}\eta_t^{(-1)} = \partial^2\psi(\theta_t^{(-1)})\frac{d}{dt}\theta_t^{(-1)} = \eta_q - \eta_p,$$

where $G(\theta_t^{(-1)}) = \partial^2\psi(\theta_t^{(-1)})$.

3 The Logit Model

For a dually flat manifold with dual affine coordinates (θ, η) and dual potentials (ψ, φ), the *canonical divergence* (or *U-divergence with* $U = \psi$) is defined as

$$D(p\|q) = \varphi(\eta_p) + \psi(\theta_q) - \eta_p \cdot \theta_q \tag{12}$$

[2, 6, 7]. In the case of the binary response model, the divergence is shown as

$$\begin{aligned} D(p\|q) &= [\eta_p \cdot \theta_p - \psi(\theta_p)] + \psi(\theta_q) - \eta_p \cdot \theta_q \\ &= E\left[\int_{x\cdot\theta_p}^{x\cdot\theta_q}\left(\int_0^v \frac{f(u)^2}{F(u)(1-F(u))}du\right)dv\right] \\ &\quad -E\left[\left(\int_0^{x\cdot\theta_p}\frac{f(u)^2}{F(u)(1-F(u))}du\right)x\cdot(\theta_q-\theta_p)\right] \end{aligned} \tag{13}$$

for each p and q in $\mathcal{P}$ because $\varphi(\eta_p) = \eta_p \cdot \theta_p - \psi(\theta_p)$ and

$$\psi(\theta_q) - \psi(\theta_p) = E\left[\int_{x\cdot\theta_p}^{x\cdot\theta_q}\left(\int_0^v \frac{f(u)^2}{F(u)(1-F(u))}du\right)dv\right].$$

The following results are standard:

Theorem 2. *Let p, q, r be in $\mathcal{P}$. Let $\theta^{(+1)}$ be the $\nabla^{(+1)}$-geodesic path connecting p and q, and let $\theta^{(-1)}$ be the $\nabla^{(-1)}$-geodesic path connecting q and r. If $\theta^{(+1)}$ and $\theta^{(-1)}$ are orthogonal at the intersection q in the sense that*

$$g_q\left(\left(\frac{d}{dt}\right)_q\theta_t^{(+1)}, \left(\frac{d}{dt}\right)_q\theta_q^{(-1)}\right) = 0,$$

then we have

$$D(p\|r) = D(p\|q) + D(q\|r). \tag{14}$$

Corollary 2. *The Pythagorean formula (14) holds if $(\eta_p - \eta_q)\cdot(\theta_q - \theta_r) = 0$.*

An alternative choice for the divergence on $\mathcal{P}$ is the KL divergence,

$$KL(p\|q) = E_p\left[\log\frac{p(y,x)}{q(y,x)}\right].$$

In the case of the binary response model, the KL divergence is

$$\begin{aligned} KL(p\|q) &= E_p\left[\log\frac{yF(x\cdot\theta_p)+(1-y)(1-F(x\cdot\theta_p))}{yF(x\cdot\theta_q)+(1-y)(1-F(x\cdot\theta_q))}\right] \\ &= E\left[F(x\cdot\theta_p)\log\left(\frac{F(x\cdot\theta_p)}{F(x\cdot\theta_q)}\right)\right] \\ &\quad +E\left[(1-F(x\cdot\theta_p))\log\left(\frac{1-F(x\cdot\theta_p)}{1-F(x\cdot\theta_q)}\right)\right] \end{aligned} \tag{15}$$

because $E_p[y|x] = F(x\cdot\theta_p)$. The canonical divergence (13) and the KL divergence (15) generally do not coincide. In a special case where F is a logistic distribution, they are equivalent.

Theorem 3. *$D = KL$ holds for arbitrary p_X if and only if F is a logistic distribution; that is,*

$$F(u) = \frac{\exp(\beta u)}{1+\exp(\beta u)} \tag{16}$$

where $\beta > 0$.

Proof. If F is a logistic distribution, $\beta F(1-F) = f$ is true. This equation is substituted on the right-hand side of (13) to obtain $D = KL$.

Now, we assume that $D \equiv KL$ holds for an arbitrary p_X. Since

$$(\partial_\theta)_p(\partial_\theta)_q D(p\|q) = -E\left[\frac{f(x\cdot\theta_p)^2}{F(x\cdot\theta_p)(1-F(x\cdot\theta_p))}xx^\top\right]$$

and

$$(\partial_\theta)_p(\partial_\theta)_q KL(p\|q) = -E\left[\frac{f(x\cdot\theta_p)f(x\cdot\theta_q)}{F(x\cdot\theta_q)(1-F(x\cdot\theta_q))}xx^\top\right],$$

$D(p\|q) \equiv KL(p\|q)$ implies that

$$\frac{f(x\cdot\theta_p)^2}{F(x\cdot\theta_p)(1-F(x\cdot\theta_p))} \equiv \frac{f(x\cdot\theta_p)f(x\cdot\theta_q)}{F(x\cdot\theta_q)(1-F(x\cdot\theta_q))}$$

for arbitrary p and q. This is possible only if there exists a positive constant β such that

$$\frac{f(u)}{F(u)(1-F(u))} \equiv \beta.$$

Therefore, F is the logistic distribution. □

In the case of $\beta = 1$, the results in the previous section are largely simplified. The Fisher information metric is given by

$$g_{ij}(\theta) = E\left[f(x \cdot \theta) x_i x_j\right]$$

for $1 \leq i, j \leq d$. The $\nabla^{(-1)}$-affine coordinate η is expressed as

$$\eta_j = E\left[F(x \cdot \theta) x_j\right]$$

for $1 \leq j \leq d$. The potential is

$$\psi(\theta) = E\left[\log\left(1 + \exp(x \cdot \theta)\right)\right]. \tag{17}$$

The canonical divergence is expressed as follows:

$$D(p\|q) = E\left[\log\left(\frac{1 + \exp(x \cdot \theta_q)}{1 + \exp(x \cdot \theta_p)}\right)\right] - E\left[\frac{\exp(x \cdot \theta_p)}{1 + \exp(x \cdot \theta_p)} x\right] \cdot (\theta_q - \theta_p), \tag{18}$$

which is equal to $KL(p\|q)$.

The logit model exhibits geometrically desirable properties not only because of the explicit integrability of F. We say that a statistical model $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ is an exponential family if it is expressed as

$$p(z, \theta) = \exp\left[C(z) + \sum_{i=1}^{d} \theta^i \beta_i(z) - \psi(\theta)\right]. \tag{19}$$

It is widely known that the (curved) exponential family possesses desirable properties such as higher-order efficiency of the maximum likelihood estimation [1, 5]. Although the logit model is not truely exponential, the conditional density $p_\theta(y|x)$ is still written as

$$p_\theta(y|x) = \exp\left((x \cdot \theta)\delta_1(y) + \delta_0(y) - \psi(\theta|x)\right), \tag{20}$$

where

$$\delta_i(y) = \begin{cases} 1 & \text{if} \quad y = i \\ 0 & \text{if} \quad y \neq i \end{cases},$$

and

$$\psi(\theta|x) = \log\left(1 + \exp(x \cdot \theta)\right).$$

Conditioned by x, the model (20) belongs to an exponential family with potential $\psi(\theta|x)$. Notably, $\psi(\theta) = E\left[\psi(\theta|x)\right]$.

The marginal density p_X does not appear in the score of the model (6). Hence, p_X plays a minor role in the estimation of θ. The statistical properties of the model are primarily determined by $p_\theta(y|x)$. In fact, the following result is obtained.

Theorem 4. *Assume that the density of z conditioned on w is given by*

$$q_\theta(z|w) = \exp\left(\theta \cdot \beta(z|w) - \psi(\theta|w)\right), \tag{21}$$

where $\theta \in \Theta \in \mathcal{O}(\mathbb{R}^d)$, $\beta(z|w)$ is an $\mathbb{R}^d$-valued function of (z, w), and

$$\psi(\theta|w) := \log \int \exp\left(\theta \cdot \beta(z|w)\right) dz. \tag{22}$$

Then, the KL divergence of $\mathcal{Q} = \{q_\theta|\theta \in \Theta\}$ is equivalent to the canonical divergence D of $\mathcal{Q}$ with potential $\psi(\theta) = E\left[\psi(\theta \mid w)\right]$.

We can generalize Theorem 3 to cover the multinomial discrete choice model. Let $\{1, \cdots, k\}$ be the choice set. Assume that the choice probability conditioned on x is now given by

$$\mathbf{P}\{y = i \mid x\} = F(x \cdot \theta_i)$$

for $1 \leq i \leq k$, where F is a smooth distribution function and $\theta = \left[\theta_1 \cdots \theta_k\right] \in (\mathbb{R}^d)^k$ with $\theta_i = (\theta_i^1, \cdots, \theta_i^d) \in \mathbb{R}^d$. Let p_X be the marginal density of x, and let $\Theta \in \mathcal{O}((\mathbb{R}^d)^k)$ be the parameter set. Then, the multinomial choice model $\{\rho_\theta \mid \theta \in \Theta\}$ is given by

$$\rho_\theta(y, x) = \sum_{i=1}^{k} \delta_i(y) F(x \cdot \theta_i) p_X(x). \tag{23}$$

In particular, when F is the standard logit distribution, the model becomes the multinomial logit model with the choice probability

$$\rho_\theta(y = i|x) = \frac{\exp(x \cdot \theta_i)}{\sum_{j=1}^{k} \exp(x \cdot \theta_j)} \tag{24}$$

for $1 \leq i \leq k$. The model is a conditional exponential family because

$$\rho_\theta(y|x) = \exp\left[\sum_{i=1}^{k} \delta_i(y) x \cdot \theta_i - \psi(\theta|x)\right]$$

with conditional potential $\psi(\theta|x) = \log \sum_{j=1}^{k} \exp(x \cdot \theta_j)$. Hence, the model set $\{\rho_\theta|\theta \in \Theta\}$ is a dually flat space with dual affine coordinates (θ, η) and potential $\psi(\theta) = E\left[\log \sum_{j=1}^{k} \exp(x \cdot \theta_j)\right]$, where $\eta = \left[\eta_1 \cdots \eta_k\right] \in (\mathbb{R}^d)^k$, $\eta_i = (\eta_{i,1}, \cdots, \eta_{i,d}) \in \mathbb{R}^d$, and

$$\eta_{i,l} = E\left[\frac{\exp(x \cdot \theta_i)}{\sum_{j=1}^{k} \exp(x \cdot \theta_j)} x_l\right]$$

for $1 \leq i \leq k$ and $1 \leq l \leq d$. Furthermore, the canonical divergence D is equivalent to the KL divergence.

4 Linearly Constraint Logistic Regression

In this section, the results of the previous section are applied to the logit model with linear constraints. In empirical applications, we often want to estimate θ under the linear constraint hypothesis, $H_0 : H^\top \theta = c$, where $H = \begin{bmatrix} h_1 & \cdots & h_m \end{bmatrix}$ is an $d \times m$ matrix with $\mathrm{rank}(H) = m < d$, and $c = (c_1, \cdots, c_m) \in \mathbb{R}^m$. Let $\mathcal{P}_\mathcal{H} = \{p_\theta \in \mathcal{P} \mid \theta \in \mathcal{H}\}$, where $\mathcal{H} = \{\theta \in \Theta \mid H^\top \theta = c\}$. Suppose that the *true* model p does not belong to $\mathcal{P}(\mathcal{H})$. Then, the KL projection $\Pi : \mathcal{P} \to \mathcal{P}(\mathcal{H})$ is given by

$$\Pi p = \arg\min \; KL(p\|q) \quad \text{subject to} \quad q \in \mathcal{P}(\mathcal{H}). \tag{25}$$

In the following, we assume that F is a standard logistic distribution.

Theorem 5. *$q = \Pi p$ if and only if $\eta_q - \eta_p \in \mathrm{Image}(H)$.*

Proof. Let $\mathcal{L}(\theta, \lambda) = KL(p\|p_\theta) - \sum_{i=1}^l \lambda^i (h_i \cdot \theta - c_i)$ be the Lagrangian corresponding to (25) with Lagrange multipliers $\lambda = (\lambda^1, \cdots, \lambda^m)$. As $\theta \mapsto KL(p\|p_\theta)$ is convex, a necessary and sufficient condition for minimization is

$$\frac{\partial}{\partial \theta} KL(p\|p_\theta) = \sum_{i=1}^m \lambda^i h_i \in \mathrm{Image}(H).$$

As $KL = D$, on the other hand,

$$\frac{\partial}{\partial \theta} KL(p\|p_\theta) = \frac{\partial}{\partial \theta} \big[\varphi(\eta_p) + \psi(\theta) - \eta_p \cdot \theta\big] = \eta_q - \eta_p.$$

□

The condition $\eta_q - \eta_p \in \mathrm{Image}(H)$ is satisfied if $\lambda^1, \cdots, \lambda^m \in \mathbb{R}$ such that $\eta_q - \eta_p = \sum_{i=1}^m \lambda^i h_i$ exist. Hence, the conditions given in the theorem are written as

$$\begin{cases} \eta - \sum_{i=1}^m \lambda^i h_i = \eta_p \\ H^\top \theta = c \\ \eta = \partial\psi(\theta) \end{cases},$$

which are $2d + m$ equations with $2d + m$ variables (θ, η, λ). The solution is given by $\theta = (\partial\psi)^{-1}(\eta_p + H\lambda)$, where λ solves $H^\top (\partial\psi)^{-1}(\eta_p + H\lambda) = c$. The solution is well approximated by $\lambda = (H^\top G(\theta_p)^{-1} H)^{-1}(c - H^\top \theta_p)$ if θ_p locates sufficiently close to $\mathcal{H}$ and might be recursively updated by the standard Newton-Raphson method.

5 Conclusions

In this study, the geometry of parametric binary response models was studied. The model was demonstrated to be a dually flat space, where the canonical coefficient parameter θ acts as an affine coordinate. The dual flat property introduces

a canonical divergence into the model. The divergence is equivalent to the KL divergence if and only if the model is a logit. As an application example, the KL projection of the logit model onto an affine linear subspace was geometrically characterized.

The dual flatness of the binary response model is caused by the single-index structure of the model, which depends on the parameter θ only through the linear index $x \cdot \theta$, making the Levi-Civita connection coefficients $\Gamma_{ij,k}$ symmetrical on (i, j, k). Therefore, the results of this study can be extended to a more general class of single-index models, including nonlinear regressions, truncated regressions, and ordered discrete response models.

References

1. Amari, S.: Differential geometry of curved exponential families–curvature and information loss. Annals of Statistics **10**(2), 357–385 (1982)
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Tokyo (2000)
3. Andrews, I., Mikusheva, A.: A geometric approach to nonlinear econometric models, Econometrica **84**(3), 1249-1264 (2016)
4. Critchley, F., Marriott P., Salmon M.: On the differential geometry of the Wald test with nonlinear restrictions, Econometrica **64**(5), 1213-1222 (1996)
5. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Annals of Statistics **11**(3), 793–803 (1983)
6. Eguchi, S., Komori, O.: Minimum Divergence Methods in Statistical Machine Learning. From an Information Geometric Viewpoint. Springer Japan KK, Tokyo (2022)
7. Eguchi, S., Komori, O., Ohara, A.: Duality of maximum entropy and minimum divergence. Entropy **16**(7), 3552–3572, (2014)
8. Kemp, G.C.R.: Invariance and the Wald test. Journal of Econometrics **104**(2), 209-217 (2001)
9. Lee, M. J.: Micro-Econometrics: Methods of Moments and Limited Dependent Variables. 2nd ed. Springer, New York (2010)
10. Marriott, P., Salmon, M.: An introduction to differential geometry in econometrics. In: Applications of Differential Geometry to Econometrics, pp. 7–63. Cambridge University Press, Cambridge (2000)
11. Train, K. E.: Discrete Choice Methods with Simulations. Cambridge University Press, New York (2003)