Information Geometry
https://doi.org/10.1007/s41884-024-00136-1

RESEARCH

Dually flat structure of binary choice models

Hisatoshi Tanaka[1]

Received: 25 December 2023 / Revised: 30 April 2024 / Accepted: 1 May 2024
© The Author(s) 2024

Abstract
In this study, we consider parametric binary choice models from the perspective of information geometry. The set of models is a dually flat manifold with dual connections, naturally derived from the Fisher information metric. Under the dual connections, the canonical divergence and the Kullback–Leibler divergence of the binary choice model coincide if and only if the model is a logit model.

Keywords Discrete choice models · Logit model · Single-index models · Hessian manifolds · Maximum likelihood estimation.

1 Introduction

Consider the following simple linear regression model:

$$y = x \cdot \theta + \epsilon, \tag{1}$$

where y is a dependent variable, x is a d-dimensional random vector, ϵ is an error term, $\theta = (\theta^1, \ldots, \theta^d) \in \mathbb{R}^d$, and $x \cdot \theta = \sum_{i=1}^d x_i \theta^i$. The model seems to be very "flat" owing to its linear appearance. If we change the parameter of the model as follows:

$$\theta^i \mapsto 1/\xi_i$$

for each $i = 1, \ldots, d$, the model becomes a nonlinear regression model,

$$y = \sum_{i=1}^{d} \frac{x_i}{\xi_i} + \epsilon,$$

Communicated by Frank Nielsen.

✉ Hisatoshi Tanaka
hstnk@waseda.jp

[1] School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku, Tokyo 169-8050, Japan

Published online: 18 May 2024

which does not appear very "flat" anymore, although the nature of the model remains unchanged. This rather simple example highlights the importance of the geometric point of view in understanding the shape of statistical models: flatness of a statistical model must be defined independent of the choice of parameters, that is, by the manner of information geometry.

In econometrics, information geometry has been used to characterize the flat nature of statistical models, including the standard linear regression model, Poisson regression, Wald tests, the ARMA model, and many other examples [4, 5, 9, 11]. The objective of this study is to explore the application potential of binary choice models.

In the binary choice model, the value of dependent variable y can be 1 or 0, based on whether or not some event occurs. The standard model is represented as

$$y = \begin{cases} 1 & \text{if} \quad x \cdot \theta \geq \epsilon \\ 0 & \text{if} \quad x \cdot \theta < \epsilon, \end{cases}$$

where x is an $\mathbb{R}^d$-valued random vector distributed according to density $p_X(x)$, $\theta \in \mathbb{R}^d$, and ϵ is a random term independent of x. The choice probability is given by

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

where F is the distribute function of ϵ. The joint density function of $(y, x) \in \{0, 1\} \times \mathbb{R}^d$ is

$$p_\theta(y, x) = F(x \cdot \theta)^y (1 - F(x \cdot \theta))^{1-y} p_X(x). \tag{2}$$

The model is commonly used in social sciences to describe the choices made by decision-makers between two alternatives. These alternatives may represent school, labor supply, marital status, or transportation choices. See [10, 15] for a list of empirical applications. In particular, the model is referred to as the *probit* model when F is the standard normal distribution, and as the *logit* model when F is the standard logistic distribution:

$$F(u) = \frac{\exp u}{1 + \exp u}, \quad u \in \mathbb{R}. \tag{3}$$

The probit model is often considered a plausible model owing to its normally distributed random errors, whereas the logit model is considered merely as a closed-form approximation of the probit. Contrary to this common belief, we think the logit model is the most natural model among the parametric binary choice models from the point of view of information geometry.

The remainder of this paper is organized as follows. In Sect. 2, the geometry of the binary choice model is formulated and the model is shown to be a dually flat space. In Sect. 3, the logit model is investigated in detail as a specified case in Sect. 2. The canonical divergence and the Kullback–Leibler (KL) divergence are introduced to the model. We demonstrate that the logit model is a unique model, whose canonical and KL divergences are equal. In Sect. 4, we offer a geometric interpretation of the

maximum likelihood estimation of the binary choice model. In Sect. 5, we summarize the conclusions of this study.

2 Geometry of the binary choice model

The model set is given by

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

where Θ is an open subset of $\mathbb{R}^d$. This study is based on the following technical assumptions:

(A1) F is an infinitely differentiable function on $\mathbb{R}$ with positive derivative $f = F' > 0$;
(A2) x has a compact support $\mathcal{X} \subset \mathbb{R}^d$ such that $\mathcal{X}^{int} \neq \emptyset$.

Therefore, model $\mathcal{P}$ is considered to be a d-dimensional C^∞ manifold with a canonical coordinate system $\Theta \to \mathcal{P}, \theta \mapsto p_\theta$.

Proposition 1 *The coordinate system $\Theta \to \mathcal{P}, \theta \mapsto p_\theta$, is bijective.*

Proof Assume that there exists $\theta \neq \theta'$ such that $p_\theta = p_{\theta'}$. Then, $F(x \cdot \theta) = F(x \cdot \theta')$ holds for every $x \in \mathcal{X}$. Because **(A1)** implies that F is strictly monotone, $\mathcal{X} \subset \{x \in \mathbb{R}^d \mid x \cdot (\theta - \theta') = 0\}$, which contradicts $\mathcal{X}^{int} \neq \emptyset$. □

Unless otherwise specified, θ is used as the (global) coordinate of manifold $\mathcal{P}$ hereafter.

For every $p = p_\theta$, let E_p be the expectation operator defined as

$$\begin{aligned} E_p\beta(y,x) &= \int \beta(y,x)p(y,x)\,dydx \\ &= \int \beta(1,x)F(x\cdot\theta)p_X(x)\,dx + \int \beta(0,x)(1-F(x\cdot\theta))p_X(x)\,dx \end{aligned}$$

for an arbitrary measurable function β of (y, x). The conditional expectation operator $E_p[\cdot \mid x]$ is also defined as follows:

$$E_p[\,\beta(y,x) \mid x\,] = \int \beta(y,x)p(y|x)\,dy = \beta(1,x)F(x\cdot\theta) + \beta(0,x)(1-F(x\cdot\theta)).$$

In particular, $E_p[\,y \mid x\,] = F(x\cdot\theta)$ holds. The expectation operator with respect to x is simply denoted by E because the value of $E\beta(x) = \int_{\mathcal{X}} \beta(x)p_X(x)\,dx$ is independent of θ.

The score function of $p = p_\theta$ is

$$\frac{\partial}{\partial\theta}\log p(y,x) = \frac{y - F(x\cdot\theta)}{F(x\cdot\theta)(1-F(x\cdot\theta))} f(x\cdot\theta)x. \tag{4}$$

Because

$$E_p[\,(y - F(x\cdot\theta))^2 \mid x\,] = E_p[\,y \mid x\,] - F(x\cdot\theta)^2 = F(x\cdot\theta)(1 - F(x\cdot\theta)),$$

the Fisher information matrix $G(\theta)$ is given as

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

For simplicity, define $r : \mathbb{R} \to \mathbb{R}_{++}$ as

$$r(u) = \frac{f(u)^2}{F(u)(1 - F(u))} \tag{5}$$

for every $u \in \mathbb{R}$ such that $G(\theta) = E\left[r(x\cdot\theta)xx^\top\right]$. By the assumptions, r is bounded on an arbitrary compact interval. Because x has a bounded support, $G(\theta)$ is finite at every $\theta \in \Theta$. In addition, we assume that

(A3) $G(\theta)$ is positive definite at every $\theta \in \Theta$.

The tangent space of $\mathcal{P}$ at $p = p_\theta$ is $T_p\mathcal{P} = \mathrm{Span}\left\{(\partial_1)_p, \ldots, (\partial_d)_p\right\}$, where $\partial_i = \frac{\partial}{\partial\theta_i}$ for $i = 1, \ldots, d$. For example, the unconditional expectation $E_p y$ is obtained as

$$E_p y = E\Big(E_\theta[\,y \mid x\,]\Big) = \int_{\mathcal{X}} F(x\cdot\theta)p_X(x)\,dx,$$

which is a smooth function on $\mathcal{P}$. A tangent vector $X = X^i(\partial_i)_p$ operates on this as

$$\begin{aligned} X(E_p y) &= X^i(\partial_i)_p \int_{\mathcal{X}} F(x\cdot\theta)p_X(x)\,dx \\ &= X^i \int_{\mathcal{X}} x_i f(x\cdot\theta)p_X(x)\,dx = \sum_{i=1}^d X^i E\left[f(x\cdot\theta)x_i\right]. \end{aligned}$$

Moreover, at every (y, x),

$$\begin{aligned} X(\log p(y,x)) &= X^i\,(\partial_i)_p\,(y\log F(x\cdot\theta) + (1-y)\log(1 - F(x\cdot\theta)) + \log p_X(x)) \\ &= X^i\left(\frac{y - F(x\cdot\theta)}{F(x\cdot\theta)(1 - F(x\cdot\theta))}x_i\right). \end{aligned}$$

The Fisher information metric g on $\mathcal{P}$ is introduced as

$$g_p(X, Y) = E_p\,(X\log p(y,x))\,(Y\log p(y,x)) = X^i Y^j g_{ij}(p),$$

where $g_{ij}(p) = E\left[r(x\cdot\theta)x_i x_j\right]$ is the (i, j) element of $G(\theta)$.

Given metric g, the binary choice model is considered to be a Riemannian manifold $(\mathcal{P}, g)$. Moreover, function $\psi : \Theta \to \mathbb{R}$ is defined as

$$\psi(\theta) = E\left[\int_0^{x\cdot\theta}\left(\int_0^{v} r(u)\,du\right)dv\right]. \tag{6}$$

Then,

$$\partial_i\partial_j\psi(\theta) = \frac{\partial}{\partial\theta^i}E\left[\left(\int_0^{x\cdot\theta} r(u)\,du\right)x_j\right] = E\left[r(x\cdot\theta)x_i x_j\right] = g_{ij}(p).$$

Hence, $(\mathcal{P}, g)$ is a Hessian manifold with potential ψ when it is equipped with flat connections [12, 14]. For later convenience, we introduce gradient $\partial\psi : \Theta \to \mathbb{R}^d$ as

$$\partial\psi(\theta) = \begin{bmatrix}\partial_1\psi(\theta)\\ \vdots \\ \partial_d\psi(\theta)\end{bmatrix} = E\left[\left(\int_0^{x\cdot\theta} r(u)\,du\right)x\right],$$

where $\partial_i\psi(\theta) = E\left[\left(\int_0^{x\cdot\theta} r(u)\,du\right)x_i\right]$ for $i = 1, \ldots, d$. Because the Hessian $\partial^2\psi(\theta) = G(\theta)$ is positive definite by **(A3)**, an inverse mapping $(\partial\psi)^{-1} : \partial\psi(\Theta) \subset \mathbb{R}^d \to \mathbb{R}^d$ exists and is continuously differentiable at every point.

Let $\mathfrak{X}(\mathcal{P})$ denote the class of C^∞ tangent vector fields on $\mathcal{P}$. For $X = X^i\partial_i$, $Y = Y^j\partial_j \in \mathfrak{X}(\mathcal{P})$, the Levi-Civita connection ∇ of $(\mathcal{P}, g)$ is introduced as

$$\nabla_X Y = X^i(\partial_i Y^j)\partial_j + X^i Y^j \Gamma_{ij}^k \partial_k, \tag{7}$$

where Γ_{ij}^k is the Christoffel symbol:

$$\begin{aligned}\Gamma_{ij}^k(\theta) &= \frac{1}{2}\Big[\partial_i g_{lj}(\theta) + \partial_j g_{il}(\theta) - \partial_l g_{ij}(\theta)\Big]g^{kl}(\theta)\\ &= \frac{1}{2}E\left[r'(x\cdot\theta)x_i x_j x_l\right]g^{kl}(\theta)\end{aligned} \tag{8}$$

for $i, j, k \in \{1, \ldots, d\}$, where g^{kl} denotes the (k, l)-element of $G(\theta)^{-1}$. Let $\Gamma_{ijk}(\theta) = \Gamma_{ij}^l(\theta)g_{kl}(\theta) = \frac{1}{2}E\left[r'(x\cdot\theta)x_i x_j x_k\right]$ such that

$$\Gamma_{ijk} = \frac{1}{2}\partial_i g_{jk} = \frac{1}{2}\partial_j g_{ki} = \frac{1}{2}\partial_k g_{ij}. \tag{9}$$

The curvature and torsion tensors $R : \mathfrak{X}(\mathcal{P}) \times \mathfrak{X}(\mathcal{P}) \times \mathfrak{X}(\mathcal{P}) \to \mathfrak{X}(\mathcal{P})$ and $T : \mathfrak{X}(\mathcal{P}) \times \mathfrak{X}(\mathcal{P}) \to \mathfrak{X}(\mathcal{P})$ of ∇ are, respectively, defined as

$$R(X, Y, Z) = \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z\nabla) - \nabla_{[X,Y]}Z \tag{10}$$

and

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y], \tag{11}$$

where $[X, Y] = X^i(\partial_i Y^j)\partial_j - Y^j(\partial_j X^i)\partial_i$.

Proposition 2 *Let ∇ be the Levi-Civita connection (7) with coefficients (8). Then,*

$$R_{ijk} := R(\partial_i, \partial_j, \partial_k) = \left(\Gamma^m_{ik}\Gamma^l_{jm} - \Gamma^m_{jk}\Gamma^l_{im}\right)\partial_l \tag{12}$$

for $i, j, k \in \{1, \ldots, d\}$ and $T \equiv 0$.

Proof Using (8), $T(\partial_i, \partial_j) = (\Gamma^k_{ij} - \Gamma^k_{ji})\partial_k = 0$ is trivially shown. Because $g_{mh}g^{hl} = 1$ if $m = l$ and 0 if $m \neq l$,

$$\partial_i(g_{mh}g^{hl}) = (\partial_i g_{mh})g^{hl} + g_{mh}(\partial_i g^{hl}) = 2\Gamma^l_{im} + g_{mh}(\partial_i g^{hl}) = 0,$$

which implies $\partial_i g^{hl} = -2\Gamma^l_{im}g^{mh}$. Using the definition of the curvature tensor,

$$\begin{aligned} R_{ijk} &= \nabla_{\partial_i}(\Gamma^l_{jk}\partial_l) - \nabla_{\partial_j}(\Gamma^l_{ik}\partial_l) \\ &= \left\{\partial_i(\Gamma_{jkh}g^{hl})\partial_l + \Gamma^l_{jk}\Gamma^h_{il}\partial_h\right\} - \left\{\partial_j(\Gamma_{ikh}g^{hl})\partial_l + \Gamma^l_{ik}\Gamma^h_{jl}\partial_h\right\} \\ &= \left(\partial_i\Gamma_{jkh} - \partial_j\Gamma_{ikh}\right)g^{hl}\partial_l \\ &\quad - 2\left(\Gamma_{jkh} - \Gamma_{ikh}\right)\Gamma^l_{im}g^{mh} + \left(\Gamma^l_{im}\Gamma^m_{jk} - \Gamma^l_{jm}\Gamma^m_{ik}\right)\partial_l. \end{aligned}$$

Because $\partial_i\Gamma_{jkh} = \partial_j\Gamma_{ikh} = \frac{1}{2}E\left[r''(x\cdot\theta)x_i x_j x_k x_h\right]$, Eq. (12) is obtained. □

Proposition 2 implies that the binary choice model with the Fisher information geometry is essentially a flat manifold. Let S be an arbitrary symmetric $(0, 3)$-tensor on $\mathcal{P}$. A family of α-connections $\{\nabla^{(\alpha)}\}_{\alpha\in\mathbb{R}}$ is defined as

$$g(\nabla^{(\alpha)}_X Y, Z) = g(\nabla_X Y, Z) - \alpha S(X, Y, Z)$$

for every $\alpha \in \mathbb{R}$. The corresponding connection coefficients are given by

$$\Gamma^{(\alpha)}_{ijk} = \Gamma_{ijk} - \alpha S_{ijk},$$

where $S_{ijk} = S(\partial_i, \partial_j, \partial_k)$. See chapter 6 of [2] for definitions and details of α-connections.

A pair $(\nabla^{(\alpha)}, \nabla^{(-\alpha)})$ of the connections provides the dual connections of $(\mathcal{P}, g)$ such that

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

because

$$\begin{aligned} X(g(Y,Z)) &= g(\nabla_X Y, Z) + g(Y, \nabla_X Z) \\ &= \{g(\nabla_X Y, Z) - \alpha S(X,Y,Z)\} + \{g(Y, \nabla_X Z) - (-\alpha S(Y,X,Z))\} \\ &= g(\nabla_X^{(\alpha)} Y, Z) + g(Y, \nabla_X^{(-\alpha)} Z) \end{aligned}$$

by symmetry $S(X,Y,Z) = S(Y,X,Z)$. Let $R^{(\alpha)}$ and $T^{(\alpha)}$ be the curvature and torsion tensors of $\nabla^{(\alpha)}$, respectively. When $R^{(\alpha)} = R^{(-\alpha)} = 0$ and $T^{(\alpha)} = T^{(-\alpha)} = 0$ hold, $(\mathcal{P}, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$ is said to be a dually flat space. A dually flat space has the dual affine coordinates (ξ, ζ), where $\xi = (\xi^1, \ldots, \xi^d)$ is the $\nabla^{(\alpha)}$-affine coordinate such that $\Gamma_{ijk}^{(\alpha)}(\xi) \equiv 0$ and $\zeta = (\zeta_1, \ldots, \zeta_d)$ is the $\nabla^{(-\alpha)}$-affine coordinate such that $\Gamma_{ijk}^{-\alpha}(\zeta) \equiv 0$. Furthermore,

$$g\left(\frac{\partial}{\partial \xi^i}, \frac{\partial}{\partial \zeta_j}\right) = \delta_i^j = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j). \end{cases}$$

Theorem 1 *Let S be a $(0,3)$-tensor on $(\mathcal{P}, g)$ given by $S(X,Y,Z) = X^i Y^j Z^k S_{ijk}$ with*

$$S_{ijk}(p) = \Gamma_{ijk}(\theta) = \frac{1}{2} E_\theta \left[r'(x \cdot \theta) x_i x_j x_k \right].$$

For $\alpha = \pm 1$, let $\nabla^{(\alpha)}$ be defined as

$$g(\nabla_X^{(\alpha)} Y, Z) = g(\nabla_X Y, Z) - \alpha S(X,Y,Z).$$

Then, $(\mathcal{P}, g, \nabla^{(+1)}, \nabla^{(-1)})$ is a dually flat space with dual affine coordinates (θ, η), where $\eta = \partial \psi(\theta)$.

Proof While the results can be immediately obtained using the general theory of the Hessian manifolds [12, 14], we give direct proof for a self-contained description of the paper.

Based on the assumption, $\Gamma_{ij}^{(+1)k}(\theta) \equiv 0$ trivially holds. To confirm that $\eta = \partial \psi(\theta)$ is the $\nabla^{(-1)}$-affine coordinate, let $\Gamma_{ab}^{(-1)c}(\eta)$ for $a, b, c \in \{1, \ldots, d\}$ be the $\nabla^{(-1)}$-connection coefficients expressed in terms of η. By the definition of η, $\partial_k \eta_l = g_{kl}$ holds. This implies that $\frac{\partial \theta^k}{\partial \eta_l} = g^{kl}$ and that

$$\frac{\partial^2 \eta_l}{\partial \theta^i \partial \theta^j} \frac{\partial \theta^k}{\partial \eta_l} = E\left[r'(x \cdot \theta) x_i x_j x_l \right] g^{kl} = 2\Gamma_{ij}^k(\theta).$$

From the definition of $\nabla^{(-1)}$, $\Gamma_{ij}^{(-1)k}(\theta) = 2\Gamma_{ij}^k(\theta)$. By the change-of-variables formula for the connection coefficients,

$$\begin{aligned}\Gamma_{ij}^{(-1)k}(\theta) &= \frac{\partial^2 \eta_l}{\partial\theta^i \partial\theta^j}\frac{\partial\theta^k}{\partial\eta_l} + \frac{\partial\eta_a}{\partial\theta^i}\frac{\partial\eta_b}{\partial\theta^j}\Gamma_{ab}^{(-1)c}(\eta)\frac{\partial\theta^k}{\partial\eta_c} \\ &= 2\Gamma_{ij}^{k}(\theta) + g_{ai}g_{bj}\Gamma_{ab}^{(-1)c}(\eta)g^{kc},\end{aligned}$$

which implies

$$\Gamma_{ab}^{(-1)c}(\eta) = g^{ai}g^{bk}\left(\Gamma_{ij}^{(-1)k}(\theta) - 2\Gamma_{ij}^{k}(\theta)\right)g_{kc} = 0.$$

Moreover,

$$g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\eta_j}\right) = \frac{\partial\theta^k}{\partial\eta_j}g\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^k}\right) = g^{jk}g_{ik} = \delta_i^j.$$

□

Corollary 1 *For $\alpha = \pm 1$, the $\nabla^{(\alpha)}$-geodesic path $\gamma^{(\alpha)} = \{\gamma_t^{(\alpha)} | \, 0 \le t \le 1\}$ connecting $p, q \in \mathcal{P}$ is given by*

$$\gamma_t^{(\alpha)}(y, x) = F(x \cdot \theta_t^{(\alpha)})^y (1 - F(x \cdot \theta_t^{(\alpha)}))^{1-y} p_X(x),$$

where

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

and

$$\theta_t^{(-1)} = (\partial\psi)^{-1}((1-t)\eta_p + t\eta_q). \tag{13}$$

In particular, (13) is a solution to the ordinary differential equation,

$$\frac{d}{dt}\theta_t^{(-1)} = G\left(\theta_t^{(-1)}\right)^{-1}(\eta_q - \eta_p),$$

with initial condition $\theta_0 = \theta_p$.

Proof Let $\eta_t^{(-1)} = \partial\psi(\theta_t^{(-1)}) = (1-t)\eta_p + t\eta_q$. Then,

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

□

3 Two divergences of the binary choice model

The dual potential φ of ψ is given as

$$\varphi(\eta) = \max_{\theta} \ \eta \cdot \theta - \psi(\theta),$$

which is the Legendre transformation of $\psi(\theta)$. Because $\theta \mapsto \eta \cdot \theta - \psi(\theta)$ is strictly concave by **(A3)**, the maximum of $\eta \cdot \theta - \psi(\theta)$ is attained at $\theta = (\partial\psi)^{-1}(\eta)$, which is a solution to the first-order condition, $\eta - \partial\psi(\theta) = 0$. Let θ_p and η_p denote the canonical coordinate and its dual at $p \in \mathcal{P}$, respectively. Then, the dual potential is explicitly given as

$$\begin{aligned}\varphi(\eta_p) &= \eta_p \cdot \theta_p - \psi(\theta_p) \\ &= E\left[\left(\int_0^{x\cdot\theta_p} r(u)\,du\right)x\right]\cdot\theta_p - E\left[\int_0^{x\cdot\theta_p}\left(\int_0^{v} r(u)\,du\right)dv\right].\end{aligned}$$

In general, for a dually flat space with dual affine coordinates (θ, η) and dual potentials (ψ, φ), the *canonical divergence* between p and q in $\mathcal{P}$ is defined as follows [3, 7, 8]:

$$D(p \,\|\, q) = \varphi(\eta_p) + \psi(\theta_q) - \eta_p \cdot \theta_q. \tag{14}$$

For binary choice model (2),

$$\begin{aligned}D(p \,\|\, q) &= \left\{E\left[\left(\int_0^{x\cdot\theta_p} r(u)\,du\right)x\right]\cdot\theta_p - E\left[\int_0^{x\cdot\theta_p}\left(\int_0^{v} r(u)\,du\right)dv\right]\right\} \\ &\quad + E\left[\int_0^{x\cdot\theta_q}\left(\int_0^{v} r(u)du\right)dv\right] - E\left[\left(\int_0^{x\cdot\theta_p} r(u)du\right)x\right]\cdot\theta_q \\ &= E\left[\int_{x\cdot\theta_p}^{x\cdot\theta_q}\left(\int_{x\cdot\theta_p}^{v} r(u)du\right)dv\right].\end{aligned} \tag{15}$$

For the given p, a function $\theta \mapsto D(p \,\|\, p_\theta)$ is strictly convex because a direct computation shows

$$\partial_i\partial_j D(p \,\|\, p_\theta) = \frac{\partial^2}{\partial\theta^i\partial\theta^j}\Big(\varphi(\eta_p) + \psi(\theta) - \eta_p\cdot\theta\Big) = g_{ij}(\theta).$$

Theorem 2 *Let p, q, $r \in \mathcal{P}$. Let $\gamma^{(+1)} = (\gamma_t^{(+1)})_{0\le t\le 1}$ be the $\nabla^{(+1)}$-geodesic path connecting p and q, and let $\gamma^{(-1)} = (\gamma_s^{(-1)})_{0\le s\le 1}$ be the $\nabla^{(-1)}$-geodesic path connecting q and r. If and only if $\gamma^{(+1)}$ and $\gamma^{(-1)}$ are orthogonal at the intersection q in the sense that*

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

we have

$$D(p \,\|\, r) = D(p \,\|\, q) + D(q \,\|\, r). \tag{16}$$

Proof The result is standard. See e.g., [3, 7, 13] for a proof. □

Corollary 2 *The Pythagorean formula* (16) *holds if and only if* $(\eta_p - \eta_q) \cdot (\theta_q - \theta_r) = 0$.

Proof From corollary 1,

$$\left(\frac{d}{dt}\right)_q \gamma_t^{(+1)} = (\theta_q^i - \theta_p^i)\left(\frac{\partial}{\partial \theta^i}\right)_q$$

and

$$\left(\frac{d}{ds}\right)_q \gamma_s^{(-1)} = g^{jk}(q)((\eta_r)_k - (\eta_q)_k)\left(\frac{\partial}{\partial \theta^j}\right)_q.$$

Therefore,

$$\begin{aligned} g_q\left(\left(\frac{d}{dt}\right)_q \gamma_t^{(+1)}, \left(\frac{d}{ds}\right)_q \gamma_s^{(-1)}\right) &= g_{ij}(q)(\theta_q^i - \theta_p^i) g^{jk}(q)((\eta_r)_k - (\eta_q)_k) \\ &= (\theta_p - \theta_q) \cdot (\eta_q - \eta_r). \end{aligned}$$

□

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

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

For the binary choice model,

$$\begin{aligned} KL(p \,\|\, q) &= E_p\left[y \log \frac{F(x \cdot \theta_p)}{F(x \cdot \theta_q)} + (1 - y) \log \frac{1 - F(x \cdot \theta_p)}{1 - F(x \cdot \theta_q)}\right] \\ &= E\left[F(x \cdot \theta_p) \log\left(\frac{F(x \cdot \theta_p)}{F(x \cdot \theta_q)}\right)\right] \\ &\quad + E\left[(1 - F(x \cdot \theta_p)) \log\left(\frac{1 - F(x \cdot \theta_p)}{1 - F(x \cdot \theta_q)}\right)\right] \end{aligned} \tag{17}$$

based on the law of iterated expectations. Canonical divergence (15) and KL divergence (17) generally do not coincide. However, in a special case where F is a logistic distribution, they are shown as equivalent.

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

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

where $\beta > 0$.

Proof If F is a logistic distribution with parameter $\beta > 0$,

$$\beta \int F(u)\, du = \log(1 + \exp(\beta u)) + C$$

and

$$f(u) = \frac{d}{du}\left(\frac{\exp(\beta u)}{1 + \exp(\beta u)}\right) = \beta F(u)(1 - F(u)).$$

Hence, $r(u) = \beta f(u)$, $\int_{x\cdot\theta_p}^{v} r(u)\, du = \beta(F(v) - F(x \cdot \theta_p))$, and

$$\begin{aligned}
D(p \,\|\, q) &= E\left[\int_{x\cdot\theta_p}^{x\cdot\theta_q} \beta\left(F(v) - F(x \cdot \theta_p)\right) dv\right] \\
&= E\left[\log\left(\frac{1 + \exp(\beta x \cdot \theta_q)}{1 + \exp(\beta x \cdot \theta_p)}\right)\right] - E\left[F(x \cdot \theta_p) \log\left(\frac{\exp(\beta x \cdot \theta_q)}{\exp(\beta x \cdot \theta_p)}\right)\right] \\
&= E\left[(1 - F(x \cdot \theta_p)) \log\left(\frac{1 - F(x \cdot \theta_p)}{1 - F(x \cdot \theta_q)}\right)\right] \\
&\quad -E\left[F(x \cdot \theta_p) \log\left(\frac{F(x \cdot \theta_q)}{F(x \cdot \theta_p)}\right)\right] \\
&= KL(p \,\|\, q).
\end{aligned}$$

On the other hand, if $D \equiv KL$ holds for an arbitrary p_X,

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

holds for arbitrary p and q because

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

and

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

From the principle of the separation of variables, 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. □

For the standard logit model ($\beta = 1$), the results presented in the preceding section are further simplified. The Fisher information metric is given as

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

The $\nabla^{(-1)}$-affine coordinate η is expressed as

$$\eta = E\left[F(x \cdot \theta)x\right] = E_p[yx].$$

The potential is $\psi(\theta) = E\left[\log\left(1 + \exp(x \cdot \theta)\right)\right]$, and the divergence is

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

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

$$\mathbf{P}\{y = i \mid x\} = \frac{F(x \cdot \theta_i)}{\sum_{j=1}^{k} F(x \cdot \theta_j)}$$

for $i \in \{1, \ldots, k\}$, where F is a smooth distribution function and $\theta = \left[\theta_1 \cdots \theta_k\right] \in (\mathbb{R}^d)^k$ with $\theta_i = (\theta_i^1, \ldots, \theta_i^d) \in \mathbb{R}^d$. Let p_X be the marginal density of x and $\Theta \subset (\mathbb{R}^d)^k$ be the set of parameters. Then, the multinomial choice model is obtained as

$$p_\theta(y, x) = \frac{\sum_{i=1}^{k} \delta_i(y) F(x \cdot \theta_i) p_X(x)}{\sum_{j=1}^{k} F(x \cdot \theta_j)}.$$

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

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

for $i \in \{1, \ldots, k\}$. The model set $\{p_\theta | \theta \in \Theta\}$ is a dually flat space with the dual affine coordinates (θ, η) and the potential

$$\psi(\theta) = E\left[\log \sum_{j=1}^{k} \exp(x \cdot \theta_j)\right],$$

where $\eta = \left[\eta_1 \cdots \eta_k\right] \in (\mathbb{R}^d)^k$, $\eta_i = (\eta_{i,1}, \ldots, \eta_{i,d}) \in \mathbb{R}^d$, 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 $i \in \{1, \ldots, k\}$ and $l \in \{1, \ldots, d\}$. Furthermore, canonical D and KL divergences of the model are equal.

4 Maximum likelihood estimation of the binary choice model

Most of the results presented in Sects. 2 and 3 are independent of the choice of p_X. Therefore, by replacing p_X with its estimates based on empirical data, we might obtain some geometric view of the statistical inference of the model. Let $\mathcal{P}$ be the set of the binary choice model (2). Let $(y_1, x_1), \ldots, (y_T, x_T)$ be i.i.d. sample from $p = p_\theta \in \mathcal{P}$. Then the *empirical* expectation operator $\hat{E}$ is given by

$$\hat{E}\beta(y, x) = \frac{1}{T} \sum_{t=1}^{T} \beta(y_t, x_t).$$

The empirical Fisher information matrix is given by $\hat{G}(\theta) = \hat{E}[r(x \cdot \theta) x x^\top]$. Again, we assume that

(A3') $\hat{G}(\theta)$ is positive definite.

The empirical versions of the Fisher information metric $\hat{g}_{ij} = \hat{E}[r(x \cdot \theta) x_i x_j]$, the potential $\hat{\psi}(\theta)$, the Levi-Civita connection $\hat{\nabla}$ and the connection coefficients $\hat{\Gamma}^k_{ij}$ are also introduced simply by replacing E with $\hat{E}$.

The "true" parameter θ of $p = p_\theta$ is well approximated by the maximum likelihood estimator,

$$\hat{\theta} = \arg\max_\theta \ \hat{E} \log p_\theta(y, x), \tag{18}$$

which is an empirical analog of the KL divergence minimization,

$$\theta = \arg\min_\theta \ KL(p \parallel p_\theta) = \arg\max_\theta \ E_p \log p_\theta$$

(see e.g., [7, 10]). The estimator is a solution to the first-order conditions of maximization (18):

$$\frac{\partial}{\partial\theta}\hat{E}\log p_\theta(y,x)=\frac{1}{T}\sum_{t=1}^{T}\frac{y_t-F(x_t\cdot\theta)}{F(x_t\cdot\theta)(1-F(x_t\cdot\theta))}f(x_t\cdot\theta)x_t=0.$$

In particular, when the logit model is considered, the condition is simplified to

$$\frac{1}{T}\sum_{t=1}^{T}\big(y_t-F(x_t\cdot\theta)\big)x_t=0,$$

which implies $\hat{E}[yx]=\hat{E}\left[F(x\cdot\hat{\theta})x\right]=(\partial\hat{\psi})(\hat{\theta})$. Therefore, in the maximum likelihood estimation of the logit model, we first estimate the dual parameter η directly as $\hat{\eta}=\hat{E}[yx]=\frac{1}{T}\sum_{t=1}^{T}y_tx_t$, and secondly we estimate the canonical parameter θ using $\hat{\theta}=(\partial\hat{\psi})^{-1}(\hat{\eta})$. This method is easily implemented because it does not involve numerical optimizations.

One objective of empirical studies of an econometric model is to test the statistical significance of its coefficients $\theta^1,\dots,\theta^d$. When we want to test a joint null hypothesis such as, say, $H_0:\theta^1=\theta^2=0$, where $d\geq 2$ is assumed, the value of θ is estimated subject to the linear constraint:

$$\begin{bmatrix}1&0&0&\cdots&0\\0&1&0&\cdots&0\end{bmatrix}\begin{bmatrix}\theta^1\\\vdots\\\theta^d\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}.$$

The restriction is generalized to the case of $H_0:H^\top\theta=c$, where $H=\begin{bmatrix}h_1&\cdots&h_m\end{bmatrix}$ is a $d\times m$ matrix with $\mathrm{rank}(H)=m<d$, and $c=(c_1,\dots,c_m)^\top\in\mathbb{R}^m$. Let $\mathcal{H}=\{\theta\in\Theta\mid H^\top\theta=c\}$ be the constraint set, and let $\mathcal{P}_\mathcal{H}=\{p_\theta\in\mathcal{P}\mid\theta\in\mathcal{H}\}$ be the constrained model. Because θ is the $\nabla^{(+1)}$-affine coordinate of $\mathcal{P}$, $\mathcal{P}_\mathcal{H}$ is an affine-flat submanifold of $\mathcal{P}$.

If the logit model is assumed, the constrained maximum likelihood estimator

$$\hat{\theta}|_\mathcal{H}=\arg\max_\theta\ \hat{E}\log p_\theta(y,x)\quad\text{subject to}\quad\theta\in\mathcal{H}$$

is found by orthogonally projecting the unconstrained estimator $\hat{\theta}$ onto the constraint set. We define the D-projection operator $\Pi:\mathcal{P}\to\mathcal{P}_\mathcal{H}$ by

$$\Pi p=\arg\min_q\ D(p\,\|\,q)\quad\text{subject to}\quad q\in\mathcal{P}_\mathcal{H}\tag{19}$$

for every $p\in\mathcal{P}$, where D is the canonical divergence (14). The operator is well defined because $D(p\,\|\,q)$ is strictly convex in θ_p.

Theorem 4 *$q=\Pi p$ holds for q, $p\in\mathcal{P}$ if and only if $\eta_q-\eta_p\in\mathrm{Image}(H)$.*

Proof Let $\mathcal{L}(\theta, \lambda) = D(p \parallel p_\theta) - \sum_{i=1}^m \lambda^i (h_i \cdot \theta - c_i)$ be the Lagrangian with multipliers $\lambda = (\lambda^1, \ldots, \lambda^m)$. As $\theta \mapsto D(p \parallel p_\theta)$ is strictly convex, a necessary and sufficient condition for minimization is given by

$$\frac{\partial}{\partial \theta} \mathcal{L}(\theta, \lambda) = \frac{\partial}{\partial \theta} D(p \parallel p_\theta) - \sum_{i=1}^m \lambda^i h_i = 0,$$

which implies

$$\frac{\partial}{\partial \theta} D(p \parallel p_\theta) = \frac{\partial}{\partial \theta} \big(\varphi(\eta_p) + \psi(\theta) - \eta_p \cdot \theta \big) = \eta - \eta_p \in \mathrm{Image}(H)$$

at a solution θ to (19). Therefore, $\eta_q - \eta_p \in \mathrm{Image}(H)$ is satisfied if and only if q solves (19). □

The empirical version of theorem 4 offers us graphical images of the maximum likelihood estimation. Let $\hat{D}$ be the empirical version of the canonical divergence, and $\hat{\Pi}$ be the $\hat{D}$-projection operator. Then, if and only if the logit model is assumed, the $\hat{D}$-projection becomes equivalent to the constraint maximum likelihood estimation; that is,

$$p_{\hat{\theta}|_{\mathcal{H}}} = \hat{\Pi} p_{\hat{\theta}}.$$

This is because

$$\begin{aligned}
\hat{\Pi} p_{\hat{\theta}} &= \arg\min_q \ \hat{D}(p_{\hat{\theta}} \parallel q) \quad \text{subject to} \quad q \in \mathcal{P}_{\mathcal{H}} \\
&= \arg\min_{p_\theta} \ \hat{E}\left[\frac{\log p_{\hat{\theta}}(y, x)}{\log p_\theta(y, x)} \right] \quad \text{subject to} \quad \theta \in \mathcal{H} \\
&= \arg\max_{p_\theta} \ \hat{E} \log p_\theta(y, x) \quad \text{subject to} \quad H^\top \theta = c \\
&= p_{\hat{\theta}|_{\mathcal{H}}}.
\end{aligned}$$

For the dual parameter $\hat{\eta}|_{\mathcal{H}} = \partial \hat{\psi}(\hat{\theta}|_{\mathcal{H}})$, the condition

$$\hat{\eta}|_{\mathcal{H}} - \hat{\eta} \in \mathrm{Image}(H)$$

holds if and only if the model is logit. Furthermore, if $\hat{\eta}|_{\mathcal{H}}$ satisfies the condition, there exists $\lambda \in \mathbb{R}^m$ such that $\hat{\eta} - \hat{\eta}|_{\mathcal{H}} = H\lambda$. Therefore, for any $\theta \in \mathcal{H}$,

$$(\hat{\eta} - \hat{\eta}|_{\mathcal{H}}) \cdot (\theta - \hat{\theta}|_{\mathcal{H}}) = \lambda^\top H^\top (\theta - \hat{\theta}|_{\mathcal{H}}) = \lambda^\top (c - c) = 0$$

is satisfied. The situation is shown in Fig. 1. The figure seems obvious, but it is to be remarked that this naive image of the orthogonal projection is consistent with the estimation with linear restrictions if and only if the logit model is assumed; in other models such as the probit model, the orthogonal projection with respect to the Fisher information metric fails to maximize the likelihood on the affine linear submodel.

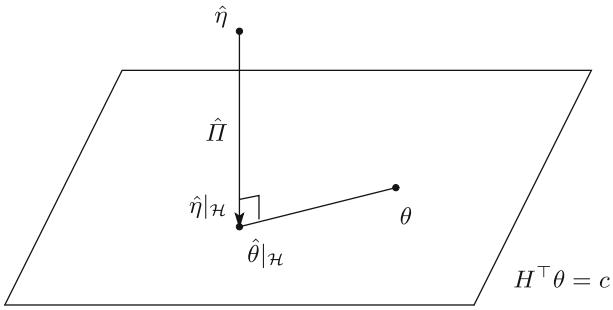

Fig. 1 Orthogonal projection $\hat{\Pi} : \hat{\eta} \mapsto \hat{\eta}|_{\mathcal{H}}$ and maximum likelihood estimator $\hat{\theta}|_{\mathcal{H}}$ under null hypothesis $H^\top\theta = c$

5 Discussion

In this study, we investigated the geometry of parametric binary choice models. The model was established as 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 model. As an example application, the projection onto an affine linear subspace was geometrically characterized.

The dual flatness of the binary choice model is caused by the single-index structure of the model, which depends on parameter θ only through linear index $x \cdot \theta$, making the Levi-Civita connection coefficients Γ_{ijk} 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 choice models. The studied model might be extended to the neural network model, which consists of connected binary response models. Each node of the network is considered as a single-index model. However, the entire structure of the model could be highly nonlinear in terms of parameter θ, which leads to the non-flatness of the model.

Among the binary choice models, the logit is shown to have good properties geometrically as well as statistically. This is not because of only the explicit integrability of the logit. In general, we say that a statistical model $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ is an exponential family if it is expressed as

$$p_\theta(z) = \exp\left[C(z) + \sum_{i=1}^{d} \theta^i \beta_i(z) - \psi(\theta)\right]. \tag{20}$$

It is widely known that the (curved) exponential family possesses desirable properties such as higher-order efficiency of the maximum likelihood estimation [1, 6]. Although the logit model is not truly exponential, the conditional density $p_\theta(y|x)$ is still written as

$$p_\theta(y|x) = \exp\left((x \cdot \theta)\delta_1(y) + \delta_0(y) - \psi(\theta|x)\right), \tag{21}$$

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, model (21) belongs to an exponential family with potential $\psi(\theta|x)$. Notably, $\psi(\theta) = E\left[\psi(\theta|x)\right]$. Because marginal density p_X does not appear in the score of model (4), statistical properties of the model are primarily determined by $p_\theta(y|x)$. Our study suggests that the logit model is a unique binary choice model, which belongs to the conditional exponential family.

Acknowledgements The author thanks to the anonymous reviewer who provided valuable comments.

Author Contributions The paper is written by a single author.

Data Availability No datasets were generated or used in this study.

Declarations

Conflict of interest The author states that there is no conflict of interest to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Amari, S.: Differential geometry of curved exponential families-curvature and information loss. Ann. Stat. **10**(2), 357–385 (1982)
2. Amari, S.: Information Geometry and Its Applications. Springer Japan KK, Tokyo (2016)
3. Amari, S., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Tokyo (2000)
4. Andrews, I., Mikusheva, A.: A geometric approach to nonlinear econometric models. Econometrica **84**(3), 1249–1264 (2016)
5. Critchley, F., Marriott, P., Salmon, M.: On the differential geometry of the Wald test with nonlinear restrictions. Econometrica **64**(5), 1213–1222 (1996)
6. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**(3), 793–803 (1983)
7. Eguchi, S., Komori, O.: Minimum Divergence Methods in Statistical Machine Learning. From an Information Geometric Viewpoint. Springer Japan KK, Tokyo (2022)
8. Eguchi, S., Komori, O., Ohara, A.: Duality of maximum entropy and minimum divergence. Entropy **16**(7), 3552–3572 (2014)

9. Kemp, G.C.R.: Invariance and the Wald test. J. Econom. **104**(2), 209–217 (2001)
10. Lee, M.J.: Micro-Econometrics: Methods of Moments and Limited Dependent Variables, 2nd edn. Springer, New York (2010)
11. 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)
12. Nakajima, N., Ohmoto, T.: The dually flat structure for singular models. Inf. Geom. **4**, 31–64 (2021)
13. Nielsen, F.: On geodesic triangles with right angles in a dually flat space. In: Progress in Information Geometry, pp. 153–190. Springer Nature Switzerland AG, Cham (2021)
14. Shima, H., Yagi, K.: Geometry of Hessian manifolds. Differ. Geom. Appl. **7**, 277–290 (1997)
15. Train, K.E.: Discrete Choice Methods with Simulations. Cambridge University Press, New York (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.