Efficiency of the Method of Generalized Moments from the Viewpoint of Information Geometry

Hisatoshi Tanaka

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

Abstract. The Generalized Method of Moments (GMM) attains the semiparametric efficiency bound when the weight matrix is optimally chosen. In this study, we characterize the efficiency of the GMM estimation from the viewpoint of information geometry. Using the convexity of the criterion function of the GMM estimation, we introduce a dually flat connection structure to the model and derive the canonical divergence. At the same time, using the asymptotic normality of the GMM estimator, we formulate statistical differentiation and define the statistical divergence. In conclusion, we prove that the two divergences coincide if and only if the optimal weight matrix is employed.

Keywords: Divergence · Statistical differentiation · Co-metric · Generalized Least Squares · Semiparametric efficiency.

1 Introduction

This paper presents a geometric intuition for the efficiency of the Generalized Method of Moments (GMM). The classical least squares method cannot be applied in regression models where the explanatory variables are correlated with the forecast errors. In econometrics, this is called the endogeneity problem. In order to obtain consistent estimates for regression with endogeneity, instrumental variables, or *instruments*, are often used. A higher dimensional vector of instruments contains more information and allows for more accurate estimation. However, it leads to the problem of over-identification, i.e., solving simultaneous equations whose dimensions are higher than the number of parameters to be estimated.

The GMM is a method proposed by [10] to deal with the over-identification problem, in which an appropriate positive definite matrix is employed as a weight matrix, and solving the high-dimensional simultaneous equations is replaced by a problem of minimizing the quadratic form of the weight matrix. [6] found an optimal weight matrix selection method that allows the GMM estimator to achieve semiparametric efficiency bounds. See also [4] and [5] for more details on the GMM estimation and its efficiency.

In this paper, we understand the efficiency of the GMM estimation from the viewpoint of information geometry. We interpret the introduction of instruments as setting an equivalence relation in the space of random variables. In the quotient space, the set of GMM models is realized as a finite-dimensional manifold. The weight matrix for the GMM estimation defines the Riemannian metric, dually flat structures, and the canonical divergence.

We also propose a new method of differentiation based on asymptotically normally distributed estimators. In the standard theory of differentiable manifolds, a smooth path on the manifold defines a tangent vector. Similarly, in this paper, a sequence of estimates converging in distribution to the true parameter defines statistical versions of the co-metric and the canonical divergence of the GMM manifold.

Hence, the GMM manifold has two different divergences, which are generally not equivalent. We prove that they coincide if and only if the weight matrix is optimal. The result implies that we can illustrate the efficiency of GMM estimation by a Pythagorean theorem between the two divergences derived from the criteria function and the asymptotically normal estimator. This insight offers some intuitions for estimation efficiency in other estimation methods, such as least squares or maximum likelihood.

The application of differential geometry to the estimation efficiency has often focused on the higher-order efficiency [3, 7]. However, this paper considers the case of the first-order efficiency, that is, the conditions for the estimator to have the least asymptotic variance. In addition, although many of the information geometric considerations of statistical models assume an exponential distribution family [1, 2, 15], the models analyzed in this paper do not make assumptions about the distribution type.

This study is based primarily on [12] and [13]: the former shows that when a convex potential function is given on a differentiable manifold, a dually flat structure is naturally derived from the potential; the latter constructs information geometry based on a co-metric rather than a metric. We also use the fact that the criterion function of the GMM estimation is convex to show the dual flatness of the GMM manifold and formulate a statistical version of the co-metric by asymptotic behaviors of the GMM estimator.

2 The Generalized Method of Moments

This section briefly explains asymptotic theory of the linear GMM estimation. Details and proofs of claims given in this section are found in Appendix A of the paper.

Suppose that the probability space $(\Omega, \mathcal{B}, P)$ is given. Denote by $L_2(P)$ the space of squared integrable random variables with norm $\|y\|_{L_2(P)} = (Ey^2)^{1/2}$; by $L_2^m(P)$ the space of m-dimensional vectors of square-integrable random variables with norm $\|x\|_{L_2^m(P)} = (E\sum_{i=1}^m x_i^2)^{1/2}$ for $x = (x_1, \cdots, x_m)^\top$.

Assume that a linear relation

$$y = x^\top \theta + \epsilon \tag{2.1}$$

between $y \in L_2(P)$ and $x \in L_2^m(P)$ holds with error term ϵ and parameter $\theta = (\theta^1, \cdots, \theta^m)^\top \in \mathbb{R}^m$. If conditions $\epsilon \perp\!\!\!\perp x$ and $\operatorname{rank} Exx^\top = m$ are satisfied, θ is consistently estimated by the Ordinary Least Squares (OLS). In applications in economics, however, some components of x are possibly correlated with the error term due to the endogeneity of the model. In such a case, classical OLS fails to be consistent for θ. To deal with the endogeneity problem, we assume the existence of *instruments* $z \in L_2^I(P)$, $I > m$, that satisfy moment conditions

(A1) $Ez\epsilon = 0$

and the rank condition

(A2) $\operatorname{rank} Ezx^\top = \operatorname{rank} Exx^\top = m.$

Under these conditions, θ is characterized as a unique solution to

$$Ez(y - x^\top \theta) = 0. \tag{2.2}$$

If n independent copies

$$(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$$

of (x, y, z) are observed, corresponding equations

$$\frac{1}{n} \sum_{i=1}^{n} z_i(y_i - x_i^\top \theta) = 0 \tag{2.3}$$

might offer an estimator $\hat{\theta}_n$ of θ. However, empirical equations (2.3) might fail to have solutions when $\dim z > \dim x$.

An equivalent way to characterize θ is to consider it as a minimizer of the criterion function

$$M(\theta) = \frac{1}{2}(Ez(y - x^\top \theta))^\top W Ez(y - x^\top \theta), \tag{2.4}$$

where W is a positive definite $I \times I$ matrix.

Let $\hat{E}_n$ be the empirical expectation operator based on the data: for an arbitrary function $f(x, y, z)$,

$$\hat{E}_n f(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} f(x_i, y_i, z_i). \tag{2.5}$$

When $\dim z > \dim x$, the Generalized Method of Moments (GMM) estimator $\hat{\theta}_n$ is defined as the minimizer of the sample M-function

$$\hat{M}(\theta) = \frac{1}{2} \hat{E}_n z(y - x^\top \theta)^\top W \hat{E}_n z(y - x^\top \theta). \tag{2.6}$$

Since $\hat{M}$ is convex with respect to θ, the estimator is obtained by solving the first-order conditions of minimization,

$$\frac{\partial}{\partial \theta} \hat{M}(\theta) = -(\hat{E}_n zx^\top)^\top W \hat{E}_n z(y - x^\top \theta) = 0, \tag{2.7}$$

which yield the formula of the GMM estimator

$$\hat{\theta}_n = ((\hat{E}_n z x^\top)^\top W \hat{E}_n z x^\top)^{-1} (\hat{E}_n z x^\top)^\top W \hat{E}_n z y. \tag{2.8}$$

Assume that

(A3) $E\epsilon^2 z z^\top < \infty$.

Let $\varLambda = E z x^\top$ and $\varSigma = E\epsilon^2 z z^\top$. By the weak law of large numbers and the central limit theorem, consistency

$$\hat{\theta}_n \xrightarrow{p} \theta \tag{2.9}$$

and asymptotic normality

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V) \tag{2.10}$$

are satisfied as $n \to \infty$, where

$$V = \left(\varLambda^\top W \varLambda\right)^{-1} \varLambda^\top W \varSigma W \varLambda \left(\varLambda^\top W \varLambda\right)^{-1}. \tag{2.11}$$

Asymptotics (2.9) and (2.10) hold for arbitrary W, but the size of the asymptotic variance V depends on the choice of W. Chemberlain (1987) proves that by the optimal choice $W_{opt} = \varSigma^{-1}$,

$$V(W) \geq V(W_{opt}) = \left(\varLambda^\top \varSigma^{-1} \varLambda\right)^{-1} \tag{2.12}$$

holds for arbitrary W, where the matrix inequality means that $V(W) - V(W_{opt})$ is positive semi-definite [6].

3 Dually flat manifold and the canonical divergence

This section summarizes [12], which illustrates how to introduce dually flat structures and the canonical divergence of a manifold via a convex function. Proofs of claims in this section are given in Appendix B of the paper.

Consider an m-dimensional C^∞-manifold $\mathcal{M}$ with a coordinate system $\theta = (\theta^1, \cdots, \theta^m) : \mathcal{M} \to \mathbb{R}^m$. Let $C^\infty(\mathcal{M})$ be the set of smooth functions on the manifold. Suppose that a function $\psi \in C^\infty(\mathcal{M})$ has the positive definite Hessian $D^2\psi = [\partial_i \partial_j \psi]$, where

$$\partial_i := \frac{\partial}{\partial \theta_i}. \tag{3.1}$$

We call the pair (θ, ψ) a *frame* of $\mathcal{M}$. If $(\tilde{\theta}, \tilde{\psi})$ satisfies

$$\tilde{\theta} = A\theta + b, \quad \tilde{\psi}(\theta) = \psi(\theta) + c^\top \theta + d \tag{3.2}$$

at every $p \in \mathcal{M}$, where A is an $m \times m$ matrix, $b, c \in \mathbb{R}^m$ and $d \in \mathbb{R}$, we regard $(\tilde{\theta}, \tilde{\psi})$ as equivalent to (θ, ψ), and write $(\theta, \psi) \sim (\tilde{\theta}, \tilde{\psi})$. Let $\mathcal{A}$ be the equivalent class of (θ, ψ).

Define $\eta = (\eta_1, \cdots, \eta_m)^\top : \mathcal{M} \to \mathbb{R}^m$ and $\varphi \in C^\infty(\mathcal{M})$ by

$$\eta_i = \partial_i \psi \quad \text{and} \quad \psi + \varphi = \theta^\top \eta. \tag{3.3}$$

Then, (η, φ) also gives a frame of $\mathcal{M}$. In particular, $\theta^i = \partial^i \varphi$ holds, where

$$\partial^i := \frac{\partial}{\partial \eta_i}. \tag{3.4}$$

The definition (3.3) implies that

$$[\partial_i \partial_j \psi(p)]^{-1} = [\partial^i \partial^j \varphi(p)]. \tag{3.5}$$

We say (η, φ) as the *dual frame* of (θ, ψ), and denote by $\mathcal{A}^*$ the set of dual frames: that is,

$$\mathcal{A}^* = \{(\eta, \varphi) \mid (\eta, \varphi) \text{ is dual of } {}^\exists(\theta, \psi) \in \mathcal{A}\}. \tag{3.6}$$

If (η, φ) is the dual frame of (θ, ψ), then (θ, ψ) is shown to be the dual frame of (η, φ). Hence, $\mathcal{A} = (\mathcal{A}^*)^*$ holds. Additionally, if $(\tilde{\eta}, \tilde{\varphi})$ is the dual frame of $(\tilde{\theta}, \tilde{\psi})$, then $(\theta, \psi) \sim (\tilde{\theta}, \tilde{\psi})$ implies $(\eta, \varphi) \sim (\tilde{\eta}, \tilde{\varphi})$. We call $(\mathcal{M}, \mathcal{A}, \mathcal{A}^*)$ the *dually flat space* for the reason stated below.

The *divergence* of $\mathcal{M}$ is a smooth function $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ such that

$$D(p\|q) \begin{cases} \geq 0, \\ = 0 \iff p = q. \end{cases} \tag{3.7}$$

When the divergence D is given, we can derive geometric properties of $\mathcal{M}$ from D [1, 2, 8, 9]. Denote by $\mathcal{X}(\mathcal{M})$ the set of smooth vector fields on $\mathcal{M}$. The Riemannian metric g of $\mathcal{M}$ is given by

$$g_p(X_p, Y_p) = X_q Y_q D(p\|q)\Big|_{q=p} \tag{3.8}$$

for every $p \in \mathcal{M}$ and $X, Y \in \mathcal{X}(\mathcal{M})$. The affine connections ∇ and ∇^* are also determined by

$$g_p((\nabla_X Y)_p, Z_p) = -X_p Y_p Z_q D(p\|q)\Big|_{q=p} \tag{3.9}$$

and

$$g_p((\nabla^*_X Y)_p, Z_p) = -X_q Y_q Z_p D(p\|q)\Big|_{q=p}. \tag{3.10}$$

To be emphasized here is that (∇, ∇^*) is a pair of the dual connections, which satisfy

$$X\,[g(Y, Z)] = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Y) \tag{3.11}$$

for any $X, Y, Z \in \mathcal{X}(\mathcal{M})$.

For the dually flat space $(\mathcal{M}, \mathcal{A}, \mathcal{A}^*)$, the *canonical divergence* is defined by

$$D(p\|q) = \varphi(p) + \psi(q) - \eta(p)^\top \theta(q), \tag{3.12}$$

where $(\eta, \varphi) \in \mathcal{A}^*$ is the dual frame of $(\theta, \psi) \in \mathcal{A}$. It is shown that (3.12) satisfies the condition (3.7). Furthermore, for any $p, q, r \in \mathcal{M}$, (3.12) satisfies

$$D(p\|q) + D(q\|r) - D(p\|r) = (\eta(p) - \eta(q))^\top (\theta(r) - \theta(q)), \tag{3.13}$$

which implies

$$D(p\|q) + D(q\|r) = D(p\|r) \tag{3.14}$$

if $(\eta(p) - \eta(q))^\top (\theta(r) - \theta(q))$ is satisfied. This is a version of the Pythagorean theorem. Therefore, we can consider $D(p\|q)$ as a squared distance between p and q.

The components of the Riemannian metric are given by

$$g_{ij}(p) := g_p(\partial_i, \partial_j) = \partial_i \partial_j \psi(p) \tag{3.15}$$

with respect to θ, and

$$g^{ij}(p) := g_p(\partial^i, \partial^j) = \partial^i \partial^j \varphi(p) \tag{3.16}$$

with respect to η. Let $G(p) = [g_{ij}(p)]$ and $G^*(p) = [g^{ij}(p)]$. Then, the relation $G(p)^{-1} = G^*(p)$ is confirmed. Furthermore, the connection coefficients of ∇ with respect to θ are

$$\varGamma_{ij,k}(p) := -(\partial_i)_p (\partial_j)_p (\partial_k)_q D(p\|q)\Big|_{q=p} \equiv 0, \tag{3.17}$$

and the coefficients of ∇^* with respect to η are

$$\varGamma^*_{ij,k}(p) := -(\partial^i)_q (\partial^j)_q (\partial^k)_p D(p\|q)\Big|_{q=p} \equiv 0. \tag{3.18}$$

Therefore, θ is the affine coordinate of $(\mathcal{M}, \nabla)$, and η is the affine coordinate of $(\mathcal{M}, \nabla^*)$. Since

$$g_p(\partial_i, \partial^j) = \partial_k \partial_i \psi(p) \partial^k \partial^j \varphi(p) = \begin{cases} 1 \ (i = j) \\ 0 \ (i \neq j) \end{cases}, \tag{3.19}$$

(θ, η) is the dual affine coordinate, and $(\mathcal{M}, g, \nabla, \nabla^*)$ is indeed the dually flat space.

Thus, the dually flat structure (g, ∇, ∇^*) is derived from the dual frames $\{(\theta, \psi), (\eta, \varphi)\}$. It is also shown that the corresponding dual frames are uniquely determined by (g, ∇, ∇^*). Therefore, there is a one-to-one correspondence between $\{(\theta, \psi), (\eta, \varphi)\}$ and (g, ∇, ∇^*).

4 Cotangent space generated by asymptotically normal estimators

In the previous section, we derive a dually flat structure from a convex function defined on a manifold. An alternative method to determine the dually flat structure is constructing the co-tangent space from asymptotically normal statistics.

Let $\mathcal{M}$ be an m-dimensional C^∞-manifold with canonical coordinate $\theta : \mathcal{M} \to \mathbb{R}^m$. Assume that there exits a sequence $\hat{\theta} = \{\hat{\theta}_n\}$ of statistics satisfying asymptotic normality

$$\sqrt{n}(\hat{\theta}_n(p) - \theta(p)) \xrightarrow{d} N(0, V_p) \tag{4.1}$$

at every $p \in \mathcal{M}$, where $V_p = [V^{ij}(p)]$ is the asymptotic variance of $\hat{\theta}(p)$.

Definition 1. *The statistical derivative $(\hat{d}f)_p$ of $f \in C^\infty(\mathcal{M})$ at p relative to $\hat{\theta} = \{\hat{\theta}_n\}$ is the limit distribution of $\sqrt{n}(f(\hat{\theta}_n(p)) - f(\theta(p)))$. That is,*

$$\sqrt{n}(f(\hat{\theta}_n(p)) - f(\theta(p))) \xrightarrow{d} (\hat{d}f)_p \tag{4.2}$$

as $n \to \infty$.

For the coordinate function $\theta : \mathcal{M} \to \mathbb{R}^m$ itself,

$$(\hat{d}\theta)_p = \begin{bmatrix} (\hat{d}\theta^1)(p) \\ \vdots \\ (\hat{d}\theta^m)(p) \end{bmatrix} \sim N(0, V_p). \tag{4.3}$$

The following properties hold for the statistical derivative.

Proposition 1. *For $f, f' \in C^\infty(\mathcal{M})$ and $a \in \mathbb{R}$,*

(i) $(\hat{d}(f + f'))_p = (\hat{d}f)_p + (\hat{d}f')_p$
(ii) $(\hat{d}(af))_p = a(\hat{d}f)_p$
(iii) $(\hat{d}(f \cdot f'))_p = f'(p)(\hat{d}f)_p + f(p)(\hat{d}f')_p$
(iv) $(\hat{d}f)_p = \partial_i f(p)(\hat{d}\theta^i)_p$

The property (iv) of the proposition is known as the Delta method in statistics [14, 16]. The properties imply that $(\hat{d}f)_p$ is identified with the ordinary derivative $(df)_p \in T_p^*\mathcal{M}$.

Definition 2. *The statistical cotangent space of $\mathcal{M}$ at p relative to $\hat{\theta} = \{\hat{\theta}_n\}$ is defined as*

$$\hat{T}_p^*\mathcal{M} = \{(\hat{d}f)_p \mid f \in C^\infty(\mathcal{M})\}. \tag{4.4}$$

Proposition 2. *The dimension of $\hat{T}_p^*\mathcal{M}$ is m, and*

$$\hat{T}_p^*\mathcal{M} = \operatorname{span}\{(\hat{d}\theta^1)_p, \cdots, (\hat{d}\theta^m)_p\}. \tag{4.5}$$

Considering that each statistical co-tangent vector is a normally distributed random variable, we define the *statistical co-metric* $\hat{g}_p$ by

$$\hat{g}_p(\hat{\alpha}_p, \hat{\beta}_p) = \operatorname{Cov}(\hat{\alpha}_p, \hat{\beta}_p) = \alpha_i \beta_j V^{ij}(p) \tag{4.6}$$

for $\hat{\alpha}_p = \alpha_i(\hat{d}\theta^i)_p$ and $\hat{\beta}_p = \beta_j(\hat{d}\theta^j)_p$ in $\hat{T}_p^*\mathcal{M}$.

Definition 3. *The statistical tangent space $\hat{T}_p\mathcal{M}$ is the dual space of $\hat{T}^*_p\mathcal{M}$.*

Let $\{(\hat{\partial}_1)_p, \ldots, (\hat{\partial}_m)_p\}$ be the dual basis of $\{(\hat{d}\theta^1)_p, \ldots, (\hat{d}\theta^m)_p\}$:

$$\hat{T}_p\mathcal{M} = \text{span}\left\{(\hat{\partial}_1)_p, \ldots, (\hat{\partial}_m)_p\right\}, \tag{4.7}$$

and

$$(\hat{\partial}_i)_p(\hat{d}\theta^j)_p = \begin{cases} 1 \ (i = j) \\ 0 \ (i \neq j) \end{cases}. \tag{4.8}$$

Each $\hat{A}_p \in \hat{T}_p\mathcal{M}$ can operate on $f \in C^\infty(\mathcal{M})$ by

$$\hat{A}_p(f) = \hat{A}_p(\hat{d}f)_p. \tag{4.9}$$

Then,

$$(\hat{\partial}_i)_p f = (\hat{\partial}_i)_p \left[\partial_j f(p)(\hat{d}\theta^j)_p\right] = (\partial_i)_p f, \tag{4.10}$$

and $\hat{T}_p\mathcal{M}$ is naturally identified with $T_p\mathcal{M}$ by $\hat{\partial}_i \mapsto \partial_i = \frac{\partial}{\partial\theta^i}$.

According to [13], given the co-metric $\hat{g}$, correspondence $\hat{A}_p \overset{\hat{g}_p}{\longleftrightarrow} \hat{\alpha}_p$ between a tangent vector $\hat{A}_p \in \hat{T}_p\mathcal{M}$ and a co-tangent vector $\hat{\alpha}_p \in \hat{T}^*_p\mathcal{M}$ is given by

$$\hat{A}_p(\hat{\beta}_p) \equiv \hat{g}_p(\hat{\alpha}_p, \hat{\beta}_p) \tag{4.11}$$

for every $\hat{\beta}_p \in \hat{T}^*_p\mathcal{M}$. Define $(\hat{\partial}^i)_p$ by

$$(\hat{\partial}^i)_p \overset{\hat{g}_p}{\longleftrightarrow} (\hat{d}\theta^i)_p. \tag{4.12}$$

For $\hat{\beta}_p = \beta_j(\hat{d}\theta^j)_p$,

$$(\hat{\partial}^i)_p(\hat{\beta}_p) = \hat{g}_p((\hat{d}\theta^i)_p, \hat{\beta}_p) = V^{ij}(p)\beta_j,$$

which implies $(\hat{\partial}_j)_p = V_{ij}(p)(\hat{\partial}^i)_p$ with $V_p^{-1} = [V_{ij}(p)]$. Therefore, the *statistical metric* of $\mathcal{M}$ is defined by

$$\hat{g}_p(\hat{\partial}_i, \hat{\partial}_j) = V_{ij}(p). \tag{4.13}$$

By identifying $\hat{\partial}_i$ with $\partial_i = \frac{\partial}{\partial\theta^i}$, the potential $\hat{\psi}$ and its dual potential $\hat{\varphi}$ are respectivelty given by

$$\begin{cases} V_{ij}(p) = \left(\frac{\partial}{\partial\theta^i}\right)_p \left(\frac{\partial}{\partial\theta^j}\right)_p \hat{\psi}(\theta) \\ \\ V^{ij}(p) = \left(\frac{\partial}{\partial\eta_i}\right)_p \left(\frac{\partial}{\partial\eta_j}\right)_p \hat{\varphi}(\eta), \end{cases} \tag{4.14}$$

where $\eta = V^{-1}\theta$.

Definition 4. *Relative to $\hat{\theta} = \{\hat{\theta}_n\}$ such that $\sqrt{n}(\hat{\theta}_n(p) - \theta(p)) \xrightarrow{d} N(0, V_p)$, the statistical divergence of $\mathcal{M}$ is defined by*

$$\hat{D}(p\|q) = \hat{\varphi}(\eta(p)) + \hat{\psi}(\theta(p)) - \eta(p)^\top\theta(p). \tag{4.15}$$

5 Geometry of the GMM model

Geometrically, the GMM is illustrated as a projection of y onto the plane spanned by x. To make this intuition more rigorous, we will define an equivalence relation on $L_2(P)$ using instruments z. We say that w and w' in $L_2(P)$ are equivalent with respect to z if $Ez(w-w')=0$ holds, and express the relation as $w \sim_z w'$. We denote by $[w]$ the equivalence class of each w and by $L_2(P)/z$ the quotient space of $L_2(P)$ under $\sim_z$. The projection $q : w \mapsto [w]$ naturally induces the quotient topology of $L_2(P)/z$, where $U \subset L_2(P)/z$ is open if $q^{-1}(U) \subset L_2(P)$ is open.

Addition and real multiplication in $L_2(P)/z$ are defined as follows:

$$\begin{cases} [w]+[w'] = [w+w'] \\ \alpha[w] = [\alpha w] \end{cases} \tag{5.1}$$

for $w, w' \in L_2(P)$ and $\alpha \in \mathbb{R}$. The operations are well-defined since, if $w_1 \sim_z w$ and $w_2 \sim_z w'$,

$$Ez((w_1+w_2)-(w+w')) = Ez(w_1-w) + Ez(w_2-w') = 0$$

and

$$Ez(\alpha w_1 - \alpha w) = \alpha Ez(w_1 - w) = 0.$$

Denote by $\mathrm{span}(x)$ the subspace of $L_2(P)$ spanned by $x=(x_1,\dots,x_m)^\top$, that is,

$$\mathrm{span}(x) = \{x^\top\theta \mid \theta \in \mathbb{R}^m\}. \tag{5.2}$$

Note that

$$\mathrm{span}(x)/z = \mathrm{span}\{[x_1],\dots,[x_m]\}$$

because $[x^\top\theta] = [x_1]\theta^1 + \cdots + [x_m]\theta^m$ for every $\theta = (\theta^1,\cdots,\theta^m)$. Let us define the GMM model set by

$$\mathcal{M} = \mathrm{span}(x)/z. \tag{5.3}$$

Then, the GMM model (2.1) is simply expressed as

$$[y] \in \mathcal{M}. \tag{5.4}$$

We introduce open sets $\mathcal{O}_\mathcal{M}$ of the model set by restricting topology of $L_2(P)/z$ onto $\mathcal{M}$. Then, the next result is shown.

Proposition 3. *Assume* ***(A1)-(A2)****. Then, $\mathcal{M}$ is a m-dimensional C^∞ manifold with the canonical coordinate $[x^\top\theta] \mapsto \theta$.*

Proof. Choose arbitrary points p and p' from $\mathcal{M}$. Then θ and θ' exist, such that $p=[x^\top\theta]$ and $p'=[x^\top\theta']$. If $\theta \neq \theta'$,

$$Ez(x^\top\theta - x^\top\theta') = Ezx^\top(\theta-\theta') \neq 0$$

because $\ker Ezx^\top = \{0\}$ by the assumption. Hence, $p \neq p'$ is shown.

For every $[U] \subset \mathcal{M}$, there exists $V \subset \mathbb{R}^m$ such that $[U] = \{[x^\top \theta] \mid \theta \in V\}$. Let $q' : \mathrm{span}(x) \to \mathrm{span}(x)/z$ be the standard projection. Since $[x^\top \theta] \mapsto \theta$ is bijection, $q'(w) = \{w\}$ holds for every $w \in \mathrm{span}(x)$. Hence, $[U] \in \mathcal{O}_\mathcal{M}$ if and only if $U := (q')^{-1}([U]) = \{x^\top \theta \mid \theta \in V\}$ is an open subset of $\mathrm{span}(x)$. Moreover, U is shown to be open if and only if V is open. Let U be open in $\mathrm{span}(x)$. There exist an open subset U_0 of $L_2(P)$ and $V \subset \mathbb{R}^m$ such that $U = U_0 \cap \mathrm{span}(x) = \{x^\top \theta \mid \theta \in V\}$. To be shown is openness of V. Choose arbitrary $\tau \in V$, and set $w = x^\top \tau$. Since $w \in U_0$, there exists sufficiently small $\delta > 0$ such that

$$B(w, \delta) := \{w' \in L_2(P) \mid \|w' - w\|_{L_2(P)} < \delta\} \subset U_0,$$

which implies

$$\begin{aligned} U_0 \cap \mathrm{span}(x) &\supset B(w, \delta) \cap \mathrm{span}(x) \\ &= \{x^\top \rho \mid \rho \in \mathbb{R}^m,\ \|x^\top \rho - x^\top \tau\|_{L_2(P)} < \delta\} \\ &= \{x^\top \rho \mid \rho \in \mathbb{R}^m,\ (\rho - \tau)^\top E x x^\top (\rho - \tau) < \delta^2\}. \end{aligned}$$

By **(A2)**, there exists $\lambda > 0$ such that $|\rho - \tau| < \lambda$ implies $(\rho - \tau)^\top E x x^\top (\rho - \tau) < \delta^2$. Hence,

$$U = \{x^\top \theta \mid \theta \in V\} \supset \{x^\top \rho \mid \rho \in \mathbb{R}^m,\ |\rho - \tau| < \lambda\},$$

which shows $\{\rho \in \mathbb{R}^m \mid |\rho - \tau| < \lambda\} \subset V$. To show the inverse is straightforward.

Again, choose arbitrary points $p = [x^\top \theta]$ and $p' = [x^\top \theta']$ from $\mathcal{M}$. If $p \neq p'$, then $\delta := |\theta - \theta'|/3 > 0$. Let U and U' be open sets respectively defined by

$$U = \{[x^\top \tau] \mid |\tau - \theta| < \delta\},\ U' = \{[x^\top \tau'] \mid |\tau' - \theta'| < \delta\}.$$

Then, $p \in U$, $p \in U'$, and $U \cap U' = \emptyset$. Hence, $\mathcal{M}$ is Hausdorff.

It is clear from the above discussion that $\mathcal{M}$ and $\mathbb{R}^m$ are homeomorphic by $[x^\top \theta] \mapsto \theta$. □

Let us define an inner product of $L_2(P)/z$ by

$$\langle [w], [w'] \rangle = (Ezw)^\top W Ezw' \tag{5.5}$$

for every $[w], [w'] \in L_2(P)/z$. A norm is also given by

$$\|[w]\| = \langle [w], [w] \rangle^{1/2}. \tag{5.6}$$

Note that $\|[w]\| = \|[w']\|$ holds if $w \sim_z w'$. Particularly, $\|[w]\| = 0$ holds if and only if $w \in [0]$. Topology generated by the norm is weaker than the standard quotient toporogy. By restricting the norm on $\mathcal{M}$, however,

$$\|p\| = \sqrt{\theta^\top \Lambda^\top W \Lambda \theta} \tag{5.7}$$

holds for every $p = [x^\top \theta]$. Since rank $\Lambda^\top W \Lambda = m$, the topology induced by the norm is equivalent to $\mathcal{O}_{\mathcal{M}}$.

A goal of the GMM estimation is to find θ solving

$$[y] = [x^\top \theta]. \tag{5.8}$$

The solution is found by minimizing

$$M(\theta) = \frac{1}{2} \| [y] - [x^\top \theta] \|^2 \tag{5.9}$$

with respect to θ. In this sense, the GMM is understood as the projection of data $[y]$ onto the model space $\mathcal{M}$.

Note that $M(\theta) \sim \psi(\theta) = \frac{1}{2}\theta^\top \Lambda^\top W \Lambda \theta$, which is the convex function of θ. Therefore, (θ, ψ) is a frame of $\mathcal{M}$, and the (i, j) component of the Riemannian metric g is given by

$$g_{ij}(p) \equiv (\Lambda^\top W \Lambda)_{ij}. \tag{5.10}$$

In fact, since $\mathcal{M}$ is a linear space, the tangent space $T_p\mathcal{M}$ at every $p = [x^\top \theta]$ is identified with $\mathcal{M}$ itself. We denote this identification $T_p\mathcal{M} \to \mathcal{M}$ by

$$A_p = a^i (\partial_i)_p \mapsto A_p^{\mathcal{M}} := [x^\top a], \tag{5.11}$$

where $a = (a^1, \cdots, a^m)^\top$. By the identification, the metric g of $\mathcal{M}$ is naturally induced by (5.7): for every $A_p = a^i(\partial_i)_p$ and $B_p = b^i(\partial_i)_p$,

$$g_p(A_p, B_p) = \left\langle A_p^{\mathcal{M}}, B_p^{\mathcal{M}} \right\rangle = a^\top \Lambda^\top W \Lambda b. \tag{5.12}$$

The correpondence $A_p \overset{g_p}{\longleftrightarrow} \alpha_p$ determines a co-metric on $\mathcal{M}$ by

$$g_p(\alpha_p, \beta_p) = g_p(A_p, B_p) \tag{5.13}$$

if $A_p \overset{g_p}{\longleftrightarrow} \alpha_p$ and $B_p \overset{g_p}{\longleftrightarrow} \beta_p$. Letting

$$g^{ij}(p) := g_p((d\theta^i)_p, (d\theta^j)_p), \tag{5.14}$$

we have $[g^{ij}(p)] = (\Lambda^\top W \Lambda)^{-1}$. In fact, the dual frame of (θ, ψ) is given by

$$\eta = \frac{\partial}{\partial \theta} \left(\frac{1}{2} \theta^\top (\Lambda^\top W \Lambda) \theta \right) = (\Lambda^\top W \Lambda) \theta \tag{5.15}$$

and

$$\varphi(\eta) = \eta^\top (\Lambda^\top W \Lambda)^{-1} \eta - \psi((\Lambda^\top W \Lambda)^{-1} \eta) = \frac{1}{2} \eta^\top (\Lambda^\top W \Lambda)^{-1} \eta. \tag{5.16}$$

Therefore, the canonical divergence of the GMM manifold is given by

$$D(p \| q) = \frac{1}{2} (\theta(p) - \theta(q))^\top (\Lambda^\top W \Lambda)(\theta(p) - \theta(q)) = \frac{1}{2} \| p - q \|^2, \tag{5.17}$$

which is the squared norm of $L_2(P)/z$.

Relative to the GMM estimator (2.8), on the other hand, the statistical divergence is obtained by

$$\hat{D}(p\|q) = \frac{1}{2}(\theta(p)-\theta(q))^\top(\varLambda^\top W\varLambda)(\varLambda^\top W\varSigma W\varLambda)^{-1}(\varLambda^\top W\varLambda)(\theta(p)-\theta(q)). \quad (5.18)$$

The main theorem of the paper is stated as follows.

Theorem 1. *The canonical divergence D of the GMM manifold $\mathcal{M} = \mathrm{span}(x)/z$ with the frame (θ, M) is identical to the statistical divergence $\hat{D}$ relative to the GMM estimator (2.8) if and only if the weight matrix is given by $W = (\varLambda^\top\varSigma^{-1}\varLambda)^{-1}$.*

Proof. Let $\psi(\theta) \sim \tilde{\psi}(\theta) = \psi(\theta)+c^\top\theta+d$. Then, the dual parameter $\tilde{\eta} = \partial\psi(\theta)+c = \eta + c$ is also equivalent to η. The dual potential is

$$\begin{aligned}\tilde{\varphi}(\tilde{\eta}) &= \tilde{\eta}^\top\theta - \tilde{\psi}(\theta)\Big|_{\theta=(\partial\psi)^{-1}(\tilde{\eta}-c)}\\ &= (\eta+c)^\top\theta - (\psi(\theta)+c^\top\theta+d)\Big|_{\theta=(\partial\psi)^{-1}(\tilde{\eta}-c),\,\eta=\tilde{\eta}-c}\\ &= \varphi(\tilde{\eta}-c) - d,\end{aligned}$$

and the divergence given by $(\theta, \tilde{\psi})$ and $(\tilde{\eta}, \tilde{\varphi})$ is

$$\begin{aligned}\tilde{D}(p\|q) &= \tilde{\varphi}(\tilde{\eta}(p)) + \tilde{\psi}(\theta(q)) - \tilde{\eta}(p)^\top\theta(q)\\ &= \big[\varphi((\eta(p)+c)-c)-d\big] + \big[\psi(\theta(q))+c^\top\theta(q)+d\big] - (\eta(p)+c)^\top\theta(q)\\ &= D(p\|q).\end{aligned}$$

Since

$$M(\theta) = \frac{1}{2}\theta^\top(\varLambda^\top W\varLambda)\theta - (Ezy)^\top W\varLambda\theta + \frac{1}{2}(Ezy)^\top E(Ezy) \sim \frac{1}{2}\theta^\top(\varLambda^\top W\varLambda)\theta,$$

the canonical divergence generated by (θ, M) is

$$D(p\|q) = \frac{1}{2}(\theta(p)-\theta(q))^\top(\varLambda^\top W\varLambda)(\theta(p)-\theta(q)). \quad (5.19)$$

Therefore, $D = \hat{D}$ holds if and only if

$$\varLambda^\top W\varLambda = (\varLambda^\top W\varLambda)(\varLambda^\top W\varSigma W\varLambda)^{-1}(\varLambda^\top W\varLambda),$$

which implies $W = \varSigma^{-1}$. □

6 Discussions

A geometric interpretation of Theorem 1 is given in Figure 6.1. As shown in (5.17), the canonical divergence D is the squared norm of $L_2(P)/z$. Hence, the canonical digergence measures the distance between data $[y]$ and the model $[x^\top \theta]$. The true model is the orthogonal projection of $[y]$ under D. On the other hand, the statistical divergence $\hat{D}$ is interpreted as a measure of the gap between the estimates $\hat{\theta}$ and true parameter value θ since the divergence is derived from the asymptotic variance of the estimator. When the two distances coincide, the generalized Pythagorean relation (3.14) justifies Figure 6.1, where estimation errors $\hat{\theta}_n - \theta$ and data $[y]$ are orthgonal. However, when $D \neq \hat{D}$, the orthogonal projection does not necessarily satisfy the Pythagorean relation, and the GMM estimator fails to be efficient.

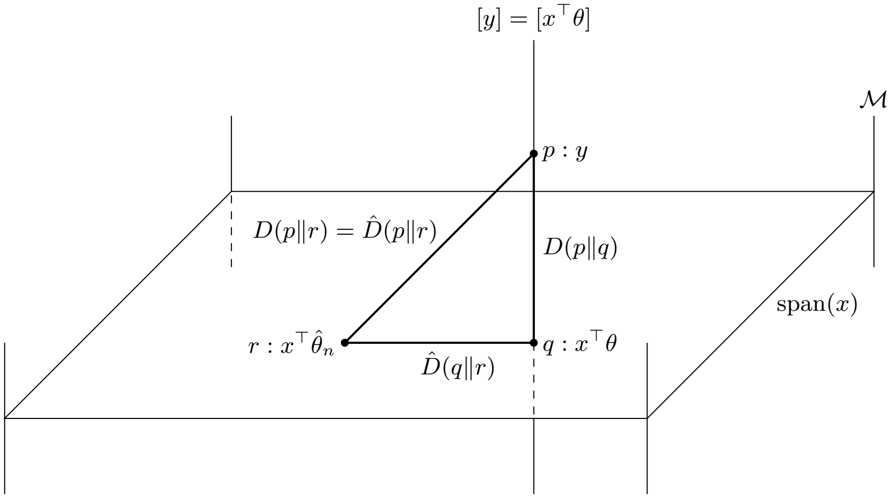

Fig. 6.1. The GMM manifold $\mathcal{M}$ when $D = \hat{D}$.

Consider the classical linear regression model

$$y = x^\top \theta + \epsilon, \ \ E(\epsilon | x) = 0 \tag{6.1}$$

to see that the above intuition is valid. Let $\mathcal{M} = \{x^\top \theta \mid \theta \in \mathbb{R}^k\}$. Then, $\mathcal{M}$ is shown to be a C^∞-manifold with canonical coordinate $p = x^\top \theta \mapsto \theta$. The criterion function corresponding to (6.1) is the sample analog of

$$M(\theta) = \frac{1}{2} E w(x)(y - x^\top \theta)^2, \tag{6.2}$$

where $w(x)$ is a positive weight function chosen by the statistician. The corresponding canonical divergence is shown to be

$$D(p\|q) = \frac{1}{2}(\theta(p) - \theta(q))^\top (E w(x) x x^\top)(\theta(p) - \theta(q)). \tag{6.3}$$

The generalized least squares estimator (GLS) for the parameters is

$$\hat{\theta}_n(p) = \left[\hat{E}_n w(x) x x^\top\right]^{-1} \hat{E}_n w(x) x (x^\top \theta(p) + \epsilon), \tag{6.4}$$

which shows asymptotic normality

$$\sqrt{n}(\hat{\theta}_n(p) - \theta(p)) \xrightarrow{d} N\left(0, V_p\right)$$

as $n \to \infty$, where $\sigma^2(x) = E(\epsilon^2 \mid x)$ and

$$V_p \equiv (E w(x) x x^\top)^{-1} (E \sigma^2(x) w(x)^2 x x^\top) (E w(x) x x^\top)^{-1}.$$

Hence, the statistical divergence derived by the GLS estimator is

$$\hat{D}(p \| q) = \frac{1}{2}(\theta(p) - \theta(q)) V_p (\theta(p) - \theta(q)). \tag{6.5}$$

Notably, $D = \hat{D}$ holds when and only when the optimal weight function

$$w(x) = \frac{1}{\sigma^2(x)}, \tag{6.6}$$

is chosen [6, 11].

Now consider the maximum likelihood estimation method. The criterion function of the maximum likelihood estimation is a sample analog of $M(\theta) = E \log p(x, \theta)$, where $p(\cdot, \theta)$ is the density function of x parametrized by θ. Since the Hessian matrix of M at the true parameter value is the information matrix

$$I(\theta) = \left[E\left(\frac{\partial_i p(x,\theta)}{p(x,\theta)}\right)\left(\frac{\partial_j p(x,\theta)}{p(x,\theta)}\right)\right], \tag{6.7}$$

$M(\tau)$ is well approximated by $\psi(\tau) = \frac{1}{2}\tau^\top I(\theta)\tau$ in a neighborhood of θ. On the other hand, the asymptotic variance of the maximum likelihood estimator $\hat{\theta}_n$ is $I(\theta)^{-1}$. Therefore, $D = \hat{D}$ locally holds, which is consistent with the fact that the maximum likelihood estimator achieves the estimation efficiency.

References

1. Amari, S.: Information Geometry and Its Applications. Springer Japan (2016)
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society (2000)
3. Andrews, I., Mikusheva, A.: A geometric approach to nonlinear econometric models. Econometrica **84**(3), 1249–1264 (2016)
4. Bates, C.E., White, H.: Determination of estimators with minimum asymptotic covariance matrices. Econometric Theory **9**(4), 633–648 (1993)
5. Carrasco, M., Jean-Pierre, F.: On the asymptotic efficiency of GMM. Econometric Theory **30**(2), 372–406 (2014)
6. Chamberlain, G.: Asymptotic efficiency in estimation with conditional moment restrictions. Journal of econometrics **34**(3), 305–334 (1987)
7. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Annals of Statistics **11**(3), 793–803 (1983)
8. Eguchi, S.: Geometry of minimum contrast. Hiroshima Mathematical Journal **22**, 631–647 (1992)
9. Eguchi, S., Komori, O.: Minimum Divergence Methods in Statistical Machine Learning. Springer (2022)
10. Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica **50**(4), 1029–1054 (1982)
11. Lee, M.J.: Micro-Econometrics: Methods of Moments and Limited Dependent Variables. Second Edition. Springer (2010)
12. Nagaoka, H.: [Dual flatness of connections and dually flat spaces: Mathematical foundation of information geometry] Setsuzoku no soutsuisei to soutsui heitan kuukan: Jouhou kikagaku no suugakuteki kiso (in Japanese). Suurikagaku **58**(11) 15–21 (2020)
13. Nagaoka, H.: The Fisher metric as a metric on the cotangent bundle. Information Geometry **7**(1), 651–677 (2024)
14. Shao, J.: Mathematical Statistics (2nd ed.). Springer (2003)
15. Tanaka, H: Dually flat structure of binary choice models. Information Geometry **7**(1), 1–18 (2024)
16. van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press (1998)