Statistical Manifolds Admitting Torsion and Partially Flat Spaces

Masayuki Henmi and Hiroshi Matsuzoe

Abstract It is well-known that a contrast function defined on a product manifold $M \times M$ induces a Riemannian metric and a pair of dual torsion-free affine connections on the manifold M. This geometrical structure is called a statistical manifold and plays a central role in information geometry. Recently, the notion of pre-contrast function has been introduced and shown to induce a similar differential geometrical structure on M, but one of the two dual affine connections is not necessarily torsion-free. This structure is called a statistical manifold admitting torsion. The notion of statistical manifolds admitting torsion has been originally introduced to study a geometrical structure which appears in a quantum statistical model. However, it has been shown that an estimating function which is used in "classical" statistics also induces a statistical manifold admitting torsion through its associated pre-contrast function. The aim of this paper is to summarize such previous results. In particular, we focus on a partially flat space, which is a statistical manifold admitting torsion where one of its dual connections is flat. In this space, it is possible to discuss some properties similar to those in a dually flat space, such as a canonical pre-contrast function and a generalized projection theorem.

1 Introduction

A statistical manifold is a Riemannian manifold with a pair of dual torsion-free affine connections and it plays a central role in information geometry. This geometrical structure is induced from an asymmetric (squared) distance-like smooth function called a contrast function by taking its second and third derivatives [1, 2]. The Kullback–Leibler divergence on a regular parametric statistical model is a typical example of contrast functions and its induced geometrical objects are the Fisher met-

M. Henmi (✉)
The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
e-mail: henmi@ism.ac.jp

H. Matsuzoe
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555, Japan
e-mail: matsuzoe@nitech.ac.jp

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_3

ric, the exponential and mixture connections. The geometrical structure determined by these objects plays an important role in the geometry of statistical inference, as is widely known [3, 4].

A statistical manifold admitting torsion (SMAT) is a Riemannian manifold with a pair of dual affine connections, where only one of them must be torsion-free but the other is *not* necessarily so. This geometrical structure naturally appears in a quantum statistical model (i.e. a set of density matrices representing quantum states) [3] and the notion of SMAT was originally introduced to study such a geometrical structure from a mathematical point of view [5]. A pre-contrast function was subsequently introduced as a generalization for the first derivative of a contrast function and it was shown that an pre-contrast function induces a SMAT by taking its first and second derivatives [6].

In statistics, an estimating function is a function defined on a direct product of parameter and sample spaces, and it is used to obtain an estimator by solving its corresponding estimating equation. Henmi and Matsuzoe [7] showed that a SMAT also appears in “classical" statistics through an estimating function. More precisely, an estimating function naturally defines a pre-contrast function on a parametric statistical model and a SMAT is induced from it.

This paper summarizes such previous results, focusing on a SMAT where one of its dual connections is flat. We call this geometrical structure a partially flat space. Although this space is different from a dually flat space in general since one of the dual connections in a SMAT possibly has torsion, some similar properties hold. For example, the canonical pre-contrast function can be naturally defined on a partially flat space, which is an analog of the canonical contrast function (or canonical divergence) in a dually flat space. In addition, a generalized projection theorem holds with respect to the canonical pre-contrast function. This theorem can be seen as a generalization of the projection theorem in a dually flat space. This paper is an extended version of the conference proceedings [8]. We consider a statistical problem to see an example of statistical manifolds admitting torsion induced from estimating functions and discuss some future problems, neither of which were included in [8].

2 Statistical Manifolds and Contrast Functions

Through this paper, we assume that all geometrical objects on differentiable manifolds are smooth and restrict our attention to Riemannian manifolds, although the most of the concepts can be defined for semi-Riemannian manifolds.

Let (M, g) be a Riemannian manifold and ∇ be an affine connection on M. The *dual connection* ∇^* of ∇ with respect to g is defined by

$$Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z) \quad (\forall X, \forall Y, \forall Z \in \mathcal{X}(M)),$$

where $\mathcal{X}(M)$ is the set of all vector fields on M.

For an affine connection ∇ on M, its curvature tensor field R and torsion tensor field T are defined by the following equations as usual:

$$R(X, Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,$$
$$T(X, Y) := \nabla_X Y - \nabla_Y X - [X, Y]$$

$(\forall X, \forall Y, \forall Z \in \mathscr{X}(M))$. It is said that an affine connection ∇ is *torsion-free* if $T = 0$. Note that for a torsion-free affine connection ∇, $\nabla^* = \nabla$ implies that ∇ is the Levi-Civita connection with respect to g. Let R^* and T^* be the curvature and torsion tensor fields of ∇^*, respectively. It is easy to see that $R = 0$ always implies $R^* = 0$, but $T = 0$ does not necessarily imply $T^* = 0$.

Let ∇ be a torsion-free affine connection on a Riemannian manifold (M, g). Following [9], we say that (M, g, ∇) is a *statistical manifold* if and only if ∇g is a symmetric $(0, 3)$-tensor field, that is

$$(\nabla_X g)(Y, Z) = (\nabla_Y g)(X, Z) \quad (\forall X, \forall Y, \forall Z \in \mathscr{X}(M)). \tag{1}$$

This condition is equivalent to $T^* = 0$ under the condition that ∇ is a torsion-free. If (M, g, ∇) is a statistical manifold, so is (M, g, ∇^*) and it is called the *dual statistical manifold* of (M, g, ∇). Since ∇ and ∇^* are both torsion-free for a statistical manifold (M, g, ∇), $R = 0$ implies that ∇ and ∇^* are both flat. In this case, (M, g, ∇, ∇^*) is called a *dually flat space* [3].

Let ϕ be a real-valued function on the direct product $M \times M$ of a manifold M and $X_1, \ldots, X_i, Y_1, \ldots, Y_j$ be vector fields on M. The functions $\phi[X_1, \ldots, X_i | Y_1, \ldots, Y_j]$, $\phi[X_1, \ldots, X_i |\ ]$ and $\phi[\ | Y_1, \ldots, Y_j]$ on M are defined by the equations

$$\phi[X_1, \ldots, X_i | Y_1, \ldots, Y_j](r) := (X_1)_p \cdots (X_i)_p (Y_1)_q \cdots (Y_j)_q \phi(p, q)|_{p=r, q=r}, \tag{2}$$

$$\phi[X_1, \ldots, X_i |\ ](r) := (X_1)_p \cdots (X_i)_p \phi(p, r)|_{p=r}, \tag{3}$$

$$\phi[\ | Y_1, \ldots, Y_j](r) := (Y_1)_q \cdots (Y_j)_q \phi(r, q)|_{q=r} \tag{4}$$

for any $r \in M$, respectively [1]. Using these notations, a *contrast function* ϕ on M is defined to be a real-valued function on $M \times M$ which satisfies the following conditions [1, 2]:

(*a*) $\phi(p, p) = 0 \quad (\forall p \in M)$,
(*b*) $\phi[X |\ ] = \phi[\ | X] = 0 \quad (\forall X \in \mathscr{X}(M))$,
(*c*) $g(X, Y) := -\phi[X | Y] \ (\forall X, \forall Y \in \mathscr{X}(M))$ is a Riemannian metric on M.

Note that these conditions imply that

$$\phi(p, q) \geq 0, \quad \phi(p, q) = 0 \Longleftrightarrow p = q$$

in some neighborhood of the diagonal set $\{(r, r)|r \in M\}$ in $M \times M$. Although a contrast function is not necessarily symmetric, this property means that a contrast function measures some discrepancy between two points on M (at least locally). For a given contrast function ϕ, the two affine connections ∇ and ∇^* are defined by

$$g(\nabla_X Y, Z) = -\phi[XY|Z], \quad g(Y, \nabla^*_X Z) = -\phi[Y|XZ]$$

$(\forall X, \forall Y, \forall Z \in \mathcal{X}(M))$. In this case, ∇ and ∇^* are both torsion-free and dual to each other with respect to g. This means that both of (M, g, ∇) and (M, g, ∇^*) are statistical manifolds. In particular, (M, g, ∇) is called the statistical manifold induced from the contrast function ϕ.

A typical example of contrast functions is the Kullback–Leibler divergence on a statistical model. Let $S = \{p(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta^1, \ldots, \theta^d) \in \Theta \subset \boldsymbol{R}^d\}$ be a regular parametric statistical model, which is a set of probability density functions with respect to a dominating measure ν on a sample space Ω. Each element is indexed by a parameter (vector) $\boldsymbol{\theta}$ in an open subset Θ of $\boldsymbol{R}^d$ and the set S satisfies some regularity conditions, under which S can be seen as a differentiable manifold. The Kullback–Leibler divergence of the two density functions $p_1(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_2)$ in S is defined to be

$$\phi_{KL}(p_1, p_2) := \int_\Omega p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})} \nu(d\boldsymbol{x}).$$

It is easy to see that the Kullback–Leibler divergence satisfies the conditions (a), (b) and (c), and so it is a contrast function on S. Its induced Riemannian metric and dual connections are Fisher metric g^F, the exponential connection $\nabla^{(e)}$ and mixture connection $\nabla^{(m)}$, respectively. They are given as follows:

$$\begin{aligned} & g^F_{jk}(\boldsymbol{\theta}) := g^F(\partial_j, \partial_k) = E_{\boldsymbol{\theta}}\{s^j(\boldsymbol{x}, \boldsymbol{\theta}) s^k(\boldsymbol{x}, \boldsymbol{\theta})\}, \\ & \begin{cases} \Gamma^{(e)}_{ij,k}(\boldsymbol{\theta}) := g^F(\nabla^{(e)}_{\partial_i} \partial_j, \partial_k) = E_{\boldsymbol{\theta}}[\{\partial_i s^j(\boldsymbol{x}, \boldsymbol{\theta})\} s^k(\boldsymbol{x}, \boldsymbol{\theta})] \\ \Gamma^{(m)}_{ik,j}(\boldsymbol{\theta}) := g^F(\partial_j, \nabla^{(m)}_{\partial_i} \partial_k) = \int_\Omega s^j(\boldsymbol{x}, \boldsymbol{\theta}) \partial_i \partial_k p(\boldsymbol{x}; \boldsymbol{\theta}) \nu(d\boldsymbol{x}) \end{cases}, \end{aligned}$$

where $E_{\boldsymbol{\theta}}$ indicates that the expectation is taken with respect to $p(\boldsymbol{x}; \boldsymbol{\theta})$, $\partial_i = \frac{\partial}{\partial \theta^i}$ and $s^i(\boldsymbol{x}; \boldsymbol{\theta}) = \partial_i \log p(\boldsymbol{x}; \boldsymbol{\theta})$ $(i = 1, \ldots, d)$. As is widely known, this geometrical structure plays the most fundamental and important role in the differential geometry of statistical inference [3, 4].

3 Statistical Manifolds Admitting Torsion and Pre-contrast Functions

A statistical manifold admitting torsion is an abstract notion for the geometrical structure where only one of the dual connections is allow to have torsion, which

naturally appears in a quantum statistical model [3]. The definition is obtained by generalizing (1) in the definition of statistical manifold as follows [5].

Let (M, g) be a Riemannian manifold and ∇ be an affine connection on M. We say that (M, g, ∇) is a *statistical manifold admitting torsion* (SMAT for short) if and only if

$$(\nabla_X g)(Y, Z) - (\nabla_Y g)(X, Z) = -g(T(X, Y), Z) \quad (\forall X, \forall Y, \forall Z \in \mathscr{X}(M)). \tag{5}$$

This condition is equivalent to $T^* = 0$ in the case where ∇ possibly has torsion, and it reduces to (1) if ∇ is torsion-free. Note that (M, g, ∇^*) is not necessarily a statistical manifold although ∇^* is torsion-free. It should be also noted that (M, g, ∇^*) is a SMAT whenever a torsion-free affine connection ∇ is given on a Riemannian manifold (M, g).

For a SMAT (M, g, ∇), $R = 0$ does not necessarily imply that ∇ is flat, but it implies that ∇^* is flat since $R^* = 0$ and $T^* = 0$. In this case, we call (M, g, ∇, ∇^*) a *partially flat space*.

Let ρ be a real-valued function on the direct product $TM \times M$ of a manifold M and its tangent bundle TM, and $X_1, \ldots, X_i, Y_1, \ldots, Y_j, Z$ be vector fields on M. The function $\rho[X_1, \ldots, X_i Z | Y_1, \ldots, Y_j]$ on M is defined by

$$\rho[X_1, \ldots, X_i Z | Y_1, \ldots, Y_j](r) := (X_1)_p \cdots (X_i)_p (Y_1)_q \cdots (Y_j)_q \rho(Z_p, q)|_{p=r, q=r}$$

for any $r \in M$. Note that the role of Z is different from those of the vector fields in the notation of (2). The functions $\rho[X_1, \ldots, X_i Z|\ ]$ and $\rho[\ |Y_1, \ldots, Y_j]$ are also defined in the similar way to (3) and (4).

We say that ρ is a *pre-contrast function* on M if and only if the following conditions are satisfied [6, 7]:

(a) $\rho(f_1 X_1 + f_2 X_2, q) = f_1 \rho(X_1, q) + f_2 \rho(X_2, q)$
$(\forall f_1, \forall f_2 \in C^\infty(M),\ \forall X_1, \forall X_2 \in \mathscr{X}(M),\ \forall q \in M)$.

(b) $\rho[X|\ ] = 0 \ \ (\forall X \in \mathscr{X}(M)) \quad \big(i.e.\ \rho(X_p, p) = 0 \ \ (\forall p \in M)\big)$.

(c) $g(X, Y) := -\rho[X|Y] \ \ (\forall X, \forall Y \in \mathscr{X}(M))$ is a Riemannian metric on M.

Note that for any contrast function ϕ on M, the function ρ_ϕ which is defined by

$$\rho_\phi(X_p, q) := X_p \phi(p, q) \quad (\forall p, \forall q \in M,\ \forall X_p \in T_p(M))$$

is a pre-contrast function on M. The notion of pre-contrast function is obtained by taking the fundamental properties of the first derivative of a contrast function as axioms. For a given pre-contrast function ρ, two affine connections ∇ and ∇^* are defined by

$$g(\nabla_X Y, Z) = -\rho[XY|Z], \quad g(Y, \nabla^*_X Z) = -\rho[Y|XZ]$$

$(\forall X, \forall Y, \forall Z \in \mathcal{X}(M))$ in the same way as for a contrast function. In this case, ∇ and ∇^* are dual to each other with respect to g and ∇^* is torsion-free. However, the affine connection ∇ possibly has torsion. This means that (M, g, ∇) is a SMAT and it is called the SMAT induced from the pre-contrast function ρ.

4 Canonical Pre-contrast Functions in Partially Flat Spaces

In a dually flat space (M, g, ∇, ∇^*), it is well-known that the canonical contrast functions (called ∇ and ∇^*- divergences) are naturally defined, and the Pythagorean theorem and the projection theorem are stated in terms of the ∇ and ∇^*- geodesics and the canonical contrast functions [3, 4]. In a partially flat space (M, g, ∇, ∇^*), where $R = R^* = 0$ and $T^* = 0$, it is possible to define a pre-contrast function which can be seen as canonical, and a projection theorem holds with respect to the "canonical" pre-contrast function and the ∇^*-geodesic.

Proposition 1 (Canonical Pre-contrast Functions) *Let (M, g, ∇, ∇^*) be a partially flat space (i.e. (M, g, ∇) is a SMAT with $R = R^* = 0$ and $T^* = 0$) and (U, η_i) be an affine coordinate neighborhood with respect to ∇^* in M. The function ρ on $TU \times U$ defined by*

$$\rho(Z_p, q) := -g_p(Z_p, \dot{\gamma}^*(0)) \quad (\forall p, \forall q \in U, \forall Z_p \in T_p(U)), \tag{6}$$

is a pre-contrast function on U, where $\gamma^* : [0, 1] \to U$ is the ∇^*-geodesic such that $\gamma^*(0) = p, \gamma^*(1) = q$ and $\dot{\gamma}^*(0)$ is the tangent vector of γ^* at p. Furthermore, the pre-contrast function ρ induces the original Riemannian metric g and the dual connections ∇ and ∇^* on U.

Proof For the function ρ defined as (6), the condition (a) in the definition of pre-contrast functions follows from the bilinearity of the inner product g_p. The condition (b) immediately follows from $\dot{\gamma}^*(0) = 0$ when $p = q$. By calculating the derivatives of ρ with the affine coordinate system (η_i), it can be shown that the condition (c) holds and that the induced Riemannian metric and dual affine connections coincide with the original g, ∇ and ∇^*. □

In particular, if (U, g, ∇, ∇^*) is a dually flat space, the pre-contrast function ρ defined in (6) coincides with the directional derivative $Z_p\phi^*(\cdot, q)$ of ∇^*-divergence $\phi^*(\cdot, q)$ with respect to Z_p (cf. [10, 11]). Hence, the definition of (6) seems to be natural one and we call the function ρ in (6) the *canonical pre-contrast function* in a partially flat space (U, g, ∇, ∇^*).

From the definition of the canonical pre-contrast function, we can immediately obtain the following theorem.

Corollary 1 (Generalized Projection Theorem) *Let (U, η_i) be an affine coordinate neighborhood in a partially flat space (M, g, ∇, ∇^*) and ρ be the canonical pre-contrast function on U. For any submanifold N in U, the following conditions are*

equivalent:

(i) *The* $\nabla^* -$ *geodesic starting at* $q \in U$ *is perpendicular to* $Nat\ p \in N$.
(ii) $\rho(Z_p, q) = 0$ *for any* Z_p *in* $T_p(N)$.

If (U, g, ∇, ∇^*) is a dually flat space, this theorem reduces to the projection theorem with respect to the ∇^*-divergence ϕ^*, since $\rho(Z_p, q) = Z_p\phi^*(p, q)$. In this sense, it can be seen as a generalized version of the projection theorem in dually flat spaces, and this is also one of the reasons why we consider the pre-contrast function ρ defined in (6) as canonical.

5 Statistical Manifolds Admitting Torsion Induced from Estimating Functions

As we mentioned in Introduction, a SMAT naturally appears through an estimating function in a "classical" statistical model as well as in a quantum statistical model. In this section, we briefly explain how a SMAT is induced on a parametric statistical model from an estimating function. See [7] for more details.

Let $S = \{p(\boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta^1, \ldots, \theta^d) \in \Theta \subset \boldsymbol{R}^d\}$ be a regular parametric statistical model. An estimating function on S, which we consider here, is a $\boldsymbol{R}^d$-valued function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ satisfying the following conditions:

$$E_{\boldsymbol{\theta}}\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\} = \boldsymbol{0}, \quad E_{\boldsymbol{\theta}}\{\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\|^2\} < \infty, \quad \det\left[E_{\boldsymbol{\theta}}\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta})\right\}\right] \neq 0 \ \ (\forall \boldsymbol{\theta} \in \Theta).$$

The first condition is called the unbiasedness of estimating functions, which is important to ensure the consistency of the estimator obtained from an estimating function. Let $X_1, \ldots, X_n$ be a random sample from an unknown probability distribution $p(\boldsymbol{x}; \boldsymbol{\theta}_0)$ in S. The estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ is called an M-estimator if it is obtained as a solution to the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \boldsymbol{0}. \tag{7}$$

The M-estimator $\hat{\boldsymbol{\theta}}$ has the consistency

$$\hat{\boldsymbol{\theta}} \longrightarrow \boldsymbol{\theta}_0 \ \ \text{(in probability)}$$

as $n \to \infty$ and the asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N\left(\boldsymbol{0}, \mathrm{Avar}\left(\hat{\boldsymbol{\theta}}\right)\right) \ \ \text{(in distribution)}$$

as $n \to \infty$ under some additional regularity conditions [12], which are also assumed in the following discussion. The matrix $\mathrm{Avar}(\hat{\boldsymbol{\theta}})$ is the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ and is given by

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \{A(\boldsymbol{\theta}_0)\}^{-1} B(\boldsymbol{\theta}_0) \{A(\boldsymbol{\theta}_0)\}^{-T}, \tag{8}$$

where $A(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}} \{(\partial \boldsymbol{u}/\partial \boldsymbol{\theta})(\boldsymbol{x}, \boldsymbol{\theta})\}$, $B(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}} \left\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}$ and $-T$ means transposing an inverse matrix (or inverting a transposed matrix).

In order to induce the structure of SMAT on S from an estimating function, we consider the notion of *standardization* of estimating functions. For an estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$, its standardization (or *standardized estimating function*) is defined by

$$\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta}) := E_{\boldsymbol{\theta}} \left\{\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^T\right\} \left[E_{\boldsymbol{\theta}} \left\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}\right]^{-1} \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}),$$

where $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta}) = (\partial/\partial \boldsymbol{\theta}) \log p(\boldsymbol{x}; \boldsymbol{\theta})$ is the score function for $\boldsymbol{\theta}$ [13]. Geometrically, the i-th component of the standardized estimating function $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ is the orthogonal projection of the i-th component of the score function $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})$ onto the linear space spanned by all components of the estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ in the Hilbert space

$$\mathscr{H}_{\boldsymbol{\theta}} := \{a(\boldsymbol{x}) \mid E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})\} = 0,\ E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})^2\} < \infty\}$$

with the inner product $< a(\boldsymbol{x}), b(\boldsymbol{x}) >_{\boldsymbol{\theta}} := E_{\boldsymbol{\theta}}\{a(\boldsymbol{x}) b(\boldsymbol{x})\}$ $(\forall a(\boldsymbol{x}), \forall b(\boldsymbol{x}) \in \mathscr{H}_{\boldsymbol{\theta}})$. The standardization $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ of $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ does not change the estimator since the estimating equation obtained from $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ is equivalent to the original estimating equation (7). In terms of the standardization, the asymptotic variance-covariance matrix (8) can be rewritten as

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \{G(\boldsymbol{\theta}_0)\}^{-1},$$

where $G(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}} \left\{\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta}) \boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})^T\right\}$. The matrix $G(\boldsymbol{\theta})$ is called a Godambe information matrix [14], which can be seen as a generalization of the Fisher information matrix.

As we have seen in Sect. 2, the Kullback–Leibler divergence ϕ_{KL} is a contrast function on S. Hence, the first derivative of ϕ_{KL} is a pre-contrast function on S and given by

$$\rho_{KL}((\partial_j)_{p_1}, p_2) := (\partial_j)_{p_1} \phi_{KL}(p_1, p_2) = -\int_{\Omega} s^j(\boldsymbol{x}, \boldsymbol{\theta}_1) p(\boldsymbol{x}; \boldsymbol{\theta}_2) \nu(d\boldsymbol{x})$$

for any two probability distributions $p_1(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_2)$ in S and $j = 1, \ldots, d$. This observation leads to the following proposition [7].

Proposition 2 (Pre-contrast Functions from Estimating Functions) *For an estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ on the parametric model S, a pre-contrast function $\rho_{\boldsymbol{u}}$:*

$TS \times S \to \boldsymbol{R}$ is defined by

$$\rho_{\boldsymbol{u}}((\partial_j)_{p_1}, p_2) := -\int_{\Omega} u_*^j(\boldsymbol{x}, \boldsymbol{\theta}_1) p(\boldsymbol{x}; \boldsymbol{\theta}_2) \nu(d\boldsymbol{x}) \tag{9}$$

for any two probability distributions $p_1(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_2)$ in S and $j = 1, \ldots, d$, where $u_*^j(\boldsymbol{x}, \boldsymbol{\theta})$ is the j-th component of the standardization $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ of $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$.

The use of the standardization $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ instead of $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ ensures that the definition of the function $\rho_{\boldsymbol{u}}$ does not depend on the choice of coordinate system (parameter) of S. In fact, for a coordinate transformation (parameter transformation) $\boldsymbol{\eta} = \boldsymbol{\Phi}(\boldsymbol{\theta})$, the estimating function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ is changed into $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\eta}) = \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}))$ and we have

$$\boldsymbol{v}_*(\boldsymbol{x}, \boldsymbol{\eta}) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}}\right)^T \boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta}).$$

This is the same as the transformation rule of coordinate bases on a tangent space of a manifold. The set of all components of the standardized estimating function $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ can be seen as a representation of the coordinate basis $\{(\partial_1)_p, \ldots, (\partial_d)_p\}$ on the tangent space $T_p(S)$ of S, where $p(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta})$.

The proof of Proposition 2 is straightforward. In particular, the condition (b) in the definition of pre-contrast function follows from the unbiasedness of the (standardized) estimating function. The Riemannian metric g, dual connections ∇ and ∇^* induced from the pre-contrast function $\rho_{\boldsymbol{u}}$ are given as follows:

$$g_{jk}(\boldsymbol{\theta}) := g(\partial_j, \partial_k) = E_{\boldsymbol{\theta}}\{u_*^j(\boldsymbol{x}, \boldsymbol{\theta}) u_*^k(\boldsymbol{x}, \boldsymbol{\theta})\} = G(\boldsymbol{\theta})_{jk},$$
$$\begin{cases} \Gamma_{ij,k}(\boldsymbol{\theta}) := g(\nabla_{\partial_i} \partial_j, \partial_k) = E_{\boldsymbol{\theta}}[\{\partial_i u_*^j(\boldsymbol{x}, \boldsymbol{\theta})\} s^k(\boldsymbol{x}, \boldsymbol{\theta})] \\ \Gamma^*_{ik,j}(\boldsymbol{\theta}) := g(\partial_j, \nabla^*_{\partial_i} \partial_k) = \int_{\Omega} u_*^j(\boldsymbol{x}, \boldsymbol{\theta}) \partial_i \partial_k p(\boldsymbol{x}; \boldsymbol{\theta}) \nu(d\boldsymbol{x}) \end{cases},$$

where $G(\boldsymbol{\theta})_{jk}$ is the (j, k) component of the Godambe information matrix $G(\boldsymbol{\theta})$. Note that ∇^* is always torsion-free since $\Gamma^*_{ik,j} = \Gamma^*_{ki,j}$, whereas ∇ is not necessarily torsion-free unless $\boldsymbol{u}_*(\boldsymbol{x}, \boldsymbol{\theta})$ is integrable with respect to $\boldsymbol{\theta}$ (*i.e.* there exists a function $\psi(\boldsymbol{x}, \boldsymbol{\theta})$ satisfying $\partial_j \psi(\boldsymbol{x}, \boldsymbol{\theta}) = u_*^j(\boldsymbol{x}, \boldsymbol{\theta})$ $(j = 1, \ldots, d)$).

If it is integrable and ∇ is torsion-free, it is possible to construct a contrast function on S, from which the pre-contrast function $\rho_{\boldsymbol{u}}$ in (9) is obtained by taking its first derivative, as follows:

$$\phi_{\boldsymbol{u}}(p_1, p_2) = \int_{\Omega} \{\psi(\boldsymbol{x}, \boldsymbol{\theta}_1) - \psi(\boldsymbol{x}, \boldsymbol{\theta}_2)\} p(\boldsymbol{x}; \boldsymbol{\theta}_2) \nu(d\boldsymbol{x}),$$

where $\partial_j \psi(\boldsymbol{x}, \boldsymbol{\theta}) = u_*^j(\boldsymbol{x}, \boldsymbol{\theta})$ $(j = 1, \ldots, d)$ and $p_l(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_l)$ $(l = 1, 2)$.

Table 1 Votes cast in the n-th constituency ($n = 1, \ldots, N$)

Party	C	L	Total
C	X_{1n}	$m_{1n} - X_{1n}$	m_{1n}
L	X_{2n}	$m_{2n} - X_{2n}$	m_{2n}
Total	X_n	$m_n - X_n$	m_n

6 Example

In this section, we consider the estimation problem of voter transition probabilities described in [15] to see an example of statistical manifolds admitting torsion (SMAT) induced from estimation functions.

Suppose that we had two successive elections which were carried out in N constituencies, and that the two political parties C and L contended in each election. The table below summarizes the numbers of voters in the n-th constituency for the respective elections. It is assumed that we can observe only the marginal totals m_{1n}, m_{2n}, X_n and $m_n - X_n$, where X_n is a random variable and the others are treated as fixed constants. Let θ^1 and θ^2 be the probabilities that a voter who votes for the parties C and L in Election 1 votes for C in Election 2, respectively. They are the parameters of interest here. Then, the random variables X_{1n} and X_{2n} in Table 1 are assumed to independently follow the binomial distributions $B(m_{1n}, \theta^1)$ and $B(m_{2n}, \theta^2)$, respectively.

In the n-th constituency, the probability function of the observation $X_n = X_{1n} + X_{2n}$ is given by

$$p_n(x_n; \boldsymbol{\theta}) = \sum_{x_{1n}=0}^{m_{1n}} \binom{m_{1n}}{x_{1n}} \binom{m_{2n}}{x_n - x_{1n}} \left(\theta^1\right)^{x_{1n}} \left(1-\theta^1\right)^{m_{1n}-x_{1n}} \left(\theta^2\right)^{x_n - x_{1n}} \left(1-\theta^2\right)^{m_{2n}-x_n+x_{1n}},$$

where $\boldsymbol{\theta} = \left(\theta^1, \theta^2\right)$. The statistical model S in this problem consists of all possible probability functions of the observed data $\boldsymbol{X} = (X_1, \ldots, X_N)$ as follows:

$$S = \left\{ p(\boldsymbol{x}; \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta} = \left(\theta^1, \theta^2\right) \in (0,1) \times (0,1) \right\},$$

where $p(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{n=1}^{N} p_n(x_n; \boldsymbol{\theta})$ $(\boldsymbol{x} = (x_1, \ldots, x_N))$ since $X_1, \ldots, X_N$ are independent.

Although the maximum likelihood estimation for $\boldsymbol{\theta}$ is possible based on the likelihood function $L(\boldsymbol{\theta}) = p(\boldsymbol{X}; \boldsymbol{\theta})$, it is a little complicated since X_{1n} and X_{2n} are not observed in each n-th constituency. An alternative approach for estimating $\boldsymbol{\theta}$ is to use the quasi-score function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta}) = (q^1(\boldsymbol{x}, \boldsymbol{\theta}), q^2(\boldsymbol{x}, \boldsymbol{\theta}))^T$ [15] as an estimating function, where

$$q^1(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{m_{1n}\{x_n - \mu_n(\boldsymbol{\theta})\}}{V_n(\boldsymbol{\theta})}, \quad q^2(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{m_{2n}\{x_n - \mu_n(\boldsymbol{\theta})\}}{V_n(\boldsymbol{\theta})}.$$

Here, $\mu_n(\boldsymbol{\theta})$ and $V_n(\boldsymbol{\theta})$ are the mean and variance of X_n, respectively, i.e.

$$\begin{aligned}\mu_n(\boldsymbol{\theta}) &= E(X_n) = m_{1n}\theta^1 + m_{2n}\theta^2 \\ V_n(\boldsymbol{\theta}) &= V(X_n) = m_{1n}\theta^1\left(1-\theta^1\right) + m_{2n}\theta^2\left(1-\theta^2\right).\end{aligned} \tag{10}$$

In this example, the random variables $X_1, \ldots, X_N$ in the observed data are independent, but not identically distributed. However, it is possible to apply the results in Sect. 5 by considering the whole of the left hand side of (7) as an estimating function and modifying the results in this case. Note that the estimating function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ is already standardized since the i-th component $q^i(\boldsymbol{x}, \boldsymbol{\theta})$ of $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ is obtained by the orthogonal projection of the i-th component of the score function $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})$ for $\boldsymbol{\theta}$ onto the linear space spanned by $\{x_1 - \mu_1(\boldsymbol{\theta}), \ldots, x_N - \mu_N(\boldsymbol{\theta})\}$. In fact, the orthogonal projection is calculated as follows:

$$\begin{aligned}&E_{\boldsymbol{\theta}}\left\{\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T\right\}\left[E_{\boldsymbol{\theta}}\left\{(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T\right\}\right]^{-1}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) \\ &= -E_{\boldsymbol{\theta}}\left\{\frac{\partial}{\partial \boldsymbol{\theta}^T}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right\}\left[E_{\boldsymbol{\theta}}\left\{(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T\right\}\right]^{-1}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) \\ &= \begin{pmatrix} m_{11} & \cdots & m_{1N} \\ m_{21} & \cdots & m_{2N} \end{pmatrix}\begin{pmatrix} V_1(\boldsymbol{\theta}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_n(\boldsymbol{\theta}) \end{pmatrix}^{-1}\begin{pmatrix} x_1 - \mu_1(\boldsymbol{\theta}) \\ \vdots \\ x_N - \mu_N(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} q^1(\boldsymbol{x}, \boldsymbol{\theta}) \\ q^2(\boldsymbol{x}, \boldsymbol{\theta}) \end{pmatrix},\end{aligned}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \ldots, \mu_N(\boldsymbol{\theta}))^T$. In addition, the estimating function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ is not integrable with respect to $\boldsymbol{\theta}$ since $\partial q^1/\partial\theta^2 \neq \partial q^2/\partial\theta^1$. From Proposition 2 and the fact that $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$ itself is a standardized estimating function, we immediately obtain the pre-contrast function $\rho_q : TS \times S \to \boldsymbol{R}$ defined by $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$, where

$$\rho_q((\partial_i)_{p_1}, p_2) = -\sum_{\boldsymbol{x}} q^i(\boldsymbol{x}, \boldsymbol{\theta}_1)p(\boldsymbol{x}; \boldsymbol{\theta}_2) = \sum_{n=1}^{N}\frac{m_{in}\{\mu_n(\boldsymbol{\theta}_1) - \mu_n(\boldsymbol{\theta}_2)\}}{V_n(\boldsymbol{\theta}_1)}$$

with $p_l(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}_l) \in S$ $(l = 1, 2)$. The pre-contrast function ρ_q induces the statistical manifold admitting torsion as follows.
Riemannian metric g:

$$g_{ij}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} q^i(\boldsymbol{x}, \boldsymbol{\theta})q^j(\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{n=1}^{N}\frac{1}{V_n(\boldsymbol{\theta})}m_{in}m_{jn}.$$

Dual affine connections ∇^* and ∇:

$$\begin{aligned}\Gamma^*_{ij,k}(\boldsymbol{\theta}) &= \sum_{\boldsymbol{x}}\{\partial_i\partial_j p(\boldsymbol{x}; \boldsymbol{\theta})\}q^k(\boldsymbol{x}, \boldsymbol{\theta}) \\ &= \sum_{n=1}^{N}\frac{m_{kn}}{V_n(\boldsymbol{\theta})}\left[\sum_{\boldsymbol{x}} x_n\{\partial_i\partial_j p(\boldsymbol{x}; \boldsymbol{\theta})\} - \mu_n(\boldsymbol{\theta})\sum_{\boldsymbol{x}}\{\partial_i\partial_j p(\boldsymbol{x}; \boldsymbol{\theta})\}\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} \frac{m_{kn}}{V_n(\boldsymbol{\theta})} \left[\partial_i \partial_j \sum_{\boldsymbol{x}} x_n p(\boldsymbol{x}; \boldsymbol{\theta}) - \mu_n(\boldsymbol{\theta}) \partial_i \partial_j \sum_{\boldsymbol{x}} p(\boldsymbol{x}; \boldsymbol{\theta}) \right] \\
&= \sum_{n=1}^{N} \frac{m_{kn}}{V_n(\boldsymbol{\theta})} \partial_i \partial_j \mu_n(\boldsymbol{\theta}) = 0 \quad \text{(from (10))} \\
\Gamma_{ij,k}(\boldsymbol{\theta}) &= \Gamma^*_{ij,k}(\boldsymbol{\theta}) - \partial_i g_{jk}(\boldsymbol{\theta}) \quad \text{(from the duality between } \nabla \text{ and } \nabla^*\text{)} \\
&= \sum_{n=1}^{N} \frac{1 - 2\theta^i}{V_n(\boldsymbol{\theta})^2} m_{in} m_{jn} m_{kn}.
\end{aligned}
$$

In this example, the statistical model S is ∇^*-flat since the coefficient of ∇^* with respect to the parameter $\boldsymbol{\theta}$ is equal to zero. Furthermore, this shows that $\boldsymbol{\theta}$ provides an affine coordinate system for ∇^*. Although the curvature tensor of ∇ vanishes because the curvature tensor of ∇^* vanishes and ∇ is dual to ∇^*, the statistical model S is not ∇-flat because ∇ is not torsion-free, which comes from the non-integrability of the estimating function $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\theta})$. Hence, this geometrical structure provides an example of partially flat spaces, which was discussed in Sect. 4.

7 Future Problems

In this paper, we have summarized existing results on statistical manifolds admitting torsion, especially focusing on partially flat spaces. Although some results that are not seen in the standard theory of information geometry have been obtained, including a generalized projection theorem in partially flat spaces and statistical manifolds admitting torsion induced from estimating functions in statistics, a lot of (essential) problems have been unsolved. We discuss some of them to conclude this paper.

(1) The canonical pre-contrast function and the generalized projection theorem in a partially flat spase (M, g, ∇, ∇^*) are described only in terms of the flat connection ∇^*. In this sense, it can be said that these are a concept and a theorem for the Riemannian manifold (M, g) with the flat connection ∇^*. What is the role of the affine connection ∇ in the partially flat space (M, g, ∇, ∇^*), especially when ∇ is not torsion-free?

(2) The canonical pre-contrast function is defined in terms of the Riemannian metric g and the ∇^*-geodesic in a partially flat space (U, g, ∇, ∇^*) *without* using the affine coordinate system (η_i) on U. Hence, this function can be defined in a general statistical manifold admitting torsion (M, g, ∇) as long as the ∇^*-geodesic uniquely exists. What is the condition under which this function is a pre-contrast function that induces the original Riemannian metric g, dual affine connections ∇ and ∇^*? What properties does the (canonical) pre-contrast function have in this case? These problems are closely related to the works by [10, 11], who try to define a canonical divergence (canonical contrast function) on a general statistical manifold beyond a dually flat space.

(3) The definition of pre-contrast functions from estimating functions is obtained by replacing the score function which appears in the pre-contrast function as the derivative of Kullback–Leibler divergence with the standardized estimating functions. However, this is not the unique way to obtain a pre-contrast function from an estimating function. For example, if we consider the β-divergence [16] (or density power divergence [17]) as a contrast function, its first derivative is also a pre-contrast function and takes the same form as (9) in Proposition 2. However, the estimating function which appears in the pre-contrast function is not standardized. Although the standardization seems to be natural, further consideration is necessary on how to define a pre-contrast function from a given estimating function.
(4) For the example considered in Sect. 6, we can show that the pre-contrast function $\rho_{\boldsymbol{q}}$ coincides with the canonical pre-contrast function in the partially flat space (S, g, ∇, ∇^*) and the generalized projection theorem (Corollary 1 in Sect. 4) can be applied. However, its statistical meaning has not been clarified yet. Although it is expected that the SMAT induced from an estimating function has something to do with statistical inference based on the estimating function, the clarification on it is a future problem.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers JP15K00064, JP15K04842.

References

1. Eguchi, S.: Geometry of minimum contrast. Hiroshima Math. J. **22**, 631–647 (1992)
2. Matsuzoe, H.: Geometry of contrast functions and conformal geometry. Hiroshima Math. J. **29**, 175–191 (1999)
3. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence; Oxford University Press, Oxford (2000)
4. Amari, S.: Information Geometry and Its Applications. Springer, Berlin (2016)
5. Kurose, T.: Statistical manifolds admitting torsion. Geometry and Something, Fukuoka University (2007)
6. Matsuzoe, H.: Statistical manifolds admitting torsion and pre-contrast functions. Information Geometry and Its Related Fields, Osaka City University (2010)
7. Henmi, M., Matsuzoe, H.: Geometry of pre-contrast functions and non-conservative estimating functions. AIP Conf. Proc. **1340**, 32–41 (2011)
8. Henmi, M.: Statistical manifolds admitting torsion, pre-contrast functions and estimating functions. Lect. Notes Comput. Sci. **10589**, 153–161 (2017)
9. Kurose, T.: On the divergences of 1-conformally flat statistical manifolds. Tohoku Math. J. **46**, 427–433 (1994)
10. Henmi, M., Kobayashi, R.: Hooke's law in statistical manifolds and divergences. Nagoya Math. J. **159**, 1–24 (2000)
11. Ay, N., Amari, S.: A novel approach to canonical divergences within information geometry. Entropy **17**, 8111–8129 (2015)
12. van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (2000)
13. Heyde, C.C.: Quasi-Likelihood and Its Application. Springer, Berlin (1997)
14. Godambe, V.: An optimum property of regular maximum likelihood estimation. Ann. Math. Stat. **31**, 1208–1211 (1960)

15. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, Boca Raton (1989)
16. Eguchi, S., Kano, Y.: Robustifying maximum likelihood estimation. In: Research Memorandum of the Institute of Statistical Mathematics, vol. 802 (2001)
17. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimizing a density power divergence. Biometrika **85**, 549–559 (1998)