Efron's curvature of the structural gradient model

Tomonari SEI

September 23, 2010

Abstract

The structural gradient model is a multivariate statistical model in order to extract various interactions of given data set. In this note, we show that Efron's statistical curvature of the structural gradient model is less than that of a competitive mixture model under a null hypothesis.

1 Introduction

Exponential families are important in statistical modeling. For example, the Gaussian family and its subfamilies are often used in multivariate analysis, time-series analysis, geostatistics and any other areas that deal with quantitative data. Using the exponential family is reasonable because it is derived from the maximum entropy criterion (see e.g. Cover and Thomas (2006)). It is also compatible with regression problem, that is, the generalized linear models (McCullagh and Nelder (1989)). A comprehensive book on exponential families is Barndorff-Nielsen (1978).

A drawback of exponential families is that the probability density function is sometimes not explicitly expressed due to the normalizing constant. For example, if one would try to find three-dimensional interaction of given data, a corresponding exponential family is not available in explicit form. Although the Markov Chain Monte Carlo procedure is available, it requires some adjustment for convergence. As an attempt to overcome the difficulty, Sei (2010) suggested a new parametric family called a structural gradient model (SGM) for multivariate quantitative data. SGM is numerically shown to have a desirable performance for such a purpose. However, it is not known whether SGM is close to an exponential family or not. In this paper, we give a partial answer to this problem.

A measure of closeness to an exponential family is Efron's statistical curvature γ^2, refered to *the Efron curvature* below. It is defined in terms of the second-order derivative of the log-likelihood function. See Section 2 for the precise definition. Efron (1975) showed that information loss of the maximum likelihood estimator is

asymptotically expanded as $\gamma^2 + o(1)$ if the sample size N goes to infinity. It is known that γ^2 vanishes if the model is an exponential family. Furthermore, γ^2 is an intrinsic quantity, that is, independent of the parameterization of the model.

Consider two statistical models M_1 and M_2, and assume that they have a common density p_0 and a common score vector at p_0. The Fisher information matrix at p_0 is common in both models. Then we can say that, without subjectivity, the model M_1 is closer to exponential family at p_0 than M_2 if the Efron curvature of M_1 is smaller than M_2.

We compare the Efron curvature of SGM and MixM, which is a competitive model with SGM. MixM is an abbreviation of *the structural mixture model.* Here we briefly describe SGM and MixM. For details, refer to Section 3 and Sei (2010). SGM is a statistical model on hypercube represented by Fourier-expanded optimal transport between the target density and the uniform density. Here the Fourier coefficients are the unknown parameter. The model is related to the optimal transport theory. See Villani (2003) and Villani (2009) for the optimal transport theory. MixM is represented by Fourier expansion of the probability density function itself. Both SGM and MixM do not need computation of normalizing constants, in contrast to the exponential family. We show that the curvature of SGM is less than MixM under the common null hypothesis. In other words, SGM is closer to exponential family than MixM. This motivates to use SGM rather than MixM for analyzing complicated dependency of given data.

The paper is organized as follows. We recall the definition of the Efron curvature in Section 2 and define SGM and MixM in Section 3. Then we state the main result of this paper in Section 4. We give some discussion in Section 5. Proofs are given in Section 6.

2   Efron's statistical curvature

We recall the Efron curvature of a general statistical model according to Efron (1975), Reeds (1975) and Amari (1985). Intuitively, the Efron curvature is the residual when the second derivative of the log-likelihood is projected onto the linear space spanned by the score functions and the constant function.

Consider a parametric family of density functions $p(x|\theta)$ with respect to a base measure dx indexed by a parameter vector $\theta = (\theta_u)_{u \in \mathcal{U}}$, where $\mathcal{U}$ is a finite set. Typically $\mathcal{U} = \{1, \ldots, d\}$ with some $d \geq 1$, but we will consider other case in the next section. The parameter space Θ of θ is an open subset of $\mathbb{R}^{\mathcal{U}}$, where $\mathbb{R}^{\mathcal{U}}$ denotes the set of all real vectors $(\theta_u)_{u \in \mathcal{U}}$ indexed by $\mathcal{U}$. Without loss of generality, we assume

$0 \in \Theta$ and define the curvature at $\theta = 0$.

Denote the first and second derivative of the log-likelihood function by

$$L_u = L_u(x) = \left. \frac{\partial}{\partial \theta_u} \log p(x|\theta) \right|_{\theta=0},$$

$$L_{uv} = L_{uv}(x) = \left. \frac{\partial^2}{\partial \theta_u \partial \theta_v} \log p(x|\theta) \right|_{\theta=0}$$

for $u, v \in \mathcal{U}$. Define the Fisher information $(J_{uv})_{u,v \in \mathcal{U}}$ and the e-connection coefficients $(\Gamma_{uv,w})_{u,v,w \in \mathcal{U}}$ and $(\Gamma_{uv}^w)_{u,v,w \in \mathcal{U}}$ by

$$J_{uv} = \int p(x|0) L_u L_v \mathrm{d}x, \quad \Gamma_{uv,w} = \int p(x|0) L_{uv} L_w \mathrm{d}x, \quad \Gamma_{uv}^w = \sum_{s \in \mathcal{U}} \Gamma_{uv,s} J^{sw},$$

where (J^{sw}) is the inverse matrix of (J_{sw}). We define a fourth-order tensor by

$$Q_{uv,wz} = \int p(x|0) \left(L_{uv} + J_{uv} - \sum_{s \in \mathcal{U}} \Gamma_{uv}^s L_s \right) \left(L_{wz} + J_{wz} - \sum_{t \in \mathcal{U}} \Gamma_{wz}^t L_t \right) \mathrm{d}x.$$

Finally, we define the Efron curvature by

$$\gamma^2 = \sum_{u,v,w,z \in \mathcal{U}} Q_{uv,wz} J^{uw} J^{vz}. \tag{1}$$

The Efron curvature is a non-negative scalar quantity independent of parameterization of $p(x|\theta)$.

The Efron curvature is related to the exponential family and information loss as stated in Section 1. Precise statements are as follows. Recall that a statistical model $p(x|\theta)$ is called an exponential family (in canonical form) if it is written as $p(x|\theta) = \exp(\sum_{u \in \mathcal{U}} \theta_u t_u(x) - \psi(\theta))$ with the sufficient statistics $t_u(x)$ and the normalizing function $\psi(\theta)$.

Lemma 1. Let Θ be an open subset of $\mathbb{R}^{\mathcal{U}}$. Then the Efron curvature vanishes over Θ if and only if $p(x|\theta)$ is an exponential family.

Lemma 2. Let $(x_1, \ldots, x_N)$ be an i.i.d. sample from a density $p(x|\theta)$. Then, under some regularity conditions, the information loss of the maximum likelihood estimator $\hat{\theta}_N$ is asymptotically

$$J_{uv}^{(x_1,\ldots,x_N)} - J_{uv}^{\hat{\theta}_N} = \sum_{w,z} Q_{uw,vz} J^{wz} + \mathrm{o}(1)$$

as $N \to \infty$, where J_{uv}^T denotes the Fisher information matrix of a statistic T. Note that $J_{uv}^{(x_1,\ldots,x_N)} = N J_{uv}$. In particular, averaged information loss is given by

$$\sum_{u,v \in \mathcal{U}} J^{uv} \left(J_{uv}^{(x_1,\ldots,x_N)} - J_{uv}^{\hat{\theta}_N} \right) = \gamma^2 + \mathrm{o}(1).$$

For the proof, refer to Efron (1975), Reeds (1975) and Amari (1985).

3

3 SGM and MixM

We prepare some notations to define SGM and MixM. Let m be a positive integer. Denote the gradient operator and Hessian operator on $\mathbb{R}^m$ by $D = (\partial/\partial x_i)_{i=1}^m$ and $D^2 = (\partial^2/\partial x_i \partial x_j)_{i,j=1}^m$, respectively. The determinant and trace of a square matrix A are denoted by $\det A$ and $\mathrm{tr} A$, respectively. For square matrices A and B, if $A - B$ is non-negative definite, we write $A \succeq B$. Let $\mathbb{Z}$ and $\mathbb{Z}_{\geq 0}$ be the set of all integers and all non-negative integers, respectively. Let $(\mathbb{Z}_{\geq 0}^m)^+ = \mathbb{Z}_{\geq 0}^m \setminus \{0\}$ be the set of all m-dimensional non-negative integer vectors except for zero vector. Define $\|u\| = (\sum_{j=1}^m u_j^2)^{1/2}$ for $u \in \mathbb{Z}^m$. The vectors are considered as column vectors unless otherwise stated.

We give the definition of SGM and MixM. Examples are given later.

Definition 1 (SGM)**.** Let $\mathcal{U}$ be a finite subset of $(\mathbb{Z}_{\geq 0}^m)^+$. *The structural gradient model* (SGM) *is a set of probability densities on the hypercube* $[0,1]^m$ *with parameter vector* $\theta = (\theta_u) \in \mathbb{R}^{\mathcal{U}}$ *defined by*

$$p^{(\mathrm{sgm})}(x|\theta) \ = \ \det(D^2 \psi(x|\theta)), \quad \psi(x|\theta) \ = \ \frac{1}{2} x^\top x - \sum_{u \in \mathcal{U}} \frac{\theta_u}{\pi^2} \prod_{j=1}^m \cos(\pi u_j x_j). \quad (2)$$

The parameter vector θ is said to be feasible if $D^2 \psi(x|\theta) \succeq 0$ for every $x \in [0,1]^m$.

Definition 2 (MixM)**.** Under the same notation as SGM, define

$$p^{(\mathrm{mix})}(x|\theta) \ = \ 1 + \sum_{u \in \mathcal{U}} \theta_u \|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j). \quad (3)$$

The set of $p^{(\mathrm{mix})}(x|\theta)$ is called MixM in this paper. The parameter vector θ is feasible if $p^{(\mathrm{mix})}(x|\theta) \geq 0$ for all $x \in [0,1]^m$.

Remark that both $p^{(\mathrm{sgm})}(x|\theta = 0)$ and $p^{(\mathrm{mix})}(x|\theta = 0)$ are the uniform density. Define a matrix $H_u(x)$ by

$$H_u(x) \ := \ D^2 \left(-\pi^{-2} \prod_{j=1}^m \cos(\pi u_j x_j) \right). \quad (4)$$

In particular,

$$\mathrm{tr}\, H_u(x) = \|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j).$$

Then we can rewrite (2) and (3) as

$$p^{(\mathrm{sgm})}(x|\theta) = \det \left(I + \sum_{u \in \mathcal{U}} \theta_u H_u(x) \right), \quad p^{(\mathrm{mix})}(x|\theta) = 1 + \sum_{u \in \mathcal{U}} \theta_u \mathrm{tr} H_u(x). \quad (5)$$

We state a fundamental lemma. For completeness, we prove it in Section 6. We denote the indicator function of a set A by 1_A.

Lemma 3 (Sei (2010) Lemma 3). *The score vector at $\theta = 0$ of both SGM and MixM is $(\operatorname{tr} H_u(x))_{u \in \mathcal{U}}$. The common Fisher information matrix $J = (J_{uv})_{u,v \in \mathcal{U}}$ at $\theta = 0$ is $J_{uv} = \|u\|^4 2^{-|\sigma(u)|} 1_{\{u=v\}}$, where $\sigma(u) = \{j \in \{1, \ldots, m\} \mid u_j > 0\}$ and $|\sigma(u)|$ denotes the cardinality of $\sigma(u)$. In particular, J_{uv} is diagonal.*

We give a few examples, where we write $(u_1, \ldots, u_m)$ instead of $(u_1, \ldots, u_m)^\top$ for simplicity.

Example 1. Let $m = 2$ and $\mathcal{U} = \{(1, 1)\}$. We abbreviate $\theta_{(1,1)}$ as θ for simplicity. Then we have

$$
\begin{aligned}
p^{(\mathrm{sgm})}(x|\theta) &= \det \begin{pmatrix} 1 + \theta \cos(\pi x_1) \cos(\pi x_2) & -\theta \sin(\pi x_1) \sin(\pi x_2) \\ -\theta \sin(\pi x_1) \sin(\pi x_2) & 1 + \theta \cos(\pi x_1) \cos(\pi x_2) \end{pmatrix} \\
&= 1 + 2\theta \cos(\pi x_1) \cos(\pi x_2) + \theta^2 \{\cos^2(\pi x_1) + \cos^2(\pi x_2) - 1\}
\end{aligned}
$$

and $p^{(\mathrm{mix})}(x|\theta) = 1 + 2\theta \cos(\pi x_1) \cos(\pi x_2)$. SGM is feasible if and only if $|\theta| \leq 1$. MixM is feasible if and only if $|\theta| \leq 1/2$.

Example 2. Let $m = 3$ and $\mathcal{U} = \{(1, 0, 0), (2, 0, 0), (1, 1, 0), (2, 1, 0), (1, 1, 1)\}$. Then the diagonal part $J_u := J_{uu}$ of the Fisher information matrix is

$$
J_{(1,0,0)} = \frac{1}{2}, \quad J_{(2,0,0)} = 8, \quad J_{(1,1,0)} = 1, \quad J_{(2,1,0)} = \frac{25}{4}, \quad J_{(1,1,1)} = \frac{9}{8}.
$$

4   Main result

Consider a finite subset $\mathcal{U}$ of $(\mathbb{Z}_{\geq 0}^m)^+$. Let $(\gamma_{\mathcal{U}}^2)^{(\mathrm{sgm})}$ and $(\gamma_{\mathcal{U}}^2)^{(\mathrm{mix})}$ be the Efron curvature (1) of SGM and MixM at $\theta = 0$, respectively. For each $i \in \{1, \ldots, m\}$, we set $\mathbb{Z}_i = \{u \in (\mathbb{Z}_{\geq 0}^m)^+ \mid u_j = 0 \text{ if } j \neq i\}$.

Our main result is the following theorem.

Theorem 4. *For any finite $\mathcal{U} \subset (\mathbb{Z}_{\geq 0}^m)^+$, the following inequality holds:*

$$
0 < (\gamma_{\mathcal{U}}^2)^{(\mathrm{sgm})} \leq (\gamma_{\mathcal{U}}^2)^{(\mathrm{mix})}. \tag{6}
$$

Equality holds if and only if there is some $i \in \{1, \ldots, m\}$ such that $\mathcal{U} \subset \mathbb{Z}_i$. If the equality holds, then the two models coincide.

We give more explicit expression of the two quantities. We prepare some additional notations. For a vector $U = (U_i) \in \mathbb{Z}^m$, its component-wise absolute value is denoted by $\mathrm{abs}(U) = (|U_i|)$. For two vectors $U = (U_i)$ and $V = (V_i)$, their

component-wise product (Hadamard product) is denoted by $U \circ V = (U_i V_i)$. Let $\beta = (\beta_i) \in \{-1, 1\}^m$ be a Bernoulli sequence, that is, β_i independently takes the value ± 1 with probability $1/2$ each. For a Bernoulli sequence β and a vector $u \in \mathcal{U}$ we call the vector $U = \beta \circ u$ *Bernoulli randomization* of u. The expectation with respect to U (inherited from β) is denoted as E_U. If Bernoulli randomization of two or more (possibly the same) vectors are considered, then they are assumed to be independent. Recall that $\|u\| = (\sum_{j=1}^m u_j^2)^{1/2}$ and $\sigma(u) = \{j \mid u_j > 0\}$.

The explicit expression of the Efron curvature is given in the following theorem. The inequality (6) is obtained as a corollary.

Theorem 5. The Efron curvature of SGM and MixM at $\theta = 0$ is given by

$$(\gamma_{\mathcal{U}}^2)^{(\text{sgm})} = \sum_{u,v \in \mathcal{U}} \mathrm{E}_{U,V,\tilde{U},\tilde{V}} \left[\omega_{\mathcal{U}}(U, V, \tilde{U}, \tilde{V}) 2^{|\sigma(u)| + |\sigma(v)|} \frac{(U^\top V)^2 (\tilde{U}^\top \tilde{V})^2}{\|u\|^4 \|v\|^4} \right], \quad (7)$$

$$(\gamma_{\mathcal{U}}^2)^{(\text{mix})} = \sum_{u,v \in \mathcal{U}} \mathrm{E}_{U,V,\tilde{U},\tilde{V}} \left[\omega_{\mathcal{U}}(U, V, \tilde{U}, \tilde{V}) 2^{|\sigma(u)| + |\sigma(v)|} \right], \quad (8)$$

where $U, V, \tilde{U}, \tilde{V}$ are Bernoulli randomization of u, v, u, v, respectively, and

$$\omega_{\mathcal{U}}(U, V, \tilde{U}, \tilde{V}) = 1_{\{U+V+\tilde{U}+\tilde{V}=0, \text{ abs}(U+V) \notin \mathcal{U} \cup \{0\}\}}.$$

In particular, $(\gamma_{\mathcal{U}}^2)^{(\text{sgm})}$ and $(\gamma_{\mathcal{U}}^2)^{(\text{mix})}$ are rational numbers.

Table 1 shows the Efron curvature for several specific cases of $\mathcal{U}$. Let $e_i = (1_{\{j \neq i\}})_{j=1}^m$, the i-th unit vector.

Table 1: The Efron curvature for several cases of $\mathcal{U}$.

$\mathcal{U}$	$(\gamma_{\mathcal{U}}^2)^{(\text{sgm})}$	$(\gamma_{\mathcal{U}}^2)^{(\text{mix})}$
$\{fe_i\}_{1 \leq f \leq d, 1 \leq i \leq m}$	$2^{-2}d(d+1)m$	$2^{-2}d(d+1)m + d^2 m(m-1)$
$\{e_i + e_j\}_{1 \leq i < j \leq m}$	$2^{-5}m(m-1)(m+2)$	$2^{-3}m(m-1)(2m^2 - 6m + 9)$
$\{e_i + e_{i+1}\}_{i=1}^{m-1}$	$2^{-4}(7m - 10)$	$2^{-2}(4m^2 - 3m - 5)$
$\{e_1 + e_i\}_{i=2}^m$	$2^{-5}(m-1)(3m+2)$	$2^{-2}(m-1)(6m-7)$

We end with an asymptotic property. For the first three examples in Table 1, it is easily confirmed that $(\gamma_{\mathcal{U}}^2)^{(\text{sgm})} / (\gamma_{\mathcal{U}}^2)^{(\text{mix})}$ converges to 0 as $m \to \infty$. This property holds in a more general setting. We define two sets $M(\mathcal{U})$ and $N(\mathcal{U})$ by

$$M(\mathcal{U}) = \{(u, v) \in \mathcal{U}^2 \mid u + v \notin \mathcal{U}\},$$
$$N(\mathcal{U}) = \{(u, v) \in \mathcal{U}^2 \mid \sigma(u) \cap \sigma(v) \neq \emptyset\}.$$

We denote cardinality of a set A by $|A|$.

Theorem 6. Let $\mathcal{U}_m$ be a finite subset of $(\mathbb{Z}_{\geq 0}^m)^+$ for each $m \in \{1, 2, \ldots\}$. Assume that $\max_{u \in \mathcal{U}_m} |\sigma(u)|$ is bounded over m. Further assume $|N(\mathcal{U}_m)|/|M(\mathcal{U}_m)| \to 0$ as $m \to \infty$. Then $(\gamma_{\mathcal{U}_m}^2)^{(\text{sgm})}/(\gamma_{\mathcal{U}_m}^2)^{(\text{mix})} \to 0$ as $m \to \infty$.

Let $\mu(\mathcal{U})$ be the set of maximal elements of $\mathcal{U}$, that is,

$$\mu(\mathcal{U}) = \{u \in \mathcal{U} \mid \forall v \in \mathcal{U} \setminus \{u\}, \ \exists i \in \{1, \ldots, m\} \text{ s.t. } v_i < u_i\}.$$

Corollary 7. Let $\mathcal{U}_m$ be a finite subset of $(\mathbb{Z}_{\geq 0}^m)^+$ for each $m \in \{1, 2, \ldots\}$. Assume that $\max_{u \in \mathcal{U}_m} |\sigma(u)|$ is bounded over m. Further assume $|N(\mathcal{U}_m)|/|\mu(\mathcal{U}_m)|^2 \to 0$ as $m \to \infty$. Then $(\gamma_{\mathcal{U}_m}^2)^{(\text{sgm})}/(\gamma_{\mathcal{U}_m}^2)^{(\text{mix})} \to 0$ as $m \to \infty$.

Table 2 shows the numbers $|N(\mathcal{U})|$ and $|\mu(\mathcal{U})|$ for the examples in Table 1. It is consistent with Corollary 7, that is, $|N(\mathcal{U})|/|\mu(\mathcal{U})|^2 \to 0$ only for the first three cases.

Table 2: The numbers $|N(\mathcal{U})|$ and $|\mu(\mathcal{U})|$.

$\mathcal{U}$	$|N(\mathcal{U})|$	$|\mu(\mathcal{U})|$
$\{fe_i\}_{1 \leq f \leq d, 1 \leq i \leq m}$	$d^2 m$	m
$\{e_i + e_j\}_{1 \leq i < j \leq m}$	$2^{-1} m(m-1)(2m-3)$	$2^{-1} m(m-1)$
$\{e_i + e_{i+1}\}_{i=1}^{m-1}$	$3m - 5$	$m - 1$
$\{e_1 + e_i\}_{i=2}^{m}$	$(m-1)^2$	$m - 1$

5 Discussion

We evaluated the Efron curvature of SGM and MixM (Theorem 5) and used it to show that SGM has smaller curvature than MixM (Theorem 4). Here we give some unsolved problems.

In Table 1, we listed explicit formulas of the Efron curvature for specific $\mathcal{U}$'s by using (7) and (8). It is challenging to derive formulas for more practical sets, such as

$$\mathcal{U} = \left\{u \in (\mathbb{Z}_{\geq 0}^m)^+ \mid \|u\|_1 \leq 3, \ \|u\|_\infty \leq 2\right\}, \quad \|u\|_1 = \sum_{j=1}^{m} u_j, \quad \|u\|_\infty = \max_j u_j.$$

Sei (2010) used this set to analyze multivariate datasets. For each small m, we can evaluate the curvature by direct computation. However, the computation needs exponential complexity with respect to the dimension m as long as one uses (7) and (8). Combinatorial methods may solve the problem.

We studied the *averaged* curvature γ^2. Instead, one can consider a tensor $H_{uv} := \sum_{w,z} Q_{uw,vz} J^{wz}$ appearing in Lemma 2, which is called the embedding e-curvature (Amari (1985)). Although an inequality $H_{uv}^{(\mathrm{sgm})} \preceq H_{uv}^{(\mathrm{mix})}$ is conjectured by numerical study, it could not be proved.

In this paper, we only considered the curvature at the origin $\theta = 0$. The reason that we restrict comes from two different kinds of difficulty. One is conceptual difficulty: the probability densities (and score vectors) of SGM and MixM are different except at $\theta = 0$. An approach may be to consider a local mixture model of SGM at each point θ (Marriott (2002)). The another kind of difficulty is computational one. The expression of the Efron curvature at $\theta \neq 0$ of SGM seems complicated. Even the Fisher information matrix J_{uv} is not written in elementary functions in general. However, the expression is written at least in terms of integration of multidimensional rational functions because $p(x|\theta)$ is a polynomial of θ_u and $z_j = \mathrm{e}^{\mathrm{i}\pi x_j}$. Algebraic methods on integration may be helpful.

6 Proofs

6.1 Proof of Lemma 3 and Theorem 5

We calculate the Efron curvature of SGM and MixM step-by-step.

For SGM, we denote the quantities $L_{uv}(x)$, $\Gamma_{uv,w}$, Γ_{uv}^w, $Q_{uv,wz}$, γ^2 in Section 2 by $L_{uv}^{(\mathrm{sgm})}(x)$, $\Gamma_{uv,w}^{(\mathrm{sgm})}$, $(\Gamma_{uv}^w)^{(\mathrm{sgm})}$, $Q_{uv,wz}^{(\mathrm{sgm})}$, $(\gamma^2)^{(\mathrm{sgm})}$, respectively. Similarly, for MixM, we denote $L_{uv}^{(\mathrm{mix})}(x)$, $\Gamma_{uv,w}^{(\mathrm{mix})}$, $(\Gamma_{uv}^w)^{(\mathrm{mix})}$, $Q_{uv,wz}^{(\mathrm{mix})}$, $(\gamma^2)^{(\mathrm{mix})}$. We use $L_u(x)$ and J_{uv} without superscripts because they are common in both models. Recall that a random matrix $H_u = H_u(x)$ is defined by (4).

Lemma 8. For any $u, v \in \mathcal{U}$, the following equality holds:

$$L_u(x) = \mathrm{tr}\, H_u, \quad L_{uv}^{(\mathrm{sgm})}(x) = -\mathrm{tr}(H_u H_v) \quad L_{uv}^{(\mathrm{mix})}(x) = -(\mathrm{tr}\, H_u)(\mathrm{tr}\, H_v).$$

Proof. By (5), the log-likelihood of SGM and MixM are expanded around $\theta = 0$ as

$$\log p^{(\mathrm{sgm})}(x|\theta) = \sum_{u \in \mathcal{U}} \theta_u \, \mathrm{tr}\, H_u - \frac{1}{2} \sum_{u,v \in \mathcal{U}} \theta_u \theta_v \, \mathrm{tr}(H_u H_v) + \mathrm{O}(\|\theta\|^3),$$

$$\log p^{(\mathrm{mix})}(x|\theta) = \sum_{u \in \mathcal{U}} \theta_u \, \mathrm{tr}\, H_u - \frac{1}{2} \sum_{u,v \in \mathcal{U}} \theta_u \theta_v (\mathrm{tr}\, H_u)(\mathrm{tr}\, H_v) + \mathrm{O}(\|\theta\|^3).$$

Then the result follows. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Since the random variables $L_u(x)$, $L_{uv}^{(\mathrm{sgm})}(x)$ and $L_{uv}^{(\mathrm{mix})}(x)$ are written in terms of H_u, it is valuable to consider moment formulas of H_u.

Lemma 9. Let $u \in \mathcal{U}$. Let U be Bernoulli randomization of u. Then H_u is written as $H_u = \mathrm{E}_U[\mathrm{e}^{\mathrm{i}\pi U^\top x} U U^\top]$. Furthermore, the random variable x can be replaced with a random variable ξ uniformly distributed on $[-1, 1]^m$, when any moment of $\mathrm{tr}(H_u)$ and $\mathrm{tr}(H_u H_v)$ is evaluated.

Proof. By Euler's formula $\cos\phi = (\mathrm{e}^{\mathrm{i}\pi\phi} + \mathrm{e}^{-\mathrm{i}\pi\phi})/2$, we obtain

$$\prod_{j=1}^m \cos(\pi u_j x_j) = \mathrm{E}_U[\mathrm{e}^{\mathrm{i}\pi U^\top x}].$$

Therefore $H_u = \mathrm{E}_U[\mathrm{e}^{\mathrm{i}\pi U^\top x} U U^\top]$. Next we consider moments. Consider, for example, expectation of $\mathrm{tr}(H_u H_v)$. The other moments are similarly evaluated. Let $\tilde{\beta}$ be a Bernoulli sequence, which is independent of x and any other Bernoulli sequences. Put $\xi = \tilde{\beta} \circ x$. Then ξ has the uniform distribution on $[-1, 1]^m$, and

$$
\begin{aligned}
\mathrm{E}_\xi[\mathrm{tr}(H_u(\xi) H_v(\xi))] &= \mathrm{E}_{\xi, U, V}[\mathrm{e}^{\mathrm{i}\pi U^\top \xi} \mathrm{e}^{\mathrm{i}\pi V^\top \xi} \mathrm{tr}(U U^\top V V^\top)] \\
&= \mathrm{E}_{\xi, U, V}[\mathrm{e}^{\mathrm{i}\pi U^\top \xi} \mathrm{e}^{\mathrm{i}\pi V^\top \xi} (U^\top V)^2] \\
&= \mathrm{E}_{x, \tilde{\beta}, U, V}[\mathrm{e}^{\mathrm{i}\pi (U \circ \tilde{\beta})^\top x} \mathrm{e}^{\mathrm{i}\pi (V \circ \tilde{\beta})^\top x} (U^\top V)^2] \\
&= \mathrm{E}_{x, \tilde{U}, \tilde{V}}[\mathrm{e}^{\mathrm{i}\pi \tilde{U}^\top x} \mathrm{e}^{\mathrm{i}\pi \tilde{V}^\top x} (\tilde{U}^\top \tilde{V})^2] \\
&= \mathrm{E}_x[\mathrm{tr}(H_u(x) H_v(x))],
\end{aligned}
$$

where we put $\tilde{U} = U \circ \tilde{\beta}$ and $\tilde{V} = V \circ \tilde{\beta}$, and used an identity $\tilde{U}^\top \tilde{V} = U^\top V$. $\qquad\square$

From Lemma 9, we simply write $H_u = \mathrm{E}_U[\mathrm{e}^{\mathrm{i}\pi U^\top \xi} U U^\top]$ below and the expectation with respect to x is replaced with the expectation with respect to ξ. Note that $\mathrm{E}_\xi[\mathrm{e}^{\mathrm{i}\pi a^\top \xi}] = 1_{\{a=0\}}$ for any $a \in \mathbb{Z}^m$.

Now the Fisher information matrix is evaluated as

$$
\begin{aligned}
J_{uv} &= \mathrm{E}_\xi[\mathrm{tr}\, H_u \,\mathrm{tr}\, H_v] \\
&= \mathrm{E}_{\xi, U, V}[\mathrm{e}^{\mathrm{i}\pi (U+V)^\top \xi} \|u\|^2 \|v\|^2] \\
&= \mathrm{E}_{U, V}[1_{\{U+V=0\}} \|u\|^2 \|v\|^2] \\
&= \mathrm{E}_{\beta, \tilde{\beta}}[1_{\{\beta \circ u = -\tilde{\beta} \circ v\}}] \|u\|^2 \|v\|^2 \\
&= \mathrm{E}_{\beta, \tilde{\beta}}\left[\prod_{i=1}^m \{1_{\{u_i = v_i = 0\}} + 1_{\{u_i = v_i > 0, \beta_i = -\tilde{\beta}_i\}}\} \right] \|u\|^2 \|v\|^2 \\
&= 1_{\{u=v\}} 2^{-|\sigma(u)|} \|u\|^4,
\end{aligned}
$$

where β and $\tilde{\beta}$ are Bernoulli sequences. This proves Lemma 3. By similar computation, we have the following lemma.

Lemma 10. Let U, V, S be Bernoulli randomization of $u, v, s \in \mathcal{U}$. Then

$$(\Gamma^w_{uv})^{(\mathrm{sgm})} = -\mathrm{E}_{U,V}[1_{\{\mathrm{abs}(U+V)=w\}}(U^\top V)^2 \|w\|^{-2}],$$

$$(\Gamma^w_{uv})^{(\mathrm{mix})} = -\mathrm{E}_{U,V}[1_{\{\mathrm{abs}(U+V)=w\}}\|u\|^2\|v\|^2\|w\|^{-2}],$$

Proof. We first calculate $\Gamma^{(\mathrm{sgm})}_{uv,s}$. By Lemma 8 and Lemma 9, we have

$$\begin{aligned}
\Gamma^{(\mathrm{sgm})}_{uv,s} &= -\mathrm{E}_\xi[\mathrm{tr}(H_u H_v)\,\mathrm{tr}\,H_s] \\
&= -\mathrm{E}_{\xi,U,V,S}\left[\mathrm{e}^{\mathrm{i}\pi(U+V+S)^\top \xi}(U^\top V)^2\|s\|^2\right] \\
&= -\mathrm{E}_{U,V,S}\left[1_{\{U+V+S=0\}}(U^\top V)^2\|s\|^2\right].
\end{aligned}$$

By using the expression of $\Gamma^{(\mathrm{sgm})}_{uv,s}$ and J^{sw}, we have

$$\begin{aligned}
(\Gamma^w_{uv})^{(\mathrm{sgm})} &= \sum_{s\in\mathcal{U}} \Gamma^{(\mathrm{sgm})}_{uv,s} J^{sw} \\
&= -\sum_{s\in\mathcal{U}} \mathrm{E}_{U,V,S}\left[1_{\{U+V+S=0\}}(U^\top V)^2\|s\|^2\right] 1_{\{s=w\}} 2^{|\sigma(s)|}\|s\|^{-4} \\
&= -\mathrm{E}_{U,V,W}\left[1_{\{U+V+W=0\}}(U^\top V)^2\|w\|^{-2}2^{|\sigma(w)|}\right] \\
&= -\mathrm{E}_{U,V,\beta}\left[1_{\{\mathrm{abs}(U+V)=w\}}1_{\{U+V=\beta\circ w\}}(U^\top V)^2\|w\|^{-2}2^{|\sigma(w)|}\right], \\
&= -\mathrm{E}_{U,V}\left[1_{\{\mathrm{abs}(U+V)=w\}}(U^\top V)^2\|w\|^{-2}\right],
\end{aligned}$$

where β is a Bernoulli sequence. The expression of $\Gamma^{(\mathrm{mix})}_{uv,s}$ and $(\Gamma^w_{uv})^{(\mathrm{mix})}$ is obtained similarly. $\qquad\square$

Lemma 11. The curvature tensor of SGM and MixM at $\theta = 0$ is

$$Q^{(\mathrm{sgm})}_{uv,wz} = \mathrm{E}_{U,V,W,Z}\left[\omega_\mathcal{U}(U,V,W,Z)(U^\top V)^2(W^\top Z)^2\right],$$

$$Q^{(\mathrm{mix})}_{uv,wz} = \mathrm{E}_{U,V,W,Z}\left[\omega_\mathcal{U}(U,V,W,Z)\|u\|^2\|v\|^2\|w\|^2\|z\|^2\right],$$

respectively, where U, V, W, Z are Bernoulli randomization of u, v, w, z and

$$\omega_\mathcal{U}(U,V,W,Z) = 1_{\{U+V+W+Z=0,\ \mathrm{abs}(U+V)\notin\mathcal{U}\cup\{0\}\}}.$$

Proof. We only derive the expression of $Q^{(\mathrm{sgm})}_{uv,wz}$. The expression of $Q^{(\mathrm{mix})}_{uv,wz}$ is obtained similarly. We first prove

$$\begin{aligned}
R^{(\mathrm{sgm})}_{uv}(x) &:= L^{(\mathrm{sgm})}_{uv}(x) + J_{uv} - \sum_{s\in\mathcal{U}}(\Gamma^s_{uv})^{(\mathrm{sgm})}L_s(x) \\
&= -\mathrm{E}_{U,V}\left[1_{\{\mathrm{abs}(U+V)\notin\mathcal{U}\cup\{0\}\}}\mathrm{e}^{\mathrm{i}\pi(U+V)^\top \xi}(U^\top V)^2\right].
\end{aligned} \tag{9}$$

The last term of $R_{uv}^{(\mathrm{sgm})}(x)$ is

$$-\sum_{s\in\mathcal{U}}(\Gamma_{uv}^s)^{(\mathrm{sgm})}L_s = \sum_{s\in\mathcal{U}}\mathrm{E}_{U,V,S}\left[1_{\{\mathrm{abs}(U+V)=s\}}\mathrm{e}^{\mathrm{i}\pi S^\top\xi}(U^\top V)^2\right]$$

$$= \sum_{s\in\mathcal{U}}\mathrm{E}_{U,V,\beta}\left[1_{\{\mathrm{abs}(U+V)=s\}}\mathrm{e}^{\mathrm{i}\pi(\beta\circ(U+V))^\top\xi}(U^\top V)^2\right]$$

$$= \mathrm{E}_{U,V,\beta}\left[1_{\{\mathrm{abs}(U+V)\in\mathcal{U}\}}\mathrm{e}^{\mathrm{i}\pi(\beta\circ(U+V))^\top\xi}(U^\top V)^2\right]$$

$$= \mathrm{E}_{U,V}\left[1_{\{\mathrm{abs}(U+V)\in\mathcal{U}\}}\mathrm{e}^{\mathrm{i}\pi(U+V)^\top\xi}(U^\top V)^2\right],$$

where β is a Bernoulli sequence. For the first and second term of $R_{uv}^{(\mathrm{sgm})}(x)$, we have

$$L_{uv}^{(\mathrm{sgm})} = -\mathrm{E}_{U,V}\left[\mathrm{e}^{\mathrm{i}\pi(U+V)^\top\xi}(U^\top V)^2\right],$$

$$J_{uv} = \mathrm{E}_{U,V}\left[1_{\{U+V=0\}}(U^\top V)^2\right] = \mathrm{E}_{U,V}\left[1_{\{\mathrm{abs}(U+V)=0\}}\mathrm{e}^{\mathrm{i}\pi(U+V)^\top\xi}(U^\top V)^2\right].$$

Hence (9) is obtained. Now the tensor $Q_{uv,wz}^{(\mathrm{sgm})}$ is calculated as follows:

$$Q_{uv,wz}^{(\mathrm{sgm})} = \mathrm{E}_\xi[R_{uv}^{(\mathrm{sgm})}R_{wz}^{(\mathrm{sgm})}]$$

$$= \mathrm{E}_{\xi,U,V,W,Z}\left[1_{\{\mathrm{abs}(U+V)\notin\mathcal{U}\cup\{0\},\mathrm{abs}(W+Z)\notin\mathcal{U}\cup\{0\}\}}\mathrm{e}^{\mathrm{i}\pi(U+V+W+Z)^\top\xi}(U^\top V)^2(W^\top Z)^2\right]$$

$$= \mathrm{E}_{U,V,W,Z}\left[1_{\{U+V+W+Z=0,\mathrm{abs}(U+V)\notin\mathcal{U}\cup\{0\}\}}(U^\top V)^2(W^\top Z)^2\right]$$

Therefore we obtain the desired expression. □

We finally prove Theorem 5. Since the Fisher information matrix is diagonal, we have

$$(\gamma^2)^{(\mathrm{sgm})} = \sum_{u,v,w,z\in\mathcal{U}} Q_{uv,wz}^{(\mathrm{sgm})}J^{uw}J^{vz}$$

$$= \sum_{u,v\in\mathcal{U}} Q_{uv,uv}^{(\mathrm{sgm})}J^{uu}J^{vv}$$

$$= \sum_{u,v\in\mathcal{U}} \mathrm{E}_{U,V,\tilde{U},\tilde{V}}\left[\omega_{\mathcal{U}}(U,V,\tilde{U},\tilde{V})(U^\top V)^2(\tilde{U}^\top\tilde{V})^2\right]\frac{2^{|\sigma(u)|+|\sigma(v)|}}{\|u\|^4\|v\|^4}.$$

Thus (7) is proved. (8) is shown similarly.

6.2 Proof of Theorem 4

We prove Theorem 4 by using the explicit expression (7) and (8) of the Efron curvature. We abbreviate $\gamma_{\mathcal{U}}^2$ as γ^2.

We prove the first inequality in (6). By the expression (7), it is sufficient to show that $\omega_{\mathcal{U}}(u,u,-u,-u)=1$ for some $u\in\mathcal{U}$. Let u be an element such that

11

$\|u\|_1 = \max_{v \in \mathcal{U}} \|v\|_1$. Then we have $u + u - u - u = 0$ and $u + u \notin \mathcal{U} \cup \{0\}$, and hence $\omega_\mathcal{U}(u, u, -u, -u) = 1$.

The second inequality in (6) follows from equations (7), (8), and

$$(U^\top V)^2 (\tilde{U}^\top \tilde{V})^2 \le \|U\|^2 \|V\|^2 \|\tilde{U}\|^2 \|\tilde{V}\|^2 = \|u\|^4 \|v\|^4.$$

We now consider the equality condition. First assume $\mathcal{U} \subset \mathbb{Z}_i$. Then $(U^\top V)^2 (\tilde{U}^\top \tilde{V})^2$ in (7) is equal to $(u_i v_i)^2 (u_i v_i)^2$, which is equal to $\|u\|^4 \|v\|^4$. Therefore $(\gamma^2)^{(\mathrm{sgm})} = (\gamma^2)^{(\mathrm{mix})}$. Conversely, assume $(\gamma^2)^{(\mathrm{sgm})} = (\gamma^2)^{(\mathrm{mix})}$. Since $\mathcal{U}$ is a non-empty finite subset, there exist some $u \in \mathcal{U}$ and some $i \in \{1, \ldots, m\}$ such that

$$u_i > 0 \text{ and } u_i \ge w_i \ (\forall w \in \mathcal{U}).$$

Fix such u and i. We show $u \in \mathbb{Z}_i$. Define an integer vector $\bar{u} \in \mathbb{Z}^m$ by $\bar{u}_i = u_i$ and $\bar{u}_j = -u_j$ for $j \neq i$. Since $|u_i + \bar{u}_i| = 2u_i > u_i$, we have $\mathrm{abs}(u + \bar{u}) \notin \mathcal{U} \cup \{0\}$ and therefore $\omega_\mathcal{U}(u, \bar{u}, -u, -\bar{u}) = 1$. Let $\{U_{(k)}\}_{k=1}^4$ be four independent Bernoulli randomization of u. Note that each $U_{(k)}$ takes u (resp. $\bar{u}$) with probability at least 2^{-m}. We evaluate

$$0 = (\gamma^2)^{(\mathrm{mix})} - (\gamma^2)^{(\mathrm{sgm})}$$

$$\ge \mathrm{E}_{U_{(1)}, U_{(2)}, U_{(3)}, U_{(4)}} \left[\omega_\mathcal{U}(U_{(1)}, U_{(2)}, U_{(3)}, U_{(4)}) \left(1 - \frac{(U_{(1)}^\top U_{(2)})^2 (U_{(3)}^\top U_{(4)})^2}{\|u\|^8} \right) \right]$$

$$\ge 2^{-4m} \omega_\mathcal{U}(u, \bar{u}, -u, -\bar{u}) \left(1 - \frac{(u^\top \bar{u})^4}{\|u\|^8} \right) \ge 0.$$

This implies $|u^\top \bar{u}| = \|u\|^2$. By equality condition of the Cauchy-Schwarz inequality, there is a real number ρ such that $u = \rho \bar{u}$. This implies $u \in \mathbb{Z}_i$. Now, by contradiction, assume that there exists some $v \in \mathcal{U} \setminus \mathbb{Z}_i$. We further assume $v_i \ge w_i$ for any $w \in \mathcal{U} \setminus \mathbb{Z}_i$ without loss of generality. Since $u_i + v_i > v_i$ and $u + v \notin \mathbb{Z}_i$, we deduce $u + v \notin \mathcal{U} \cup \{0\}$. Hence $\omega_\mathcal{U}(u, v, -u, -v) = 1$. Then we have

$$0 = (\gamma^2)^{(\mathrm{mix})} - (\gamma^2)^{(\mathrm{sgm})} \ge 2^{-4m} \omega_\mathcal{U}(u, v, -u, -v) \left(1 - \frac{(u^\top v)^4}{\|u\|^4 \|v\|^4} \right) \ge 0.$$

This implies $|u^\top v| = \|u\| \|v\|$. By equality condition of the Cauchy-Schwarz inequality, there is a real number $\tilde{\rho}$ such that $v = \tilde{\rho} u$. This implies $v \in \mathbb{Z}_i$ and contradict the definition of v. Thus we have $\mathcal{U} \subset \mathbb{Z}_i$.

6.3 Proof of Theorem 6 and Corollary 7

We first prove Theorem 6. Put $d = \max_m \max_{u \in \mathcal{U}_m} |\sigma(u)| < \infty$. We abbreviate $\mathcal{U}_m$ by $\mathcal{U}$ below. It is sufficient to prove that $(\gamma_\mathcal{U}^2)^{(\mathrm{sgm})} \le |N(\mathcal{U})|$ and $(\gamma_\mathcal{U}^2)^{(\mathrm{mix})} \ge c|M(\mathcal{U})|$

with a positive constant c. If $(u, v) \notin N(\mathcal{U})$, then $U^\top V = 0$ in (7). Hence

$$(\gamma_\mathcal{U}^2)^{(\text{sgm})} \leq \sum_{(u,v) \in N(\mathcal{U})} \mathrm{E}_{U,V,\tilde{U},\tilde{V}} \left[\omega_\mathcal{U}(U, V, \tilde{U}, \tilde{V}) \frac{(U^\top V)^2 (\tilde{U}^\top \tilde{V})^2}{\|u\|^4 \|v\|^4} \right] \leq |N(\mathcal{U})|.$$

We next evaluate (8). If $(u, v) \in M(\mathcal{U})$, then $\omega_\mathcal{U}(u, v, -u, -v) = 1$. Since u has at most d non-zero elements, the event $U = u$ happens with probability at least 2^{-d}, where U is a Bernoulli randomization of u. Therefore

$$(\gamma_\mathcal{U}^2)^{(\text{mix})} \geq \sum_{(u,v) \in M(\mathcal{U})} \mathrm{E}_{U,V,\tilde{U},\tilde{V}} \left[\omega_\mathcal{U}(U, V, \tilde{U}, \tilde{V}) \right] \geq 2^{-4d} |M(\mathcal{U})|.$$

This proves Theorem 6.

Next we prove Corollary 7. Assume $|N(\mathcal{U})|/|\mu(\mathcal{U})|^2 \to 0$. Note that $|\mu(\mathcal{U})| \to \infty$ since $|N(\mathcal{U})| \geq |\mathcal{U}| \geq 1$. From the definition of $M(\mathcal{U})$ and $\mu(\mathcal{U})$, the set $\{(u, v) \in \mathcal{U}^2 \mid u, v \in \mu(\mathcal{U}), u \neq v\}$ is a subset of $M(\mathcal{U})$. Then we have $|M(\mathcal{U})| \geq |\mu(\mathcal{U})|(|\mu(\mathcal{U})| - 1)$. Thus

$$\frac{|N(\mathcal{U})|}{|M(\mathcal{U})|} \leq \frac{|N(\mathcal{U})|}{|\mu(\mathcal{U})|^2 (1 - |\mu(\mathcal{U})|^{-1})} \to 0$$

and the proof is completed.

Acknowledgements

References

S. Amari. *Differential-Geometrical Methods in Statistics*. Springer, New York, 1985.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, New York, 1978.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, second edition, 2006.

B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.*, 3(6):1189–1242, 1975.

P. Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.

P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall/CRC, second edition, 1989.

J. Reeds. Discussion to Efron's paper. *Ann. Statist.*, 3(6):1234–1238, 1975.

T. Sei. A structural model on a hypercube represented by optimal transport. *Statistica Sinica*, 2010. To appear. (Preprint: arXiv:0901.4715).

C. Villani. *Topics in Optimal Transportation.* AMS, Providence, 2003.

C. Villani. *Optimal Transport, Old and New.* Springer, Berlin, 2009.