A COMMENT ON A REGULARITY CONDITION IN A CURVED EXPONENTIAL FAMILY

Etsuo Kumagai

Received February 2, 2014; revised March 28, 2014

Abstract. Based on a curved exponential family, there is a regularity condition that the score function with random variables is the linear independence, which is commonly used in the information geometry. An equivalence relation to the regularity condition is that the Fisher information is positive definite under the curved exponential family. We investigate a key condition for two regularity conditions and we recognize it as the linear independence for the first derivative of natural parameter with respect to the parameter.

1 Introduction [3] introduced the ideas of the statistical curvature with respect to the asymptotic information loss. [1] introduced the statistical differential manifold and developed α-connection and m-connection in the curved exponential family. A lot of researchers have investigated the information geometry and there are a lot of fruitful and valuable results for the asymptotic.

In the framework of the information geometry, the statistical manifold $\{p(\boldsymbol{x};\boldsymbol{\theta})\}$ is based on the family of distributions with a parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbf{R}^k$. Among the regularity conditions in this framework, a regularity condition which we consider is that the derivatives $\{\partial\ell(\boldsymbol{\theta})/\partial\theta_i\}$ are linearly independent where $\ell(\boldsymbol{\theta})$ is the log-likelihood ([4](page 76), [2](page 29)). It seems that the linear independence on the derivatives is reasonable for constructing a tangent space in the statistical manifold, but we wonder whether this assumption is rational with respect to the underlying distribution of the random variable. Remark that we do not intend to consider singular models that do not satisfy the usual regularity conditions.

Based on a curved exponential family, an equivalence relation to the regularity condition that the score function with random variables is the linear independence is the regularity condition that the Fisher information is positive definite ([2](pages 25–29)). We investigate a key condition for above two regularity conditions and we recognize it as the linear independence for the first derivative of natural parameter with respect to the parameter.

2 The regularity condition [2](pages 25–29) considers a family of probability distribution on $\mathcal{X}$, i.e., $\mathcal{S} = \{p_\theta = p(\boldsymbol{x};\boldsymbol{\theta}) \,|\, \boldsymbol{\theta} \in \Theta \subseteq \mathbf{R}^k\}$ as k-dimensional statistical model on a set $\mathcal{X}$ which is a discrete set or $\mathbf{R}^m$ $(k \leq m)$. Letting p be a probability (density) function on $\mathcal{X}$, the support of p is defined by $\mathrm{supp}(p) \stackrel{\mathrm{def}}{=} \{\boldsymbol{x} \,|\, p(\boldsymbol{x}) > 0\}$ which is assumed to be constant with respect to $\boldsymbol{\theta}$, and $\mathcal{X}$ is redefined as $\mathrm{supp}(p)$, so that the statistical model $\mathcal{S}$ is a subset of

$$\mathcal{P}(\mathcal{X}) \stackrel{\mathrm{def}}{=} \left\{ p : \mathcal{X} \to \mathbf{R} \;\middle|\; p(\boldsymbol{x}) > 0 \ (\forall x \in \mathcal{X}),\ a\int_{\mathcal{X}} p(\boldsymbol{x})d\boldsymbol{x} = 1 \right\}.$$

Note that the integral is interpreted as the summation if $\mathcal{X}$ is a discrete case. For the statistical model $\mathcal{S} = \{p_\theta\}$, defining the mapping $\varphi : \mathcal{S} \to \mathbf{R}^k$ by $\varphi(p_\theta) = \boldsymbol{\theta}$ implies a

2010 *Mathematics Subject Classification.* 62B10.

Key words and phrases. Curved exponential family, regularity condition, linear independence, information geometry.

coordinate system for $\mathcal{S}$. Furthermore suppose that there is a C^∞ diffeomorphism $\psi : \Theta \to \psi(\Theta)(\subset \mathbf{R}^k)$, so that, if we use $\boldsymbol{\eta} = \psi(\boldsymbol{\theta})$ as another parameter, then it holds that $\mathcal{S} = \{p_\theta \,|\, \boldsymbol{\theta} \in \Theta\} = \{p_{\psi^{-1}(\eta)} \,|\, \boldsymbol{\eta} \in \psi(\Theta)\}$. Thus a parameterization of $\mathcal{S}$ is a coordinate system of $\mathcal{S}$ as a C^∞ differentiable manifold.

Letting $[\theta^i]$ be a coordinate system in the statistical manifold $\mathcal{S}$ implies the vector fields formed by the natural bases $\{\partial_i\} \in T_\theta(\mathcal{S})$ which is the tangent space. Note that ∂_i means $\dfrac{\partial}{\partial\theta^i}$ and $\{\partial_i\}$ are vector fields. The Fisher information matrix at $\boldsymbol{\theta}$ in $\mathcal{S}$ is defined by the $k \times k$ matrix $G(\boldsymbol{\theta}) = (g_{ij}(\boldsymbol{\theta}))$ where the (i,j)-th element of $G(\boldsymbol{\theta})$ is, for $i, j = 1, \ldots, k$,

$$g_{ij}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} E_\theta\Big[\partial_i \ell \; \partial_j \ell\Big] = \int_{\mathcal{X}} \partial_i \ell(\boldsymbol{x};\boldsymbol{\theta})\, \partial_j \ell(\boldsymbol{x};\boldsymbol{\theta})\, p(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x} \quad (\in \mathbf{R}),$$

where $\ell = \ell(\boldsymbol{x};\boldsymbol{\theta}) = \log p(\boldsymbol{x};\boldsymbol{\theta})$ is the log-likelihood function and E_θ means the expectation with respect to the distribution p_θ. [2] shows the assumptions as follows:

> **(page 28, line 3 from below to page 29, line 5)** *The matrix $G(\boldsymbol{\theta})$ is symmetric ($g_{ij}(\boldsymbol{\theta}) = g_{ji}(\boldsymbol{\theta})$), and since for any k-dimensional vector $\boldsymbol{c} = (c^1, \ldots, c^k)^t$ (^{t} denotes transpose),*
>
> $$\boldsymbol{c}^t G(\boldsymbol{\theta}) \boldsymbol{c} = \int \left\{ \sum_{i=1}^k c^i\, \partial_i \ell(\boldsymbol{x};\boldsymbol{\theta}) \right\}^2 p(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x} \geq 0, \tag{1}$$
>
> *it is also positive semidefinite. We assume further that $G(\boldsymbol{\theta})$ is positive definite. From the equation above, we see that this is equivalent to stating that the elements of $\{\partial_1 \ell, \ldots, \partial_k \ell\}$ when viewed as functions on $\mathcal{X}$ are linearly independent, which, in turn, is equivalent to stating that the elements of $\{\partial_1 p_\theta, \ldots, \partial_k p_\theta\}$ are linearly independent.*
>
> **(page 29, lines 13–16, 18)** *Now suppose that the assumption above hold, and define the inner product of the natural basis of the coordinate system $[\theta^i]$ by $g_{ij}(\boldsymbol{\theta}) = \langle \partial_i, \partial_j \rangle$. This uniquely determines a Riemannian metric $g(\boldsymbol{\theta}) = \langle , \rangle$. We call this the Fisher metric, or alternatively, the information metric.*
>
> **(page 29, line 18)** *Indeed we may write $\langle X, Y \rangle_\theta = E_\theta[(X\ell)(Y\ell)]$ for all tangent vectors $X, Y \in T_\theta(\mathcal{S})$.*

The above assumptions seem to be the key to the regularity conditions in the information geometry. For convenience sake, we shall define the following conditions:

Condition A The matrix $G(\boldsymbol{\theta}) = \Big(\big\langle \partial_i, \partial_j \big\rangle_\theta\Big) = \Big(E_\theta\big[\partial_i \ell \; \partial_j \ell\big]\Big)$ in (1) is positive definite where ∂_i, ∂_j are the natural bases in $T_\theta(\mathcal{S})$ and ℓ is the log-likelihood function.

Condition B The elements of $\big\{\partial_1 \ell, \ldots, \partial_k \ell\big\}$ are linearly independent.

In the curved exponential family with a parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbf{R}^k$, the density of the random variable $\boldsymbol{X} \in \mathbf{R}^m$ is $p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\langle \boldsymbol{\alpha}(\boldsymbol{\theta}), \boldsymbol{x} \rangle - \psi(\boldsymbol{\alpha}(\boldsymbol{\theta}))\} p_0(\boldsymbol{x})$, where $p_0(\boldsymbol{x})$ is a pivotal probability measure, $\boldsymbol{\alpha}(\boldsymbol{\theta}) \in \mathbf{R}^m$ is a curved natural parameter parametrized by $\boldsymbol{\theta}$, the Euclidean inner product $\langle , \rangle$ in the exponent is the usual product of two vectors, and $\psi(\boldsymbol{\alpha}(\boldsymbol{\theta})) \in \mathbf{R}$ is the cumulant generating function. Note that the natural parameter space $\mathcal{A}$ is defined by $\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbf{R}^m | \int \exp\{\langle \boldsymbol{\alpha}, \boldsymbol{x} \rangle\} p_0(\boldsymbol{x}) d\boldsymbol{x} < \infty\}$ and $\{\boldsymbol{\alpha}(\boldsymbol{\theta})\} \in \mathcal{A}$.

Now we assume that $\mathrm{supp}(p) = \mathbf{R}^m$ which is independent of the parameter $\boldsymbol{\theta}$. The usual derivative of the log-likelihood $\ell = \log p(\boldsymbol{x};\boldsymbol{\theta})$ with respect to the vector $\boldsymbol{\theta}$ is defined by

$$\partial_{\boldsymbol{\theta}}\, \ell \stackrel{\text{def}}{=} \frac{\partial\, \ell}{\partial\, \boldsymbol{\theta}} = \big(\partial_1 \ell, \ldots, \partial_k \ell\big)^T = \big\langle \dot{\boldsymbol{\alpha}}(\boldsymbol{\theta}),\; \boldsymbol{X} - \boldsymbol{\beta}(\boldsymbol{\theta}) \big\rangle \quad (\in \mathbf{R}^k),$$

where $\dot{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \partial\boldsymbol{\alpha}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ and $\boldsymbol{\beta}(\boldsymbol{\theta}) = E_\theta[\boldsymbol{X}]$, that is, the i-th element is

$$\partial_i \ell \;=\; \langle \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}),\; \boldsymbol{X} - \boldsymbol{\beta}(\boldsymbol{\theta}) \rangle \quad (\in \mathbf{R}) \quad (i = 1, \ldots, k),$$

where $\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta^i}\boldsymbol{\alpha}(\boldsymbol{\theta})$. The matrix $G(\boldsymbol{\theta})$ is obtained by

$$G(\boldsymbol{\theta}) \;=\; E_\theta \left[\langle \partial_{\boldsymbol{\theta}}\, \ell, \partial_{\boldsymbol{\theta}}\, \ell \rangle \right] \;=\; \dot{\boldsymbol{\alpha}}(\boldsymbol{\theta})^t\; \boldsymbol{\Sigma}(\boldsymbol{\theta})\; \dot{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \quad (\in \mathbf{R}^{k \times k}), \tag{2}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix of $\boldsymbol{X}$ under the probability $p(\boldsymbol{x}; \boldsymbol{\theta})$.

Since the derivative of ℓ is $\boldsymbol{\partial}\ell = \langle \dot{\boldsymbol{\alpha}}(\boldsymbol{\theta}),\, \boldsymbol{X} - \boldsymbol{\beta}(\boldsymbol{\theta}) \rangle$ for $k < m$, Condition B means the following relationship:

$$\sum_{i=1}^{k} c_i\, \partial_i \ell \;=\; \Big\langle \sum_{i=1}^{k} c_i\, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}),\, \boldsymbol{X} - \boldsymbol{\beta}(\boldsymbol{\theta}) \Big\rangle \;=\; 0 \;\Longrightarrow\; \forall\, c_i \;=\; 0. \tag{3}$$

Although the derivatives $\{\partial_i \ell\}$ $(i = 1, \ldots, k)$ are supposed to be linearly independent, they are also random variables based on the distribution p_θ and the matrix $G(\boldsymbol{\theta})$ is assumed to be calculated by both the random variables $\{\partial_i \ell\}$ and their distribution p_θ. Note that, in the exponential family $(k = m)$, since $\boldsymbol{\partial}\ell = \boldsymbol{X} - \boldsymbol{\beta}$, Condition B means the following relationship:

$$\sum_{i=1}^{m} c_i\, \partial_i \ell \;=\; \sum_{i=1}^{m} c_i\, (X_i - \beta_i) \;=\; 0 \;\Longrightarrow\; \forall\, c_i \;=\; 0. \tag{4}$$

3 An underlying condition for those regularity conditions With respect to the two conditions in the previous section, we investigate what are an underlying condition if Condition A is equivalent to Condition B under the curved exponential family.

Lemma 3.1 *In the curved exponential family, assume that the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite. Then Condition A is equivalent to the linear independence of $\{\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta})\}$, i.e.,*

$$\sum_{i=1}^{k} c_i\, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) \;=\; \mathbf{0} \quad\Longrightarrow\quad \forall c_i = 0. \tag{5}$$

Proof: Since the covariance matrix is positive definite, we decompose the matrix as follows: $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}$. Then the matrix $G(\boldsymbol{\theta})$ in (2) is decomposed by

$$G(\boldsymbol{\theta}) \;=\; \Big(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\dot{\boldsymbol{\alpha}}(\boldsymbol{\theta})\Big)^t \Big(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\dot{\boldsymbol{\alpha}}(\boldsymbol{\theta})\Big)$$

and is considered as the Gram matrix, so that, by its property, Condition A is equivalent to that the k components $\{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta})\}$ of $m \times 1$ vectors in $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\dot{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ are linearly independent, i.e.,

$$\sum_{i=1}^{k} c_i\, \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) \;=\; \boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2} \left(\sum_{i=1}^{k} c_i\, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) \right) \;=\; \mathbf{0} \quad\Longrightarrow\quad \forall c_i = 0,$$

so that, since the matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{1/2}$ has the inverse, Condition A equals that $\{\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta})\}$ are linearly independent. □

This lemma means that Condition A depends on only the derivatives of the natural parameter $\boldsymbol{\alpha}(\boldsymbol{\theta})$, not the log-likelihood function $\ell(\boldsymbol{\theta})$ with the random variable $\boldsymbol{X}$ directly.

Note that, since $\boldsymbol{\alpha}(\boldsymbol{\theta}) = \boldsymbol{\alpha}$ in an exponential family, it holds that $\partial_i \boldsymbol{\alpha} = \boldsymbol{e}_i$ which is the i-th unit vector, so that the equivalent condition in Lemma 3.1 is

$$\sum_{i=1}^{m} c_i \, \boldsymbol{e}_i \;=\; \mathbf{0} \quad \Longrightarrow \quad \forall c_i = 0, \tag{6}$$

which is trivial because of the property of unit vectors. Next we consider the equivalent condition for Condition B.

Lemma 3.2 *In the curved exponential family, Condition B is equivalent to the condition (5), i.e., the linear independence of* $\{\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta})\}$.

Proof: Since Condition B is that the derivatives $\{\partial_i \ell\}$ are linearly independent, i.e., (3), for the equation

$$\left\langle \sum_{i=1}^{k} c_i \, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}),\; \boldsymbol{X} - \boldsymbol{\beta}(\boldsymbol{\theta}) \right\rangle = 0 \tag{7}$$

in Condition B, we consider two cases as follows: Case (i) $\sum_{i=1}^{k} c_i \, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) = \mathbf{0}$ and Case (ii) $\sum_{i=1}^{k} c_i \, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) \neq \mathbf{0}$. For the Case (i), the equation (7) always holds without reference to the random variable $\boldsymbol{X}$, so that the Condition (5) implies Condition B.

On the other hand, for the Case (ii), it holds under the condition as follows:

$$\boldsymbol{X} - \boldsymbol{\beta}(\boldsymbol{\theta}) \;\in\; \mathcal{N}\left(\sum_{i=1}^{k} c_i \, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) \right), \tag{8}$$

which is a normal space against the vector $\sum_{i=1}^{k} c_i \, \partial_i \boldsymbol{\alpha}(\boldsymbol{\theta}) (\neq \mathbf{0})$. If the random variable $\boldsymbol{X}$ satisfies the condition (8) for the Case (ii), then the equation (7) holds, but the necessary condition $\forall c_i = 0$ in Condition B contradicts the Case (ii).

Therefore the equation (7) in Condition B is equivalent to the Case (i), i.e., the sufficient condition in (5) without reference to the random variable $\boldsymbol{X}$ in $\{\partial_i \ell\}$, so that we have the required result. □

Note that, since $\boldsymbol{\alpha}(\boldsymbol{\theta}) = \boldsymbol{\alpha}$ in an exponential family, it holds that, in the same fashion before, the equivalent condition in Lemma 3.2 is (4), which is equivalent to the condition (6) because of $\langle \boldsymbol{e}_i, \boldsymbol{X} - \boldsymbol{\beta} \rangle = X_i - \beta_i$ for $i = 1, \ldots, m$. The previous two lemmas imply the following theorem:

Theorem 3.1 *If the first derivative* $\dot{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ *of the natural parameter is full rank in the curved exponential family and the covariance matrix* $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ *is positive definite, then, without reference to the random variable* $\boldsymbol{X}$ *in the derivatives of log-likelihood function, Condition A with respect to the Fisher information matrix* $\boldsymbol{G}(\boldsymbol{\theta})$ *is equivalent to Condition B with respect to the log-likelihood function* $\ell(\boldsymbol{\theta})$. □

Therefore, from Theorem 3.1, what [2](page 29) stated that *"We assume further that* $G(\boldsymbol{\theta})$ *is positive definite. From the equation above, we see that this is equivalent to stating that the elements of* $\{\partial_1 \ell, \ldots, \partial_k \ell\}$ *when viewed as functions on* $\mathcal{X}$ *are linearly independent."*

just means that the derivatives $\{\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta})\}$ are linearly independent under the positive definite covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

Because a relationship in the curved exponential family

$$\dot{\boldsymbol{\beta}}(\boldsymbol{\theta}) \;=\; \boldsymbol{\Sigma}(\boldsymbol{\theta})\, \dot{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \tag{9}$$

holds, we have the following corollary:

COROLLARY **3.1** *If the first derivative $\dot{\boldsymbol{\beta}}(\boldsymbol{\theta})$ of the expectation parameter is full rank in the curved exponential family and the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite, then, without reference to the random variable $\boldsymbol{X}$ in the derivatives of log-likelihood function, Condition A is equivalent to Condition B.*

Proof: Because of the relationship (9), if the first derivative $\dot{\boldsymbol{\beta}}(\boldsymbol{\theta})$ of the expectation parameter is full rank in a curved exponential family and the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite, then the first derivative $\dot{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ of the natural parameter is full rank, so that the derivatives $\{\partial_i \boldsymbol{\alpha}(\boldsymbol{\theta})\}$ are linearly independent and we have the required result by Theorem 3.1. □

4 Conclusion Based on the curved exponential family, we investigate the regularity conditions of Condition A and Condition B with respect to the linear independence and we conclude they are equivalent to the linear independence for the first derivative of the natural parameter with respect to the parameter under the condition that the covariance matrix is positive definite.

Acknowledgement: The author would like to express his appreciation to the referee.

REFERENCES

[1] S.Amari (1985), *Differential-Geometrical Methods in Statistics*, Springer.

[2] S.Amari and H.Nagaoka (2000), *Methods of Information Geometry*, American Mathematical Society.

[3] B.Efron (1975), Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency), *Ann. Stat.*, **3**, 1189–1242.

[4] M.K.Murray and J.W.Rice (1993), *Differential Geometry and Statistics*, Chapman & Hall.

Communicated by *Hisao Nagao*

Division of Mathematical Science,
Graduate School of Engineering Science,
Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan.