Inference on the eigenvalues of the covariance matrix of a multivariate normal distribution–geometrical view–

Yo Sheena[*]

September 2012

Abstract

We consider inference on the eigenvalues of the covariance matrix of a multivariate normal distribution. The family of multivariate normal distributions with a fixed mean is seen as a Riemannian manifold with Fisher information metric. Two submanifolds naturally arises; one is the submanifold given by the fixed eigenvectors of the covariance matrix, the other is the one given by the fixed eigenvalues. We analyze the geometrical structures of these manifolds such as metric, embedding curvature under e-connection or m-connection. Based on these results, we study 1) the bias of the sample eigenvalues, 2) asymptotic variance of estimators, 3) the asymptotic information loss caused by neglecting the sample eigenvectors, 4) the derivation of a new estimator that is natural from a geometrical point of view.

MSC(2010) *Subject Classification*: Primary 62H05; Secondary 62F12
Key words and phrases: curved exponential family, information loss, Fisher information metric, embedding curvature, affine connection, positive definite matrix.

1 Introduction

Consider a normal distribution with zero mean and an unknown covariance matrix, $N(\mathbf{0}, \boldsymbol{\Sigma})$. Let denote the eigenvalues of $\boldsymbol{\Sigma}$ by

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p), \quad \lambda_1 > \ldots > \lambda_p$$

and eigenvectors matrix by $\boldsymbol{\Gamma}$, hence we have the spectral decomposition

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t, \quad \boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}), \tag{1}$$

where $\mathrm{diag}(\boldsymbol{\lambda})$ means the diagonal matrix with the ith diagonal element λ_i. It is needless to say that the inference on $\boldsymbol{\Sigma}$ is an important task in

[*] Faculty of Economics, Shinshu University

many practical situations in such a diversity of fields as engineering, biology, chemistry, finance, psychology etc. Especially we often encounter the cases where the property of interest depends on $\boldsymbol{\Sigma}$ only through its eigenvalues $\boldsymbol{\lambda}$. We treat an inference problem on the eigenvalues $\boldsymbol{\lambda}$ from a geometrical point of view.

Treating the family of normal distributions $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{\mu}$ is not necessarily zero) as a Riemmanian manifold has been done by several authors. For example, see Fletcher and Joshi, [12], Lenglet et al. [18], Skovgaard [25], Smith [26], Yoshizawa and Tanabe [29]. When μ euqals zero, the family of normal distributions $N(\mathbf{0}, \boldsymbol{\Sigma})$ can be taken as a manifold (say $\mathcal{S}$) with a single coordinate system $\boldsymbol{\Sigma}$. Hence, $\mathcal{S}$ is identified with the space of symmetric positive definite matrices. Geometrically analyzing the space of symmetric positive definite matrices has been an interesting topic in a mathematical or engineering point of view. Refer to Moakher and Zéraï [20], Ohara et al. [23] and Zhang et al. [30] as well as the above literature.

In this paper, we analyze $\mathcal{S}$ from the standpoint of information geometry while focusing on the inference on the eigenvalues of $\boldsymbol{\Sigma}$. The paper is aimed to make a contribution in two regards: 1) The geometrical structure of $\mathcal{S}$ is analyzed in view of the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$; 2) Some statistical problems on the inference for $\boldsymbol{\lambda}$ are explained in the geometrical terms.

We summarize the inference problem for $\boldsymbol{\lambda}$. Based on independent n samples $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})', \ i = 1, \ldots, n$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$, we want to make inference on the unknown $\boldsymbol{\lambda}$. We confine ourselves to the classical case where $n \geq p$. It is well-known that the product-sum matrix

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t$$

is sufficient statistic for both unknown $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$. The spectral decomposition of $\boldsymbol{S}$ is given by

$$\boldsymbol{S} = \boldsymbol{H}\boldsymbol{L}\boldsymbol{H}^t, \quad \boldsymbol{L} = \mathrm{diag}(\boldsymbol{l}),$$

where

$$\boldsymbol{l} = (l_1, \ldots, l_p), \quad l_1 > \ldots > l_p > 0 \ a.e.$$

are the eigenvalues of $\boldsymbol{S}$, and $\boldsymbol{H}$ is the corresponding eigenvectors matrix. This decomposition gives us two statistics available, i.e. the sample eigenvalues $\boldsymbol{l}$ and the sample eigenvectors $\boldsymbol{H}$. However it is almost customary that we only use the sample eigenvalues, discarding the information contained in $\boldsymbol{H}$. In the past literature on the inference for the population eigenvalues, every notable estimator is based simply on the sample eigenvalues. See Takemura [27], Dey and Srinivasan [9], Haff [13], Yang and Berger [28] for orthogonally invariant estimators of $\boldsymbol{\Sigma}$; Dey [8], Hydorn and Muirhead [14], Jin [15], Sheena and Takemura [24] for direct estimators of $\boldsymbol{\lambda}$. Since we do not have enough space to state the concrete form of each estimator, we just mention Stein's estimator as a pioneering work for "shrinkage" estimator of $\boldsymbol{\Sigma}$. In general, an orthogonally invariant estimator of $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{H}\boldsymbol{\Phi}\boldsymbol{H}^t, \quad \boldsymbol{\Phi} = \mathrm{diag}(\phi_1(\boldsymbol{l}), \ldots, \phi_p(\boldsymbol{l})). \tag{2}$$

The estimator of $\boldsymbol{\lambda}$ is given by the eigenvalues of $\hat{\boldsymbol{\Sigma}}$, that is, $(\phi_1(\boldsymbol{l}), \ldots, \phi_p(\boldsymbol{l}))$. The sample covariance matrix (M.L.E. estimator) $\bar{\boldsymbol{S}} \triangleq n^{-1}\boldsymbol{S}$ gives the estimator of $\boldsymbol{\lambda}$ as $\phi_i(\boldsymbol{l}) = n^{-1}l_i,\ i = 1, \ldots, p$, while Stein's "shrinkage" estimator gives birth to

$$\phi_i(\boldsymbol{l}) = l_i/(n + p + 1 - 2i), \quad i = 1, \ldots, p. \tag{3}$$

Stein's estimator assigns the lighter (heavier) weight to the larger (smaller) sample eigenvalues, hence the diversity of $\boldsymbol{l}$ is shrunk. This estimator is quite simple and performs much better than M.L.E. (see [9]). Unlike Stein's estimator, many estimators in the above literature are not explicitly given or too complicated for immediate use. Nonetheless they all have one common feature. The derived estimators of $\boldsymbol{\lambda}$ only depends on $\boldsymbol{l}$.

In a sense it is natural to implicitly associate the sample eigenvalues to the population eigenvalues, and the sample eigenvectors to the population counterpart. However the sample eigenvalues are not sufficient for the unknown population eigenvalues. Therefore it is important to evaluate how much information is lost by neglecting the sample eigenvectors. Following Amari [1], we gain an understanding of the asymptotic information loss with geometric terms such as Fisher information metric and embedding curvatures.

Another statistically interesting topic is the bias of $n^{-1}\boldsymbol{l}$. It is well known that $n^{-1}\boldsymbol{l}$ is largely biased and the estimators mentioned above are all modification of $n^{-1}\boldsymbol{l}$ to correct the bias, that is, "shrinkage estimators." We show that the bias is closely related to the embedding curvatures. Moreover the geometric structure of $\mathcal{S}$ naturally leads us to a new estimator, which is also a shrinkage estimator.

The organization of this paper is as follows: In the former part (Section 2 and Section 3), we describe the geometrical structure of $\mathcal{S}$ in view of the spectral decomposition (1). In Section 2, we observe $\mathcal{S}$ as a Riemannian manifold endowed with Fisher information metrics. In Section 3, we treat two submanifolds of $\mathcal{S}$, a submanifold given by the fixed eigenvectors and the one given by the fixed eigenvalues. The embedding curvatures of these submanifolds are explicitly given. We will show that the bias of $\boldsymbol{l}$ is closely related to the curvatures. In the latter part (Section 4 and 5), we consider the estimation problem of $\boldsymbol{\lambda}$. In Section 4, we describe the asymptotic variance of estimators when $\boldsymbol{\Gamma}$ is known (Section 4.1) and the asymptotic information loss caused by discarding the sample eigenvectors $\boldsymbol{H}$ (Section 4.2). The asymptotic information loss could be measured by the difference in the asymptotic variance between two certain estimators. In Section 5 for the case when $\boldsymbol{\Gamma}$ is unknown, we propose a new estimator of $\boldsymbol{\lambda}$, which is naturally derived from a geometric point of view. In the last section, some comments are made for further research. All the proofs are collected in Appendix.

Unfortunately we do not have enough space to explain the geometrical concepts used in this paper. Please refer to Boothby [6], Amari [2], Amari and Nagaoka [3].

2 Riemannian Manifold and Metric

The density of the normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ is given by

$$f_{\boldsymbol{\Sigma}}(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right), \quad \boldsymbol{x} = (x_1, \ldots, x_p) \in R^p$$

If we let σ_{ij} and σ^{ij} denote the (i, j) element of respectively $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$, then the log likelihood equals

$$\begin{aligned} \log f_{\boldsymbol{\Sigma}}(\boldsymbol{x}) &= \sum_i x_i^2 \left(-\sigma^{ii}/2\right) + \sum_{i<j} x_i x_j \left(-\sigma^{ij}\right) - (p/2)\log 2\pi - (1/2)\log|\boldsymbol{\Sigma}| \\ &= \sum_i y_{ii}\theta^{ii} + \sum_{i<j} y_{ij}\theta^{ij} - \psi(\Theta) \text{ (say } l(\boldsymbol{y}; \Theta)), \end{aligned} \tag{4}$$

where $\Theta = (\theta^{ij})_{i\leq j}$ and $\boldsymbol{y} = (y_{ij})_{i\leq j}$ are given by

$$\begin{cases} \theta^{ii} = (-1/2)\sigma^{ii}, & i = 1, \ldots, p, \\ \theta^{ij} = -\sigma^{ij}, & 1 \leq i < j \leq p, \\ y_{ii} = x_i^2, & i = 1, \ldots, p, \\ y_{ij} = x_i x_j, & 1 \leq i < j \leq p, \end{cases} \tag{5}$$

and

$$\psi(\Theta) = (p/2)\log 2\pi + (1/2)\log|\boldsymbol{\Sigma}(\Theta)|. \tag{6}$$

The summations Σ_i, $\Sigma_{i<j}$ in the equation (4) are abbreviations respectively for $\sum_{i=1}^p$ and $\sum_{1\leq i<j\leq p}$, and we will use these kinds of notations implicitly hereafter.

The expression (4) gives natural coordinate system Θ of the manifold $\mathcal{S}$ as a full exponential family. Another coordinate system, so called expectation parameters, is also useful, which is defined as;

$$\sigma_{ij} = E(y_{ij}), \qquad 1 \leq i \leq j \leq p. \tag{7}$$

For the analysis of the information carried by $\boldsymbol{l}$ and $\boldsymbol{H}$, we need to prepare another coordinate system. The matrix exponential expression of an orthogonal matrix $\boldsymbol{O}$ is given by

$$\boldsymbol{O} = \exp \boldsymbol{U} = \boldsymbol{I}_p + \boldsymbol{U} + \frac{1}{2}\boldsymbol{U}^2 + \frac{1}{3!}\boldsymbol{U}^3 + \cdots, \tag{8}$$

where $\boldsymbol{I}_p$ is the p-dimensional unit matrix, $\boldsymbol{U}$ is a skew-symmetric matrix and parametrized by $\boldsymbol{u} = (u_{ij})_{1\leq i<j\leq p}$ as

$$(\boldsymbol{U})_{ij} = \begin{cases} u_{ij}, & \text{if } 1 \leq i < j \leq p, \\ -u_{ij}, & \text{if } 1 \leq j < i \leq p, \\ 0, & \text{if } 1 \leq i = j \leq p. \end{cases}$$

The function $\exp \boldsymbol{U}$ is diffeomorphic, and $\boldsymbol{u}$ gives "normal coordinate" for the group of orthogonal matrices (see (6.7) in Boothby[6] or Th. A9.11 of Muirhead[21]). We can use this coordinate as local system around $\boldsymbol{I}_p$ and construct an atlas for the entire space of p-dimensional orthogonal matrices (note this space is compact); for each $\boldsymbol{\Gamma}$, there exists an open neighborhood and some open ball B in $R^{p(p-1)/2}$ around the origin such that these spaces are diffeomorphic by the function $\boldsymbol{\Gamma} \exp \boldsymbol{U}(\boldsymbol{u})$ on B.

We will use $(\boldsymbol{\lambda}, \boldsymbol{u})$ as the third coordinate system of $\mathcal{S}$ and call it "spectral coordinate (system)". Notice that this coordinate system is associated with the following submanifolds in $\mathcal{S}$. If we fix $\boldsymbol{\Gamma}$ in (1), then we get a submanifold $\mathcal{M}(\boldsymbol{\Gamma})$ embedded in $\mathcal{S}$ with a coordinate system $\boldsymbol{\lambda}$. This is a subfamily in $N(\mathbf{0}, \boldsymbol{\Sigma})$ and called curved exponential family. Its log-likelihood is expressed, as we emphasize it as a function of $\boldsymbol{\lambda}$, to be

$$l(\boldsymbol{y}; \Theta(\boldsymbol{\lambda})) = \sum_i y_{ii}\theta^{ii}(\boldsymbol{\lambda}) + \sum_{i<j} y_{ij}\theta^{ij}(\boldsymbol{\lambda}) - \psi(\Theta(\boldsymbol{\lambda})). \tag{9}$$

On the contrary, if we fix $\boldsymbol{\lambda}$ in (1), we get another submanifold $\mathcal{A}(\boldsymbol{\lambda})$ in $\mathcal{S}$, whose coordinate system is given by $\boldsymbol{u}$ in a neighborhood of each point of $\mathcal{A}(\boldsymbol{\lambda})$. Its log-likelihood expression is given by

$$l(\boldsymbol{y}; \Theta(\boldsymbol{u})) = \sum_i y_{ii}\theta^{ii}(\boldsymbol{u}) + \sum_{i<j} y_{ij}\theta^{ij}(\boldsymbol{u}) - \psi(\Theta(\boldsymbol{u})). \tag{10}$$

First we consider a metric, that is, a field of symmetric, positive definite, bilinear form on $\mathcal{S}$. The statistically most natural metric is Fisher information metric. Suppose $\{f(\boldsymbol{x}; \boldsymbol{\theta})\}$ is a parametric family of probability density functions, whose coordinate as a manifold is given by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Then the (i, j) component of Fisher information metric with respect to $\boldsymbol{\theta}$ is given by

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_i} \log f(\boldsymbol{x}; \theta) \frac{\partial}{\partial\theta_j} \log f(\boldsymbol{x}; \theta)\right].$$

For the multivariate normal distribution family, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{\mu}$, the mean parameter is also included), Skovgaard [25] gives a clear form of Fisher information metric. The tangent vector space at a fixed point $\boldsymbol{\Sigma}$ w.r.t. $(\sigma_{ij})_{i\le j}$ coordinate can be identified with the space of symmetric matrices. For any symmetric matrix $\boldsymbol{A}$, $\boldsymbol{B}$, the metric with respect to the $\boldsymbol{\Sigma} = (\sigma_{ij})$ coordinate system is given by

$$\frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\Big). \tag{11}$$

We are interested in Fisher information metric with respect to the spectral coordinate $(\boldsymbol{\lambda}, \boldsymbol{u})$. Let ∂_a, ∂_b, $\cdots$ denote the tangent vectors w.r.t. the $\boldsymbol{\lambda}$ coordinate, $\partial_{(s,t)}$, $\partial_{(u,v)}, \cdots$ denote the tangent vectors w.r.t. the $\boldsymbol{u}$ coordinate. Namely

$$\partial_a \triangleq \frac{\partial}{\partial\lambda_a}, \qquad \partial_{(s,t)} \triangleq \frac{\partial}{\partial u_{st}}.$$

These tangent vectors (exactly speaking, vector fields) are invariant with respect to any orthogonal transformation of $\boldsymbol{\Sigma}$; For some orthogonal matrix $\boldsymbol{O}$, an orthogonal transformation $F$of $\mathcal{S}$ is defined as

$$F(\boldsymbol{\Sigma}) = \boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{O}^t \tag{12}$$

For any $\boldsymbol{O}$,

$$\begin{aligned} F_*(\partial_a) &= \partial_a, & 1 \le a \le p, \qquad (13)\\ F_*(\partial_{(s,t)}) &= \partial_{(s,t)}, & 1 \le s < t \le p, \qquad (14) \end{aligned}$$

where F_* is the derivative of F.

Proposition 1 *Let $\langle\ ,\ \rangle$ denote Fisher information metric based on $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, then the components of the metric with respect to $(\boldsymbol{\lambda}, \boldsymbol{u})$ is given as follows;*

$$\begin{aligned} & g_{ab} \triangleq \langle \partial_a, \partial_b \rangle = (1/2)\lambda_a^{-2}\, \delta\big(a = b\big) && 1 \le a, b \le p,\\ & g_{a(s,t)} \triangleq \langle \partial_a, \partial_{(s,t)} \rangle = 0 && 1 \le a \le p,\ 1 \le s < t \le p,\\ & g_{(s,t)(u,v)} \triangleq \langle \partial_{(s,t)}, \partial_{(u,v)} \rangle &&\\ & \quad = (\lambda_s - \lambda_t)^2 \lambda_s^{-1} \lambda_t^{-1}\, \delta\big((s,t) = (u,v)\big) && 1 \le s < t \le p,\ 1 \le u < v \le p. \end{aligned}$$

$\delta(\cdot)$ equals one if the logic inside the parenthesis is correct, otherwise zero.

There are two remarkable properties of the metric for the spectral coordinate. First note that since the metric components matrix is diagonal, $(\boldsymbol{\lambda}, \boldsymbol{u})$ is an orthogonal coordinate system, especially that the submanifolds $\mathcal{M}(\boldsymbol{\Gamma})$ and $\mathcal{A}(\boldsymbol{\lambda})$ are orthogonal to each other for any $\boldsymbol{\lambda}$ and $\boldsymbol{\Gamma}$. Second it is independent of $\boldsymbol{\Gamma}$, hence the metric stays constant with respect to the orthogonal transformation F in (12) for any orthogonal matrix $\boldsymbol{O}$. (Second property is instantly derived from the expression (11).)

Theoretically, other metrics could be naturally implemented. Calvo and Oller [7] introduced Sigel metric. Lovrić et al. [19] considered the natural invariant metric from the standpoint of Riemannian symmetric space. The concrete forms of the both metrics are given by (3.4) and (3.2) in [19]. (The information metric (11) corresponds to (3.3) in [19]. See also Theorem 1 of Zhang [30].)

Once a metric is given on the manifold $\mathcal{S}$, a connection is needed for further geometrical analysis. Connection is an important "rule" which defines how a tangent space is shifted with an infinitesimal move in a differential manifold. Although connection has an infinite variation, the most commonly used one is Levi-Civita connection. It is characterized as a unique torsion-free, metric-preserving connection. This connection is essential to consider a distance function on the manifold. Skovgaard [25] , Calvo and Oller [7], Fletcher and Joshi [12], Lenglet et al. [18], Lovrić et al. [19], Moakhaer and Zéraï [20] analyze the manifold of the normal distributions under Levi-Civita connection.

On the other hand, Amari [1] showed that "α-connection" is suitable for statistical manifolds in general. He also found that e-connection ($\alpha = 1$) and m-connection ($\alpha = -1$) are especially important for the asymptotic analysis of information loss for a curved exponential family. Amari and Kumon [4], Kumon, Amari [16] and Eguchi [11] gave further development along this line. Specifically in the relation with the multivariate normal distribution or $\mathcal{S}$, Ohara et al. [23], Yoshizawa and Tanabe [29] and Zhang et al. [30] considered the dual geometry (α and $-\alpha$ connections) of the manifolds. Notice that Levi-Civita connection is 0-connection and the "mean" between e-connection and m-connection. Therefore, using the results on geometric properties of $\mathcal{S}$ under e-connection and m-connection, we could also derive those under Levi-Civita connection.

Since this paper is aimed for the statistical inference on $\boldsymbol{\Sigma}$, we adopt α-connections, especially e- and m-connections, hereafter. We conclude this section by mentioning the important fact that $\mathcal{S}$ is e-flat and m-flat, and corresponding affine coordinates are given respectively by (σ^{ij}) and (σ_{ij}).

3 Embedding Curvatures

Curvature, which is important property for an geometrical analysis, is defined based on a given connection. A submanifold has both intrinsic and extrinsic curvatures. The latter describes how the submanifold is placed in the whole manifold, and called an embedding curvature or the second fundamental form. (The first fundamental form is the metric.)

In this section, we observe the embedding curvatures of $\mathcal{M}$ and $\mathcal{A}$ for the analysis of the distribution $(\boldsymbol{l}, \boldsymbol{H})$. Specifically we consider the following embedding curvatures;

1. Embedding curvature of $\mathcal{M}$ with respect to e-connection or m-connection. Its components w.r.t the spectral coordinate are given by

$$\overset{e}{H}_{ab(s,t)} \triangleq \langle \overset{e}{\nabla}_{\partial_a}\partial_b \, , \partial_{(s,t)} \rangle, \qquad \overset{m}{H}_{ab(s,t)} \triangleq \langle \overset{m}{\nabla}_{\partial_a}\partial_b \, , \partial_{(s,t)} \rangle, \tag{15}$$

where $\overset{e}{\nabla}_{\partial_a}\partial_b$ is the covariant derivative of ∂_b in the direction of ∂_a with respect to e-connection. $\overset{m}{\nabla}_{\partial_a}\partial_b$ is similarly defined.

2. Embedding curvature of $\mathcal{A}$ with respect to m-connection. Its components w.r.t the spectral coordinate are given by

$$\overset{m}{H}_{(s,t)(u,v)a} \triangleq \langle \overset{m}{\nabla}_{\partial_{(s,t)}}\partial_{(u,v)} \, , \partial_a \rangle, \tag{16}$$

where $\overset{m}{\nabla}_{\partial_{(s,t)}}\partial_{(u,v)}$ is the covariant derivative of $\partial_{(s,t)}$ in the direction of $\partial_{(u,v)}$ with respect to m-connection.

On these curvatures at the point $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$, we have the following results.

Proposition 2 *For* $1 \leq a, b \leq p, \; 1 \leq s < t \leq p$,

$$\overset{e}{H}_{ab(s,t)} = \overset{m}{H}_{ab(s,t)} = 0. \tag{17}$$

For $1 \leq a \leq p,\ 1 \leq s < t \leq p,\ 1 \leq u < v \leq p,$

$$\overset{m}{H}_{(s,t)(u,v)a}= \begin{cases} \lambda_a^{-2}(\lambda_t - \lambda_a), & \textit{if } s = u = a,\ t = v, \\ \lambda_a^{-2}(\lambda_s - \lambda_a), & \textit{if } s = u,\ t = v = a, \\ 0, & \textit{otherwise.} \end{cases} \tag{18}$$

Another expression of the embedding curvature of $\mathcal{A}$ is given by

$$\overset{m}{H}{}^{a}_{(s,t)(u,v)} \triangleq \sum_b \overset{m}{H}_{(s,t)(u,v)b}\ g^{ba}, \tag{19}$$

With this notation, the orthogonal projection of the covariant derivative

$$\overset{m}{\nabla}_{\partial_{(s,t)}} \partial_{(u,v)}$$

onto the tangent space of $\mathcal{M}$ is given by

$$\sum_a \overset{m}{H}{}^{a}_{(s,t)(u,v)}\ \partial_a.$$

From Proposition 1, 2, we have

$$\overset{m}{H}{}^{a}_{(s,t)(u,v)}= 2(\lambda_t - \lambda_a)\delta(s = u = a,\ t = v) + 2(\lambda_s - \lambda_a)\delta(s = u,\ t = v = a), \tag{20}$$

hence

$$\sum_a \overset{m}{H}{}^{a}_{(s,t)(u,v)}\ \partial_a = \begin{cases} 2(\lambda_t - \lambda_s)\partial_s + 2(\lambda_s - \lambda_t)\partial_t, & \text{if } (s,t) = (u,v), \\ 0, & \text{otherwise.} \end{cases}$$

Similarly another embedding curvature components $\overset{e}{H}{}^{(s,t)}_{ce}$ is defined as

$$\overset{e}{H}{}^{(s,t)}_{ab}= \sum_{u<v} \overset{e}{H}_{ab(u,v)}\ g^{(u,v)(s,t)} \tag{21}$$

and actually it vanishes

$$\overset{e}{H}{}^{(s,t)}_{ab}= 0, \quad 1 \leq a, b \leq p,\ 1 \leq s < t \leq p. \tag{22}$$

An embedding curvature has full information about the "extrinsic curvature" of the embedded submanifold in any direction. Sometimes it is convenient to compress it into a scalar measure of the curvature. "Statistical curvature" by Efron (see Efron [10], Murray and Rice [22]) is such a measure; For $\mathcal{A}$, it is defined by (see p.159 of Amari [2])

$$\gamma(\mathcal{A}) \triangleq \sum_{1 \leq a,b \leq p} \ \sum_{s<t,u<v,o<p,q<r} \overset{m}{H}_{(s,t)(u,v)a}\ \overset{m}{H}_{(o,p)(q,r)b}\ g^{(s,t)(o,p)}\ g^{(u,v)(q,r)}\ g^{ab},$$

which attains the following value at the point $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$.

Corollary 1

$$\gamma(\mathcal{A}) = 2\sum_{a<b} \frac{\lambda_a^2 + \lambda_b^2}{(\lambda_a - \lambda_b)^2}$$

From these results, we notice that if $\mathcal{S}$ is endowed with m-connection, then 1) the embedding curvatures and the statistical curvatures of $\mathcal{A}$ are independent of $\boldsymbol{\Gamma}$, 2) any one-parameter curve $(\boldsymbol{\lambda}, \boldsymbol{\Gamma}(\boldsymbol{u}))$ given by a parameter $u_{(s,t)},\ s < t$, where $\boldsymbol{\lambda}$ and the other elements of $\boldsymbol{u}$ are fixed, is curved in the direction of $\partial_t - \partial_s$ and contained in a two-dimensional plane composed by $\partial_{(s,t)}$ and $\partial_t - \partial_s$, 3) the statistical curvature of $\mathcal{A}$ could be quite large when $\boldsymbol{\lambda}$ are close to each other, while $\mathcal{M}$ is flat everywhere.

Here we introduce another submanifold $\tilde{\mathcal{A}}$ which is contrasting to $\mathcal{A}$ in the sense that $\tilde{\mathcal{A}}$ is flat with respect to m-connection. For a point $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$, let

$$\tilde{\mathcal{A}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}) \triangleq \{\boldsymbol{\Sigma} \in \mathcal{S} \mid (\boldsymbol{\Gamma}^t \boldsymbol{\Sigma} \boldsymbol{\Gamma})_{ii} = \lambda_i,\ 1 \le \forall i \le p\}.$$

We easily notice that $\tilde{\mathcal{A}}$ is the minimum distance points with respect to Kullback-Leibler divergence. That is,

$$\tilde{\mathcal{A}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}) = \{\boldsymbol{\Sigma} \in \mathcal{S} \mid \operatorname{argmin}_{\tilde{\boldsymbol{\lambda}}} KL(\boldsymbol{\Sigma}, \boldsymbol{\Gamma}\mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p)\boldsymbol{\Gamma}^t) = \boldsymbol{\lambda}\},$$

where $KL(\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}})$ is the Kullback-Leibler divergence between $N(\mathbf{0}, \boldsymbol{\Sigma})$ and $N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$, which is specifically given by

$$\mathrm{tr}(\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}) - \log|\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}| - p.$$

The minimum distance points with respect to the Kullback-Leibler divergence consists of all the points on the m-geodesics which pass through the point $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ and are orthogonal to $\mathcal{M}(\boldsymbol{\Gamma})$ at that point. (See Theorem in A2 of Amari [1]).

We can visualize the structure of $\mathcal{S}$ endowed with m-connection for the two dimensional case. See Figure 1, where $\mathcal{M}_i \triangleq \mathcal{M}(\boldsymbol{\Gamma}_i),\ i = 1, \ldots, 3$, $\mathcal{A}_i \triangleq \mathcal{A}(\boldsymbol{\lambda}_i),\ i = 1, 2$ and $\tilde{\mathcal{A}}_1 \triangleq \tilde{\mathcal{A}}(\boldsymbol{\lambda}_1, \boldsymbol{\Gamma}_1)$ are drawn. When $p = 2$, $\mathcal{M}$ is a two-dimensional autoparallel submanifold with the affine coordinate (λ_1, λ_2), while $\mathcal{A}$ is a one-dimensional submanifold with an coordinate $u_{(1,2)}$. As it is seen in Proposition 1, all the tangent vectors $\partial_1(\triangleq \frac{\partial}{\partial\lambda_1})$, $\partial_2(\triangleq \frac{\partial}{\partial\lambda_2})$, $\partial_{(1,2)}(\triangleq \frac{\partial}{\partial u_{(1,2)}})$ are orthogonal to each other. $\tilde{\mathcal{A}}$ is a "straight" line which is also orthogonal to $\mathcal{M}$. The arrow on $\mathcal{M}$ is the line $\{\boldsymbol{\lambda} | \lambda_1 + \lambda_2 \text{ is constant}\}$, and the arrow head indicates the direction in which $c \triangleq \lambda_2/\lambda_1$ increases. The statistical curvature turns out to be the increasing function of c ;

$$\gamma(\mathcal{A}) = 2\frac{1 + c^2}{(1 - c)^2}.$$

We can analyze the bias of $\bar{l}_i \triangleq n^{-1} l_i,\ i = 1, \ldots, p$ from the geometrical structure of $\mathcal{S}$. It is well known that $E[\,\bar{l}_i\,]$ $(i = 1, \ldots, p)$ majorizes λ_i $(i = 1, \ldots, p)$, that is,

$$\sum_{i=1}^{j} E[\,\bar{l}_i\,] \ge \sum_{i=1}^{j} \lambda_i, \quad 1 \le \forall j \le p - 1, \qquad \sum_{i=1}^{p} E[\,\bar{l}_i\,] = \sum_{i=1}^{p} \lambda_i. \tag{23}$$

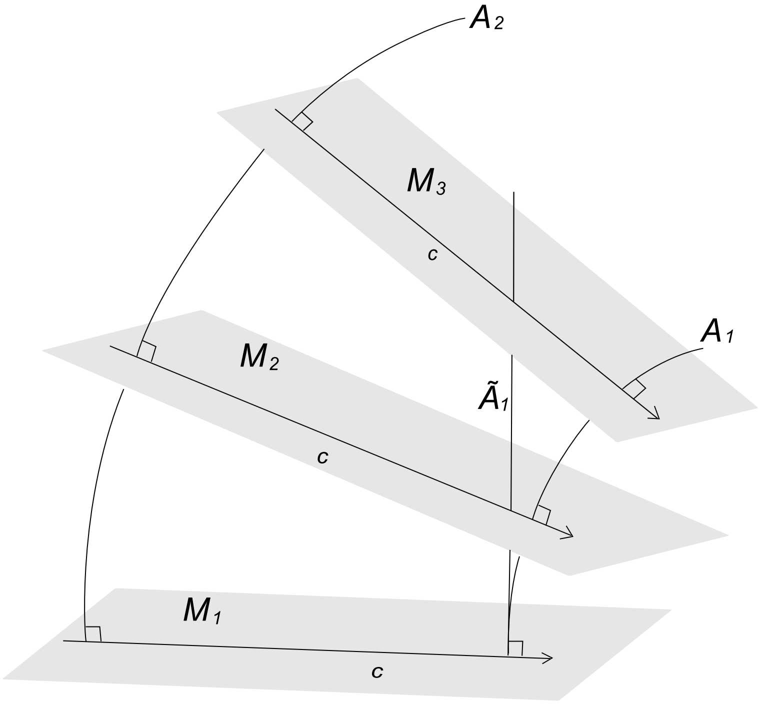

Figure 1: Submanifolds of $\mathcal{S}$ when $p = 2$, $\mathcal{M}$, $\mathcal{A}$ and $\tilde{\mathcal{A}}$

The bias $E[\,\bar{l}_i\,]$ is quite large when n is small and λ_i's are close to each other (see Lawley [17], Anderson [5]). For the case $p = 2$,

$$E[\,\bar{l}_1\,] \geq \lambda_1, \qquad E[\,\bar{l}_2\,] \leq \lambda_2, \qquad E[\,\bar{l}_1\,] + E[\,\bar{l}_2\,] = \lambda_1 + \lambda_2. \tag{24}$$

Suppose a sample $\bar{\boldsymbol{S}} \triangleq n^{-1}\boldsymbol{S}$ takes the value at a point $s \in \mathcal{S}$. Let s_1 denote the point on $\mathcal{M}(\boldsymbol{\Gamma})$ designated by the eigenvalues of $\bar{\boldsymbol{S}}$, namely $\bar{\boldsymbol{l}} \triangleq (\bar{l}_1, \bar{l}_2)$. The curve $\mathcal{A}(\bar{\boldsymbol{l}})$ connects s and s_1. If we define s_2 as the point on $\mathcal{M}(\boldsymbol{\Gamma})$ designated by $\hat{\boldsymbol{\lambda}} \triangleq (\hat{\lambda}_1, \hat{\lambda}_2) \triangleq ((\boldsymbol{\Gamma}^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma})_{11}, (\boldsymbol{\Gamma}^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma})_{22})$, then $\tilde{\mathcal{A}}(\hat{\boldsymbol{\lambda}}, \boldsymbol{\Gamma})$ connects s and s_2. The three points s, s_1 and s_2 are on the same plane, and if we move from s_1 in the direction to s_2, then the statistical curvature of $\mathcal{A}$ increases (see Figure 2). If we estimate (λ_1, λ_2) by $\bar{\boldsymbol{l}}$, then the estimate is the point s_1, while for the unbiased estimator $\hat{\boldsymbol{\lambda}}$, the estimate is the point s_2. Since the c-coordinate of s_1 is always smaller than that of s_2, the estimator $(\bar{l}_1, \bar{l}_2)$ is likely to estimate λ_1 and λ_2 too apart, which causes the bias (24). It is also seen that the bias gets larger when c approaches to one, that is, λ_1 and λ_2 get closer to each other.

Though the exact magnitude of the bias $E(\bar{l}_a) - \lambda_a$ is hard to evaluate, the asymptotic bias can be evaluated. This can be also described with embedding curvatures (see (5.4) of Amari [2]);

$$E(\bar{l}_a - \lambda_a) = -\frac{1}{2n}C^a + O(n^{-3/2}),$$

where

$$C^a = \sum_{c,d} \overset{m}{\Gamma}{}^a_{cd}\; g^{cd} + \sum_{s<t,u<v} \overset{m}{H}{}^a_{(s,t)(u,v)}\; g^{(s,t)(u,v)},$$

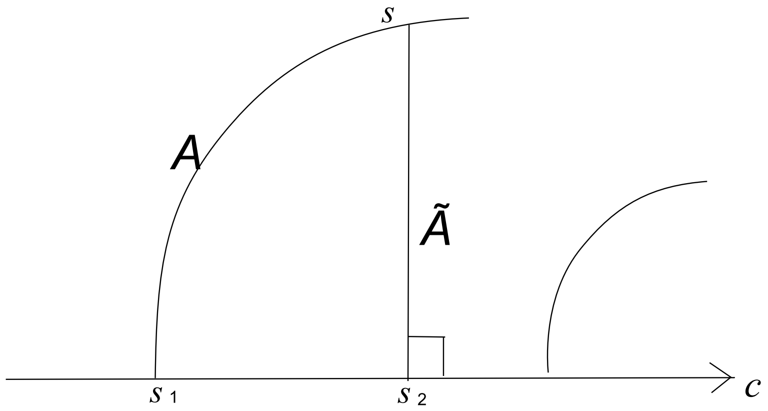

Figure 2: Horizontal perspective of $\mathcal{A}$ and $\tilde{\mathcal{A}}$ on the plane $\mathcal{M}$ when $p = 2$

and $\overset{m}{\Gamma}{}^a_{cd}$ is a m-connection coefficients of $\mathcal{M}$, which is defined by

$$\overset{m}{\Gamma}{}^a_{cd} = \overset{m}{\Gamma}_{cdb}\, g^{ba}, \quad \overset{m}{\Gamma}_{cdb} \triangleq \langle \overset{m}{\nabla}_{\partial_c} \partial_d \,, \partial_b \rangle. \tag{25}$$

Since $\mathcal{M}$ is autoparallel in m-flat $\mathcal{S}$,

$$\overset{m}{\Gamma}{}^a_{cd} = \overset{m}{\Gamma}_{cdb} = 0, \quad 1 \leq a, b, c, d \leq p. \tag{26}$$

Hence we have the following equation from Proposition 1 and (20).

$$\begin{aligned} C^a(\boldsymbol{\lambda}) &= \sum_{a<t} \overset{m}{H}{}^a_{(a,t)(a,t)}\, g^{(a,t)(a,t)} + \sum_{s<a} \overset{m}{H}{}^a_{(s,a)(s,a)}\, g^{(s,a)(s,a)} \\ &= 2 \sum_{t \neq a} \frac{\lambda_a \lambda_t}{\lambda_t - \lambda_a}. \end{aligned} \tag{27}$$

This bias was originally derived by the perturbation method in Lawley [17].

4 Estimation of $\boldsymbol{\lambda}$ when $\boldsymbol{\Gamma}$ is known

We consider an estimation problem when $\boldsymbol{\Gamma}$ is known to be $\boldsymbol{\Gamma}^0$. From a practical point of view, the case when $\boldsymbol{\Gamma}$ is known is not of much interest compared to the general case where both $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$ are unknown. However as we will show in this section, the asymptotic information loss caused by discarding the sample eigenvectors (Section 4.2) are closely related to the asymptotic variance difference between two certain estimators (Section 4.1). Both asymptotic variance and information loss are described with geometrical terms.

4.1 Asymptotic variance of the estimators of $\boldsymbol{\lambda}$

In a general term, the subfamily (submanifold) $\mathcal{M}(\boldsymbol{\Gamma}^0)(\triangleq \{\boldsymbol{\Sigma} \in \mathcal{S} | \boldsymbol{\Gamma}(\boldsymbol{\Sigma}) = \boldsymbol{\Gamma}^0\})$ in $\mathcal{S}$ is a "curved" exponential family, since it is a subfamily in an exponential family $\mathcal{S}$. In a usual case, a subfamily is not "flat", hence the term "curved" is used. However as you can see from (17), $\mathcal{M}(\boldsymbol{\Gamma}^0)$ is autoparallel in $m(e)$-flat $\mathcal{S}$, and intrinsically $m(e)$-flat (see e.g. Theorem 1.1 in [3]).

We are supposed to estimate unknown coordinate $\boldsymbol{\lambda}$ of $\mathcal{M}(\boldsymbol{\Gamma}^0)$ using an estimator $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ of some kind. An estimator $\hat{\boldsymbol{\lambda}}(\boldsymbol{S})$ is specified by its inverse image $\hat{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\lambda})$

$$\hat{\mathcal{A}}(\boldsymbol{\lambda}) \triangleq \hat{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\lambda}) = \{\boldsymbol{\Sigma} \in \mathcal{S} \,|\, \hat{\boldsymbol{\lambda}}(\boldsymbol{\Sigma}) = \boldsymbol{\lambda}\}. \tag{28}$$

This is another submanifold in $\mathcal{S}$, where we will use $\boldsymbol{u}$ as a coordinate system.

A consistent estimator $\hat{\boldsymbol{\lambda}}$ is called first-order (Fisher) efficient if the first order term (i.e. $O(n^{-1})$ order term) w.r.t. the asymptotic expansion of the variance (covariance) in n is minimized among all (regular) estimators. Correct the bias of the first-order efficient estimator $\hat{\boldsymbol{\lambda}}$ up to the term of order n^{-1}, and let it be denoted by $\hat{\boldsymbol{\lambda}}^* \triangleq (\hat{\lambda}_1^*, \ldots, \hat{\lambda}_p^*)$. Amari showed (see e.g. Theorem 4.4 in [3]) that its asymptotic variance can be described by the geometrical properties such as the metric and the embedding curvatures of $\mathcal{M}(\boldsymbol{\Gamma}^0)$ and $\hat{\mathcal{A}}$; For $1 \leq a, b \leq p$,

$$E[(\hat{\lambda}_a^* - \lambda_a)(\hat{\lambda}_b^* - \lambda_b)] = \frac{1}{n} g^{ab} + \frac{1}{2n^2}\{(\Gamma^{m}_{M})^{2ab} + 2(H^{e}_{M})^{2ab} + (H^{m}_{\hat{A}})^{2ab}\} + O(n^{-3}) \tag{29}$$

where

$$\begin{aligned}
(\Gamma^{m}_{M})^{2ab} &= \sum_{c,d,e,f} \overset{m}{\Gamma}{}^a_{cd} \, \overset{m}{\Gamma}{}^b_{ef} \, g^{ce} g^{df}, \\
(H^{e}_{M})^{2ab} &= \sum_{c,d,e,f,s<t,u<v} \overset{e}{H}{}^{(s,t)}_{ce} \, \overset{e}{H}{}^{(u,v)}_{df} \, g_{(s,t)(u,v)} \, g^{cd} g^{ea} g^{fb}, \\
(H^{m}_{\hat{A}})^{2ab} &= \sum_{s<t,u<v,o<p,q<r} \overset{m}{H}{}^a_{(s,t)(u,v)} \, \overset{m}{H}{}^b_{(o,p)(q,r)} \, g^{(s,t)(o,p)} g^{(u,v)(q,r)},
\end{aligned}$$

$\overset{m}{\Gamma}{}^a_{cd}$ and $\overset{e}{H}{}^{(s,t)}_{ce}$ are already defined in the previous section as the connection coefficients (see (25)) or the embedding curvature components (see (21)) of $\mathcal{M}$. They are defined independently of the particular estimator. $\overset{m}{H}{}^a_{(s,t)(u,v)}$ are the components of the embedding m-curvature of $\hat{\mathcal{A}}$, which differ among the estimators.

We apply this formula to the following two estimators, $\boldsymbol{l}^* = (l_1^*, \ldots, l_p^*)$ and $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$. The former is the bias-corrected sample eigenvalues, which is given, using (27), by

$$l_a^* = \bar{l}_a + \frac{1}{2n} C^a(\boldsymbol{l}) = \bar{l}_a + \frac{1}{n} \sum_{t \neq a} \frac{l_a l_t}{l_t - l_a}, \quad a = 1, \ldots, p, \tag{30}$$

and the latter is defined by

$$\hat{\lambda}_a = ((\boldsymbol{\Gamma}^0)^t \bar{\boldsymbol{S}} \boldsymbol{\Gamma}^0)_{aa}, \quad a = 1, \ldots, p, \tag{31}$$

which is (exactly) unbiased. In fact $\hat{\boldsymbol{\lambda}}$ is the maximum likelihood estimator for the case $\boldsymbol{\Gamma}$ is known. Notice that for $\boldsymbol{l}$, $\hat{\mathcal{A}}(\boldsymbol{\lambda}) = \mathcal{A}(\boldsymbol{\lambda})$ and that for

$\hat{\boldsymbol{\lambda}}$, $\hat{\mathcal{A}}(\boldsymbol{\lambda}) = \tilde{\mathcal{A}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}^0)$. The first-order efficiency of the both estimators are guaranteed by the orthogonality to $\mathcal{M}(\boldsymbol{\Gamma}^0)$ of $\mathcal{A}(\boldsymbol{\lambda})$ and $\tilde{\mathcal{A}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}^0)$.

The terms $(\Gamma^m_M)^{2ab}$ and $(H^e_M)^{2ab}$, which are related to the submanifold $\mathcal{M}$, hence common to the both estimators, vanish, because of (22) and (26). The term $(H^m_{\hat{A}})^{2ab}$ is different between the two estimators. As we observed in the previous section, $\mathcal{A}(\boldsymbol{\lambda})$ is not autoparallel in $\mathcal{S}$ (see (18)). On the other hand, $\tilde{\mathcal{A}}(\boldsymbol{\lambda}, \boldsymbol{\Gamma}^0)$ is autoparallel in $\mathcal{S}$, hence $(H^m_{\hat{A}})^{2ab}$ vanishes. Consequently the following results are gained.

Proposition 3 *For* $1 \le a, b \le p$,

$$\begin{aligned} &E[(l^*_a - \lambda_a)(l^*_b - \lambda_b)](\triangleq V^{ab}(\boldsymbol{l}^*)) \\ &= \begin{cases} \dfrac{2}{n}\lambda_a^2 + \dfrac{2}{n^2}\displaystyle\sum_{t \ne a} \dfrac{\lambda_a^2 \lambda_t^2}{(\lambda_t - \lambda_a)^2} + O(n^{-3}), & \textit{if } a = b, \\ \\ -\dfrac{2}{n^2}\dfrac{\lambda_a^2 \lambda_b^2}{(\lambda_a - \lambda_b)^2} + O(n^{-3}), & \textit{if } a \ne b. \end{cases} \end{aligned} \tag{32}$$

$$\begin{aligned} &E[(\hat{\lambda}_a - \lambda_a)(\hat{\lambda}_b - \lambda_b)](\triangleq V^{ab}(\hat{\boldsymbol{\lambda}})) \\ &= \begin{cases} \dfrac{2}{n}\lambda_a^2 + O(n^{-5/2}), & \textit{if } a = b, \\ \\ O(n^{-5/2}), & \textit{if } a \ne b. \end{cases} \end{aligned} \tag{33}$$

This result says that $\hat{\boldsymbol{\lambda}}$ is the second-order efficient (among the bias-corrected first-order efficient estimators), but the bias-corrected sample eigenvalues are not. The difference in the asymptotic performance between the two estimators is due to the fact $\boldsymbol{l}^*$ do not use the prior information $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^0$, while $\hat{\boldsymbol{\lambda}}$ does. In contrast to $\boldsymbol{l}^*$, which does not use $\boldsymbol{H}$, $\hat{\boldsymbol{\lambda}}$ incorporates the information of $\boldsymbol{H}$ with the aid of the prior knowledge $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^0$. In fact, as we will see in the next subsection, the difference between (32) and (33) is closely related to the asymptotic information loss caused by discarding $\boldsymbol{H}$.

4.2 Asymptotic Information Loss

In this subsection, we consider the asymptotic information loss caused by ignoring $\boldsymbol{H}$ for the estimation of $\boldsymbol{\lambda}$. Information loss matrix $(\Delta g_{ab}(\boldsymbol{l}))$, $1 \le a, b \le p$ at a fixed point $\boldsymbol{\Sigma} = (\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ is given by

$$\Delta g_{ab}(\boldsymbol{l}) \triangleq E[g_{ab}(\boldsymbol{S}|\boldsymbol{l})] = g_{ab}(\boldsymbol{S}) - g_{ab}(\boldsymbol{l}),$$

where $g_{ab}(\boldsymbol{S}), g_{ab}(\boldsymbol{l}), g_{ab}(\boldsymbol{S}|\boldsymbol{l})$ are the components of the metrics w.r.t. ∂_a and ∂_b based on respectively the distributions $\boldsymbol{S}$, $\boldsymbol{l}$ and the conditional distribution of $\boldsymbol{S}$ given $\boldsymbol{l}$, all of which are measured at the point $\boldsymbol{\Sigma} = (\boldsymbol{\lambda}, \boldsymbol{\Gamma})$.

Amari [1] found that the asymptotic information loss can be expressed in terms of the metric and the embedding curvatures;

$$\begin{aligned}\Delta g_{ab}(\boldsymbol{l}) = n &\sum_{s<t,u<v} g_{a(s,t)} g_{b(u,v)} g^{(s,t)(u,v)} \\ &+ \sum_{c,d,s<t,u<v} \overset{e}{H}_{ac(s,t)}\ \overset{e}{H}_{bd(u,v)}\ g^{cd}\ g^{(s,t)(u,v)} \\ &+ (1/2) \sum_{s<t,u<v,o<p,q<r} \overset{m}{H}_{(s,t)(u,v)a}\ \overset{m}{H}_{(o,p)(q,r)b}\ g^{(s,t)(o,p)}\ g^{(u,v)(q,r)} \\ &+ O(n^{-1}). \qquad (34)\end{aligned}$$

Straightforward calculation leads us to the following result:

Proposition 4

$$\Delta g_{ab}(\boldsymbol{l}) = B_{ab} + O(n^{-1}),$$

where

$$B_{ab} = \begin{cases} \dfrac{1}{2\lambda_a^2} \sum_{t \neq a} \dfrac{\lambda_t^2}{(\lambda_t - \lambda_a)^2}, & \textit{if } a = b, \\ -\dfrac{1}{2(\lambda_a - \lambda_b)^2}, & \textit{if } a \neq b. \end{cases}$$

B_{ab} at the point $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ depends only on $\boldsymbol{\lambda}$. When the information loss of a statistic has the order $O(n^{-q+1})$, we call the statistic is the qth order sufficient. Consequently the statistic $\boldsymbol{l}$ is the first order sufficient, but not the second order sufficient.

B_{ab}, the information loss in the second order term ($O(1)$) could be quite large when the population eigenvalues are close to each other. Note that the information carried by $\boldsymbol{l}$ is given by the formula;

$$\begin{aligned} g_{ab}(\boldsymbol{l}) &= g_{ab}(\boldsymbol{S}) - \Delta g_{ab}(\boldsymbol{l}) \\ &= n g_{ab}(\boldsymbol{x}) - \Delta g_{ab}(\boldsymbol{l}) \\ &= (n/2)\lambda_a^{-2}\delta(a = b) - \Delta g_{ab}(\boldsymbol{l}). \end{aligned}$$

Since $(g_{ab}(\boldsymbol{l}))$ is positive definite, $\mathrm{diag}(n2^{-1}\lambda_1^{-2}, \ldots, n2^{-1}\lambda_p^{-2}) > (\Delta g_{ab})$. This holds true even in the neighborhood of a point $\lambda_1 = \cdots = \lambda_p$ where B_{ab} diverges. This indicates that the term of order $O(n^{-1})$ in $\Delta g_{ab}(\boldsymbol{l})$ is also unbounded in such a neighborhood. Hence the expansion of the information loss with respect to n is not useful when the population eigenvalues are close to each other.

Except for the case where the population eigenvalues are close to each other, Proposition 4 tells us approximately how much information is lost by ignoring the sample eigenvectors for the inference on the population eigenvalues. If we contract Δg_{ab}, then we could get a scalar measure on the information loss;

$$IL \triangleq \sum_{a,b} g^{ab} \Delta g_{ab} = \sum_a 2\lambda_a^2 B_{aa} + O(n^{-1}) = \sum_{a<b} \frac{\lambda_a^2 + \lambda_b^2}{(\lambda_a - \lambda_b)^2} + O(n^{-1})$$

Table 1: Simulate risk of $\boldsymbol{l}^*$ when $p = 2$ as c varies

c : Second Eigenvalue	1.0	0.8	0.6	0.4	0.2
Simulated Risk of $\boldsymbol{l}^*$	0.85	0.83	0.70	0.60	0.50
Standard Deviation	0.24	0.48	0.15	0.09	0.22
100×(Risk Difference/Risk of $\hat{\boldsymbol{\lambda}}$)	111	107	75	49	24

Asymptotic information loss is closely related to the asymptotic variance of the two estimators $\boldsymbol{l}^*$ and $\hat{\lambda}$ in the previous subsection. Actually if we contract the asymptotic performance difference between the two estimators $V^{ab}(\boldsymbol{l}^*) - V^{ab}(\hat{\lambda})$, then it equals $n^{-2}IL$, that is,

$$\begin{aligned}
&\sum_{a,b}(V^{ab}(\boldsymbol{l}^*) - V^{ab}(\hat{\boldsymbol{\lambda}}))g_{ab} \\
&= 2^{-1}E[\sum_{a}(l^*_a/\lambda_a - 1)^2] - 2^{-1}E[\sum_{a}(\hat{\lambda}_a/\lambda_a - 1)^2] \\
&= n^{-2}\sum_{a<b}\frac{\lambda_a^2 + \lambda_b^2}{(\lambda_a - \lambda_b)^2} + O(n^{-3}) = n^{-2}IL. \qquad (35)
\end{aligned}$$

As a numerical example, we made a simulation for the case $p = 2,\ n = 20$. Taking the relationship (35) into account, we could measure an information loss as the normalized quadratic risk difference between $\boldsymbol{l}^*$ and $\hat{\boldsymbol{\lambda}}$. We randomly generated a two-dimensional normal vector under the following conditions, $\boldsymbol{\Sigma} = \mathrm{diag}(1.0,\ c),\ c = 0.2, 0.4, 0.6, 0.8, 1.0$. We made 10^8 times repetition and took the average for each condition. The Table 1 shows the result. (Note: 1)The risk of $\hat{\boldsymbol{\lambda}}$ theoretically equals 0.4. 2)The simulated risk of $\boldsymbol{l}^*$ is quite unstable as its large s.d. shows.) We notice that information loss is not negligible. The risk of $\boldsymbol{l}^*$ is larger than that of $\hat{\boldsymbol{\lambda}}$ by 24–111 %. The risk difference is quite large especially when the population eigenvalues are close to each other.

5 Estimation of $\boldsymbol{\lambda}$ when $\boldsymbol{\Gamma}$ is unknown

In this section, we consider the more practical case where $\boldsymbol{\Gamma}$ is unknown. The derivation of a new estimator for this case will be done in view of the modification of the bias of $\bar{\boldsymbol{l}}$. Actually almost all the literature on the estimation of $\boldsymbol{\lambda}$ we mentioned in Section 1 modify the bias of $\bar{\boldsymbol{l}}$ by so called "shrinkage" method, that is, decreasing the dispersion of $\bar{\boldsymbol{l}}$. Though the concrete methods of shrinkage differ for each estimator, they are proposed mainly from analytical motivations. Here we consider another shrinkage estimator from a geometrical point of view.

Suppose that we have a sample $\bar{\boldsymbol{S}} \triangleq n^{-1}\boldsymbol{S}$ which takes the point $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ in $\mathcal{S}$, that is, $\boldsymbol{\lambda} = \bar{\boldsymbol{l}}, \boldsymbol{\Gamma} = \boldsymbol{H}$. (See Figure 3.) Take the orthogonal projection of this point onto the submanifold $\mathcal{M}(\boldsymbol{\Gamma}_i) \triangleq \mathcal{M}_i (i = 1, 2)$, where the projected

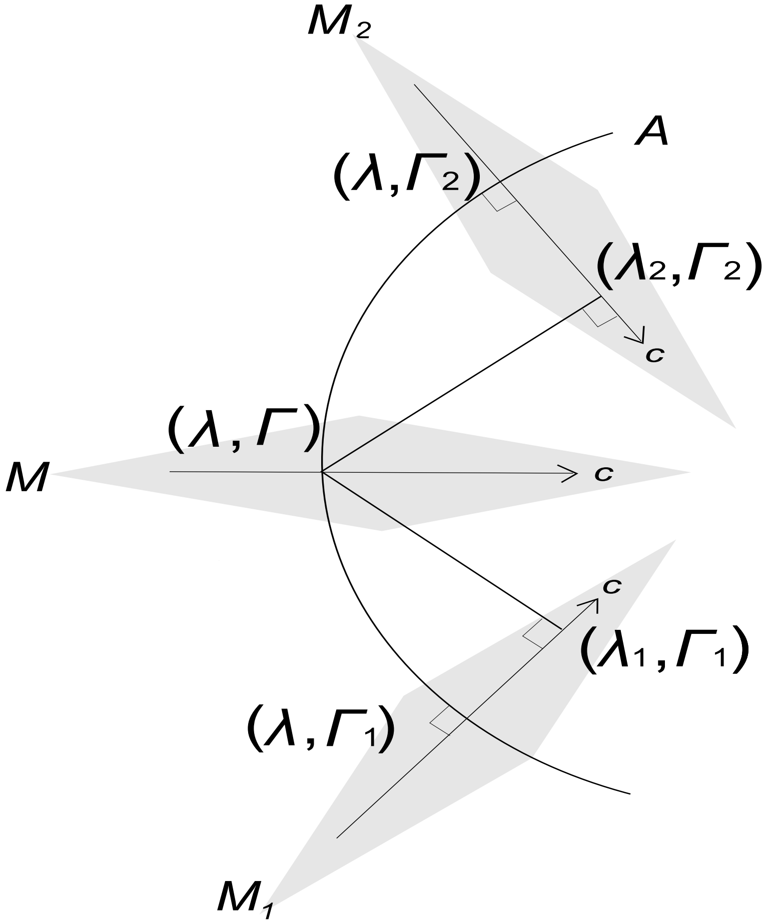

Figure 3: The shrinkage effect of the projection $(\boldsymbol{\lambda},\boldsymbol{\Gamma})$ onto $\mathcal{M}_i,\ i=1,2$

point $(\boldsymbol{\lambda}_i,\boldsymbol{\Gamma}_i)$ is given by $\boldsymbol{\lambda}_i=((\boldsymbol{\Gamma}_i^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma}_i)_{11},\ldots,(\boldsymbol{\Gamma}_i^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma}_i)_{pp})$. As we mentioned in Section 3, $(\boldsymbol{\lambda}_i,\boldsymbol{\Gamma}_i)$ is the minimum distance point on $\mathcal{M}_i$ from $(\boldsymbol{\lambda},\boldsymbol{\Gamma})$ with respect to Kullback-Leibler divergence. It is clearly understood that this projection has the shrinkage effect. If we have an appropriate probability measure of $\boldsymbol{\Gamma}$ on the group of p-dimensional orthogonal matrices $\mathcal{O}(p)$, the expectation of $(\boldsymbol{\Gamma}^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma})_{ii}, i=1,\ldots,p$ for that measure would give birth to a natural shrinkage estimator.

We choose the conditional distribution of $\boldsymbol{H}$ when $\boldsymbol{l}$ is given for the probability measure on $\mathcal{O}(p)$. Since $\boldsymbol{S}=\boldsymbol{H}\boldsymbol{L}\boldsymbol{H}^t$ is distributed as Wishart matrix $W_p(n,\boldsymbol{\Sigma})$, its density w.r.t. the uniform probability $d\mu(\boldsymbol{H})$ on $\mathcal{O}(p)$ equals

$$f(\boldsymbol{H}|\boldsymbol{l}\,;\boldsymbol{\Sigma})=K(\boldsymbol{l}\,;\boldsymbol{\Sigma})^{-1}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{H}\boldsymbol{L}\boldsymbol{H}^t\boldsymbol{\Sigma}^{-1}\Big), \qquad (36)$$

where normalizing constant $K(\boldsymbol{l}\,;\boldsymbol{\Sigma})$ is given by

$$K(\boldsymbol{l}\,;\boldsymbol{\Sigma})=\int_{\mathcal{O}(p)}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{H}\boldsymbol{L}\boldsymbol{H}^t\boldsymbol{\Sigma}^{-1}\Big)d\mu(\boldsymbol{H}).$$

This conditional distribution depends on $\boldsymbol{\Sigma}$. If we substitute $\boldsymbol{\Sigma}$ with an estimator $\hat{\boldsymbol{\Sigma}}(\bar{\boldsymbol{S}})$, it gives a distribution on $\mathcal{O}(p)$, whose density with respect to $d\mu(\boldsymbol{\Gamma})$ is given by

$$f(\boldsymbol{\Gamma}|\boldsymbol{l}\,;\hat{\boldsymbol{\Sigma}}) = K(\boldsymbol{l}\,;\bar{\boldsymbol{S}})^{-1}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}^t\hat{\boldsymbol{\Sigma}}^{-1}\Big), \tag{37}$$

where

$$K(\boldsymbol{l}\,;\hat{\boldsymbol{\Sigma}}) = \int_{\mathcal{O}(p)}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}^t\hat{\boldsymbol{\Sigma}}^{-1}\Big)d\mu(\boldsymbol{\Gamma}).$$

Take the expectation of $(\boldsymbol{\Gamma}^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma})_{ii}$ w.r.t. the density (37), then we have

$$\hat{\lambda}_i^* \triangleq K(\boldsymbol{l}\,;\bar{\boldsymbol{S}})^{-1}\int_{\mathcal{O}(p)}(\boldsymbol{\Gamma}^t\bar{\boldsymbol{S}}\boldsymbol{\Gamma})_{ii}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}^t\hat{\boldsymbol{\Sigma}}^{-1}\Big)d\mu(\boldsymbol{\Gamma}), \quad i = 1,\dots,p. \tag{38}$$

We propose $\hat{\boldsymbol{\lambda}}^* \triangleq (\hat{\lambda}_1^*,\dots,\hat{\lambda}_p^*)$ as a new estimator of $\boldsymbol{\lambda}$.

If $\hat{\boldsymbol{\Sigma}}$ is given by an orthogonally invariant estimator (2), $\hat{\lambda}_i^*$ can be more specifically described. Let $\bar{\boldsymbol{L}}$ denote diag$(\bar{\boldsymbol{l}})$. Because of the invariance of $d\mu$, it turns out that

$$\begin{aligned}\hat{\lambda}_i^* &= K(\boldsymbol{l})^{-1}\int_{\mathcal{O}(p)}(\boldsymbol{\Gamma}^t\boldsymbol{H}\bar{\boldsymbol{L}}\boldsymbol{H}^t\boldsymbol{\Gamma})_{ii}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{L}\boldsymbol{\Gamma}^t\boldsymbol{H}\boldsymbol{\Phi}^{-1}\boldsymbol{H}^t\boldsymbol{\Gamma}\Big)d\mu(\boldsymbol{\Gamma})\\ &= K(\boldsymbol{l})^{-1}\int_{\mathcal{O}(p)}(\boldsymbol{\Gamma}^t\bar{\boldsymbol{L}}\boldsymbol{\Gamma})_{ii}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{L}\boldsymbol{\Gamma}^t\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}\Big)d\mu(\boldsymbol{\Gamma}), \qquad (39)\end{aligned}$$

where

$$K(\boldsymbol{l}) \triangleq \int_{\mathcal{O}(p)}\exp\Big(-(1/2)\mathrm{tr}\boldsymbol{L}\boldsymbol{\Gamma}^t\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}\Big)d\mu(\boldsymbol{\Gamma}). \tag{40}$$

The analytic evaluation of this estimator's performance seems difficult even for the large sample case. Instead we show the numerical result comparing $\bar{\boldsymbol{l}}$, $\hat{\boldsymbol{\lambda}}^*$ and Stein's estimator (3). Our new estimator $\hat{\lambda}_i^*$ is also equipped with the same ϕ's in (3). We simulated the risks of three estimators for the case $p = 2,\ n = 10$ w.r.t. K-L loss, which is given by

$$\sum_{i=1}^p \hat{\lambda}_i\lambda_i^{-1} - \sum_{i=1}^p \log(\hat{\lambda}_i\lambda_i^{-1}) - p,$$

where $\hat{\lambda}_i = \bar{l}_i,\ \hat{\lambda}_i^*,\ \phi_i,\ i = 1,\dots,p.$ Since all the estimators are functions of $\boldsymbol{l}$ and scale invariant, it is enough to measure the risks for $\boldsymbol{\Sigma} = \mathrm{diag}(1,c),\ 0 < c \le 1$. We varied c from 0.04 to 1.00 by the increment 0.04, and for each c we repeated the risk evaluation 10^5 times and took the average. For the integral calculation of (39) and (40), we picked up 50 points from $\mathcal{O}(2)$ in an equidistant manner. Figure 4 shows the result. The new estimator performs better compared to $\bar{\boldsymbol{l}}$, especially $\boldsymbol{\lambda}$ are close to each other, though it seems that $\hat{\boldsymbol{\lambda}}^*$ does not dominate $\bar{\boldsymbol{l}}$ as Stein's estimator does. Unfortunately we do not have any theoretical explanation of the risk behavior of the new estimator. We could only guess that the shrinkage effect works well when c is close to one, while its effect is too strong elsewhere. We also simulated the risk of the new estimator equipped with M.L.E. instead of Stein's estimator. Since its performance is almost the same as the above new estimator, we skip the result.

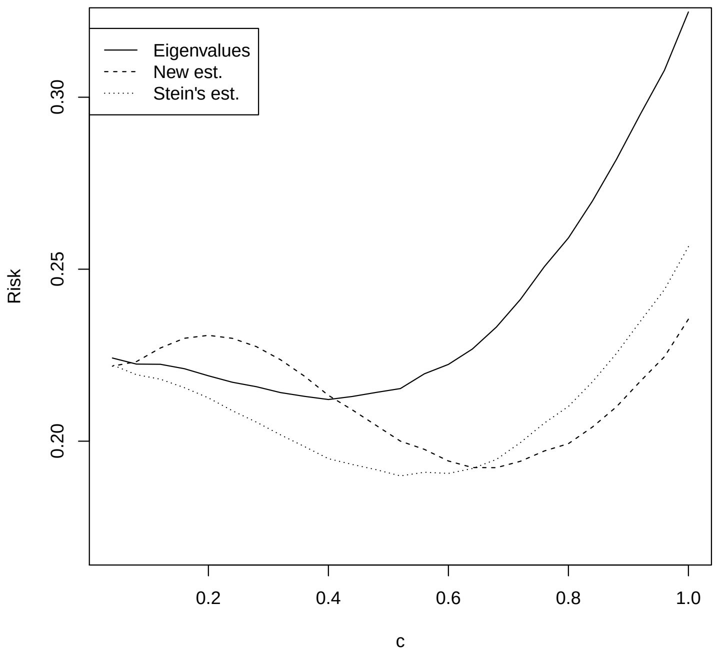

Figure 4: Risks of the three estimators as c changes

6 Remark

1. We treated the estimation problem of the eigenvalues $\boldsymbol{\lambda}$ in the latter half of the paper. The estimation on the eigenvectors $\boldsymbol{\Gamma}$ seems rather untouched in the classical situation $n \geq p$. Corollary 1 on the statistical curvatures of $\mathcal{A}$ or (27) on the asymptotic bias tells us that the point where $\boldsymbol{\lambda}$ has some multiplicity is a statistically singular point. Around these points, inference on $\boldsymbol{\Gamma}$ are considered to need subtle treatment. Especially the eigenvectors are not well identified around the multiplicity point, hence the information contained in $\boldsymbol{H}$ vanishes there (see $g_{(s,t)(u,v)}$ in Proposition 1). This indicates that the inference using only $\boldsymbol{H}$ is not appropriate.

2. We proposed a new estimator for $\boldsymbol{\lambda}$ in Section 5 . However this belongs to the same category as most estimators in the past literature in that it uses sample eigenvalues $\boldsymbol{\lambda}$ only. It is still unclear how we can use the sample eigenvalues $\boldsymbol{H}$ for the inference of $\boldsymbol{\lambda}$.

7 Appendix

7.1 Proof of Proposition 1

As a base for the vector space of real symmetric matrices, we consider $\boldsymbol{E}_{ij}$ $(1 \le i \le j \le p)$ which is a $p \times p$ matrix defined by

$$\boldsymbol{E}_{ij} = \begin{cases} \boldsymbol{I}_{ii}, & \text{if } i = j, \\ \boldsymbol{I}_{ij} + \boldsymbol{I}_{ji}, & \text{if } i < j, \end{cases}$$

where $\boldsymbol{I}_{ij}$ $(1 \le i, j \le p)$ is the $p \times p$ matrix whose (i, j) element equals one, and all the other elements are zero. The one to one correspondence

$$\partial^{(i,j)} \triangleq \frac{\partial}{\partial \sigma_{ij}} \longleftrightarrow \boldsymbol{E}_{ij}, \qquad 1 \le i \le j \le p,$$

gives the component expression of (11)

$$\langle \partial^{(i,j)}, \partial^{(k,l)} \rangle = \frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_{ij}\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_{kl}\Big), \qquad 1 \le i \le j \le p,\ 1 \le k \le l \le p.$$

Since

$$\partial_a \triangleq \frac{\partial}{\partial \lambda_a} = \sum_{i \le j} \frac{\partial \sigma_{ij}}{\partial \lambda_a}\frac{\partial}{\partial \sigma_{ij}} = \sum_{i \le j} \frac{\partial \sigma_{ij}}{\partial \lambda_a}\partial^{(i,j)} \quad 1 \le a \le p, \tag{41}$$

$$\partial_{(s,t)} \triangleq \frac{\partial}{\partial u_{st}} = \sum_{i \le j} \frac{\partial \sigma_{ij}}{\partial u_{st}}\frac{\partial}{\partial \sigma_{ij}} = \sum_{i \le j} \frac{\partial \sigma_{ij}}{\partial u_{st}}\partial^{(i,j)} \quad 1 \le s < t \le p, \tag{42}$$

we have the following relations

$$g_{ab} = \frac{1}{2}\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\sum_{i \le j}\frac{\partial \sigma_{ij}}{\partial \lambda_a}E_{ij}\Big)\boldsymbol{\Sigma}^{-1}\Big(\sum_{k \le l}\frac{\partial \sigma_{kl}}{\partial \lambda_b}E_{kl}\Big)\Big\}, \tag{43}$$

$$g_{a(s,t)} = \frac{1}{2}\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\sum_{i \le j}\frac{\partial \sigma_{ij}}{\partial \lambda_a}E_{ij}\Big)\boldsymbol{\Sigma}^{-1}\Big(\sum_{k \le l}\frac{\partial \sigma_{kl}}{\partial u_{st}}E_{kl}\Big)\Big\}, \tag{44}$$

$$g_{(s,t)(u,v)} = \frac{1}{2}\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\sum_{i \le j}\frac{\partial \sigma_{ij}}{\partial u_{st}}E_{ij}\Big)\boldsymbol{\Sigma}^{-1}\Big(\sum_{k \le l}\frac{\partial \sigma_{kl}}{\partial u_{uv}}E_{kl}\Big)\Big\}, \tag{45}$$

where $1 \le a, b \le p$, $1 \le s < t \le p$, $1 \le u < v \le p$.

For the first order derivative at $\boldsymbol{u} = \boldsymbol{0}$, we only have to consider $\boldsymbol{\Sigma}$ up to the term to the first power w.r.t. $\boldsymbol{u}$, hence we put $\boldsymbol{\Sigma}(\boldsymbol{\lambda}, \boldsymbol{u})$ as

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{\lambda}, \boldsymbol{u}) &= \boldsymbol{\Gamma}(\boldsymbol{I}_p + \boldsymbol{U})\boldsymbol{\Lambda}(\boldsymbol{I}_p + \boldsymbol{U})^t\boldsymbol{\Gamma}^t + O(||\boldsymbol{u}||^2) \\ &= \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t + \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{U}^t\boldsymbol{\Gamma}^t + \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t + O(||\boldsymbol{u}||^2). \end{aligned} \tag{46}$$

Therefore we have

$$\sigma_{ij} = \sum_k \gamma_{ik}\gamma_{jk}\lambda_k + \sum_{k,l}\gamma_{ik}\gamma_{jl}\Big(\lambda_k u_{lk} + \lambda_l u_{kl}\Big) + O(||\boldsymbol{u}||^2), \quad 1 \le i \le j \le p,$$

where $u_{ii} \triangleq 0$ $(1 \leq i \leq p)$, $u_{ij} \triangleq -u_{ji}$ $(1 \leq j < i \leq p)$, which leads to

$$\left.\frac{\partial \sigma_{ij}}{\partial \lambda_a}\right|_{\boldsymbol{u}=\boldsymbol{0}} = \gamma_{ia}\gamma_{ja}, \tag{47}$$

and

$$\left.\frac{\partial \sigma_{ij}}{\partial u_{st}}\right|_{\boldsymbol{u}=\boldsymbol{0}} = \lambda_t \gamma_{it}\gamma_{js} - \lambda_s \gamma_{is}\gamma_{jt} + \lambda_t \gamma_{is}\gamma_{jt} - \lambda_s \gamma_{it}\gamma_{js}. \tag{48}$$

From (47) and (48), we have the following results on tangent vectors;

$$\sum_{i \leq j} \frac{\partial \sigma_{ij}}{\partial \lambda_a} E_{ij} = \sum_{i \leq j} \gamma_{ia}\gamma_{ja}\boldsymbol{E}_{ij} = \boldsymbol{\gamma}_a \boldsymbol{\gamma}_a^t, \tag{49}$$

where $\boldsymbol{\gamma}_a$ is the ath column of $\boldsymbol{\Gamma}$, and

$$\sum_{i \leq j} \frac{\partial \sigma_{ij}}{\partial u_{st}} E_{ij} = \lambda_t \boldsymbol{\gamma}_t \boldsymbol{\gamma}_s^t - \lambda_s \boldsymbol{\gamma}_s \boldsymbol{\gamma}_t^t + \lambda_t \boldsymbol{\gamma}_s \boldsymbol{\gamma}_t^t - \lambda_s \boldsymbol{\gamma}_t \boldsymbol{\gamma}_s^t. \tag{50}$$

If we substitute (49) and (50) into (43), (44) and (45), we get the results as follows;

$$\begin{aligned} 2g_{ab} &= \operatorname{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_b\boldsymbol{\gamma}_b^t\Big) \\ &= \operatorname{tr}\Big\{\Big(\boldsymbol{\gamma}_b^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_a\Big)\Big\{\Big(\boldsymbol{\gamma}_a^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_b\Big)\Big\} \\ &= \operatorname{tr}\Big(\lambda_a^{-1}\delta(a=b)\lambda_b^{-1}\delta(a=b)\Big) \\ &= \lambda_a^{-2}\delta(a=b), \end{aligned}$$

$$
\begin{aligned}
2g_{a(s,t)} &= \mathrm{tr}\Big\{ \boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t\boldsymbol{\Sigma}^{-1}\Big(\lambda_t\boldsymbol{\gamma}_t\boldsymbol{\gamma}_s^t - \lambda_s\boldsymbol{\gamma}_s\boldsymbol{\gamma}_t^t + \lambda_t\boldsymbol{\gamma}_s\boldsymbol{\gamma}_t^t - \lambda_s\boldsymbol{\gamma}_t\boldsymbol{\gamma}_s^t\Big)\Big\} \\
&= \lambda_t\lambda_a^{-2}\delta(a=s=t) - \lambda_s\lambda_a^{-2}\delta(a=s=t) \\
&\qquad +\lambda_t\lambda_a^{-2}\delta(a=s=t) - \lambda_s\lambda_a^{-2}\delta(a=s=t) \\
&= 0, \\
2g_{(s,t)(u,v)} &= \mathrm{tr}\Big\{ \boldsymbol{\Sigma}^{-1}\Big(\lambda_t\boldsymbol{\gamma}_t\boldsymbol{\gamma}_s^t - \lambda_s\boldsymbol{\gamma}_s\boldsymbol{\gamma}_t^t + \lambda_t\boldsymbol{\gamma}_s\boldsymbol{\gamma}_t^t - \lambda_s\boldsymbol{\gamma}_t\boldsymbol{\gamma}_s^t\Big) \\
&\qquad \times \boldsymbol{\Sigma}^{-1}\Big(\lambda_v\boldsymbol{\gamma}_v\boldsymbol{\gamma}_u^t - \lambda_u\boldsymbol{\gamma}_u\boldsymbol{\gamma}_v^t + \lambda_v\boldsymbol{\gamma}_u\boldsymbol{\gamma}_v^t - \lambda_u\boldsymbol{\gamma}_v\boldsymbol{\gamma}_u^t\Big)\Big\} \\
&= \lambda_t\lambda_v\lambda_t^{-1}\delta(u=t)\lambda_s^{-1}\delta(s=v) - \lambda_t\lambda_u\lambda_t^{-1}\delta(v=t)\lambda_s^{-1}\delta(s=u) \\
&\qquad + \lambda_t\lambda_v\lambda_t^{-1}\delta(v=t)\lambda_s^{-1}\delta(s=u) - \lambda_t\lambda_u\lambda_t^{-1}\delta(u=t)\lambda_s^{-1}\delta(s=v) \\
&\qquad - \lambda_s\lambda_v\lambda_s^{-1}\delta(u=s)\lambda_t^{-1}\delta(t=v) + \lambda_s\lambda_u\lambda_s^{-1}\delta(s=v)\lambda_t^{-1}\delta(t=u) \\
&\qquad - \lambda_s\lambda_v\lambda_s^{-1}\delta(s=v)\lambda_t^{-1}\delta(t=u) + \lambda_s\lambda_u\lambda_s^{-1}\delta(u=s)\lambda_t^{-1}\delta(t=v) \\
&\qquad + \lambda_t\lambda_v\lambda_s^{-1}\delta(u=s)\lambda_t^{-1}\delta(t=v) - \lambda_t\lambda_u\lambda_s^{-1}\delta(v=s)\lambda_t^{-1}\delta(t=u) \\
&\qquad + \lambda_t\lambda_v\lambda_s^{-1}\delta(v=s)\lambda_t^{-1}\delta(t=u) - \lambda_t\lambda_u\lambda_s^{-1}\delta(u=s)\lambda_t^{-1}\delta(t=v) \\
&\qquad - \lambda_s\lambda_v\lambda_t^{-1}\delta(u=t)\lambda_s^{-1}\delta(s=v) + \lambda_s\lambda_u\lambda_t^{-1}\delta(v=t)\lambda_s^{-1}\delta(u=s) \\
&\qquad - \lambda_s\lambda_v\lambda_t^{-1}\delta(t=v)\lambda_s^{-1}\delta(u=s) + \lambda_s\lambda_u\lambda_t^{-1}\delta(u=t)\lambda_s^{-1}\delta(s=v) \\
&= (-1+\lambda_t\lambda_s^{-1} - 1 + \lambda_s\lambda_t^{-1} + \lambda_t\lambda_s^{-1} - 1 + \lambda_s\lambda_t^{-1} - 1)\delta(s=u,t=v) \\
&= 2(\lambda_s^{-1}(\lambda_t-\lambda_s) + \lambda_t^{-1}(\lambda_s-\lambda_t))\delta(s=u,t=v) \\
&= 2(\lambda_t-\lambda_s)(\lambda_s^{-1}-\lambda_t^{-1})\delta(s=u,t=v) \\
&= 2(\lambda_t-\lambda_s)^2(\lambda_s\lambda_t)^{-1}\delta(s=u,t=v).
\end{aligned}
$$

7.2 Proof of Proposition 2

Note that $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}\mathbf{\Lambda}^{-1}\boldsymbol{\Gamma}^t$, hence

$$
\theta^{ij} = \begin{cases} -\sum_k \gamma_{ik}\gamma_{jk}\lambda_k^{-1} & \text{if } i<j, \\ -2^{-1}\sum_k \gamma_{ik}^2\lambda_k^{-1} & \text{if } i=j. \end{cases}
$$

This means $\mathcal{M}$ is an affine subspace of $\mathcal{S}$ w.r.t. an Θ, which is an affine coordinate system of $\mathcal{S}$ with e-connection. Consequently $\mathcal{M}$ is e-flat, i.e. $\overset{e}{H}_{ab(s,t)}=0$. $\overset{m}{H}_{ab(s,t)}=0$ is similarly proved. See Theorem 1.1 in Amari and Nagaoka [3].

Now we consider $\overset{m}{H}_{(s,t)(u,v)a}$. Using (4.14) in Amari [2], it is calculated as

$$
\begin{aligned}
\overset{m}{H}_{(s,t)(u,v)a} &= \sum_{i\le j} \frac{\partial^2\sigma_{ij}}{\partial u_{st}\partial u_{uv}}\bigg|_{\boldsymbol{u}=\mathbf{0}} \frac{\partial\theta^{ij}}{\partial\lambda_a}\bigg|_{\boldsymbol{u}=\mathbf{0}} \\
&= -2^{-1}\sum_{1\le i,j\le p} \frac{\partial^2\sigma_{ij}}{\partial u_{st}\partial u_{uv}}\bigg|_{\boldsymbol{u}=\mathbf{0}} \frac{\partial\sigma^{ij}}{\partial\lambda_a}\bigg|_{\boldsymbol{u}=\mathbf{0}} \\
&= -2^{-1}\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}),
\end{aligned}
$$

where $p \times p$ matrices $\boldsymbol{A}$, $\boldsymbol{B}$ are given by

$$(\boldsymbol{A})_{ij} \triangleq \left. \frac{\partial^2 \sigma_{ij}}{\partial u_{st} \partial u_{uv}} \right|_{\boldsymbol{u}=\boldsymbol{0}}, \ (\boldsymbol{B})_{ij} \triangleq \left. \frac{\partial \sigma^{ij}}{\partial \lambda_a} \right|_{\boldsymbol{u}=\boldsymbol{0}}, \quad 1 \le i, j \le p.$$

In order to calculate $\boldsymbol{A}$, we only have to consider $\boldsymbol{\Sigma}$ up to the terms powered by two w.r.t. $\boldsymbol{u}$;

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\Big(\boldsymbol{I}_p + \boldsymbol{U} + 2^{-1}\boldsymbol{U}^2\Big)\boldsymbol{\Lambda}\Big(\boldsymbol{I}_p + \boldsymbol{U} + 2^{-1}\boldsymbol{U}^2\Big)^t \boldsymbol{\Gamma}^t + O(||\boldsymbol{u}||^3) \\ &= \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t + \boldsymbol{\Gamma}(\boldsymbol{U}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{U}^t)\boldsymbol{\Gamma}^t + 2^{-1}\boldsymbol{\Gamma}(\boldsymbol{U}^2\boldsymbol{\Lambda} + \boldsymbol{\Lambda}(\boldsymbol{U}^2)^t)\boldsymbol{\Gamma}^t + \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^t\boldsymbol{\Gamma}^t \\ &\qquad + O(||\boldsymbol{u}||^3). \end{aligned}$$

Therefore σ_{ij} is expressed as

$$\sigma_{ij} = 2^{-1} \sum_{k,l} \gamma_{ik}\gamma_{jl}\Big((\boldsymbol{U}^2\boldsymbol{\Lambda} + \boldsymbol{\Lambda}(\boldsymbol{U}^2)^t)_{kl} + 2(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^t)_{kl}\Big) + \boldsymbol{R}_{ij} + O(||\boldsymbol{u}||^3), \quad (51)$$

where $\boldsymbol{R} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^t + \boldsymbol{\Gamma}(\boldsymbol{U}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{U}^t)\boldsymbol{\Gamma}^t$. Since

$$\begin{aligned} (\boldsymbol{U}^2\boldsymbol{\Lambda} + \boldsymbol{\Lambda}(\boldsymbol{U}^2)^t)_{kl} &= (\boldsymbol{U}^2\boldsymbol{\Lambda})_{kl} + (\boldsymbol{U}^2\boldsymbol{\Lambda})_{lk} \\ &= \sum_b u_{kb}u_{bl}\lambda_l + \sum_b u_{lb}u_{bk}\lambda_k, \\ 2(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^t)_{kl} &= 2\sum_b u_{kb}u_{lb}\lambda_b, \end{aligned}$$

(51) truns out to be

$$\sigma_{ij} = 2^{-1} \sum_{k,l,b} \gamma_{ik}\gamma_{jl}(u_{kb}u_{bl}\lambda_l + u_{lb}u_{bk}\lambda_k + 2u_{kb}u_{lb}\lambda_b) + \boldsymbol{R}_{ij} + O(||\boldsymbol{u}||^3). \quad (52)$$

From this we have

$$\left. \frac{\partial^2 \sigma_{ij}}{\partial u_{st} \partial u_{uv}} \right|_{\boldsymbol{u}=\boldsymbol{0}} \times 2 = \sum_{k,l,b}(a_{ij}^{(1)} + a_{ji}^{(1)} + a_{ij}^{(2)} + a_{ji}^{(2)} + a_{ij}^{(3)} + a_{ij}^{(4)}), \quad (53)$$

where

$$\begin{aligned} a_{ij}^{(1)} = \ & \gamma_{is}\gamma_{jv}\lambda_v\delta\{(k,b) = (s,t), (b,l) = (u,v), (s,t) \ne (u,v)\} \\ & - \gamma_{it}\gamma_{jv}\lambda_v\delta\{(k,b) = (t,s), (b,l) = (u,v), (s,t) \ne (u,v)\} \\ & - \gamma_{is}\gamma_{ju}\lambda_u\delta\{(k,b) = (s,t), (b,l) = (v,u), (s,t) \ne (u,v)\} \\ & + \gamma_{it}\gamma_{ju}\lambda_u\delta\{(k,b) = (t,s), (b,l) = (v,u), (s,t) \ne (u,v)\} \\ & + \gamma_{iu}\gamma_{jt}\lambda_t\delta\{(k,b) = (u,v), (b,l) = (s,t), (s,t) \ne (u,v)\} \\ & - \gamma_{iv}\gamma_{jt}\lambda_t\delta\{(k,b) = (v,u), (b,l) = (s,t), (s,t) \ne (u,v)\} \\ & - \gamma_{iu}\gamma_{js}\lambda_s\delta\{(k,b) = (u,v), (b,l) = (t,s), (s,t) \ne (u,v)\} \\ & + \gamma_{iv}\gamma_{js}\lambda_s\delta\{(k,b) = (v,u), (b,l) = (t,s), (s,t) \ne (u,v)\}, \end{aligned}$$

$$
\begin{aligned}
a_{ij}^{(2)} &= 2\gamma_{is}\gamma_{jt}\lambda_t\delta\{(k,b)=(s,t),(b,l)=(s,t),(s,t)=(u,v)\} \\
&\quad + 2\gamma_{it}\gamma_{js}\lambda_s\delta\{(k,b)=(t,s),(b,l)=(t,s),(s,t)=(u,v)\} \\
&\quad - 2\gamma_{is}\gamma_{js}\lambda_s\delta\{(k,b)=(s,t),(b,l)=(t,s),(s,t)=(u,v)\} \\
&\quad - 2\gamma_{it}\gamma_{jt}\lambda_t\delta\{(k,b)=(t,s),(b,l)=(s,t),(s,t)=(u,v)\} \\
&= -2\gamma_{is}\gamma_{js}\lambda_s\delta\{(k,b)=(s,t),(b,l)=(t,s),(s,t)=(u,v)\} \\
&\quad - 2\gamma_{it}\gamma_{jt}\lambda_t\delta\{(k,b)=(t,s),(b,l)=(s,t),(s,t)=(u,v)\},
\end{aligned}
$$

$$
\begin{aligned}
a_{ij}^{(3)} &= 2\gamma_{is}\gamma_{ju}\lambda_t\delta\{(k,b)=(s,t),(l,b)=(u,v),(s,t)\neq(u,v)\} \\
&\quad - 2\gamma_{it}\gamma_{ju}\lambda_s\delta\{(k,b)=(t,s),(l,b)=(u,v),(s,t)\neq(u,v)\} \\
&\quad - 2\gamma_{is}\gamma_{jv}\lambda_t\delta\{(k,b)=(s,t),(l,b)=(v,u),(s,t)\neq(u,v)\} \\
&\quad + 2\gamma_{it}\gamma_{jv}\lambda_s\delta\{(k,b)=(t,s),(l,b)=(v,u),(s,t)\neq(u,v)\} \\
&\quad + 2\gamma_{iu}\gamma_{js}\lambda_t\delta\{(k,b)=(u,v),(l,b)=(s,t),(s,t)\neq(u,v)\} \\
&\quad - 2\gamma_{iv}\gamma_{js}\lambda_t\delta\{(k,b)=(v,u),(l,b)=(s,t),(s,t)\neq(u,v)\} \\
&\quad - 2\gamma_{iu}\gamma_{jt}\lambda_s\delta\{(k,b)=(u,v),(l,b)=(t,s),(s,t)\neq(u,v)\} \\
&\quad + 2\gamma_{iv}\gamma_{jt}\lambda_s\delta\{(k,b)=(v,u),(l,b)=(t,s),(s,t)\neq(u,v)\},
\end{aligned}
$$

$$
\begin{aligned}
a_{ij}^{(4)} &= 4\gamma_{is}\gamma_{js}\lambda_t\delta\{(k,b)=(l,b)=(s,t)=(u,v)\} \\
&\quad + 4\gamma_{it}\gamma_{jt}\lambda_s\delta\{(k,b)=(l,b)=(t,s)=(v,u)\} \\
&\quad - 4\gamma_{is}\gamma_{jt}\lambda_t\delta\{(k,b)=(s,t),(l,b)=(t,s),(s,t)=(u,v)\} \\
&\quad - 4\gamma_{it}\gamma_{js}\lambda_s\delta\{(k,b)=(t,s),(l,b)=(s,t),(s,t)=(u,v)\} \\
&= 4\gamma_{is}\gamma_{js}\lambda_t\delta\{(k,b)=(l,b)=(s,t)=(u,v)\} \\
&\quad + 4\gamma_{it}\gamma_{jt}\lambda_s\delta\{(k,b)=(l,b)=(t,s)=(v,u)\}.
\end{aligned}
$$

Furthermore we have

$$
2\boldsymbol{A} = \boldsymbol{A}^{(1)} + (\boldsymbol{A}^{(1)})^t + \boldsymbol{A}^{(2)} + (\boldsymbol{A}^{(2)})^t + \boldsymbol{A}^{(3)} + \boldsymbol{A}^{(4)}, \tag{54}
$$

where

$$
\begin{aligned}
\boldsymbol{A}^{(1)} &= \boldsymbol{\gamma}_s\boldsymbol{\gamma}_v^t\lambda_v\delta(t=u) + \boldsymbol{\gamma}_t\boldsymbol{\gamma}_u^t\lambda_u\delta(s=v) \\
&\quad + \boldsymbol{\gamma}_u\boldsymbol{\gamma}_t^t\lambda_t\delta(s=v) + \boldsymbol{\gamma}_v\boldsymbol{\gamma}_s^t\lambda_s\delta(u=t) \\
&\quad - \boldsymbol{\gamma}_t\boldsymbol{\gamma}_v^t\lambda_v\delta(s=u,t\neq v) - \boldsymbol{\gamma}_s\boldsymbol{\gamma}_u^t\lambda_u\delta(t=v,s\neq u) \\
&\quad - \boldsymbol{\gamma}_v\boldsymbol{\gamma}_t^t\lambda_t\delta(u=s,t\neq v) - \boldsymbol{\gamma}_u\boldsymbol{\gamma}_s^t\lambda_s\delta(t=v,s\neq u), \\
\boldsymbol{A}^{(2)} &= -2(\boldsymbol{\gamma}_s\boldsymbol{\gamma}_s^t\lambda_s\delta(s=u,t=v) + \boldsymbol{\gamma}_t\boldsymbol{\gamma}_t^t\lambda_t\delta(s=u,t=v)), \\
\boldsymbol{A}^{(3)} &= 2\Big(\boldsymbol{\gamma}_s\boldsymbol{\gamma}_u^t\lambda_t\delta(t=v,s\neq u) + \boldsymbol{\gamma}_t\boldsymbol{\gamma}_v^t\lambda_s\delta(s=u,t\neq v) \\
&\qquad + \boldsymbol{\gamma}_u\boldsymbol{\gamma}_s^t\lambda_t\delta(v=t,s\neq u) + \boldsymbol{\gamma}_v\boldsymbol{\gamma}_t^t\lambda_s\delta(u=s,t\neq v)\Big) \\
&\qquad - 2\Big(\boldsymbol{\gamma}_t\boldsymbol{\gamma}_u^t\lambda_s\delta(s=v) + \boldsymbol{\gamma}_s\boldsymbol{\gamma}_v^t\lambda_t\delta(t=u) + \boldsymbol{\gamma}_v\boldsymbol{\gamma}_s^t\lambda_t\delta(u=t) + \boldsymbol{\gamma}_u\boldsymbol{\gamma}_t^t\lambda_s\delta(s=v)\Big), \\
\boldsymbol{A}^{(4)} &= 4\Big(\boldsymbol{\gamma}_s\boldsymbol{\gamma}_s^t\lambda_t\delta(s=u,t=v) + \boldsymbol{\gamma}_t\boldsymbol{\gamma}_t^t\lambda_s\delta(s=u,t=v)\Big).
\end{aligned}
$$

Since

$$\left.\frac{\partial\sigma^{ij}}{\partial\lambda_a}\right|_{\boldsymbol{u}=\boldsymbol{0}} = -\lambda_a^{-2}\gamma_{ia}\gamma_{ja},$$

we have

$$\boldsymbol{B} = -\lambda_a^{-2}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t. \tag{55}$$

From (54) and (55), we have

$$\begin{aligned}\overset{m}{H}_{(s,t)(u,v)a} &= -4^{-1}\mathrm{tr}(2\boldsymbol{AB})\\ &= -4^{-1}\mathrm{tr}\{(\boldsymbol{A}^{(1)} + (\boldsymbol{A}^{(1)})^t + \boldsymbol{A}^{(2)} + (\boldsymbol{A}^{(2)})^t + \boldsymbol{A}^{(3)} + \boldsymbol{A}^{(4)})\boldsymbol{B}\}\\ &= 4^{-1}\lambda_a^{-2}\mathrm{tr}\{(\boldsymbol{A}^{(1)} + (\boldsymbol{A}^{(1)})^t + \boldsymbol{A}^{(2)} + (\boldsymbol{A}^{(2)})^t + \boldsymbol{A}^{(3)} + \boldsymbol{A}^{(4)})\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t\}.\end{aligned}$$

The following equalities hold;

$$\begin{aligned}\mathrm{tr}(\boldsymbol{A}^{(1)}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t) &= \lambda_a\delta(s=v=a, t=u) + \lambda_a\delta(t=u=a, s=v)\\ &\quad + \lambda_a\delta(t=u=a, s=v) + \lambda_a\delta(s=v=a, t=u)\\ &\quad - \lambda_a\delta(t=v=a, s=u, t\neq v) - \lambda_a\delta(s=u=a, t=v, s\neq u)\\ &\quad - \lambda_a\delta(t=v=a, s=u, t\neq v) - \lambda_a\delta(s=u=a, t=v, s\neq u)\\ &= 0.\\ \mathrm{tr}((\boldsymbol{A}^{(1)})^t\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t) &= 0.\\ \mathrm{tr}(\boldsymbol{A}^{(2)}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t) &= -2(\lambda_a\delta(s=u=a, t=v) + \lambda_a\delta(s=u, t=v=a)).\\ \mathrm{tr}((\boldsymbol{A}^{(2)})^t\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t) &= -2(\lambda_a\delta(s=u=a, t=v) + \lambda_a\delta(s=u, t=v=a)).\\ \mathrm{tr}(\boldsymbol{A}^{(3)}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t) &= 2\{\lambda_t\delta(s=u=a)\delta(t=v, s\neq u) + \lambda_s\delta(t=v=a)\delta(s=u, t\neq v)\\ &\quad + \lambda_t\delta(s=u=a)\delta(t=v, s\neq u) + \lambda_s\delta(t=v=a)\delta(s=u, t\neq v)\}\\ &\quad - 2\{\lambda_s\delta(t=u=a)\delta(s=v) + \lambda_t\delta(s=v=a)\delta(t=u)\\ &\quad + \lambda_t\delta(s=v=a)\delta(t=u) + \lambda_s\delta(u=t=a)\delta(s=v)\}\\ &= 0.\\ \mathrm{tr}(\boldsymbol{A}^{(4)}\boldsymbol{\gamma}_a\boldsymbol{\gamma}_a^t) &= 4\lambda_t\delta(s=u=a, t=v) + 4\lambda_s\delta(t=v=a, s=u).\end{aligned}$$

Consequently

$$\begin{aligned}\overset{m}{H}_{(s,t)(u,v)a} &= -\lambda_a^{-1}\delta(s=u=a, t=v) - \lambda_a^{-1}\delta(s=u, t=v=a)\\ &\quad + \lambda_a^{-2}\lambda_t\delta(s=u=a, t=v) + \lambda_a^{-2}\lambda_s\delta(t=v=a, s=u)\\ &= \begin{cases}\lambda_a^{-2}(\lambda_t-\lambda_a), & \text{if } s=u=a,\ t=v,\\ \lambda_a^{-2}(\lambda_s-\lambda_a), & \text{if } s=u,\ t=v=a,\\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

7.3 Proof of Corollary 1

As we will see in the next subsection,

$$\sum_{s<t,u<v,o<p,q<r} \overset{m}{H}_{(s,t)(u,v)a} \overset{m}{H}_{(o,p)(q,r)b} \; g^{(s,t)(o,p)} \; g^{(u,v)(q,r)}$$

$$= \begin{cases} \dfrac{1}{\lambda_a^2} \displaystyle\sum_{t \neq a} \dfrac{\lambda_t^2}{(\lambda_t - \lambda_a)^2}, \text{ if } a = b, \\ -\dfrac{1}{(\lambda_a - \lambda_a)^2}, \text{ if } a \neq b. \end{cases}$$

Combine this with Proposition 1, we have

$$\begin{aligned} \gamma(\mathcal{A}) &= 2 \sum_a \sum_{t \neq a} \frac{\lambda_t^2}{(\lambda_t - \lambda_a)^2} \\ &= 2 \sum_{a<b} \frac{\lambda_a^2 + \lambda_b^2}{(\lambda_a - \lambda_b)^2}. \end{aligned}$$

7.4 Proof of Proposition 3

We calculate each term in (29). $g^{ab} = \delta(a = b) 2\lambda_a^2$ from Proposition 1. Because of (22) and (26),

$$(\Gamma^m_M)^{2ab} = (H^e_M)^{2ab} = 0.$$

For $\boldsymbol{l}^*$, $(H^m_{\hat{A}})^{2ab} = (H^m_A)^{2ab}$.

$$\begin{aligned} (H^m_A)^{2ab} &= \sum_{s<t,u<v,o<p,q<r} \overset{m}{H}{}^a_{(s,t)(u,v)} \overset{m}{H}{}^b_{(o,p)(q,r)} \; g^{(s,t)(o,p)} g^{(u,v)(q,r)} \\ &= \sum_{t>a,p>b} \overset{m}{H}{}^a_{(a,t)(a,t)} \overset{m}{H}{}^b_{(b,p)(b,p)} \; (g^{(a,t)(b,p)})^2 \\ &+ \sum_{t>a,p<b} \overset{m}{H}{}^a_{(a,t)(a,t)} \overset{m}{H}{}^b_{(p,b)(p,b)} \; (g^{(a,t)(p,b)})^2 \\ &+ \sum_{t<a,p>b} \overset{m}{H}{}^a_{(t,a)(t,a)} \overset{m}{H}{}^b_{(b,p)(b,p)} \; (g^{(t,a)(b,p)})^2 \\ &+ \sum_{t<a,p<b} \overset{m}{H}{}^a_{(t,a)(t,a)} \overset{m}{H}{}^b_{(p,b)(p,b)} \; (g^{(t,a)(p,b)})^2 \end{aligned} \tag{56}$$

If $a = b$, then the r.h.s of (56) equals

$$\begin{aligned} &\sum_{t>a} (\overset{m}{H}{}^a_{(a,t)(a,t)})^2 (g^{(a,t)(a,t)})^2 + \sum_{t<a} (\overset{m}{H}{}^a_{(t,a)(t,a)})^2 (g^{(t,a)(t,a)})^2 \\ &= \sum_{t>a} (2(\lambda_t - \lambda_a))^2 \Big(\frac{\lambda_a \lambda_t}{(\lambda_a - \lambda_t)^2} \Big)^2 + \sum_{t<a} (2(\lambda_t - \lambda_a))^2 \Big(\frac{\lambda_a \lambda_t}{(\lambda_a - \lambda_t)^2} \Big)^2 \\ &= 4 \sum_{t \neq a} \frac{\lambda_a^2 \lambda_t^2}{(\lambda_a - \lambda_t)^2}. \end{aligned}$$

If $a \neq b$, then the r.h.s of (56) equals

$$\begin{aligned}
&\overset{m}{H}{}^{a}_{(a,b)(a,b)}\ \overset{m}{H}{}^{b}_{(a,b)(a,b)}\ (g^{(a,b)(a,b)})^2 \\
&= 4(\lambda_b - \lambda_a)(\lambda_a - \lambda_b)\Big(\frac{\lambda_a\lambda_b}{(\lambda_a - \lambda_b)^2}\Big)^2 \\
&= -\frac{4\lambda_a^2\lambda_b^2}{(\lambda_a - \lambda_b)^2}.
\end{aligned}$$

7.5 Proof of Proposition 4

The term of the order n in (34) vanishes since $g_{a(s,t)}$ equals zero for $1 \leq a \leq p,\ 1 \leq s < t \leq p$. We consider the term of order $O(1)$. Since $\overset{e}{H}_{ac(s,t)}$ also vanishes for $1 \leq a, c \leq p,\ 1 \leq s < t \leq p$, we only have to consider the term

$$(1/2) \sum_{s<t,u<v,o<p,q<r} \overset{m}{H}_{(s,t)(u,v)a}\ \overset{m}{H}_{(o,p)(q,r)b}\ \ g^{(s,t)(o,p)}\ g^{(u,v)(q,r)}.$$

Because of (18), the above term equals

$$\begin{aligned}
&2^{-1}\sum_{t>a,p>b} \overset{m}{H}_{(a,t)(a,t)a}\overset{m}{H}_{(b,p)(b,p)b}\ (g^{(a,t)(b,p)})^2 \\
&+ 2^{-1}\sum_{t>a,p<b} \overset{m}{H}_{(a,t)(a,t)a}\overset{m}{H}_{(p,b)(p,b)b}\ (g^{(a,t)(p,b)})^2 \\
&+ 2^{-1}\sum_{t<a,p>b} \overset{m}{H}_{(t,a)(t,a)a}\overset{m}{H}_{(b,p)(b,p)b}\ (g^{(t,a)(b,p)})^2 \\
&+ 2^{-1}\sum_{t<a,p<b} \overset{m}{H}_{(t,a)(t,a)a}\overset{m}{H}_{(p,b)(p,b)b}\ (g^{(t,a)(p,b)})^2 \qquad (57)
\end{aligned}$$

If $a = b$, then (57) equals

$$\begin{aligned}
&2^{-1}\sum_{t>a}(\overset{m}{H}_{(a,t)(a,t)a})^2(g^{(a,t)(a,t)})^2 + 2^{-1}\sum_{t<a}(\overset{m}{H}_{(t,a)(t,a)a})^2(g^{(t,a)(t,a)})^2 \\
&= 2^{-1}\Big\{\sum_{t>a}(\lambda_a^{-2}(\lambda_t - \lambda_a))^2\Big(\frac{\lambda_a\lambda_t}{(\lambda_a - \lambda_t)^2}\Big)^2 + \sum_{t<a}(\lambda_a^{-2}(\lambda_t - \lambda_a))^2\Big(\frac{\lambda_a\lambda_t}{(\lambda_a - \lambda_t)^2}\Big)^2\Big\} \\
&= 2^{-1}\sum_{t\neq a}\frac{\lambda_t^2}{\lambda_a^2(\lambda_a - \lambda_t)^2}.
\end{aligned}$$

If $a < b$, then (57) equals

$$\begin{aligned}
&2^{-1}\ \overset{m}{H}_{(a,b)(a,b)a}\overset{m}{H}_{(a,b)(a,b)b}\ (g^{(a,b)(a,b)})^2 \\
&= 2^{-1}\lambda_a^{-2}(\lambda_b - \lambda_a)\lambda_b^{-2}(\lambda_a - \lambda_b)\Big(\frac{\lambda_a\lambda_b}{(\lambda_a - \lambda_b)^2}\Big)^2 \\
&= -\frac{1}{2(\lambda_a - \lambda_b)^2}.
\end{aligned}$$

Acknowledgment

The author really appreciates Dr. M. Kumon kindly answering his question on a basic fact of the information loss. He also shows deep gratitude to anonymous referees for their valuable comments which improved the quality of the paper.

References

[1] S. Amari. Differential geometry of curved exponential families–curvature and information loss– *The Annals of Statistics*, 10:357-385, 1982.

[2] S. Amari. *Differential-Geometrical Methods in Statistics.* Lecture Notes in Statistics 28. Springer-Verlag, 1985.

[3] S. Amari and H. Nagaoka. *Methods of Information Geometry.* Translations of Mathematical Monographs 191. American Mathematical Society, 2000.

[4] S. Amari and M. Kumon. Differential geometry of Edgeworth expansions in curved exponential family. *Annals of the Institute of Statistical Mathematics*, 35: 1-24, 1983.

[5] G. A. Anderson. An asymptotic expansion for the distribution of the latent roots of the estimated covariance matrix. *The Annals of Mathematical Statistics*, 36: 1153-1173, 1965.

[6] W. M. Boothby *An introduction to differentiable manifolds and Riemannian geometry, revised 2nd. edition.* Academic Press, 2002.

[7] M. Calvo and J. M. Oller A distance between multivariate normal distributions based on an embedding into Siegel group. *Journal of Multivariate Analysis*, 35: 223-242, 1990.

[8] D. K. Dey. Simultaneous estimation of eigenvalues. *Annals of the Institute of Statistical Mathematics*, 40: 137-147, 1988.

[9] D. K. Dey and C. Srinivasan. Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, 13: 1581-1591, 1985.

[10] B. Efron. Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion). *The Annals of Statistics*, 3: 1189-1242, 1975.

[11] S. Eguchi. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Mathematical Journal*, 15: 341-391, 1985.

[12] P. T. Fletcher and S. Joshi Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87: 250-262, 2007.

[13] L. R. Haff. The variational form of certain Bayes estimators. *The Annals of Statistics*, 19:1163-1190, 1991.

[14] D. L. Hydorn and R. J. Muirhead. Polynomial estimation of eigenvalues. *Communications in Statistics. Theory and Methods*, 28: 581-596, 1999.

[15] C. Jin. A note on simultaneous estimation of eigenvalues of a multivariate normal covariance matrix. *Statistics and Probability Letters*, 16: 197-203, 1993.

[16] M. Kumon and S. Amari. Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference. *Proceedings of the Royal Society of London*, A387: 429-458, 1983.

[17] D. N. Lawley. Test of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43: 128-136, 1956.

[18] C. Lenglet, M. Rousson, R. Deriche and O. Faugeras. Statistics on the manifold of multivariate normal distributions: theory and application to diffusion tensor MRI processing. *Journal of Mathematical Imaging and Vision*, 25: 423-444, 2006

[19] M. Lovrić, M. Min-Oo and E. A. Ruh. Multivariate normal distributions parametrized as a Riemannian symmetric space. *Journal of Multivariate Analysis*, 74: 36-48, 2000.

[20] Maher Moakher and Mourad Zéraï. The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision*, 40; 171-187, 2011.

[21] R. J. Muirhead. *Aspects of Multivariate Statistical Theory.* Wiley, 1982.

[22] M. K. Murray and J. W. Rice. *Differential Geometry and Statistics.* Chapman, 1993.

[23] A. Ohara, N. Suda and S. Amari. Dualistic differential geometry of positive definite matrices and its applications to related problems. *Linear Algebra and its Applications*, 247: 31-53, 1996.

[24] Y. Sheena and A. Takemura. Admissible estimator of the eigenvalues of the variance-covariance matrix for multivariate normal distributions. *Journal of Multivariate Analysis*, 102: 801-815, 2011.

[25] L. T. Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11: 211-233, 1984.

[26] S. T. Smith Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Transactions on Signal Processing*, 53: 1610-1630, 2005

[27] A. Takemura. An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba Journal of Mathematics*, 8: 367-376, 1984.

[28] R. Yang and J. O. Berger. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 22: 1195-1221, 1994.

[29] S. Yoshizawa and K. Tanabe. Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT journal of mathematics*, 35: 113-137, 1999.

[30] S. Zhang, H. Sun and C. Li. Information geometry of positive definite matrices. *Journal of Beijing Institute of Technology*, 18: 484-487, 2009.