Linear Algebra and its Applications 548 (2018) 293–304

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Differentiating the pseudo determinant ☆

Andrew Holbrook

Department of Statistics, University of California, Irvine, United States

ARTICLE INFO

Article history:
Received 13 February 2018
Accepted 8 March 2018
Available online 13 March 2018
Submitted by R. Brualdi

MSC:
primary 15A15
secondary 62H12

Keywords:
Pseudo determinant
Pseudo inverse
Maximum likelihood
Degenerate Gaussian
Singular covariance

ABSTRACT

A class of derivatives is defined for the pseudo determinant $\mathrm{Det}(A)$ of a Hermitian matrix A. This class is shown to be non-empty and to have a unique, canonical member $\nabla\mathbf{Det}(A) = \mathrm{Det}(A)A^{+}$, where A^{+} is the Moore–Penrose pseudo inverse. The classic identity for the gradient of the determinant is thus reproduced. Examples are provided, including the maximum likelihood problem for the rank-deficient covariance matrix of the degenerate multivariate Gaussian distribution.

1. Introduction

We derive the class of derivatives of the pseudo determinant with respect to Hermitian matrices, placing an emphasis on understanding the forms taken by this class and their relationship to established results in linear algebra. In particular, care must be taken to address the discontinuous nature of the pseudo derivative. The contributions in

☆ This work was supported by the UC Irvine Graduate Dean's Dissertation Fellowship. I thank Professor Oliver Knill for his encouragement and helpful discussion, and I am grateful to an anonymous reviewer for further helpful recommendations.

E-mail address: aholbroo@uci.edu.

https://doi.org/10.1016/j.laa.2018.03.018
0024-3795/

this paper are primarily of a linear algebraic nature but are well motivated in fields of application.

The pseudo determinant arises in graph theory within Kirchoff's matrix tree theorem [1] and in statistics, in the definition of the degenerate Gaussian distribution. The degenerate Gaussian has been useful in image segmentation [2], communications [3], and as the asymptotic distribution for multinomial samples [4]. Despite these appearances, knowledge of how to differentiate the distribution's density function is conspicuously absent from the literature, and—since differentiation is often essential for maximization—the lack of this knowledge is a plausible barrier to the distribution's wider use.

Specifically, to obtain the maximum likelihood (ML) estimator for the singular covariance matrix of the degenerate Gaussian, one must be able to calculate the derivative of the log likelihood and hence the pseudo determinant of the covariance. Although [5] firmly establishes the subject of ML estimation for multivariate Gaussians, the authors never directly address singular covariance estimation. This problem is explored in Section 3. In Section 2, the pseudo determinant is introduced, and its derivative with respect to Hermitian matrices is derived.

2. The canonical derivative

We begin by introducing the pseudo determinant both as a product of eigenvalues and as a limiting form.

Definition 2.1. The pseudo determinant Det of a square matrix A is defined as the product of its non-zero eigenvalues. If a matrix has no non-zero eigenvalues, then we say $\mathrm{Det}(0) = 1$.

See [1] for an equivalent definition of the pseudo determinant in terms of the characteristic polynomial. In deriving its derivative, it will be useful to write the pseudo determinant as a limit.

Proposition 2.2. *If A is an $n \times n$ matrix of rank k, then* $\mathrm{Det}(A)$ *is the limit*

$$\mathrm{Det}(A) = \lim_{\delta \to 0} \frac{\det(A + \delta I)}{\delta^{n-k}} \tag{2.1}$$

for $\det(\cdot)$ *the regular determinant.*

Whereas this result is known [6], we were unable to find its proof, so it is given here in the spirit of completeness.

Proof. We use the identity

$$\det(X + ZYZ^*) = \det(Y^{-1} + Z^* X^{-1} Z)\, \det(Y)\, \det(X)\,. \tag{2.2}$$

Replacing X with $k\,I_n$ and letting $A = U\Lambda U^* = ZYZ^*$, we have

$$\begin{aligned}\lim_{\delta\to 0}\frac{\det(A+\delta I)}{\delta^{n-k}} &= \lim_{\delta\to 0}\frac{k^n}{k^{n-r}}\det(\Lambda^{-1}+\frac{1}{k}I_r)\,\det(\Lambda) \\ &= \mathrm{Det}(A)\lim_{k\to 0}k^r\det(\Lambda^{-1}+\frac{1}{k}I_r) \\ &= \mathrm{Det}(A)\lim_{k\to 0}\det(k\Lambda^{-1}+I_r) \\ &= \mathrm{Det}(A)\,. \quad \square\end{aligned} \tag{2.3}$$

Next, we define the Moore–Penrose pseudo inverse [7], an important object involved in the derivative of the pseudo determinant.

Definition 2.3. The pseudo inverse A^+ of a matrix A is also defined in terms of a limit:

$$A^+ = \lim_{\delta\to 0} A^*(AA^*+\delta I)^{-1} = \lim_{\delta\to 0}(A^*A+\delta I)^{-1}A^*\,. \tag{2.4}$$

A^+ exists in general and is unique. It may also be defined as the matrix satisfying all the following criteria:

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $(AA^+)^* = AA^+$
4. $(A^+A)^* = A^+A$

For Hermitian matrices, the pseudo inverse is obtained by inverting the matrix eigenvalues.

As is the case for the pseudo inverse [7], the pseudo determinant is discontinuous. For an example, consider the two matrices

$$A = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, \quad \text{and} \qquad B_j = \begin{pmatrix}0 & 0\\ 0 & j\end{pmatrix}. \tag{2.5}$$

Note that $\mathrm{Det}(A) = 1$ and $\mathrm{Det}(A+B_j) = j$, but that

$$\lim_{j\to 0}\mathrm{Det}(A+B_j) = 0 \neq 1 = \mathrm{Det}(\lim_{j\to 0}A+B_j)\,. \tag{2.6}$$

As one might gather from this example, the pseudo determinant is discontinuous between sets of matrices of differing ranks. This discontinuity will effect the way we define the derivative of the pseudo determinant. We now turn to deriving this derivative.

For matrix A in the space of $n\times n$ matrices $M_{n\times n}$, the matrix derivative of a function $h : M_{n\times n}\to\mathbb{R}$ is given by the matrix $\nabla h(A)$ satisfying

$$\nabla_B h(A) = \operatorname{tr}\left(B \nabla h(A)\right) = \lim_{\tau \to 0} \frac{h(A + \tau B) - h(A)}{\tau} \tag{2.7}$$

for any matrix $B \in M_{n\times n}$, where $\nabla_B h(A)$ is the directional derivative. We use the directional derivative to define the derivative of the pseudo determinant, but, on account of the discontinuity of the pseudo determinant, we must restrict the directions B in which the directional derivative is defined. For this reason, we may define the derivative at a point only in certain directions and must modify the common definition of the directional derivative.

Definition 2.4. (Definition 1) For a matrix A in the space of Hermitian $n \times n$, rank k matrices $M^k_{n\times n}$, the directional derivative $\nabla_B \operatorname{Det}(A)$ of the pseudo determinant $\operatorname{Det} : M_{n\times n} \to \mathbb{R}$ is defined in directions $B \in M^k_{n\times n}$ *that share the same kernel as* A, i.e. for which $Ker(A) = Ker(B)$. Then the derivative $\nabla \operatorname{Det}(A)$ is given by any matrix satisfying

$$\nabla_B \operatorname{Det}(A) = \operatorname{tr}\left(B \nabla \operatorname{Det}(A)\right) = \lim_{\tau \to 0} \frac{\operatorname{Det}(A + \tau B) - \operatorname{Det}(A)}{\tau}\,. \tag{2.8}$$

Note that, according to this definition, $\nabla \operatorname{Det}(A)$ is not unique, since it can take on different values along the kernel of B. This non-uniqueness can also be seen using the following class equations for the class of derivatives $\nabla \operatorname{Det}(A)$ of the pseudo determinant at a matrix A.

Definition 2.5. (Definition 2) A derivative of the pseudo determinant at a point $A \in M^k_{n\times n}$ is any non-zero matrix $\nabla \operatorname{Det}(A) \in M^k_{n\times n}$ satisfying

$$A \nabla \operatorname{Det}(A) = A\, A^+ \operatorname{Det}(A) \tag{2.9}$$

$$\nabla \operatorname{Det}(A)\, A = A^+ A \operatorname{Det}(A)\,. \tag{2.10}$$

We demonstrate that this is a natural definition using the facts that $A(A^2)^+ = A^+$ and $(A^2)^+ A = A^+$ for any Hermitian A and assuming one may interchange limits:

$$\begin{aligned} A^{1/2} \nabla \operatorname{Det}(A) &= A^{1/2} \nabla \lim_{\delta \to 0} \frac{\det(A + \delta I)}{\delta^{n-k}} \\ &= A^{1/2} \lim_{\delta \to 0} \frac{1}{\delta^{n-k}} \nabla \det(A + \delta I) \\ &= \operatorname{Det}(A) \lim_{\delta \to 0} A^{1/2} (A + \delta I)^{-1} \\ &= \operatorname{Det}(A)\, (A^{1/2})^+ \\ &= \operatorname{Det}(A)\, A^{1/2} A^+\,. \end{aligned} \tag{2.11}$$

Multiplying both sides by $A^{1/2}$ and rearranging gives the first class equation. The derivation of the second equation is symmetric. We illustrate the preceding definitions—and that they do not define unique derivatives—with a few examples.

Example 2.6. Consider the 2×2 matrix

$$A = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.12}$$

It is clear that $\mathrm{Det}(A) = a$ and A^+ is obtained by taking the reciprocal of the first element of A. The above result renders

$$A\nabla \,\mathrm{Det}(A) = a\, AA^+ = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} = a\, A^+A = A\nabla \,\mathrm{Det}(A)\,. \tag{2.13}$$

Note that multiple matrices solve this equation. Two examples are the identity and the matrix A/a.

Example 2.7. Now consider the 2×2 matrix pair

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad A^+ = \begin{pmatrix} .25 & .25 \\ .25 & .25 \end{pmatrix}. \tag{2.14}$$

One can check that $\mathrm{Det}(A) = 2$. Then it follows from the result that

$$A\nabla \,\mathrm{Det}(A) = 2\, AA^+ = 2\,\frac{1}{2}A = A = \cdots = \nabla \,\mathrm{Det}(A)A\,. \tag{2.15}$$

Again, multiple matrices satisfy Equation (2.15): take for example

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix}. \tag{2.16}$$

It turns out that the matrix A in the class equations of Definition (2.5) may be replaced by any Hermitian B such that $Ker(B) = Ker(A)$. This is easily shown using the fact that $BA^+A = B = BAA^+ = B = A^+AB = B = AA^+B$ for any such B.

Proposition 2.8. *The derivative of the pseudo determinant is any matrix* $\nabla \,\mathrm{Det}(A)$ *satisfying the equations*

$$B\,\nabla \,\mathrm{Det}(A) = B\, A^+ \,\mathrm{Det}(A) \tag{2.17}$$

$$\nabla \,\mathrm{Det}(A)\, B = A^+B\ \mathrm{Det}(A)\,, \tag{2.18}$$

for any matrix B *for which* $Ker(B) = Ker(A)$.

This result may be combined with the directional derivative based definition of $\nabla \,\mathrm{Det}(A)$.

Proposition 2.9. *The derivative of the pseudo determinant is any matrix* $\nabla \operatorname{Det}(A)$ *satisfying the equations*

$$\operatorname{tr}\left(B\nabla \operatorname{Det}(A)\right) = \operatorname{Det}(A)\operatorname{tr}(BA^{+}), \tag{2.19}$$

for any matrix B *for which* $Ker(B) = Ker(A)$.

In practice, one may obtain the canonical element $\nabla\mathbf{Det}(A)$ of class $\nabla \operatorname{Det}(A)$ directly from a corollary to the following Pythagorean theorem.

Theorem 2.10. *(Knill 2014 [1]) For Hermitian* A *of rank* k,

$$\operatorname{Det}^2(A) = \operatorname{Det}(A^2) = \sum_P \det^2(A_P) \tag{2.20}$$

where P *indexes all* $k \times k$ *minors of* A *satisfying* $\det(A_P) \neq 0$.

As a corollary, the canonical gradient $\nabla\mathbf{Det}$ is directly obtainable.

Corollary 2.11. *For Hermitian* A *of rank* k, *one has*

$$\nabla \operatorname{Det}(A) = \frac{1}{\operatorname{Det}(A)} \sum_P \det^2(A_P) A_P^{-1} = \frac{\sum_P \det^2(A_P) A_P^{-1}}{\sqrt{\sum_P \det^2(A_P)}} := \nabla\mathbf{Det}(A)\,. \tag{2.21}$$

This $\nabla\mathbf{Det}(A)$ satisfies the class equations as well as Equation (2.19). Before proving this claim, we illustrate by revisiting our examples.

Example 2.12. We again consider matrix

$$A = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.22}$$

This time we use Formula (2.21). Here, the rank k minors are simply the elements of A. Since only the first element is non-zero, we have

$$\nabla\mathbf{Det}(A) = \frac{\det^2(A_{11})\, A_{11}^{-1}}{\operatorname{Det}(A)} = \frac{a^2}{a} \begin{pmatrix} a^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.23}$$

This, of course, agrees with the original example.

Example 2.13. Again, consider the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{2.24}$$

The gradient of the pseudo determinant may be found using Formula (2.21):

$$\nabla \mathbf{Det}(A) = \frac{1}{2}\sum_{ij} \det^2(A_{ij}) A_{ij}^{-1} = \frac{1}{2} A\,. \tag{2.25}$$

The reader may check that

$$A\nabla \mathbf{Det}(A) = \frac{1}{2}A^2 = A = \cdots = \nabla \operatorname{Det}(\mathbf{A})A\,, \tag{2.26}$$

as expected from Equation (2.15).

The above examples suggest that $\nabla \mathbf{Det}(A)$ should satisfy the class equations in general. To show this, we first cite a result.

Theorem 2.14. *(Berg 1986 [8]) The pseudo inverse of a Hermitian, rank k matrix A takes the following form:*

$$A^+ = \frac{\sum_P \det^2(A_P) A_P^{-1}}{\operatorname{Det}^2(A)} = \frac{\sum_P \det^2(A_P) A_P^{-1}}{\sum_P \det^2(A_P)}\,. \tag{2.27}$$

Theorem 2.15.

$$\nabla \mathbf{Det}(A) = \operatorname{Det}(A) A^+ \tag{2.28}$$

Thus $\nabla \mathbf{Det}(A)$ *satisfies the class equations and belongs to the equivalence class* $\nabla \operatorname{Det}(A)$. *Moreover,* $\nabla \mathbf{Det}(A)$ *is the unique member of the equivalence class that has the same kernel as A. In this sense, it may be considered the canonical gradient of the pseudo determinant.*

Proof. That $\nabla \mathbf{Det}(A) = \operatorname{Det}(A) A^+$ is a simple result of Corollary 2.11 and Theorem 2.14. As a result, it immediately satisfies the two propositions as well.

We now consider the uniqueness claim. In general, $A : Ker(A)^\perp \to Im(A)$ is an isomorphism, and $A : Im(A) \to Ker(A)^\perp$ is its inverse. Since A is Hermitian, $Ker(A) \oplus Im(A) = \mathbb{C}^n$, and so $A : Im(A) \to Im(A)$, $A^+ : Im(A) \to Im(A)$ is the isomorphism pair. Clearly $Ker(A) = Ker(A^+)$, and so $Ker\big(\nabla \mathbf{Det}(A)\big) = Ker(A)$.

We proceed by contradiction. Suppose that there exists another matrix $B \neq \nabla \mathbf{Det}(A)$ satisfying $Ker(A) = Ker(B)$ and

$$\begin{aligned} AB &= A\, A^+ \operatorname{Det}(A) \\ BA &= A^+ A \operatorname{Det}(A)\,. \end{aligned} \tag{2.29}$$

Since $B \neq A$, there exists at least one element $y \in \mathbb{C}^n$ such that $By \neq \nabla \mathbf{Det}(A) y$. Since $\mathbb{C}^n = Im(A) \oplus Ker(A)$, we may consider y in each subspace separately. If $y \in Ker(A)$, then $By = 0 = \nabla \mathbf{Det}(A) y$. Therefore y must be an element of $Im(A)$. Then,

$$\begin{aligned}(B - \nabla\mathbf{Det}(A))y &= (B - \nabla\mathbf{Det}(A))(AA^+)y \\ &= (BA - \nabla\mathbf{Det}(A)A)A^+y \\ &= (A^+A\ \mathrm{Det}(A) - A^+A\ \mathrm{Det}(A))A^+y \\ &= 0\,.\end{aligned} \tag{2.30}$$

Then $By = \nabla\mathbf{Det}(A)y$, thus establishing a contradiction. □

We round out this section with a few examples demonstrating applications of Formula (2.28).

Example 2.16. Let A be the constant, $n \times n$ matrix satisfying $A_{ij} = 1, \forall i, j = 1, \ldots, n$. Then it is true that

$$\mathrm{Det}(A) = n\,, \quad \text{and} \quad A^+ = \frac{1}{n^2}A\,. \tag{2.31}$$

Hence

$$\nabla\mathbf{Det}(A) = \mathrm{Det}(A)A^+ = \frac{1}{n}A\,. \tag{2.32}$$

Example 2.17. Let $A = 0$ be the $n \times n$ zero matrix for arbitrary integer n. The reader can check that $A^+ = A = 0$ by observing the four criteria in the definition of the pseudo inverse. Recall also that $\mathrm{Det}(0) = 1$ for any square matrix with no non-zero eigenvalues. It follows that

$$\nabla\mathbf{Det}(A) = \mathrm{Det}(A)A^+ = A = 0\,. \tag{2.33}$$

This basic result is more appealing using the shorthand $\nabla\mathbf{Det}(0) = 0$.

Example 2.18. Consider the projection–dilation matrix

$$A = \begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix} \tag{2.34}$$

that maps a point $v \in \mathbb{R}^2$ onto the line through the origin containing the unit vector $u = (a, b)^T / \sqrt{(a^2 + b^2)}$ while scaling by $a^2 + b^2$. The reader may check that

$$\mathrm{Det}(A) = a^2 + b^2\,, \quad \text{and} \quad A^+ = \frac{1}{(a^2 + b^2)^2}A\,. \tag{2.35}$$

We thus obtain the intriguing result

$$\nabla\mathbf{Det}(A) = \frac{1}{a^2 + b^2}A = \frac{1}{a^2 + b^2}\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix} = \frac{1}{(a, b)(a, b)^T}(a, b)^T(a, b)\,, \tag{2.36}$$

where the last form is meant to make clear that the result is the projection onto the subspace spanned by $(a, b)^T$.

The previous example touches on graph theory if we let $(a, b) = (\sqrt{c}, -\sqrt{c})$.

Example 2.19. Let L denote the Laplacian $L = D - A$ of a weighted graph, where A is the weighted adjacency matrix having zeros down the diagonal and off-diagonal elements A_{ij} equal to the value associated with the edge connecting nodes i and j. The matrix D is diagonal and has elements satisfying $D_{ij} = \sum_i A_{ij} = \sum_j A_{ij}$.

In the special case of a connected, two node graph with edge value c, the Laplacian is

$$L = \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} - \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} = c \cdot \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{2.37}$$

Noting that L is a projection-dilation matrix (see prior example), we get

$$\mathrm{Det}(L) = \sqrt{c}^2 + (-\sqrt{c})^2 = 2c\,, \quad \text{and} \quad L^+ = \frac{1}{4c^2} L\,, \tag{2.38}$$

and thus, by Formula (2.28),

$$\nabla \mathbf{Det}(L) = \frac{2c}{4c^2} L = \frac{1}{2c} L = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{2.39}$$

The last term is half the Laplacian associated to the simple, *unweighted* graph obtained by removing the weight c. Hence, $\nabla \mathbf{Det}(L)$ takes graph connectivity into account but not scale.

2.1. The matrix differential

When obtaining matrix derivatives, it is often easiest to calculate the matrix differential dA and then relate back to the gradient using the formula [9]

$$dh(A) = \mathrm{tr}\left((dA)\, G\right) \iff \nabla h(A) = G\,. \tag{2.40}$$

Combining this identity with directional derivative Formula (2.7), we see that $Ker(dA)$ must equal $Ker(A)$ for the special case of the derivative of the pseudo determinant. It follows that the matrix differential of the pseudo determinant is

$$d\,\mathrm{Det}(A) = \mathrm{Det}(A)\ \mathrm{tr}\left(A^+(dA)\right), \tag{2.41}$$

where we are implicitly selecting for the canonical gradient $\nabla \mathbf{Det}(A)$ in order to satisfy $Ker(dA) = Ker(A)$. Equation (2.41) may also be derived directly using the spectral

decomposition $A = U\Lambda U^* = \sum_{j=1}^{k} \lambda_j \, u_j u_j^*$ for rank k, Hermitian A. The differential of an eigenvalue of a Hermitian matrix A may be written in terms of the matrix differential itself [9]:

$$d\lambda = \operatorname{tr}\left(uu^* \, (dA)\right). \tag{2.42}$$

Theorem 2.20. *The matrix differential of the pseudo determinant of Hermitian* $A \in M_{n\times n}^{k}$ *is*

$$d\operatorname{Det}(A) = \operatorname{Det}(A) \, \operatorname{tr}\left(A^{+}(dA)\right). \tag{2.43}$$

Proof. The result is proven directly using Formula (2.42).

$$\begin{aligned} d\operatorname{Det}(A) &= d\prod_{j=1}^{k} \lambda_j \\ &= \sum_{j=1}^{k} d\lambda_j \prod_{i\neq j} \lambda_i \\ &= \sum_{j=1}^{k} \operatorname{tr}\left(u_j u_j^* \, (dA)\right) \prod_{i\neq j} \lambda_i \\ &= \sum_{j=1}^{k} \operatorname{tr}\left(\frac{1}{\lambda_j} u_j u_j^* \, (dA)\right) \prod_{i=1}^{k} \lambda_i \\ &= \operatorname{Det}(A) \sum_{j=1}^{k} \operatorname{tr}\left(\frac{1}{\lambda_j} u_j u_j^* \, (dA)\right) \\ &= \operatorname{Det}(A) \, \operatorname{tr}\left(\sum_{j=1}^{k} \frac{1}{\lambda_j} u_j u_j^* \, (dA)\right) \\ &= \operatorname{Det}(A) \, \operatorname{tr}\left(A^{+}(dA)\right) \quad \square \end{aligned} \tag{2.44}$$

The reader should note that Theorem 2.20 could also be used to derive the canonical gradient $\nabla\mathbf{Det}(A)$ via Formula (2.40).

3. An example from statistics

We now derive the maximum likelihood estimator (MLE) for the singular covariance of the degenerate multivariate Gaussian distribution. Thus, this section may be considered an extension of the results found in [5]. The MLE may be incorporated into more advanced statistical algorithms such as expectation maximization for image segmentation [2]. The formulas derived in the following are also potentially useful in a Hamiltonian

Monte Carlo algorithm for Bayesian inference over reduced-rank covariance matrices (cf. [10]).

Let $x_1, \ldots, x_N$ follow a degenerate Gaussian distribution with mean μ and singular covariance Σ. The probability density function of such a random variable x_i is given by

$$f(x_i; \mu, \Sigma) = \mathrm{Det}(2\pi\Sigma)^{-1/2} \exp\big(-\frac{1}{2}(x_i - \mu)^T \Sigma^+ (x_i - \mu)\big)\,. \tag{3.1}$$

Assuming that μ is known, the log-likelihood $\ell(\Sigma)$ of Σ is proportional to

$$-N \log\big(\mathrm{Det}(\Sigma)\big) - \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^+ (x_i - \mu) = -N \log\big(\mathrm{Det}(\Sigma)\big) - \mathrm{tr}\left(\Sigma^+ R\right), \tag{3.2}$$

where R is the matrix of residuals.

To obtain the MLE $\hat{\Sigma}$, we obtain the gradient of $\ell(\Sigma)$ and set it to zero, just as in the case of a full-rank covariance matrix. To calculate the second term in the log-likelihood, we need the formula for the matrix differential of the pseudo inverse [7]:

$$d\Sigma^+ = -\Sigma^+ (d\Sigma)\Sigma^+ + \Sigma^+\Sigma^+ (d\Sigma)(I - \Sigma\Sigma^+) + (I - \Sigma^+\Sigma)(d\Sigma)\Sigma^+\Sigma^+\,. \tag{3.3}$$

It follows that

$$\begin{aligned} d\ell(\Sigma) &= -N\,\mathrm{tr}\left(\Sigma^+ (d\Sigma)\right) + \mathrm{tr}\left(\Sigma^+ (d\Sigma)\Sigma^+ R\right) \\ &\quad - \mathrm{tr}\left(\Sigma^+\Sigma^+ (d\Sigma)(I - \Sigma\Sigma^+)R\right) - \mathrm{tr}\left((I - \Sigma^+\Sigma)(d\Sigma)\Sigma^+\Sigma^+ R\right) \\ &= -N\,\mathrm{tr}\left(\Sigma^+ (d\Sigma)\right) + \mathrm{tr}\left(\Sigma^+ R\Sigma^+ (d\Sigma)\right) \\ &\quad - \mathrm{tr}\left((I - \Sigma\Sigma^+)R\Sigma^+\Sigma^+ (d\Sigma)\right) - \mathrm{tr}\left(\Sigma^+\Sigma^+ R(I - \Sigma^+\Sigma)(d\Sigma)\right). \end{aligned} \tag{3.4}$$

Setting $d\ell(\hat{\Sigma}) = 0$ and applying Formula (2.40), we get

$$N\hat{\Sigma}^+ = \hat{\Sigma}^+ R\hat{\Sigma}^+ - (I - \hat{\Sigma}\hat{\Sigma}^+)R\hat{\Sigma}^+\hat{\Sigma}^+ - \hat{\Sigma}^+\hat{\Sigma}^+ R(I - \hat{\Sigma}^+\hat{\Sigma})\,, \tag{3.5}$$

and multiplying both sides by $\hat{\Sigma}$ on the right and the left gives

$$\begin{aligned} N\hat{\Sigma} &= \hat{\Sigma}\hat{\Sigma}^+ R\hat{\Sigma}^+\hat{\Sigma} - \hat{\Sigma}(I - \hat{\Sigma}\hat{\Sigma}^+)R\hat{\Sigma}^+\hat{\Sigma}^+\hat{\Sigma} - \hat{\Sigma}\hat{\Sigma}^+\hat{\Sigma}^+ R(I - \hat{\Sigma}^+\hat{\Sigma})\hat{\Sigma} \\ &= \hat{\Sigma}\hat{\Sigma}^+ R\hat{\Sigma}^+\hat{\Sigma}\,. \end{aligned} \tag{3.6}$$

This last line follows because the matrices $\Sigma\Sigma^+$ and $\Sigma^+\Sigma$ are projections onto the range of Σ and Σ^+, and therefore $(I - \Sigma^+\Sigma)$ and $(I - \Sigma\Sigma^+)$ annihilate Σ. For the same reason, *if we are willing to assume that* $Ker(R) = Ker(\Sigma)$, this last equation is solved by

$$\hat{\Sigma} = \frac{1}{N}\hat{\Sigma}\hat{\Sigma}^+ R\hat{\Sigma}^+\hat{\Sigma} = \frac{1}{N}R\,. \tag{3.7}$$

Thus only with that key assumption are we able to reproduce the classical result for full rank Σ. If we are not willing to make this assumption, i.e. if we have prior belief that, or have set up our model in such a way that, the range of Σ is a predetermined subspace, then the above equation may be written

$$\hat{\Sigma} = \frac{1}{N}\hat{\Sigma}\hat{\Sigma}^{+}R\hat{\Sigma}^{+}\hat{\Sigma} = \hat{\Sigma} = \frac{1}{N}\Sigma\Sigma^{+}R\Sigma^{+}\Sigma\,. \tag{3.8}$$

Then $\hat{\Sigma}$ is precisely the projection of the residual matrix R/N onto the range of Σ.

References

[1] Oliver Knill, Cauchy–Binet for pseudo-determinants, Linear Algebra Appl. 459 (2014) 522–547.
[2] Allen Y. Yang, et al., Unsupervised segmentation of natural images via lossy data compression, Comput. Vis. Image Underst. 110 (2) (2008) 212–225.
[3] Mario H. Castaneda, Josef A. Nossek, Estimation of rank deficient covariance matrices with Kronecker structure, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2014, pp. 394–398.
[4] Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer Science & Business Media, 2013.
[5] Theodore Wilbur Anderson, Ingram Olkin, Maximum-likelihood estimation of the parameters of a multivariate normal distribution, Linear Algebra Appl. 70 (1985) 147–171.
[6] Thomas P. Minka, Inferring a Gaussian Distribution, 1998.
[7] Gene H. Golub, Victor Pereyra, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, SIAM J. Numer. Anal. 10 (2) (1973) 413–432.
[8] Lothar Berg, Three results in connection with inverse matrices, Linear Algebra Appl. 84 (1986) 63–77.
[9] Jan R. Magnus, Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley Series in Probability and Mathematical Statistics, 1988.
[10] Andrew Holbrook, et al., Geodesic Lagrangian Monte Carlo over the space of positive definite matrices: with application to Bayesian spectral density estimation, J. Stat. Comput. Simul. 88 (5) (2018) 982–1002.