*The Annals of Statistics*
2021, Vol. 49, No. 5, 2721–2754
https://doi.org/10.1214/21-AOS2052
© Institute of Mathematical Statistics, 2021

# EXISTENCE AND UNIQUENESS OF THE KRONECKER COVARIANCE MLE

BY MATHIAS DRTON[1], SATOSHI KURIKI[2] AND PETER HOFF[3]

[1]*Department of Mathematics, Technical University of Munich, mathias.drton@tum.de*
[2]*The Institute of Statistical Mathematics, kuriki@ism.ac.jp*
[3]*Department of Statistical Science, Duke University, peter.hoff@duke.edu*

In matrix-valued datasets the sampled matrices often exhibit correlations among both their rows and their columns. A useful and parsimonious model of such dependence is the matrix normal model, in which the covariances among the elements of a random matrix are parameterized in terms of the Kronecker product of two covariance matrices, one representing row covariances and one representing column covariance. An appealing feature of such a matrix normal model is that the Kronecker covariance structure allows for standard likelihood inference even when only a very small number of data matrices is available. For instance, in some cases a likelihood ratio test of dependence may be performed with a sample size of one. However, more generally the sample size required to ensure boundedness of the matrix normal likelihood or the existence of a unique maximizer depends in a complicated way on the matrix dimensions. This motivates the study of how large a sample size is needed to ensure that maximum likelihood estimators exist, and exist uniquely with probability one. Our main result gives precise sample size thresholds in the paradigm where the number of rows and the number of columns of the data matrices differ by at most a factor of two. Our proof uses invariance properties that allow us to consider data matrices in canonical form, as obtained from the Kronecker canonical form for matrix pencils.

## 1. Introduction.

1.1. *Kronecker covariances and matrix normal models*. A matrix-valued dataset consists of a sample of matrices $Y_1, \dots, Y_n$, each taking values in $\mathbb{R}^{m_1 \times m_2}$ for integers $m_1, m_2 \geq 2$. Such data arise in spatial statistics [16] as well as in a variety of experimental settings where outcomes are obtained under combinations of two conditions, such as international trade between pairs of countries [25], studies involving multivariate time-series of EEG measurements on multiple individuals [18], age by period human mortality data [12] and factorial experiments arising in genomics [2], to name a few. In these applications, the matrices $Y_1, \dots, Y_n$ often exhibit substantial covariance among their rows and covariance among their columns, in the sense that (after appropriately subtracting out any mean or regression effects) the two empirical covariance matrices $\frac{1}{n}\sum_{i=1}^n Y_i Y_i^T$ and $\frac{1}{n}\sum_{i=1}^n Y_i^T Y_i$ are substantially nondiagonal. This motivates the use of a statistical model that can represent such data features.

One suitable and widely-used model is based on matrix normal distributions; see [6, 13, 14, 21] or also the work on graphical modeling in [1, 5, 27–29]. Let $Z$ be an $m_1 \times m_2$ random matrix with i.i.d. standard normal entries. A random matrix $Y$ taking values in $\mathbb{R}^{m_1 \times m_2}$ is said to have a matrix normal distribution if there exist (deterministic) matrices $M \in \mathbb{R}^{m_1 \times m_2}$, $A \in \mathbb{R}^{m_1 \times m_1}$ and $B \in \mathbb{R}^{m_2 \times m_2}$ such that $Y \stackrel{d}{=} M + AZB^T$. In this case, we write $Y \sim \mathcal{N}(M, \Sigma_2 \otimes \Sigma_1)$, where $\Sigma_1 = AA^T$ and $\Sigma_2 = BB^T$ and "$\otimes$" denotes the Kronecker product. This notation

Received March 2020; revised October 2020.
*MSC2020 subject classifications.* Primary 62H12; secondary 62R01.
*Key words and phrases.* Gaussian distribution, Kronecker canonical form, matrix normal model, maximum likelihood estimation, separable covariance.

reflects that $\Sigma_2 \otimes \Sigma_1$ is the covariance matrix of the vectorization of $Y$. If $A$ and $B$, and thus also $\Sigma_1$ and $\Sigma_2$, are invertible then the matrix normal distribution $\mathcal{N}(M, \Sigma_2 \otimes \Sigma_1)$ is regular and has a Lebesgue density.

For fixed $m_1$ and $m_2$, the matrix normal model is the set of all regular matrix normal distributions. The number of its covariance parameters is on the order of $m_1^2 + m_2^2$, which is a substantial reduction as compared to the normal model with unrestricted covariance and on the order of $m_1^2 m_2^2$ parameters. As a result, matrix normal models can be used to make likelihood-based inference with sample sizes that would preclude use of an unrestricted covariance model. For example, in [25], matrix normal models were used to construct nontrivial tests of dependence for square data matrices even when the sample size was one.

1.2. *Practical relevance of sample size conditions.* Numerical experiments indicate that the sample size required for a bounded likelihood or unique maximum likelihood estimator (MLE) depends in subtle ways on the dimensions $m_1$ and $m_2$ of the data matrices; see Example 1.1 below. Despite recent progress on sufficient sample size conditions, the precise behavior of the matrix normal likelihood function in settings with small sample size $n$ is not fully understood [23]. Specifically, from prior literature it is not known when precisely the *Kronecker MLE*, that is, MLE of the covariance matrix $\Sigma_2 \otimes \Sigma_1$, exists or exists uniquely. This is the problem we consider in this article, which gives new sample size conditions that provide a full solution to the problem for matrices whose dimensions $m_1$ and $m_2$ differ by a factor of at most 2; this is the regime where prior results leave the largest gaps. Precise sample size conditions are of great practical relevance as they are useful for the development of numerical methods to obtain MLEs, the design of experiments and as a guide to alternative data analysis strategies when conditions for existence or uniqueness are not met.

Regarding numerical methods, matrix normal models generally do not admit a closed-form MLE. Hence, a data analyst relies on iterative methods of computation, for which suitable convergence criteria need to be set [9]. However, as we will describe below, iteration of such an algorithm will either converge to a unique maximizer, converge to a nonunique maximizer, or not converge. Absent knowledge of the sample size conditions for existence of the Kronecker MLE, at each iteration the algorithm would have to distinguish in an ad-hoc manner between a situation in which the procedure has not yet sufficiently converged but will converge eventually, and one in which it will never converge. Additionally, in cases where the MLE exists, the analyst would certainly want to know if it exists uniquely, as nonuniqueness implies multiple explanations of the data generating mechanism that are equally valid (based on likelihood). Absent knowledge of the sample size conditions for uniqueness, the data analyst might not be aware that an MLE obtained via the iterative procedure was not unique. Even if the procedure was run from several different starting values, an assessment of nonuniqueness could be imprecise, as differences from different runs could be due to nonuniqueness, or just an indication that the algorithm has not been sufficiently iterated. In contrast, precise sample size conditions allow one to side-step these numerical issues.

Knowledge of sample size conditions can also assist with study design. For example, consider an experiment in which each replication produces a matrix of gene expression levels for a set of tissue types. If interest is inferring dependencies among genes and among types using the matrix normal model (see, e.g., [11]), then knowledge of the sample size requirement for existence and uniqueness of the Kronecker MLE would certainly be useful. Additionally, in cases where the study has been completed and the sample size conditions have not been met, our results still provide a guide to alternative methods for making inference. For example, suppose we wish to test a null hypothesis $H : \Sigma_2 = I$ of exchangeability within each row. If the sample size is insufficient for existence of the MLE, then the likelihood ratio statistic is infinity with probability one and a nontrivial level-$\alpha$ likelihood ratio test is unavailable.

However, suppose the maximum number $m_2'$ of columns for which the maximized likelihood is bounded, and thus for which a likelihood ratio test with nontrivial power may be obtained, were known. In this case, the null hypothesis $H$ could be evaluated by properly combining the results of multiple tests of exchangeability on subsets of $m_2'$ columns, even at sample sizes as small as $n = 1, 2$.

1.3. *Maximum likelihood thresholds and known results.* For notational simplicity, we obtain results for samples of size $n$ from a mean-zero matrix normal model, that is,

$$Y_1, \dots, Y_n \sim \text{i.i.d. } \mathcal{N}(0, \Sigma_2 \otimes \Sigma_1). \tag{1.1}$$

As noted in Remark 1.1, sample-size thresholds for the case of an unknown mean will be equal to those in this mean-zero case, plus one additional data matrix. Let $\mathrm{PD}(m)$ be the cone of positive definite $m \times m$ matrices. Let $\Psi_1 = \Sigma_1^{-1}$ and $\Psi_2 = \Sigma_2^{-1}$ be the precision matrices. Ignoring an additive constant, two times the log-likelihood function for the matrix normal model can be written as

$$\ell(\Psi_1, \Psi_2) = nm_2 \log\det(\Psi_1) + nm_1 \log\det(\Psi_2) - \operatorname{tr}\left(\Psi_1 \sum_{i=1}^{n} Y_i \Psi_2 Y_i^T\right). \tag{1.2}$$

The log-likelihood function only depends on the Kronecker product $\Psi_2 \otimes \Psi_1$ and

$$\ell\big(c\Psi_1, c^{-1}\Psi_2\big) = \ell(\Psi_1, \Psi_2) \tag{1.3}$$

for all scalars $c > 0$.

A standard method to compute the MLE of $\Psi_2 \otimes \Psi_1$ is block-coordinate descent, also referred to as the "flip-flop algorithm" [9]. If $\Psi_2$ is fixed to be a value $\tilde{\Psi}_2$, the likelihood function is maximized by $\tilde{\Psi}_1 = \tilde{\Sigma}_1^{-1}$ with $\tilde{\Sigma}_1 = \frac{1}{nm_2} \sum_i Y_i \tilde{\Psi}_2 Y_i^T$. Similarly, since the trace term in (1.2) can be alternatively written as $\operatorname{tr}(\Psi_2 \sum_i Y_i^T \Psi_1 Y_i)$, the maximizer when fixing $\Psi_1 = \tilde{\Psi}_1$ is $\tilde{\Psi}_2 = \tilde{\Sigma}_2^{-1}$ with $\tilde{\Sigma}_2 = \frac{1}{nm_1} \sum_i Y_i^T \tilde{\Psi}_1 Y_i$. The flip-flop algorithm proceeds by iteratively updating $\tilde{\Psi}_1$ and $\tilde{\Psi}_2$ (or equivalently $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$) using these formulas. Evidently, the updates are well defined as long as both $\sum_i Y_i \tilde{\Psi}_2 Y_i^T$ and $\sum_i Y_i^T \tilde{\Psi}_1 Y_i$ are invertible. As will be made precise later in Lemma 2.3, this condition will be met a.s. as long as the "row sample size" $nm_2$ is as big as the number of rows $m_1$, and similarly $nm_1$ is as big as $m_2$.

EXAMPLE 1.1. Consider the case of $n = 2$ data matrices of size $(m_1, m_2)$, where we fix $m_2 = 4$ and consider $m_1 = 5, 6, 7, 8$. In each case we take the data to be in a canonical form (as specified in Theorem 5.1) and run the flip-flop algorithm starting from a random choice for $\Psi_1$. Figure 1 depicts the behavior of the algorithm in terms of the function $g$ defined as $-2/n$ times the log-likelihood (solid line), where we also omitted additive constants from the log-likelihood; see (2.2). In addition, the figure shows the difference in $g$ from one iteration to the next (dashed line). The top left figure, which is for $(m_1, m_2) = (5, 4)$, shows the function $g = -(2/n)\ell$ converging to its minimum. The top right figure, for $(m_1, m_2) = (6, 4)$, shows the function $g$ converging similarly. In the bottom left figure, for $(m_1, m_2) = (7, 4)$, the function $g$ diverges to $-\infty$; the dashed line stays at a nonzero negative value. Finally, in the bottom right figure, for $(m_1, m_2) = (8, 4)$, the algorithm converges in one step. These four settings correspond to the cases where the MLE exists uniquely (top left), MLEs exist nonuniquely (top and bottom right), and an MLE does not exist (bottom left). We emphasize that while our figure pertains to one particular initialization the observed behavior is similar for other random starting values.

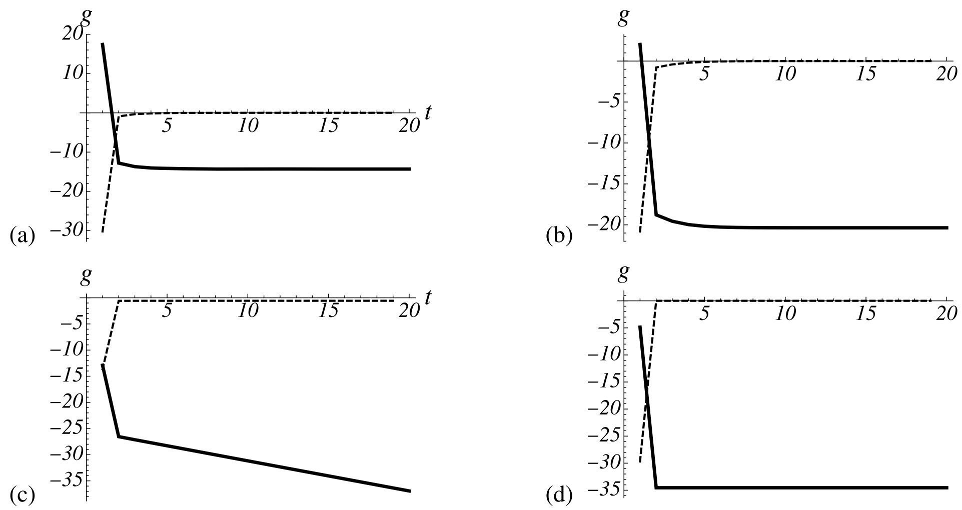


FIG. 1. *Plots of profile log-likelihood function g (solid line) and its one-step differences (dashed line) against the number of iterations t of the flip-flop algorithm*: (*a*) $(m_1, m_2) = (5, 4)$ *with unique MLE*, (*b*) $(m_1, m_2) = (6, 4)$ *with the MLE existing nonuniquely*, (*c*) $(m_1, m_2) = (7, 4)$ *with the MLE nonexisting*, (*d*) $(m_1, m_2) = (8, 4)$ *with the MLE existing nonuniquely.*

As explained in Section 5, the behavior of the log-likelihood function with respect to existence and uniqueness of the Kronecker MLE is essentially independent of the realizations $Y_i$, and merely depends on the triple $(m_1, m_2, n)$. It is well known that for large enough $n$ the Kronecker MLE exists uniquely a.s.; this is simply a consequence of the properties of the usual (vector) normal model. However, special and at times somewhat paradoxical properties of the matrix normal model emerge for small sample size $n$. Indeed, as we noted in Example 1.1, for fixed $m_2$ minor differences of $m_1$ may cause substantially different behavior of the log-likelihood function when the sample size $n$ is small. To capture this behavior, we define in this paper three types of sample size thresholds, which are critical sample sizes at which the a.s. behavior of the likelihood function changes. Our terminology follows [8, 15].

DEFINITION 1.1.

(i) The Kronecker MLE *exists* if the function $\ell$ achieves its maximum over the domain of definition $\mathrm{PD}(m_1) \times \mathrm{PD}(m_2)$. It *exists uniquely* if in addition all local maxima of $\ell$ have the same Kronecker product.

(ii) We define three positive integer thresholds for the sample size. The *existence threshold* $N_e(m_1, m_2)$ is the integer such that the Kronecker MLE exists a.s. if and only if $n \geq N_e(m_1, m_2)$. The *uniqueness threshold* $N_u(m_1, m_2)$ is the integer such that the Kronecker MLE exist uniquely a.s. if and only if $n \geq N_u(m_1, m_2)$. Finally, the *boundedness threshold* $N_b(m_1, m_2)$ is the integer such that $\ell$ is bounded a.s. if and only if $n \geq N_b(m_1, m_2)$.

REMARK 1.1. Throughout the paper, we assume the expectation of $Y_i$ to be zero and known. Standard results yield that $N_b(m_1, m_2) + 1$, $N_e(m_1, m_2) + 1$, and $N_u(m_1, m_2) + 1$ are the relevant thresholds for the case where the mean matrix in $\mathbb{R}^{m_1 \times m_2}$ is unknown and also estimated by maximum likelihood [4], Section 3.3.

The three thresholds from Definition 1.1 are finite and no larger than $m_1 m_2$. Specifically, the well-known results on estimation of an unconstrained Gaussian covariance matrix collected in [4], Section 3.2, together with Lemma 2.2 below yield that $n \geq m_1 m_2$ is sufficient for a.s. unique existence of the Kronecker MLE. However, this condition is far from necessary. Indeed, the main theorem of [23] states that unique existence holds a.s. under the much weaker requirement that

$$n > \frac{m_1}{m_2} + \frac{m_2}{m_1}.$$

To our knowledge, this is the best known sufficient condition for a.s. (unique) existence.

The condition

$$n \geq \max\left\{\frac{m_1}{m_2}, \frac{m_2}{m_1}\right\} \tag{1.4}$$

is necessary for existence; compare also Theorem 1(1) in [23]. This is a consequence of the following simple lemma, for which we include a proof in the Appendix.

LEMMA 1.2. *Suppose $m_1 \geq m_2$. If $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^{m_1 \times nm_2}$ has row rank smaller than $m_1$, then $\ell(\Psi_1, \Psi_2)$ is not bounded above on* $\mathrm{PD}(m_1) \times \mathrm{PD}(m_2)$.

Note from our earlier discussion that the condition from (1.4) is precisely the requirement needed for the flip-flop algorithm to have well-defined iterative update steps. Somewhat confusingly, some of the literature refers to (1.4) as the necessary and sufficient condition for existence of the Kronecker MLE; see, for example, [9]. However, even when (1.4) holds, the likelihood function need not achieve its maximum, or even be bounded (recall Example 1.1).

In terms of the thresholds we defined, the known results from the literature may be summarized as follows. The floor and ceiling functions are denoted by $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$, respectively.

PROPOSITION 1.3. *The three ML thresholds satisfy*

$$\max\left\{\frac{m_1}{m_2}, \frac{m_2}{m_1}\right\} \leq N_b(m_1, m_2) \leq N_e(m_1, m_2) \leq N_u(m_1, m_2) \leq \left\lfloor \frac{m_1}{m_2} + \frac{m_2}{m_1} \right\rfloor + 1.$$

We remark that $\frac{m_1}{m_2} + \frac{m_2}{m_1}$ is integer if and only if $m_1 = m_2$. So, for rectangular matrices ($m_1 \neq m_2$), the upper bound on $N_u(m_1, m_2)$ can be written as $\lceil \frac{m_1}{m_2} + \frac{m_2}{m_1} \rceil$.

1.4. *New contributions.* Our interest is in precise formulas for the thresholds. The case where one matrix dimension divides the other is the simplest. We derive the following result in the Appendix. The result considers $m_2 \geq 2$. The case $m_2 = 1$ reduces to the vector case in which the MLE exists uniquely a.s. for $n \geq m_1$ and does not exist (with unbounded likelihood) if $n < m_1$ [4], Section 3.2; so $N_b(m_1, 1) = N_e(m_1, 1) = N_u(m_1, 1) = m_1$.

THEOREM 1.4. *Suppose $m_1 \geq m_2 \geq 2$ and $m_2 | m_1$, that is, $m_1$ is divisible by $m_2$. Then*

$$N_b(m_1, m_2) = N_e(m_1, m_2) = \frac{m_1}{m_2}, \qquad N_u(m_1, m_2) = \begin{cases} 3 & \text{if } m_1 = m_2, \\ \dfrac{m_1}{m_2} + 1 & \text{if } m_1 > m_2. \end{cases}$$

For the case where the matrix dimensions do not divide each other, Proposition 1.3 yields a solution when one matrix dimension is sufficiently large compared to the other.

COROLLARY 1.5. *Suppose that $m_1 > m_2 \geq 2$. Let $r = m_1 \bmod m_2$ be the remainder in integer division, so $r \in \{0, \dots, m_2 - 1\}$. If $r \geq 1$ and*

$$\left\lfloor \frac{m_1}{m_2} \right\rfloor > 1 + \frac{r^2}{m_2(m_2 - r)},$$

*then*

$$N_b(m_1, m_2) = N_e(m_1, m_2) = N_u(m_1, m_2) = \left\lfloor \frac{m_1}{m_2} \right\rfloor + 1.$$

*This formula holds, in particular, when* $m_1 \geq m_2^2$ *and* $r \geq 1$.

Our new results in Theorem 1.4 and Corollary 1.5 can be roughly interpreted as saying that, as long as $m_1$ is divisible or approximately divisible by $m_2$, then the sample size requirements for existence of an MLE are the same as the sample size requirements for the steps of the flip-flop algorithm to be well defined. This equivalence is perhaps the reason why in the early literature on Kronecker MLEs, these two sample sizes were conflated. However, as illustrated in Example 1.1, the flip-flop algorithm can diverge even though each step is well defined. The main additional results in our paper are to understand and describe this discrepancy as well as the difference between existence and unique existence of the MLE. Specifically, we provide formulas for the three thresholds we defined in the regime where the upper bound from Proposition 1.3 leaves the largest gap, namely, the case where

$$2m_2 \geq m_1 \geq m_2.$$

We state our main theorem here.

THEOREM 1.6. *Suppose that* $2m_2 \geq m_1 \geq m_2$. *Then*

$$N_u(m_1, m_2) = \begin{cases} 3 & \text{if } m_1 = m_2, \\ 2 & \text{if } m_1 = m_2 + 1, \\ 3 & \text{otherwise}, \end{cases}$$

*and*

$$N_e(m_1, m_2) = N_b(m_1, m_2) = \begin{cases} 1 & \text{if } m_1 = m_2, \\ 2 & \text{if } m_1 > m_2 \text{ and } m_1 - m_2 | m_2, \\ 3 & \text{otherwise}. \end{cases}$$

The ingredients needed to establish Theorem 1.6 will be developed in the remainder of the paper. How they fit together is also outlined in the proof of Theorem 1.6 that we include in the Appendix. When combined with additional calculations using Gröbner basis methods to check algebraic conditions given in Theorems 3.1 and 3.3, Theorem 1.6 provides $N_u(m_1, m_2)$, $N_e(m_1, m_2)$ and $N_b(m_1, m_2)$ for small $m_1$ and $m_2$; see Table 1.

The remainder of the paper is organized as follows. We begin by recalling preliminary results concerning convexity properties of the negated log-likelihood function, and we introduce a profile likelihood function (Section 2). We then give algebraic conditions for existence and uniqueness of the Kronecker MLE (Section 3). These are formulated in terms of rank drops, meaning the extent to which the rank of a data matrix may be reduced through certain linear transformations. The sufficient condition also appears in a more geometric form in the proofs in [23]. The key ingredients for the proof of Theorem 1.4 are derived in Section 4. The proof of Theorem 1.6 requires a study of the case of sample size $n = 2$ and is developed in Section 5. Our arguments use invariance of the likelihood surface under group actions, and we use certain canonical forms for generic data matrices under the group action. For data in such canonical form, we are able to explicitly give the critical points of the likelihood function (Section 7). We end with a brief conclusion (Section 8).

TABLE 1
*ML thresholds: Unique existence* $N_u(m_1, m_2)$ (*left*);
*Bounded likelihood/existence* $N_e(m_1, m_2) = N_b(m_1, m_2)$ (*right*)

| $m_1 \setminus m_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | |
| 2 | 2 | 3 | | | | | | | | |
| 3 | 3 | 2 | 3 | | | | | | | |
| 4 | 4 | 3 | 2 | 3 | | | | | | |
| 5 | 5 | 3 | 3 | 2 | 3 | | | | | |
| 6 | 6 | 4 | 3 | 3 | 2 | 3 | | | | |
| 7 | 7 | 4 | 3 | 3 | 3 | 2 | 3 | | | |
| 8 | 8 | 5 | 3 | 3 | 3 | 3 | 2 | 3 | | |
| 9 | 9 | 5 | 4 | 3 | 3 | 3 | 3 | 2 | 3 | |
| 10 | 10 | 6 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 3 |

| $m_1 \setminus m_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | |
| 2 | 2 | 1 | | | | | | | | |
| 3 | 3 | 2 | 1 | | | | | | | |
| 4 | 4 | 2 | 2 | 1 | | | | | | |
| 5 | 5 | 3 | 3 | 2 | 1 | | | | | |
| 6 | 6 | 3 | 2 | 2 | 2 | 1 | | | | |
| 7 | 7 | 4 | 3 | 3 | 3 | 2 | 1 | | | |
| 8 | 8 | 4 | 3 | 2 | 3 | 2 | 2 | 1 | | |
| 9 | 9 | 5 | 3 | 3 | 3 | 2 | 3 | 2 | 1 | |
| 10 | 10 | 5 | 4 | 3 | 2 | 3 | 3 | 2 | 2 | 1 |

## 2. Preliminaries.

2.1. *Geodesic convexity and uniqueness of MLE.* The log-likelihood function $\ell$ defined in (1.2) is not concave. However, it can be shown to be g-concave, that is, $\ell$ is concave along suitable geodesics between any pair of matrices in $\mathrm{PD}(m_1) \times \mathrm{PD}(m_2)$. The geodesics are obtained from the geodesics in $\mathrm{PD}(m_1)$ and in $\mathrm{PD}(m_2)$, which take the form

$$\gamma_t^{(j)}(Q_0, Q_1) = Q_0^{\frac{1}{2}}\left(Q_0^{-\frac{1}{2}} Q_1 Q_0^{-\frac{1}{2}}\right)^t Q_0^{\frac{1}{2}}, \quad t \in [0, 1], \tag{2.1}$$

when linking two matrices $Q_0, Q_1 \in \mathrm{PD}(m_j)$, $j = 1, 2$. The g-concavity of $\ell$ amounts to

$$\ell\left(\gamma_t^{(1)}\left(Q_0^{(1)}, Q_1^{(1)}\right), \gamma_t^{(2)}\left(Q_0^{(2)}, Q_1^{(2)}\right)\right) \geq t \cdot \ell\left(Q_0^{(1)}, Q_0^{(2)}\right) + (1-t) \cdot \ell\left(Q_1^{(1)}, Q_1^{(2)}\right)$$

for all $t \in [0, 1]$, $Q_0^{(j)}, Q_1^{(j)} \in \mathrm{PD}(m_j)$, $j = 1, 2$. This property of $\ell$ was observed in [26] and yields the following fact; see also [22], Chapter 6.

LEMMA 2.1. *Every critical point* (*i.e.*, *point of zero gradient*) *and*, *in particular*, *every local maximum of* $\ell$ *is a global maximum.*

Suppose $\ell$ has two distinct maxima. Then, by concavity, $\ell$ is constant along the geodesic linking them. We may extend the geodesic by considering $t \in \mathbb{R}$ in (2.1). Along this extended geodesic, $\ell$ remains constant even as the underlying matrices diverge or approach the boundary of $\mathrm{PD}(m_1)$ or $\mathrm{PD}(m_2)$. This can be used to guarantee uniqueness of the Kronecker MLE as noted in [23]. Call the log-likelihood function $\ell$ *coercive* if

$$\lim_{t \to \infty} \ell\left(\Psi_1^{(t)}, \Psi_2^{(t)}\right) = -\infty$$

for all sequences $\left(\Psi_1^{(t)}, \Psi_2^{(t)}\right)$ that diverge or approach the boundary of $\mathrm{PD}(m_1) \times \mathrm{PD}(m_2)$. If $\ell$ is coercive, then clearly the Kronecker MLE exists and it exists uniquely based on our above discussion. However, more is true according to the following lemma that is proven as part of Lemma 4 in [23].

LEMMA 2.2. *The Kronecker MLE exists uniquely if and only if the log-likelihood function* $\ell$ *is coercive.*

In the following, we will often consider properties that the normal data matrices $Y_1, \dots, Y_n$ possess almost surely. More precisely, the properties will hold as long as the data lie outside certain, not further specified lower-dimensional sets that are defined by polynomial equations. We indicate this fact by speaking of data matrices that are *generic*.

2.2. *Profile likelihood.* For any $\Psi_2 \in \mathrm{PD}(m_2)$, the log-likelihood function admits a section $\ell_{\Psi_2} : \mathrm{PD}(m_1) \to \mathbb{R}$ given by $\ell_{\Psi_2}(\Psi_1) = \ell(\Psi_1, \Psi_2)$.

LEMMA 2.3. *Suppose $nm_2 \geq m_1 \geq m_2$. For generic data $Y_1, \dots, Y_n$, every section $\ell_{\Psi_2}$, $\Psi_2 \in \mathrm{PD}(m_2)$, achieves its maximum on $\mathrm{PD}(m_1)$ uniquely at*

$$\Psi_1(\Psi_2) = \left( \frac{1}{nm_2} \sum_{i=1}^{n} Y_i \Psi_2 Y_i^T \right)^{-1}.$$

PROOF. According to the well-known results for ML estimation of a Gaussian covariance matrix [4], Section 3.2, the claim is true for a particular restriction $\ell_{\Psi_2}$ if the positive semi-definite $m_1 \times m_1$ matrix

$$\sum_{i=1}^{n} Y_i \Psi_2 Y_i^T$$

is nonsingular. The matrix is a sum of positive semidefinite matrices. Hence, a vector $v \in \mathbb{R}^{m_2}$ is in its kernel if and only if $v^T Y_i \Psi_2 Y_i^T v = 0$ for all $i = 1, \dots, n$. Since $\Psi_2$ is positive definite, this holds if and only if $v$ is in the kernel of each matrix $Y_i^T$ if and only if $v$ is in the kernel of

$$\sum_{i=1}^{n} Y_i Y_i^T.$$

Being a sum of $nm_2 \geq m_1$ generic rank 1 matrices, this $m_1 \times m_1$ matrix has full rank $m_1$. The kernel is thus zero. □

In the regime of interest, with $nm_2 \geq m_1 \geq m_2$, Lemma 2.3 ensures that for generic data $Y_1, \dots, Y_n$ the profile log-likelihood function

$$\ell_{\mathrm{prof}} : \Psi_2 \mapsto \ell\big(\Psi_1(\Psi_2), \Psi_2\big)$$

is well-defined on $\mathrm{PD}(m_2)$. Now, $\ell$ is bounded above/achieves its maximum on $\mathrm{PD}(m_1) \times \mathrm{PD}(m_2)$ if and only if $\ell_{\mathrm{prof}}$ is bounded above/achieves its maximum on $\mathrm{PD}(m_2)$. This in turn is equivalent to the function

$$g(\Psi) = m_2 \log\det\left( \sum_{i=1}^{n} Y_i \Psi Y_i^T \right) - m_1 \log\det(\Psi) \tag{2.2}$$

being bounded below/achieving its minimum on $\mathrm{PD}(m_2)$. We note that

$$g(\Psi) = m_2 \log\det\big(Y[I_n \otimes \Psi]Y^T\big) - m_1 \log\det(\Psi), \tag{2.3}$$

where $Y = \big(Y_1, \dots, Y_n\big) \in \mathbb{R}^{m_1 \times nm_2}$. The function $g$ is geodesically convex (g-convex). This follows from its definition as negatived profile of a g-concave log-likelihood function. It can also be seen directly by observing that $\log\det(\cdot)$ is linear along the geodesics from (2.1) and verifying that the first term in (2.2) is g-convex; see Lemma 2 in [26].

Call $g$ coercive if $g(\Psi^{(t)})$ tends to $+\infty$ for any sequence $\Psi^{(t)}$ that diverges or approaches the boundary of $\mathrm{PD}(m_2)$. Then our observations may be summarized as follows.

LEMMA 2.4. *The Kronecker MLE exists if and only if the function $g$ from (2.2) achieves its minimum on $\mathrm{PD}(m_2)$. The Kronecker MLE exists uniquely if and only if $g$ is coercive.*

PROOF. The first assertion of this lemma is clear from the definition of $g$, and the second claim follows from Lemma 2.2. □

2.3. *Group action.* An important ingredient to our later analysis is the fact that a group action allows one to consider data in canonical position. Let $\mathrm{GL}(m,\mathbb{R})$ be the general linear group of $m \times m$ real invertible matrices. The direct product $\mathrm{GL}(m_1,\mathbb{R}) \times \mathrm{GL}(m_2,\mathbb{R})$ acts naturally on a data set comprised of matrices $Y_1,\dots,Y_n \in \mathbb{R}^{m_1\times m_2}$. For $A \in \mathrm{GL}(m_1,\mathbb{R})$ and $B \in \mathrm{GL}(m_2,\mathbb{R})$, the action is

$$Y_i \mapsto AY_iB, \quad i = 1,\dots,n.$$

Now two data sets $(Y_1,\dots,Y_n)$ and $(Y_1',\dots,Y_n')$ are in the same orbit under the group action if one can be transformed into the other using a pair $(A,B) \in \mathrm{GL}(m_1,\mathbb{R}) \times \mathrm{GL}(m_2,\mathbb{R})$.

Recall that the log-likelihood surface of a model is the graph of its log-likelihood function.

LEMMA 2.5. *If two data sets are in the same* $\mathrm{GL}(m_1,\mathbb{R}) \times \mathrm{GL}(m_2,\mathbb{R})$*-orbit, then their log-likelihood surfaces are translations of one another.*

PROOF. Let $\ell$ be the log-likelihood function from (1.2) for data $Y_1,\dots,Y_n \in \mathbb{R}^{m_1\times m_2}$. Let $A \in \mathrm{GL}(m_1,\mathbb{R})$ and $B \in \mathrm{GL}(m_2,\mathbb{R})$. Define $Y_i' = AY_iB$ for $i = 1,\dots,n$. The log-likelihood function for the data $(Y_1',\dots,Y_n')$ is

$$\begin{aligned}\ell'(\Psi_1,\Psi_2) &= nm_2 \log\det(\Psi_1) + nm_1 \log\det(\Psi_2) - \operatorname{tr}\left(\Psi_1 A \sum_{i=1}^n Y_i B\Psi_2 B^T Y_i^T A^T\right)\\ &= \ell\big(A^T\Psi_1 A, B\Psi_2 B^T\big) + c,\end{aligned}$$

where $c$ is a constant that depends on $n, m_1, m_2, \det(A)$ and $\det(B)$. The maps $\Psi_1 \mapsto A^T\Psi_1 A$ and $\Psi_2 \mapsto B\Psi_2 B^T$ are bijections from $\mathrm{PD}(m_1)$ to $\mathrm{PD}(m_1)$ and $\mathrm{PD}(m_2)$ to $\mathrm{PD}(m_2)$, respectively. Therefore, subtracting $c$ from each function value translates the graph of $\ell'$ into the graph of $\ell$. □

**3. Algebraic conditions for existence and uniqueness.** In this section, we prove a necessary condition for existence of the Kronecker MLE as well as a sufficient condition for its unique existence. Both conditions involve the rank of

$$Y = (Y_1,\dots,Y_n) \in \mathbb{R}^{m_1\times nm_2} \tag{3.1}$$

after linear transformation using certain $nm_2 \times nm_2$ matrices in Kronecker product form.

3.1. *Necessary condition for existence.* The following condition strengthens the necessary condition from Lemma 1.2.

THEOREM 3.1.

(i) *If there exists a matrix* $X \in \mathbb{R}^{m_2\times m_2}$ *such that*

$$\operatorname{rank}\big(Y_1X,\dots,Y_nX\big) < \frac{m_1}{m_2}\operatorname{rank}(X),$$

*then the log-likelihood function $\ell$ is unbounded and the Kronecker MLE does not exist.*

(ii) *If there exists a nonzero and singular matrix* $X \in \mathbb{R}^{m_2\times m_2}$ *such that*

$$\operatorname{rank}\big(Y_1X,\dots,Y_nX\big) \le \frac{m_1}{m_2}\operatorname{rank}(X),$$

*then the log-likelihood function $\ell$ is not coercive and the Kronecker MLE does not exist uniquely.*

PROOF. (i) Let $X \in \mathbb{R}^{m_2 \times m_2}$ satisfy the assumed rank condition. If $X$ is invertible, then the data matrix $Y$ from (3.1) has $\operatorname{rank}(Y) < m_1$ and the likelihood function is unbounded by Lemma 1.2. Hence, only the case where $\operatorname{rank}(X) = k < m_1$ needs to be considered.

Let $q_1, \dots, q_{m_2}$ be the eigenvectors of $XX^T$, with corresponding eigenvalues $d_1, \dots, d_{m_2}$. Since $\operatorname{rank}(XX^T) = \operatorname{rank}(X) = k$, we may assume that $d_j = 0$ for $j \geq k+1$, in which case

$$XX^T = \sum_{j=1}^{k} d_j q_j q_j^T.$$

Define the positive semidefinite matrix

$$\Psi = \sum_{j=k+1}^{m_2} q_j q_j^T.$$

Then $XX^T + \Psi$ is positive definite. We claim that

$$\lim_{\lambda \to \infty} g\bigl(\lambda XX^T + \Psi\bigr) = -\infty, \tag{3.2}$$

which implies the theorem's assertion via Lemma 2.4.

Let $r$ be the rank of

$$(Y_1 X, \dots, Y_n X) = Y(I_n \otimes X).$$

Then $r$ is also the rank of

$$Y(I_n \otimes X)(I_n \otimes X)^T Y^T = Y\bigl(I_n \otimes XX^T\bigr)Y^T.$$

By Lemma 3.2 below, the determinant of

$$Y\bigl[I_n \otimes \bigl(\lambda XX^T + \Psi\bigr)\bigr]Y^T = \lambda \cdot Y\bigl[I_n \otimes XX^T\bigr]Y^T + Y[I_n \otimes \Psi]Y^T$$

is a polynomial of degree $r$ in $\lambda$. Lemma 3.2 also yields that $\det(\lambda XX^T + \Psi)$ is a polynomial of degree $k$ in $\lambda$. By assumption $rm_2 < km_1$. Therefore,

$$\lim_{\lambda \to \infty} e^{g(\lambda XX^T + \Psi)} = \lim_{\lambda \to \infty} \frac{\{\det(Y[I_n \otimes (\lambda XX^T + \Psi)]Y^T)\}^{m_2}}{\{\det(\lambda XX^T + \Psi)\}^{m_1}} = 0. \tag{3.3}$$

Taking the logarithm yields the claim from (3.2).

(ii) Proceeding as in case (i), we obtain in (3.3) a ratio of two polynomials of equal degree and a finite and positive limit. It follows that $g$ converges to a finite limit, and thus is not coercive. An application of Lemma 2.4 yields the claim. $\square$

LEMMA 3.2. *Let $A, B \in \mathbb{R}^{m \times m}$ be two positive semidefinite matrices whose sum $A + B$ is positive definite. Let* $\operatorname{rank}(A) = r$. *Then* $\det(\gamma A + B)$ *is a degree $r$ polynomial in $\gamma$.*

PROOF. Choose an invertible matrix $C$ such that $C^T C = A + B$. Let $Q^T D Q$ be the spectral decomposition of $C^{-T} A C^{-1}$ with $D = \operatorname{diag}(d_1, \dots, d_r, 0, \dots, 0)$ and $d_j > 0$ for $j \leq r$. Then

$$\begin{aligned} \det(\gamma A + B) &= \det\bigl((\gamma - 1)A + (A + B)\bigr) \\ &= \det(C)^2 \det\bigl((\gamma - 1)D + I\bigr) = \det(A + B)^2 \prod_{j=1}^{r} (d_j \gamma + 1 - d_j). \end{aligned}$$

$\square$

REMARK 3.1. The eigenvalues $d_j$ in the above proof are also the eigenvalues of $(A + B)^{-1}A$. If $v \in \ker(B)$ then

$$(A + B)^{-1} A v = (A + B)^{-1}(A + B)v = v,$$

so that $v$ is an eigenvector for eigenvalue 1.

3.2. *Sufficient conditions for existence and uniqueness.*

THEOREM 3.3. *Let* $\operatorname{rank}(Y_1, \dots, Y_n) = m_1$.

(i) *If all singular matrices* $X \in \mathbb{R}^{m_2 \times m_2}$ *satisfy*

$$\operatorname{rank}\bigl(Y_1 X, \dots, Y_n X\bigr) \geq \frac{m_1}{m_2} \operatorname{rank}(X),$$

*then the log-likelihood function* $\ell$ *is bounded from above.*

(ii) *If all nonzero singular matrices* $X \in \mathbb{R}^{m_2 \times m_2}$ *satisfy*

$$\operatorname{rank}\bigl(Y_1 X, \dots, Y_n X\bigr) > \frac{m_1}{m_2} \operatorname{rank}(X),$$

*then the log-likelihood function* $\ell$ *is coercive and the Kronecker MLE exists uniquely.*

PROOF. (i) Assume that the log-likelihood function $\ell$ is not bounded from above. Then there exists a sequence $\Psi^{(t)}$, $t = 1, 2, \dots$, in $\mathrm{PD}(m_2)$ such that $g(\Psi^{(t)}) \to -\infty$ as $t \to \infty$. Let $\Psi^{(t)} = Q^{(t)} \Lambda^{(t)} (Q^{(t)})^T$ be the spectral decomposition. The set of orthogonal matrices is compact and, passing to a subsequence if necessary, we may assume the sequence $Q^{(t)}$ to be convergent. By Lemma A.1 in the Appendix, again passing to a subsequence if necessary, we may assume the diagonal elements of $\Lambda^{(t)}$ to be such that the resulting sequence $\Psi^{(t)}$ is of the form

$$\Psi^{(t)} = \epsilon_1^{(t)} \Psi_1^{(t)} + \epsilon_2^{(t)} \Psi_2^{(t)} + \cdots + \epsilon_K^{(t)} \Psi_K^{(t)}, \quad K \leq m_2, \tag{3.4}$$

where $\Psi_k^{(t)}$ is a sequence of positive semidefinite matrices that converges to a limit $\Psi_k$ and such that for all $t$ it holds that $\operatorname{rank} \Psi_k^{(t)} = \operatorname{rank} \Psi_k$ and

$$\operatorname{Im} \Psi_1^{(t)} \oplus \cdots \oplus \operatorname{Im} \Psi_K^{(t)} = \mathbb{R}^{m_2}. \tag{3.5}$$

Moreover, the scalars in (3.4) are positive, $\epsilon_k^{(t)} > 0$, and satisfy $\epsilon_{k+1}^{(t)} / \epsilon_k^{(t)} \to 0$. Note also that in (3.4) we must have $K \geq 2$ because if $K = 1$, then $\lim_{t \to \infty} \Psi_1^{(t)} = \Psi_1 \in \mathrm{PD}(m_2)$ and $\lim_{t \to \infty} g(\Psi^{(t)}) = \lim_{t \to \infty} g(\epsilon_1^{(t)} \Psi_1^{(t)}) = \lim_{t \to \infty} g(\Psi_1^{(t)}) = \lim_{t \to \infty} g(\Psi_1) > -\infty$.

Next, we redefine $\Psi_k^{(t)}$ and $\epsilon_k^{(t)}$ as

$$\epsilon_k^{(t)} := \epsilon_k^{(t)} - \epsilon_{k+1}^{(t)}, \quad k = 1, \dots, K-1, \qquad \epsilon_K^{(t)} := \epsilon_K^{(t)},$$

$$\Psi_k^{(t)} := \Psi_1^{(t)} + \cdots + \Psi_k^{(t)}, \qquad \Psi_k := \Psi_1 + \cdots + \Psi_k, \quad k = 1, \dots, K.$$

The new $\Psi^{(t)}$ remains of the form (3.4) with $\Psi_k^{(t)} \to \Psi_k$, $\operatorname{rank} \Psi_k^{(t)} = \operatorname{rank} \Psi_k$. Similarly, the new $\epsilon_k^{(t)}$ remains positive with $\epsilon_{k+1}^{(t)} / \epsilon_k^{(t)} \to 0$. However, instead of (3.5), we now have

$$\operatorname{Im} \Psi_1^{(t)} \subset \cdots \subset \operatorname{Im} \Psi_K^{(t)} = \mathbb{R}^{m_2}. \tag{3.6}$$

Write $r_k = \operatorname{rank} \Psi_k$. Then, by (3.6) and Lemma 3.2,

$$\begin{aligned}
\det\bigl(\Psi^{(t)}\bigr) &= \det\bigl(\epsilon_1^{(t)} \Psi_1^{(t)} + \epsilon_2^{(t)} \Psi_2^{(t)} + \cdots + \epsilon_K^{(t)} \Psi_K^{(t)}\bigr) \\
&= \bigl(\epsilon_K^{(t)}\bigr)^{m_2} \det\bigl(\bigl(\epsilon_1^{(t)} / \epsilon_K^{(t)}\bigr) \Psi_1^{(t)} + \bigl(\epsilon_2^{(t)} / \epsilon_K^{(t)}\bigr) \Psi_2^{(t)} + \cdots + \Psi_K^{(t)}\bigr) \\
&\asymp \bigl(\epsilon_K^{(t)}\bigr)^{m_2} \bigl(\epsilon_1^{(t)} / \epsilon_K^{(t)}\bigr)^{r_1} \bigl(\epsilon_2^{(t)} / \epsilon_K^{(t)}\bigr)^{r_2 - r_1} \cdots \bigl(\epsilon_{K-1}^{(t)} / \epsilon_K^{(t)}\bigr)^{r_{K-1} - r_{K-2}} \\
&= \bigl(\epsilon_K^{(t)}\bigr)^{m_2} \bigl(\epsilon_1^{(t)} / \epsilon_2^{(t)}\bigr)^{r_1} \bigl(\epsilon_2^{(t)} / \epsilon_3^{(t)}\bigr)^{r_2} \cdots \bigl(\epsilon_{K-1}^{(t)} / \epsilon_K^{(t)}\bigr)^{r_{K-1}} \\
&= \bigl(\epsilon_K^{(t)}\bigr)^{m_2} \bigl(\gamma_1^{(t)}\bigr)^{r_1} \bigl(\gamma_2^{(t)}\bigr)^{r_2} \cdots \bigl(\gamma_{K-1}^{(t)}\bigr)^{r_{K-1}},
\end{aligned}$$

where $\gamma_k^{(t)} = \epsilon_k^{(t)} / \epsilon_{k+1}^{(t)}$. Note that $\gamma_k^{(t)} \to \infty$ as $t \to \infty$.

Let $M^{(t)} = \sum_{i=1}^n Y_i \Psi^{(t)} Y_i^T$ and $M_k^{(t)} = \sum_{i=1}^n Y_i \Psi_k^{(t)} Y_i^T$. Then, as $t \to \infty$, $M_k^{(t)} \to M_k = \sum_{i=1}^n Y_i \Psi_k Y_i^T$. The monotonicity property (3.6) is inherited as

$$\operatorname{Im} M_1^{(t)} \subset \cdots \subset \operatorname{Im} M_K^{(t)} = \mathbb{R}^{m_1},$$

and, therefore,

$$\det(M^{(t)}) \asymp (\epsilon_K^{(t)})^{m_2} (\gamma_1^{(t)})^{\operatorname{rank}(M_1^{(t)})} (\gamma_2^{(t)})^{\operatorname{rank}(M_2^{(t)})} \cdots (\gamma_{K-1}^{(t)})^{\operatorname{rank}(M_{K-1}^{(t)})}.$$

Define

$$R(k) := \min_{X \in \mathbb{R}^{m_2 \times m_2}, \operatorname{rank} X = k} \operatorname{rank}(Y_1 X, \dots, Y_n X).$$

The assumption in part (i) of the theorem is that

$$R(r) - (m_1/m_2) r \geq 0, \quad r = 1, \dots, m_2 - 1. \tag{3.7}$$

Then the order of $\det(M^{(t)})$ is bounded from below by

$$(\epsilon_K^{(t)})^{m_1} (\gamma_0^{(t)})^{R(r_0)} (\gamma_1^{(t)})^{R(r_1)} \cdots (\gamma_{K-1}^{(t)})^{R(r_{K-1})},$$

and the order of

$$e^{g(\Psi^{(t)})} = \frac{\det(M^{(t)})^{m_2}}{\det(\Psi^{(t)})^{m_1}}$$

is bounded from below by

$$(\gamma_0^{(t)})^{m_2 R(r_0) - m_1 r_0} (\gamma_1^{(t)})^{m_2 R(r_1) - m_1 r_1} \cdots (\gamma_{K-1}^{(t)})^{m_2 R(r_{K-1}) - m_1 r_{K-1}}. \tag{3.8}$$

Under the condition (3.7), the product in (3.8), and thus also $e^{g(\Psi^{(t)})}$ does not converge to 0. This means that $g(\Psi^{(t)})$ is bounded from below and cannot diverge to $-\infty$. This is a contradiction.

(ii) The assumption is now that

$$R(r) - (m_1/m_2) r > 0, \quad r = 1, \dots, m_2 - 1. \tag{3.9}$$

Let $\Psi^{(t)}$ be a sequence in $\mathrm{PD}(m_2)$ such that $\Psi^{(t)} \to \Psi^0$, where $\Psi^0$ is singular. Assume that the likelihood function $\ell$ is not coercive and $g(\Psi^{(t)})$ is bounded from above. As in the proof of (i), we can take a subsequence of the form (3.4) with $\Psi_1 = \Psi^0$ and $\epsilon_1^{(t)} \to 1$. Under the assumption (3.9), the lower bound (3.8) of $e^{g(\Psi^{(t)})}$ always diverges to infinity, which is a contradiction. □

3.3. *Minimal ranks.* Assume, as throughout, that we have data matrices $Y_1, \dots, Y_n \in \mathbb{R}^{m_1 \times m_2}$ with $m_1 \geq m_2$. For $k = 1, \dots, m_2$, define the minimal rank

$$r_n(m_1, m_2, k) = \min\{\operatorname{rank}(Y_1 X, \dots, Y_n X) : X \in \mathbb{R}^{m_2 \times k}, \operatorname{rank}(X) = k\}. \tag{3.10}$$

Now define

$$S_n(m_1, m_2) = \min_{1 \leq k < m_2} \{m_2 r_n(m_1, m_2, k) - m_1 k\}. \tag{3.11}$$

For generic data matrices, the results in this section can be summarized as follows.

THEOREM 3.4. *The likelihood function is a.s. bounded if and only if a.s. $S_n(m_1, m_2) \geq 0$. The Kronecker MLE exists uniquely a.s. if and only if a.s. $S_n(m_1, m_2) > 0$.*

**4. Square matrices.** This section treats the case of square data matrices. So, $m_1 = m_2$ and we denote this common value also by $m$. The results we develop, specifically, Corollaries 4.3, 4.6 and 4.9, yield the following statement about the three sample size thresholds.

THEOREM 4.1. *For square data matrices of any size $m \geq 2$, it holds that $N_e(m,m) = N_b(m,m) = 1$ and $N_u(m,m) = 3$.*

Throughout the remainder of this section, we tacitly assume that each one of the data matrices $Y_1, \dots, Y_n$ is invertible, as is the case almost surely.

4.1. *One square data matrix.* We begin with an observation utilized in [25].

PROPOSITION 4.2. *If $m_1 = m_2$ and $n = 1$, then the profile likelihood function $g$ from* (2.2) *is constant.*

PROOF. The single $m \times m$ data matrix $Y_1$ being invertible, we have

$$g(\Psi) = m \log \det\bigl(Y_1 \Psi Y_1^T\bigr) - m \log \det(\Psi) = 2m \log\bigl|\det(Y_1)\bigr|,$$

which does not depend on $\Psi$. $\square$

The proposition implies that for $n = 1$ the likelihood function achieves its maximum but not uniquely so. We may deduce from the rank conditions in Section 3 that $r_1(m,m,k) \geq k$ for all $m \geq 2$ and $k = 1, \dots, m$. As $r_n(m_1, m_2, k)$ is nondecreasing in $n$, we obtain that $r_n(m,m,k) \geq k$ always, which implies $S_n(m,m) \geq 0$. By Theorem 3.4(i), we have the following.

COROLLARY 4.3. *The boundedness threshold of square matrices of size $m \geq 2$ is $N_b(m,m) = 1$.*

4.2. *Two square data matrices.* Moving to the case of $n = 2$ square data matrices, we first provide detail on the ranks $r_2(m,m,k)$.

LEMMA 4.4. *Let $Y_1, Y_2 \in \mathbb{R}^{m \times m}$ be generic, and let $1 \leq k \leq m$. If $Y_1^{-1} Y_2$ has a real eigenvalue, or if $k$ is even, then*

$$r_2(m,m,k) = k.$$

*If all eigenvalues of $Y_1^{-1} Y_2$ are complex and if $k$ is odd, then*

$$r_2(m,m,k) = k + 1.$$

PROOF. Let $W = Y_1^{-1} Y_2$. We evaluate

$$r_2(m,m,k) = \min_{\operatorname{rank}(X)=k} \operatorname{rank}\bigl(Y_1 X, Y_2 X\bigr) = \min_{\operatorname{rank}(X)=k} \operatorname{rank}\bigl(X, WX\bigr).$$

Since $W$ is real and generic, its characteristic function does not have multiple zeros and all of its (complex) Jordan blocks are of size 1. For any real eigenvalue $\lambda_j$, let $x_j \in \mathbb{R}^m \setminus \{0\}$ be an associated real eigenvector such that

$$W x_j = \lambda_j x_j.$$

For any pair of complex eigenvalues $\mu_j \pm i\nu_j$, let $y_j, z_j \in \mathbb{R}^m \setminus \{0\}$ be real vectors such that $y_j \pm i z_j \in \mathbb{C}^m$ are eigenvectors corresponding to $\mu_j \pm i\nu_j$, so

$$W(y_j \pm i z_j) = (\mu_j \pm i\nu_j)(y_j \pm i z_j) \quad \iff \quad W\bigl(y_j, z_j\bigr) = \bigl(y_j, z_j\bigr)\begin{pmatrix} \mu_j & \nu_j \\ -\nu_j & \mu_j \end{pmatrix}.$$

Altogether the vectors $x_j$, $y_j$, and $z_j$ form a basis of $\mathbb{R}^m$.

Suppose now that $W$ has at least one real eigenvalue or $k$ is even. Choose a full rank matrix $X^* \in \mathbb{R}^{m\times k}$ by selecting its columns as individual vectors $x_j$ or pairs $(y_j, z_j)$, that is,

$$X^* = (x_1, \dots, x_s, y_1, z_1, \dots, y_t, z_t), \quad s + 2t = k.$$

Then

$$r_2(m, m, k) \le \operatorname{rank}\left(X^*, WX^*\right) = k.$$

On the other hand, trivially we have

$$r_2(m, m, k) = \min_{\operatorname{rank}(X)=k} \operatorname{rank}\left(X, WX\right) \ge \min_{\operatorname{rank}(X)=k} \operatorname{rank}(X) = k.$$

In the remaining case where $k$ is odd and all eigenvalues complex, we may reduce the rank of $(X, WX)$ to $k + 1$ by choosing choose $k - 1$ columns of $X$ based on eigenvectors. However, we cannot reduce rank further as this would contradict the linear independence of eigenvectors. □

LEMMA 4.5. *Square matrices of size $m \ge 2$ have*

$$S_2(m, m) = \begin{cases} 2 & \text{if } m = 2 \text{ and } Y_1^{-1}Y_2 \text{ has complex eigenvalues}, \\ 0 & \text{otherwise}. \end{cases}$$

PROOF. When $m = 2$ and $Y_1^{-1}Y_2$ has complex eigenvalues, $r_2(2, 2, 1) = 1 + 1$, and $S_2(2, 2) = \min_{1\le k\le 2-1}\{2r_2(2, 2, k) - 2k\} = 2r_2(2, 2, 1) - 2 \times 1 = 2$.

When $m = 2$ and $Y_1^{-1}Y_2$ has real eigenvalues, $r_2(2, 2, 1) = 1$, and $S_2(2, 2) = 2r_2(2, 2, 1) - 2 \times 1 = 0$.

When $m \ge 3$, because of $r_2(m, m, k) \ge k$, $S_2(m, m) = \min_{1\le k\le m-1}\{mr_2(m, m, k) - mk\} \ge 0$ holds, and the equality always attains at $k = 2$, $r_2(m, m, k) = r_2(m, m, 2) = 2$. □

COROLLARY 4.6. *The uniqueness threshold of square matrices of size $m \ge 2$ is $N_u(m, m) = 3$.*

PROOF. We know from Proposition 1.3 that $N_u(m, m) \le 3$. If $m \ge 3$, then Lemma 4.4 yields that $S_2(m, m) = 0$ generically, and thus a sample size of $n = 2$ is not sufficient for a.s. unique existence of the Kronecker MLE. Hence, $N_u(m, m) = 3$.

If $m = 2$, then there is the subtlety that $S_2(m, m) = 0$ if $Y_1^{-1}Y_2$ has real eigenvalues, and $S_2(m, m) = 1$ if $Y_1^{-1}Y_2$ has complex eigenvalues. However, as either case arises with positive probability, $N_u(m, m) = 3$ also for $m = 2$. □

REMARK 4.1. Contrasting cases (i) and (iii) in Proposition 4.7, we see that the minimal rank $r_n(m_1, m_2, k)$ may change when we minimize over complex instead of real matrices.

4.3. *Achieving maximum likelihood for two square data matrices.* As we know that $N_u(m, m) = 3$ and $N_b(m, m) = 1$, with maximum achieved, there only remains the question whether the (bounded) likelihood function achieves its maximum for a sample of $n = 2$ invertible data matrices. We begin with the smallest case of $m = 2$, which exhibits exceptional behavior as noted in the proof of Corollary 4.6; see also [23], Section 4.3. The following proposition considers all possible cases and gives their probabilities.

PROPOSITION 4.7. *Suppose $n = 2$ with invertible data matrices $Y_1$ and $Y_2$ of size $2 \times 2$. Three cases are possible*:

(i) *The matrix* $Y_1^{-1}Y_2$ *has real eigenvalues and is diagonalizable. The likelihood function is then bounded and achieves its maximum, but not uniquely so.*

(ii) *The matrix* $Y_1^{-1}Y_2$ *has real eigenvalues but is not diagonalizable. The likelihood function is then bounded but fails to achieve its maximum.*

(iii) *The eigenvalues of* $W = (w_{jk}) = Y_1^{-1}Y_2$ *are complex. The Kronecker MLE then exists uniquely and is given by any positive definite matrix of the form*

$$\Psi = \lambda \begin{pmatrix} w_{12} & \frac{1}{2}(w_{22} - w_{11}) \\ \frac{1}{2}(w_{22} - w_{11}) & -w_{21} \end{pmatrix}, \qquad \lambda \in \mathbb{R}.$$

*Only cases* (*i*) *and* (*iii*) *occur with positive probability. If the entries of* $Y_1$ *and* $Y_2$ *are i.i.d.* $\mathcal{N}(0, 1)$, *then case* (*i*) *has probability* $\pi/4 \approx 0.7854$.

PROOF. We may put $Y_1$ and $Y_2$ in special position through the action of $\mathrm{GL}(m, \mathbb{R}) \times \mathrm{GL}(m, \mathbb{R})$ discussed in Section 2.3. With $A = B^{-1}Y_1^{-1}$, we have

$$AY_1B = I_m, \qquad AY_2B = B^{-1}Y_1^{-1}Y_2B.$$

Now choose $B$ such that $AY_2B$ becomes the real-valued Jordan form of $W = Y_1^{-1}Y_2$.

Case (iii): If the two eigenvalues are complex then $S_2(2, 2) = 1$ as noted in the proof of Corollary 4.6. By Theorem 3.4 the Kronecker MLE exists uniquely. In special form, our data matrices take the form

$$Y_1 = I_2, \qquad Y_2 = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \tag{4.1}$$

with $a, b \in \mathbb{R}$. Then the negated profile log-likelihood function from (2.2) is

$$g(\Psi) = 2\log\left(1 + 2a^2 + a^4 + b^4 + 2a^2b^2 + b^2 \frac{\|\Psi\|_F^2}{\det(\Psi)}\right).$$

Let $\lambda_1 \geq \lambda_2 > 0$ be the two eigenvalues of the positive definite $2 \times 2$ matrix $\Psi$. Then

$$\frac{\|\Psi\|_F^2}{\det(\Psi)} = \left(\frac{\lambda_1}{\lambda_2}\right)^2 + \left(\frac{\lambda_2}{\lambda_1}\right)^2$$

is minimal when $\lambda_1 = \lambda_2$, which occurs if and only if $\Psi = \lambda I_2$ for $\lambda > 0$. Translating back to the original data yields the claimed formula for the MLE.

Case (i): By the diagonalizability assumption, the special form of our data matrices is

$$Y_1 = I_2, \qquad Y_2 = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

with $a, b \in \mathbb{R}$. The negated profile log-likelihood function from (2.2) now equals

$$g(\Psi) = 2\log\left((1 + ab)^2 + (a - b)^2 \frac{\psi_{11}\psi_{22}}{\det(\Psi)}\right). \tag{4.2}$$

Let $\rho \equiv \rho(\Psi) = \psi_{12}/\sqrt{\psi_{11}\psi_{22}}$ be the correlation. Then

$$\frac{\psi_{11}\psi_{22}}{\det(\Psi)} = \frac{1}{1 - \rho^2}$$

is minimized uniquely for $\rho = 0$. Hence, the function $g$ from (4.2) is minimized by all diagonal matrices. The likelihood function achieves its maximum but not uniquely so.

Case (ii): The special form of our data matrices is now

$$Y_1 = I_2, \qquad Y_2 = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}$$

with $a \in \mathbb{R}$. The negated profile log-likelihood function from (2.2) is

$$g(\Psi) = 2\log\left((1+a^2)^2 + \frac{\psi_{22}^2}{\det(\Psi)}\right) > 2\log\bigl((1+a^2)^2\bigr) \tag{4.3}$$

as $\psi_{22}, \det(\Psi) > 0$. If we fix the values $\psi_{11} = 1$ and $\psi_{12} = 0$, and let $\psi_{22} \to 0$, then $g(\Psi)$ converges to $2\log((1+a^2)^2)$. Hence,

$$2\log\bigl((1+a^2)^2\bigr) = \inf\{g(\Psi) : \Psi \in \mathrm{PD}(2)\}$$

but this infimum is not achieved.

Finally, case (ii) occurs with probability zero as $Y_1^{-1}Y_2$ has to have an eigenvalue of multiplicity two. The probability of case (i) is found in Lemma A.2 in the Appendix. $\square$

The dichotomy from the case of $2 \times 2$ matrices disappears for larger matrices.

PROPOSITION 4.8. *If $n = 2$ with square data matrices of size $m \geq 3$, then the likelihood function is a.s. bounded and achieves its maximum, but not uniquely so.*

PROOF. We prove the claim by exhibiting an at least two-dimensional set of critical points, which must all be global optima by Lemma 2.1.

As in the proof of Proposition 4.7, assume that $Y_1 = I_m$ is the identity and that $Y_2$ is in real Jordan form for the almost surely occurring case of all eigenvalues being distinct. Then $Y_2$ is block-diagonal with blocks of size 1 or 2; the $2 \times 2$ blocks are as in (4.1). As the matrix size is $m \geq 3$, there are $k \geq 2$ blocks, which we denote by $Y_{21}, \dots, Y_{2k}$. Let $b_1, \dots, b_k \in \{1, 2\}$ be the sizes of these blocks.

The profile function we minimize is

$$g(\Psi) = m \log\det\bigl(\Psi + Y_2 \Psi Y_2^T\bigr) - m \log\det(\Psi), \qquad \Psi \in \mathrm{PD}(m). \tag{4.4}$$

For each block, define an analogous function

$$g_l(\Psi_l) = b_l \log\det\bigl(\Psi_l + Y_{2l}\Psi_l Y_{2l}^T\bigr) - b_l \log\det(\Psi_l), \qquad \Psi_l \in \mathrm{PD}(b_l).$$

The logarithm of the determinant has differential

$$\mathrm{d}\log\det(\Psi) = \mathrm{tr}\bigl(\Psi^{-1}\mathrm{d}\Psi\bigr).$$

It follows that the differential of $g$ in (4.4) is

$$\mathrm{d}g(\Psi; U) = m\,\mathrm{tr}\bigl\{\bigl(\Psi + Y_2\Psi Y_2^T\bigr)^{-1}U\bigr\} - m\,\mathrm{tr}\bigl(\Psi^{-1}U\bigr).$$

As candidates, consider block-diagonal matrices $\Psi_0$ with $k$ blocks $\Psi_{01}, \dots, \Psi_{0k}$ of sizes $b_1, \dots, b_k$, respectively. Then $\Psi_0 + Y_2\Psi_0 Y_2^T$ is block-diagonal, and we have

$$\frac{1}{m}\mathrm{d}g(\Psi_0; U) = \sum_{l=1}^{k} \frac{1}{b_l}\mathrm{d}g_l(\Psi_{0l}; U_l).$$

Now, take each block of $\Psi_0$ to be a multiple of the identity, so $\Psi_{0l} = \lambda_l I_{b_l}$ for $l = 1, \dots, k$. If $b_l = 1$, then $\mathrm{d}g_l(\Psi_{0l}; U_l) = 0$ trivially because $g_l$ is then constant. If $b_l = 2$, then $\mathrm{d}g_l(\Psi_{0l}; U_l) = 0$ as we showed in the proof of Proposition 4.7 that $g_l$ is minimized by multiples of $I_2$. We conclude that block-diagonal matrices with blocks equal to multiples of the identity are critical points. As there are $k \geq 2$ blocks, the critical points we exhibited form a set of dimension at least 2. Hence, the likelihood function achieves its maximum, but not uniquely so. $\square$

COROLLARY 4.9. *The existence threshold of square matrices of size* $m \geq 2$ *is* $N_e(m,m) = 1$.

**5. Rectangular matrices.** In this section, we consider $n = 2$ rectangular matrices $Y_1$ and $Y_2$ of size $m_1 \times m_2$ with $m_1 > m_2$. As discussed in Section 1, the nontrivial case is then $nm_2 = 2m_2 > m_1 > m_2$. For this case, we derive explicit solutions for the minimal rank $r_2(m_1, m_2, k)$ in (3.10) and $S_2(m_1, m_2)$ in (3.11).

5.1. *Kronecker canonical form.* As discussed in Section 2.3, our problem is invariant with respect to the group action $Y_i \mapsto AY_iB$, $(A, B) \in \mathrm{GL}(m_1, \mathbb{R}) \times \mathrm{GL}(m_2, \mathbb{R})$. The theorem below states that when $Y_1$ and $Y_2$ are generic, by choosing $A$ and $B$ appropriately (depending on the data $Y_i$), the problem can be reduced into a simplified canonical form.

In the sequel, we write $0_{k,l}$ for the $k \times l$ matrix with all entries zero.

THEOREM 5.1. *Let* $Y_1$ *and* $Y_2$ *be generic rectangular matrices of size* $m_1 \times m_2$ *with* $2m_2 > m_1 > m_2$. *There exist* $A \in \mathrm{GL}(m_1, \mathbb{R})$ *and* $B \in \mathrm{GL}(m_2, \mathbb{R})$ *such that*

$$AY_1B = \begin{pmatrix} I_{m_2} \\ 0_{m_1-m_2,m_2} \end{pmatrix}, \qquad AY_2B = \begin{pmatrix} 0_{m_1-m_2,m_2} \\ I_{m_2} \end{pmatrix}. \tag{5.1}$$

PROOF. This is a variation of the Kronecker canonical form (see Remark 5.1). For constructive proofs, see [24] and [20], Theorem 5.1.8. $\square$

REMARK 5.1. Let

$$U_l = \begin{pmatrix} I_l \\ 0_{1,l} \end{pmatrix} \in \mathbb{R}^{(l+1)\times l}, \qquad L_l = \begin{pmatrix} 0_{1,l} \\ I_l \end{pmatrix} \in \mathbb{R}^{(l+1)\times l}.$$

It is known that for generic matrices $Y_1$ and $Y_2$ in Theorem 5.1, there exist $A \in \mathrm{GL}(m_1, \mathbb{R})$ and $B \in \mathrm{GL}(m_2, \mathbb{R})$ such that

$$\begin{aligned} AY_1B &= \mathrm{diag}(\underbrace{U_{l+1}, \dots, U_{l+1}}_{n_a}, \underbrace{U_l, \dots, U_l}_{n_b}), \\ AY_2B &= \mathrm{diag}(\underbrace{L_{l+1}, \dots, L_{l+1}}_{n_a}, \underbrace{L_l, \dots, L_l}_{n_b}), \end{aligned} \tag{5.2}$$

where $l$, $n_a$ and $n_b$ are functions of $(m_1, m_2)$ defined in (5.6) and (5.7); see [10], Section 3.3. The pair of block diagonal matrices in (5.2) is referred to as the Kronecker canonical form. We easily see that the form in (5.1) may be obtained from that in (5.2) by permuting rows and columns.

In the remainder of this section, we assume without loss of generality that $Y_1$ and $Y_2$ are already in the canonical form in (5.1). The minimal rank (3.10) we will determine then becomes

$$\begin{aligned} r_2(m_1, m_2, k) &= \min\left\{\mathrm{rank}\begin{pmatrix} XY_1^T \\ XY_2^T \end{pmatrix} : X \in \mathbb{R}^{k\times m_2}, \mathrm{rank}(X) = k\right\} \\ &= \min\left\{\mathrm{rank}\left(\begin{array}{c|c} X & 0 \\ \hline 0 & X \end{array}\right)_{2k\times m_1} : X \in \mathbb{R}^{k\times m_2}, \mathrm{rank}(X) = k\right\}, \end{aligned}$$

where for ease of presentation the matrix whose rank we consider has been transposed.

5.2. *Gröbner basis computation.* When $m_2$ is small, the minimal rank $r_2(m_1, m_2, k)$ can be found by algebraic computations. Since

$$\operatorname{rank}\left(\begin{array}{c|c} X & 0 \\ \hline 0 & X \end{array}\right) = \operatorname{rank}\left(\begin{array}{c|c} SX & 0 \\ \hline 0 & SX \end{array}\right), \quad S \in \mathrm{GL}(k, \mathbb{R}),$$

we can set $k$ columns of $X \in \mathbb{R}^{k\times m_2}$ to form an identity matrix. We thus proceed through the following steps:

Step 1. For each set $\{l_1, \dots, l_k\}$ with $1 \le l_1 < \cdots < l_k \le m_2$, repeat (i) and (ii) below:

(i) Let

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1m_2} \\ \vdots & & \vdots \\ x_{k1} & \cdots & x_{km_2} \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} x_{1l_1} & \cdots & x_{1l_k} \\ \vdots & & \vdots \\ x_{kl_1} & \cdots & x_{kl_k} \end{pmatrix} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

(ii) For $i = 0, 1, \dots$, try to solve the polynomial system

$$\text{all } (2k-i)\times(2k-i) \text{ minors of } \begin{pmatrix} X & 0_{k,m_1-m_2} \\ 0_{k,m_1-m_2} & X \end{pmatrix} = 0 \tag{5.3}$$

by computing and inspecting a Gröbner basis. If a real solution $X^*$ exists for $i = i^*$ but not for $i = i^* + 1$, let

$$\texttt{Rank}[\{l_1, \dots, l_k\}] := \operatorname{rank}\left(\begin{array}{c|c} X^* & 0 \\ \hline 0 & X^* \end{array}\right). \tag{5.4}$$

Step 2. Take the minimum for all possible $1 \le l_1 < \cdots < l_k \le m_2$:

$$r_2(m_1, m_2, k) = \min_{1\le l_1<\cdots<l_k\le m_2} \texttt{Rank}[\{l_1, \dots, l_k\}].$$

EXAMPLE 5.1. Suppose $m_1 = 5$, $m_2 = 3$, $k = 2$. For $\{l_1, l_2\} = \{1, 3\}$, $X = \left(\begin{smallmatrix} 1 & x_{12} & 0 \\ 0 & x_{22} & 1 \end{smallmatrix}\right)$, and

$$\left(\begin{array}{c|c} X & 0 \\ \hline 0 & X \end{array}\right)_{4\times 5} = \begin{pmatrix} 1 & x_{12} & 0 & 0 & 0 \\ 0 & x_{22} & 1 & 0 & 0 \\ 0 & 0 & 1 & x_{12} & 0 \\ 0 & 0 & 0 & x_{22} & 1 \end{pmatrix}. \tag{5.5}$$

The first, third and fifth column of this $4 \times 5$ matrix are linearly independent so its rank cannot drop below 3. This is reflected in the $3 \times 3$ minors being $\{x_{12}, x_{22}, x_{12}^2, x_{22}^2, x_{12}x_{22}, 1, 0\}$, with Gröbner basis $\{1\}$ and no solution (real or complex) for (5.3) when $i = 1$. However, for $i = 0$, the set of $4 \times 4$ minors of the matrix is $\{x_{12}, x_{22}, x_{12}^2, x_{22}^2, x_{12}x_{22}\}$ with a Gröbner basis being $\{x_{12}, x_{22}\}$. This confirms that $x_{12} = x_{22} = 0$ is the (evident) solution for (5.3) when $i = 0$. Our procedure concludes

$$\texttt{Rank}[\{1,3\}] = \operatorname{rank}\left(\begin{array}{c|c} X & 0 \\ \hline 0 & X \end{array}\right)\Bigg|_{x_{12}=x_{22}=0} = \operatorname{rank}\left(\begin{array}{ccc|cc} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right) = 3 < 2k = 4.$$

We observe a drop in rank. For the other combinations $\{l_1, l_2\} = \{1, 2\}$ and $\{2, 3\}$, no rank drop occurs, that is, $\texttt{Rank}[\{l_1, l_2\}] = 4$. Hence, $r_2(m_1, m_2, k) = r_2(5, 3, 2) = 3$.

In the example just given a well-devised 0–1 matrix $X$ attains the minimal rank. We shall see that such a matrix exists for general $(m_1, m_2, k)$; see the construction in (5.10).

5.3. *Evaluation of the minimal rank.* Our strategy to determine $r_2(m_1, m_2, k)$ is to first provide an upper bound by specifying a special 0–1 matrix $X$. We then prove that no other matrix can achieve lower rank than $X$. In order to state our results, some further notation is needed.

Let

$$l(m_1, m_2) = \max\{l \in \mathbb{N} \mid (l+1)m_2 - lm_1 > 0\} = \left\lceil \frac{m_2}{m_1 - m_2} \right\rceil - 1 \geq 1. \tag{5.6}$$

Based on the value $l(m_1, m_2)$, the set of pairs $(m_1, m_2)$ of interest is disjointly divided as

$$\begin{aligned}\{(m_1, m_2) \mid m_1 < 2m_2\} &= \bigsqcup_{l \geq 1} \{(m_1, m_2) \mid l = l(m_1, m_2)\} \\ &= \bigsqcup_{l \geq 1} \{(m_1, m_2) \mid (l+1)m_2 - lm_1 > 0, (l+1)m_1 - (l+2)m_2 \geq 0\}.\end{aligned}$$

Let

$$n_a = (l+1)m_2 - lm_1 > 0, \qquad n_b = (l+1)m_1 - (l+2)m_2 \geq 0 \quad \text{with } l = l(m_1, m_2). \tag{5.7}$$

Now we partition the columns of $X$ as

$$X = (\underbrace{X_1}_{n_a}, \underbrace{X_2}_{n_b}, \underbrace{X_3}_{n_a}, \underbrace{X_4}_{n_b}, \dots, \underbrace{X_{2l+1}}_{n_a})_{k \times m_2}.$$

Indeed, the number of columns of $X$ is

$$(l+1)n_a + ln_b = m_2.$$

Accordingly,

$$\left(\begin{array}{c|c} X & 0 \\ \hline 0 & X \end{array}\right)_{2k \times m_1} = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & \cdots & X_{2l+1} & 0 & 0 \\ 0 & 0 & X_1 & X_2 & \cdots & X_{2l-1} & X_{2l} & X_{2l+1} \end{pmatrix}_{2k \times m_1}. \tag{5.8}$$

THEOREM 5.2. *For given $(m_1, m_2)$ with $2m_2 > m_1 > m_2$, let $l = l(m_1, m_2)$, $n_a$, and $n_b$ be defined as in (5.6) and (5.7). Then the minimal rank $r_2(m_1, m_2, k)$ is the solution of the integer programming problem*

$$r_2(m_1, m_2, k) = \min\{a_1 + b_1 + k \mid k \leq a_1(l+1) + b_1 l, 0 \leq a_1 \leq n_a, 0 \leq b_1 \leq n_b\}, \tag{5.9}$$

*where $a_1$ and $b_1$ are nonnegative integers.*

PROOF. [Upper bound] We first show that the right-hand side of (5.9) is an upper bound on $r_2(m_1, m_2, k)$. To do this, we specify a particular matrix $X$ that gives a rank equal to the right-hand side of (5.9).

Let $n_a \geq a_1 \geq \cdots \geq a_{l+1} \geq 0$ and $n_b \geq b_1 \geq \cdots \geq b_l \geq 0$ such that $k_a = \sum_{j=1}^{l+1} a_j$, $k_b = \sum_{j=1}^{l} b_j$ with $k_a + k_b = k$. For integer $i \geq 1$, define the matrices

$$X_{2i-1} = \begin{pmatrix} 0_{\sum_{j=1}^{i-1} a_j, n_a} \\ (I_{a_i}, 0_{a_i, n_a - a_i}) \\ 0_{\sum_{j=i+1}^{l+1} a_j, n_a} \\ 0_{k_b, n_a} \end{pmatrix}, \qquad X_{2i} = \begin{pmatrix} 0_{k_a, n_b} \\ 0_{\sum_{j=1}^{i-1} b_j, n_b} \\ (I_{b_i}, 0_{b_i, n_b - b_i}) \\ 0_{\sum_{j=i+1}^{l} b_j, n_b} \end{pmatrix}. \tag{5.10}$$

The matrices $X_{2i-1}$ are of size $k \times n_a$ and defined for $i \le l+1$. The matrices $X_{2i}$ are of size $k \times n_b$ and defined for $i \le l$. Then

$$\operatorname{rank}\begin{pmatrix} X_1 & X_2 & X_3 & X_4 & \cdots & X_{2l+1} & 0 & 0 \\ 0 & 0 & X_1 & X_2 & \cdots & X_{2l-1} & X_{2l} & X_{2l+1} \end{pmatrix}_{2k\times m_1} = a_1 + b_1 + k.$$

We may now minimize this rank $a_1+b_1+k$ by varying $a_1,\dots,a_{l+1}$ and $b_1,\dots,b_l$. Our claim is then that this optimization over $2l+1$ variables gives the minimum on the right-hand side of (5.9). To show this, we first show that

$$\begin{aligned} &\Bigl\{(a_1,b_1)\Bigm| n_a \ge a_1 \ge \cdots \ge a_{l+1} \ge 0, n_b \ge b_1 \ge \cdots \ge b_l \ge 0, k = \sum_i a_i + \sum_j b_j\Bigr\} \\ &\quad = \bigl\{(a_1,b_1) \mid 0 \le a_1 \le n_a, 0 \le b_1 \le n_b, a_1 + b_1 \le k \le a_1(l+1) + b_1 l\bigr\}. \end{aligned} \tag{5.11}$$

The inclusion "$\subset$" for the two sets in (5.11) is obvious. To prove "$\supset$," let $(a_1,b_1)$ be a point in the set on the right-hand side of (5.11). We then need to argue that there exist $a_2,\dots,a_{l+1},b_2,\dots,b_l$ such that $a_1 \ge a_2 \ge \cdots \ge a_{l+1} \ge 0$, $b_1 \ge b_2 \ge \cdots \ge b_l \ge 0$, and $\sum_{i=2}^{l+1} a_i + \sum_{i=2}^{l} b_i = k - a_1 - b_1$ hold. But this is obvious because

$$\begin{aligned} &\Biggl\{\sum_{i=2}^{l+1} a_i + \sum_{j=2}^{l} b_j \Biggm| a_1 \ge a_2 \ge \cdots \ge a_{l+1} \ge 0, b_1 \ge b_2 \ge \cdots \ge b_l \ge 0\Biggr\} \\ &\quad = \Biggl\{\sum_{i=2}^{l+1} a_i + \sum_{j=2}^{l} b_j \mid 0 \le a_2,\dots,a_{l+1} \le a_1, 0 \le b_2,\dots,b_l \le b_1\Biggr\} \\ &\quad = \{0,1,\dots,a_1 l + b_1(l-1)\} \end{aligned}$$

covers all possible values for $k - a_1 - b_1$.

Finally, the feasible set for the minimization in (5.9) differs from the set on the right-hand side of (5.11) only by the constraint $a_1 + b_1 \le k$. However, a pair $(a_1,b_1)$ cannot be a minimizer for (5.9) if $a_1 + b_1 > k$. Hence, the minimum in (5.9) equals the minimum over the right-hand side of (5.11).

[Lower bound] To obtain a lower bound, it is convenient to rearrange the columns of (5.8) as

$$\begin{pmatrix} X_1 & X_3 & \cdots & X_{2l+1} & 0 & X_2 & X_4 & \cdots & X_{2l} & 0 \\ 0 & X_1 & \cdots & X_{2l-1} & X_{2l+1} & 0 & X_2 & \cdots & X_{2l-2} & X_{2l} \end{pmatrix}_{2k\times m_1}. \tag{5.12}$$

We first find a lower bound of the rank of (5.12) for a fixed matrix $X$, and then obtain a lower bound for all possible $X$.

Let $k_a = \operatorname{rank}(X_1, X_3, \dots, X_{2l+1})$, and $k_b = k - k_a$. There is a $(k-k_a)\times k$ matrix $T$ such that $TX = T(X_1,X_3,\dots,X_{2l+1}) = 0$. Let $S$ be a $k_a \times k$ matrix such that $\binom{S}{T} \in \mathrm{GL}(k,\mathbb{R})$. Multiplying the matrix

$$\begin{pmatrix} S & 0 \\ 0 & S \\ T & 0 \\ 0 & T \end{pmatrix} \in \mathrm{GL}(2k,\mathbb{R})$$

to (5.12) from the left yields

$$\begin{pmatrix} X_1^{(1)} & X_3^{(1)} & \cdots & X_{2l+1}^{(1)} & 0 & * & * & \cdots & * & 0 \\ 0 & X_1^{(1)} & \cdots & X_{2l-1}^{(1)} & X_{2l+1}^{(1)} & 0 & * & \cdots & * & * \\ 0 & 0 & \cdots & 0 & 0 & X_2^{(1)} & X_4^{(1)} & \cdots & X_{2l}^{(1)} & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & X_2^{(1)} & \cdots & X_{2l-2}^{(1)} & X_{2l}^{(1)} \end{pmatrix}_{2k \times m_1},$$

where we define $X_{2i-1}^{(1)} = SX_{2i-1}$ and $X_{2i}^{(1)} = TX_{2i}$. The rank of (5.12) is bounded below by the sum of the ranks of the two matrices:

$$\begin{pmatrix} X_1^{(1)} & X_3^{(1)} & \cdots & X_{2l+1}^{(1)} & 0 \\ 0 & X_1^{(1)} & \cdots & X_{2l-1}^{(1)} & X_{2l+1}^{(1)} \end{pmatrix}_{2k_a \times n_a(l+2)} \tag{5.13}$$

and

$$\begin{pmatrix} X_2^{(1)} & X_4^{(1)} & \cdots & X_{2l}^{(1)} & 0 \\ 0 & X_2^{(1)} & \cdots & X_{2l-2}^{(1)} & X_{2l}^{(1)} \end{pmatrix}_{2k_b \times n_b(l+1)}. \tag{5.14}$$

Note here that

$$\operatorname{rank}(X_1^{(1)}, \ldots, X_{2l+1}^{(1)}) = k_a, \qquad \operatorname{rank}(X_2^{(1)}, \ldots, X_{2l}^{(1)}) = k_b,$$

because

$$\begin{pmatrix} S \\ T \end{pmatrix} (X_1, \ldots, X_{2l+1}, X_2, \ldots, X_{2l}) = \begin{pmatrix} X_1^{(1)} & \cdots & X_{2l+1}^{(1)} & * & \cdots & * \\ 0 & \cdots & 0 & X_2^{(1)} & \cdots & X_{2l}^{(1)} \end{pmatrix}$$

is of row full rank $k = k_a + k_b$.

Let $r_a^{(1)}$ be the rank of (5.13), and let $a_1 = \operatorname{rank}(X_1^{(1)})$. Then we obviously have

$$r_a^{(1)} \geq a_1 + k_a. \tag{5.15}$$

Take $A \in \mathrm{GL}(k_a, \mathbb{R})$ and $P \in \mathrm{GL}(n_a, \mathbb{R})$ such that $AX_1^{(1)}P = \binom{I_{a_1}}{0}$; these are the first $k_a$ rows of $X_1$ in (5.10). Consider then transforming (5.13) by multiplying the $2k_a \times 2k_a$ block-diagonal matrix $\operatorname{diag}(A, A)$ from the left and the $(l+2)n_a \times (l+2)n_a$ block-diagonal matrix $\operatorname{diag}(P, \ldots, P)$ from the right. This transformation preserves rank and turns the matrix in (5.13) into

$$\begin{pmatrix} \begin{pmatrix} I_{a_1} \\ 0 \end{pmatrix} & \widetilde{X}_3 & \widetilde{X}_5 & \cdots & \widetilde{X}_{2l+1} & 0 \\ 0 & \begin{pmatrix} I_{a_1} \\ 0 \end{pmatrix} & \widetilde{X}_3 & \cdots & \widetilde{X}_{2l-1} & \widetilde{X}_{2l+1} \end{pmatrix}_{2k_a \times n_a(l+2)}, \tag{5.16}$$

where $\widetilde{X}_{2i-1} = AX_{2i-1}^{(1)}P$. For each $i$, let $X_{2i-1}^{(2)}$ be the submatrix consisting of the $(a_1+1)$st to $k_a$th row of $\widetilde{X}_{2i-1}$. Note that

$$\operatorname{rank}(X_1^{(1)}, \ldots, X_{2l+1}^{(1)}) = \operatorname{rank} \begin{pmatrix} I_{a_1} & * & \cdots & * \\ 0 & X_3^{(2)} & \cdots & X_{2l+1}^{(2)} \end{pmatrix},$$

and hence

$$\operatorname{rank}(X_3^{(2)}, \ldots, X_{2l+1}^{(2)}) = \operatorname{rank}(X_1^{(1)}, \ldots, X_{2l+1}^{(1)}) - a_1 = k_a - a_1.$$

Deleting the rows with indices between $k_a+1$ and $k_a+a_1$ from (5.16), we have the inequality

$$\begin{aligned} r_a^{(1)} &\geq \operatorname{rank}\begin{pmatrix} \begin{pmatrix} I_{a_1} \\ 0 \end{pmatrix} & \widetilde{X}_3 & \widetilde{X}_5 & \cdots & \widetilde{X}_{2l+1} & 0 \\ 0 & 0 & X_3^{(2)} & \cdots & X_{2l-1}^{(2)} & X_{2l+1}^{(2)} \end{pmatrix}_{(2k_a-a_1)\times n_a(l+2)} \\ &= a_1 + \operatorname{rank}\begin{pmatrix} X_3^{(2)} & X_5^{(2)} & \cdots & X_{2l+1}^{(2)} & 0 \\ 0 & X_3^{(2)} & \cdots & X_{2l-1}^{(2)} & X_{2l+1}^{(2)} \end{pmatrix}_{2(k_a-a_1)\times n_a(l+1)} \\ &=: a_1 + r_a^{(2)}. \end{aligned} \tag{5.17}$$

Repeating the procedure in the preceding paragraph $(i-1)$ times, we obtain the matrices $X_{2i-1}^{(i)}, \ldots, X_{2l+1}^{(i)}$. Let

$$a_i = \operatorname{rank}\bigl(X_{2i-1}^{(i)}\bigr)$$

and

$$r_a^{(i)} = \operatorname{rank}\begin{pmatrix} X_{2i-1}^{(i)} & X_{2i+1}^{(i)} & \cdots & X_{2l+1}^{(i)} & 0 \\ 0 & X_{2i-1}^{(i)} & \cdots & X_{2l-1}^{(i)} & X_{2l+1}^{(i)} \end{pmatrix}_{2(k_a-\sum_{j=1}^{i-1} a_j)\times n_a(l+3-i)}.$$

Noting that

$$\operatorname{rank}\bigl(X_{2i-1}^{(i)}, \ldots, X_{2l+1}^{(i)}\bigr) = \operatorname{rank}\bigl(X_{2i-3}^{(i-1)}, \ldots, X_{2l+1}^{(i-1)}\bigr) - a_{i-1} = k_a - \sum_{j=1}^{i-1} a_j, \tag{5.18}$$

we obtain that the inequalities in (5.15) and (5.17) extend to

$$r_a^{(i)} \geq a_i + \left(k_a - \sum_{j=1}^{i-1} a_j\right), \tag{5.19}$$

$$r_a^{(i)} \geq a_i + r_a^{(i+1)}, \tag{5.20}$$

respectively. Here, we let $r_a^{(l+2)} = 0$.

From (5.19) and (5.20), we find that

$$r_a^{(1)} = \sum_{j=1}^{i-1}\bigl(r_a^{(j)} - r_a^{(j+1)}\bigr) + r_a^{(i)} \geq \sum_{j=1}^{i-1} a_j + a_i + \left(k_a - \sum_{j=1}^{i-1} a_j\right) = a_i + k_a$$

for all $i$. This is equivalent to

$$r_a^{(1)} \geq \max_i a_i + k_a. \tag{5.21}$$

From the construction (5.18), we have

$$k_a = \sum_{i=1}^{l+1} a_i. \tag{5.22}$$

Applying the same arguments, the rank $r_b^{(1)}$ of (5.14) is seen to satisfy the inequality

$$r_b^{(1)} \geq \max_j b_j + k_b, \tag{5.23}$$

where the $b_j = \operatorname{rank}(X_{2j}^{(j)})$ are defined in analogy to the $a_i$ and satisfy the identity

$$k_b = \sum_{j=1}^{l} b_j. \tag{5.24}$$

Combining (5.21) and (5.23), the rank of the matrix (5.12) can be bounded from below as

$$\text{rank of (5.12)} \geq \max_i a_i + \max_j b_j + k. \tag{5.25}$$

The bound in (5.25) is for a given fixed matrix $X$ and in terms of the ranks $a_i$ and $b_j$ the matrix determines. To obtain a lower bound for all possible $X$, we may minimize the right-hand side of (5.25) under the constraints the $a_i$ and $b_j$ should satisfy. These constraints are given by (5.22) and (5.24). We thus minimize over the set

$$\begin{aligned}
&\left\{(\max a_i, \max b_j) \,\middle|\, 0 \leq a_i \leq n_a, 0 \leq b_j \leq n_b, k = \sum_i a_i + \sum_j b_j\right\} \\
&\quad = \{(\max a_i, \max b_j) \mid 0 \leq \min a_i \leq \max a_i \leq n_a, 0 \leq \min b_j \leq \max b_j \leq n_b, \\
&\qquad \max a_i + \max b_j \leq k \leq \max a_i (l+1) + \max b_j l\}.
\end{aligned}$$

As this set is contained in

$$\{(\max a_i, \max b_j) \mid 0 \leq \max a_i \leq n_a, 0 \leq \max b_j \leq n_b, k \leq \max a_i (l+1) + \max b_j l\},$$

we see that $r_2(m_1, m_2, k)$ is bounded from below by the right-hand side of (5.9). $\square$

5.4. *Evaluation of $S_2(m_1, m_2)$.* Our next goal is to evaluate the quantity

$$S_2(m_1, m_2) = \min_{1 \leq k \leq m_2 - 1} \{m_2 r_2(m_1, m_2, k) - m_1 k\} \tag{5.26}$$

whose sign determines the (in)existence of the MLE. As in the previous subsection, we refer to the numbers $l(m_1, m_2)$, $n_a$, and $n_b$ defined in (5.6) and (5.7), respectively. Note first that

$$\frac{m_2}{m_1 - m_2} \text{ is an integer} \quad \Longleftrightarrow \quad n_b = 0.$$

THEOREM 5.3.

(i) *If $m_1 = m_2 + 1$, then $n_a = 1$, $n_b = 0$, and*

$$S_2(m_1, m_2) = 1,$$

*with the minimum in (5.26) attained iff $k = m_2 - 1$.*

(ii) *If $m_1 > m_2 + 1$ and $\frac{m_2}{m_1 - m_2}$ is an integer, then $n_a \geq 2$, $n_b = 0$, and*

$$S_2(m_1, m_2) = 0,$$

*with the minimum in (5.26) attained iff $k$ is an integer multiple of $\frac{m_2}{m_1 - m_2}$.*

(iii) *If $m_1 > m_2 + 1$ and $\frac{m_2}{m_1 - m_2}$ is not an integer, then $n_a \geq 1$, $n_b \geq 1$, and*

$$S_2(m_1, m_2) = -n_a n_b,$$

*with the minimum in (5.26) attained iff $k = (l+1) n_a$.*

PROOF. (i) When $m_1 = m_2 + 1$, we have $l = m_2 - 1$, $n_a = 1$, and $n_b = 0$. It follows from Theorem 5.2 that $r_2(m_1, m_2, k) = k + 1$. We obtain that

$$S_2(m_1, m_2) = \min_{1 \le k \le m_2 - 1} \{m_2(k+1) - (m_2+1)k\} = \min_{1 \le k \le m_2 - 1} (m_2 - k) = 1.$$

(ii) When $m_1 > m_2 + 1$ and $\frac{m_2}{m_1 - m_2}$ is an integer, it holds that $l = \frac{m_2}{m_1 - m_2} - 1$, $n_a = m_1 - m_2$ and $n_b = 0$. Theorem 5.2 yields that

$$r_2(m_1, m_2, k) = k + \left\lceil k \frac{m_1 - m_2}{m_2} \right\rceil.$$

Consequently,

$$S_2(m_1, m_2) = \min_{1 \le k \le m_2 - 1} m_2 \left( \left\lceil k \frac{m_1 - m_2}{m_2} \right\rceil - k \frac{m_1 - m_2}{m_2} \right) \ge 0.$$

The lower bound 0 is attained when $k \frac{m_1 - m_2}{m_2}$ is an integer.

(iii) Finally, consider the case where $\frac{m_2}{m_1 - m_2}$ is not an integer (trivially $m_1 - m_2 > 1$). Then $S_2(m_1, m_2)$ equals the minimum of the function

$$h(k, a, b) = m_2(a + b + k) - k m_1 = m_2(a + b) - (m_1 - m_2)k \tag{5.27}$$

over the set

$$\{(k, a, b) \mid 1 \le k \le m_2 - 1, k \le a(l+1) + bl, 0 \le a \le n_a, 0 \le b \le n_b\}.$$

We distinguish two cases of how the minimum may be attained, namely, case 1 with $m_2 - 1 \ge a(l+1) + bl$, and case 2 with $m_2 - 1 < a(l+1) + bl$. Accordingly,

$$S_2(m_1, m_2) = \min\{S_2'(m_1, m_2), S_2''(m_1, m_2)\},$$

where $S_2'(m_1, m_2)$ and $S_2''(m_1, m_2)$ are the minima of $h(k, a, b)$ from (5.27) over the sets

$$\{(k, a, b) \mid 1 \le k \le a(l+1) + bl \le m_2 - 1, 0 \le a \le n_a, 0 \le b \le n_b\}$$

and

$$\{(k, a, b) \mid 1 \le k \le m_2 - 1 < a(l+1) + bl, 0 \le a \le n_a, 0 \le b \le n_b\}, \tag{5.28}$$

respectively.

Case 1: The minimum $S_2'(m_1, m_2)$ is attained iff $k = a(l+1) + bl$, in which case

$$h(k, a, b) = m_2(a + b) - (m_1 - m_2)\{a(l+1) + bl\} = -n_b a + n_a b.$$

Therefore,

$$S_2'(m_1, m_2) = \min\{-n_b a + n_a b \mid a(l+1) + bl \le m_2 - 1, 0 \le a \le n_a, 0 \le b \le n_b\}.$$

This minimum is achieved by taking $b$ as small as possible, so $b = 0$, and $a$ is large as possible. Indeed, $(a, b) = (n_a, 0)$ is feasible as

$$(m_2 - 1) - [n_a(l+1) + 0 \cdot l] = l n_b - 1 \ge 0.$$

We conclude that $S_2'(m_1, m_2) = -n_a n_b$, with the minimum attained at $k = n_a(l+1)$.

Case 2: Because

$$m_2 - 1 < a(l+1) + bl \le n_a(l+1) + n_b l = m_2,$$

the set (5.28) is

$$\{(k, a, b) \mid 1 \le k \le m_2 - 1, a = n_a, b = n_b\}.$$

The minimum $S_2''(m_1, m_2)$ is attained iff $k = m_2 - 1$, in which case

$$h(k, a, b) = m_2(n_a + n_b) - (m_1 - m_2)(m_2 - 1) = m_1 - m_2 > 0.$$

In summary, $S_2(m_1, m_2) = \min\{S_2'(m_1, m_2), S_2''(m_1, m_2)\} = -n_a n_b$, and this minimum is attained iff $k = n_a(l+1) + 0 = n_a(l+1)$. $\square$

TABLE 2
$S_2(m_1, m_2)$ $(m_1 \geq m_2)$

| $m_1 \setminus m_2$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 0/2 | | | | | | | | | | | | | | |
| 3 | 1 | 0 | | | | | | | | | | | | | |
| 4 | 0 | 1 | 0 | | | | | | | | | | | | |
| 5 | | –1 | 1 | 0 | | | | | | | | | | | |
| 6 | | 0 | 0 | 1 | 0 | | | | | | | | | | |
| 7 | | | –2 | –1 | 1 | 0 | | | | | | | | | |
| 8 | | | 0 | –2 | 0 | 1 | 0 | | | | | | | | |
| 9 | | | | –3 | 0 | –1 | 1 | 0 | | | | | | | |
| 10 | | | | 0 | –4 | –2 | 0 | 1 | 0 | | | | | | |
| 11 | | | | | –4 | –3 | –2 | –1 | 1 | 0 | | | | | |
| 12 | | | | | 0 | –6 | 0 | 0 | 0 | 1 | 0 | | | | |
| 13 | | | | | | –5 | –6 | –3 | –2 | –1 | 1 | 0 | | | |
| 14 | | | | | | 0 | –8 | –4 | –4 | –2 | 0 | 1 | 0 | | |
| 15 | | | | | | | –6 | –9 | 0 | –3 | 0 | –1 | 1 | 0 | |
| 16 | | | | | | | 0 | –10 | –8 | –4 | 0 | –2 | 0 | 1 | 0 |
| 17 | | | | | | | | –7 | –12 | –5 | –6 | –3 | –2 | –1 | 1 |

“ ”: MLE exists uniquely. “ ”: MLE exists nonuniquely. “ ”: MLE does not exist.

REMARK 5.2. When $S_2(m_1, m_2) = 0$, neither $m_1$ nor $m_2$ is a prime number.

Including the square case, the possible values of $S_2(m_1, m_2)$ may be summarized as follows. The values are tabulated up to $m_1 \leq 17$ in Table 2.

PROPOSITION 5.4. *For $n = 2$ generic data matrices $Y_1, Y_2 \in \mathbb{R}^{m_1 \times m_2}$ with $m_2 \leq m_1 < 2m_2$,*

$$S_2(m_1, m_2) = \begin{cases} 0 & (m_1 = m_2 \geq 3 \textit{ or} \\ & m_1 = m_2 = 2 \textit{ and } Y_1^{-1} Y_2 \textit{ has real eigenvalues}), \\ 2 & (m_1 = m_2 = 2 \textit{ and } Y_1^{-1} Y_2 \textit{ has complex eigenvalues}), \\ 1 & (m_1 = m_2 + 1), \\ 0 & (m_1 > m_2 + 1, m_1 - m_2 | m_2), \\ -n_a n_b & (m_1 - m_2 \nmid m_2). \end{cases}$$

We recall that the MLE exists uniquely if $S_2(m_1, m_2) > 0$, exists nonuniquely if $S_2(m_1, m_2) = 0$, and does not exist if $S_2(m_1, m_2) < 0$.

## 6. When the column size of matrices is 2.

When the matrix size of $Y_i$ is $m_1 \times 2$, that is, $m_2 = 2$, we can find $S_n(m_1, 2)$ as a byproduct of the case $n = 2$ discussed in Sections 4 and 5. Indeed, from the definitions (3.11) and (3.10),

$$S_n(m_1, 2) = \min_{1 \leq k < 2} \{2 r_n(m_1, 2, k) - m_1 k\} = 2 r_n(m_1, 2, 1) - m_1,$$

and

$$\begin{aligned} r_n(m_1, 2, 1) &= \min_{X \in \mathbb{R}^{2 \times 1} : \operatorname{rank}(X) = 1} \operatorname{rank}(Y_1 X, \ldots, Y_n X) \\ &= \min_{(x_1, x_2) \in \mathbb{R}^2 \setminus \{0\}} \operatorname{rank}(x_1 Y_{(1)} + x_2 Y_{(2)}), \end{aligned}$$

where $Y_{(j)} = (y_{1j}, \dots, y_{nj})$, $j = 1, 2$, with $y_{ij}$ the $j$th column vector of $Y_i$. The matrices $Y_{(j)}$ are $m_1 \times n$ generic matrices. We examine the cases (i) $m_1 = n$ and (ii) $m_1 \neq n$ separately.

(i) Let $m = m_1 = n$. Since the $Y_{(j)}$ are nonsingular, we may define $W = Y_{(1)}^{-1} Y_{(2)}$. Then $r_n(m_1, 2, 1) = \min_{x \in \mathbb{R}} \operatorname{rank}(x I_m + W) = m - 1$ if $W$ has a real eigenvalue, and $r_n(m_1, 2, 1) = m$ otherwise. Therefore, $S_n(m, 2) = 2r_m(m, 2, 1) - m = m - 2$ if $W$ has a real eigenvalue, and $S_n(m, 2) = m$ otherwise.

(ii) Suppose that $m_1 > n$. Then, by multiplying a suitable $(A, C) \in \mathrm{GL}(m_1) \times \mathrm{GL}(n)$ from left and right to get the Kronecker canonical form, we have

$$r_n(m_1, 2, 1) = \min_{(x_1, x_2) \neq 0} \operatorname{rank}\left( x_1 \begin{pmatrix} I_n \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ I_n \end{pmatrix} \right) = n. \tag{6.1}$$

Hence, $S_n(m_1, 2) = 2n - m_1$. Similarly, when $m_1 < n$, we have $r_n(m_1, 2, 1) = m_1$ and $S_n(m_1, 2) = m_1$. Recall that we are examining the region $m_2 \le m_1 < nm_2$, or equivalently $1 \le m_1/2 < n$ when $m_2 = 2$.

PROPOSITION 6.1. *For generic data matrices $Y_1, \dots, Y_n \in \mathbb{R}^{m_1 \times 2}$ with $1 \le m_1/2 < n$,*

$$S_n(m_1, 2) = \begin{cases} 2n - m_1 & (n < m_1), \\ m_1 & (m_1 < n), \\ m & (m_1 = n = m, \textit{W does not have real eigenvalues}), \\ m - 2 & (m_1 = n = m, \textit{W has a real eigenvalue}). \end{cases}$$

*That is*, (*i*) *$S_n(m_1, 2) > 0$ when $m_1 \neq n$, or when $m_1 = n > 2$, or when $m_1 = n = 2$ and $W$ does not have real eigenvalues*; (*ii*) *$S_n(m_1, 2) = 0$ when $m_1 = n = 2$ and $W$ has a real eigenvalue.*

REMARK 6.1. The transformation from $Y_i$'s to the Kronecker form in (6.1) is written as $(Y_1, \dots, Y_n) \mapsto A(Y_1, \dots, Y_n)(C \otimes I_{m_2})$, where $C$ is an $n \times n$ nonsingular matrix. Here, we used the fact that $\operatorname{rank}(Y_1, \dots, Y_n)(C \otimes I_n)$ is invariant as long as $C \in \mathrm{GL}(n)$. Combined with the group action discussed in Section 2.3, the group action

$$(Y_1, \dots, Y_n) \mapsto A(Y_1, \dots, Y_n)(C \otimes B), \qquad (A, B, C) \in \mathrm{GL}(m_1) \times \mathrm{GL}(m_2) \times \mathrm{GL}(n)$$

keeps the values $r_n(m_1, m_2, k)$, and hence also $S_n(m_1, m_2)$, invariant.

**7. Maximum likelihood estimation for two data matrices.** In this section, we derive the precise form of maximizers of the likelihood function for $n = 2$ rectangular data matrices of size $m_1 \times m_2$ with $2m_2 \ge m_1 > m_2$. We first give the MLE when it exists uniquely. This then allows us to show existence of maximizers in the cases where the study of ranks in Section 5 implies that the likelihood function is bounded.

7.1. *Closed-form MLEs for $m_1 = m_2 + 1$.* Consider the case where $m_1 = m_2 + 1$ and $n = 2$, so that a unique MLE exists almost surely. For simpler notation, let $m := m_2$. By Theorem 5.1, we may assume that the two data matrices are

$$Y_1 = \begin{pmatrix} I_m \\ 0_{1,m} \end{pmatrix}, \qquad Y_2 = \begin{pmatrix} 0_{1,m} \\ I_m \end{pmatrix}. \tag{7.1}$$

Then the negated profile log-likelihood function takes the form

$$g_0(\Phi) = m \log\det\left( \begin{pmatrix} \Phi & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Phi \end{pmatrix} \right) - (m + 1) \log\det(\Phi).$$

PROPOSITION 7.1. *When restricted to matrices $\Phi = (\phi_{jk})$ with $\phi_{11} = 1$, the function $g_0$ is uniquely minimized by the diagonal matrix*

$$\Phi_0 = \operatorname{diag}\left(\binom{m-1}{j-1} : j = 1, \dots, m\right).$$

PROOF. The existence of a unique minimizer is clear from Theorem 5.3. It thus suffices to show that $\Phi_0$ is a critical point of $g_0$.

The logarithm of the determinant has differential

$$\mathrm{d} \log \det(\Phi) = \operatorname{tr}(\Phi^{-1} \mathrm{d}\Phi).$$

It follows that the differential of $g_0$ is

$$\begin{aligned} &\mathrm{d}g_0(\Phi; U) \\ &\quad = m \operatorname{tr}\left\{\left[\begin{pmatrix} \Phi & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Phi \end{pmatrix}\right]^{-1} \left[\begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & U \end{pmatrix}\right]\right\} - (m+1) \operatorname{tr}(\Phi^{-1} U). \end{aligned}$$

Since our candidate $\Phi_0$ is diagonal,

$$\mathrm{d}g_0(\Phi_0; U) = m\left(\frac{u_{11}}{\phi_1} + \frac{u_{mm}}{\phi_m} + \sum_{j=1}^{m-1} \frac{u_{jj} + u_{j+1,j+1}}{\phi_j + \phi_{j+1}}\right) - (m+1) \sum_{j=1}^{m} \frac{u_{jj}}{\phi_j},$$

where $\phi_j$ is the $j$th diagonal entry of $\Phi_0$, and $U = (u_{jk})$. The differential $\mathrm{d}g_0(\Phi_0; U)$ is zero if

$$\frac{m+1}{\phi_1} = m\left(\frac{1}{\phi_1} + \frac{1}{\phi_1 + \phi_2}\right), \tag{7.2}$$

$$\frac{m+1}{\phi_j} = m\left(\frac{1}{\phi_j + \phi_{j-1}} + \frac{1}{\phi_j + \phi_{j+1}}\right), \quad j = 2, \dots, m-1, \tag{7.3}$$

$$\frac{m+1}{\phi_m} = m\left(\frac{1}{\phi_m} + \frac{1}{\phi_m + \phi_{m-1}}\right). \tag{7.4}$$

It is easy to see that the first equation, that in (7.2), holds for our choice of $\Phi_0$. Indeed, after clearing denominators, the equation becomes

$$(m+1)(\phi_1 + \phi_2) = m(2\phi_1 + \phi_2) \quad \iff \quad \phi_2 = (m-1)\phi_1,$$

which clearly holds when $\phi_1 = \binom{m-1}{0} = 1$ and $\phi_2 = \binom{m-1}{1} = m-1$. The last equation in (7.4) holds similarly as we have $\phi_m = \binom{m-1}{m-1} = 1$ and $\phi_{m-1} = \binom{m-1}{m-2} = m-1$.

Let $2 \le j \le m-1$. Solving the $j$th equation in (7.3) for $\phi_{j+1}$, we obtain

$$\phi_{j+1} = \phi_j \cdot \frac{(m-1)\phi_j - \phi_{j-1}}{\phi_j + (m+1)\phi_{j-1}}. \tag{7.5}$$

Using that

$$\binom{m-1}{j-1} = \frac{m-j+1}{j-1} \binom{m-1}{j-2},$$

we derive that

$$\frac{(m-1)\phi_j - \phi_{j-1}}{\phi_j + (m+1)\phi_{j-1}} = \frac{(m-1)\binom{m-1}{j-1} - \binom{m-1}{j-2}}{\binom{m-1}{j-1} + (m+1)\binom{m-1}{j-2}} = \frac{(m-1)\frac{m-j+1}{j-1} - 1}{\frac{m-j+1}{j-1} + (m+1)} = \frac{m-j}{j}.$$

We see that (7.5) holds because

$$\phi_{j+1} = \binom{m-1}{j} = \frac{m-j}{j}\binom{m-1}{j-1} = \frac{m-j}{j}\phi_j.$$

We have thus shown that our choice of $\Phi_0$ satisfies $\mathrm{d}g_0(\Phi_0; U) \equiv 0$. $\square$

Simple calculations yield that, at the critical point,

$$g_0(\Phi_0) = m \log d(m) - (m+1)\log e(m), \tag{7.6}$$

where

$$d(m) = \det\left(\sum_{i=1}^{2} Y_i \Phi_0^{-1} Y_i^T\right) = \left(\frac{m}{m-1}\right)^{m-1} \frac{1}{\prod_{j=1}^{m-1}\binom{m-2}{j-1}}$$

and

$$e(m) = \det \Psi(m) = \frac{1}{\prod_{j=1}^{m}\binom{m-1}{j-1}}.$$

7.2. *Critical points when MLEs exist nonuniquely.* Nonunique existence corresponds to case (ii) in Theorem 5.3. So, $n_b = 0$ and $n_a = m_1 - m_2 \geq 2$. Moreover, $m_1 = (l+2)n_a$ and $m_2 = (l+1)n_a$. Assume, as before, that the two $m_1 \times m_2$ data matrices are

$$Y_1 = \begin{pmatrix} I_{m_2} \\ 0_{m_1-m_2,m_2} \end{pmatrix}, \qquad Y_2 = \begin{pmatrix} 0_{m_1-m_2,m_2} \\ I_{m_2} \end{pmatrix}. \tag{7.7}$$

As seen in Remark 5.1, by permuting the rows and the columns of $Y_1$ and $Y_2$ simultaneously, we may transform $Y_1$ and $Y_2$ into

$$Y_1 = \begin{pmatrix} U_{l+1} & & 0 \\ & \ddots & \\ 0 & & U_{l+1} \end{pmatrix}, \qquad U_{l+1} = \begin{pmatrix} I_{l+1} \\ 0_{1,l+1} \end{pmatrix} \in \mathbb{R}^{(l+2)\times(l+1)}$$

and

$$Y_2 = \begin{pmatrix} L_{l+1} & & 0 \\ & \ddots & \\ 0 & & L_{l+1} \end{pmatrix}, \qquad L_{l+1} = \begin{pmatrix} 0_{1,l+1} \\ I_{l+1} \end{pmatrix} \in \mathbb{R}^{(l+2)\times(l+1)}.$$

Let $\Phi = \mathrm{diag}(\Phi_1, \dots, \Phi_{n_a})$ with $\Phi_j \in \mathbb{R}^{(l+1)\times(l+1)}$. Then the negated profile log-likelihood function takes the form

$$g_0(\Phi) = \sum_{j=1}^{n_a}\left[m_2 \log\det\left(\begin{pmatrix} \Phi_j & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Phi_j \end{pmatrix}\right) - m_1 \log\det(\Phi_j)\right].$$

Applying Proposition 7.1 to each summand we may determine critical points $\Phi_{0j}$ for each summand in $g_0(\Phi)$. These may then be combined to obtain critical points of $g_0$.

PROPOSITION 7.2. *When restricted to matrices $\Phi = (\phi_{jk})$ with $\phi_{11} = 1$, the function $g_0$ has a critical point at every diagonal matrix $\Phi_0 = (\phi_{jk}^0)$ whose diagonal entries are*

$$\phi^0_{(j-1)n_a+k,(j-1)n_a+k} = c_j \binom{l}{k-1}, \quad j = 1, \dots, n_a, k = 1, \dots, l+1,$$

*where $c_1 = 1$ and $c_j > 0$ arbitrary for $j = 2, \dots, n_a$.*

Note that for the critical points $\Phi_0$ from Proposition 7.2 it holds that

$$g_0(\Phi_0) = \sum_{j=1}^{n_a} [m_2 \log c_j^{l+2} d(l+1) - m_1 \log c_j^{l+1} e(l+1)]$$
$$= \sum_{j=1}^{n_a} [m_2 \log d(l+1) - m_1 \log e(l+1)],$$

which is independent of $c_j$'s. This is a confirmation of the fact that all of the critical points define minima of $g_0$.

**8. Conclusion.** In this paper, we considered uniqueness and existence of the maximum likelihood estimator in the matrix normal model. In other words, we considered Gaussian models for i.i.d. matrix-valued observations $Y_1, \dots, Y_n$ that posit a Kronecker product for the joint covariance matrix of the entries of each random matrix $Y_i$. Our goal was to give precise formulas for maximum likelihood thresholds, which are defined to be the sample sizes that are minimally needed for almost sure existence of the MLE, unique existence of the MLE, or possibly mere boundedness of the likelihood function. Our main result solves this problem for data matrices whose dimensions $m_1$ and $m_2$ differ at most by a factor of two. Our solution exhibits subtle dependencies on $m_1$ and $m_2$. From a statistical perspective, our work clarifies that very small sample sizes are sufficient to make matrix normal models amenable to likelihood inference.

As observed in the Introduction, prior work of [23] can be used to determine the maximum likelihood thresholds for settings where one matrix dimension is sufficiently large compared to the other or, more precisely put, where dividing one matrix dimension by the other leaves a sufficiently small remainder. In intermediate settings, however, the maximum likelihood thresholds remain unknown. Although good bounds exist, it would be of obvious interest to determine the thresholds in full generality. Here, it should be noted that our solution for the setting $2m_2 \geq m_1 \geq m_2$ crucially relies on invariance properties that allowed us to exploit the Kroncker canonical form for matrix pencils. For larger sample sizes, new additional ideas are needed as a similarly simple canonical form does not exist [17], Chapter 10.

In all cases covered by our results almost sure boundedness of the likelihood function implies almost sure existence of a maximizer. We conjecture this to be true in general. This said, there do exist individual data sets for which the likelihood function is bounded but does not achieve its maximum; recall Proposition 4.7.

Finally, we would like to note that since our paper was submitted a new preprint was posted on arXiv, which makes a connection between maximum likelihood estimation and computational invariant theory for a series of problems including the matrix normal models considered here [3]. A further preprint appeared later announcing that a full solution to the (unique) existence problem can be achieved via quiver representation theory [7].

## APPENDIX: SOME LEMMAS AND PROOFS

PROOF OF LEMMA 1.2. The $m_1 \times m_1$ matrix $\sum_{i=1}^n Y_i Y_i^T$ is positive semidefinite of rank at most $nm_2 < m_1$. By spectral decomposition,

$$\sum_{i=1}^n Y_i Y_i^T = Q^T D Q,$$

where $Q = Q(Y)$ is $m_1 \times m_1$ orthogonal, and $D$ is diagonal with $D_{m_1 m_1} = 0$. Let $I_m$ be the $m \times m$ identity matrix, and let $e_m = (0, \dots, 0, 1)^T$ the $m$th canonical basis vector of $\mathbb{R}^m$.

Define $\Psi_1^{(t)} = Q(I_{m_1} + t \cdot e_{m_1} e_{m_1}^T) Q^T \in \mathrm{PD}(m_1)$. Then

$$\begin{aligned}\ell(\Psi_1^{(t)}, I_{m_2}) &= nm_2 \log\det(I_{m_1} + t \cdot e_{m_2} e_{m_2}^T) - \operatorname{tr}[(I_{m_1} + t \cdot e_{m_2} e_{m_2}^T) D] \\ &= nm_2 \log(1+t) - \operatorname{tr}(D)\end{aligned}$$

tends to $\infty$ as $t \to \infty$. □

PROOF OF THEOREM 1.4. We are assuming that $m_1/m_2$ is an integer. Take $n = m_1/m_2$. Then $Y = (Y_1, \dots, Y_n)$ is $n \times n$, and the profile log-likelihood function $g(\Psi)$ in (2.3) is easily seen to be constant. Indeed,

$$g(\Psi) = m_2 \log\det(Y)^2 + m_2 \log\det(\Psi)^n - m_1 \log\det(\Psi) = 2m_2 \log|\det(Y)|.$$

Hence, the function is maximized by any matrix $\Psi \in \mathrm{PD}(m_2)$, and we obtain that $N_b(m_1, m_2) = N_e(m_1, m_2) \le m_1/m_2$ and $N_u(m_1, m_2) > m_1/m_2$. The lower bound in Proposition 1.3 now implies that $N_b(m_1, m_2) = N_e(m_1, m_2) = m_1/m_2$.

When $m_1 = m_2$, Corollary 4.6 gives $N_u(m_1, m_2) = 3$. When $m_1 > m_2$, Proposition 1.3 gives the upper bound $N_u(m_1, m_2) \le m_1/m_2 + 1$, which implies $N_u(m_1, m_2) = m_1/m_2 + 1$. □

PROOF OF COROLLARY 1.5. Let $m_1 = hm_2 + r$ with quotient $h = \lfloor m_1/m_2 \rfloor \ge 1$ and remainder $r = m_1 \bmod m_2$. Applying Proposition 1.3, we have

$$h + 1 \le N_b(m_1, m_2) \le N_e(m_1, m_2) \le N_u(m_1, m_2).$$

The upper bound from Proposition 1.3 implies that all thresholds are equal to $h + 1$ if

$$\frac{m_1}{m_2} + \frac{m_2}{m_1} < h + 1 \quad \iff \quad m_1^2 + m_2^2 < (h+1) m_1 m_2.$$

Substituting $m_1 = hm_2 + r$ and simplifying, this condition is equivalent to

$$m_2(m_2 - r)h - (m_2^2 - m_2 r + r^2) > 0,$$

and so equivalent to the claimed inequality

$$h > \frac{m_2^2 - m_2 r + r^2}{m_2(m_2 - r)}.$$

The right-hand side of the inequality just given is increasing in the remainder $r$. Thus, for fixed $m_2$, it never exceeds

$$\frac{m_2^2 - m_2(m_2 - 1) + (m_2 - 1)^2}{m_2(m_2 - m_2 + 1)} = m_2 - \frac{m_2 - 1}{m_2} < m_2.$$

Hence, $h \ge m_2$ is sufficient for all thresholds being equal to $h + 1$. □

PROOF OF THEOREM 1.6. Proposition 4.2 gives $N_e(m_1, m_2) = N_b(m_1, m_2)$ for $m_1 = m_2$. Corollary 4.6 gives $N_u(m_1, m_2)$ for $m_1 = m_2$. Theorem 1.4 yields $N_e(m_1, m_2) = N_b(m_1, m_2)$ and $N_u(m_1, m_2)$ when $2m_2 = m_1$. When $2m_2 > m_1 > m_2$, by Theorem 5.2, we have that $N_u(m_1, m_2) = 2$ if $m_1 = m_2 + 1$, and $N_u(m_1, m_2) > 2$ otherwise, and that $N_b(m_1, m_2) = N_e(m_1, m_2) = 2$ if $m_1 - m_2 | m_2$, and $N_u(m_1, m_2) > 2$ otherwise. On the other hand, by Proposition 1.3, $N_b(m_1, m_2) \le N_e(m_1, m_2) \le N_u(m_1, m_2) \le \lfloor m_1/m_2 + m_2/m_1 \rfloor + 1 \le 3$. □

LEMMA A.1. *Let $x_i^{(t)} > 0$, $1 \le i \le m$, $t = 1, 2, \ldots$, be positive sequences. After suitable relabeling of the indices $i$ of $x_i^{(t)}$, we can take a subsequence of the form*

$$
\begin{pmatrix} x_1^{(t)} \\ \vdots \\ x_{r_1}^{(t)} \\ x_{r_1+1}^{(t)} \\ \vdots \\ x_{r_1+r_2}^{(t)} \\ \vdots \\ x_{m-r_K+1}^{(t)} \\ \vdots \\ x_m^{(t)} \end{pmatrix} = \epsilon_1^{(t)} \begin{pmatrix} y_{11}^{(t)} \\ \vdots \\ y_{1r_1}^{(t)} \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \epsilon_2^{(t)} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y_{21}^{(t)} \\ \vdots \\ y_{2r_2}^{(t)} \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \epsilon_K^{(t)} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ y_{K1}^{(t)} \\ \vdots \\ y_{Kr_K}^{(t)} \end{pmatrix}, \tag{A.1}
$$

*where $r_1 + \cdots + r_K = m$, each sequence $y_{ij}^{(t)}$ converges to a limit $y_{ij}^0 > 0$, and $\epsilon_{i+1}^{(t)}/\epsilon_i^{(t)} \to 0$, $1 \le i \le K-1$, as $t \to \infty$.*

PROOF. When comparing two sequences $x_i^{(t)}$ and $x_j^{(t)}$, at least one of the two statements "$x_i^{(t)} \ge x_j^{(t)}$ for infinitely many $t$" or "$x_i^{(t)} \le x_j^{(t)}$ for infinitely many $t$" holds. By relabeling the indices, we can assume that $x_1^{(t)} \ge x_i^{(t)}$ $(1 < i \le m)$ for infinitely many $t$. Take subsequences so that $x_i^{(t)}/x_1^{(t)} \in (0, 1]$ have finite limits for all $i$. Suppose that the limits for $i = 2, \ldots, r_1$ are positive, and the others are zero. Then we set $\epsilon_1^{(t)} = x_1^{(t)}$, $y_{1i}^{(t)} = x_i^{(t)}/x_1^{(t)}$, $i = 1, \ldots, r_1$.

Next, we apply the same procedure to $x_{r_1+1}^{(t)}, \ldots, x_m^{(t)}$. Suppose that $x_{r_1+1}^{(t)} \ge x_i^{(t)}$ $(r_1 + 1 < i \le m)$ for infinitely many $t$. Take subsequences so that $x_i^{(t)}/x_{r_1+1}^{(t)} \in (0, 1]$ have finite limits for all $r_1 + 1 < i \le m$ again, and classify the limits by their signs (i.e., positive or zero). Suppose that the limits for $i = r_1 + 2, \ldots, r_1 + r_2$ are positive, and the others are zero. Then we take $\epsilon_2^{(t)} = x_{r_1+1}^{(t)}$ and $y_{2i}^{(t)} = x_{r_1+i}^{(t)}/x_{r_1+1}^{(t)}$, $i = r_1 + 1, \ldots, r_1 + r_2$. Note that $\epsilon_2^{(t)}/\epsilon_1^{(t)} = x_{r_1+1}^{(t)}/x_1^{(t)} \to 0$.

By repeating this procedure, we get the form (A.1). □

LEMMA A.2. *Let $Y_1$ and $Y_2$ be two independent random $2 \times 2$ matrices whose entries are i.i.d. $\mathcal{N}(0, 1)$. Then the matrix $Y_1^{-1} Y_2$ has real eigenvalues with probability $\pi/4 \approx 0.7854$.*

PROOF. We first obtain the distribution of $Z = Y_1^{-1} Y_2$, which we do for general matrix size $m$. We use $|A|$ as a shorthand for the determinant of a matrix $A$. Then noting that $Y_2 = Y_1 Z$ and $dY_2 = |Y_1|^m \, dZ$, the joint density of $(Y_1, Y_2)$ is

$$
\begin{aligned}
&\frac{1}{(2\pi)^{m^2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\bigl(Y_1 Y_1^T + Y_2 Y_2^T\bigr)\right\} dY_1 \, dY_2 \\
&= \frac{1}{(2\pi)^{m^2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\bigl(Y_1 Y_1^T + Y_1 Z Z^T Y_1^T\bigr)\right\} dY_1 |Y_1|^m \, dZ \\
&= \frac{1}{(2\pi)^{m^2}} \exp\left\{-\frac{1}{2} \operatorname{tr}\bigl(Y_1^T Y_1 (I + Z Z^T)\bigr)\right\} |Y_1|^m \, dY_1 \, dZ.
\end{aligned}
$$

Let $Y_1 = HT$ be the QR decomposition. That is, $H \in O(m)$ and $T = (t_{ij})$ is an upper-triangular matrix with $t_{ii} > 0$. Let $S = Y_1^T Y_1 = T^T T$. Then the mapping $Y_1 \mapsto (S, H)$ is one to one, and according to Theorem 2.1.14 in [19], page 66, its Jacobian is

$$dY_1 = 2^{-m}|S|^{-1/2}\,dS(dH), \qquad (dH) = \bigwedge_{1\le i<j\le m} h_j^T\,dh_i,$$

where $h_i$ is the $i$th column of $H$. By Theorems 2.1.12 and 2.1.15 in [19],

$$\int_{O(m)} (dH) = \frac{2^m \pi^{m^2/2}}{\Gamma_m(\frac{m}{2})}, \qquad \Gamma_m(a) = \pi^{m(m-1)/4}\prod_{i=1}^m \Gamma\biggl(\frac{2a-i+1}{2}\biggr),$$

and we have the joint density of $(Z, S)$ as

$$\frac{\pi^{m^2/2}}{(2\pi)^{m^2}\Gamma_m(\frac{m}{2})}\exp\biggl\{-\frac{1}{2}\operatorname{tr} S(I+ZZ^T)\biggr\}|S|^{(m-1)/2}\,dZ\,dS.$$

Moreover, by letting $n = 2m$ in the Wishart integral

$$\int \exp\biggl\{-\frac{1}{2}\operatorname{tr} S\Sigma^{-1}\biggr\}|S|^{(n-m-1)/2}\,dS = 2^{mn/2}\Gamma_m\biggl(\frac{n}{2}\biggr)|\Sigma|^{n/2},$$

we have

$$\int \exp\biggl\{-\frac{1}{2}\operatorname{tr} S\Sigma^{-1}\biggr\}|S|^{(m-1)/2}\,dS = 2^{m^2}\Gamma_m(m)|\Sigma|^m.$$

Hence, the marginal of $Z$ is

$$\frac{\pi^{m^2/2}2^{m^2}\Gamma_m(m)}{(2\pi)^{m^2}\Gamma_m(\frac{m}{2})}|I+ZZ^T|^{-m}\,dZ = \frac{\prod_{i=1}^m \Gamma(\frac{2m-i+1}{2})}{\pi^{m^2/2}\prod_{i=1}^m \Gamma(\frac{m-i+1}{2})}|I+ZZ^T|^{-m}\,dZ.$$

We now restrict our attentions to the case $m = 2$. The density of $Z$ is then

$$\frac{1}{\pi^2}\frac{\Gamma(\frac{4}{2})\Gamma(\frac{3}{2})}{\Gamma(\frac{2}{2})\Gamma(\frac{1}{2})}|I+ZZ^T|^{-2}\,dZ = \frac{1}{2\pi^2}|I+ZZ^T|^{-2}\,dZ.$$

Suppose that $Z$ has real eigenvalues. For such $Z$, we have the decomposition $Z = PLP^{-1}$, where $L = \operatorname{diag}(l_1, l_2)$, $l_1 > l_2$ and $P = (p_{ij})_{2\times 2}$ is a nonsingular matrix. Without loss of generality, we may assume that the eigenvectors $\binom{p_{1i}}{p_{2i}}$ are unit vectors with $p_{1i} > 0$. Then the map $Z \mapsto (L, P)$ is one to one. However, it will be convenient to also allow $p_{i1} < 0$ in the below calculations. In this parameterization, $Z \mapsto (L, P)$ is 1 to $2^2$.

Write

$$\begin{pmatrix} p_{1i} \\ p_{2i} \end{pmatrix} = \begin{pmatrix} \cos\theta_i \\ \sin\theta_i \end{pmatrix}, \qquad i = 1, 2.$$

The Jacobian of $Z \mapsto (L, P)$ is

$$dZ = \frac{(l_1 - l_2)^2}{\sin(\theta_1 - \theta_2)^2}\,dL\,d\theta_1\,d\theta_2.$$

Integrating we find that

$$\begin{aligned} &\frac{1}{2\pi^2}|I+ZZ^T|^{-2}\,dZ \\ &\quad = \frac{1}{2\pi^2}\frac{\sin(\theta_1-\theta_2)^2}{\{(1+l_1^2)(1+l_2^2) - (1+l_1 l_2)^2\cos(\theta_1-\theta_2)^2\}^2}\frac{(l_1-l_2)^2}{\sin(\theta_1-\theta_2)^2}\,dL\,d\theta_1\,d\theta_2 \end{aligned}$$

over $\theta_1, \theta_2 \in [0, 2\pi)$. Dividing by $2^2$, we obtain the density (not probability density) of $(l_1, l_2)$ as

$$\frac{1}{4}\frac{l_1 - l_2}{(1+l_1^2)^{3/2}(1+l_2^2)^{3/2}}\,dL.$$

Taking the integral over $-\infty < l_2 < l_1 < \infty$,

$$\int_{-\infty<l_2<l_1<\infty} \frac{1}{4}\frac{l_1 - l_2}{(1+l_1^2)^{3/2}(1+l_2^2)^{3/2}}\,dl_1\,dl_2 = \frac{\pi}{4}.$$

This is the integral over the space where $Z$ has real eigenvalues, and our proof is complete. □

**Acknowledgments.** We are grateful to Satoru Iwata, Lek-Heng Lim and Fumihiro Sato for their comments on the Kronecker canonical form.

**Funding.** SK was partially supported by JSPS KAKENHI Grant Number JP16H02792.

## REFERENCES

[1] ALLEN, G. I. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Stat.* **4** 764–790. MR2758420 https://doi.org/10.1214/09-AOAS314

[2] ALLEN, G. I. and TIBSHIRANI, R. (2012). Inference with transposable data: Modelling the effects of row and column correlations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 721–743. MR2965957 https://doi.org/10.1111/j.1467-9868.2011.01027.x

[3] AMÉNDOLA, C., KOHN, K., REICHENBACH, P. and SEIGAL, A. (2020). Invariant theory and scaling algorithms for maximum likelihood estimation.

[4] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1990662

[5] CHEN, X. and LIU, W. (2019). Graph estimation for matrix-variate Gaussian data. *Statist. Sinica* **29** 479–504. MR3889377

[6] DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963 https://doi.org/10.1093/biomet/68.1.265

[7] DERKSEN, H. and MAKAM, V. (2020). Maximum likelihood estimation for matrix normal models via quiver representations.

[8] DRTON, M., FOX, C., KÄUFL, A. and POULIOT, G. (2019). The maximum likelihood threshold of a path diagram. *Ann. Statist.* **47** 1536–1553. MR3911121 https://doi.org/10.1214/18-AOS1724

[9] DUTILLEUL, P. (1999). The mle algorithm for the matrix normal distribution. *J. Stat. Comput. Simul.* **64** 105–123. https://doi.org/10.1080/00949659908811970

[10] EDELMAN, A., ELMROTH, E. and KÅGSTRÖM, B. (1997). A geometric approach to perturbation theory of matrices and matrix pencils. I. Versal deformations. *SIAM J. Matrix Anal. Appl.* **18** 653–692. MR1453545 https://doi.org/10.1137/S0895479895284634

[11] EFRON, B. (2009). Are a set of microarrays independent of each other? *Ann. Appl. Stat.* **3** 922–942. MR2750220 https://doi.org/10.1214/09-AOAS236

[12] FOSDICK, B. K. and HOFF, P. D. (2014). Separable factor analysis with applications to mortality data. *Ann. Appl. Stat.* **8** 120–147. MR3191985 https://doi.org/10.1214/13-AOAS694

[13] GLANZ, H. and CARVALHO, L. (2018). An expectation-maximization algorithm for the matrix normal distribution with an application in remote sensing. *J. Multivariate Anal.* **167** 31–48. MR3830632 https://doi.org/10.1016/j.jmva.2018.03.010

[14] GREENEWALD, K. and HERO, A. O. III (2015). Robust Kronecker product PCA for spatio-temporal covariance estimation. *IEEE Trans. Signal Process.* **63** 6368–6378. MR3421045 https://doi.org/10.1109/TSP.2015.2472364

[15] GROSS, E. and SULLIVANT, S. (2018). The maximum likelihood threshold of a graph. *Bernoulli* **24** 386–407. MR3706762 https://doi.org/10.3150/16-BEJ881

[16] KOCH, D., LELE, S. and LEWIS, M. A. (2020). Computationally simple anisotropic lattice covariograms. *Environ. Ecol. Stat.* **27** 665–688.

[17] LANDSBERG, J. M. (2012). *Tensors*: *Geometry and Applications*. *Graduate Studies in Mathematics* **128**. Amer. Math. Soc., Providence, RI. MR2865915 https://doi.org/10.1090/gsm/128

[18] MAKEIG, S., KOTHE, C., MULLEN, T., BIGDELY-SHAMLO, N., ZHANG, Z. and KREUTZ-DELGADO, K. (2012). Evolving signal processing for brain–computer interfaces. *Proc*. *IEEE* **100** 1567–1584.

[19] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0652932

[20] MUROTA, K. (2000). *Matrices and Matroids for Systems Analysis*. *Algorithms and Combinatorics* **20**. Springer, Berlin. MR1739147

[21] OHLSON, M., RAUF AHMAD, M. and VON ROSEN, D. (2013). The multilinear normal distribution: Introduction and some basic properties. *J*. *Multivariate Anal*. **113** 37–47. MR2984354 https://doi.org/10.1016/j.jmva.2011.05.015

[22] RAPCSÁK, T. (1997). *Smooth Nonlinear Optimization in* $\mathbf{R}^N$. *Nonconvex Optimization and Its Applications* **19**. Kluwer Academic, Dordrecht. MR1480415 https://doi.org/10.1007/978-1-4615-6357-0

[23] SOLOVEYCHIK, I. and TRUSHIN, D. (2016). Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness. *J*. *Multivariate Anal*. **149** 92–113. MR3507317 https://doi.org/10.1016/j.jmva.2016.04.001

[24] TEN BERGE, J. M. F. and KIERS, H. A. L. (1999). Simplicity of core arrays in three-way principal component analysis and the typical rank of $p \times q \times 2$ arrays. *Linear Algebra Appl*. **294** 169–179. MR1693919 https://doi.org/10.1016/S0024-3795(99)00057-9

[25] VOLFOVSKY, A. and HOFF, P. D. (2015). Testing for nodal dependence in relational data matrices. *J*. *Amer*. *Statist*. *Assoc*. **110** 1037–1046. MR3420682 https://doi.org/10.1080/01621459.2014.965777

[26] WIESEL, A. (2012). Geodesic convexity and covariance estimation. *IEEE Trans*. *Signal Process*. **60** 6182–6189. MR3006411 https://doi.org/10.1109/TSP.2012.2218241

[27] YIN, J. and LI, H. (2012). Model selection and estimation in the matrix normal graphical model. *J*. *Multivariate Anal*. **107** 119–140. MR2890437 https://doi.org/10.1016/j.jmva.2012.01.005

[28] ZHOU, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *Ann*. *Statist*. **42** 532–562. MR3210978 https://doi.org/10.1214/13-AOS1187

[29] ZHU, Y. and LI, L. (2018). Multiple matrix Gaussian graphs estimation. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **80** 927–950. MR3874304 https://doi.org/10.1111/rssb.12278