Sparsity induced by covariance transformation: some deterministic and probabilistic results

Jakub Rybak* H. S. Battey*

January 26, 2021

Abstract
Motivated by statistical challenges arising in modern scientific fields, notably genomics, this paper seeks embeddings in which relevant covariance models are sparse. The work exploits a bijective mapping between a strictly positive definite matrix and its orthonormal eigen-decomposition, and between an orthonormal eigenvector matrix and its principle matrix logarithm. This leads to a representation of covariance matrices in terms of skew-symmetric matrices, for which there is a natural basis representation, and through which sparsity is conveniently explored. This theoretical work establishes the possibility of exploiting sparsity in the new parameterisation and converting the conclusion back to the one of interest, a prospect of high relevance in statistics. The statistical aspects associated with this operation, while not a focus of the present work, are briefly discussed.

Some key words: genomics, high-dimensionality, matrix embeddings, sparsity, statistical inference, structured random matrices.

1 Introduction
Genomics studies routinely produce data of a different character to that which shaped the development of statistical theory throughout the 20th century. Typically measurements are made on the expression levels of tens of thousands of genes, for a comparatively small number of individuals. An implicit or explicit assumption of sparsity – by which we mean the existence of many zeros or near-zeros – is essential for the provision of stable statistical procedures in such contexts. It is therefore of high relevance to determine embeddings in which the relevant probabilistic models are sparse. We study this challenging problem for covariance models.

Covariance matrices and their inverses (concentration matrices) arise in numerous statistical procedures, sometimes implicitly and almost always as nuisance parameters, needed to complete the specification of the probabilistic model but not of primary concern as far as the subject-matter is concerned. It is widespread statistical practice to estimate nuisance parameters, although this often affects estimation of interest parameters in a damagingly relevant sense when the dimension,

*Department of Mathematics, Imperial College London, 180 Queen’s Gate, South Kensington, London SW7 2AZ. JR: jakub.rybak18@imperial.ac.uk; HSB: h.battey@imperial.ac.uk.
$p$, of the nuisance parameter is large relative to the sample size, $n$. This has been known since Bartlett (1937), who provided a simple example. Ideally such nuisance parameters would be evaded. However this is typically difficult and in some settings seems impossible. An elegant example is the use of partial likelihood for estimation of the regression coefficients in the proportional hazards model (Cox, 1972, 1975). There, the unknown baseline hazard function, an infinite-dimensional nuisance parameter, is evaded completely. Other examples, exploiting special structure in the likelihood function, are discussed by Battey and Cox (2020).

Unstructured estimation of covariance matrices, when embedded within a broader statistical endeavour, typically leads to highly unstable and sometimes misleading statistical analyses. It is, for instance, well documented that the empirical covariance matrix based on $n$ independent observations and denoted by $S_n$, is not a sensible estimator of $\Sigma$ when the dimension $p$ of $\Sigma$ is large relative to $n$. While the elements of $S_n$ are individually consistent, the matrix is not consistent in the matrix norms relevant for applications. See e.g. Marchenko and Pastur (1967), Fan and Fan (2008) and Cacciapuoti et al. (2013) for a discussion of some of the adverse consequences of using $S_n$ in place of $\Sigma$ when $p \gg n$.

Although sparsity restores the ability to obtain consistent estimators, consistency is only interesting insofar as the assumptions made are satisfied to an adequate order of approximation. This motivates a search for relevant sparse parameterisations: a coordinate transformation that induces sparsity in the new coordinate system and that has a plausible interpretation in terms of the original random variables. As with any scientific model, the idea that covariance structure in a complex physical, biological or sociological system can be exactly described is unreasonably optimistic. Probabilistic models instead constitute idealised representations, capturing important stable aspects of the data generating mechanism (Cox, 1995).

There are several examples of such structure in wide use. One is the covariance graph models, which correspond to sparsity of the covariance matrix itself; another is the partial correlation graph models, otherwise known as Gaussian graphical models because of their interpretation in terms of conditional independencies when the distribution is Gaussian (Lauritzen, 1996). The latter class of models comprises covariance matrices whose inverses are sparse. It is notable that covariance matrices associated with Gaussian graphical models have no zeros as long as the underlying conditional independence graph is connected, while the inverse covariance matrix is sparse by construction. This paper aims at understanding different scenarios, where sparsity in another coordinate system may be more suitable. The questions addressed are of the following form. For a given, relevant, structure on the original covariance matrix (a covariance model), is there an alternative parameterisation in which the model is sparse? If one imposes sparsity in the aforementioned parameterisation, is the covariance model necessarily induced on the associated covariance matrix? Is the sparsity of a form that can be exploited for statistical purposes?

A similar structure of questions was considered by Battey (2017), but in a form somewhat less relevant for the genomics applications we have in mind. Specifically, the key question was of the form: if one imposes sparsity in a seemingly unnatural coordinate system, what structure is induced on the covariance matrix? To our knowledge, that work and the present paper are the only ones to consider sparse covariance embeddings in this way.

The conclusion from this earlier work was that the structure induced on the covariance matrix corresponds to a potentially large set of arbitrarily correlated random variables, and a separate set
of variables, uncorrelated with those in the first set and with each other. The latter also have unit variances.

While the structure induced through the sparse matrix logarithm considered by Battey (2017) is sometimes plausible for correlation matrices, at least as an approximation, for covariance matrices it is desirable to allow arbitrary variances for the small number of uncorrelated variables. The present paper establishes the parameterisation that produces the desired structure when sparsity is imposed, together with the converse result. The mapping proposed is substantially different to that studied by Battey (2017). Specifically, the representation is one in which the matrix logarithm of the orthonormal matrix of eigenvectors of $\Sigma$ possesses a particular form of sparsity, induced via a sparse basis representation for skew-symmetric matrices. In contrast, Battey (2017) considers matrix logarithmic transformation of the covariance matrix, corresponding to an elementwise logarithm of the eigenvalue matrix. As an incidental product of this work, we also prove the converse direction of a result in the aforementioned paper. Specifically, Battey (2017) only explicitly proves that a sparse matrix logarithm induces a particular structure on the covariance matrix. The reverse implication, arguably the direction of key practical relevance, is presented in Section 2.

Three aspects demand further comment.

We show that the sparsity class considered has a clear probabilistic interpretation. As alluded to above, it may in spite of this be too great an ask to require the variables to have exactly the structure specified. However, for nuisance parameters that cannot be eliminated by conditioning and other likelihood-based manipulations (see e.g. Barndorff-Nielsen and Cox, 1994), attention shifts to consistency of estimators in suitable norms rather than operational interpretation. One would then typically accept that a sparsity structure imposed through estimation be a reasonable approximation to the truth on the basis that variance is appreciably reduced over unstructured estimation.

A natural question is whether the work could be extended to study broader classes of relevant embeddings for covariance matrices where the interesting models are sparse. While such an endeavour would be of high relevance in statistics, producing formulations that are amenable to detailed mathematical analysis is challenging. Our proof strategies rely critically on the basis representation for skew-symmetric matrices.

An important issue not discussed in detail is estimation theory within our proposed sparsity class. We outline a draft proposal for this in Section 4. In the present paper we emphasise the deterministic and probabilistic aspects which, while motivated by statistical considerations, are of self-standing interest.

The following notation is used throughout the paper. Let $M$ be an $n \times n$ matrix. Then $|M|$ denotes its determinant and $M^T$ its transpose. Sets are denoted by calligraphic letters and $\{\}$ is used when defining finite sets. The set of indices $\{1, \ldots, n\}$ is written $[n]$. For a set $S$, $|S|$ denotes the cardinality of $S$. The $i$th column of $M$ is denoted by $m_{i \cdot}$, and the $j$th row by $m_{\cdot j}$. If a matrix has a subscript, e.g. $M_1$, the $i$th row is denoted by $m_{i1}$, and columns are denoted analogously. A canonical basis vector with 1 at the $i$th coordinate and zeros elsewhere is denoted $e_i$. The Frobenius, spectral and elementwise norms of a matrix $M$ are written as $\|M\|_F$, $\|M\|_{op}$ and $\|M\|_{\max}$ respectively. The Euclidean, $\ell_1$ and $\ell_0$ norms of a vector $x$ are written $\|x\|_2$, $\|x\|_1$ and $\|x\|_0$, where the latter counts the number of non-zero elements of $x$. 

3
2 A preliminary extension of earlier work

We first summarise the most closely related work in the literature and extend a key result from it. Our main contribution in Section 3 studies a question of similar structure, although the formulation and parameterisation considered there are different in obvious ways.

Throughout this paper observations are treated as $n$ realisations of independent and identically distributed random vectors $V$ of dimension $p$, where $p \gg n$. We denote such random vectors by $V_1, \ldots, V_n$, all valued in $\mathbb{R}^p$. The associated $p \times p$ covariance matrix is defined as $\Sigma = \mathbb{E}((V - \mathbb{E}V)(V - \mathbb{E}V)^T)$ and we assume for simplicity that $\mathbb{E}V = 0$.

The matrix logarithm of $\Sigma$ is the unique real symmetric matrix $L$ such that $\Sigma = \exp(L) = \sum_{k=0}^{\infty} L^k / k!$. By the positive definiteness of $\Sigma$, a more convenient representation of $L$ is available in the form $L = \Gamma(\log \Lambda)\Gamma^T$, where $\log \Lambda = \text{diag}(\log \lambda_1, \ldots, \log \lambda_p)$, with $\lambda_1 \geq \cdots \geq \lambda_p$ the ordered eigenvalues of $\Sigma$, and $\Gamma$ the matrix of corresponding orthonormal eigenvectors.

The usefulness of the matrix logarithm, exploited by Battey (2017), is that it belongs to the vector space of $p \times p$ symmetric matrices,

$$\mathcal{V}(p, \mathbb{R}) \triangleq \{ S \in \mathcal{M}_p(\mathbb{R}) : S = S^T \}. \quad (2.1)$$

In equation (2.1), $\mathcal{M}_p(\mathbb{R})$ is the space of $p \times p$ matrices with elements in $\mathbb{R}$. There is a natural symmetrised indicator basis for $\mathcal{V}(p, \mathbb{R})$, written $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$. Here $\mathcal{B}_1$ consists of $p$ diagonal matrices of the form

$$\mathcal{B}_1 = \{ B : B = e_j e_j^T, \ j \in [p] \},$$

for $e_1, \ldots, e_p$ the canonical basis vectors for $\mathbb{R}^p$, and $\mathcal{B}_2$ consists of $p(p - 1)/2$ symmetric matrices of the form

$$\mathcal{B}_2 = \{ B : B = e_j e_k^T + e_k e_j^T, \ j, k \in [p], j \neq k \}.$$  

The set $\mathcal{B}$ is such that $\mathcal{V}(p, \mathbb{R}) = \text{span}_\mathbb{R}\{ \mathcal{B}_1, \mathcal{B}_2 \} \triangleq \text{span}_\mathbb{R}\{ \mathcal{B}_1, \ldots, \mathcal{B}_{p(p-1)/2} \}$ and consists of linearly independent elements of $\mathcal{V}(p, \mathbb{R})$.

Suppose that $\alpha = (\alpha_1, \ldots, \alpha_{p(p+1)/2})^T$ in the basis expansion $L(\alpha) = \sum_{m=1}^{p(p+1)/2} \alpha_m B_m$ is sparse in the sense that only $s^*$ of its elements are non-zero, where $s^* \ll p$. Battey (2017) established the implications of this on the ordered eigenvalues and corresponding eigenvectors of $\Sigma(\alpha) = \exp(L(\alpha))$ and subsequently derived Proposition 2.1, whose interpretation requires some notation. We write $\mathcal{B}_S \triangleq \{ B_m \in \mathcal{B} : m \in S \}$ where $S$ is a function of $s^*$ via $\alpha$, specifically

$$S \triangleq S(s^*(\alpha)) = \{ m \in [p(p+1)/2] : \alpha_m \neq 0, \|\alpha\|_0 = s^* \}, \quad (2.2)$$

so that, by construction $|S(s^*(\alpha))| = s^*$. We denote the set of non-zero column vectors of the matrix entries of $\mathcal{B}_S$ by

$$\mathcal{L}(s^*(\alpha)) \triangleq \{ b_{m,j} \neq 0 \in \mathbb{R}^p : m \in S, \ j \in [p] \}, \quad (2.3)$$

and the number of distinct elements of $\mathcal{L}(s^*(\alpha))$ by $d^*(\alpha)$. The latter appears in Proposition 2.1.

**Proposition 2.1.** A positive definite symmetric matrix $\Sigma(\alpha)$ that is logarithmically sparse in the
sense that $\Sigma(\alpha) = \exp(L(\alpha))$ with

$$L(\alpha) = \sum_{m=1}^{p} (\alpha_m B_m, \|\alpha\|_0 = s^*$$

is necessarily of the form $\Sigma(\alpha) = P\Sigma^0 P^{-1}$ with $P$ a permutation matrix and $\Sigma^0$ a block diagonal matrix with blocks $\Sigma_1^0$ and $\Sigma_2^0 = I_{p-d^*(\alpha)}$.

Thus in terms of the random vector $V$ itself, sparsity of $\alpha$ implies that $V$ consists of two subsets of variables, $V_1$ and $V_2$, such that the elements of $V_1$ have covariance structure $\Sigma_1^0$, while those of $V_2$ are uncorrelated with each other and with the elements of $V_1$. Moreover, the elements of $V_2$ all have unit variances. Battey (2017) illustrates the possibility of exploiting appreciable sparsity in the $L$ domain that is not in general present in the $\Sigma = \exp(L)$ and $\Sigma^{-1} = \exp(-L)$ parameterisations.

For completeness, in Section 5 we prove the following stronger form of Proposition 2.1.

**Proposition 2.2.** A positive definite symmetric matrix $\Sigma(\alpha)$ is logarithmically sparse in the sense that $\Sigma(\alpha) = \exp(L(\alpha))$ with

$$L(\alpha) = \sum_{m=1}^{p} (\alpha_m B_m, \|\alpha\|_0 = s^*$$

if and only if $\Sigma(\alpha) = P\Sigma^0 P^{-1}$ with $P$ a permutation matrix and $\Sigma^0$ a block diagonal matrix with blocks $\Sigma_1^0$ and $\Sigma_2^0 = I_{p-d^*(\alpha)}$. Moreover, $\Sigma_1^0$ is of maximum dimension in the sense that it is not possible to find an alternative permutation matrix $P$ such that the second diagonal block is an identity matrix of larger dimension.

A concern is that the unit variance restriction on the elements of $V_2$ will be too strong in many situations. While this log-sparsity model would sometimes be reasonable for correlation matrices, standardisation would entail estimation of the variances of all $p$ random variables, and this is likely to destroy any theoretical guarantees that result from exploiting sparsity in the logarithmic domain. Moreover, if one writes the correlation matrix as $R = D^{-1/2} \Sigma D^{-1/2}$, where $D = \text{diag}(\Sigma_{11}, \ldots, \Sigma_{pp})$, there is no simple general relationship between the matrix logarithms of the two matrices because $\log(R) \not= D^{-1/2} \log(\Sigma) D^{-1/2}$ in general. Motivated by this, we seek a broader sparsity class of matrices: one that can capture a similar type of permuted-block structure to the logarithmic sparsity model discussed in this section, but allows for arbitrary variances of the variables in $V_2$. While the latter structure would seem sufficiently close to that induced by the logarithmic sparsity model to be achieved without too much difficulty, it turns out to require appreciable changes to the model and analysis, as we outline next.

### 3 A new sparse embedding for covariance models

#### 3.1 Specification

It is convenient for comparison purposes to use parallel notation to that in Section 2, although this entails redefining the sets $S$, $L(s^*(\alpha))$ and various related quantities. As far as the important mathematical properties are concerned, definitions always parallel those in Section 2. However, since the model presented here differs in important ways, there are considerable differences in the substantive details of these objects.
Consider the decomposition \( \Sigma = \Gamma \Lambda \Gamma^T = \sum_{j=1}^{p} \lambda_j \gamma_j \gamma_j^T \), where \( \Gamma \) is an orthonormal matrix of eigenvectors \( \gamma_1, \ldots, \gamma_p \) and \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p) \) is the corresponding diagonal matrix of eigenvalues. With the unit length restriction on the columns of \( \Gamma \), this decomposition is unique up to multiplication of columns of \( \Gamma \) by \(-1\), which does not change the associated eigenvalue. Any \( \Gamma \) satisfying this decomposition is either a special orthogonal matrix, i.e. has unit determinant, or can be made special orthogonal by multiplying a single arbitrary column of \( \Gamma \) by \(-1\). In what follows, we will assume that a special orthogonal decomposition has been chosen so that there exists a skew-symmetric matrix \( A = -A^T \) such that \( \Gamma = \exp(A) \). This relationship between special orthogonal matrices and skew symmetric matrices is reasonably well-known (e.g. Chevalley, 1946).

Unlike the space of orthogonal matrices, the space of \( p \)-dimensional skew-symmetric matrices is a vector space of dimension \( p(p-1)/2 \) and thus possesses a basis. Using the natural basis, \( B \triangleq \{ B : B = e_i e_j^T - e_j e_i^T, i < j \} \), we write \( A = A(\alpha) = \sum_{m=1}^{p(p-1)/2} \alpha_m B_m \) for \( B_m \in B \). This representation allows us to impose sparsity on \( A \) through \( \alpha \). Consider a sparsity assumption of the form \( s^* = \sum_{m=1}^{p(p-1)/2} 1\{\alpha_m \neq 0\} \), i.e. the number of non-zero coefficients in \( \alpha \). A size restriction on \( s^* \) induces sparsity on \( A(\alpha) \) and thereby induces structure on \( \Sigma \). The goal of this section is to establish the form of such structure and the stronger converse direction, i.e. the analogue of Proposition 2.2.

### 3.2 Some deterministic results

We first consider the structure induced on the eigenvector matrix \( \Gamma \) by the restriction on \( \alpha \). This does not follow directly from the results of Battey (2017) because \( \Gamma \) is not positive definite, which appreciably complicates the derivations. We nevertheless obtain a similar result.

In direct analogue to the notation used in Section 2, let

\[
S = S(s^*(\alpha)) = \{ m \in [p(p-1)/2] : \alpha_m \neq 0, \|\alpha\|_0 = s^* \}. \tag{3.1}
\]

and let \( d^*(\alpha) \) be the number of distinct elements of the set of non-zero column vectors of the matrices in \( B_S = \{ B_m \in B : m \in S \} \). In particular, it will be useful to define \( \mathcal{L}(s^*(\alpha)) \) in an analogous way to equation (2.3) as

\[
\mathcal{L}(s^*(\alpha)) \triangleq \{ b_{m,j} \neq 0 \in \mathbb{R}^p : m \in S, j \in [p] \}. \tag{3.2}
\]

Also let \( A \triangleq \{ j \in [p] : \tau_j \neq 1 \} \), where \( \tau_j \) is the \( j \)th eigenvalue of \( \Gamma \).

**Proposition 3.1.** For any \( \alpha \in \mathbb{R}^{p(p-1)/2} \), \( |A| \leq d^*(\alpha) \) and the inequality is strict for odd values of \( d^*(\alpha) \).

Proposition 3.1, while clearly different in detail, only differs structurally from Lemma 2.1 of Battey (2017) by the weak inequality sign, which was an equality there. While for the deterministic results of the present section the inequality is needed, when we consider random matrices in Section 3.3, leading to a random analogue of \( d^*(\alpha) \) denoted by \( D^*(\alpha) \), the event \( \{|A| < D^*(\alpha)\} \) is of negligible probability when the non-zero entries of \( \alpha \) are drawn from a continuous distribution. Hence in these settings the weak inequality in Proposition 3.1 can be treated as equality with
probability one. This is intuitive since $A(\alpha)$ has $p - D^*(\alpha)$ zero rows and columns, thus the probability of the event $\{|A| < D^*(\alpha)\}$ is equal to the probability that a random $D^* \times D^*$ matrix with no zero rows or columns is rank deficient. When the non-zero entries of $\alpha$ are drawn from a continuous distribution the determinant of the resulting matrix is continuously distributed and thus is zero on a set of measure zero.

**Proposition 3.2.** For any $\alpha \in \mathbb{R}^{p(p-1)/2}$, $\Gamma = \Gamma(\alpha)$ has eigenvectors $\xi_1, \ldots, \xi_p$, of which $p - d^*(\alpha)$ are of the form $\xi_i = e_{v(i)}$ for $e_{v(i)} \notin \mathcal{L}(s^*(\alpha))$ and correspond to $p - d^*(\alpha)$ unit eigenvalues of $\Gamma$. The remaining $d^*(\alpha)$ eigenvectors have at least $p - d^*(\alpha)$ zero entries.

From Propositions 3.1 and 3.2 we demonstrate the following result about the structure of the covariance matrix $\Sigma$.

**Proposition 3.3.** For any $\alpha \in \mathbb{R}^{p(p-1)/2}$, $\Sigma = \Sigma(\alpha) = \Gamma(\alpha) \Lambda \Gamma(\alpha)^T$, where $\Gamma(\alpha) = \exp(A(\alpha))$, $A(\alpha) = \sum_{m=1}^{p(p-1)/2} \alpha_m B_m$ and $\|\alpha\|_0 = s^* < p$ if and only if $\Sigma = P \Sigma^o P^{-1}$, where $P$ is a permutation matrix and $\Sigma^o$ is a block diagonal matrix with blocks $\Sigma_1^o$ and $\Sigma_2^o$, where $\Sigma_2^o$ is an arbitrary dense positive definite matrix with no rows or columns that are zero once the corresponding element of the main diagonal is removed, and $\Sigma_3^o$ is a $(p - d^*)$-dimensional square matrix with arbitrary entries on the diagonal and zeros elsewhere.

In terms of the random vector $V$ with covariance matrix $\Sigma$, the implication of Proposition 3.3 is that the above structure is present whenever $V$ consists of two subsets of variables, $\mathcal{V}_1$ and $\mathcal{V}_2$ such that the elements of $\mathcal{V}_1$ have covariance structure $\Sigma_1^o$, while the elements of $\mathcal{V}_2$ are uncorrelated with each other and with the elements of $\mathcal{V}_1$. Our new embedding is therefore one in which a rich and relevant class of models is sparse. Much more so than the class considered by Battey (2017), although the transformation in this earlier paper was simpler and probably easier to exploit statistically using the estimation theory studied by Battey (2019).

Although Proposition 3.3 is stated in terms of the covariance matrix $\Sigma$, the same structure, in the sense of positions of exact zeros, is present in the concentration matrix $\Sigma^{-1}$. This can be seen from the representation $\Sigma^{-1}(\alpha) = \Gamma(\alpha) \Lambda^{-1} \Gamma(\alpha)^T$ on noting that any sparsity constraint on $\alpha$ only affects the matrix $\Gamma(\alpha)$ of eigenvectors. The result also follows by the formula for the inverse of a block diagonal matrix. For Gaussian $V$, a zero in the $(i,j)$th component of $\Sigma^{-1}$ means that the $i$th and $j$th component of $V$ are conditionally independent, given all the other variables. See Lauritzen (1996) or Drton et al. (2009) for a detailed exposition of conditional independence graphs and Gaussian graphical models. The joint interpretation of the structure induced on $\Sigma$ and $\Sigma^{-1}$ by sparsity of $\alpha$ is that the conditional independence graph associated with $V$ has a dense subgraph in $d^*$ variables, which are independent of the remaining $p - d^*$ variables. The latter are also independent of one another.

### 3.3 Some results for random matrices with the specified structure

The covariance matrix associated with our sparse reparameterisation, whose structure is detailed in Proposition 3.3, may be rather dense or rather sparse, as governed by $d^*$. While it is clear that $d^*$ and $s^*$ are strongly related, a functional relationship between the two is not available from the deterministic analysis, because the positions of the non-zero entries of $\alpha$ are relevant. It is instructive, nevertheless, to see how $d^*$ is likely to change with $p$ and $s^*$ when the positions of the non-zero entries of $\alpha$ are chosen at random. Thus, while our results in Propositions 3.1 and 3.2
and Proposition 3.3 are deterministic, these can be made more interpretable by considering random matrices with the corresponding structure. Consider sparse random matrices $A(\alpha)$ obtained through the basis representation $A(\alpha) = \sum_{m=1}^{p(p-1)/2} \alpha_m B_m$ by letting the support of $\alpha$ be a simple random sample of size $s^*$ from the index set $[p(p-1)/2]$. Then $d^*(\alpha)$ of Propositions 3.1 – 3.3 is a random variable and, to emphasise this, we replace it by $D^*(\alpha)$, with realisation $d^*(\alpha)$. The distribution of non-zero entries of $\alpha_m$ is irrelevant to this discussion.

To obtain an approximation to the expected value of $D^*(\alpha)$ as a function of $p$ and $s^*$, consider initially the simpler problem of approximating the expected number of random draws $N_d$ from $\mathcal{B}$ required to obtain $d$ distinct elements of $\mathcal{L}(N_d(\alpha))$. On replacing the expected number of draws by $s^*$ and solving for $d$ we obtain an approximation to the expected value of $D^*(\alpha)$ and, in view of Proposition 3.1, an upper bound on the cardinality of the set $\mathcal{A}$.

The following argument follows that of Battey (2017) but requires alteration due to the different structure of the basis $\mathcal{B}$. For ease of reference we reproduce the argument with the appropriate modifications here.

**Figure 1:** $p = 100$ and sparsity of $\alpha$ is $s^* \in [100]$. (A) averaged (over 100 Monte Carlo replications) number of non-unit eigenvalues of $\Gamma(\alpha)$ as a function of $D^*(\alpha)$. The upper bound of Proposition 3.1 is seen to be a close approximation. (B) Monte Carlo average of $D^*$ as a function of $s^*$ and the analytic approximation to this relationship.

Introduce the nested sequence of random sets $\mathcal{L}(1) \subset \mathcal{L}(2) \subset \cdots$, somewhere among which $\mathcal{L}(s^*) \triangleq \mathcal{L}(s^*(\alpha))$ from equation (3.2) appears. Let $X_1 = 1$ denote the number of draws from $\mathcal{B}$ required to obtain the first new non-zero column vector in $\mathcal{L}(1)$. For $i \geq 2$, let $X_i$ be the number of additional draws required to obtain the $i$th new column vector of $\mathcal{L}(\sum_{j=1}^{i-1} X_j)$ after the $(i-1)$th new column has been drawn. Then the expected number of draws, $N_d$, such that $\mathcal{L}(N_d)$ contains $d$ distinct columns satisfies

$$
\mathbb{E}(N_d) \approx \sum_{i=1}^{d} \mathbb{E}(X_i) = \frac{p}{2} \sum_{i=1}^{d} \frac{1}{p - (i - 1)}.
$$

The last line follows from the fact that $X_i$ is a geometric random variable with parameter $2\{p - (i - 1)/p$. For this result to hold, the distribution from which the non-zero values of $\alpha_1, \ldots, \alpha_{p(p-1)/2}$ are drawn
is immaterial. The approximation in equation (3.3) arises from approximating sampling without replacement from $\mathcal{B}$ by sampling with replacement, which is accurate when $p(p-1)/2$ is large relative to $s^*$. An approximate expression for $\mathbb{E}(D^*(\alpha))$ is thus obtained by setting the left hand side of equation (3.3) to $s^*$ and solving for $d$. Changing variables to $j = p - (i - 1)$, we have

$$
\mathbb{E}(N_d) \approx \frac{p}{2} \left( \log p + \gamma + \varepsilon_p - (\log(d) + \gamma + \varepsilon_{p-d}) \right)
$$

where $\gamma$ is the Euler-Mascheroni constant and $\varepsilon_p \simeq 1/2p$ and $\varepsilon_{p-d} \simeq 1/(p-d)$. It follows that $\mathbb{E}(D^*(\alpha))$, which appears in deterministic form in Proposition 3.1 is well approximated by

$$
d^* = \text{root} \left( \frac{p}{2} \left( \log \left( \frac{p}{p-d} \right) - \frac{d}{2p(p-d)} \right) - s^* \right).
$$

We plot this solution for $p = 100$ in Figure 1A, observing that the analytic approximation to the expectation of $D^*$, although only an upper bound on the expected cardinality of $\mathcal{A}(\alpha)$, coincides almost perfectly with the Monte Carlo expected value of $|\mathcal{A}(\alpha)|$, indicating that the event $\{|\mathcal{A}| < D^*(\alpha)\}$ is of negligible probability under this data generating process. This is explained below Proposition 3.1.

### 3.4 Connection to the sparse matrix transform

A referee has drawn our attention to the so-called sparse matrix transform (SMT) of Cao and Bouman (2008), which has found wide practical usage. In that work the eigenvector matrix is decomposed as a product of pairwise coordinate rotations, known as Givens rotations (Givens, 1958). In particular, a SMT representation of order $K$ has $\Sigma = \Gamma \Lambda \Gamma^T$ and $\Gamma = \prod_{k=1}^K \Gamma_k$, where each $\Gamma_k$ is a Givens rotation operating on a pair of coordinate indices $(i_k, j_k)$, with $i_k < j_k$. These Givens rotations can be written as $\Gamma_k = I + \Theta(i_k, j_k, \theta_k)$, where for $u < v$

$$
\Theta(u, v, \theta)_{ij} = \begin{cases} 
\cos(\theta) - 1 & \text{if } i = j = u \text{ or } i = j = v, \\
\sin(\theta) & \text{if } i = u \text{ and } j = v, \\
-\sin(\theta) & \text{if } i = v \text{ and } j = u, \\
0 & \text{otherwise.}
\end{cases}
$$

With $K = p(p-1)/2$, any orthonormal matrix can be represented in this way, while $K < p(p-1)/2$ imposes sparsity on the eigenvector matrix $\Gamma$.

It is clear that the approach of Cao and Bouman (2008) is similar to the parameterisation considered in the present paper, in the sense that eigenvalues of the covariance matrix are left intact, while sparsity is imposed on the eigenvector matrix. Exact equivalence is harder to establish. To this end, note that Givens rotations are orthonormal with determinant one, that is, special orthogonal. Then for each Givens rotation $\Gamma_k$ there exists a skew-symmetric matrix $\mathcal{A}_k$ such that
\[ G_k = \exp(A_k). \] Thus, expanding in the natural basis matrices for the space of skew symmetric matrices, we write \[ A_k = \sum_{m=1}^{p(p-1)/2} \alpha_m^{(k)} B_m, \] where \( B_m \in \mathcal{B} \) for all \( m \), as defined in section 3.1. For any \( \Gamma_k = I + \Theta(i_k, j_k, \theta_k) \) with \( i_k < j_k \), its matrix logarithm \( A_k \) has the following form:

\[
(A_k)_{uv} = \begin{cases} 
\theta_k & \text{if } u = i_k \text{ and } v = j_k, \\
-\theta_k & \text{if } u = j_k \text{ and } v = i_k, \\
0 & \text{otherwise}.
\end{cases}
\]

Hence \( \Gamma_k = \exp(A_k) = \exp(\theta_k B_k) \). On writing \( \theta_k = \alpha_k \) and assuming, without loss of generality, that the basis elements have been ordered in such a way that \( \alpha_{K+1} = \ldots = \alpha_{p(p-1)/2} = 0 \) if \( K < p(p-1)/2 \), we have the following equivalence between the sparse matrix transform of Cao and Bouman (2008) and the sparse skew symmetric basis expansion elucidated in Propositions 3.1-3.3, with \( s^* \triangleq ||\alpha||_0 = K \):

\[
\Gamma = \prod_{k=1}^{K} \Gamma_k = \prod_{k=1}^{K} \exp(A_k) = \prod_{k=1}^{K} \exp(\alpha_k B_k) = \exp \left( \sum_{k=1}^{K} \alpha_k B_k \right) = \exp \left( \sum_{k=1}^{p(p-1)/2} \alpha_k B_k \right).
\]

The equivalence of the two structures is fortuitous, as the present work and that of Cao and Bouman (2008) are complementary. In particular, Cao and Bouman (2008) discussed estimation in greater detail than our section 4, proposing an estimator that exploits the assumed structure and providing some empirical support for it. On the other hand, the Proposition 3.3 provides greater transparency over the structure induced on the covariance and concentration matrices by the SMT, and thereby the implied relationships between the underlying random variables. See the discussion below Proposition 3.3 for further details.

### 4 Statistical implications

We have studied an embedding in which certain interesting covariance models, interpretable in terms of the original random vector \( V \), are sparse. As previously noted, the work is particularly relevant in statistics, where covariance and concentration matrices arise throughout classical multivariate analysis. Since they are invariably estimated, it is important to explore a range of possible sparsity assumptions and to understand their implications.

To illustrate the potential usefulness of considering sparsity on the non-standard scale studied in the present paper, Figure 2A plots the indicator function \( 1\{|A_{ij}| > 0\} \) against the coordinates \( i \) and \( j \) for one realisation of \( A(\alpha) \), generated as described in Section 3.3 with \( ||\alpha||_0 = 100 \). Figure 2B plots \( 1\{|\Sigma_{ij}| > 0\} \) for the corresponding covariance matrix \( \Sigma \). As with the Gaussian graphical models alluded to in Section 1, the covariance matrix \( \Sigma \) can be itself dense while possessing a sparse representation in an alternative coordinate system. In the case considered here, the sparse representation is present for the matrix logarithm of the eigenvector matrix. This underpins the potential value of our new representation for statistical purposes, revealing it as a parameterisation in which one of the interesting models is sparse. This suggests imposing sparsity via the basis coefficients of Section 3 and converting the conclusions back to the scale of interest.
Figure 2: For \( p = 100 \) and sparsity \( s^* = 100 \), plot of \( \mathbb{1}\{|M_{ij}| > 0\} \) for one realisation of \( M = (M_{ij}) \), where (A) \( M = A(\alpha) \) and (B) \( M = \Sigma(\alpha) = \exp\{A(\alpha)\} \Lambda \exp\{A(\alpha)^T\}, \) where the entries of the diagonal matrix \( \Lambda \) are independent draws from a unit exponential distribution.

It remains an open challenge to provide theoretical guarantees for estimators that exploit the structure studied in the present paper. A natural candidate is the estimator of Cao and Bouman (2008). Here we briefly outline another, somewhat similar, estimator that we would expect to be reasonable if \( V \) was normally distributed of zero mean and unknown covariance matrix \( \Sigma \) possessing the structure of Proposition 3.3. Given a sample \( V_1, \ldots, V_n \) of independent random vectors from the same distribution as \( V \), the logarithm of the likelihood function for \( \Sigma \) is

\[
\ell(\Sigma) \propto -\frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{i=1}^{n} V_i^T \Sigma^{-1} V_i
\]

where \( S_n = n^{-1} \sum_{i=1}^{n} V_i V_i^T \) is the empirical covariance matrix. In the first term of the previous display we have used that the logarithm of the determinant of a matrix is the trace of its logarithm, and for the second term we have used the cyclic permutation property of the trace. This representation can be traced back to at least Anderson and Olkin (1985). Cao and Bouman (2008) reparameterise \( 4.1 \) in terms of \( \Lambda \) and \( \Gamma \) by noting that \( \log \det(\Sigma) = \log \det(\Lambda) \) and

\[
\text{tr}(S_n \Sigma^{-1}) = \text{tr}(S_n \Gamma \Lambda^{-1} \Gamma^T) = \text{tr}(\text{diag}(\Gamma^T S_n \Gamma) \Lambda^{-1}).
\]

Differentiation of \( \ell(\Lambda, \Gamma) \) with respect to \( \Lambda \) shows that the maximum likelihood estimator of \( \Lambda \), with \( \Gamma \) assumed known, is \( \text{diag}(\Gamma^T S_n \Gamma) \), and replacement of \( \Lambda \) by its maximum likelihood estimator in \( \ell(\Lambda, \Gamma) \) gives the profile loglikelihood function:

\[
\ell_p(\Gamma) \propto \log \det(\text{diag}(\Gamma^T S_n \Gamma)). \tag{4.2}
\]

A natural way to exploit the structure discussed in the present paper is to reparameterise again by
replacing $\Gamma$ in equation (4.2) by $\sum_{m=1}^{p(p-1)/2} \alpha_m B_m$ for $B_m \in B$ and to exploit the sparsity of $\alpha$ by penalisation. Specifically, on letting $\ell_p(\alpha)$ denote the reparameterised expression (4.2), we propose to estimate $\alpha$ by

$$\hat{\alpha} \triangleq \operatorname{argsup}_{\alpha \in \mathbb{R}^{p(p-1)/2}} \left( \ell_p(\alpha) - \eta \|\alpha\|_1 \right),$$

(4.3)

for some tuning parameter $\eta$. The latter is usually chosen by cross-validation in penalised optimisation problems such as these.

Neither the theoretical properties of this estimator, nor the computational aspects associated with fitting the estimator have been explored in the present work. There are further key questions, such as existence and uniqueness of solutions, whether there are any advantages or disadvantages relative to the approach of Cao and Bouman (2008), and over robustness to departures from the Gaussianity assumption.

5 Proofs

5.1 Proofs of main results

Proof. [Proposition 2.2]. One direction of Proposition 2.2 is Proposition 2.1, proved by Battey (2017). For the converse direction, suppose that $\Sigma(\alpha) = P\Sigma^o P^{-1}$ with block diagonal component $\Sigma^o_2$ of maximum dimension in the sense that is not possible to construct an identity block of larger dimension than $p - d^*$ from $\Sigma^o_1$ through a different choice of $P$. Then $\Sigma^o = P^{-1}\Sigma P$ and $\Sigma$ are orthogonally similar and hence share the same eigenvalues.

By the Spectral Theorem $\Sigma^o = \Gamma^o \Lambda \Gamma^o_T$, where $\Gamma^o$ is an orthonormal matrix. Since $\Sigma^o_2$ is also positive definite and symmetric, we have: $\Sigma^o_1 = \Gamma^o_1 \Lambda_1 \Gamma^o_1^T$, where $\Gamma^o_1$ is orthonormal. We can thus write

$$\Sigma^o = \begin{pmatrix} \Gamma^o_1 & 0 \\ 0 & I_{p-d^*} \end{pmatrix} \begin{pmatrix} \Lambda_1 & 0 \\ 0 & I_{p-d^*} \end{pmatrix} \begin{pmatrix} \Gamma^o_1 & 0 \\ 0 & I_{p-d^*} \end{pmatrix}^T.$$

Since $\Gamma^o_1$ is orthonormal, the above is a spectral decomposition of $\Sigma^o$, unique up to permutations in the last $p - d^*$ columns of $\Sigma^o$ and $\Lambda$. The matrix logarithm $L^o = \log(\Sigma^o)$ is therefore

$$L^o = \begin{pmatrix} \Gamma^o_1 & 0 \\ 0 & I_{p-d^*} \end{pmatrix} \begin{pmatrix} \log \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Gamma^o_1 & 0 \\ 0 & I_{p-d^*} \end{pmatrix}^T = \begin{pmatrix} \log \Sigma^o_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Lemma 5.1. Suppose that $\Sigma^o_2$ is of maximum dimension $v$, in the sense defined above. Then the number of zero rows of $\log \Sigma^o$ is $v$, regardless of the number of unit eigenvalues of $\Sigma^o_1$.

Proof. Let $k$ denote the number of unit eigenvalues of $\Sigma^o_1$. First consider the case of $k = 0$. Then $\log \Sigma^o_1$ has no zero eigenvalues and is therefore of full rank, implying that $\log \Sigma^o_1$ cannot have any zero rows or columns. If follows that there are exactly $v$ zero rows of $L^o$.

Suppose now that there are $p - v \geq k \geq 1$ unit eigenvalues and hence let $\Gamma^o_1$ be such that $\Lambda_1$ is diagonal with diagonal blocks $\Lambda_{11} = \text{diag}(\lambda_1, \ldots, \lambda_{p-v-k})$, say and $I_k$. Then $\log(\Lambda_1)$ has a single non-zero diagonal block $\log(\Lambda_{11})$. The other elements of the diagonal of $\log(\Lambda_1)$ are zero. Partition $\Gamma^o_1$ analogously into four blocks, $\Gamma^o_{11}, \Gamma^o_{12}, \Gamma^o_{13},$ and $\Gamma^o_{14}$, say, where $\Gamma^o_{11} \in \mathbb{R}^{p-v-k \times p-v-k}$,
Let \( \Gamma^o_{14} \in \mathbb{R}^{k \times k} \), \( \Gamma^o_{13} \in \mathbb{R}^{k \times p - v - k} \), and \( \Gamma^o_{12} \in \mathbb{R}^{p - v - k \times k} \). Then

\[
\Gamma^o_{12}T = \begin{pmatrix}
\Gamma^o_{11}T \Gamma^o_{11} + \Gamma^o_{13}T \Gamma^o_{13} & \Gamma^o_{12}T \Gamma^o_{12} + \Gamma^o_{14}T \Gamma^o_{14} \\
(\Gamma^o_{12}T \Gamma^o_{11} + \Gamma^o_{14}T \Gamma^o_{13}) & \Gamma^o_{12}T \Gamma^o_{12} + \Gamma^o_{14}T \Gamma^o_{14}
\end{pmatrix} = I_{p-v},
\]

\[
\Gamma^o_{13}T = \begin{pmatrix}
\Gamma^o_{11}T \Gamma^o_{11} + \Gamma^o_{13}T \Gamma^o_{13} & \Gamma^o_{12}T \Gamma^o_{12} + \Gamma^o_{14}T \Gamma^o_{14} \\
(\Gamma^o_{13}T \Gamma^o_{11} + \Gamma^o_{14}T \Gamma^o_{13}) & \Gamma^o_{13}T \Gamma^o_{13} + \Gamma^o_{14}T \Gamma^o_{14}
\end{pmatrix} = I_{p-v}.
\]

The orthogonality of \( \Gamma^o_{11} \) implies that if \( \Gamma^o_{12} \) is 0, columns of \( \Gamma^o_{13} \) are orthogonal to all columns of \( \Gamma^o_{14} \), where \( \Gamma^o_{14} \) is a \( k \times k \) orthonormal matrix. This implies that \( \Gamma^o_{13} \) is zero. An analogous argument implies that if \( \Gamma^o_{13} \) is zero, \( \Gamma^o_{12} \) is also zero. But if both \( \Gamma^o_{13} \) and \( \Gamma^o_{12} \) are zero, \( \Gamma^o_{14} = I_k \) by the orthonormality of \( \Gamma^o_1 \). There is contradiction with the statement that \( I_v \) is of maximum dimension unless \( \Gamma^o_{12} \) and \( \Gamma^o_{13} \) are non-zero. Thus we proceed on the basis of non-zero \( \Gamma^o_{12} \) and \( \Gamma^o_{13} \).

Using the formula for matrix multiplication of partitioned matrices we obtain

\[
L^o_1 = \log \Sigma^o = \begin{pmatrix}
\Gamma^o_{11} \log \Lambda_{11} \Gamma^o_{11} & \Gamma^o_{11} \log \Lambda_{11} \Gamma^o_{13} \\
\Gamma^o_{13} \log \Lambda_{11} \Gamma^o_{11} & \Gamma^o_{13} \log \Lambda_{11} \Gamma^o_{13}
\end{pmatrix}.
\]

We will show that \( L^o_1 \) has no zero rows by first showing this for its first \( k \) rows and then for its first \( p-v-k \) rows.

The \( j \)th rows of \( \Gamma^o_{11} \Lambda_{11} \Gamma^o_{11} \) and \( \Gamma^o_{13} \log \Lambda_{11} \Gamma^o_{13} \) are \( \sum_i \log(\lambda_i) \gamma^o_{11,[j]} \gamma^o_{11,i} \) and \( \sum_i \log(\lambda_i) \gamma^o_{13,[j]} \gamma^o_{13,i} \), respectively for \( j = 1, \ldots, k \). Similarly, the \( j \)th rows of \( \Gamma^o_{13} \log \Lambda_{11} \Gamma^o_{13} \) and \( \Gamma^o_{11} \log \Lambda_{11} \Gamma^o_{11} \) are \( \sum_i \log(\lambda_i) \gamma^o_{13,[j]} \gamma^o_{13,i} \) and \( \sum_i \log(\lambda_i) \gamma^o_{11,[j]} \gamma^o_{11,i} \), respectively for \( j = 1, \ldots, p-v-k \). The \( j \)th column of \( \Gamma^o_1 \) has component vectors \( \gamma^o_{11,j} \) and \( \gamma^o_{13,j} \). Suppose for a contradiction that the \( j \)th row of

\[
\left( \Gamma^o_{13} \log \Lambda_{11} \Gamma^o_{11} \quad \Gamma^o_{13} \log \Lambda_{11} \Gamma^o_{13} \right)
\]

is zero. Then there exists a non-trivial combination of the columns of \( \Gamma^o_1 \) that give the zero vector and hence \( \Gamma^o_1 \) is singular, contradicting its orthogonality. The linear combination is non-trivial only if the \( j \)th row of \( \Gamma^o_{13} \) is non-zero. Suppose for a contradiction that the \( j \)th row of \( \Gamma^o_{13} \) is zero, then the \( j \)th row of \( \Gamma^o_{14} \) would be of unit length and orthogonal to the remaining rows of \( \Gamma^o_{14} \) and rows of \( \Gamma^o_{12} \) by the orthonormality of \( \Gamma^o_1 \). This is in contradiction with \( I_v \) being of maximum dimension, which can be seen through the matrix multiplication \( \Gamma^o_1 \Lambda_{11} \Gamma^o_1 \). This proves that the last \( k \) rows of \( \log \Sigma^o_1 \) are non-zero.

Consider the first \( p-k-v \) columns of \( L^o_1 \). Since \( L^o_1 \) is symmetric, this is equivalent to studying the first \( p-k-v \) rows. The \( j \)th column of \( \Gamma^o_{11} \log \Lambda_{11} \Gamma^o_{11} \) is \( \sum_i \log(\lambda_i) \gamma^o_{11,[j]} \gamma^o_{11,i} \) and the \( j \)th column of \( \log(\Lambda_1) \Gamma^o_{11} \) is \( \sum_i \log(\lambda_i) \gamma^o_{11,[j]} \gamma^o_{11,i} \). Hence the \( j \)th column of \( L^o_1 \) for \( j = 1, \ldots, p-k-v \) is

\[
\sum_i \log(\lambda_i) \gamma^o_{11,[j]} \left( \begin{array}{c}
\gamma^o_{11,i} \\
\gamma^o_{13,i}
\end{array} \right).
\]

The vectors in the sum are columns of \( \Gamma^o_1 \) so that a zero column of \( L^o_1 \) would imply a singular \( \Gamma^o_1 \) so long as \( \gamma^o_{11,j} \) is not zero, this is a contradiction to the orthonormality of \( \Gamma^o_1 \). Suppose for a contradiction that the \( j \)th row of \( \Gamma^o_1 \) is zero. Then \( \gamma^o_{12,j} \) is of unit length and by the orthogonality
of \( \Gamma_0 \) all other rows \( \gamma_{2l}^i, l \neq i \) are either orthogonal to it or zero. This contradicts with \( \Sigma_2 \) being of maximum dimension. It follows that the first \( p - v - k \) columns of \( L_1^0 \) are non-zero, and since \( L_1^0 \) is symmetric, the same holds for its rows. The number of zero rows in \( L^0 \) is thus \( v \) for any \( k \) \((0 \leq k \leq p - v)\).

Since \( \Sigma = P\Sigma^0 P^{-1} \), its matrix logarithm satisfies

\[
L = PL^0 P^{-1} = P \left( \log \Sigma_1^0 \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right) P^{-1}.
\]

Since this is a symmetric transformation of \( L^0 \), \( L \) also has \( v \) zero rows. Since \( L \) is symmetric it has a basis expansion \( L(\alpha) = \sum_{m=1}^{M} \alpha_m B_m \). Let \( s^* = \|\alpha\|_0 \). Since the columns of basis matrices are of the form \( e_j \) or zero, the number of non-zero rows of \( L^0 \) corresponds to the number of unique elements in the set \( L(s^*(\alpha)) \), denoted by \( d^* \). By the statement of the proposition \( v = p - d^* \), completing the proof.

**Proof.** [Proposition 3.1] Let \( \alpha \in \mathbb{R}^{p(p-1)/2} \) and let \( \Gamma = \exp(A) \), where \( A = \sum_{m=1}^{p(p-1)/2} \alpha_m B_m \). By the real Schur decomposition, \( \Gamma = QTQ^T \), where \( Q \in \mathbb{R}^{p \times p} \) is orthonormal and \( T \) is the real Schur form of \( \Gamma \), which is block diagonal with \( \alpha \)-th block of the form \( T_{ii} = \pm 1 \) or

\[
T_{ii} = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}.
\]

There are zero or an even number of blocks of type \( T_{ii} = -1 \). Then \( A = \log(\Gamma) = Q \log(T)Q^T \). Let \( \mathcal{A}^c = \{ j \in [p] : \tau_j = 1 \} \), where \( \tau_j \) denotes the \( j \)-th eigenvalue of \( \Gamma \). Then \( \log(\tau_j) = 0 \) for all \( j \in \mathcal{A}^c \), resulting in a zero diagonal term of \( log(T) \).

Let \( q_j \) be the \( j \)-th column vector of \( Q \). We first show that \( \text{span}\{q_j : j \in \mathcal{A}^c\} = \ker(A) \) by showing containment on both sides. Consider the \( q_j \) for which \( j \in \mathcal{A}^c \). Then

\[
Aq_j = Q \log(T)Q^T q_j = Q \log(T)(q_1, \ldots, q_p)^T q_j = Q \log(T)e_j = Q \log(\tau_j) = 0
\]

since \( \tau_j = 1 \) for all \( j \in \mathcal{A}^c \). As this holds for any \( q_j, j \in \mathcal{A}^c \), it holds for any linear combination of them, implying that all \( x \) in \( \text{span}\{q_j : j \in \mathcal{A}^c\} \) are also in the kernel of \( A \), or equivalently \( \text{span}\{q_j : j \in \mathcal{A}^c\} \subseteq \ker(A) \).

To prove the converse containment, suppose for a contradiction that \( \ker(A) \not\subseteq \text{span}\{q_j : j \in \mathcal{A}^c\} \). Then there exists a vector \( v \in \mathbb{R}^p \) such that \( v \in \ker(A) \) and \( v \not\in \text{span}\{q_j : j \in \mathcal{A}^c\} \). For notational convenience, let \( \Psi \) be a matrix consisting of the columns of \( Q \), reordered such the the first \( k \) entries are the \( q_j \) such that \( j \in \mathcal{A}^c \). Thus introduce \( \Psi_{\mathcal{A}^c} = (\psi_1, \ldots, \psi_k) \) and \( \Psi_{\mathcal{A}} = (\psi_{k+1}, \ldots, \psi_p) \).

Since \( Q \) is orthonormal, its columns are orthogonal and hence the set of vectors \( (\psi_1, \ldots, \psi_k) \) is linearly independent. Since \( \text{span}\{q_j : j \in \mathcal{A}^c\} \subseteq \ker(A) \), each column vector of \( \Psi_{\mathcal{A}^c} \) belongs to \( \ker(A) \) and hence the set can be extended to a basis of the vector space \( \ker(A) \). Let \( (\psi_1, \ldots, \psi_k, \psi'_{k+1}, \ldots, \psi'_p) \) be an extended basis of \( \ker(A) \). Since \( v \in \ker(A) \), there exists \( \omega_1, \ldots, \omega_p \in \mathbb{R} \) such that \( v = \omega_1 \psi_1 + \ldots + \omega_k \psi_k + \omega_{k+1} \psi'_{k+1} + \ldots + \omega_p \psi'_p \). For all \( i \in [k] \), \( Aq_i = 0 \). Hence \( Av = \omega_{k+1} \psi'_{k+1} + \ldots + \omega_p \psi'_p \). Since \( \psi_1, \ldots, \psi_p \) are linearly independent, they form a basis of \( \mathbb{R}^p \). Thus, for all \( i \in \{k+1, \ldots, p\} \) there exists \( \kappa_1^i, \ldots, \kappa_p^i \) such that \( \psi'_i = \kappa_1^i \psi_1 + \ldots + \kappa_p^i \psi_p \). Substituting
in the basis expansion for \( v \), we obtain

\[
Av = \omega_{k+1} \sum_{j=1}^{p} \kappa_{j}^{k+1} A\psi_{j} + \cdots + \omega_{p} \sum_{j=1}^{p} \kappa_{j}^{p} A\psi_{j} \\
= \omega_{k+1} \sum_{j=k+1}^{p} \kappa_{j}^{k+1} A\psi_{j} + \cdots + \omega_{p} \sum_{j=k+1}^{p} \kappa_{j}^{p} A\psi_{j} = 0,
\]

where the last equality follows because \( v \in \ker(A) \).

For \( q_{i} \in \Psi_{A} \), \( Aq_{i} = Q \log(T)Q^{T}q_{i} = Qc_{i}e_{i} = q_{i}c_{i} \), where \( c_{i} \in \mathbb{R} \) corresponds to the non-zero element of the \( i \)th column vector of \( \log(T) \). The form of \( \log(T) \) can be obtained from Lemmas 5.3 and 5.4 thus we obtain, since \( q_{i} \in \Psi_{A} \), that there exists exactly one such element. Hence

\[
Av = \left( \sum_{i=k+1}^{p} \omega_{i} \kappa_{i}^{i} \right) \psi_{k+1} + \cdots + \left( \sum_{i=k+1}^{p} \omega_{i} \kappa_{i}^{i} \right) c_{p} \psi_{p} = 0,
\]
i.e. a linear combination of the columns of \( \Psi_{A} \). Since columns of \( \Psi_{A} \) are linearly independent, the linear combination must be the trivial one in which all coefficients are zero. Since \( c_{i} \neq 0 \) for all \( i \in \{k+1, \ldots, p\} \), we have that for all \( \ell \in \{k+1, \ldots, p\} \), \( \sum_{i=k+1}^{p} \omega_{i} \kappa_{i}^{\ell} = 0 \).

Introduce the \((p-k)\)-dimensional square matrix

\[
K = \begin{pmatrix}
\kappa_{k+1}^{k+1} & \cdots & \kappa_{k+1}^{p} \\
\kappa_{k+1}^{k+2} & \cdots & \kappa_{k+1}^{k+2} \\
\vdots & \ddots & \vdots \\
\kappa_{p}^{k+1} & \cdots & \kappa_{p}^{p}
\end{pmatrix},
\]

and the vector \( \omega = (\omega_{k+1}, \ldots, \omega_{p})^{T} \). Then we can rewrite this system of equations as \( K\omega = 0 \). Note that the rows of \( K^{T} \) are (subsets of) coefficients of the basis expansion of \( \psi_{k+1}', \psi_{p}' \), i.e. of the elements of the extended basis of \( \ker(A) \).

Suppose for a contradiction that \( K \) is not of full rank. Then, one of the columns of \( K \) can be expressed as a linear combination of the remaining columns. Without loss of generality let this be the first column, \( K_{1} \). Then there exists \( \beta_{k+2}, \ldots, \beta_{p} \in \mathbb{R} \) such that \( K_{1} = \beta_{k+2}K_{2} + \cdots + \beta_{p}K_{p-k} \), and the basis expansion of \( \psi_{k+1}' \) can be written as

\[
\psi_{k+1}' = \kappa_{1}^{k+1} \psi_{1} + \cdots + \kappa_{k}^{k+1} \psi_{k} + \sum_{i=k+1}^{p} \kappa_{i}^{k+1} \psi_{i} \\
= \kappa_{1}^{k+1} \psi_{1} + \cdots + \kappa_{k}^{k+1} \psi_{k} + \sum_{i=k+1}^{p} \left( \sum_{l=k+2}^{p} \beta_{l} \kappa_{i}^{l} \right) \psi_{i} \\
= \kappa_{1}^{k+1} \psi_{1} + \cdots + \kappa_{k}^{k+1} \psi_{k} + \sum_{i=k+1}^{p} \beta_{l} \left( \sum_{i=k+1}^{p} \kappa_{i}^{l} \psi_{i} \right) \\
= \sum_{i=k+1}^{p} \beta_{l} \psi_{1}' + \left( \gamma_{1}^{k+1} - \sum_{i=k+2}^{p} \beta_{l} \kappa_{i}^{l} \right) \psi_{1} + \cdots + \left( \kappa_{k}^{k+1} - \sum_{i=k+2}^{p} \beta_{l} \kappa_{i}^{l} \right) \psi_{k}.
\]

It follows that \( \psi_{k+1}' \) is a linear combination of \( \psi_{1}, \ldots, \psi_{k}, \psi_{k+1}', \ldots, \psi_{p}' \). Hence, the set of vectors \( \langle \psi_{1}, \ldots, \psi_{k}, \psi_{k+1}', \ldots, \psi_{p}' \rangle \) is not linearly independent and hence not a basis for \( \ker(A) \), a contradiction. We conclude that \( K \) is full rank. This implies that the system of equations has a unique
solution $\omega = K^{-1} 0 = 0$. Combining the result that $\omega = (\omega_{k+1}, \ldots, \omega_p)^T = 0$ with the expansion $v = \omega_1 \psi_1 + \cdots + \omega_k \psi_k + \omega_{k+1} \psi_{k+1} + \cdots + \omega_p \psi_p$, we see that $v$ is a linear combination of $\psi_1, \ldots, \psi_k$ and hence $v \in \text{span} \{q_j : j \in \mathcal{A}^c\}$, a contradiction. Hence, $\ker(A) \subseteq \text{span} \{q_j : j \in \mathcal{A}^c\}$.

We next demonstrate that $\dim \{\text{Im}(A)\} = |\mathcal{A}|$. Since $A \in \mathbb{R}^{p \times p}$, $p = |\mathcal{A} \cup \mathcal{A}^c| = |\mathcal{A}| + |\mathcal{A}^c|$. By the Rank Nullity Theorem $p = \dim(\text{Im}(A)) + \dim(\ker(A))$. Since $\ker(A) = \text{span} \{q_j : j \in \mathcal{A}^c\}$, $\dim \{\ker(A)\} = |\mathcal{A}^c|$. Combining the three equations yields $\dim \{\text{Im}(A)\} = |\mathcal{A}|$.

$A$ has $p - d^*(\alpha)$ zero rows, so $\text{rank}(A) \leq d^*(\alpha)$. Since $\dim \{\text{Im}(A)\} = \text{rank}(A)$, we obtain $\dim \{\text{Im}(A)\} \leq d^*(\alpha)$. The inequality is strict for odd values of $d^*(\alpha)$, since all real eigenvalues of skew-symmetric matrices are 0, and complex eigenvalues come in conjugate pairs. We have shown that $|\mathcal{A}| = \dim \{\text{Im}(A)\} \leq d^*$.

Proof. [Proposition 3.2] Recall that $\Psi$ is a matrix consisting of the columns of $Q$, reordered such that the first $k$ entries are the $q_j$ such that $j \in \mathcal{A}^c$. Thus $\Psi_{\mathcal{A}^c} = (\psi_1, \ldots, \psi_k)$ and $\Psi_A = (\psi_{k+1}, \ldots, \psi_p)$. By Proposition 3.1, $q_i$ is a column of $\Psi_{\mathcal{A}^c}$, i.e. a column of $Q$ corresponding to $T_{ii} = 1$ if and only if $q_i \in \ker(A)$. Then

$$A q_i = A_1 q_{1i} + \cdots + A_p q_{pi} = q_{1i} \left( \sum_{m \in S} \alpha_m b_1^m \right) + \cdots + q_{pi} \left( \sum_{m \in S} \alpha_m b_p^m \right) = 0.$$  \hfill (5.1)

Let $\mathcal{V} \subseteq [p]$ be a set of non-zero row coordinates of $\mathcal{L}\{s^*(\alpha)\}$, i.e. rows for which there is at least one non-zero entry in the set of basis matrices with non-zero coefficients. Then for any $\kappa_i \neq 0$, $q_i = \kappa_i e_{v(i)}$ for $v(i) \in \mathcal{V}$ clearly satisfies equation (5.1). Let $k = |\mathcal{A}^c|$, i.e. the number of unit eigenvalues of $A$. There are $k$ unit diagonal entries of $T$ and by Proposition 3.1, $k = \dim \{\ker(A)\} \geq p - d^*$. We have $|\mathcal{V}| = p - d^*$. Hence for $p - d^*$ unit eigenvalues, we can set $q_i = \kappa_i e_{v(i)}$, $q_j = \kappa_j e_{v(j)}$ with $\kappa_i, \kappa_j \neq 0$ and $v(i) \neq v(j)$ for $i \neq j$, $v(i), v(j) \in \mathcal{V}$. The orthonormality then requires that for all $i$, $\kappa_i = \pm 1$ and we take the special orthogonal representation with $\kappa_i = 1$ for all $i$.

The eigenvectors corresponding to the unit eigenvalues of $\Gamma$ thus fall into two categories and we can write $\Psi_{\mathcal{A}^c} = \Psi_{\mathcal{A}_1} \cup \Psi_{\mathcal{A}_2}$, where $\Psi_{\mathcal{A}_1} \in \mathbb{R}^{p \times p - d^*}$ consists of column vectors of the form $e_i$, $i \in \mathcal{V}$ and $\Psi_{\mathcal{A}_2} \in \mathbb{R}^{p \times k - p + d^*}$ consists of the eigenvectors corresponding to the remaining $(k - p + d^*)$ unit eigenvalues.

Note that $\Xi_A \in \mathbb{R}^{p \times p - k}$. Consider $\Psi_{\cup \mathcal{A}_2}$, the $p \times d^*$ matrix with columns given by $\Psi_{\mathcal{A}}$ and $\Psi_{\mathcal{A}_2}$. By linear independence, which follows from the orthogonality of $Q$, $\Psi_{\cup \mathcal{A}_2}$ is of full rank. Since $p \geq d^*$, there are $p - d^*$ linearly dependent rows in $\Psi_{\cup \mathcal{A}_2}$. To achieve orthogonality to the columns of $\Psi_{\mathcal{A}_1}$ any column of $\Psi_{\cup \mathcal{A}_2}$ must have zero entries at coordinates $\mathcal{V}$, hence the linearly dependent rows of $\Psi_{\cup \mathcal{A}_2}$ are 0.

Since $A = \sum_{m=1}^{p(p-1)/2} \alpha_m B_m$ with rows $A_i$ equal to zero for $i \in \mathcal{V}$, the corresponding columns $A_i$ are zero too, as $A$ is skew-symmetric. Apply a permutation $\tilde{A} = \tilde{P} A \tilde{P}^T$ such that the zero rows and columns are the last ones. Skew-symmetry is preserved. Further $A$ and $\tilde{A}$ are orthogonally similar (because permutation matrices are orthogonal) and hence share the same eigenvalues. Thus, $\exp(A)$ and $\exp(\tilde{A})$ have the same eigenvalues. It follows that there exists an orthogonal matrix $\tilde{Q} \in \mathbb{R}^{p \times p}$ such that $\exp(\tilde{A}) = \tilde{Q} \tilde{T} \tilde{Q}^T$, and $T$ is, by Lemma 5.4, block diagonal with a block denoted by $T_1$ and a block of the form $I_{p-d^*}$ in the position of its last $p-d^*$ rows and columns. This follows because $\tilde{A}$ has zeros in the position of its last $p-d^*$ rows and columns. Setting columns $\tilde{q}_j$ of $\tilde{Q}$
corresponding to the $p - d^*$ diagonal entries of $\log(T)$ to $e_{\nu(j)}$, we obtain $\tilde{Q}$ block diagonal, with a block denoted $\tilde{Q}_1$ and a block of the form $I_{p - d^*}$ in the position of its last $p - d^*$ rows and columns. Further

$$\tilde{\Gamma} \triangleq \tilde{Q} T \tilde{Q}^T = \begin{pmatrix} \tilde{Q}_1 & 0 & 0 \\ 0 & I_{p - d^*} & 0 \\ 0 & 0 & I_{p - d^*} \end{pmatrix} \begin{pmatrix} T_1 & 0 & 0 \\ 0 & I_{p - d^*} & 0 \\ 0 & 0 & I_{p - d^*} \end{pmatrix} = \begin{pmatrix} \tilde{Q}_1 T_1 \tilde{Q}_1^T & 0 & 0 \\ 0 & I_{p - d^*} & 0 \\ 0 & 0 & I_{p - d^*} \end{pmatrix},$$

where $T_1$ is a matrix of eigenvalues of $\tilde{A}_1$, the $d^*$-dimensional non-zero block of $\tilde{A}$ and $\tilde{\Gamma}_1 \triangleq \tilde{Q}_1 T_1 \tilde{Q}_1^T \in R^{d^* \times d^*}$ is orthogonal. Such a $\tilde{Q}_1$ thus exists as it corresponds to the orthogonal matrix in the real Schur decomposition of $\tilde{\Gamma}_1$.

Using this result and the definition of $\tilde{A}$ we obtain

$$A = \tilde{P}^T \tilde{A} \tilde{P} = \tilde{P}^T \tilde{Q} \log(T) \tilde{Q}^T \tilde{P} = (\tilde{P}^T \tilde{Q}) \log(T) (\tilde{P}^T \tilde{Q})^T.$$

Setting $Q = \tilde{P}^T \tilde{Q}$, we thus complete the specification of the decomposition of $\Gamma$.

**Proof.** [Proposition 3.3] We will first prove that a covariance matrix with the sparse skew-symmetric structure in the matrix logarithm of its eigenvector matrix is necessarily of the form described.

Consider $\tilde{A}$ as defined in the proof of Proposition 3.1 with associated real Schur decomposition $\tilde{\Gamma} \triangleq \exp \tilde{A} = \tilde{Q} T \tilde{Q}^T$. Recall that $\tilde{A}$ and $A = A(\alpha) = \sum_{m=1}^{p(p-1)/2} \alpha_m B_m$ share the same eigenvalues and with $Q = \tilde{P}^T \tilde{Q}$, where $\tilde{P}$ is as in the proof of Proposition 3.1, $\Gamma = \exp A = Q T Q^T$, where $T$ is a real Schur form of $\exp(A)$. By Proposition 3.1 and the representation used in its proof we have

$$\Gamma(\alpha) = \exp \{A(\alpha)\} = Q T Q^T = (\tilde{P}^T \tilde{Q}) T (\tilde{P}^T \tilde{Q})^T \triangleq \tilde{P}^T \begin{pmatrix} \tilde{\Gamma}_1 & 0 \\ 0 & I_{p - d^*} \end{pmatrix} \tilde{P}. $$

Hence, by direct calculation, $\Sigma(\alpha) = \Gamma(\alpha) \Lambda \Gamma(\alpha)^T$ is of the form $\tilde{P}^T \Sigma^o \tilde{P}$, where $\Sigma^o$ is a block diagonal matrix with blocks $\Sigma^o = \tilde{\Gamma}_1 \Lambda_1 \tilde{\Gamma}_1^T$ and $\Sigma^o_2 = \Lambda_2$, where $\Lambda \triangleq \text{blockdiag}(\Lambda_1, \Lambda_2)$ and $\tilde{\Lambda} = \tilde{P} \Lambda \tilde{P}^T$. With $P = \tilde{P}^T$, noting that $P^{-1} = P^T$, we have the required result.

For the converse direction suppose that $\Sigma = P \Sigma^o P^{-1}$, where $\Sigma^o$ is block diagonal with blocks $\Sigma^o_1$ and $\Sigma^o_2$ whose dimensions and structures are as described in the statement of Proposition 3.3. By the statement of the proposition, the diagonal matrix $\Sigma^o_2$ is of maximal dimension $p - d^*$, in the sense that a diagonal block of larger dimension cannot be obtained through a different choice of permutation matrix $P$. Consider $\Sigma^o = P^{-1} \Sigma P$, which is orthogonally similar to $\Sigma$ and thus has the same eigenvalues.

Let $\Gamma^o \Lambda^o \Gamma^o^T$ denote the spectral decomposition of $\Sigma^o$. By the specification of $\Sigma^o$, $\Gamma^o$ is block diagonal with upper block $\Gamma^o_1$ and lower block equal to $I_{p - d^*}$, and $\Lambda^o = \text{blockdiag}(\Lambda^o_1, \Lambda^o_2)$, where $\Sigma^o$ is diagonal with positive entries and dimension $p - d^*$, as described above.

Since $\Gamma^o_1$ corresponds to the eigenvalue matrix from the spectral decomposition of $\Sigma^o_1$, it can be chosen to be special orthogonal, and hence $\Gamma^o$ is special orthogonal too and there exists $Q^o_1$ and $T^o_1$ such that $\Gamma^o = Q^o_1 T^o_1 Q^o_1^T$. Then

$$\Gamma^o = \begin{pmatrix} Q^o_1 & 0 & 0 \\ 0 & I_{p - d^*} & 0 \\ 0 & 0 & I_{p - d^*} \end{pmatrix} T^o_1 \begin{pmatrix} Q^o_1 & 0 & 0 \\ 0 & I_{p - d^*} & 0 \\ 0 & 0 & I_{p - d^*} \end{pmatrix}^T.$$
Since $\Gamma^o$ is special orthogonal, $A^o = \log(\Gamma^o)$ is skew-symmetric and given by

$$A = \begin{pmatrix} Q_1^o & 0 \\ 0 & I_{p-d^*} \end{pmatrix} \begin{pmatrix} \log T_1^o & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q_1^o & 0 \\ 0 & I_{p-d^*} \end{pmatrix}^T. $$

Since $A^o$ is skew-symmetric, there exists an $\alpha^o \in \mathbb{R}^{p(p-1)/2}$ such that $A^o = \sum_{m=1}^{p(p-1)/2} \alpha^o_m B_m$. Because the columns of the basis matrices are either zero or one of the canonical basis vectors for $\mathbb{R}^p$, the number of non-zero rows of $A^o$ corresponds to the number $d^{o*}$ of unique elements, up to sign, in the set $\mathcal{L}\{s^*(\alpha^o)\}$. To complete the proof, we need to show that $d^{o*} = d^*$. We proceed along the same lines as in the proof of Proposition 2.2. The main difference is that $T^o$ is not diagonal as it has a number of $2 \times 2$ blocks corresponding to complex conjugate pairs of eigenvalues of $\Gamma^o$. Its blocks of the form $\pm 1$ correspond to the real eigenvalues of $\Gamma^o$. By Lemma 5.4, any $2 \times 2$ block is orthogonal. Specifically it has diagonal entries $c$, upper off-diagonal entry $b$ and lower off-diagonal entry $-b$, where $b, c \in \mathbb{R}$ and $b^2 + c^2 = 1$.

Let $k$ denote the number of unit eigenvalues of $\Gamma^o_1$. We will consider two cases separately: $k = 0$ and $k \geq 1$. Suppose that $k = 0$. Then $\log T^o_1$ has no zero diagonal entries and $A^o_1 \triangleq \log \Gamma^o_1$ has no zero eigenvalues. It follows that $A^o_1$ is of full rank and hence has no zero rows. Thus $d^{o*} = d^*$.

Now suppose that $d^* \geq k \geq 1$. Hence $T^o_1$ has $(d^* - k)/2$, $2 \times 2$ blocks of the form described above, i.e.

$$T^o_1 = \begin{pmatrix} T^o_{11} & 0 \\ 0 & I_k \end{pmatrix},$$

where $T^o_{11} = \text{blockdiag} \left( T^{(d^* - k)/2}_{11}, \ldots, T^{(d^* - k)/2}_{11} \right)$, $T^o_{11} \in \mathbb{R}^{2 \times 2}$. Partition $Q^o_1$ analogously:

$$Q^o_1 = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{13} & Q_{14} \end{pmatrix},$$

where $Q_{11} \in \mathbb{R}^{d^* - k \times d^* - k}$, $Q_{14} \in \mathbb{R}^{k \times k}$, $Q_{12} \in \mathbb{R}^{d^* - k \times k}$, and $Q_{13} \in \mathbb{R}^{k \times d^* - k}$. Then

$$\log \Gamma^o_1 = \begin{pmatrix} Q_{11} \log(T^o_{11})Q^T_{11} & Q_{11} \log(T^o_{11})Q^T_{13} \\ Q_{13} \log(T^o_{11})Q^T_{11} & Q_{13} \log(T^o_{11})Q^T_{13} \end{pmatrix};$$

Firstly, we show that the last $k$ rows of $\log \Gamma^o_1$ are non-zero. Hence we consider the matrix

$$\begin{pmatrix} Q_{13} \log(T^o_{11})Q^T_{11} \\ Q_{13} \log(T^o_{11})Q^T_{13} \end{pmatrix};$$

Note that, due to the form of diagonal blocks of $\log T^o_{11}$ described above

$$Q_{13} \log(T^o_{11})Q^T_{11} = \sum_{i=1}^{(d^* - k)/2} \left( -c_i q_{13\cdot \cdot 2i}^T q_{11\cdot \cdot 2i}^T + c_i q_{13\cdot \cdot 2i}^T q_{11\cdot \cdot 2i}^T \right),$$

$$Q_{13} \log(T^o_{11})Q^T_{13} = \sum_{i=1}^{(d^* - k)/2} \left( -c_i q_{13\cdot \cdot 2i}^T q_{13\cdot \cdot 2i}^T + c_i q_{13\cdot \cdot 2i}^T q_{13\cdot \cdot 2i}^T \right).$$

The last $k$ rows of $\log(\Gamma^o_1)$ are a linear combination of the columns of $Q_1$. For example, the $l$th row
can be written as
\[(d^*-k)/2 \sum_{i=1}^{(d^*-k)/2} (-c_iq_{13\|,2i|,2i-1} + c_iq_{13\|,2i-1,q_{11\|,2i}})^T.\]

Such linear combination is trivial only if \(q_{13\|,l} = 0\), i.e. if the \(l^{th}\) row of \(Q_{13}\) is zero. If the \(l^{th}\) row of \(Q_{13}\) was zero, the \(l^{th}\) row of \(Q_{14}\) would be of unit length and orthogonal to the remaining rows of \(Q_{14}\) and rows of \(Q_{12}\), by the orthonormality of \(Q_1\). This would be in contradiction with \(\Sigma_2^o\) being of a maximum dimension. The linear combination is thus non-trivial and hence must be non-zero as \(Q_1\) is orthogonal. The last \(k\) rows of \(\log(\Gamma_1^o)\) are thus non-zero.

Consider now the first \(d^* - k\) columns of \(\log(\Gamma_1^o)\), i.e. the matrix
\[
\begin{pmatrix}
(Q_{11} \log(T_{11}^o) Q_{11}^T) \\
(Q_{13} \log(T_{11}^o) Q_{13}^T)
\end{pmatrix}
\]

The \(l^{th}\) column of this is
\[
\begin{pmatrix}
\sum_{i=1}^{(d^*-k)/2} -c_iq_{11\|,2i|,2i-1} + c_iq_{11\|,2i-1,q_{11\|,2i}} \\
\sum_{i=1}^{(d^*-k)/2} -c_iq_{13\|,2i|,2i-1} + c_iq_{13\|,2i-1,q_{11\|,2i}} \\
(d^*-k)/2 \\
\end{pmatrix}
\]
which represents a linear combination of the columns of \(Q_1\). The linear combination is trivial only if \(q_{11\|,l}\) is zero. However, if this was the case, \(\Sigma_2^o\) would not be of a maximum dimension. Hence the linear combination in (5.2) is non-trivial. Since \(Q_1\) is orthogonal the sum must be non-zero, and thus the first \(d^* - k\) columns of \(\log(\Gamma_1^o)\) are non-zero. Since \(\log(\Gamma_1^o)\) is skew-symmetric, its first \(d^* - k\) rows are non-zero too.

Thus \(\log(\Gamma_1^o)\) has no zero rows, implying that \(A^o\) has \(p - d^*\) zero rows. Hence \(d^{os} = d^*\), as required.

\[\square\]

### 5.2 Auxiliary lemmas

The key to the relationship between special orthogonal matrices and skew-symmetric matrices is the real Schur decomposition, which holds for any real square matrix.

**Lemma 5.2** (Real Schur Decomposition (Golub and Van Loan, 2013)). If \(M \in \mathbb{R}^{n \times n}\), then there exists an orthogonal \(Q \in \mathbb{R}^{n \times n}\), such that:

\[
Q^T MQ = \begin{pmatrix}
R_{11} & R_{12} & \cdots & R_{1M} \\
R_{22} & \cdots & R_{2M} \\
\vdots & \ddots & \ddots \\
R_{MM} & \cdots & R_{MM}
\end{pmatrix} \equiv R,
\]

where each \(R_{ii}\) is either a 1-by-1 matrix or a 2-by-2 matrix having complex conjugate eigenvalues. We refer to \(R\) as the real Schur form of \(M\).
Since the eigenvector matrix $\Gamma$ is orthogonal, so too is the real Schur form of $\Gamma$. This leads to the following refinement of the real Schur decomposition for orthogonal matrices.

**Lemma 5.3.** Let $\Gamma \in \mathbb{R}^{p \times p}$ be an orthogonal matrix. Then the real Schur form of $\Gamma$ is block-diagonal, with diagonal blocks either $1 \times 1$ or $2 \times 2$ matrices with complex conjugate eigenvalues.

**Lemma 5.4.** Let $\Gamma \in \mathbb{R}^{p \times p}$ be a special orthogonal matrix. Let $T$ be the real Schur form of $\Gamma$. Then the diagonal blocks of $T$ can be of the following types: (i) $T_{ii} = 1$; (ii) $T_{ii} = -1$; (iii) $T_{ii} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$.

There are an even number of blocks of type (ii).

**Acknowledgement:** We are grateful to two referees for their valuable comments.

**Funding statement:** The work was partially supported by the U.K. Engineering & Physical Sciences Research Council.

**References**


