Biometrika (0000), **000**, 0, *p.* 1
Printed in Great Britain Advance Access publication on 16 June 1904

Regression graphs and sparsity-inducing reparametrizations

By J. RYBAK

Department of Mathematics, Imperial College London,
180 Queen's Gate, London SW7 2AZ, U.K.
jakub.rybak18@imperial.ac.uk

H. S. BATTEY

Department of Mathematics, Imperial College London,
180 Queen's Gate, London SW7 2AZ, U.K.
h.battey@imperial.ac.uk

and K. BHARATH

School of Mathematical Sciences, University of Nottingham,
University Park, Nottingham NG7 2RD, U.K.
karthik.bharath@nottingham.ac.uk

Summary

That parametrization and sparsity are inherently linked raises the possibility that relevant models, not obviously sparse in their natural formulation, exhibit a population-level sparsity after reparametrization. In covariance models, positive-definiteness enforces additional constraints on how sparsity can legitimately manifest. It is therefore natural to consider reparametrization maps in which sparsity respects positive definiteness. The paper provides insight into structures on the physically-natural scale that induce and are induced by sparsity after reparametrization. Of the four structures initially uncovered, the richest can be generated, under a causal ordering, by the joint-response graphs studied by Wermuth & Cox (2004). This connection leads to an interpretation of approximate zeros and explains modelling implications of enforcing sparsity after reparameterization: in effect, the relation between two variables would be declared null if relatively direct regression effects were negligible and other effects manifested through long paths. The Iwasawa decomposition of the general linear group, combined with the graphical-models interpretation, points to a class of reparametrizations for the chain-graph models (Andersson et al., 2001), with undirected and directed acyclic graphs as special cases. The insights have a bearing on methodology, some aspects of which are developed. An extensive simulation uses the theoretical insights to further explore regimes under which reparametrization is beneficial.

Some key words: Causality; Chain graphs; Graphical models; Matrix logarithm; Reparametrization; Sparsity.

1. Introduction

Sparsity, the existence of many zeros or near-zeros in some domain, plays at least two roles in statistics, depending on context: to aid interpretation and to prevent accumulation of error incurred through estimation of nuisance parameters. There is now a large literature concerned with enforcing sparsity on sample quantities, having assumed that the corresponding population-level object is sparse. The present paper is concerned with the more fundamental question of whether there are parametrizations enjoying a population-level sparsity not present to the same extent in the original formulation. In other words, from a parametrization that is natural from a modelling point of view, we seek a sparsity-inducing reparametrization.

Inducement of population-level sparsity, through a traversal of parametrization space or data-transformation space, is a relatively unexplored area. Battey (2023) unified four isolated examples from this perspective, starting from the work on parameter orthogonalization (Cox and Reid, 1987). The development of Gaussian graphical models (e.g. Lauritzen, 1996; Cox & Wermuth, 1996) and graphical models for extremes (Engelke & Hitz, 2020) is also somewhat in this vein. In the same spirit, we focus on interpretation and insight at the population level, leaving for future development the important question of how to deduce the sparsity scale empirically.

The motivating question for this paper is whether, for a broad enough class of covariance structures, not obviously sparse in their natural parameter domain, a non-trivial sparsity-inducing reparametrization can be deduced in which sparsity respects positive definiteness. By non-trivial, we mean that it is possible to discriminate more effectively on the new scale between elements that are large and elements that are small. This rules out artificially sparse reparametrizations such as $\Sigma \mapsto c\Sigma$ for $c > 0$ close to zero. Battey (2017) and Rybak & Battey (2021) provided a proof of concept. Their position was that covariance matrices and their inverses are often nuisance parameters, and it is therefore arguably more important that the sparsity holds to an adequate order of approximation in an arbitrary parametrization, than that the sparse parametrization has interpretable zeros. An example of this type is linear discriminant analysis, where the interest parameter is the linear discriminant. In the case of undirected Gaussian graphical models, the precision matrix is the interest parameter by virtue of the interpretation ascribed to its zeros. Thus, both aspects are of interest and are addressed here. A third and different type of situation is when the covariance matrix is a nuisance parameter that has a known structure up to a low-dimensional parameter. This is common in some settings, for instance in the analysis of split-plot or Latin square designs with block effects treated as random.

The starting point for the paper is the identification of new parametrizations in which sparsity conveniently manifests in a vector space. For these, we uncover the structure induced on the original scale through zeros in the new parameter domain, as well as the converse result: that matrices encoding such structure possess exact zeros after reparametrization. The scope is considerably broadened through the possibility of approximate zeros, of which there may be many more in the new parameter domain than in the original or inverse domains. An important insight is therefore the interpretation of approximate zeros, as this explains the modelling implications of enforcing sparsity after reparameterization. Under a, perhaps notional, causal ordering, the relation between two variables would be declared null if relatively direct regression effects were negligible and other effects manifested through long paths. Section 7 unifies old and new parametrizations via a class of matrix decompositions representing the chain graphs, allowing for both directed and undirected edges, and recovering the four fundamental parametrizations as special cases.

Because the population-level sparsity manifests in a vector space, any sensible estimator exploiting the sparsity will respect positive definiteness. We present one approach with high-dimensional statistical guarantees in §8.

2. Notation

Table 1 indicates subsets of the vector space $\mathsf{M}(p)$ of $p \times p$ real matrices. Most are matrix Lie groups with matrix multiplication as the group operation; those that are vector spaces have a natural matrix basis. A generic vector subspace of $\mathsf{M}(p)$ is written $\mathsf{V}(p)$.

Notation	Matrix subset	Basis notation
$\mathsf{PD}(p)$	symmetric positive definite matrices	
$\mathsf{Sym}(p)$	symmetric matrices	$\mathcal{B}_{sym}$
$\mathsf{D}(p)$	diagonal matrices	$\mathcal{B}_{diag}$
$\mathsf{GL}(p)$	nonsingular matrices	
$\mathsf{P}(p)$	permutation matrices	
$\mathsf{O}(p)$	orthogonal matrices	
$\mathsf{SO}(p)$	special orthogonal matrices (determinant +1)	
$\mathsf{Sk}(p)$	skew-symmetric matrices	$\mathcal{B}_{sk}$
$\mathsf{LT}(p)$	lower-triangular matrices	$\mathcal{B}_{lt}$
$\mathsf{LT}_{\mathsf{u}}(p)$	lower-triangular with unit diagonal entries	
$\mathsf{LT}_{\mathsf{s}}(p)$	strictly lower-triangular matrices	$\mathcal{B}_{lts}$

Table 1: Matrix subsets of $\mathsf{M}(p)$.

Also extensively referenced is $\mathsf{Cone}_p \subset \mathsf{Sym}(p)$, the interior of a convex cone within $\mathsf{Sym}(p)$ excluding the origin, as formalized in Appendix B of the supplementary material. For the purpose of the present paper, Cone_p can be thought of as the constrained set of $p(p+1)/2$ elements constituting the upper triangular part of a positive definite matrix.

Diagonal and lower triangular matrices with positive diagonal elements are differentiated using the subscript +. The symbol $\oplus$ denotes the direct sum of two vector spaces; $A \oplus B$ also represents a block-diagonal matrix with blocks A and B. The index set $\{1, \dots, p\}$ is written $[p]$. The length of a vector v is written $\dim(v)$ and the cardinality of a finite set $\mathcal{A}$ is written $|\mathcal{A}|$.

The sets of basis matrices in Table 1 are constructed from the canonical basis vectors $e_1, \dots, e_p$ for $\mathbb{R}^p$, where $e_i \in \mathbb{R}^p$ is a zero vector with 1 as its ith component. Specifically $\mathcal{B}_{sym} := \{B_1, \dots, B_{p(p+1)/2}\}$ consists of $p(p-1)/2$ non-diagonal matrices $e_j e_k^{\mathrm{T}} + e_k e_j^{\mathrm{T}}$ for $j < k$ and p diagonal matrices of the form $e_j e_j^{\mathrm{T}}$, the latter also constituting $\mathcal{B}_{diag}$; $\mathcal{B}_{sk} := \{B_1, \dots, B_{p(p-1)/2}\}$ consists of skew symmetric matrices $e_j e_k^{\mathrm{T}} - e_k e_j^{\mathrm{T}}$ for $j < k$; $\mathcal{B}_{lt} := \{B_1, \dots, B_{p(p+1)/2}\}$, consists of lower triangular matrices $e_k e_j^{\mathrm{T}}$, $j \leq k$; and $\mathcal{B}_{lts} := \{B_1, \dots, B_{p(p-1)/2}\}$ consists of strictly lower triangular matrices $e_k e_j^{\mathrm{T}}$ with $j < k$. The matrix exponential of a square matrix A is defined as $e^A = \sum_{k=0}^{\infty} A^k/k!$. Conversely, if a matrix logarithm L of a square matrix M exists, then $M = e^L$. See Appendix C for existence and uniqueness conditions.

For random variables X_1, X_2 and X_3, the statement that X_1 is conditionally independent of X_2 given X_3 is notated by $X_1 \perp\!\!\!\perp X_2 | X_3$, unconditional independence notated by $X_1 \perp\!\!\!\perp X_2$.

3. Reparametrization

The set Cone_p is, from one perspective, the natural parameter domain for parametrizing the manifold $\mathsf{PD}(p)$. The question we seek to address is whether there is another parameter domain that is less direct, but in which a population-level sparsity is present, ideally with interpretable zeros or near-zeros. This is most compelling in the absence of considerable sparsity on the original or inverse scales; in that case, the reparametrization is said to be sparsity-inducing. The problem is initially addressed from the opposite direction, by considering the parameter domains in which sparsity can be fruitfully represented, and then studying the form of multivariate dependencies that are implied by exact zeros in these non-standard parametrizations. We clarify in subsequent sections the extent to which, and manner in which, a covariance or precision matrix might be sparser after reparametrization, and the implications for estimation and interpretation.

In a sense to be clarified, a construction based on the chain graph representations of §7 subsumes the dependence structures identified in §4 into a broader sparsity class. This ultimately points to a class of structures in which sparsity manifests in a vector space, and that enjoy a graphical models interpretation on the physically natural scale when a full or partial causal ordering is present.

From an initial parameterization of $\mathsf{PD}(p)$, formalized in Appendix B, we study four reparameterizations arising from maps $\mathsf{Cone}_p \to \mathbb{R}^{p(p+1)/2}$ such that sparsity in the new domain $\mathbb{R}^{p(p+1)/2}$ respects the positive definiteness constraint on Σ. In other words, an arbitrary configuration of zeros in the new parameter domain $\mathbb{R}^{p(p+1)/2}$ does not violate positive definiteness of Σ or Σ^{-1}. The four fundamental parametrizations discussed in this work are, with $D(d) = \mathrm{diag}(d_1, \ldots, d_p)$,

$$\begin{aligned}
\alpha \mapsto \Sigma_{pd}(\alpha) &:= e^{L(\alpha)}, & L(\alpha) &\in \mathsf{Sym}(p), & \alpha &\in \mathbb{R}^{p(p+1)/2};\\
(\alpha, d) \mapsto \Sigma_{o}(\alpha, d) &:= e^{L(\alpha)} e^{D(d)} (e^{L(\alpha)})^{\mathrm{T}}, & L(\alpha) &\in \mathsf{Sk}(p), & \alpha &\in \mathbb{R}^{p(p-1)/2},\ d \in \mathbb{R}^{p};\\
\alpha \mapsto \Sigma_{lt}(\alpha) &:= e^{L(\alpha)} (e^{L(\alpha)})^{\mathrm{T}}, & L(\alpha) &\in \mathsf{LT}(p), & \alpha &\in \mathbb{R}^{p(p+1)/2};\\
(\alpha, d) \mapsto \Sigma_{ltu}(\alpha, d) &:= e^{L(\alpha)} e^{D(d)} (e^{L(\alpha)})^{\mathrm{T}}, & L(\alpha) &\in \mathsf{LT}_{\mathsf{s}}(p), & \alpha &\in \mathbb{R}^{p(p-1)/2},\ d \in \mathbb{R}^{p}.
\end{aligned} \tag{1}$$

That the four maps (1) are fundamental emerges from the Iwasawa decomposition of the group of nonsingular matrices, owing to which the paper has an enlightening group-theoretic underpinning. We have placed most of this discussion in the supplementary material in favour of a more broadly accessible exposition, but we return briefly to the Iwasawa decomposition in §7.2.

For each of the four reparametrization maps, the parameter domain $\mathbb{R}^{p(p+1)/2}$ is identified with a different vector space of the same dimension. These are, respectively, $\mathsf{Sym}(p)$, $\mathsf{Sk}(p) \times \mathsf{D}(p)$, $\mathsf{LT}(p)$, and $\mathsf{LT}_{\mathsf{s}}(p) \times \mathsf{D}(p)$. The subscripts on Σ in the parametrizations indicate which of the matrix sets, $\mathsf{PD}(p)$, $\mathsf{SO}(p)$, $\mathsf{LT}_{+}(p)$ and $\mathsf{LT}_{\mathsf{u}}(p)$ respectively, prescribed coordinates in terms of α, are represented as the image of the matrix exponential. In each case $L(\alpha) \in \mathsf{V}(p)$ depends on α through its expansion

$$L(\alpha) = \alpha_1 B_1 + \cdots + \alpha_m B_m \tag{2}$$

in the canonical basis for $\mathsf{V}(p)$, as specified in §2. The canonical basis is part of the definition of the reparametrization maps. Appendices B and C establish the legitimacy of the maps, this hinging in the case of Σ_o and Σ_{lt} on some constraints on α or conditions on the covariance matrices. Thus, Σ_{pd} and Σ_{ltu} are favourable in this respect.

Another parametrization in which sparsity respects positive definiteness is in terms of the Cholesky factors themselves, rather than the matrix logarithm of the Cholesky factors. This is related to the Σ_{lt} and Σ_{ltu} parametrizations as discussed in §5. While sparsity in the Cholesky factors has not been explicitly considered, its implications can be deduced from Wermuth & Cox (2004). Several authors have modelled the Cholesky components in terms of covariates; Pourahmadi (1999) appears to have started this line of enquiry.

4. Sparsity structures of $\Sigma(\alpha)$ induced by, and inducing, exact zeros in α

Consider the matrices $L \in \mathsf{V}(p) \subset \mathsf{M}(p)$ from (1), all of which can be written in terms of a canonical basis $\mathcal{B}$ of dimension m as $L(\alpha) = \alpha_1 B_1 + \cdots + \alpha_m B_m$. Suppose now that $\alpha = (\alpha_1, \ldots, \alpha_m)$ is sparse in the sense that $\|\alpha\|_0 = \sum_j \mathbb{I}\{\alpha_j \neq 0\} = s^* \ll m$. In general, the sparsity of $L(\alpha) = \log(M)$ is different from the sparsity of M. However, certain sparsity structures are necessarily preserved in both directions, in the sense that particular arrangements of exact zeros in M and its logarithmic transformation coincide. These structures, and the corresponding structures of Σ, are specified in Corollaries 1-4. One might call these zeros *structural zeros*: ones that are preserved through transformation regardless of the values of the non-zero entries. There is also the possibility of *coincidental zeros* in the logarithmic domain that are not present in the original domain. This can be seen most easily from the Σ_{pd} parametrisation. For $s^* < p$, Battey (2017) showed that $\Sigma = \Sigma_{pd}(\alpha)$ is necessarily of the form

$$\Sigma = \sum_{j \in \mathcal{A}} \lambda_j o_j o_j^{\mathrm{T}} + \sum_{j \in \mathcal{A}^c} e_j e_j^{\mathrm{T}}, \quad \|o_j\|_0 = p - |\mathcal{A}^c|, \tag{3}$$

where $(\lambda_j, o_j)_{j=1}^p$ are the pairs of eigenvalues of Σ and $\mathcal{A}$ is a set specified by the configuration of zeros in α. The implication of (3), since the second sum only specifies unit entries in diagonal positions corresponding to $\mathcal{A}$, is that if

$$L_{ik} = \sum_{j \in \mathcal{A}} \log \lambda_j o_{ij} o_{kj} = 0 \tag{4}$$

for positions i and k such that o_{ij} and o_{kj} are not identically zero over $j \in \mathcal{A}$, then the corresponding entry Σ_{ik} is necessarily non-zero. This coincidental zero in the logarithm comes from cancellation, as distinct from the structural zeros in the eigenvectors. Equation (4) illustrates clearly that nothing is lost by transforming to the matrix logarithmic domain, discarding eigenpairs used for the calculation, and transforming back: even when coincidental zeros are encountered in $L = \log(\Sigma)$, the corresponding entries of Σ are recovered from the ensemble.

In Gaussian graphical models, the structures expounded in Corollaries 1-4 correspond to conditional and unconditional independencies between specific sets of variables. Since identical patterns of zeros are present in transformed matrices, these sparsity structures, when present in the transformed domain, imply equivalent statements about conditional and unconditional independence. It is possible, however, to make additional statements from approximate zeros. We revisit this aspect in §5 and §6.

Theorem 1 is a general result, whose application to the four cases results in Corollaries 1-4.

THEOREM 1. *Let $M = e^L \in \mathsf{M}(p)$ where $L \in \mathsf{V}(p)$, a vector space with canonical basis $\mathcal{B}$ of dimension m. The matrix L has d_r^* and d_c^* non-zero rows and columns, of which d^* coincide after transposition, if and only if M has $p - d_r^*$ rows of the form e_j^{T} for some $j \in [p]$, all distinct, $p - d_c^*$ columns of the form e_j, and of these, $p - d^*$ coincide after transposition. If M is normal, i.e. $M^{\mathrm{T}}M = MM^{\mathrm{T}}$, then $d_r^* = d_c^* = d^*$.*

The quantities d_r^*, d_c^* and d^* are related to s^* when α is sparse. A loose bound is $\max\{d_r^*, d_c^*\} \leq 2s^*$, but $\max\{d_r^*, d_c^*\}$ can be considerably smaller than this, as it depends on the configuration of basis elements picked out by the sparse α. Indeed, $\max\{d_r^*, d_c^*\} \ll p$ is possible even when s^* exceeds p, provided that the configuration of non-zero elements of α produces zero rows or columns of L. Figure 1 shows an example of a structure of M established by Theorem 1. In particular settings, where the form of $\mathsf{V}(p)$ is made explicit, there may be additional structure, e.g. lower triangular, that is not reflected in Figure 1.

Corollaries 1 and 2 to be presented are not new. However, their proofs in Appendix G are new, and presented in terms of the general formulation of Theorem 1.

COROLLARY 1. *Let Σ be parametrized as $\Sigma_{pd}(\alpha) = e^{L(\alpha)}$ and let $d < p$. Then, $\Sigma_{pd}(\alpha)$ is of the form $\Sigma = P(\Sigma_1 \oplus D_{p-d})P^{\mathrm{T}}$, where $P \in \mathsf{P}(p)$ is a permutation matrix, $\Sigma_1 \in \mathsf{PD}(d)$ and $D_{p-d} \in \mathsf{D}(p-d)$, if and only if $L(\alpha) = P(L_1 \oplus \Delta_{p-d})P^{\mathrm{T}}$, where $L_1 \in \mathsf{Sym}(d)$ and $\Delta_{p-d} \in \mathsf{D}(p-d)$.*

Corollary 1 as stated emerges from the properties of the matrix logarithm applied to block diagonal matrices. The version of Battey (2017) gives a stronger restriction in that s^* is required

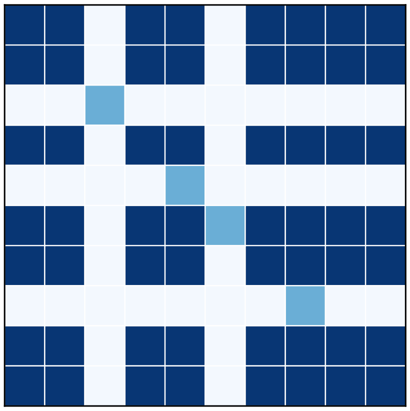


Fig. 1: Example of a structure of M as established in Theorem 1 with $p = 10$, $d_r^* = 7$, $d_c^* = 8$ and $d^* = 9$. The entries that are zero by Theorem 1 are light blue, those equal to one are medium blue, and the remaining entries, whose values are unconstrained, are dark blue.

to be less than p which also reduces the rank of the matrix logarithm. In that case D_{p-d} and Δ_{p-d} from Corollary 1 are replaced by I_{p-d^*} and 0_{p-d^*}, as reflected in (3).

Corollary 2. *For an arbitrary diagonal component $D = \text{diag}\{d_1, \ldots, d_p\}$, let Σ be parametrized as $\Sigma_o(\alpha) = e^{L(\alpha)} e^{D} (e^{L(\alpha)})^{\mathrm{T}}$. Then, $\Sigma_o(\alpha)$ is of the form $\Sigma = P(\Sigma_1 \oplus D_{p-d^*})P^{\mathrm{T}}$, where $P \in \mathsf{P}(p)$ is a permutation matrix, $D_{p-d^*} \in \mathsf{D}(p - d^*)$ and $\Sigma_1 \in \mathsf{PD}(d^*)$, if and only if $L(\alpha) = P(L_1 \oplus 0_{p-d^*})P^{\mathrm{T}}$, where $L_1 \in \mathsf{Sym}(d^*)$.*

Corollary 3. *Let Σ be parametrized as $\Sigma_{lt}(\alpha) = e^{L(\alpha)}(e^{L(\alpha)})^{\mathrm{T}}$. Then, $\Sigma_{lt}(\alpha)$ is of the form $\Sigma = VV^{\mathrm{T}}$, where $V = I_p + \Theta$ with $\Theta \in \mathsf{LT}_+(p)$. The ith row of Θ is zero if and only if the ith row of $L(\alpha)$ is zero. Similarly, the jth column of Θ is zero if and only if the jth column of $L(\alpha)$ is zero.*

Corollary 4. *For an arbitrary diagonal component $D = \text{diag}\{d_1, \ldots, d_p\}$, let Σ be parametrized as $\Sigma_{ltu}(\alpha) = e^{L(\alpha)} e^{D} (e^{L(\alpha)})^{\mathrm{T}}$. Then, $\Sigma_{ltu}(\alpha)$ is of the form $\Sigma = U\Psi U^{\mathrm{T}}$, where $\Psi = e^{D} \in \mathsf{D}_+(p)$, $U = I_p + \Theta$. The ith row of Θ is zero if and only if the ith row of $L(\alpha)$ is zero. The jth column of Θ is zero if and only if the jth column of $L(\alpha)$ is zero.*

Zeros in α produce structured patterns of zeros in Σ and Σ^{-1} in parametrizations Σ_{pd} and Σ_o, and structured patterns of zeros in the Cholesky factors in parametrizations Σ_{lt} and Σ_{ltu} respectively. These structures are interpreted in §5.

Although constraints on s^* are avoided in Theorem 1, a small value, e.g. $s^* < p/2$, is guaranteed both to generate and to be implied by a simplification in the underlying conditional independence graph, under a notional Gaussian model. Relatively large values of s^* can also entail graphical reduction in many cases, in the sense of introducing conditional independence relations relative to the saturated case. As an illustration, there are $p(p + 1)/2$ basis elements for L in the Σ_{pd} parametrization. For α to induce a pattern of zeros in Σ of the type discussed in Corollary 1, L needs to have a zero row, which requires only p zero coefficients in the basis expansion of L. Thus, s^* can be as large as $s^* = p(p - 1)/2$ for the structure to hold.

To make a comparison between different structures of Σ more explicit, we consider a simple example with $p = 5$. For the parametrizations Σ_{pd} and Σ_o we set $d^* = 3$, corresponding to $s^* = 6$ and $s^* = 3$ respectively. For Σ_{ltu} we consider two cases: Σ_{ltu}^r, for which $d_r^* < p$, $d_c^* = p$ (this serving as the definition of Σ_{ltu}^r), and Σ_{ltu}^c, for which $d_r^* = p$, $d_c^* < p$; in both cases $s^* = 6$. The

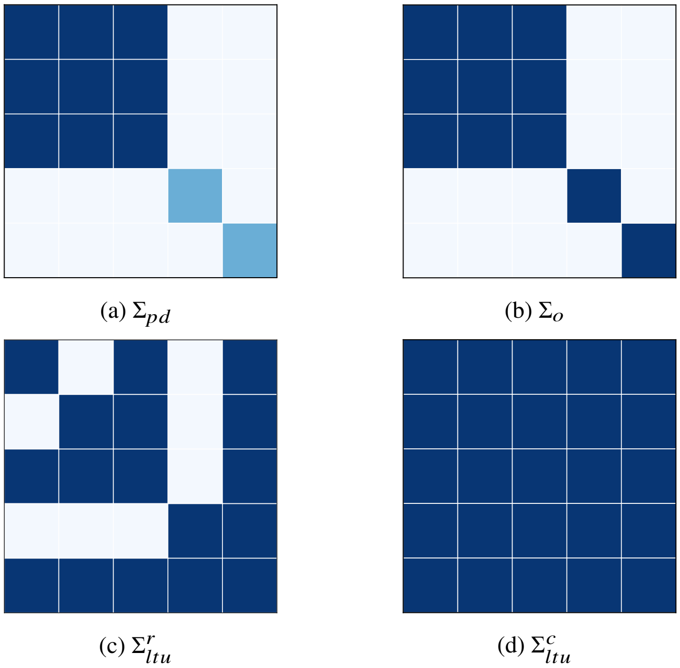

(a) Σ_{pd}
(b) Σ_o
(c) Σ^r_{ltu}
(d) Σ^c_{ltu}

Fig. 2: Structure of $\Sigma(\alpha)$ induced by sparsity of α. Zero entries are depicted by light blue, unit entries by medium blue, and the unrestricted entries by dark blue.

quantities d^*, d^*_r and d^*_c are defined in Theorem 1. The resulting covariance structure is depicted in Figure 2. If, for an arbitrary diagonal component, the map $\alpha \mapsto \Sigma_{ltu}(\alpha)$ is instead replaced by an essentially equivalent representation $\alpha \mapsto \Sigma_{utu}(\alpha)$ in terms of upper triangular matrices, the analogous structures $\Sigma^r_{utu}(\alpha)$ and $\Sigma^c_{utu}(\alpha)$ are the same as for $\Sigma^c_{ltu}(\alpha)$ and $\Sigma^r_{ltu}(\alpha)$ respectively.

For the Σ_{ltu} parametrization, Figure 2 illustrates that, unlike a Θ with zero rows, a Θ with zero columns can generate a dense covariance matrix. Intuitively, for the same restriction on the sparsity of α, the corresponding covariance matrices Σ^r_{ltu} and Σ^c_{ltu} should represent relationships of similar inherent structural complexity. The following result confirms this intuition. Specifically, Lemma 1 shows that, although Σ^c_{ltu} might have no zeros, the sparsity restriction on α induces a low-rank structure on a submatrix of Σ^c_{ltu}. The existence of a low-rank structure has a statistical interpretation in terms of latent variables (e.g. Fan et al, 2013).

Lemma 1. *Consider a random vector* $Y = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}}, Y_3^{\mathrm{T}})^{\mathrm{T}}$ *with covariance matrix* Σ. *The columns of* Θ *in Corollary 4 corresponding to* Y_2 *are zero if and only if the submatrix*

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{23} \end{pmatrix}$$

of Σ *has rank* $\dim(Y_1)$.

If zeros in d are allowed, the basis coefficients of the $\alpha \mapsto \Sigma_{lt}(\alpha)$ and $(\alpha, d) \mapsto \Sigma_{ltu}(\alpha, d)$ parametrizations are related. To see this, write

$$\Sigma_{lt} = \exp(L)\exp(L^{\mathrm{T}}) = \exp(L_u D_1)\exp((L_u D_1)^{\mathrm{T}}),$$

where $D_1 \in \mathsf{D}(p)$ and $L_u \in \mathsf{LT}_\mathsf{u}(p)$. On writing $L_u D_1 = L_s + D_1$, where $L_s \in \mathsf{LT}_\mathsf{s}(p)$ contains the strictly lower triangular part of $L_u D_1$,

$$\Sigma_{lt} = \exp(L)\exp(L^{\mathrm{T}}) = \exp(L_s)\exp(D_1 + D_1^{\mathrm{T}})\exp(L_s^{\mathrm{T}}), \tag{5}$$

by the properties of the matrix exponential, which recovers the Σ_{ltu} parametrization with $D = 2D_1$. Thus if d is allowed to have zeros, there is an exact relationship between the coefficients of the expansion of Σ_{ltu} and those of Σ_{lt}. Since the transformation Σ_{ltu} provides a more convenient way of parametrizing regression graphs, subsequent discussion focuses on Σ_{ltu}. Most insights derived for the Σ_{ltu} parametrization extend directly to Σ_{lt}.

5. Sparsity under the Σ_{ltu} parametrization

5.1. *Causal ordering*

A familiar result interprets zeros in a precision matrix as conditional independencies under a Gaussianity assumption. The less familiar directed graphical models have important differences, both mathematically and conceptually. For instance, many different causal models may be compatible with the same structure of zeros in the precision matrix, and an undirected graph whose associated Gaussian model has a sparse precision matrix could be appreciably less sparse in Σ^{-1} when the undirected edges are replaced by directed ones. The key factor determining this is whether there are common response variables occurring later in the causal ordering.

Whether directed or undirected edges are more natural depends on context. The present section is concerned with directed edges. By postulating a, perhaps notional or provisional, causal ordering among the underlying random variables, substantive understanding can be attached to the interpretation of sparsity on the transformed scale. Through this route we develop insight into the implicit assumptions involved in enforcing sparsity when it is only approximately present, broadening the scope of the work.

5.2. *The matrix logarithm and weighted causal paths*

With $[p] = \{1, \ldots, p\}$, let $a \subset [p]$ and $b = [p] \backslash a$ be disjoint subsets of variable indices. As a consequence of a block-diagonalization identity for symmetric matrices (Cox & Wermuth, 1993; Wermuth & Cox, 2004),

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} = \begin{pmatrix} I_{|a|} & 0 \\ \Sigma_{ba}\Sigma_{aa}^{-1} & I_{|b|} \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & 0 \\ 0 & \Sigma_{bb.a} \end{pmatrix} \begin{pmatrix} I_{|a|} & \Sigma_{aa}^{-1}\Sigma_{ab} \\ 0 & I_{|b|} \end{pmatrix}. \tag{6}$$

The components $\Pi_{b|a} := \Sigma_{ba}\Sigma_{aa}^{-1} \in \mathbb{R}^{|b|\times|a|}$, $\Sigma_{aa} \in \mathsf{PD}(|a|)$ and $\Sigma_{bb.a} := \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \in \mathsf{PD}(|b|)$ are the so-called partial Iwasawa coordinates for $\mathsf{PD}(p)$ based on a two-component partition $|a| + |b| = p$ of p. For a statistical interpretation, let $Y = (Y_a^{\mathrm{T}}, Y_b^{\mathrm{T}})^{\mathrm{T}}$ be a mean-zero random vector with covariance matrix Σ. Then $\Pi_{b|a}$ is the matrix of regression coefficients of Y_a in a linear regression of Y_b on Y_a, and $\Sigma_{bb.a}$ is the error covariance matrix, i.e. $Y_b = \Pi_{b|a}Y_a + \varepsilon_b$ and $\Sigma_{bb.a} = \mathrm{var}(\varepsilon_b)$. Applying the block-diagonalization identity recursively results in a block-diagonalization in 1×1 blocks, which corresponds to the LDL decomposition of Σ inherent to the Σ_{ltu} parametrization. Specifically, $\Sigma = U\Psi U^{\mathrm{T}}$ with $\Psi \in \mathsf{D}_+(p)$ and

$$U = I_p + \Theta = (I_p - B)^{-1}, \tag{7}$$

where $\Theta \in \mathsf{LT}_\mathsf{s}(p)$ and B_{ij} is the regression coefficient of Y_j in a regression of Y_i on its predecessors $Y_1, \ldots, Y_{i-1}$. Although in principle an arbitrary ordering can be chosen, it is natural to use a postulated causal ordering, if one is available. In the corresponding representation of Y as a directed acyclic graph with nodes $Y_1, \ldots, Y_p$, a directed edge $Y_j \longrightarrow Y_i$ can exist only if $j < i$, in

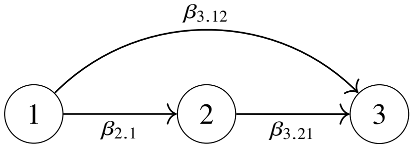


Fig. 3: Directed acyclic graph with edge weights corresponding to regression coefficients.

which case the total effect of Y_j on Y_i can be expressed in terms of the regression coefficients. An example gives intuition prior to a formal statement in Proposition 1.

Example 1. Consider a set of three variables (Y_1, Y_2, Y_3). The total effect of Y_1 on Y_3 is related to the conditional effects through Cochran's recursion (Cochran, 1938), also known as the trek rule,

$$\beta_{3.1} = \beta_{3.12} + \beta_{3.21}\beta_{2.1}. \tag{8}$$

The coefficient $\beta_{3.1}$ is the regression coefficient of Y_1 in a regression of Y_3 on Y_1 only, having marginalized over Y_2, while $\beta_{3.12}$ is the coefficient of Y_1 in a regression of Y_3 on Y_1 and Y_2. To make this concrete at the population level, $\beta_{3.1}$ is the total derivative of

$$f(y_1, \bar{y}_2) := \mathbb{E}(Y_3 \mid Y_1 = y_1, Y_2 = \bar{y}_2),$$

treating y_1 and $\bar{y}_2 = \bar{y}_2(y_1) = \mathbb{E}(Y_2 \mid Y_1 = y_1)$ as free variables, i.e.

$$\beta_{3.1} = \frac{Df(y_1, \bar{y}_2)}{Dy_1} = \frac{\partial f(y_1, \bar{y}_2)}{\partial y_1} + \frac{\partial f(y_1, \bar{y}_2)}{\partial \bar{y}_2}\frac{d\bar{y}_2(y_1)}{dy_1}.$$

The right-hand side of (8) corresponds to tracing the effects of Y_3 on Y_1 along two paths connecting the nodes in a recursive system of random variables (Y_1, Y_2, Y_3), with edge weights given by the corresponding regression coefficients, as depicted in Figure 3.

A directed edge in a recursive system can exist from node j to node i only if $j < i$. Thus, there are two possible paths from Y_1 to Y_3: $Y_1 \longrightarrow Y_3$ and $Y_1 \longrightarrow Y_2 \longrightarrow Y_3$, which correspond to the first and second term in (8) respectively. Let $\upsilon_{ij}(l)$ denote the effect of Y_j on Y_i along all paths of length l, specified for this three-dimensional system as

$$\begin{aligned} \upsilon_{21}(1) &= \beta_{2.1}, & \upsilon_{31}(1) &= \beta_{3.12}, \\ \upsilon_{32}(1) &= \beta_{3.21}, & \upsilon_{31}(2) &= \beta_{3.21}\beta_{2.1}. \end{aligned}$$

The lower-triangular matrices U and $L = \log(U)$ have the form,

$$U = \begin{pmatrix} 1 & 0 & 0 \\ \beta_{2.1} & 1 & 0 \\ \beta_{3.12} + \beta_{3.21}\beta_{2.1} & \beta_{3.21} & 1 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{2.1} & 0 & 0 \\ \beta_{3.12} + \frac{\beta_{3.21}\beta_{2.1}}{2} & \beta_{3.21} & 0 \end{pmatrix}. \tag{9}$$

More generally, the following proposition establishes the interpretation of entries of the lower-triangular matrix U and its matrix logarithm, L.

PROPOSITION 1. *Consider a parametrization* $\Sigma_{ltu}(\alpha, d) = e^{L(\alpha)}e^{D(d)}(e^{L(\alpha)})^{\mathrm{T}}$ *and let* $U = \exp(L(\alpha))$. *Let* $\beta_{i.j[k]}$ *denote the regression coefficient on* Y_j *in a regression of* Y_i *on* Y_j *and*

$Y_1, \ldots, Y_k$. The (i, j)th elements of U and L have the form,

$$U_{ij} = \begin{cases} 0 & \textit{if} \quad i < j, \\ 1 & \textit{if} \quad i = j, \\ \sum_{l=1}^{p-1} \upsilon_{ij}(l) & \textit{if} \quad i > j, \end{cases} \qquad L_{ij} = \begin{cases} 0 & \textit{if} \quad i \leq j \\ \sum_{l=1}^{p-1} \frac{\upsilon_{ij}(l)}{l} & \textit{if} \quad i > j, \end{cases}$$

where

$$\upsilon_{ij}(l) = \begin{cases} \sum_{k=j+l-1}^{i-1} \beta_{i.k[i-1]} \upsilon_{kj}(l-1) & \textit{if} \quad i - j \geq l, \\ 0 & \textit{otherwise}. \end{cases}$$

The entry U_{ij} for $j < i$ thus corresponds to the sum of effects of Y_j on Y_i along all paths connecting the two nodes, with edge weights given by regression coefficients. In contrast, L_{ij}, and by extension, the corresponding coefficient in the basis expansion of L, is equal to the weighted sum of effects of Y_j on Y_i along all paths connecting the two nodes, with weights inversely proportional to the length of the corresponding path.

Proposition 1 provides insight into the effect of logarithmic transformation relative to the identity transformation and the inverse transformation, whose resulting zeros encapsulate conditional independencies in a Gaussian model. Specifically, the entries of B can be viewed as representing paths of length 1, corresponding to a complete discounting of longer paths, while $U = (I - B)^{-1}$ has entries aggregating the contributions along all paths, with weights equal to one, i.e. no discounting of longer paths. In between these two extremes, logarithmic transformation weights a path of length l by a factor $1/l$, as reflected in Proposition 1. Moreover, the weights in the logarithmic transformation are such that the off-diagonal entries of $\log(I - B)$ and $\log((I - B)^{-1})$ are equal in absolute value, since $\log((I - B)^{-1}) = -\log(I - B)$. Section 5.4 discusses some of the implications of these distinctions, following a discussion of exact zeros in §5.3.

5.3. *Exact zeros*

The previous discussion makes clear that there can be configurations of zeros in α that do not produce whole rows or columns of zeros in L. In that case, no structural insights are available from Corollary 4, although an interpretation is still available for any exact zero of L via Proposition 1. Corollary 5 provides the relationship between exact zeros in U and L under a causal ordering.

COROLLARY 5. *If no directed path exists from node j to node i, $j < i$, then $B_{ij} = U_{ij} = L_{ij} = 0$. If $U_{ij} = 0$ or $L_{ij} = 0$ for $j < i$, then either effects of Y_j on Y_i along different paths cancel, in which case B_{ij} need not be zero, or there exists no directed path from j to i, in which case B_{ij} is zero.*

Under an assumption of no path cancellations, Corollary 5 generalizes Corollary 4 to situations in which the configuration of zeros in α does not produce a zero row or column of L. To see this, note that a zero jth row of Θ implies that there are no directed paths between nodes $Y_1, \ldots, Y_{j-1}$ and Y_j, while a zero ith column implies that there are no directed paths between node Y_i and nodes $Y_{i+1}, \ldots, Y_p$. An example of a graph whose sparsity structure is described by Corollary 5 but not by Corollary 4 is depicted in Figure 4.

Unlike the sparsity structures identified in the more general Corollary 5, the sparsity patterns described in Corollary 4 can be interpreted in terms of conditional independencies under an additional assumption of Gaussianity.

PROPOSITION 2. *Consider a Gaussian random vector $Y = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ with zero mean and covariance matrix $\Sigma = U\Psi U^T$, where $U = I + \Theta$, $\Theta \in \mathsf{LT}_\mathsf{s}(p)$ and $\Psi \in \mathsf{D}_+$. Then,*

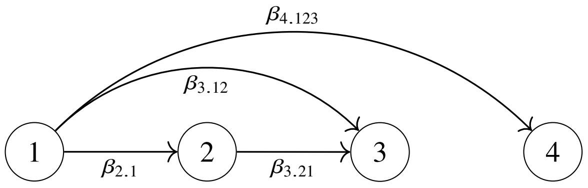


Fig. 4: Directed acyclic graph corresponding to U and L satisfying $U_{42} = 0$ and $L_{42} = 0$.

1. *If the* j*th column of* Θ *is zero, then* $Y_j \perp\!\!\!\perp Y_{j+1}, \ldots, Y_p | Y_1, \ldots, Y_{j-1}$. *Consequently,* $\Sigma^{-1}_{ji} = 0$ *for* $i \in \{j+1, \ldots, p\}$.
2. *If the* j*th row of* Θ *is zero, then* $Y_j \perp\!\!\!\perp Y_1, \ldots, Y_{j-1}$ *and* $\Sigma_{ij} = 0$ *for* $i \in \{1, \ldots, j-1\}$.

For the same number of edges in a graph, the number of zeros in the Gaussian precision matrix depends on the directions of the arrows relative to the configuration of arrows; no reordering of variables can produce a sparser representation. This arises because conditioning in the interpretation of precision matrices is on all variables, rather than only those that occur earlier in the ordering. Given a pair of variables i and j, marginalization over a third variable induces an edge between i and j if the marginalized variable is a transition node or a source node. By contrast, conditioning is edge-inducing if the conditioning variable is a sink node. In a diagrammatic representation due to Cox & Wermuth (1996), with $\not\!\not\circ$ and $\boxed{\circ}$ representing marginalization and conditioning respectively,

$$\begin{array}{ccc} i \longleftarrow \not\!\not\circ \longrightarrow j, & i \longleftarrow \not\!\not\circ \longleftarrow j, & i \longrightarrow \boxed{\circ} \longleftarrow j \\ i \text{ ---- } j, & i \longleftarrow j, & i \text{ ---- } j. \end{array}$$

where ---- indicates that no direction in the induced edge is implied.

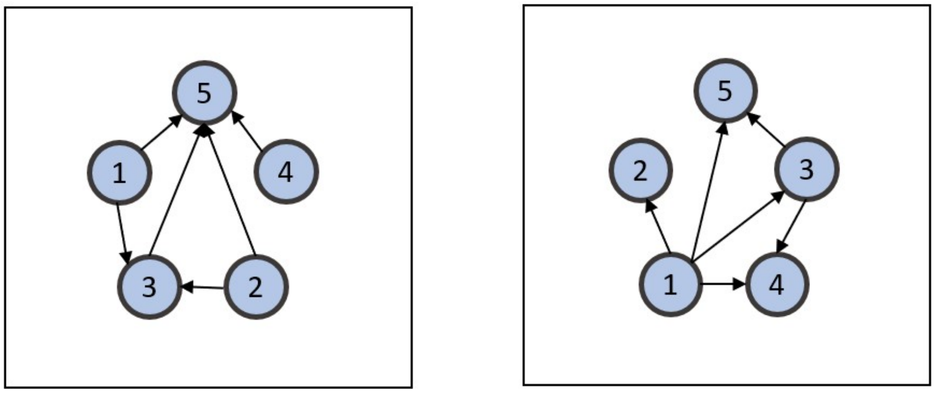
These marginalization and conditioning identities also imply that U will typically be sparser in the correct causal ordering than in an erroneous ordering, modulo coincidental path cancellations, as by definition, any source or transition nodes can only be present among the conditioning sets, and there can be no conditioning on sink nodes.

Figure 5 depicts two examples of directed graphs that are compatible with the covariance matrices from Figure 2 (c) and (d) respectively. The indices of non-zero off-diagonal entries of the corresponding lower-triangular matrices are $\{(3,1),(3,2),(5,1),(5,2),(5,3),(5,4)\}$ and $\{(2,1),(3,1),(4,1),(5,1),(4,3),(5,3)\}$. These are non-zero entries of both B and Θ, as a convergent Taylor representation of the matrix inverse in (7) shows that if Θ has zero rows or columns, the corresponding rows and columns of B are also zero.

Interpretation of the precision matrix is more appropriate for undirected graphs. For instance, if the edges in Figure 5 (a) were undirected, it would hold that $4 \perp\!\!\!\perp \{1,2,3\} \mid 5$ and Σ^{-1}_{4j} would be zero for $j = 1, 2, 3$. That variable 5 is a sink node, however, invalidates this result, as conditioning on the common sink node induces an edge between variable 4 and all other variables.

5.4. *Approximate zeros*

By Proposition 1, the element (i, j) of $L = \log(U)$, and by extension, the corresponding coefficient in the basis expansion of L, is equal to the weighted sum of effects of Y_j on Y_i along all paths connecting the two nodes, with weights inversely proportional to the length of a given path. As a result, in the absence of cancellations of effects along paths of different lengths, which produces an exact zero, a logarithmic transformation reduces the contribution of long paths relative to short



(a) $\Sigma^{r}_{ltu}(\alpha)$ (b) $\Sigma^{c}_{ltu}(\alpha)$

Fig. 5: Directed acyclic graphs corresponding to the example of Figure 2. Arrows indicate directed edges and nodes correspond to random variables.

paths in the absolute entries of the matrix. This leads to an interpretation of a near-zero as meaning that short paths between variables are associated with small conditional effects, while any large conditional effects are mediated by a string of intermediate variables, where conditioning is on all variables that occur earlier in the causal ordering. The approximation inherent to any statistical algorithm that sets small values of α to zero is thus as follows: the relation between nodes i and $j < i$ would be declared null if relatively direct regression effects were negligible and other effects manifested through long paths.

All of B, U and L contain the same information in different guises, that in B being the most easily interpretable. Once sparsity is sought, however, the sparse approximations to B, U and L place emphasis on different aspects.

Since components of B are the direct effects, thresholding on this scale (i.e. setting absolute entries below a given threshold to zero) implicitly assumes that the direct effects are the most important to recover. Consider three variables with connections $1 \rightarrow 2 \rightarrow 3$ and no direct edge between 1 and 3. Suppose that edge weight $1 \rightarrow 2$ is very large, while that of $2 \rightarrow 3$ is small and hence thresholded to zero. The effect of 1 on 3 is not reflected in the resulting thresholded approximation to B, even though this effect may be appreciable in view of the large $1 \rightarrow 2$ effect.

At the other extreme, the entries of U represent the sum of effects along all paths, in which the information about short paths is absorbed in a composite. The cost, potentially, is a small number of near-zeros, and recovery of distant effects, as thresholding implicitly assumes that paths of all lengths are equally important. Consider a simplistic example in three variables to illustrate a particular point, ignoring other aspects. These variables are “parents smoke”, “individual smokes”, “individual has lung cancer”. Since longer paths are not discounted, it may superficially appear that “parents smoke” has a larger effect on “individual has lung cancer” than “individual smokes”, as it has a positive direct effect as well as a positive indirect effect via the intermediate variable “individual smokes”.

Thresholding on the scale of L is a compromise between these two extremes. In the first example we can still recover, after sparse approximation, the effect $1 \rightarrow 3$ that would be lost to thresholding on the original scale, while in the second example, we can still identify “individual smokes” as the main cause of cancer.

Interestingly, the above conclusion bears some resemblance to a suggestion from influential work in computer science about learning in networks. Grover & Leskovec (2016) define two types of neighbourhoods of a node in a graph: one consisting of direct neighbours, and another that involves traversing long paths in the graph. The authors argue that information from both types should be combined, with a hyperparameter specifying the relative importance of paths. The Σ_{ltu} parametrization specifies the analogue of this hyperparameter as a discount rate on long paths, given by the inverse path length.

6. Sparsity under the Σ_{pd} parametrization

6.1. *Sparsity-induced structures*

The ordering is immaterial for the Σ_{pd} and Σ_o parametrizations. The structures elucidated in Corollaries 1 and 2 imply the same structure on the scale of the precision matrix, and therefore correspond to conditional independencies. We do not put forward that such structures are likely to hold exactly. Their purpose is instead to approximate a more complex reality, so as to aid interpretation or limit the accumulation of estimation error in procedures like linear discriminant analysis, where the covariance matrix is a nuisance parameter. With this in mind, §6.2 establishes the interpretation of exact and approximate zeros under the Σ_{pd} parametrization, while §7 explains the relevance of the Σ_{pd} parametrization in the context of chain graphs, broadening its scope.

6.2. *Interpretation of zeros under the Σ_{pd} parametrization*

The interpretation of basis coefficients in the transformation Σ_{pd} is less straightforward than in the case of Σ_{ltu}. Under additional assumptions we can recover an interpretation of zero entries in the matrix logarithm of $\Sigma \in \mathsf{PD}(p)$.

Consider an undirected Gaussian graphical model for $Y_1, \dots, Y_p$, with no self-loops. Write $V = \Sigma^{-1}$ for the precision matrix. The regression coefficients are entries of the matrix $\tilde{V} = \mathrm{diag}(V)^{-1}V$, where $\mathrm{diag}(A)$ denotes a diagonal matrix whose diagonal entries are equal to those of A. The form of $\tilde{V}$ is analogous to that of normalized graph Laplacian in spectral graph theory (e.g. Von Luxburg , 2007). Let $\upsilon^{u}_{ij}(l)$ denote the total effect of a unit change in Y_j on Y_i along all paths of length l, the superscript u in $\upsilon^{u}_{ij}(l)$ distinguishing this quantity for an undirected graph. As compared with $\upsilon_{ij}(l)$ from §5.2, in the undirected case, a given node can be connected to all others. The following proposition establishes the interpretation of entries of $\log(\Sigma\mathrm{diag}(V)) = -\log(\tilde{V})$. This result can be seen as an adaptation of Proposition 1 for undirected graphs.

Proposition 3. *For a matrix $\Sigma \in PD(p)$, let $\tilde{V} = diag(V)^{-1}V$ and $\tilde{\Sigma} = \Sigma diag(V)$. Then the elements (i, j) of $\tilde{\Sigma}$ and of its matrix logarithm have the form,*

$$\tilde{\Sigma}_{ij} = \sum_{l=1}^{\infty} \upsilon^{u}_{ij}(l), \qquad \log(\tilde{\Sigma})_{ij} = -\log(\tilde{V})_{ij} = \sum_{l=1}^{\infty} \frac{\upsilon^{u}_{ij}(l)}{l}$$

if the infinite sums converge.

Unlike in Proposition 1, expressions for elements of transformed matrices in Proposition 3 involve infinite sums. As a result, Proposition 3 requires an additional assumption of convergence. The interpretation established in Proposition 3 holds for the matrix logarithm of a suitably scaled covariance matrix, rather than for $\log(\Sigma)$. However, a direct calculation yields the following result, which is also implicit in Corollary 1.

Corollary 6. *Assume that no cancellation of effects of Y_j on Y_i occurs along different paths. Then $L_{ij} = 0$ if and only if $\tilde{V}_{ij} = 0$.*

Thus, in the absence of cancellation of effects, an element (i, j) of $L = \log(\Sigma)$, as well as the corresponding basis coefficient, is zero if and only if there is no path between nodes i and j. It is often the case, in spite of the previous sentence, that L is sparser than Σ and Σ^{-1} under more general notions of sparsity. This is probed by simulation in §9.

7. Chain graphs and the Iwasawa decomposition

7.1. *A unifying parametrization for chain graphs*

The interpretation of (6) in terms of partial Iwasawa coordinates points, via a group-theoretic treatment, to an encompassing formulation. This is first developed from a statistical perspective.

Corollary 7 to be presented is a unifying result in which the interaction of sparsity on the transformed scale with structure on the original scale is elucidated, recovering three of the results of §3 as special cases in which connected components consist of either all undirected edges (Corollary 1) or all directed edges (Corollaries 3 and 4).

A graph $G = (V, E)$ is a chain graph if it contains both directed and undirected edges, but no semi-directed cycles (Drton & Eichler, 2006, p. 83). When two nodes $v, w \in V$ are connected by a path consisting solely of undirected edges, we say that u and w are equivalent. Let $\mathcal{U}$ be a set of equivalence classes, called chain components, of this equivalence relation. Define a new graph $\mathcal{D} = (\mathcal{U}, \mathcal{E})$ with nodes $\mathcal{U}$ and edges $\mathcal{E}$ between chain components. Since we assume that there are no semi-directed cycles in G, the graph of $\mathcal{D}$ is a directed acyclic graph.

Chain graphs are usually characterized by the alternative Markov property (Andersson et al., 2001), satisfied under a mean-zero Gaussianity assumption if and only if the precision matrix can be written as

$$\Sigma^{-1} = (I - B^{\mathrm{T}})\Omega^{-1}(I - B), \tag{10}$$

where Ω^{-1}_{uv} is zero if an undirected edge (u, v) is not in the graph, and $B_{vu} = 0$ if a directed edge $u \longrightarrow v$ is not in the graph (Drton & Eichler, 2006). Since there exists at most one edge between any pair of nodes, $\Omega_{uv} \neq 0$ implies $B_{vu} = 0$, and vice-versa. Every directed acyclic graph can be represented by a triangular matrix, so we can always find a permutation matrix $P \in \mathsf{P}(p)$ such that decomposition (10) of $P\Sigma^{-1}P^{\mathrm{T}}$ yields a strictly lower-triangular matrix B and a block-diagonal matrix Ω^{-1}. That there exists a $P \in \mathsf{P}(p)$ that simultaneously rearranges B and Ω^{-1} follows from the assumption that there is an underlying chain graph. From now on we assume that an ordering of variables has been chosen such that B is triangular and Ω^{-1} is block-diagonal. The factorization (10) then represents the precision matrix as a product of block-diagonal and block-triangular matrices. The block-triangular matrix captures connections between chain components, while the block-diagonal matrix describes connections within the chain components.

Suppose that $Y \sim N(0, \Sigma)$ is partitioned into c blocks, $Y = (Y_1, \ldots, Y_c)$, where each block Y_i constitutes a chain component, that is, the variables within Y_i form a connected undirected graphical model. Let p_i denote the dimension of the sub-vector Y_i. The decomposition of the precision matrix (10) implies the decomposition of Σ in terms of a block-diagonal component Ω:

$$\Sigma = T\Omega T^{\mathrm{T}}, \qquad \Omega = \Omega_1 \oplus \Omega_2 \oplus \cdots \oplus \Omega_c, \qquad \Omega_i \in \mathsf{PD}(p_i), \tag{11}$$

where $T = (I - B)^{-1} \in \mathsf{LT}_s(p)$ has diagonal blocks $I_{p_1}, \ldots, I_{p_c}$. Factorization (11) can be obtained by successive block-triangularization of Σ, and the corresponding block-triangular transformation can be parametrized as

$$(\alpha, \delta) \mapsto \Sigma_{bt}(\alpha, \delta) := e^{L(\alpha)}e^{D(\delta)}(e^{L(\alpha)})^{\mathrm{T}},$$

where $D(\delta) = D_1(\delta_1) \oplus \cdots \oplus D_c(\delta_c)$, $D_i(\delta_i) \in \mathsf{Sym}(p_i)$, $\delta_i \in \mathbb{R}^{p_i(p_i+1)/2}$ and $\delta = (\delta_1^{\mathrm{T}} \ldots \delta_c^{\mathrm{T}})^{\mathrm{T}}$. The matrix logarithm $L(\alpha)$ is block-triangular with c diagonal blocks, each equal to a $p_i \times p_i$ identity matrix. With p_δ the dimension of δ, the dimension of α is $p(p+1)/2 - p_\delta$.

That Σ_{bt} represents a unifying structure is seen on noting that when $D(\delta)$ is diagonal we recover the parameterization $\Sigma_{ltu}(\alpha, \delta)$, and therefore also $\Sigma_{lt}(\alpha)$ by the discussion surrounding equation (5). At the other extreme, when $D(\delta)$ consists of a single block of dimension $p \times p$, we recover $\Sigma_{pd}(\delta)$. The remaining parametrization Σ_o is not directly recoverable as a special case of Σ_{bt}, although there is an indirect connection because $L \in \mathsf{Sk}(p)$ can be decomposed as $L = L_s - L_s^{\mathrm{T}}$ with $L_s \in \mathsf{LT_s}(p)$. Additional details are in Appendix E.

Corollary 7. *Let Σ be parametrized as $\Sigma_{bt}(\alpha, \delta) = e^{L(\alpha)}e^{D(\delta)}(e^{L(\alpha)})^{\mathrm{T}}$ and let $d \leq p$. Then, $\Sigma_{bt}(\alpha, \delta) = T\Omega T^{\mathrm{T}}$, and,*

1. *(Sparsity of DAG): $T = I + A$ and the ith row of A is zero if and only if the ith row of $L(\alpha)$ is zero. The jth column of A is zero if and only if the jth column of $L(\alpha)$ is zero.*
2. *(Sparsity of chain components): Ω is of the form $\Omega = P\Omega^{(0)}P^{\mathrm{T}}$, where P is a permutation matrix and $\Omega^{(0)} = \Omega_1^{(0)} \oplus \Phi_{p-d}$, where $\Phi_{p-d} \in \mathsf{D}(p-d)$ and $\Omega_1^{(0)} \in \mathsf{PD}(d)$ is block-diagonal if and only if $D(\delta) = P(D_1^{(0)} \oplus \Delta_{p-d})P^{\mathrm{T}}$ where $D_1^{(0)}$ is block-diagonal and $\Delta_{p-d} \in \mathsf{D}(p-d)$.*

Since $L(\alpha) = \log(T) = -\log(I - B)$, the coefficients of $\log(I - B)$ in the appropriate basis are equal to $-\alpha$. Thus, the sparsity indices of $I - B$ and T on the logarithmic scale coincide. In contrast, no obvious relationship exists between the sparsity of $I - B$ and T, since a zero entry in $I - B$ does not imply a zero entry in T, and vice versa. A similar point applies to Ω and Ω^{-1} since $D(\delta) = \log(\Omega) = -\log(\Omega^{-1})$.

A result analogous to Corollary 7 can be obtained for the precision matrix, since zero rows and columns of $L(\alpha)$ and $-L(\alpha) = \log(I - B)$ coincide. Part (1) of Corollary 7 thus describes structures arising when the sparsity patterns of T and $I - B$ coincide. For example, suppose that the ith column of $I - B$ is zero, that is, node i has no descendants. This is reflected on the transformed scale by a zero ith row of $L(\alpha)$.

7.2. *Connection to the Iwasawa decomposition of the general linear group*

The parametrization Σ_{bt} represents the general form of the Iwasawa decomposition of the general linear group, pertaining to the partition $p_1 + \cdots + p_c = p$ of p. This provides a group-theoretic perspective on how the parametrization Σ_{bt} unifies and generalizes three of the four parameterizations from (1), detached from any consideration of causal ordering. Appendix D provides further discussion. The identification $\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p)$ characterizes a positive definite matrix as a non-singular one whose 'orthogonal component' has been discounted. Explicitly, $\Sigma_A = A^{\mathrm{T}}A$ is positive definite for every $A \in \mathsf{GL}(p)$, and left-multiplication by any orthogonal matrix gives an equivalence class $[A] = \{OA : O \in \mathsf{O}(p)\} \subset \mathsf{GL}(p)$ such that $\Sigma_A = V^{\mathrm{T}}V$ for any $V \in [A]$. Relatedly, Draisma & Zwiernik (2017) identified the subgroup of $\mathsf{GL}(p)$ acting on Σ, and studied corresponding equivariant estimators that preserve the chain graph property.

The group theoretic perspective provides a new way to interpret the information geometry of the zero-mean multivariate Gaussian (Skovgaard, 1984). The Fisher information metric tensor coincides with the quotient geometry of $\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p)$ under a Riemannian metric that is invariant to the transitive action of $\mathsf{GL}(p)$. Upon representing a positive definite covariance matrix Σ in the partial Iwasawa coordinates $(\Sigma_{aa}, \Sigma_{ba}\Sigma_{aa}^{-1}, \Sigma_{bb.a})$, the metric is endowed with an interpretation compatible with a suitable graphical model and a corresponding Σ_{ltu} parametrization.

8. Propagation of error and estimation

8.1. *Existing results for Σ_{pd}*

A question of practical relevance is whether increased sparsity on the transformed scale translates to an inferential advantage. Intuition in the case of the Σ_{pd} parametrization can be obtained through consideration of the simplest estimator exploiting sparsity, namely the thresholding estimator of Bickel & Levina (2008a,b) applied on the transformed scale. For the Σ_{pd} parametrization, thresholding sets to zero the entries of a pilot estimator $\hat{L} = \log \hat{\Sigma}$ that are below a threshold in absolute value. Success of the approach hinges on the elementwise consistency of $\hat{L}$ for L. We are thus interested in how the estimation error $\hat{\Sigma} - \Sigma$ propagates to the scale of the matrix logarithm. To simplify notation and isolate the considerations involved, we outline the argument for a deliberately oversimplified setting, before highlighting modifications for the analysis of $\hat{L} - L$.

Consider a small perturbation of Σ, of the form $\Sigma + \varepsilon I$ for $\varepsilon > 0$, which preserves the eigenvectors. The argument in Appendix I shows that, using a complex-variable representation of the matrix logarithm, the error propagates to the (j, k)th entry on the logarithmic scale as

$$[\log(\Sigma + \varepsilon I) - \log(\Sigma)]_{j,k} = \varepsilon \sum_{r,v} \Big(\frac{1}{2\pi i} \oint_\gamma \frac{\log(z)}{(z - (\lambda_r + \varepsilon))(z - \lambda_v)} dz \Big) o_{jr} o_{kv} \sum_{\ell,s} o_{\ell r} o_{sv}, \qquad (12)$$

where o_{ij} denotes the ith entry of the jth eigenvector of Σ. Consider the summation over ℓ and s in (12). For $r = v$,

$$\sum_{\ell,s} o_{\ell v} o_{sv} = \sum_{s} o_{sv} o_{sv} + \sum_{s,\ell \neq s} o_{\ell v} o_{sv} = 1$$

by the orthonormality identity $O^{\mathrm{T}} O = O O^{\mathrm{T}} = I$. For $r \neq v$, the double summation is approximately zero by the observation that cross-products $o_{\ell r} o_{sv}$ are of order $1/p$ and zero on average for large p. Subject to this last approximation, (12) simplifies to

$$[\log(\Sigma + \varepsilon I) - \log(\Sigma)]_{j,k} = \varepsilon \sum_{v} \Big(\frac{1}{2\pi i} \oint_\gamma \frac{\log(z)}{(z - (\lambda_v + \varepsilon))(z - \lambda_v)} dz \Big) o_{jv} o_{kv}, \qquad (13)$$

where the term in parenthesis is given by the sum of the residues at the two singularities,

$$\frac{1}{2\pi i} \oint_\gamma \frac{\log(z)}{(z - (\lambda_v + \varepsilon))(z - \lambda_v)} dz = \frac{\log(\lambda_v + \varepsilon) - \log(\lambda_v)}{\lambda_v + \varepsilon - \lambda_v} = \frac{\log(\lambda_v + \varepsilon) - \log(\lambda_v)}{\varepsilon},$$

whose first-order Taylor expansion around $\varepsilon = 0$ is λ_v^{-1}, i.e. the derivative of $\log \lambda_v$. The perturbation ε thus propagates to the scale of the matrix logarithm as

$$[\log(\Sigma + \varepsilon I) - \log(\Sigma)]_{j,k} = \varepsilon \sum_{v} \lambda_v^{-1} o_{jv} o_{kv} + O(\varepsilon^2) = \varepsilon [\Sigma^{-1}]_{j,k} + O(\varepsilon^2),$$

which, as expected, is the directional derivative of the matrix logarithm at Σ in the direction εI.

Realistic pilot estimators of Σ entail perturbations of both eigenvectors and eigenvalues, and the previous argument then requires that pilot estimators provide consistent estimates $\hat{o}_v$ of eigenvectors in the sense that $\hat{o}_r^{\mathrm{T}} o_v \to_p 0$ for $r \neq v$ and $\hat{o}_v^{\mathrm{T}} o_v \to_p 1$. In the more general argument, the constant ε is replaced by elements of $\hat{\Sigma} - \Sigma$ in a summation on the right hand side. A more complete development for specific pilot estimators can be found in Battey (2019), where results are also presented for the propagation of error in the converse direction under the spectral norm, having exploited sparsity on the scale of the matrix logarithm.

8.2. *New results for* Σ_{ltu}

A broadly analogous scheme applies to estimation under a sparse Σ_{ltu} parameterization. Consider a decomposition (11), i.e. $\Sigma = T\Omega T^{\mathrm{T}}$, where Ω is block-diagonal and T is triangular. The Σ_{ltu} parametrization arises as a special case when each block contains a single variable. When a causal ordering of variables or blocks of variables is available, a natural pilot estimator for T regresses each variable on its causal predecessors, called parents, and a corresponding pilot estimator of Ω is the sample covariance matrix of the resulting residuals. If the variables are not generated by a causal model, pilot estimators for both T and Ω can be obtained through an LDL decomposition of the sample covariance matrix. The resulting pilot estimators, $\hat{T}$ and $\hat{\Omega}$, are elementwise consistent. As in Section 8.1, sparsity on the transformed scale is exploited by thresholding $\log(\hat{T})$ and $\log(\hat{\Omega})$, and converting the resulting quantities back to the original scale by applying the matrix exponential. The resulting estimators, $\tilde{T}$ and $\tilde{\Omega}$, are consistent in the spectral norm. A natural estimator of Σ is then $\tilde{\Sigma} = \tilde{T}\tilde{\Omega}\tilde{T}^{\mathrm{T}}$, which is also consistent in the spectral norm. A more detailed discussion of the estimator and its properties can be found in Appendix J.

Proposition 4 establishes spectral-norm consistency of the proposed estimator under conditions detailed in Appendix J. These include Condition 1, which characterizes the sparsity of L and D.

Condition 1. Assume that $L \in \mathcal{U}(q_l, s_l(p)) \cap \mathsf{LT_s}(p)$ and $D \in \mathcal{U}(q_\omega, s_\omega(p)) \cap \mathsf{PD}(p)$, where $q_l, q_\omega \in [0,1]$, $s_l(p)/p \to 0$, $s_\omega(p)/p \to 0$ and

$$\mathcal{U}(q, s(p)) = \left\{ A \in \mathsf{M}(p) : \max_i \sum_{j=1}^{p} |A_{ij}|^q = s(p) \right\}. \tag{14}$$

Proposition 4. *Suppose that the tuning parameters of equations* (J.4) *and* (J.8) *of Appendix J satisfy* $\tau_l \asymp (n^{-1}\log p)^{1/2}$ *and* $\tau_\omega \asymp s_l(p)^2 (n^{-1}\log p)^{(3/2-q_l)(1-q_\omega)}$. *Under Condition 1 and Conditions J.1-J.4 of Appendix J.2, with* $\varphi \geq 1/2$, *the estimator* $\tilde{\Sigma} = \tilde{T}\tilde{\Omega}\tilde{T}^{\mathrm{T}}$ *of* $\Sigma = T\Omega T^{\mathrm{T}}$ *satisfies*

$$\|\tilde{\Sigma} - \Sigma\|_2 = O_p\big(\max\{r_t, r_\omega\}\big),$$

where

$$r_t = s_l(p)^2 (n^{-1}\log p)^{3/2-q_l}, \quad r_\omega = s_\omega(p) s_l(p)^{2-2q_\omega} (n^{-1}\log p)^{(3/2-q_l)(1-q_\omega)}.$$

An important question concerns the implications of misspecification of the causal ordering. Although this would annul the interpretation of §5, the role of the causal ordering in Proposition 4 is via the degree of sparsity present, which is reflected in the rates in Proposition 4. Thus, to the extent that the conditions are still satisfied, Proposition 4 remains valid.

9. Some numerical insights

9.1. *Approximate sparsity in the four logarithmic domains*

The prospect of routinely inducing sparsity through logarithmic transformation under the four maps (1) is only realistic under a notion of approximate sparsity that allows for slight departures from zero. Simulations in Appendix K give an indication of how approximate sparsity in the four logarithmic domains associated with Σ_{pd}, Σ_o, Σ_{lt} and Σ_{ltu} transfers to a commensurate notion of approximate sparsity in the inverse domain, this being the most widely used parameter domain in which to perform sparse estimation. Tables K.1-K.4 of Appendix K also compare the performance of thresholding estimators on the different scales, suggesting in all cases except Σ_o that exploiting sparsity on the most sparse scale, i.e. the logarithmic scale under the relevant parametrization, transfers substantial benefits to estimation of Σ^{-1}. In the case of Σ_o, the simple thresholding approach of Appendix K appears too simplistic, presumably owing to the constraints on α needed to make the parametrization injective.

9.2. *Exploration of sparsity regimes*

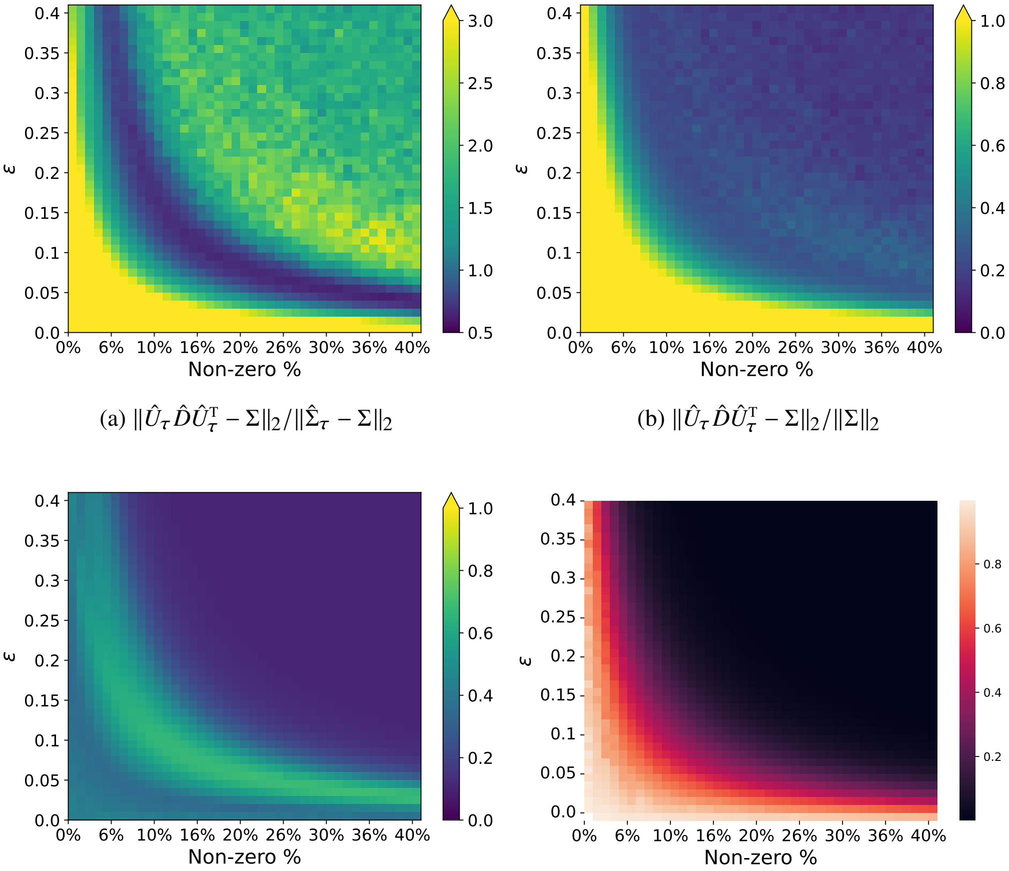
In §9.1, the matrices on the transformed scale were sparse by construction. We now investigate whether the logarithmic transformation can be useful in less idealized situations.

Consider a Gaussian directed acyclic graph with covariance matrix $\Sigma = (I - B)^{-1}D^{-1}(I - B^{\mathrm{T}})^{-1}$. When B contains many zeros or near-zeros, Σ is also likely to be sparse, and the estimator $\hat{\Sigma}_\tau$ of Bickel & Levina (2008b) would be a natural choice. Sparsity of Σ typically decreases both as the number of non-zero elements of B increases, and as the magnitude of non-zero entries (the weights of directed edges) becomes large. This follows from Proposition 1, whereby the entry (i, j) of the matrix $(I - B)^{-1}$ corresponds to the sum of effects of node i on node j along all paths connecting the two nodes. As the number, or magnitude, of non-zero edge weights increases, cumulative effects inevitably increase. Although a similar phenomenon is expected for $L = -\log(I - B)$, Proposition 1 suggests that the sparsity of L should decrease more slowly, due to the discounting of longer paths. Eventually, as the number of entries of B below a threshold increases, or as the absolute value of these entries increase, we expect a strong accumulation of effects on variables ordered last, as these will have the largest number of incoming paths. Since the number of possible paths increases exponentially with the number of nodes, the accumulation of effects can cause entries of Σ to be unbounded.

There are thus three regimes. When the edge matrix B is sparse or its non-zero entries have small absolute values, thresholding in the original domain will typically yield better results. As the number of non-zero entries of B increases, or the edge weights increase, thresholding in the logarithmic domain is expected to be advantageous. With a further decrease in the sparsity of B or increase in the edge weights, there is no approximate sparsity in either domain.

To verify this empirically, we compared the performance of thresholding on the original and logarithmic scales under the Σ_{ltu} parametrization, for different values of edge weights and different levels of sparsity for B. Specifically, we took D as the identity matrix and generated an edge matrix B by randomly selecting a prespecified percentage of its entries, and assigning a fixed value $\epsilon > 0$ to those entries. Positivity of ϵ avoids cancellations of effects along different paths. For each covariance matrix, we generated a sample of size $n = 150$ from the corresponding multivariate normal distribution and constructed thresholding estimates on the scales of interest, following the recommendation of Bickel & Levina (2008b) for selecting the threshold τ. Specifically, a sample covariance matrix was estimated on two disjoint subsets of the data, of size $n/3$ and $2n/3$. The estimate $\tilde{\Sigma}$ based on the larger sample was treated for the purpose of tuning as the target covariance matrix. Thresholding was applied to the matrix estimated on the smaller sample, yielding a sparse estimate $\bar{\Sigma}_\tau$. The threshold was then chosen to minimize the relative ℓ_2-norm error, $\|\tilde{\Sigma} - \bar{\Sigma}_\tau\|_2/\|\tilde{\Sigma}\|_2$ across 5 random splits. The final estimate $\hat{\Sigma}_\tau$ was based on the selected threshold and the full sample. The same procedure, with the obvious modifications, was used to select the threshold used for sparse estimation on the logarithmic scale under the Σ_{ltu} parametrization, resulting in estimates $\hat{U}_\tau$ of U and $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$ of Σ.

Results in Figure 6 (a) show that thesholding on the original scale outperforms thresholding on the logarithmic scale under the Σ_{ltu} parametrization for high levels of sparsity of B and small values of ϵ, while the opposite is true for medium levels of sparsity and ϵ. For large values, the covariance matrix is highly non-sparse and neither sparsity scale is suitable. The standalone performance of $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$ and $\hat{\Sigma}_\tau$ is shown in Figures 6 (b) and (c). Results indicate that when $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$ outperforms $\hat{\Sigma}_\tau$ it is due both to the poorer performance of the latter estimate and improved performance of the former. The performance of $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$ exhibits a sharp transition as



(a) $\|\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}} - \Sigma\|_2 / \|\hat{\Sigma}_\tau - \Sigma\|_2$

(b) $\|\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}} - \Sigma\|_2 / \|\Sigma\|_2$

(c) $\|\hat{\Sigma}_\tau - \Sigma\|_2 / \|\Sigma\|_2$

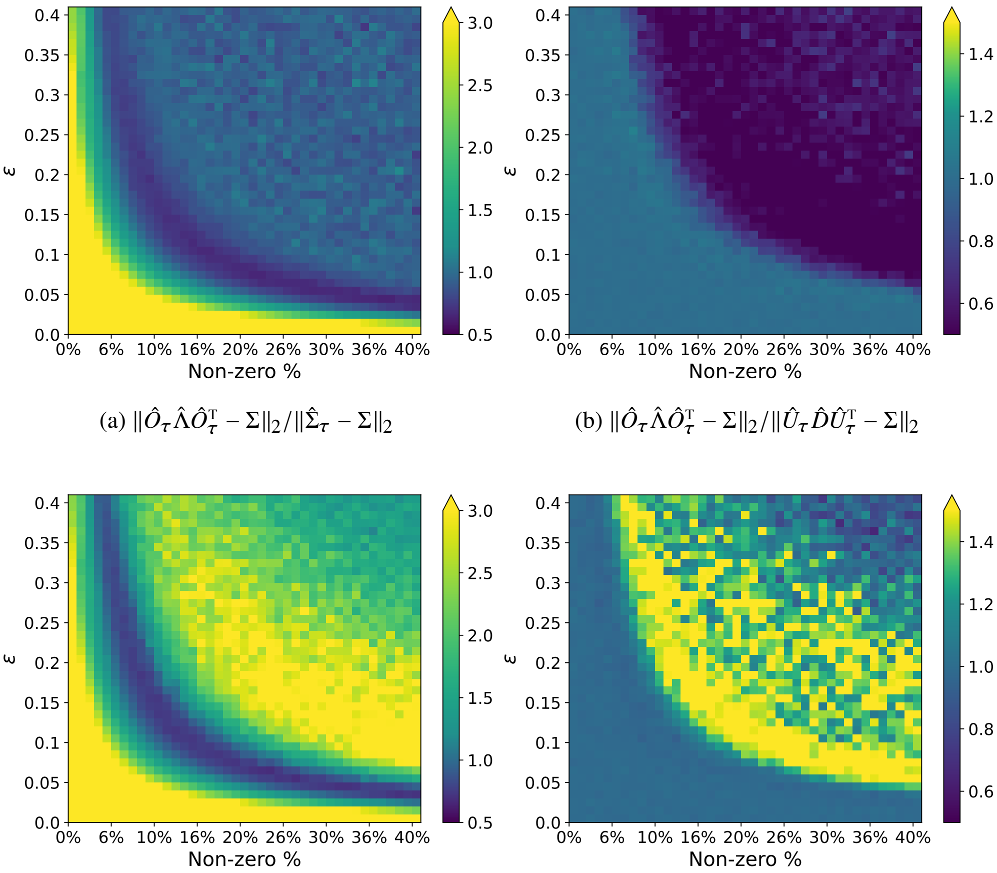
(d) $r(L)/r(\Sigma)$

Fig. 6: Relative ℓ_2 errors (a, b and c) and relative ℓ_1 row norms (d) for different combinations of ϵ (y-axis) and levels of sparsity of B, measured by the percentage of non-zero entries (x-axis). Pixels are median values over 100 simulations with $n = 150$, $p = 100$ for different combinations of ϵ (y-axis) and levels of sparsity of B, measured by the percentage of non-zero entries (x-axis).

the sparsity of B decreases and ϵ increases. This may suggest that the logarithmic transformation is detrimentally distorting for very sparse covariance matrices. Since the sparsity conditions for thresholding are closely related to row-wise norms, we show in Figure 6 (d), for comparison to (a), the ratio of maximum ℓ_1 row norm of the two matrices, defined for $A \in \mathsf{M}(p)$ as

$$r(A) = \max_{i \in \{2,\ldots,p\}} \sum_{j=1}^{i-1} |A_{ij}|. \tag{15}$$

In order to make the metric comparable for lower-triangular and symmetric matrices, $r(A)$ only considers the entries of A below the diagonal. The contours of equal $r(L)/r(\Sigma)$ in (b) closely resemble those of the relative errors in (a). We probe this relationship further in Figure K.4 of Appendix K.1. The performance of $\hat{O}_\tau \hat{\Lambda} \hat{O}_\tau^{\mathrm{T}}$ and $\exp(\hat{L}_\tau)$ relative to $\hat{\Sigma}_\tau$ is shown in Figure 7. Interestingly, the pattern of relative behaviour mirrors that of $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$.



(a) $\|\hat{O}_\tau \hat{\Lambda} \hat{O}_\tau^{\mathrm{T}} - \Sigma\|_2 / \|\hat{\Sigma}_\tau - \Sigma\|_2$

(b) $\|\hat{O}_\tau \hat{\Lambda} \hat{O}_\tau^{\mathrm{T}} - \Sigma\|_2 / \|\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}} - \Sigma\|_2$

(c) $\|\exp(\hat{L}_\tau) - \Sigma\|_2 / \|\hat{\Sigma}_\tau - \Sigma\|_2$

(d) $\|\exp(\hat{L}_\tau) - \Sigma\|_2 / \|\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}} - \Sigma\|_2$

Fig. 7: Relative ℓ_2 errors for different combinations of ϵ (y-axis) and levels of sparsity of B, measured by the percentage of non-zero entries (x-axis). Each entry corresponds to the ratio of median errors over 100 Monte Carlo simulations with $n = 150$ and $p = 100$.

	Sample size n		
Estimator	30	50	100
$\hat{\Sigma}_\tau$	60% (6.4%)	62.6% (5.2%)	66.9% (7.7%)
$\exp(\hat{L}_\tau)$	64.5% (3.7%)	66.8% (3.6%)	70.6% (2.5%)
$\hat{U}_\tau \hat{D} \hat{U}^{\mathrm{T}}$	65.5% (3.1%)	67.3% (3.6%)	69.7% (2.4%)

Table 2: Median (standard deviation) accuracy over 20 simulations using arrhythmia dataset.

9.3. *Classification of leukemia and arrhythmia patients*

We assessed the use of the new sparsity scales for classification of leukemia and arrhythmia patients from high-dimensional observations. The details are described in Appendix L. For leukemia patients ($p = 3571$), the results show a slight improvement in accuracy, from median of 95.5% on the original scale to 97.7% for Σ_{ltu} transformation. However, the difference is insignificant, with standard deviations of errors around 5%. For arrhythmia patients ($p = 164$), we observe an improvement in performance for a range of sample sizes, as shown in Table 2.

10. Closing discussion

The work has uncovered insights into the interpretation of sparsity on non-standard scales, identifying situations in which an assumption of sparsity might be more reasonable on a transformed scale. Open questions concern how one might test for sparsity across several different scales, or find the best sparsity scale empirically. The work also points to the development of more sophisticated estimators than those used in the simulations of §9, perhaps in the vein of Zwiernik (2025), who proposed an elegant formulation covering constraints in the Σ_{pd} parametrization.

Acknowledgements

We are grateful to two anonymous referees for their careful reading and probing questions, and to Sergey Oblezin for helpful discussions. The EPSRC, NSF and NIH provided research funding.

Supplementary material

The supplementary material contains Appendices A–L referenced in the main text.

References

Al-Mohy, A. H. and Higham, D. J. (2012). Improved inverse scaling and squaring algorithms for the matrix logarithm. *SIAM J. Sci. Comput*, 34, C153–C169.

Andersson, S. A., Madigan, D. and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.*, 28, 33–85.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*, Chapman and Hall.

Battey, H. S. (2017). Eigen structure of a new class of structured covariance and inverse covariance matrices. *Bernoulli*, 23, 3166–3177.

Battey, H. S. (2019). On sparsity scales and covariance matrix transformations. *Biometrika*, 106, 605–617.

Battey, H. S. (2023). Inducement of population sparsity *Canad. J. Statist.* (Festschrift in honour of Nancy Reid), 51, 760–768.

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36, 199–227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.*, 36, 2577–2604.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. B*, 26, 211–252.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.*, 77, 209–210.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the J. Roy. Statist. Soc.*, 5, 171–176.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. B*, 49, 1–39.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.*, 8, 204–218.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.

Draisma, J. and Zwiernik, P.(2017). Automorphism groups of Gaussian Bayesian networks. *Bernoulli*, 23, 1102-1129.

Drton, M. and Eichler, M. (2006) Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist.*, 33, 247–257

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes (with discussion). *J. Roy. Statist. Soc. B*, 82, 871–932.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. Roy. Statist. Soc. B*, 75, 603–680.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proc. 22nd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, 855–864.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Pavlov, D. , Sturmfels, B. and Telen, S. (2024). Gibbs manifolds. *Info. Geo.*, 7 (Suppl. 2), 691–717.

Pavlov, D. (2024). Logarithmically sparse symmetric matrices. *Beitr. Algebra Geom.*, 65, 907–922.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86, 577–90.

Rybak, J. and Battey, H. S. (2021). Sparsity induced by covariance transformation: some deterministic and probabilistic results. *Proc. Roy. Soc. London, A.*, 477, 20200756.

SKOVGAARD, L. T.(1984). A Riemannian geometry of the multivariate normal model. *Scand J Statist*, 11, 211-223.

WERMUTH, N. AND COX, D. R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Statist. Soc. B*, 66, 687–717.

VON LUXBURG, U. (2004). A tutorial on spectral clustering. *Statistics and computing*, 17, 395–416.

ZWIERNIK, P. (2025). Entropic covariance models. *Ann. Statist.*, to appear.

[*Received on* 1 *July* 2024. *Editorial decision on* 16 *June* 1904]

Biometrika (0000), **000**, 0, *p.* 1
Printed in Great Britain Advance Access publication on 16 June 1904

Supplementary material for 'Regression graphs and sparsity-inducing reparametrizations'

BY J. RYBAK

Department of Mathematics, Imperial College London,
180 Queen's Gate, London SW7 2AZ, U.K.
jakub.rybak18@imperial.ac.uk

H. S. BATTEY

Department of Mathematics, Imperial College London,
180 Queen's Gate, London SW7 2AZ, U.K.
h.battey@imperial.ac.uk

AND K. BHARATH

School of Mathematical Sciences, University of Nottingham,
University Park, Nottingham NG7 2RD, U.K.
karthik.bharath@nottingham.ac.uk

SUMMARY

The supplementary material contains Appendices A–K referenced in the main text. These include proofs and technical details, details of the estimator discussed in §8.2, assumptions and proofs for Proposition 4 and extensive numerical work.

Some key words: Causality; Chain graphs; Graphical models; Matrix logarithm; Reparametrization; Sparsity.

A. MATRIX GROUPS

The following concepts from group theory are relevant to parts of the exposition and proofs. The action of a group G on a set X is a continuous map $G \times X \longrightarrow X, (g, x) \longrightarrow gx$. The *orbit* $[x]$ of $x \in X$ is the equivalence class $\{gx : g \in G\}$, and the set of orbits $X/G := \{[x] : x \in X\}$ is a partition of X known as the *quotient* of X under the action of G. A group action is said to be *transitive* if between any pair $x_1, x_2 \in X$ there exists a $g \in G$ such that $x_2 = gx_1$; in other words, orbits of all $x \in X$ coincide. The subset $G_x = \{g \in G : gx = x\}$ of G that fixes x is known as the *isotropy group* of x and if G_x equals the identity element for every x, the group action is said to be *free*.

For every subgroup $H \subset G$ we can consider the (right) coset, or the quotient, $G/H = \{Hg : g \in G\}$ consisting of equivalence classes of g, where $\tilde{g} \sim g$ if $h\tilde{g} = g$ for some $h \in H$. For groups G that act transitively on X, the map $X \longrightarrow G/G_0$ is a bijection, where G_0 is the isotropy group of the identity element.

A group G that is also a differential manifold is a Lie group. Lie groups thus enjoy a rich structure given by both algebraic and geometric operations. The tangent space at the identity element, denoted by $\mathfrak{g}$, has a special role in that, together with the group operation, it generates the entire group, i.e. every element $g \in G$ can be accessed through elements of $\mathfrak{g}$ and the group operation. It is referred to as the Lie algebra and is a vector subspace of the same dimension as the group. Relevant to the matrices introduced in §2, we have:

(i) If $G = \mathsf{GL}(p)$ with matrix multiplication as the group action, then $\mathfrak{g} = \mathsf{M}(p)$.
(ii) If $G = \mathsf{O}(p)$ or $\mathsf{SO}(p)$ with matrix multiplication as the group action, then $\mathfrak{g} = \mathsf{Sk}(p)$, the set of skew-symmetric matrices.
(iii) If $G = \mathsf{LT}_+(p)$ with matrix multiplication as the group action, then $\mathfrak{g}$ is the set $\mathsf{LT}(p)$ of lower triangular matrices.
(iv) If $G = \mathsf{LT}_{\mathsf{u}}(p)$ with matrix multiplication as the group action, then $\mathfrak{g}$ is the set $\mathsf{LT}_{\mathsf{s}}(p)$ of lower triangular matrices with zeros along the diagonal.
(iv) If $G = \mathsf{PD}(p)$ with logarithmic addition as the group action (Arsigny et al., 2017) then $\mathfrak{g} = \mathsf{Sym}(p)$.

When G is a matrix Lie group, the usual matrix exponential can be related to the map $\mathbf{expo} : \mathfrak{g} \to G$ such that $\mathbf{expo}(A) = e^A$. Properties of $\mathbf{expo}$ (e.g., injectivity) will depend on that of the matrix exponential, and hence on the topology of G. If G is compact and connected then e is surjective. If it is injective in a small neighbourhood around the origin in $\mathfrak{g}$, then it bijective. Since G is also a differentiable manifold, a geometric characterization of the matrix exponential e is that it will coincide with the Riemannian exponential map under a bi-invariant Riemannian metric on G.

A good source of reference for matrix groups is Baker (2001).

B. Formalized definition of reparametrization

Let $\mathbf{vec} : \mathsf{Sym}(p) \to \mathbb{R}^{p^2}$ be the vectorization map taking a symmetric matrix to a p^2-dimensional column vector. Define the half-vectorization map

$$\mathbf{vech} : \mathsf{Sym}(p) \to \mathbb{R}^{p(p+1)/2}, \quad \mathbf{vech}(x) := A\,\mathbf{vec}(x),$$

where the matrix

$$A := \sum_{i\geq j}(u_{ij} \otimes e_j^{\mathrm{T}} \otimes e_i^{\mathrm{T}}) \in \mathbb{R}^{p(p+1)/2\times p^2}$$

picks out the upper triangular part of the vectorization, and u_{ij} is a $p(p+1)/2$-dimensional unit vector with 1 in position $(j-1)p + i - j(j-1)/2$ and 0 elsewhere. Its inverse

$$\mathbb{R}^{p(p+1)/2} \ni x \mapsto \mathbf{vech}^{-1}(x) := (\mathbf{vech}(I_p)^{\mathrm{T}} \otimes I_p)(I_p \otimes x) \in \mathsf{Sym}(p)$$

exists through the Moore-Penrose inverse of A. Let

$$\mathsf{Cone}_p := \left\{\sigma \in \mathbb{R}^{p(p+1)/2} : \left\langle \mathbf{vech}^{-1}(\sigma)y, y\right\rangle > 0,\ y \in \mathbb{R}^p\right\},$$

be a constrained set within $\mathbb{R}^{p(p+1)/2}$. The definition of Cone_p implicitly engenders an injective parametrization

$$f : \mathsf{Cone}_p \to \mathsf{Sym}(p), f(\sigma) = \mathbf{vech}^{-1}(\sigma),$$

with image $f(\mathsf{Cone}_p) = \mathsf{PD}(p) \subset \mathsf{Sym}(p)$. Since $\mathsf{PD}(p)$ is open in $\mathsf{Sym}(p)$, with respect to f it is a parametrized submanifold of $\mathsf{Sym}(p)$.

A *reparametrization* of $\mathsf{PD}(p)$ corresponds to an injective map $h : N \to \mathsf{Sym}(p)$ from a domain N with non-singular derivative such that there is a diffeomorphism $\psi : N \to \mathsf{Cone}_p$ with $h = f \circ \psi$. The derivative condition ensures that N is of dimension $p(p+1)/2$. The two parametrizations f and h are said to be equivalent since $f(\mathsf{Cone}_p) = h(N) = \mathsf{PD}(p)$, and h is a reparametrization of f (and vice versa). The commutative diagram in the left half of Figure B.1 illustrates the type of reparametrization used in this paper.

In contrast, the statistical model or manifold is determined via an injective map $g : \mathsf{PD}(p) \to \left\{\mathbb{P}_\Gamma : \Gamma \in \mathsf{PD}(p)\right\}$ that maps a covariance matrix Γ to a parametric probability measure $\mathbb{P}_\Gamma$ on some sample space. Every Σ is obtained from unique points in Cone_p and N, and the statistical model given by g is impervious to reparametrization of the manifold $\mathsf{PD}(p)$. Reparametrization of the statistical model amounts to applying

$$\begin{tikzcd}
N \arrow[r, "\psi"] \arrow[dr, "h"'] & \mathsf{Cone}_p \arrow[d, "f"] & \\
& \mathsf{PD}(p) \arrow[r, "g"] & \left\{\mathbb{P}_\Gamma : \Gamma \in \mathsf{PD}(p)\right\}
\end{tikzcd}$$

Fig. B.1: The map h is a reparameterization of $\mathsf{PD}(p)$ initially parameterized using f. In contrast, map g parametrizes a statistical manifold of parametric probability measures.

a diffeomorphism $\mathsf{PD}(p) \to \mathsf{PD}(p)$ such that correspondences $\Gamma \mapsto \mathbb{P}_\Gamma$ change but not the image $g(\mathsf{PD}(p))$. This form of reparametrization is not considered in the present work.

A matrix $\Sigma \in \mathsf{PD}(p)$ with respect to the parametrization f is sparse if $\sigma \in \mathsf{Cone}_p$ is sparse. On the other hand, the structure of sparsity in Σ with respect to the domain N depends on the map h and how N is prescribed coordinates. We will use sparsity to refer to the domain or to the range of a parametrization interchangeably, with context disambiguating the two.

C. Legitimacy of the four maps from Section 3

Consider first the Σ_{pd} parametrization. Starting from the natural parametrization $\sigma \mapsto f(\sigma) = \Sigma$, where $\sigma \in \mathsf{Cone}_p$, the same Σ is reached via the more circuitous route $\sigma \mapsto e \circ b_{sym} \circ \phi_{pd}(\sigma)$ involving the composition of three maps (see the upper left panel of Figure C.2). This composition consists of a diffeomorphism $\phi_{pd} : \mathsf{Cone}_p \to \mathbb{R}^{p(p+1)/2}$ that takes σ to α, a bijective map $b_{sym} : \mathbb{R}^{p(p+1)/2} \to \mathsf{Sym}(p)$ that maps $\alpha \in \mathbb{R}^{p(p+1)/2}$ to the symmetric matrix $L(\alpha)$ via the expansion (2) in the canonical basis $\mathcal{B}_{sym}$, and the matrix exponential $e : \mathsf{Sym}(p) \to \mathsf{PD}(p)$. The legitimacy of this parametrization is ensured by Propositions C.2 and C.3 below.

The second parameterization Σ_o is also not new: this was considered by Rybak & Battey (2021) who applied the matrix logarithm to O in the spectral decomposition $\Sigma = O\Lambda O^{\mathrm{T}}$, where $O \in \mathsf{O}(p)$ is an orthonormal matrix of eigenvectors and $\Lambda = e^D \in \mathsf{D}_+(p)$ is a diagonal matrix of corresponding eigenvalues. Without loss of generality Rybak & Battey (2021) took the representation in which $O \in \mathsf{SO}(p)$ and considered the map $\log : \mathsf{SO}(p) \to \mathsf{Sk}(p)$, yielding a different vector space from that in Σ_{pd} in which to study sparsity. Allowance for additional sparsity via $d = \mathrm{diag}(D)$ can be easily incorporated and corresponds to further structure. The composition of three maps described in Figure 1 (top right) consists of a map $\phi_o : \mathsf{Cone}_p \to \mathbb{R}^{p(p+1)/2}$, a map b_{sk} from (α, d) to $D(d)$ and $L(\alpha)$ via (2) in the canonical basis $\mathcal{B}_{sk}$, and the matrix exponential $e : \mathsf{Sk}(p) \to \mathsf{PD}(p)$ and $e : \mathsf{D}(p) \to \mathsf{D}_+(p)$. The situation regarding invertibility of the maps is more nuanced than for Σ_{pd}, owing to the non-uniqueness of the decomposition $\Sigma = O\Lambda O^{\mathrm{T}}$ and multivaluedness of the matrix logarithm of $O \in \mathsf{SO}(p)$. For the purpose of the present paper, the implications are negligible, as we can make the parametrization injective under some conditions in Σ. This is clarified in Proposition C.4.

The situation is analogous for the two new reparametrization maps Σ_{lt} and Σ_{ltu}, depicted in the bottom row of Figure C.2. The constructions can alternatively be expressed in terms of upper triangular matrices with analogous parametrizations Σ_{ut} and Σ_{utu} and there are no substantive differences in the conclusions of section 4 and section 5. As with Σ_o, invertibility of Σ_{lt} is not guaranteed without further constraints, since $e : \mathsf{LT}(p) \to \mathsf{LT}_+(p)$ is not injective, while the Σ_{ltu} parametrization enjoys invertibility without any restrictions on the parameter domain (Proposition C.4). The so-called LDL decomposition of Σ is $\Sigma = U\Psi U^{\mathrm{T}}$, $U \in \mathsf{LT}_{\mathsf{u}}(p)$ where $\Psi = e^D \in \mathsf{D}_+(p)$. Analogously to the previous cases, the matrix logarithm $\log : \mathsf{LT}_{\mathsf{u}}(p) \to \mathsf{LT}_{\mathsf{s}}(p)$ is applied to U and represented in the canonical basis $\mathcal{B}_{ltu}$.

$$\begin{tikzcd}
\alpha \arrow[d, "b_{sym}"] & \sigma \arrow[l, "\phi_{pd}"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L \in \mathsf{Sym}(p) \arrow[r, "e^L"] & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}$$

$$\begin{tikzcd}
(\alpha, d) \arrow[d, shift right=1.5ex] \arrow[d, shift left=1.5ex, "b_{sk}"'] & \sigma \arrow[l, "\phi_o"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L, D \in \mathsf{Sk}(p) \times \mathsf{D}(p) \arrow[r, "e^L e^D (e^L)^{\mathrm{T}}"] & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}$$

$$\begin{tikzcd}
\alpha \arrow[d, "b_{lt}"] & \sigma \arrow[l, "\phi_{lt}"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L \in \mathsf{LT}(p) \arrow[r, "e^L (e^L)^{\mathrm{T}}"] & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}$$

$$\begin{tikzcd}
(\alpha, d) \arrow[d, shift right=1.5ex] \arrow[d, shift left=1.5ex, "b_{ltu}"'] & \sigma \arrow[l, "\phi_{ltu}"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L, D \in \mathsf{LT_s}(p) \times \mathsf{D}(p) \arrow[r, "e^L e^D (e^L)^{\mathrm{T}}"] & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}$$

Fig. C.2: Reparametrization maps for Σ_{pd} (top left), Σ_o (top right), Σ_{lt} (bottom left), and Σ_{ltu} (bottom right).

Proposition C.1 establishes existence of the maps $\phi_\bullet$ introduced above.

Proposition C.1. *The convex set Cone_p is diffeomorphic to $\mathbb{R}^{p(p+1)/2}$.*

Proof. The set $\mathsf{PD}(p)$ is a symmetric space of dimension $p(p+1)/2$ of noncompact type and can thus has nonpositive (sectional) curvature when equipped with a Riemannian structure (Helgason, 2001). The map $\mathbf{vech}^{-1} : \mathsf{Cone}_p \to \mathsf{PD}(p)$ is injective, and we can thus pullback the metric from $\mathsf{PD}(p)$ to Cone_p making it non-positively curved. The set Cone_p is simply connected and complete, and by the Cartan-Hadamard theorem (Helgason, 2001) it is diffeomorphic to $\mathbb{R}^{p(p+1)/2}$. □

The inverses $\phi_\bullet^{-1}$ determine precisely how sparsity in α or (α, d) manifests in a point in the convex cone, and thus, quite straightforwardly, in the covariance matrix $\Sigma(\alpha)$. However, they are difficult to determine in closed form. The maps $\Sigma_\bullet$ prescribe a path from α to $\Sigma(\alpha)$ (and similarly for (α, d)) and can be viewed as suitable surrogates, but need not be diffeomorphisms even when injectivity is guaranteed.

Injectivity of $\Sigma_{pd}, \Sigma_o, \Sigma_{lt}$ and Σ_{ltu} hinge on injectivity of the matrix exponential, and uniqueness of eigen, Cholesky and LDL decompositions for the latter three. We first consider the matrix exponential.

Propositions C.2 and C.3 describe conditions for existence and uniqueness of the matrix logarithm, which affect invertibility of the four reparametrizations in §3.

Proposition C.2 (Culver (1966)). *Let $M \in M(p)$. There exists an $L \in M(p)$ such that $M = e^L$ if and only if $M \in GL(p)$ and each Jordan block of M corresponding to a negative eigenvalue occurs an even number of times.*

Proposition C.3 (Culver (1966)). *Let $M \in M(p)$ and suppose that a matrix logarithm exists. Then $M = e^L$ has a unique real solution L if and only if all eigenvalues of M are positive and real, and no elementary divisor (Jordan block) of M corresponding to any eigenvalue appears more than once.*

Proposition C.2 covers all four logarithm maps $\log : \mathsf{PD}(p) \to \mathsf{Sym}(p)$, $\log : \mathsf{SO}(p) \to \mathsf{Sk}(p)$, $\log : \mathsf{LT}_+(p) \to \mathsf{LT}(p)$ and $\log : \mathsf{LT_u}(p) \to \mathsf{LT_s}(p)$. Conditions that ensures uniqueness in Proposition C.3 are satisfied only by $\log : \mathsf{PD}(p) \to \mathsf{Sym}(p)$ and $\log : \mathsf{LT_u}(p) \to \mathsf{LT_s}(p)$. A geometric version of the sufficient condition ("if" part) in Proposition C.3 claims uniqueness if M lies in the ball $\mathscr{B}_{I_p}(1) := \{X \in \mathsf{M}(p) : \|X - I_p\|_2 < 1\}$ around I_p, where $\|X\|_2$ is the spectral norm of X. This provides a sufficient (not necessary) condition to ensure that $\log(\exp Y) = Y$.

Relatedly, perhaps more appropriate from the perspective of reparametrization of $\mathsf{PD}(p)$, are conditions that ensure injectivity of the matrix exponential $e : \mathsf{M}(p) \to \mathsf{GL}(p)$. As a consequence of Proposition C.3, e is injective when restricted to Lie subalgebras $\mathsf{Sym}(p)$ and $\mathsf{LT_s}(p)$, but not $\mathsf{LT}(p)$ and $\mathsf{Sk}(p)$. The geometric version of the sufficient condition in Proposition C.3 then asserts that the matrix exponential is injective when restricted to the ball $\mathscr{B}_0(\ln 2) := \{L \in \mathsf{M}(p) : \|L\|_2 < \ln 2\}$ around the origin 0 (zero matrix) within $\mathsf{M}(p)$ (e.g. Baker, 2001, Proposition 2.4). This provides a sufficient (not necessary) condition to ensure that $\exp(\log X) = X$. The condition is close to being necessary for $e : \mathsf{Sk}(p) \to \mathsf{SO}(p)$ and $e : \mathsf{LT}(p) \to \mathsf{LT}_+(p)$.

For example, $e : \mathsf{Sk}(p) \rightarrow \mathsf{SO}(p)$ with $p = 2$ maps

$$2\pi n B_1 = 2\pi n \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2\pi n \\ -2\pi n & 0 \end{pmatrix}, \quad n \in \mathbb{Z}$$

to the identity I_2; thus, $\Sigma_o((2\pi, d)) = \Sigma_o((4\pi, d))$ for any fixed $d \in \mathbb{R}^2$. The issue arises because skew-symmetric matrices of the form considered comprise the kernel of $e : \mathsf{Sk}(p) \rightarrow \mathsf{SO}(p)$. We see that $\|2\pi B_1\|_2 > \ln 2$ and thus violates the sufficient condition.

Moving on to the decompositions, the following proposition elucidates on conditions that ensure injectivity of the four maps from §3 and legitimize them as reparametrizations of $\mathsf{PD}(p)$.

Proposition C.4.

(i) The maps $\alpha \mapsto \Sigma_{pd}(\alpha)$ and $\alpha \mapsto \Sigma_{ltu}$ are injective on $\mathbb{R}^{p(p+1)/2}$.

(iii) Assume that the elements of d are distinct. The map $(\alpha, d) \mapsto \Sigma_o(\alpha, d)$ is injective when α is restricted to $N_o \subset \mathbb{R}^{p(p-1)/2}$ such that the image $b_{sk}(N_o \times \mathbb{R}^p) \subseteq \mathscr{B}_0(\ln 2) \times \mathsf{D}(p)$ within $\mathsf{Sk}(p) \times \mathsf{D}(p)$, and upon choosing ORP^{T} for a particular permutation $P \in \mathsf{P}(p)$ of and combination of signs $R \in \mathsf{D}(p) \cap \mathsf{O}(p)$ for the columns of O and a permutation $P\Lambda P^{\mathsf{T}}$ of elements of Λ, where b_{sk} is as in §3.

(ii) The map $\alpha \mapsto \Sigma_{lt}(\alpha)$ is injective when restricted to $N_{lt} \subset \mathbb{R}^{p(p+1)/2}$ such that the image $b_{lt}(N_{lt}) \subseteq \mathscr{B}_0(\ln 2)$ within $\mathsf{LT}(p)$, where b_{lt} is as in §3.

Proof. Injectivity of Σ_{pd} follows from Proposition C.3. It is well-known that the Cholesky and LDL decompositions as used in the definitions of Σ_{lt} and Σ_{ltu} respectively are unique (Golub & Van Loan, 2013). Uniqueness of the LDL decomposition also stems from uniqueness of the Iwasawa decomposition of $\mathsf{GL}(p)$ (Terras, 1988) through the identification $\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p)$. The maps $L(\alpha) \mapsto e^{L(\alpha)}(e^{L(\alpha)})^{\mathsf{T}}$ and $(L(\alpha), D(d)) \mapsto e^{L(\alpha)}e^{D(d)}(e^{L(\alpha)})^{\mathsf{T}}$ are thus injective. When combined with injectivity of the exponential map $e : \mathsf{LT}_{\mathsf{s}} \rightarrow \mathsf{LT}_{\mathsf{u}}$ from Proposition C.4, the parameterization Σ_{ltu} is injective.

The situation concerning uniqueness of the eigen decomposition $\Sigma = O\Lambda O^{\mathsf{T}}$ is involved, even after restriction to a subset of $\mathsf{Sk}(p)$ that renders the exponential $e : \mathsf{Sk}(p) \rightarrow \mathsf{SO}(p)$ injective. First note that $O \in \mathsf{SO}(p)$ under our parameterization using the exponential map. Then, observe that $ORR^{\mathsf{T}}\Lambda RR^{\mathsf{T}}O^{\mathsf{T}} = O\Lambda O^{\mathsf{T}}$ for any $R \in \mathsf{SO}(p)$, and thus pairs $(OR, R^{\mathsf{T}}\Lambda R)$ map to the same Σ for *every* $R \in \mathsf{SO}(p)$ for which $R^{\mathsf{T}}\Lambda R = \Lambda$, since $OR \in \mathsf{SO}(p)$. Indeed, $\{R \in \mathsf{SO}(p) : R^{\mathsf{T}}\Lambda R = \Lambda\}$ fixes Λ and is the isotropy subgroup $\mathsf{SO}(p)_\Lambda$ in $\mathsf{SO}(p)$ (see Supplementary Material A). In addition to $\mathsf{SO}(p)_\Lambda$, another source of indeterminacy comes from permutations OP and $P\Lambda P^{\mathsf{T}}$, with $P \in \mathsf{P}(p)$ and $|P| = 1$ so that $P \in \mathsf{SO}(p)$. Put together, this implies that every pair $(ORP, P\Lambda P^{\mathsf{T}})$ maps to the same Σ as long as $R \in \mathsf{SO}(p)_\Lambda$ and $P \in \mathsf{SO}(p)$.

The situation can be salvaged if the positive elements of Λ are all distinct so that $\mathsf{SO}(p)_\Lambda$ reduces to the set $\mathsf{D}(p) \cap \mathsf{SO}(p)$ of diagonal rotation matrices with ±1 entries (Grossier et al., 2021, Theorem 3.3). In this case, the map $\pi : \mathsf{SO}(p) \times \mathsf{D}_+(p) \rightarrow \mathsf{PD}(p)$ is a $2^{p-1}p!$ covering map with fibers $\pi^{-1}(\Sigma)$ consisting of matrices obtained by $p!$ permutations of elements of Λ, and a similar permutation of eigenvectors of Σ, and 2^{p-1} matrices in the set $\mathsf{D}(p) \cap \mathsf{SO}(p)$ of diagonal matrices mentioned above with unit determinant, which determine signs of the eigenvectors of Σ; there are 2^{p-1} such diagonal matrices and not 2^p owing to the unit determinant constraint.

Uniqueness can be ensured upon identifying a global *cross section* $S_o \subset \mathsf{SO}(p) \times \mathsf{D}_+(p)$ that picks out one element from every fiber such that $\pi^{-1}(\Sigma) \cap S_o$ is a singleton for every $\Sigma \in \mathsf{PD}(p)$. For example, S_o can be defined by selecting a particular permutation $P\Lambda P^{\mathsf{T}}$ and ORP of the eigenvalues and eigenvectors of Σ (e.g., elements of Λ arranged in a decreasing order); since $R \in \mathsf{D}(p) \cap \mathsf{SO}(p)$, a fixed rule for choosing signs of the eigenvectors determines a unique R. Then, S_o contains pairs $(ORP^{\mathsf{T}}, P\Lambda P^{\mathsf{T}})$ for a fixed permutation $P \in \mathsf{SO}(p)$. The cross section S_o is bijective with the quotient $(\mathsf{SO}(p) \times \mathsf{D}_+(p))/\sim$ under the equivalence relation $\sim$ that identifies any two pairs (O, Λ) that map to the same Σ.

The proof for injectivity of Σ_{lt} follows upon noting that the exponential map $e : \mathsf{LT}(p) \rightarrow \mathsf{LT}_+(p)$ is injective when restricted to the given ball within $\mathsf{LT}(p)$. This completes the proof.

D. The Iwasawa decomposition of $\mathsf{GL}(p)$ and its Lie algebra

The matrix $X^{\mathrm{T}}X$ is positive definite for every $X \in \mathsf{GL}(p)$, and the map $X \mapsto X^{\mathrm{T}}X$ is invariant to the action $(X, O) \to OX$ for $O \in \mathsf{O}(p)$ of the orthogonal group. A positive definite matrix S can be transformed to any other under the transitive action

$$(X, S) \to XSX^{\mathrm{T}}, \quad X \in \mathsf{GL}(p),\ S \in \mathsf{PD}(p)$$

of $\mathsf{GL}(p)$, and $\mathsf{PD}(p)$ is hence a homogeneous space: a differentiable manifold with a transitive differentiable action of $\mathsf{GL}(p)$. For example, between any pair $S_1, S_2 \in \mathsf{PD}(p)$ the invertible matrix $X = S_2^{1/2}S_1^{-1/2}$ transforms S_1 to S_2 under the above action. The orthogonal group $\mathsf{O}(p)$ is the stabilizer of $X = I_p$ and fixes $X \in \mathsf{GL}(p)$, and we thus obtain the identification with $\mathsf{GL}(p)$ via the group isomorphism

$$\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p),$$

where $\mathsf{GL}(p)/\mathsf{O}(p)$ is the set of equivalence classes $[X] := \{OX : O \in \mathsf{O}(p)\}$ or orbits of elements $X \in \mathsf{GL}(p)$. The benefit with this representation of $\mathsf{PD}(p)$ lies in the use of the Iwasawa decompositions of the group $\mathsf{GL}(p)$, and its lie algebra $\mathsf{M}(p)$, to define new parametrizations of $\mathsf{PD}(p)$; the decomposition of $\mathsf{GL}(p)$ corresponds to the LDL decomposition of $\mathsf{GL}(p)$.

The *Iwasawa decomposition* of $X \in \mathsf{GL}(p)$ determines a unique triple $(O, D, U) \in \mathsf{O}(p) \times \mathsf{D}_+(p) \times \mathsf{LT}_{\mathsf{u}}(p)$ such that $X = ODU$ (see e.g. Terras, 1988, Ch. 4). Since DU is lower triangular with positive diagonal entries, we also recover the well-known QR decomposition. From the Iwasawa decomposition we have $X^{\mathrm{T}}X = UD^2U^{\mathrm{T}} \in \mathsf{PD}(p)$, and we recover the unique LDL decomposition of the positive definite matrix $X^{\mathrm{T}}X$ (Golub & Van Loan, 2013). Additionally, the Iwasawa decomposition into $\mathsf{O}(p)$, $\mathsf{D}_+(p)$ and $\mathsf{LT}_{\mathsf{u}}(p)$ at the group level ($\mathsf{GL}(p)$) has a corresponding decomposition of the Lie algebra of $\mathsf{GL}(p)$:

$$\mathsf{M}(p) = \mathsf{Sk}(p) \oplus \mathsf{D}(p) \oplus \mathsf{LT}_{\mathsf{s}}(p), \tag{D.1}$$

where $\mathsf{Sk}(p)$, $\mathsf{D}(p)$ and $\mathsf{LT}_{\mathsf{s}}(p)$ are the Lie algebras of $\mathsf{O}(p)$, $\mathsf{D}_+(p)$ and $\mathsf{LT}_{\mathsf{u}}(p)$, respectively.

E. Unification of the four fundamental parametrizations

The four parametrizations considered in this work are based on the Iwasawa decomposition of $\mathsf{GL}(p)$ and its Lie algebra $\mathsf{M}(p)$, since the matrices $L(\alpha)$ and $D(d)$ are elements of the Lie algebras in (D.1). The map $\alpha \mapsto \Sigma_{lt}(\alpha)$ is based on the sum $\mathsf{LT}_s(p) \oplus \mathsf{D}(p)$ of two constituent Lie subalgebras from (D.1), which coincides with another Lie subalgebra $\mathsf{LT}(p)$ of $\mathsf{M}(p)$ consisting of all lower triangular matrices. The parametrization Σ_{ltu} represents a full use of the Iwasawa decomposition (D.1).

The parameterization Σ_{pd} relates to the Iwasawa decomposition via the *Cartan decomposition* (Terras, 1988, p.268) of the the Lie algebra $\mathsf{M}(p)$ of $\mathsf{GL}(p)$:

$$\mathsf{M}(p) = \mathsf{Sk}(p) \otimes \mathsf{Sym}(p),$$

which at the group level corresponds to the singular value decomposition of an invertible matrix. By further decomposing the symmetric part of the Cartan decomposition, the Iwasawa decomposition represents a refinement. In other words, since every $L \in \mathsf{Sym}(p)$ can be decomposed as $L = L_s + L_s^{\mathrm{T}} + D$ for $L_s \in \mathsf{LT}_{\mathsf{s}}(p)$ and $D \in \mathsf{D}(p)$, we have that

$$\mathsf{Sym}(p) = \mathsf{LT}_{\mathsf{s}}(p) \oplus \mathsf{D}(p).$$

The identification $\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p)$ implies that the orthogonal component of $\mathsf{GL}(p)$ is ignored in Σ_{pd}, and the Lie algebra $\mathsf{Sk}(p)$ of the orthogonal group $\mathsf{O}(p)$ containing the skew-symmetric parts of $\mathsf{GL}(p)$ is thus unused.

The Σ_o parametrization, on the other hand, uses the Lie algebras $\mathsf{Sk}(p)$ and $\mathsf{D}(p)$ in (D.1). However, since every skew symmetric $L \in \mathsf{Sk}(p)$ can be decomposed as $L = L_s - L_s^{\mathrm{T}}$ with $L_s \in \mathsf{LT}_{\mathsf{s}}(p)$, the Lie algebra $\mathsf{Sk}(p)$ can be generated from the Lie algebra $\mathsf{LT}(p)$, and thus links the parameterization Σ_o with the Iwasawa decomposition of $\mathsf{GL}(p)$.

F. Change of basis

A change of a matrix basis $\mathcal{B} = \{B_1, \dots, B_d\}$ is achieved by the action of a nonsingular $W \in \mathsf{GL}(p)$ as $W\mathcal{B}W^{-1} := \{WB_1W^{-1}, \dots, WB_dW^{-1}\}$. The group $\mathsf{GL}(p)$ acts equivariantly on the map $\sum_j \alpha_j B_j \mapsto \Sigma(\alpha) = e^{\sum_j \alpha_j B_j}$, since $e^{WAW^{-1}} = We^AW^{-1}$ for every $A \in \mathsf{M}(p)$ and $W \in \mathsf{GL}(p)$. Hence,

$$e^{W(\sum_j \alpha_j B_j)W^{-1}} = We^{\sum_j \alpha_j B_j}W^{-1} = W\Sigma(\alpha)W^{-1},$$

may belong to $\mathsf{PD}(p)$ depending on the W chosen. The four maps $\Sigma_{pd}, \Sigma_o, \Sigma_{lt}, \Sigma_{ltu}$ are thus well-defined only upon fixing a basis $\mathcal{B}$ for the considered Lie subalgebra.

G. Proofs for Section 4

G.1. *Preliminary lemmas*

Lemma G.1 (Axler, 2015). *Let V be a real inner-product space and let $T : V \longrightarrow V$ be a linear operator on V with matrix representation $M = M(T)$. The following are equivalent: (i) M is normal; (ii) there exists an orthonormal basis of V such that $M = O\tilde{B}O^{-1}$ where O is orthogonal and the blocks of the block-diagonal matrix $\tilde{B}$ are either 1×1 or 2×2 of the form*

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{G.1}$$

where $a, b \in \mathbb{R}$, $b > 0$, $\rho > 0$ and $\theta \in [0, 2\pi]$. Each 1×1 block λ is an eigenvalue of M, and for each 2×2 block (G.1)*, $a + bi$ and $a - bi$ are eigenvalues of M.*

The representation G.1 in terms of polar coordinates is convenient for subsequent calculations involving the matrix logarithm.

Lemma G.2. *Let $M \in M(p)$ be a normal matrix. The matrix logarithm L, if it exists, takes the form OBO^{-1}, where $O \in \mathsf{O}(p)$ is orthonormal and B is block diagonal with blocks of the form described in Lemma G.1.*

Proof. By Lemma G.1, $M = O\tilde{B}O^{-1}$, where $O \in \mathsf{O}(p)$ is orthonormal and $\tilde{B}$ is block diagonal. Let λ be an eigenvalue of M. From Proposition C.2, existence of a logarithm requires that any negative real eigenvalues have associated with them an even number of blocks. By Lemma G.1 negative eigenvalues appear in 1×1 blocks, since 2×2 blocks correspond to complex conjugate pairs of eigenvalues. It follows that the matrix logarithm of a normal matrix exists if and only if negative eigenvalues have even multiplicity, in which case, we can without loss of generality construct blocks of size 2×2 for a negative eigenvalue λ of the form $\tilde{B}_\lambda = \lambda I_2 = -|\lambda| I_2$. Then $\log(\tilde{B}_\lambda) = \log\{(|\lambda| I_2)(-I_2)\}$, and since I_2 and $-I_2$ commute, $\log(\tilde{B}_\lambda) = \log(|\lambda| I_2) + \log(-I_2)$, where

$$\log(-I_2) = \pi \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Thus

$$\log(\tilde{B}_\lambda) = \begin{pmatrix} \log|\lambda| & -\pi \\ \pi & \log|\lambda| \end{pmatrix},$$

which is of the form in Lemma G.1. For 2×2 blocks $\tilde{B}_{\mathbb{C}}$ corresponding to complex conjugate pairs of eigenvalues of M, a similar argument together with

$$\log \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}$$

shows that

$$\log(\tilde{B}_{\mathbb{C}}) = \begin{pmatrix} \log\rho & -\theta \\ \theta & \log\rho \end{pmatrix}$$

which is also of the form of Lemma G.1. □

LEMMA G.3. *Let $M = e^L \in \mathsf{M}(p)$. Then M is normal if and only if L is normal.*

Proof. Suppose that L is normal, that is $L^{\mathrm{T}}L = LL^{\mathrm{T}}$. By the Jordan decomposition $L = QJQ^{-1}$, normality of L implies normality of J. The matrix exponential $M = \exp(L) = Q\exp(J)Q^{-1}$ is normal if and only if $\exp(J)^{\mathrm{T}}\exp(J) = \exp(J)\exp(J)^{\mathrm{T}}$. Two general properties of the matrix exponential are that for matrices $A, B \in \mathsf{M}(p)$ such that $AB = BA$, $\exp(A)^{\mathrm{T}} = (\exp(A))^{\mathrm{T}}$ and $\exp(A)\exp(B) = \exp(A + B)$. Thus $\exp(J^{\mathrm{T}})\exp(J) = \exp(J + J^{\mathrm{T}}) = \exp(J)\exp(J^{\mathrm{T}})$ showing that M is normal. The converse statement follows by Lemmas G.1 and G.2. □

LEMMA G.4 (WEIERSTRASS'S M-TEST, E.G. WHITTAKER AND WATSON, 1965, P.49). *Let $S_k(x) = s_1(x) + \cdots + s_k(x)$ be a sequence of functions such that, for all x within some region $\mathcal{R}(x)$, $S_k(x) \leq T_k = t_1 + \cdots + t_k$, where $(t_j)_{j\in\mathbb{N}}$ are independent of x and T_k is a positive convergent sequence. Then $S_k(x)$ converges to some limit, $S(x)$ say, uniformly over $\mathcal{R}(x)$.*

LEMMA G.5. *For $A \in \mathsf{M}(p)$, define $\psi(A) = \sum_{k=0}^{\infty} A^k/(k+1)!$. Then for any operator norm $\|\cdot\|_{op}$, provided that $\|A\|_{op}$ is bounded, $\|\psi(A)\|_{op} \leq \|\exp(A)\|_{op}$. Additionally, $\psi(A) \in \mathsf{GL}(p)$ for any $A \in \mathsf{M}(p)$.*

Proof. Let $\psi_k(A) = \sum_{n=0}^{k} A^n/(n+1)!$ and let $r \geq 0$ be such that $\|A\|_{\mathrm{op}} \leq r$. Since the operator norm is subadditive and submultiplicative

$$\|\psi_k(A)\|_{\mathrm{op}} \leq \sum_{n=0}^{k} \frac{\|A^n\|_{\mathrm{op}}}{(n+1)!} \leq \sum_{n=0}^{k} \frac{\|A\|_{\mathrm{op}}^n}{n!} \leq \sum_{n=0}^{k} \frac{r^n}{n!} \to e^r$$

as $k \to \infty$. Lemma G.4 applies with $S_k(A) = \|\psi_k(A)\|_{\mathrm{op}}$.

For the second statement, it suffices by the Jordan decomposition of A to show that $\tilde{\psi}(\lambda) \neq 0$ where $\lambda \in \mathbb{C}$ is an eigenvalue of A and $\tilde{\psi}(x) = \sum_{n=0}^{k} x^n/(n+1)!$ for $x \in \mathbb{C}$. For $\lambda = 0$, $\psi(\lambda) = 1$ by definition, whereas for $z \in \mathbb{C}$, $z \neq 0$, $\psi(z) = (e^z - 1)/z \neq 0$. □

LEMMA G.6. *Consider $M = e^L \in \mathsf{M}(p)$ where $L \in \mathsf{V}(p)$, a vector space with canonical basis $\mathcal{B}$ of dimension m. The matrix M is logarithmically sparse in the sense that $L = L(\alpha) = \alpha_1 B_1 + \cdots + \alpha_m B_m$, $B_j \in \mathcal{B}$ with $\|\alpha\|_0 = s^*$ if and only if $M = P\tilde{M}P^{\mathrm{T}}$, where $P \in \mathsf{P}(p)$ is a permutation matrix and*

$$\tilde{M} = \begin{pmatrix} O_{11} & 0 \\ 0 & I_q \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ 0 & I_q \end{pmatrix} \begin{pmatrix} O_{11}^{\mathrm{T}} & 0 \\ 0 & I_q \end{pmatrix}, \tag{G.2}$$

where $q \geq p - d_r^*$ and $O_{11} \in \mathsf{O}(p-q)$, $C_{11} \in \mathsf{M}(p-q)$. Moreover, if M is normal, C_{11} is also normal and $C_{12} = 0$ in (G.2).

Proof. Suppose first that $M = e^L$ is logarithmically sparse. By definition, L has $p - d_r^*$ zero rows. Thus, there exists a permutation matrix P such that the last $p - d_r^*$ rows of $P^T LP$ are zero. Thus write $PLP^{\top}$ as a partitioned matrix with upper blocks L_1, L_2 of dimensions $d_r^* \times d_r^*$ and $d_r^* \times p - d_r^*$, the remaining blocks being zero. From the definition of a matrix exponential,

$$e^{PLP^{\top}} = \begin{pmatrix} \exp(L_1) & \sum_{k=0}^{\infty} L_1^k L_2/(k+1)! \\ 0 & I_{p-d_r^*} \end{pmatrix}$$

which is of the form given in equation (G.2) with $O_{11} = I_{d_r^*}$. The result follows since $e^{PLP^{\top}} = Pe^L P^{\top}$.

To prove the reverse direction, we need to construct L such that,

$$\exp(L) = \begin{pmatrix} O_{11}C_{11}O_{11}^{\mathrm{T}} & O_{11}C_{12} \\ 0 & I \end{pmatrix}, \quad L = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}, \tag{G.3}$$

where the number of zero rows of L is greater or equal to $p - d_r^*$. Let $\Gamma_1 = O_{11}C_{11}O_{11}^{\mathrm{T}}$ and $\Gamma_2 = O_{11}C_{12}$. Set $A = \log(\Gamma_1)$, which exists by assumption since eigenvalues of Γ_1 are eigenvalues of $\tilde{M}$. By the matrix

Taylor series expansion of the matrix exponential,

$$\exp(L) = \begin{pmatrix} \Gamma_1 & \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!} B \\ 0 & I \end{pmatrix}.$$

It thus remains to be shown that $\sum_{k=0}^{\infty} \frac{A^k}{(k+1)!} B = \Gamma_2$ some $d_r^* \times p - d_r^*$ matrix B. Let $\psi(A) = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}$. By Lemma G.5, $\psi(A)$ is invertible, and we can take $B = \psi(A)^{-1}\Gamma_2$. Note that the matrix logarithm of $\tilde{M}$ is not unique. However, for a real eigenvalue λ of $\tilde{M}$, the logarithm of the Jordan block $J_k(\lambda)$ has periodicity $i2\pi q I$, $q \in \mathbb{Z}$ (Culver, 1966). Thus, for $\lambda = 1$, every real matrix L, $L = \log(\tilde{M})$ will have the form (G.3) in the sense of the last $p - d_r^*$ rows being equal to canonical basis vectors, with a non-zero diagonal element. The result follows by observing that for $M = P\tilde{M}P^{\mathrm{T}}$, $\log(M) = PLP^{\mathrm{T}}$. □

Lemma G.7. *Let $M \in \mathsf{M}(p)$. With $M = e^L$ and $L \in \mathsf{V}(p) \subset \mathsf{M}(p)$, a vector space of dimension m, let $M = QJQ^{-1}$ be a real Jordan decomposition of M (e.g. Horn and Johnson, 2012, p. 202) and let $\mathcal{A} \subset [p]$ denote the set of indices for columns of Q corresponding to eigenvectors whose eigenvalues are not equal to one. Thus, the cardinality $|\mathcal{A}^c|$ of the complementary set is the geometric multiplicity of the unit eigenvalue of M, and $|\mathcal{A}| = p - |\mathcal{A}^c|$. The dimension of $\mathcal{A}$ satisfies $|\mathcal{A}| \leq \max\{d_r^*, d_c^*\}$.*

Proof. Suppose that $|\mathcal{A}| \leq d_r^*$. Since $\mathbf{rank}(L) = \mathbf{rank}(L^{\mathrm{T}})$, the geometric multiplicities of the unit eigenvalues of M and M^{T} are equal. Thus, $|\mathcal{A}| \leq d_c^*$ and therefore $|\mathcal{A}| \leq \max\{d_r^*, d_c^*\}$.

To prove that $|\mathcal{A}| \leq d_r^*$, consider the real Jordan decomposition $M = QJQ^{-1}$. Let q_j denote the jth column of Q. We show that $\mathbf{span}\{q_j : j \in \mathcal{A}^c\} = \mathsf{ker}(L)$ by establishing containment on both sides. Let $v \in \mathbf{span}\{q_j : j \in \mathcal{A}^c\}$. Then there exist coefficients $\beta_j \in \mathbb{R}$ such that

$$Lv = \sum_{j \in \mathcal{A}^c} \beta_j L q_j = \sum_{j \in \mathcal{A}^c} \beta_j Q \log(J) Q^{-1} q_j = \sum_{j \in \mathcal{A}^c} \beta_j Q \log(J) e_j = 0$$

where the final equality follows since $\lambda_j = 1$ for all $j \in \mathcal{A}^c$, so the jth diagonal entry of $\log(J)$ is zero. It follows that $\mathbf{span}\{q_j : j \in \mathcal{A}^c\} \subseteq \mathsf{ker}(L)$..

For the converse containment, suppose for a contradiction that there exists $v \in \mathsf{ker}(L)$ such that $v \notin \mathbf{span}\{q_j : j \in \mathcal{A}^c\}$. Since Q has full rank, its columns are linearly independent and there exist coefficients $\beta_1, \ldots, \beta_p$, each in $\mathbb{R}$ such that $v = \beta_1 q_1 + \cdots + \beta_p q_p$. Since $q_j \in \mathsf{ker}(L)$ for $j \in \mathcal{A}^c$ by the previous argument,

$$0 = Lv = \sum_{j \in \mathcal{A}} \beta_j L q_j = \sum_{j \in \mathcal{A}} \beta_j Q \log(J) e_j.$$

By definition of $\mathcal{A}$, $J_{jj} \neq 1$ for any $j \in \mathcal{A}$, thus the equality $Lv = 0$ implies $\beta_j = 0$ for all $j \in \mathcal{A}$, a contraction, since the columns $Q\log(J)e_j$, $j \in \mathcal{A}$ are linearly independent. □

For normal matrices in $\mathsf{M}(p)$, i.e. those satisfying $M^{\mathrm{T}}M = MM^{\mathrm{T}}$, $d_r^* = d_c^*$ and Lemma G.7 recovers Lemma 2.1 of Battey (2017) and Proposition 3.1 of Rybak and Battey (2021).

G.2. *Proof of Theorem 1*

Proof. From Lemma G.6, $p - d_r^*$ rows of M are of the canonical form e_j^{T}, and since zero columns of L are zero rows of L^{T}, it is also true by Lemma G.6 applied to M^{T} that $p - d_c^*$ columns of M are of canonical form e_j. If d^* rows and columns of M are of the canonical form, then $M = P(V \oplus I_{p-d^*})P^{\mathrm{T}}$, where $P \in \mathsf{P}(p)$ and $V \in \mathsf{M}(d^*)$. The matrix logarithm of M is $L = P(\log(V) \oplus 0_{p-d^*})P^{\mathrm{T}}$. The converse direction follows by applying the exponential map to $L = P(\log(V) \oplus 0_{p-d^*})P^{\mathrm{T}}$ and invoking Lemma G.6 in the converse direction. □

G.3. *Proof of Lemma 1*

Proof. Consider a random vector $(Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}}, Y_3^{\mathrm{T}})^{\mathrm{T}}$ with a covariance matrix Σ. By the assumptions of Lemma 1

$$\Sigma = \begin{pmatrix} A & 0 & 0 \\ B & I & 0 \\ C & 0 & D \end{pmatrix} \begin{pmatrix} A^{\mathrm{T}} & B^{\mathrm{T}} & C^{\mathrm{T}} \\ 0 & I & 0 \\ 0 & 0 & D^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} AA^{\mathrm{T}} & AB^{\mathrm{T}} & AC^{\mathrm{T}} \\ BA^{\mathrm{T}} & BB^{\mathrm{T}} + I & BC^{\mathrm{T}} \\ CA^{\mathrm{T}} & CB^{\mathrm{T}} & CC^{\mathrm{T}} + DD^{\mathrm{T}} \end{pmatrix} =: \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix},$$

where the matrices A and D are lower triangular with unit entries on the diagonal. Then,

$$\Sigma_{23} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{13} = BC^{\mathrm{T}} - BA^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}AC^{\mathrm{T}} = 0,$$

since A is full-rank by definition. Consider a submatrix

$$\Sigma_{bc|a} = \begin{pmatrix} \Sigma_{23} & \Sigma_{21} \\ \Sigma_{13} & \Sigma_{11} \end{pmatrix} = \begin{pmatrix} I & \Sigma_{21}\Sigma_{11}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{23} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{13} & 0 \\ 0 & \Sigma_{11} \end{pmatrix} \begin{pmatrix} I & 0 \\ \Sigma_{11}^{-1}\Sigma_{13} & I \end{pmatrix}.$$

Since Σ_{11} is invertible, and for any two matrices M, N, with compatible dimensions $\mathrm{rank}(MN) \leq \min\{\mathrm{rank}(M), \mathrm{rank}(N)\}$, the result follows. □

H. PROOFS FOR SECTIONS 5 AND 6

H.1. *Preliminary lemmas*

LEMMA H.8. *Let $\beta_{i.jk}$ denote a regression coefficient of Y_j in a regression of Y_i on Y_j and Y_k. Let $\upsilon_{ij}(l)$ denote the effect of Y_j on Y_i along all paths of length l in a recursive directed acyclic graph. Then,*

$$\upsilon_{ij}(l) = \begin{cases} \sum_{k=j+l-1}^{i-1} \beta_{i.k[i-1]}\upsilon_{kj}(l-1) & \text{if} \quad i - j \geq l, \\ 0 \quad \text{otherwise}. \end{cases}$$

Proof Proof of Lemma H.8. We use proof by induction. Consider $l = 1$ and take any pair (i, j) such that $i \geq j + 1$. Then, $\upsilon_{ij}(1) = \beta_{i.j[i-1]}$ as claimed. Now consider $l > 1$. Every path from node j to node i can be decomposed into a path of length $l - 1$ from j to node k for some $k \in \{j + 1, \ldots, i - 1\}$ and a path with length one from k to i. The total effect of Y_j on Y_i along such a path is equal to $\beta_{i.k[i-1]}\upsilon_{kj}(l - 1)$. The total effect along all paths of length l is the sum of single paths over all nodes $k \in \{j + 1, \ldots, i - 1\}$, which yields the result. □

LEMMA H.9. *Let $\beta_{i.k[p]\setminus\{i\}}$ denote a coefficient of Y_k in a regression of Y_i on $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_p$. Let $\upsilon_{ij}^u(l)$ denote the total effect of a unit change in Y_j on Y_i along all paths of length l in an undirected graphical model with edge weights given by regression coefficients. Then,*

$$\upsilon_{ij}^u(l) = \sum_{k \neq i} \beta_{i.k[p]\setminus\{i\}}\upsilon_{kj}^u(l - 1), \quad \beta_{i.k[p]\setminus\{i\}} = -V_{ij}/V_{ii}.$$

Proof. The proof is analogous to that of Lemma H.8. The only difference is that an edge can exist between any pair of nodes (i, j), $i \neq j$ and the effect of Y_j on Y_i is given by a regression coefficient $\beta_{i.j[p]\setminus\{i\}} = -\tilde{V}_{ij} = -V_{ij}/V_{ii}$ (Lauritzen, 1996). □

H.2. *Proofs of main results*

Proof Proof of Proposition 1. For an arbitrary partition $Y = (Y_a^{\mathrm{T}}, Y_b^{\mathrm{T}})^{\mathrm{T}}$, let Σ^{-1} be partitioned accordingly as

$$\Sigma^{-1} = \begin{pmatrix} \Sigma^{aa} & \Sigma^{ab} \\ \Sigma^{ba} & \Sigma^{bb} \end{pmatrix}.$$

The block upper-triangular decomposition takes the form,

$$\Sigma^{-1} = \Upsilon\Gamma\Upsilon^{\mathrm{T}} = \begin{pmatrix} I_{aa} & \Sigma^{ab}(\Sigma^{bb})^{-1} \\ 0 & I_{bb} \end{pmatrix} \begin{pmatrix} \Sigma^{aa.b} & 0 \\ 0 & \Sigma^{bb} \end{pmatrix} \begin{pmatrix} I_{aa} & 0 \\ (\Sigma^{bb})^{-1}\Sigma^{ba} & I_{bb} \end{pmatrix},$$

where $\Sigma^{aa.b} = \Sigma^{aa} - \Sigma^{ab}(\Sigma^{bb})^{-1}\Sigma^{ba}$. Then,

$$U^{-1} = \Upsilon^{\mathrm{T}} = \begin{pmatrix} I_{aa} & 0 \\ -(\Sigma^{bb})^{-1}\Sigma^{ba} & I_{bb} \end{pmatrix}.$$

The matrix of regression coefficients of Y_b in a regression of Y_a on Y_b is equal to $(\Sigma^{bb})^{-1}\Sigma^{ba}$ (Wermuth & Cox, 2004), which is the negative non-zero off-diagonal block of U^{-1}.

By partitioning Σ^{-1} recursively until Υ is upper-triangular, we obtain that the ith row of U^{-1} contains minus the regression coefficients of Y_i on $Y_1, \ldots, Y_{i-1}$. Let $\bar{U} = I - \Upsilon^{\mathrm{T}}$. Then, $\bar{U} \in \mathsf{LT}_{\mathsf{s}}(p)$ and $\bar{U}_{ij} = \beta_{i.j[i-1]}$ for $j < i$, i.e., the element (i, j) of $\bar{U}$ equals the coefficient of Y_j in a regression of Y_i on $Y_1, \ldots, Y_{i-1}$. Then,

$$U = (I - \bar{U})^{-1} = I + \sum_{j=1}^{p-1} \bar{U}^j, \tag{H.1}$$

where we used that, for a nilpotent matrix N of degree k, $(I + N)^{-1} = I + \sum_{j=1}^{k-1}(-1)^j N^j$. The result for U_{ij} follows from (H.1).

Using the properties of the matrix logarithm,

$$L = \log(U) = \log[(I - \bar{U})^{-1}] = -\log(I - \bar{U}) = \sum_{k=1}^{p-1} \frac{\bar{U}^k}{k},$$

which establishes the claim about L_{ij}. □

Proof Proof of Proposition 3. The matrix $\tilde{V}$ has entries

$$\tilde{V}_{ij} = \begin{cases} -\beta_{i.j[p]\setminus\{i\}} & \text{for} \quad i \neq j, \\ 1 & \text{for} \quad i = j. \end{cases}$$

Thus, the element (i, j) of matrix $I - \tilde{V}$ is equal to the effect of Y_j on Y_i along a path of length one. Since $(I - \tilde{V})^l = (I - \tilde{V})(I - \tilde{V})^{l-1}$, the element (i, j) of $(I - \tilde{V})^l$ is equal to the effect of Y_j on Y_i along all paths of length l. Provided that the sum on the right hand side converges, the power expansion of the matrix inverse and logarithm gives

$$\tilde{\Sigma} = \tilde{V}^{-1} = \sum_{k=0}^{\infty}(I - \tilde{V})^k$$

$$\log(\tilde{\Sigma}) = \log(\tilde{V}^{-1}) = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k}(I - \tilde{V})^k$$

Proof Proof of Lemma 2. The result follows from a power series expansion of matrix inverse, Lemma 1, and Proposition 2.1 and Corollary 2.2 of Uhler (2019). □

I. Derivation of equation (12)

A version of the following derivation appears in Battey (2019). The argument is more complicated than is necessary for the oversimplified case presented here, but the representation is helpful for showing the considerations involved in the generalisation.

A function f of a $p \times p$ matrix A satisfies (Kato, 1976, p.44)

$$f(A) = \frac{1}{2\pi i} \oint_{\gamma_A} f(z)(zI - A)^{-1} dz, \tag{I.1}$$

where I is the identity matrix and γ_A is a simple closed curve lying in the region of analyticity of f and enclosing all the eigenvalues of A in its interior.

From (I.1), the error on the scale of the matrix logarithm is

$$\log(\Sigma + \varepsilon I) - \log(\Sigma) = \frac{1}{2\pi i}\left(\oint_{\gamma_\varepsilon} \log(z)(zI - (\Sigma + \varepsilon I))^{-1} dz - \oint_{\gamma} \log(z)(zI - \Sigma)^{-1} dz\right),$$

where γ_ε must enclose γ by positivity of ε. Then provided that the eigenvalues of Σ are bounded away from zero, γ_ε can be chosen so as not to cross the imaginary axis and the previous display simplifies to

$$\begin{aligned} \log(\Sigma + \varepsilon I) - \log(\Sigma) &= \frac{1}{2\pi i}\oint_{\gamma_\varepsilon} \log(z)\{(zI - (\Sigma + \varepsilon I))^{-1} - (zI - \Sigma)^{-1}\} dz \\ &= \frac{\varepsilon}{2\pi i}\oint_{\gamma_\varepsilon} \log(z)(zI - (\Sigma + \varepsilon I))^{-1}(zI - \Sigma)^{-1} dz \end{aligned} \tag{I.2}$$

by Cauchy's theorem, where we have used that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for invertible matrices A and B, where $B - A = \varepsilon I$. Let $\Sigma = O\Lambda O^{\mathrm{T}}$ be the spectral decomposition of Σ, where O have orthonormal columns $o_1, \ldots, o_p$ and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$. Then $(zI - \Sigma)^{-1} = O(zI - \Lambda)^{-1}O^{\mathrm{T}}$ and similarly for the expression involving $\Sigma + \varepsilon I$. It follows that the (j, k)th entry of the difference in log transformations is

$$[\log(\Sigma + \varepsilon I) - \log(\Sigma)]_{j,k} = \varepsilon \sum_{r,v}\left(\frac{1}{2\pi i}\oint_{\gamma} \frac{\log(z)}{(z - (\lambda_r + \varepsilon))(z - \lambda_v)} dz\right) o_{jr} o_{kv} \sum_{\ell,s} o_{\ell r} o_{sv}, \tag{I.3}$$

which is equation (12).

J. ESTIMATION UNDER THE SPARSE Σ_{ltu} PARAMETRIZATION

J.1. *Construction of estimator*

Recall the notation $\Sigma_{ltu} = T\Omega T$, $L = \log(T) = -\log(I - B)$ and $D = \log(\Omega)$. We are primarily interested in situations where L and D are sparse. The guarantee that is typically sought for high-dimensional covariance estimators is consistency in the spectral norm under a notional double-asymptotic regime in dimension $p = p(n)$ and sample size n. Different approaches and asymptotic regimes might be considered, giving for instance, faster rates of convergence with slower permissible scaling of p with n, or vice versa. Here we show one possible estimator and derive its convergence rates in spectral norm, under the scaling $\log p/n \longrightarrow 0$. The theoretical properties are detailed in section J.2 and proved in section J.4.

The broad scheme involves constructing pilot estimators of the relevant quantities which have an elementwise consistency property, before exploiting sparsity on the transformed scale to obtain guarantees in the stronger norm. Suppose that $\tilde{T}$ and $\tilde{\Omega}$ are estimators of T and Ω that have exploited sparsity on the transformed scale, and have been shown to be consistent in spectral norm. A natural estimator of Σ is then $\tilde{\Sigma} = \tilde{T}\tilde{\Omega}\tilde{T}^{\mathrm{T}}$, which is also consistent in spectral norm.

In order to construct $\tilde{T}$ and $\tilde{\Omega}$, pilot estimators $\hat{T} = (I - \hat{B})^{-1}$ and $\hat{\Omega}$ are needed that are consistent in an elementwise sense. From these, let $\hat{L} = -\log(I - \hat{B})$ and $\hat{D} = \log(\hat{\Omega})$. The simplest way to exploit sparsity of L and D is to use a thresholding operator (Bickel & Levina, 2008b), which sets entries of $\hat{L}$ and $\hat{D}$ to zero if their absolute values are below a specified threshold. The spectral-norm consistent estimators $\tilde{T}$ and $\tilde{\Omega}$ are then obtained by defining $\tilde{T} := \exp(\tilde{L})$ and $\tilde{\Omega} := \exp(\tilde{D})$, where $\tilde{L}$ and $\tilde{D}$ are the thresholded versions of $\hat{L}$ and $\hat{D}$.

To construct elementwise-consistent estimators $\hat{B}$ and $\hat{\Omega}$, note that (10) from the main text implies,

$$(I - B)X \sim N(0, \Omega).$$

For a chain component c, let $\mathrm{pa}(c)$ denote the set of parent nodes of c. Then,

$$X_c | X_{\mathrm{pa}(c)} = N(B_c X_{\mathrm{pa}(c)}, \Omega_c). \tag{J.1}$$

The factorization of joint density implies factorization of the parameter space (see also Drton & Eichler, 2006). As a result, we can estimate B_c and Ω_c separately for each chain component. From now on we omit the subscript c to simplify the notation. Equation (J.1) suggests estimating B by regressing each node on its parents, which yields an elementwise-consistent estimator $\hat{B}$. The estimator $\hat{\Omega}$ can be obtained as a sample covariance matrix of residuals for all nodes in a given chain component (see equation (J.1)). For this, we can use estimates of regression coefficients $\hat{B}$ or alternatively, a version $\tilde{B}$ that exploits any sparsity on the transformed scale. Both result in an elementwise-consistent estimator of Ω, although $\tilde{B}$ offers some advantage in high-dimensional settings.

J.2. *Theoretical guarantees*

For an $m \times m$ matrix M, let $\|M\|_{\max} = \max_{i,j} |M_{i,j}|$, where $M_{i,j}$ denotes an entry (i, j) of M, and $\|M\|_2 = \sup_{\|w\|_2=1} \|Mw\|_2$. The largest eigenvalue of M is denoted by $\lambda_{\max}(M)$. The size of a random vector X is denoted by $|X|$. The set of parent nodes of node i and chain component c are denoted, respectively, by $\mathrm{pa}(i)$ and $\mathrm{pa}(c)$. Let $\hat{\Sigma}$ be a sample covariance matrix of X. For two sets of indices, s_1, $s_2 \subseteq [p]$, let $\hat{\Sigma}_{s_1,s_2}$ be the matrix obtained by selecting rows s_1 and columns s_2 of $\hat{\Sigma}$.

The results presented in this section are valid under a weaker assumption of sub-Gaussian rather than Gaussian distributions.

Condition J.1. For every chain component c, $X_c | X_{\mathrm{pa}(c)}$ is sub-Gaussian with a variance proxy σ_ε^2.

In addition, we assume that the covariance matrix Σ of X satisfies conditions J.2 and J.3.

Condition J.2. The quantities $\|\Sigma\|_{\max}$, $\|\Sigma^{-1}\|_{\max}$, $\|L\|_2$ and $\|\Omega\|_2$ are bounded as $n, p \to \infty$.

Condition J.3. The sequence of smallest eigenvalues of $\hat{\Sigma}$ is bounded away from zero as $p \to \infty$.

Equation J.1 suggests estimating the ith row of B, denoted by β^i, by regressing X_i on $X_{\mathrm{pa}(i)}$. Lemma J.10 establishes elementwise consistency of the resulting estimator, $\hat{\beta}^i$, which implies the consistency of $\hat{B} = (\hat{\beta}^1, \ldots, \hat{\beta}^p)^{\mathrm{T}}$.

LEMMA J.10. *Let $X_j = X_{pa(j)}\beta + \varepsilon$, where $X_j \in \mathbb{R}^n$, $X_{pa(j)} \in \mathbb{R}^{n \times |pa(j)|}$, $\beta \in \mathbb{R}^{|pa(j)|}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is sub-Gaussian with zero mean and variance proxy σ_ε^2. Then,*

$$\max_{j \in [p]} \|\hat{\beta}^j - \beta^j\|_{\max} = O_p((\log p/n)^{1/2}).$$

We now seek an estimator $\hat{L}$ of L that inherits the elementwise consistency of $\hat{B}$. As discussed in Section 5, the element (i, j) of the matrix logarithm of T corresponds to the effects of node i on node j along all directed paths connecting the two nodes. The following condition assumes that there is some length, say l^*, such that effects along paths of longer length are negligible when weighted inversely by the path length.

Condition J.4. There exists $l^* \in \mathbb{N}$, such that for any pair of nodes (i, j), $j < i$, $\sum_{l=l^*+1}^{p} \frac{1}{l} \delta_{i|j}(l) = C(\log p/n)^{\varphi/(2(\varphi+1))}$ for $\varphi > 0$, where $\delta_{i|j}(l)$ denotes the sum of effects of node j on node i along all paths of length l.

Since $L = -\log(I - B)$, a natural way of exploiting Condition J.4 is to approximate the matrix logarithm by a truncated power expansion of order l^*. Specifically, for a matrix A, define a truncated matrix logarithm of the lth order as

$$\log_{|l}(A) := \sum_{k=1}^{l} (-1)^{k+1} \frac{(A - I)^k}{k}$$

and let $\hat{L} = -\log_{|l^*}(I - \hat{B})$. Lemma J.11 establishes elementwise consistency of $\hat{L}$ under Condition J.4.

LEMMA J.11. *Let $\hat{B}$ be an estimator of B such that $\|\hat{B} - B\|_{\max} = O_p((n^{-1} \log p)^{1/2})$. Assume that Condition J.4 holds. Then,*

$$\|\hat{L} - L\|_{\max} = O_p\big((n^{-1} \log p)^{1/2}\big). \tag{J.2}$$

Simulations presented in Section J.3 suggest that the assumptions of Lemma J.11 are not necessary for bound (J.2) to hold. In particular, the rate of convergence in (J.2) is valid also for $\hat{L} = -\log(I - \hat{B})$ and in the absence of Condition J.4.

We now use the elementwise-consistent estimator $\hat{L}$ to construct an estimator $\tilde{L}$ of L consistent in the spectral norm. The key assumption to achieve consistency in a high-dimensional regime is that of sparsity. Specifically, assume that matrix L belongs to a sparse class of matrices, as stated in Condition 1, which generalizes the notion of sparsity used in most of the main paper by allowing approximate zeros. Thus,

$$L = \log(T) \in \left\{ L \in \mathsf{LT_s}(p) : \max_i \sum_{j=1}^{p} |L_{ij}|^{q_l} = s_l(p) \right\},$$

where $0 \leq q_l \leq 1$ and $s_l(p)/p \to 0$. The estimator $\tilde{L}$ is obtained by elementwise thresholding of $\hat{L}$. In particular, for a $p \times p$ matrix A, an elementwise thresholding operator $\mathcal{T}(A)$, introduced by Bickel & Levina (2008b), has the form,

$$\mathcal{T}(A)_{ij} = \mathcal{T}(A_{ij}) = A_{ij}\mathbb{I}\{|A_{ij}| > \tau\}. \tag{J.3}$$

Thus, $\tilde{L}$ has the form,

$$\tilde{L} = \mathcal{T}(\hat{L}), \quad \mathcal{T}(\hat{L})_{ij} = \hat{L}_{ij}\mathbb{I}\{|\hat{L}_{ij}| > \tau_l\}. \tag{J.4}$$

Under Condition 1, the following result follows from Theorem 1 in Bickel & Levina (2008b).

Corollary 1. *Suppose that $L \in \mathcal{U}(q_l, s_l(p))$ and $\|\hat{L} - L\|_{\max} = O_p(r_{n,p})$. Let $\tau_l \asymp r_{n,p}$ in* (J.4). *Then, $\|\tilde{L} - L\|_2 = O_p\left(s_l(p) r_{n,p}^{1-q_l}\right)$ as n, $p \to \infty$.*

The consistency of $\tilde{L}$ is sufficient to obtain an spectral-norm consistent estimator of T, as shown in Lemma J.12.

Lemma J.12. *Let $\tilde{B} = I - \exp(-\tilde{L}^{\mathrm{T}})$ and $\tilde{T} = \exp(\tilde{L})$. Then,*

$$\begin{aligned} \|B - \tilde{B}\|_2 &\leq \exp(\lambda_{\max}(L^{\mathrm{T}}L))\|L - \tilde{L}\|_2 \exp(\|L - \tilde{L}\|_2), \\ \|T - \tilde{T}\|_2 &\leq \exp(\lambda_{\max}(L^{\mathrm{T}}L))\|L - \tilde{L}\|_2 \exp(\|L - \tilde{L}\|_2). \end{aligned}$$

A direct consequence of Lemma J.12 is that thresholding in the transformed domain yields an ℓ_2-norm consistent estimator of regression coefficients β, which constitute the rows of B.

Corollary 2. *Let $\tilde{\beta}^i$ and β^i denote the ith row of $\tilde{B}$ and B respectively. Then,*

$$\|\beta^i - \tilde{\beta}^i\|_2 \leq \exp(\lambda_{\max}(L^{\mathrm{T}}L))\|L - \tilde{L}\|_2 \exp(\|L - \tilde{L}\|_2).$$

Lemma J.12, together with Corollary 1 and Lemma J.11 imply that

$$\|T - \tilde{T}\|_2 = O_p\left(s_l(p) \left(n^{-1} \log p \right)^{(1-q_l)/2} \right). \tag{J.5}$$

Since Ω is positive-definite, a spectral-norm consistent estimator $\tilde{\Omega}$ can be obtained using the approaches of Battey (2019) or Zwiernik (2025). The former requires an elementwise-consistent estimator of Ω. Given an estimator $\bar{B}$ of B, let $\hat{\Omega}$ denote a sample covariance matrix of residuals for a chain component c. Lemma J.13 establishes elementwise consistency of $\hat{\Omega}$.

Lemma J.13. *Assume Condition J.2 holds. Let $\bar{B}$ denote an estimator of B.*

1. *If $\|B - \bar{B}\|_2 = O_p(r_\beta(n, p))$ then,*

$$\|\hat{\Omega} - \Omega\|_{\max} = O_p\left(r_\beta^2(n, p)\sqrt{\frac{\log \rho_{pa}}{n}} + \sqrt{\frac{\log \rho}{n}} \right), \tag{J.6}$$

2. *If $\|B - \bar{B}\|_{\max} = O_p(r_\beta(n, p))$ then,*

$$\|\hat{\Omega} - \Omega\|_{\max} = O_p\left(\rho_{pa}^2 r_\beta^2(n, \rho)\sqrt{\frac{\log \rho_{pa}}{n}} \right), \tag{J.7}$$

where $\rho_{pa} = \max_{i \in c} |X_{pa(i)}|$*,* $\rho = |X_c|$ *and c denotes a chain component.*

If the size of chain components grows at the same rate as p, $\rho \asymp p$ and $\rho_{\mathrm{pa}} \asymp p$, under conditions of Lemma J.10, the rate of convergence in (J.6) is more advantageous than in (J.7). This suggests using $\tilde{B}$ rather than $\hat{B}$ to obtain the residual covariance matrix $\hat{\Omega}$, which yields,

$$\begin{aligned}\|\hat{\Omega} - \Omega\|_{\max} &= O_p\left(s_l(p)^2(n^{-1}\log p)^{3/2-q_l} + (n^{-1}\log p)^{1/2}\right) \\ &= O_p\left(s_l(p)^2(n^{-1}\log p)^{3/2-q_l}\right),\end{aligned}$$

where we have assumed that the first term dominates the convergence rate. Let

$$\tilde{\Omega} = \exp(\mathcal{T}(\log(\hat{\Omega}))), \quad \mathcal{T}(\hat{\Omega})_{ij} = \hat{\Omega}_{ij}\mathbb{I}\{|\hat{\Omega}_{ij}| > \tau_\omega\}. \tag{J.8}$$

Then, under Condition 1, for $\tau_\omega \asymp s_l(p)^2(n^{-1}\log p)^{1/2}$, Theorem 2 of Battey (2019) implies,

$$\|\Omega - \tilde{\Omega}\|_2 = O_p\left(s_\omega(p)s_l(p)^{2-2q_\omega}(n^{-1}\log p)^{(3/2-q_l)(1-q_\omega)}\right), \tag{J.9}$$

where $s_\omega(p)/p \to 0$ and $0 \leq q_\omega \leq 1$.

Given estimators $\tilde{T}$ and $\tilde{\Omega}$, the estimator of the covariance matrix Σ can be obtained by $\tilde{\Sigma} = \tilde{T}\tilde{\Omega}\tilde{T}^{\mathrm{T}}$. Proposition 4 in the paper establishes the spectral-norm consistency of $\tilde{\Sigma}$ as p, $n \to \infty$, provided that $\log p/n \to 0$.

In the absence of a causal ordering of variables, a natural pilot estimator of T is a triangular matrix obtained by the LDL decomposition of $\hat{\Sigma}$. Since $T = (I - B)^{-1}$, we can use the proof strategy used to establish the elementwise consistency of the matrix logarithm above. Specifically, under Condition J.5, an elementwise consistency of a truncated matrix inverse, defined for a matrix A as $A^{-1}_{|l} := \sum_{k=1}^{l}(-1)^{k+1}A^k$ is established by Lemma J.14 below.

Condition J.5. There exists $l^* \in \mathbb{N}$, such that for any pair of nodes (i, j), $j < i$, $\sum_{l=l^*+1}^{p} \delta_{i|j}(l) = C(\log p/n)^{\varphi_2/(2(\varphi_2+1))}$ for $\varphi_2 > 0$, where $\delta_{i|j}(l)$ denotes the sum of effects of node j on node i along all paths of length l.

Lemma J.14. *Let* $\hat{B}$ *be an estimator of* B *such that* $\|\hat{B} - B\|_{\max} = O_p((n^{-1}\log p)^{1/2})$*. Assume that Condition J.5 holds and let* $\bar{T} = (I - \hat{B})^{-1}_{|l}$*. Then,*

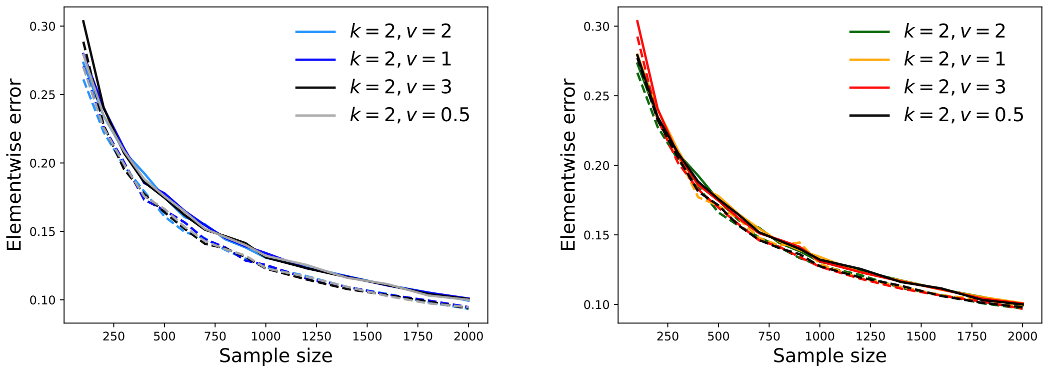
$$\|\bar{T} - T\|_{\max} = O_p\left((n^{-1}\log p)^{1/2}\right). \tag{J.10}$$

Simulations presented in Figure 3 suggest that Condition J.14 and the restriction to a truncated inverse are not necessary for Lemma J.14 to hold. This suggests that the rate $(\log p/n)^{1/2}$ is also valid for a pilot estimator $\hat{T} = (I - \hat{B})^{-1}$, which corresponds to the triangular matrix obtained by the LDL decomposition of the sample covariance matrix.

J.3. *Simulations*

Lemma J.11 shows that Condition J.4 is sufficient to establish elementwise convergence of $\hat{L}$, where $\hat{L} = -\log_{|l^*}(I - \hat{B})$. Using simulations, we now compare the rate of convergence of $\|\bar{L} - L\|_{\max}$, $\bar{L} = -\log(I - \hat{B})$, and $\|\hat{B} - B\|_{\max}$ in the absence of Condition J.4. The results, presented in Figure J.3 (a), suggest that Condition J.4 is not necessary for the equation (J.2) to hold. In addition, Lemma J.11 holds when $\hat{L}$ is replaced by $\bar{L}$. An analogous analysis is performed to assess the necessity of Condition J.5 for the validity of Lemma J.14 in Figure J.3 (b), which compares the rate of convergence of $\|\bar{T} - T\|_{\max}$, $\bar{T} = (I - \hat{B})^{-1}$, and $\|\hat{B} - B\|_{\max}$.

For each simulation, $\Sigma = O\Lambda O^{\mathrm{T}}$, where O is an orthogonal matrix obtained by a QR decomposition of a $p \times p$ matrix with iid standard normal entries, and elements of Λ are drawn from a gamma distribution with a shape parameter k and a scale parameter v.



(a) $\|\bar{L}-L\|_{\max}$

(b) $\|\bar{T}-T\|_{\max}$

Fig. J.3: Average elementwise errors $\|\hat{B}-B\|_{\max}$ (solid lines), $\|\bar{L}-L\|_{\max}$ (dotted lines, left plot) and $\|\bar{T}-T\|_{\max}$ (dotted lines, right plot) for 100 simulations for each n, with $p = n/10$.

J.4. *Proofs of results in Appendix J.2*

Proof of Lemma J.10. The estimation error for the ith element of $\hat{\beta}^j$ has the following form,

$$\Delta_i^j \equiv \hat{\beta}_i^j - \beta_i^j = e_i^T (X_{\mathrm{pa}(j)}^T X_{\mathrm{pa}(j)})^{-1} X_{\mathrm{pa}(j)}^T \varepsilon.$$

Since ε is sub-Gaussian, Δ_i^j is a linear combination of sub-Gaussian random variables. Thus, Δ_i^j is sub-Gaussian with a variance proxy $\|e_i^T (X_{\mathrm{pa}(j)}^T X_{\mathrm{pa}(j)})^{-1} X_{\mathrm{pa}(j)}^T\|_2^2 \sigma_\varepsilon$. Now,

$$\|e_i^T (X_{\mathrm{pa}(j)}^T X_{\mathrm{pa}(j)})^{-1} X_{\mathrm{pa}(j)}^T\|_2^2 = e_i^T (X_{\mathrm{pa}(j)}^T X_{\mathrm{pa}(j)})^{-1} e_i = \widehat{\mathrm{var}}(\mathrm{pa}(j))_{ii}^{-1}/(n-1).$$

Under Condition J.3 the maximum eigenvalue of $\widehat{\mathrm{var}}(\mathrm{pa}(j))^{-1}$ is upper-bounded. By the definition of the operator norm, for any column v of $\widehat{\mathrm{var}}(\mathrm{pa}(j))^{-1}$ we have $\|v\|_2 \leq M$. By the Cauchy-Schwartz inequality $|v_k| = |\langle e_k, v\rangle| \leq \|v\|_2 \leq M$, where v_k is the kth entry of v and e_k is a canonical basis vector with kth element equal to one. Thus, $\max_{i,j\in[p]} \widehat{\mathrm{var}}(\mathrm{pa}(j))_{ii}^{-1} \leq M$ for some constant M. Hence, $\mathbb{E}\exp(\Delta_i^j) \leq \exp(\widehat{\mathrm{var}}(\mathrm{pa}(j))_{ii}^{-1}\sigma_\varepsilon^2/2(n-1)) \leq \exp(M\sigma_\varepsilon^2/2n)$ for every $i, j \in [p]$. Then,

$$\mathbb{P}(|\Delta_i^j| \geq t) \leq 2\exp\left(-\frac{nt^2}{2M\sigma_\varepsilon^2}\right).$$

On setting $t = (2M\sigma_\varepsilon^2 \log(2p^2/\delta)/n)^{1/2}$ we obtain

$$\mathbb{P}\left(|\Delta_i| \geq \left(\frac{2M\sigma_\varepsilon^2 \log(2p^2/\delta)}{n}\right)^{1/2}\right) \leq 2\exp\left(-\frac{nt^2}{2M\sigma_\varepsilon^2}\right) = \frac{\delta}{p^2}.$$

The union bound gives

$$\mathbb{P}\left(\max_{j\in[p]} \|\hat{\beta}^j - \beta^j\|_\infty \geq t\right) \leq \sum_{j=1}^{p}\sum_{i=1}^{p} \mathbb{P}(|\Delta_i^j| \geq t) \leq \delta.$$

Hence, for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\max_{j\in[p]} \|\hat{\beta}^j - \beta^j\|_\infty \geq \left(\frac{2M\sigma_\varepsilon^2 \log(2p^2/\delta)}{n}\right)^{1/2}\right) \leq 1-\delta.$$

□

Proof of Lemma J.11. Let $r_\beta(n,p) \triangleq (n^{-1}\log p)^{\frac{\kappa}{2(\kappa+1)}}$. Recall that $\delta_{i|j}(l)$ denotes the total effect of node j on node i along all paths of length l. Specifically, let $S(j,i,l)$ be a set of subsets of indices $\{j+1,\dots,i-1\}$ of length $l-2$. For each $s \in S(j,i,l)$, denote the corresponding indices by $s_1 < s_2 < \dots < s_{l-2}$. Then, for each l, $\delta_{k|i}(l)$ has the form,

$$\delta_{k|i}(l) = \sum_{s\in S(j,i,l)} \gamma_{i|s_{l-1}}\gamma_{j|s_1}\prod_{v=1}^{l-2}\gamma_{s_{v+1}|s_v},$$

where $\gamma_{t,v}$ denotes a regression coefficient of X_v from regression of X_t on $X_1,\dots,X_v$.

Now consider,

$$\Delta(j,i,s) = \Big|\gamma_{i,s_l}\gamma_{j|s_1}\prod_{v=1}^{l-2}\gamma_{s_{r+1}|s_r} - \hat{\gamma}_{i,s_l}\hat{\gamma}_{j|s_1}\prod_{v=1}^{l-2}\hat{\gamma}_{s_{v+1}|s_v}\Big|.$$

This expression has the form $\left|\prod_{v=1}^{l} a_v - \prod_{t=1}^{l} c_l\right|$ with $|a_v - c_v| = O_p(r_\beta(p,n))$. Let $A_v = \prod_{i=1}^{v} a_v$ and $C_v = \prod_{i=1}^{v} c_v$. Then, by the triangle inequality,

$$\begin{aligned}|A_{v+1} - C_{v+1}| &= |a_{v+1}A_v - c_{v+1}C_v| \\ &\le |A_v||a_{v+1} - c_{v+1}| + |c_{v+1}||A_v - C_v|.\end{aligned}$$

Applying the inequality recursively we obtain $|A_{v+1} - C_{v+1}| = O_p((v+1)r_\beta(n,p))$. Thus, $\Delta(j,i,s) = O_p(lr_\beta(n,p))$, which represents an estimation error for a single path of length l connecting nodes i and j. By the binomial theorem there are 2^{i-j-2} directed paths between i and j. Then,

$$\|\log_{|l^*}(I-B) - \log_{|l^*}(I-\hat{B})\|_{\max} = O_p(2^{l^*-2}r_\beta(n,p)).$$

As a result,

$$\|\log(I-B) - \log_{|l^*}(I-\hat{B})\|_{\max} = O_p\left(r_\beta(n,p)\right) + \max_{i,j}\sum_{l=l^*+1}^{p}\frac{1}{l}|\Delta(j,i,s)|.$$

By Condition J.4,

$$\max_{i,j}\sum_{l=l^*+1}^{p}\frac{1}{l}|\Delta(j,i,s)| = O_p\left((n^{-1}\log p)^{\varphi}\right).$$

The result follows since $\varphi \ge 1/2$. □

Proof of Corollary 1. Except for the change in the object being thresholded, the proof is that of Theorem 1 in Bickel & Levina (2008b). □

Proof of Lemma J.12. Consider $B - \tilde{B} = \exp(-L) - \exp(-\mathcal{T}(\hat{L}))$. By Corollary 6.2.32 in Horn & Johnson (1994),

$$\|B - \tilde{B}\|_2 = \|\exp(-L) - \exp(-\mathcal{T}(\hat{L}))\|_2 \le \|L - \mathcal{T}(\hat{L})\|_2\exp(\|L\|_2)\exp(\|L - \mathcal{T}(\hat{L})\|_2).$$

By the definition of the spectral norm, $\|L\|_2 = \lambda_{\max}(L^{\mathrm{T}}L)$. Similarly, since $T = \exp(L)$,

$$\begin{aligned}\|T - \tilde{T}\|_2 &= \|\exp(L) - \exp(\mathcal{T}(\hat{L}))\|_2 \\ &\le \|L - \mathcal{T}(\hat{L})\|_2\exp(\|L\|_2)\exp(\|L - \mathcal{T}(\hat{L})\|_2),\end{aligned}$$

where the inequality follows from Corollary 6.2.32 in Horn & Johnson (1994). □

Proof of Lemma J.13. For a chain component c, let $\mathcal{E} = X_c - BX_{\mathrm{pa}(c)}$ and $\hat{\mathcal{E}} = X_c - \hat{B}X_{\mathrm{pa}(c)}$. Then, $\hat{\Omega} = \widehat{\mathrm{var}}(\mathcal{E})$ and

$$\mathrm{var}(\mathcal{E}) = \mathrm{var}(X^c) - \mathrm{var}(BX_{\mathrm{pa}(c)}).$$

Using the triangle inequality and the equality above,

$$\|\widehat{\mathrm{var}}(\mathcal{E}) - \mathrm{var}(\mathcal{E})\|_{\max} = \|(\widehat{\mathrm{var}}(X_c) - \widehat{\mathrm{var}}(\hat{B}X_{\mathrm{pa}(c)})) - (\mathrm{var}(X_c) - \mathrm{var}(BX_{\mathrm{pa}(c)}))\|_{\max} \tag{J.11}$$

$$= \|\widehat{\mathrm{var}}(X_c) - \mathrm{var}(X_c)\|_{\max} + \|\widehat{\mathrm{var}}((\hat{B} - B)X_{\mathrm{pa}(c)})\|_{\max} + \|\widehat{\mathrm{var}}(BX_{\mathrm{pa}(c)}) - \mathrm{var}(BX_{\mathrm{pa}(c)})\|_{\max}. \tag{J.12}$$

By Lemma A.3 in Bickel & Levina (2008a), $\|\widehat{\mathrm{var}}(X_c) - \mathrm{var}(X_c)\|_{\max} = O_p(\sqrt{\log \rho_c / n})$. Now consider the second term in (J.12) and let $X^i_{\mathrm{pa}(c)}$ denote the ith sample of $X_{\mathrm{pa}(c)}$.

$$\widehat{\mathrm{var}}((\hat{B} - B)X_{\mathrm{pa}(c)}) = \frac{1}{n-1}\sum_{i=1}^{n}(B - \hat{B})X^i_{\mathrm{pa}(c)}X^i_{\mathrm{pa}(c)}(B - \hat{B})^{\mathrm{T}} \tag{J.13}$$

$$= (B - \hat{B})\left(\frac{1}{n-1}\sum_{i=1}^{n}X^i_{\mathrm{pa}(c)}X^i_{\mathrm{pa}(c)}\right)(B - \hat{B})^{\mathrm{T}} \tag{J.14}$$

$$= (B - \hat{B})\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})(B - \hat{B})^{\mathrm{T}}. \tag{J.15}$$

Let $[A]_{ij}$ denote an element (i, j) of matrix A. Then, from equation (J.15),

$$\begin{aligned}
|[\widehat{\mathrm{var}}(\hat{B}X_{\mathrm{pa}(c)} - BX_{\mathrm{pa}(c)})]_{ij}| &= |[(B - \hat{B})\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})(B - \hat{B})^{\mathrm{T}}]_{ij}| \\
&= \left|\sum_{l=1}^{|pa|}\sum_{k=1}^{|pa|}(B - \hat{B})_{il}\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})_{lk}(B - \hat{B})_{jk}\right| \\
&\leq \|\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})\|_{\max}\left|\sum_{l=1}^{|pa|}(B - \hat{B})_{il}\right|\left|\sum_{k=1}^{|pa|}(B - \hat{B})_{jk}\right| \\
&\leq \|\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})\|_{\max}\left(\sum_{l=1}^{|pa|}(B - \hat{B})^2_{il}\right)^{1/2}\left(\sum_{k=1}^{|pa|}(B - \hat{B})^2_{jk}\right)^{1/2} \\
&= \|\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})\|_{\max}\|\hat{\Delta}_i\|_2\|\hat{\Delta}_j\|_2
\end{aligned}$$

where $\hat{\Delta}_i$ denotes an ith row of $B - \hat{B}$. The elementwise consistency of the covariance matrix, together with Condition J.2, the spectral-norm consistency of $\hat{B}$ and the triangle inequality imply,

$$\|\widehat{\mathrm{var}}((\hat{B} - B)X_{\mathrm{pa}(c)})\|_{\max} = O_p\left(r_\beta(n, p)^2 + r_\beta(n, p)^2\sqrt{\log \rho_{pa}/n}\right).$$

To upper-bound the third term in equation (J.12) note that

$$\|\widehat{\mathrm{var}}(BX_{\mathrm{pa}(c)}) - \mathrm{var}(BX_{\mathrm{pa}(c)})\|_{\max} = \max_{ij}\left|[\widehat{\mathrm{var}}(BX_{\mathrm{pa}(c)}) - \mathrm{var}(BX_{\mathrm{pa}(c)})]_{ij}\right|$$

where $BX_{pa(c)}$ is sub-Gaussian with zero mean. Thus, we can upper-bound this term using the elementwise consistency of the covariance matrix estimator, which yields

$$\|\widehat{\mathrm{var}}(BX_{\mathrm{pa}(c)}) - \mathrm{var}(BX_{\mathrm{pa}(c)})\|_{\max} = O_p(\sqrt{\log \rho_c / n}).$$

Overall, we obtain,

$$\|\widehat{\mathrm{var}}(\mathcal{E}) - \mathrm{var}(\mathcal{E})\|_{\max} = O_p\left(r_\beta(n, p)^2 + r_\beta(n, p)^2\sqrt{\log \rho_{pa}/n} + \sqrt{\log \rho_c / n}\right),$$

which establishes the first claim in Lemma J.13. The proof for the second claim is identical, except for the upper bound for the second term in (J.12), which we address now. The proof is similar to that of Bickel &

Levina (2008a). By Cauchy-Schwartz inequality and the fact that $\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})_{lk} \le \widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})_{ll}\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})_{kk}$,

$$\begin{aligned}
|[\widehat{\mathrm{var}}((\hat{B} - B)X_{\mathrm{pa}(c)})]_{ij}| &= |[(B - \hat{B})\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})(B - \hat{B})^{\mathrm{T}}]_{ij}| \\
&= |\sum_{l=1}^{|pa|}\sum_{k=1}^{|pa|}(B - \hat{B})_{il}\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})_{lk}(B - \hat{B})_{jk}| \\
&\le \left(\sum_{l=1}^{|pa|}|(B - \hat{B})_{il}|\right)^2 \widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})_{ll} \\
&\le \rho_{pa}^2\|\widehat{\mathrm{var}}(X_{\mathrm{pa}(c)})\|_{\max}\left(\max_{il}|B_{il} - \hat{B}_{il}|\right)^2
\end{aligned}$$

□

J.5. *Proof of Proposition 4*

Recall the inequality,

$$\begin{aligned}
&\|A_1A_2A_3 - C_1C_2C_3\|_2 \\
&\le \sum_{j=1}^{3}\|A_j - C_j\|_2\prod_{k\neq j}\|C_k\|_2 + \sum_{j=1}^{3}\|C_j\|_2\prod_{k\neq j}\|A_k - C_k\|_2 + \prod_{j=1}^{3}\|A_j - C_j\|_2.
\end{aligned}$$

Let $A_1 = A_3^{\mathrm{T}} = \tilde{T}$, $C_1 = C_3^{\mathrm{T}} = T$, $A_2 = \hat{\Omega}$ and $C_2 = \Omega^{-1}$. Let r_t and r_ω denote the convergence rates of $\tilde{T}$ and $\hat{\Omega}$ respectively. In particular, $\|\tilde{T} - T\|_2 = O_p(r_t)$ and $\|\hat{\Omega} - \Omega\|_2 = O_p(r_\omega)$. Then,

$$\|\hat{\Sigma} - \Sigma\|_2 \le 2r_t\|T\|_2\|\Omega\|_2 + r_\omega\|T\|_2^2 + 2r_tr_\omega\|T\|_2 + r_t^2\|\Omega_c\|_2^2 + r_t^2r_\omega.$$

The result follows from equations (J.5) and (J.9). □

K. Simulation results for §9.1

A notion of approximate sparsity that allows for slight departures from zero is, for any matrix A,

$$s_\tau(A) = \sum_{i,j<i}\mathbb{I}(|A_{ij}| > \tau). \qquad \text{(K.1)}$$

This replaces elements by 1 and 0 according to their values relative to τ, and thus is more suitable than (14) for comparison across scales.

For each of the four parametrizations of (1), we explore the extent to which L is sparser than Σ^{-1} according to equation (K.1), and the implications for estimation. For tables K.1–K.4, random matrices L of dimension $p = 60$ were generated using the appropriate basis in equation (2) by randomly drawing $s^*/2$ entries of α from a uniform distribution on $[-4, -2]$, $s^*/2$ entries from a uniform distribution on $[2, 4]$ and $m - s^*$ entries from a uniform distribution on $[-0.01, 0.01]$, where m is the number of elements in the basis. The resulting matrix L was converted to the relevant matrix space $\mathsf{PD}(p)$, $\mathsf{SO}(p)$, $\mathsf{LT}(p)$ or $\mathsf{LT}_{\mathsf{u}}(p)$ by taking the matrix exponential. The positive diagonal entries needed to complete the specification for the Σ_o and Σ_{ltu} parametrizations were drawn from an exponential distribution of rate ρ. In producing the simulations of this section, we have used R functions to implement the LDL, Cholesky, and LU decompositions, mainly to avoid the complications arising from pivoting operations used in the corresponding implementations in Matlab.

The estimation error in non-trivial matrix norms is most relevant when the matrix object is a nuisance parameter, and the numerical results presented here are motivated by that setting. Since it is usually the precision matrix that is the nuisance parameter in procedures of multivariate analysis, rather than the covariance matrix, we focus on estimation of Σ^{-1}.

			Estimator E	
s^*	$\frac{\|E-\Sigma^{-1}\|_\bullet}{\|\Sigma^{-1}\|_\bullet}$	$s_\tau(\Sigma^{-1})$	$\hat{\Sigma}_\tau^{-1}$	$\exp(-\hat{L}_\tau)$
6	$\bullet = 2$	45.6	0.561	0.203
6	$\bullet = \mathrm{F}$	45.6	0.503	0.187
10	$\bullet = 2$	54.7	0.551	0.214
10	$\bullet = \mathrm{F}$	54.7	0.504	0.197
20	$\bullet = 2$	94.3	0.540	0.215
20	$\bullet = \mathrm{F}$	94.3	0.505	0.202
40	$\bullet = 2$	436	0.496	0.231
40	$\bullet = \mathrm{F}$	436	0.477	0.221
Largest std. err.		130	0.189	0.076

Table K.1: Simulation averages of $s_\tau(\Sigma^{-1})$ and the relative estimation errors for estimators exploiting an assumption of sparsity on the inverse and logarithmic scales under the Σ_{pd} parametrization.

For each of 200 simulation replicates, $n = 200$ p-dimensional random vectors were generated from a mean-zero normal distribution with covariance matrix as specified above. Three estimators of the precision matrix were compared in terms of their average estimation errors in the spectral and Frobenius norms.

The simplest type of estimator exploiting sparsity sets entries of a preliminary estimate to zero if they are below a threshold τ. For the four parametrizations of equation (1), the simplest preliminary estimate is the matrix logarithm of the relevant sample quantity, constructed from the eigen-, Cholesky, or LDL decomposition of the sample covariance matrix. The matrix logarithm was computed using the algorithm of Al-Mohy & Higham (2012), whose implementation is part of Matlab's standard distribution and R's `expm` package. Let $\hat{L}_\tau$ denote the thresholded estimator on the logarithmic scale, so that an estimator of Σ^{-1} under the Σ_{pd} parametrization is $\exp(-\hat{L}_\tau)$ and the analogous quantities for the other three parametrizations are $\hat{O}_\tau = \exp(\hat{L}_\tau) \in \mathsf{SO}(p)$, $\hat{V}_\tau = \exp(\hat{L}_\tau) \in \mathsf{LT}(p)$ and $\hat{U}_\tau = \exp(\hat{L}_\tau) \in \mathsf{LT}_{\mathsf{u}}(p)$, from which an estimator of Σ^{-1} is constructed in the obvious way. A comparable estimator based on an assumption of sparsity directly on the inverse scale is $\hat{\Sigma}_\tau^{-1}$, the inverse sample covariance matrix thresholded at τ. In all cases, the threshold $\tau = 1$ was used as the level below which entries were set to zero, implying that $s_\tau(L)$ from equation (K.1) is s^* by the simulation design. The estimator $\hat{\Sigma}_\tau^{-1}$ typically violates positive definiteness, which may or may not be problematic, depending on context. The results for the three parametrizations are reported in Tables K.1–K.4.

For the Σ_o parametrization, an additional step checked whether the matrix of orthonormal eigenvectors $\hat{O}$ of the sample covariance matrix was special orthogonal, and if not, converted it to special orthogonal by multiplying the first row of $\hat{O}$ by minus one. This step ensures that the matrix logarithm is skew-symmetric and real-valued.

Thresholding on the logarithmic scale was justified by Battey (2019) under the Σ_{pd} parametrization, and in Proposition 4 under the Σ_{ltu} parametrization. We have not in these simulations attempted to optimize tuning constants, and it is likely that the results could be improved through a data-adaptive tuning, nevertheless, several of the results suggest a benefit from exploiting sparsity on the logarithmic scale as opposed to on the inverse scale.

The performance of $\hat{O}_\tau \hat{\Lambda}^{-1} \hat{O}_\tau^{\mathrm{T}}$, as reported in Table K.2, is relatively poor, suggesting that the thresholding approach is too simplistic for this case. One issue concerns the constraints on α needed to make the Σ_o parametrization injective (see Proposition C.4), which are not naturally accommodated by the thresholding estimator. Another aspect is the distortion of the distribution of matrix entries by the matrix logarithm, and its possible effect on the estimation error, which has not been formally studied for the Σ_o parametrization. Rybak & Battey (2021) noted a different estimator that does not involve taking matrix

				Estimator E	
s^*	ρ	$\frac{\|E-\Sigma^{-1}\|_\bullet}{\|\Sigma^{-1}\|_\bullet}$	$s_\tau(\Sigma^{-1})$	$\hat{\Sigma}_\tau^{-1}$	$\hat{O}_\tau\hat{\Lambda}^{-1}\hat{O}_\tau^{\mathrm{T}}$
6	2	$\bullet = 2$	91.3	0.561	1.264
6	2	$\bullet = \mathrm{F}$	91.3	0.540	1.451
6	4	$\bullet = 2$	132	0.565	1.281
6	4	$\bullet = \mathrm{F}$	132	0.551	1.469
10	2	$\bullet = 2$	91.7	0.599	1.305
10	2	$\bullet = \mathrm{F}$	91.7	0.561	1.467
10	4	$\bullet = 2$	131	0.604	1.308
10	4	$\bullet = \mathrm{F}$	131	0.571	1.480
20	2	$\bullet = 2$	110	0.597	1.340
20	2	$\bullet = \mathrm{F}$	110	0.564	1.521
20	4	$\bullet = 2$	155	0.602	1.355
20	4	$\bullet = \mathrm{F}$	155	0.576	1.539
Largest standard error			96.2	0.205	0.328

Table K.2: Simulation averages of $s_\tau(\Sigma^{-1})$ and the relative estimation errors for estimators exploiting an assumption of sparsity on the inverse and logarithmic scales under the Σ_o parametrization.

			Estimator E	
s^*	$\frac{\|E-\Sigma^{-1}\|_\bullet}{\|\Sigma^{-1}\|_\bullet}$	$s_\tau(\Sigma^{-1})$	$\hat{\Sigma}_\tau^{-1}$	$(\hat{V}_\tau\hat{V}_\tau^{\mathrm{T}})^{-1}$
6	$\bullet = 2$	42.6	0.590	0.148
6	$\bullet = \mathrm{F}$	42.6	0.508	0.129
10	$\bullet = 2$	50.4	0.581	0.147
10	$\bullet = \mathrm{F}$	50.4	0.514	0.131
20	$\bullet = 2$	74.5	0.551	0.164
20	$\bullet = \mathrm{F}$	74.5	0.516	0.153
40	$\bullet = 2$	162	0.503	0.203
40	$\bullet = \mathrm{F}$	162	0.481	0.191
Largest std. err.		47.7	0.198	0.112

Table K.3: Simulation averages of $s_\tau(\Sigma^{-1})$ and the relative estimation errors for estimators exploiting an assumption of sparsity on the inverse and logarithmic scales under the Σ_{lt} parametrization.

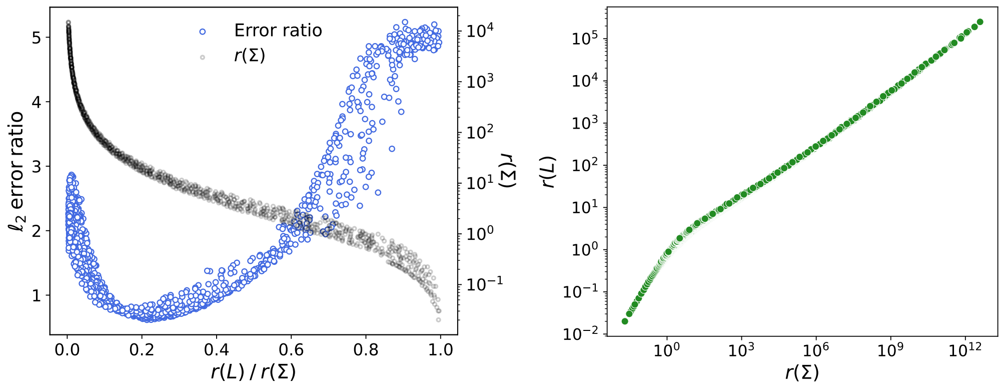
logarithms of sample quantities and that accommodates constraints on α. The formal implementation and theoretical justification of that approach requires major work not taken up here.

K.1. *Additional Figures for §9.2*

The simulation setting is that described in §9.2 of the main text. Figure K.4 explores the relationship between the relative performance of the two sparse estimators and their relative row norms, as quantified by equation (15) of the main text. Specifically, Figure K.4 (B) shows that the metrics $r(\Sigma)$ and $r(L)$ are closely related. Thus, when $r(L)$ is low, so is $r(\Sigma)$. Thresholding $\hat{\Sigma}$ yields a significantly lower ℓ_2 error when the ratio $r(L)/r(\Sigma)$, and $r(\Sigma)$, are either large, or very small, while $\hat{U}_\tau\hat{D}\hat{U}_\tau^{\mathrm{T}}$ seems advantageous for medium values of $r(L)/r(\Sigma)$, as depicted in Figure K.4 (A).

				Estimator E	
s^*	ρ	$\frac{\Vert E-\Sigma^{-1}\Vert_\bullet}{\Vert\Sigma^{-1}\Vert_\bullet}$	$s_\tau(\Sigma^{-1})$	$\hat{\Sigma}_\tau^{-1}$	$(\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}})^{-1}$
6	2	$\bullet = 2$	90.7	0.526	0.405
6	2	$\bullet = \mathrm{F}$	90.7	0.503	0.372
6	4	$\bullet = 2$	130	0.527	0.405
6	4	$\bullet = \mathrm{F}$	130	0.506	0.373
10	2	$\bullet = 2$	110	0.547	0.431
10	2	$\bullet = \mathrm{F}$	110	0.520	0.396
10	4	$\bullet = 2$	157	0.548	0.425
10	4	$\bullet = \mathrm{F}$	157	0.522	0.392
20	2	$\bullet = 2$	163	0.548	0.482
20	2	$\bullet = \mathrm{F}$	163	0.514	0.448
20	4	$\bullet = 2$	235	0.548	0.486
20	4	$\bullet = \mathrm{F}$	235	0.514	0.453
Largest standard error			83.2	0.189	0.310

Table K.4: Simulation averages of $s_\tau(\Sigma^{-1})$ and the relative estimation errors for estimators exploiting an assumption of sparsity on the inverse and logarithmic scales under the Σ_{ltu} parametrization.



(a)

(b)

Fig. K.4: (a) ℓ_2 error ratio $\Vert\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}} - \Sigma\Vert_2 / \Vert\hat{\Sigma}_\tau - \Sigma\Vert_2$ (left-axis, blue) and $r(\Sigma)$ (right axis, black) plotted against the ratio $r(L)/r(\Sigma)$. (b) Maximum row-sum of L versus maximum row sum of the lower-triangular part of Σ. Each point in both plots corresponds to a median over 100 simulations, with $n = 150$, $p = 100$, for each combination of ϵ and percentage of non-zero entries of B.

L. Application to leukemia and arrhythmia data

The data (Efron & Hastie, 2016, §19.1) consist of 3571 features for 72 patients. Of these, 47 have acute lymphoblastic leukaemia and 25 have acute myeloid leukemia. We used linear discriminant analysis with the sample covariance matrix replaced by a thresholded estimator on each of the scales considered in the paper, in order to assess the ultimate classification performance. Since the estimator $\exp(\hat{L}_\tau)$ requires the sample covariance matrix to be positive definite, which fails to hold if $n < p$, we replace $\hat{\Sigma}$ by $\hat{\Sigma} + \delta_{p,n}\mathrm{diag}(\hat{\Sigma})$, where $\delta_{p,n} = (\log(p)/n)^{1/2}$; this choice was justified by Battey (2019). For estimators $\hat{\Sigma}_\tau$, $\hat{O}_\tau \hat{\Lambda}^{-1} \hat{O}_\tau^{\mathrm{T}}$ and $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$ we considered both $\hat{\Sigma}$ and $\hat{\Sigma} + \delta_{p,n}\mathrm{diag}(\hat{\Sigma})$ as pilot estimators, and report the higher of the two accuracy rates in Table L.5.

The accuracy rates were obtained by randomly splitting the sample into two sets, consisting of 50 and 22 patients respectively, the smaller subset serving as a hold-out for testing classification performance on the basis of the larger training set. To select a threshold, the larger subset is itself split into a training (80%) and a validation set (20%) ten times. For each method, we select a threshold that minimizes validation error over the ten splits. The final classifier is estimated using all 50 patients and its out-of-sample performance is calculated using the hold-out sample. The procedure is repeated 50 times, which results in a set of 50 out-of-sample accuracy rates for each method. Results are reported in Table L.5.

Test error	$\hat{\Sigma}_\tau$	$\hat{O}_\tau \hat{\Lambda}^{-1} \hat{O}_\tau^{\mathrm{T}}$	$\exp(\hat{L}_\tau)$	$\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$
Median	95.5%	95.5%	95.5%	97.7%
s. e.	6.3%	4.0%	4.5%	5.1%

Table L.5: Median and standard error of accuracy scores on a hold-out dataset. Calculated over 50 randomly chosen test sets.

The Arrhythmia dataset from the UCI Machine Learning Repository (Guvenir et al. , 1997) has 452 observations, each representing a different patient. There are 16 classes, one representing normal ECG, the remaining ones corresponding to different types of arrhythmia. We convert this to a binary classification problem by pooling all arrhythmia classes together. The resulting dataset consists of 245 healthy patients, and 207 patients with arrhythmia. We omit categorical features with fewer than 10 categories, resulting in 164 explanatory variables.

The accuracy was calculated using the same approach as for leukemia data, based on 20 different splits of the data into a training and a hold-out set. The thresholding hyperparameter was chosen using a five-fold cross-validation.

References

Al-Mohy, A. H. and Higham, D. J. (2012). Improved inverse scaling and squaring algorithms for the matrix logarithm. *SIAM J. Sci. Comput*, 34, C153–C169.

Andersson, S. A., Madigan, D. and Perlman, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.*, 28, 33–85.

Arsigny, V., Fillard, P., Pennac, X. and Ayache, N. (2017). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Analy. Applic.*, 29, 328–347.

Baker, A(2001). *Matrix Groups: an introduction to Lie group theory*, Springer.

Battey, H. S. (2019). On sparsity scales and covariance matrix transformations. *Biometrika*, 106, 605–617.

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36, 199–227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.*, 36, 2577–2604.

Culver, W. J.(1966). On the existence and uniqueness of the real logarithm of a matrix. *Proc. Am. Math. Soc.*, 17, 1146-1151.

Drton, M. and Eichler, M. (2006) Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist.*, 33, 247–257.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*, Cambridge University Press.

Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist.*, 17, 333–353.

Golub, G. H. and Van Loan, C. F.(2013). *Matrix Computations*. Fourth edition. The Johns Hopkins University Press.

Grossier, D., Jung, S. and Schwartzman, A.(2021). Uniqueness questions in a scaling-rotation geometry on the space of symmetric positive-definite matrices. *Differ. Geom. Appl.*, 79, 101798.

Guvenir, H., Acar, B., Muderrisoglu, H., and Quinlan, R. Arrhythmia *UCI Machine Learning Repository.*

Helgason, S.(2001). *Differential geometry, Lie groups, and symmetric spaces*. American Mathematical Society.

Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Mnalysis*. Cambridge University Press.

Kato, T. (1976). *Perturbation Theory for Linear Operators.* Second edition. Springer-Verlag: Berlin.

Lauritzen, S. L. and Wermtuh, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17, 31–57.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

MAATHUIS, M., DRTON, M., LAURITZEN, S. L. AND WAINWRIGHT, M. (2019). *Handbook of Graphical Models*, CRC Press.

RYBAK, J. AND BATTEY, H. S. (2021). Sparsity induced by covariance transformation: some deterministic and probabilistic results. *Proc. Roy. Soc. London, A.*, 477, 20200756.

TERRAS, A. (1988). *Harmonic Analysis on Symmetric Spaces and Applications II*. Springer.

UHLER, C.(2009). Gaussian graphical models: An algebraic and geometric perspective. *Chapter in Handbook of Graphical Models*, CRC Press.

WERMUTH, N. AND COX, D. R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Statist. Soc. B*, 66, 687–717.

WHITTAKER, E. T. AND WATSON, G. N. (1965). *A Course of Modern Analysis*. Cambridge University Press.

ZWIERNIK, P. (2025). Entropic covariance models. *Ann. Statist.*, to appear.

[*Received on* 1 *July* 2024. *Editorial decision on* 16 *June* 1904]