Biometrika (2026), **113**, 1, asaf071 https://doi.org/10.1093/biomet/asaf071
Advance Access publication 24 October 2025

Regression graphs and sparsity-inducing reparameterizations

By J. RYBAK, H. S. BATTEY

Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, U.K.
jakub.rybak18@imperial.ac.uk 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 parameterization and sparsity are inherently linked raises the possibility that relevant models, not obviously sparse in their natural formulation, exhibit a population-level sparsity after reparameterization. In covariance models, positive definiteness enforces additional constraints on how sparsity can legitimately manifest. It is therefore natural to consider reparameterization maps in which sparsity respects positive definiteness. This paper provides insight into structures on the physically natural scale that induce and are induced by sparsity after reparameterization. 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-model interpretation, points to a class of reparameterizations 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 reparameterization is beneficial.

Some key words: Causality; Chain graph; Graphical model; Matrix logarithm; Reparameterization; 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 parameterizations enjoying a population-level sparsity not present to the same extent in the original formulation. In other words, from a parameterization that is natural from a modelling point of view, we seek a sparsity-inducing reparameterization.

Inducement of population-level sparsity, through a traversal of parameterization 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 & Reid, 1987). The development of Gaussian graphical models (e.g., Cox & Wermuth, 1996; Lauritzen, 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 nontrivial sparsity-inducing reparameterization can be deduced in which sparsity respects positive definiteness. By nontrivial, 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 reparameterizations 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 parameterization than that the sparse parameterization 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 parameterizations 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 reparameterization. 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 parameterizations via a class of matrix decompositions representing the chain graphs, allowing for both directed and undirected edges, and recovering the four fundamental parameterizations 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.

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

Notation	Matrix subset	Basis notation
$\mathsf{PD}(p)$	Symmetric positive-definite matrices	
$\mathsf{Sym}(p)$	Symmetric matrices	$\mathcal{B}_{\text{sym}}$
$\mathsf{D}(p)$	Diagonal matrices	$\mathcal{B}_{\text{diag}}$
$\mathsf{GL}(p)$	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	$\mathcal{B}_{\text{sk}}$
$\mathsf{LT}(p)$	Lower-triangular matrices	$\mathcal{B}_{\text{lt}}$
$\mathsf{LT}_{\text{u}}(p)$	Lower triangular with unit diagonal entries	
$\mathsf{LT}_{\text{s}}(p)$	Strictly lower-triangular matrices	$\mathcal{B}_{\text{lts}}$

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 $\mathsf{V}(p)$.

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 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 $\{1, \ldots, p\}$ 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 $e_1, \ldots, e_p$ for $\mathbb{R}^p$, where $e_i \in \mathbb{R}^p$ is a zero vector with 1 as its ith component. Specifically, $\mathcal{B}_{\text{sym}} := \{B_1, \ldots, B_{p(p+1)/2}\}$ consists of $p(p-1)/2$ nondiagonal 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 $\mathcal{B}_{\text{diag}}$; $\mathcal{B}_{\text{sk}} := \{B_1, \ldots, B_{p(p-1)/2}\}$ consists of skew symmetric matrices $e_j e_k^{\mathrm{T}} - e_k e_j^{\mathrm{T}}$ for $j < k$; $\mathcal{B}_{\text{lt}} := \{B_1, \ldots, B_{p(p+1)/2}\}$ consists of lower-triangular matrices $e_k e_j^{\mathrm{T}}$, $j \leqslant k$; and $\mathcal{B}_{\text{lts}} := \{B_1, \ldots, B_{p(p-1)/2}\}$ 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 $\mathrm{e}^A = \sum_{k=0}^{\infty} A^k/k!$. Conversely, if a matrix logarithm L of a square matrix M exists then $M = \mathrm{e}^L$. See the Supplementary Material 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 $X_1 \perp\!\!\!\perp X_2 \mid X_3$, unconditional independence is notated by $X_1 \perp\!\!\!\perp X_2$.

3. Reparameterization

The set Cone_p is, from one perspective, the natural parameter domain for parameterizing 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 reparameterization 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 nonstandard parameterizations. We clarify in subsequent sections the extent to which, and manner in which, a covariance or precision matrix might be sparser after reparameterization, 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 which enjoy a graphical-model 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 the Supplementary Material, 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 parameterizations discussed in this work are, with $D(d) = \mathrm{diag}(d_1, \ldots, d_p)$,

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

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 reparameterization 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 parameterizations 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

$$L(\alpha) = \alpha_1 B_1 + \cdots + \alpha_m B_m$$

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

Another parameterization 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} parameterizations, 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 below. One might call these zeros *structural zeros*: ones that are preserved through transformation regardless of the values of the nonzero 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} parametrization. For $s^* < p$, Battey (2017) showed that $\Sigma = \Sigma_{\mathrm{pd}}(\alpha)$ is necessarily of the form

$$\Sigma = \sum_{j \in \mathcal{A}} \lambda_j o_j o_j^{\mathrm{T}} + \sum_{j \in \mathcal{A}^c} e_j e_j^{\mathrm{T}}, \qquad \|o_j\|_0 = p - |\mathcal{A}^c|, \tag{2}$$

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 (2), since the second sum only specifies unit entries in diagonal positions corresponding to $\mathcal{A}$, is that if

$$L_{ik} = \sum_{j \in \mathcal{A}} \log \lambda_j o_{ij} o_{kj} = 0 \tag{3}$$

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 nonzero. This coincidental zero in the logarithm comes from cancellation, as distinct from the structural zeros in the eigenvectors. Equation (3) 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.

The following theorem is a general result, whose application to the four cases results in Corollaries 1–4.

Theorem 1. *Let $M = \mathrm{e}^L \in \mathsf{M}(p)$, 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^* nonzero 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,*

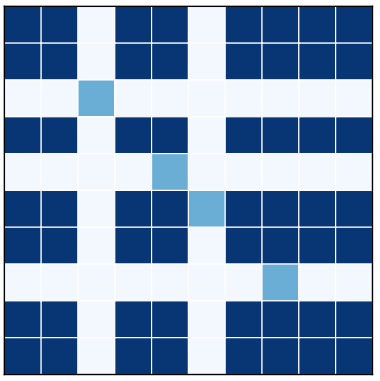

Fig. 1. Example of a structure of M as established in Theorem 1 with $p = 10$, $d_r^* = 7$, $d_c^* = 8$ and $d^* = 9$. The entries that are zero by Theorem 1 are light blue, those equal to one are medium blue and the remaining entries, whose values are unconstrained, are dark blue.

$p - d_c^*$ 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 $\max\{d_r^*, d_c^*\} \leqslant 2s^*$, 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 nonzero 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 Fig. 1.

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

COROLLARY 1. *Let Σ be parameterized as $\Sigma_{\mathrm{pd}}(\alpha) = \mathrm{e}^{L(\alpha)}$, and let $d < p$. Then, $\Sigma_{\mathrm{pd}}(\alpha)$ 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 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 (2).

COROLLARY 2. *For an arbitrary diagonal component $D = \mathrm{diag}\{d_1, \ldots, d_p\}$, let Σ be parameterized as $\Sigma_o(\alpha) = \mathrm{e}^{L(\alpha)}\,\mathrm{e}^{D}(\mathrm{e}^{L(\alpha)})^{\mathrm{T}}$. 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 parameterized as $\Sigma_{\mathrm{lt}}(\alpha) = \mathrm{e}^{L(\alpha)}(\mathrm{e}^{L(\alpha)})^{\mathrm{T}}$. Then, $\Sigma_{\mathrm{lt}}(\alpha)$ 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* $D = \mathrm{diag}\{d_1, \ldots, d_p\}$, *let* Σ *be parameterized as* $\Sigma_{\mathrm{ltu}}(\alpha) = \mathrm{e}^{L(\alpha)}\,\mathrm{e}^{D}(\mathrm{e}^{L(\alpha)})^{\mathrm{T}}$. *Then,* $\Sigma_{\mathrm{ltu}}(\alpha)$ *is of the form* $\Sigma = U\Psi U^{\mathrm{T}}$, *where* $\Psi = \mathrm{e}^{D} \in \mathsf{D}_{+}(p)$, $U = I_p + \Theta$. *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 parameterizations Σ_{pd} and Σ_o, and structured patterns of zeros in the Cholesky factors in parameterizations 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} parameterization; 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 parameterizations 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^r_{ltu}, for which $d^*_r < p$, $d^*_c = p$ (this serving as the definition of Σ^r_{ltu}), and Σ^c_{ltu}, for which $d^*_r = p$, $d^*_c < p$; in both cases $s^* = 6$. The quantities d^*, d^*_r and d^*_c are defined in Theorem 1. The resulting covariance structure is depicted in Fig. 2. If, for an arbitrary diagonal component, the map $\alpha \mapsto \Sigma_{\mathrm{ltu}}(\alpha)$ is instead replaced by an essentially equivalent representation $\alpha \mapsto \Sigma_{\mathrm{utu}}(\alpha)$ in terms of upper-triangular matrices, the analogous structures $\Sigma^r_{\mathrm{utu}}(\alpha)$ and $\Sigma^c_{\mathrm{utu}}(\alpha)$ are the same as for $\Sigma^c_{\mathrm{ltu}}(\alpha)$ and $\Sigma^r_{\mathrm{ltu}}(\alpha)$, respectively.

For the Σ_{ltu} parameterization, Fig. 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 $\alpha \mapsto \Sigma_{\mathrm{lt}}(\alpha)$ and $(\alpha, d) \mapsto \Sigma_{\mathrm{ltu}}(\alpha, d)$ parameterizations are related. To see this, write

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

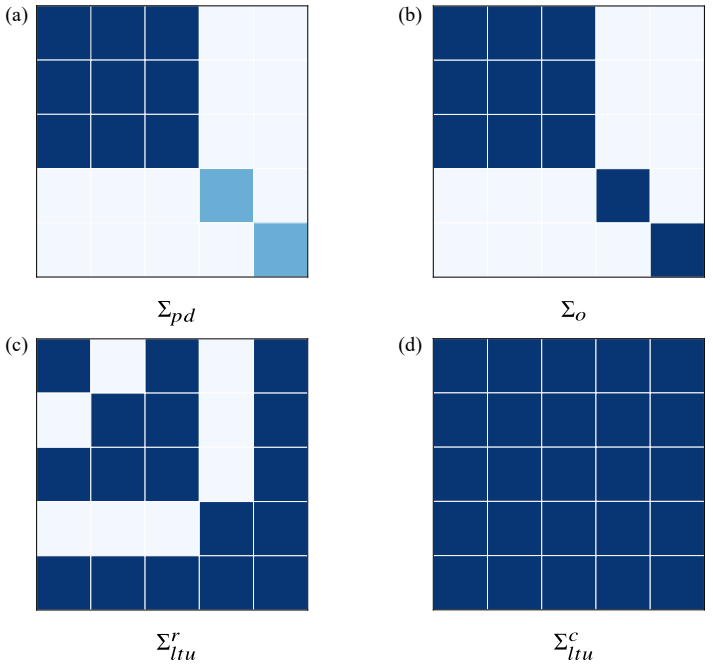

Fig. 2. Structure of $\Sigma(\alpha)$ induced by sparsity of α. Zero entries are depicted by light blue, unit entries by medium blue and the unrestricted entries by dark blue.

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

$$\Sigma_{\mathrm{lt}} = \exp(L)\exp(L^{\mathrm{T}}) = \exp(L_s)\exp(D_1 + D_1^{\mathrm{T}})\exp(L_s^{\mathrm{T}}), \tag{4}$$

by the properties of the matrix exponential, which recovers the Σ_{ltu} parameterization 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 parameterizing regression graphs, subsequent discussion focuses on Σ_{ltu}. Most insights derived for the Σ_{ltu} parameterization extend directly to Σ_{lt}.

5. SPARSITY UNDER THE Σ_{ltu} PARAMETERIZATION

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),

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} = \begin{pmatrix} I_{|a|} & 0 \\ \Sigma_{ba}\Sigma_{aa}^{-1} & I_{|b|} \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & 0 \\ 0 & \Sigma_{bb.a} \end{pmatrix} \begin{pmatrix} I_{|a|} & \Sigma_{aa}^{-1}\Sigma_{ab} \\ 0 & I_{|b|} \end{pmatrix}. \tag{5}$$

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} parameterization. Specifically, $\Sigma = U\Psi U^{\mathrm{T}}$ with $\Psi \in \mathsf{D}_+(p)$ and

$$U = I_p + \Theta = (I_p - B)^{-1}, \tag{6}$$

where $\Theta \in \mathsf{LT_s}(p)$ 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 $Y_j \to Y_i$ can exist only if $j < i$, in 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,

$$\beta_{3.1} = \beta_{3.12} + \beta_{3.21}\beta_{2.1}. \tag{7}$$

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 (7) 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 Fig. 3.

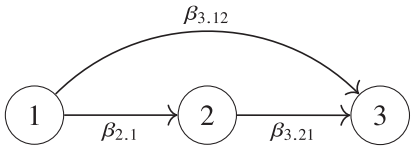

Fig. 3. Directed acyclic graph with edge weights corresponding to regression coefficients.

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: $Y_1 \to Y_3$ and $Y_1 \to Y_2 \to Y_3$, which correspond to the first and second terms in (7), respectively. Let $v_{ij}(l)$ denote the effect of Y_j on Y_i along all paths of length l, specified for this three-dimensional system as

$$v_{21}(1) = \beta_{2.1}, \qquad v_{31}(1) = \beta_{3.12}, \qquad v_{32}(1) = \beta_{3.21}, \qquad v_{31}(2) = \beta_{3.21}\beta_{2.1}.$$

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

$$U = \begin{pmatrix} 1 & 0 & 0 \\ \beta_{2.1} & 1 & 0 \\ \beta_{3.12} + \beta_{3.21}\beta_{2.1} & \beta_{3.21} & 1 \end{pmatrix}, \qquad L = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{2.1} & 0 & 0 \\ \beta_{3.12} + \beta_{3.21}\beta_{2.1}/2 & \beta_{3.21} & 0 \end{pmatrix}.$$

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 parameterization* $\Sigma_{\mathrm{ltu}}(\alpha, d) = \mathrm{e}^{L(\alpha)}\,\mathrm{e}^{D(d)}(\mathrm{e}^{L(\alpha)})^{\mathrm{T}}$, *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*

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

where

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

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,

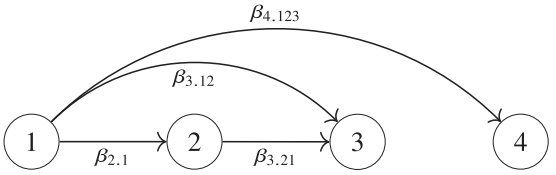

Fig. 4. Directed acyclic graph corresponding to U and L satisfying $U_{42} = 0$ and $L_{42} = 0$.

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 the next subsection.

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 the 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 Fig. 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 $U = I + \Theta$, $\Theta \in \mathsf{LT_s}(p)$ and $\Psi \in \mathsf{D}_+$. Then the following statements hold.*

(i). *If the jth column of Θ is zero then $Y_j \perp\!\!\!\perp Y_{j+1}, \ldots, Y_p \mid Y_1, \ldots, Y_{j-1}$. Consequently, $\Sigma^{-1}_{ji} = 0$ for $i \in \{j+1, \ldots, p\}$.*
(ii). *If the jth 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

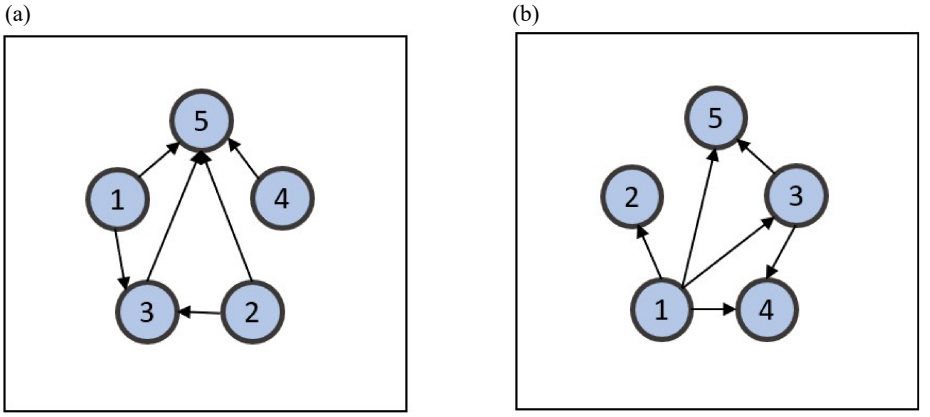

Fig. 5. Directed acyclic graphs corresponding to the example of Fig. 2: (a) $\Sigma^{r}_{\text{ltu}}(\alpha)$, (b) $\Sigma^{c}_{\text{ltu}}(\alpha)$. Arrows indicate directed edges and nodes correspond to random variables.

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 $\not\circ$ and $\boxed{\circ}$ representing marginalization and conditioning, respectively,

$$i \longleftarrow \not\circ \longrightarrow j, \qquad i \longleftarrow \not\circ \longleftarrow j, \qquad i \longrightarrow \boxed{\circ} \longleftarrow j,$$
$$i \text{ ---- } j, \qquad i \longleftarrow j, \qquad i \text{ ---- } j,$$

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 (c) and (d) of Fig. 2, respectively. The indices of nonzero 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 nonzero entries of both B and Θ, as a convergent Taylor representation of the matrix inverse in (6) 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 Fig. 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 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, with 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 $1 \to 2 \to 3$ and no direct edge between 1 and 3. Suppose that edge weight $1 \to 2$ is very large, while that of $2 \to 3$ 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 $1 \to 2$ 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 $1 \to 3$ 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.

The above conclusion bears some resemblance to a suggestion from influential work in computer science about learning in networks. Grover & Leskovec (2016) defined 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 argued that information from both types should be combined, with a hyperparameter specifying the relative importance of paths. The Σ_{ltu} parameterization 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} parameterization

6.1. *Sparsity-induced structures*

The ordering is immaterial for the Σ_{pd} and Σ_o parameterizations. 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} parameterization, while § 7 explains the relevance of the Σ_{pd} parameterization in the context of chain graphs, broadening its scope.

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

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 $Y_1, \ldots, Y_p$, 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 the normalized graph Laplacian in spectral graph theory (e.g., Von Luxburg, 2007). Let $v_{ij}^u(l)$ denote the total effect of a unit change in Y_j on Y_i along all paths of length l, with the superscript u in $v_{ij}^u(l)$ distinguishing this quantity for an undirected graph. As compared with $v_{ij}(l)$ 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 $\Sigma \in \mathsf{PD}(p)$, let $\tilde{V} = \mathrm{diag}(V)^{-1}V$ and $\tilde{\Sigma} = \Sigma\,\mathrm{diag}(V)$. Then the elements (i,j) of $\tilde{\Sigma}$ and of its matrix logarithm have the form*

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

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 parameterization for chain graphs*

The interpretation of (5) 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 semidirected 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 semidirected 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

$$\Sigma^{-1} = (I - B^{\mathrm{T}})\Omega^{-1}(I - B), \tag{8}$$

where Ω_{uv}^{-1} is zero if an undirected edge (u, v) is not in the graph, and $B_{vu} = 0$ if a directed edge $u \to v$ 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 that $B_{vu} = 0$, 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 (8) 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. Factorization (8) 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 subvector Y_i. The decomposition of the precision matrix (8) implies the decomposition of Σ in terms of a block-diagonal component Ω:

$$\Sigma = T\Omega T^{\mathrm{T}}, \qquad \Omega = \Omega_1 \oplus \Omega_2 \oplus \cdots \oplus \Omega_c, \qquad \Omega_i \in \mathsf{PD}(p_i). \tag{9}$$

Here $T = (I - B)^{-1} \in \mathsf{LT}_s(p)$ has diagonal blocks $I_{p_1}, \ldots, I_{p_c}$. Factorization (9) can be obtained by successive block triangularization of Σ, and the corresponding block-triangular transformation can be parameterized as

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

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 $\Sigma_{\mathrm{ltu}}(\alpha, \delta)$, and therefore also $\Sigma_{\mathrm{lt}}(\alpha)$ by the discussion surrounding (4). At the other extreme, when $D(\delta)$ consists of a single block of dimension $p \times p$, we recover $\Sigma_{\mathrm{pd}}(\delta)$. The remaining parameterization 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 given in the Supplementary Material.

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

(i). *(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;*

(ii). *(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 parameterization 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 parameterization Σ_{bt} unifies and generalizes three of the four parameterizations from (1), detached from any consideration of causal ordering; the Supplementary Material provides further discussion. The identification $\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p)$ characterizes a positive-definite matrix as a nonsingular 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 $[A] = \{OA\colon O \in \mathsf{O}(p)\} \subset \mathsf{GL}(p)$ 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} parameterization.

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} parameterization 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} parameterization, 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 § I of the Supplementary Material 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

$$[\log(\Sigma + \varepsilon I) - \log(\Sigma)]_{j,k} = \varepsilon \sum_{r,v} \left(\frac{1}{2\pi i} \oint_{\gamma} \frac{\log(z)}{(z - (\lambda_r + \varepsilon))\{z - \lambda_v\}} \, \mathrm{d}z \right) o_{jr} o_{kv} \sum_{\ell,s} o_{\ell r} o_{sv}, \tag{10}$$

where o_{ij} denotes the ith entry of the jth eigenvector of Σ. Consider the summation over ℓ and s in (10). 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 $O^{\mathrm{T}}O = OO^{\mathrm{T}} = I$. 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, (10) simplifies to

$$\big[\log(\Sigma + \varepsilon I) - \log(\Sigma)\big]_{j,k} = \varepsilon \sum_{v} \left(\frac{1}{2\pi i} \oint_{\text{anticlockwise}} \frac{\log(z)}{\{z - (\lambda_v + \varepsilon)\}(z - \lambda_v)} \, dz \right) o_{jv} o_{kv}$$

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

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

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 decomposition (9), i.e., $\Sigma = T\Omega T^{\mathrm{T}}$, where Ω is block diagonal and T is triangular. The Σ_{ltu} parameterization 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 § 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 the Supplementary Material.

Proposition 4 below establishes spectral-norm consistency of the proposed estimator under conditions detailed in the Supplementary Material. These include the following condition, 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

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

PROPOSITION 4. *Suppose that the tuning parameters of equations (J.4) and (J.5) in the Supplementary Material 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 the Supplementary Material, with* $\varphi \geqslant 1/2$, *the estimator* $\tilde{\Sigma} = \tilde{T}\tilde{\Omega}\tilde{T}^{\mathrm{T}}$ *of* $\Sigma = T\Omega T^{\mathrm{T}}$ *satisfies*

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

where

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

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 the Supplementary Material 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 in the Supplementary Material 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 parameterization, transfers substantial benefits to estimation of Σ^{-1}. In the case of Σ_o, the simple thresholding approach in § K of the Supplementary Material appears too simplistic, presumably owing to the constraints on α needed to make the parameterization 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 nonzero elements of B increases, and as the magnitude of nonzero entries (the weights of directed edges) becomes large. This follows from Proposition 1, whereby entry (i,j) of 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 nonzero 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 nonzero entries have small absolute values, thresholding in the original domain will typically yield better

results. As the number of nonzero 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} parameterization, 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 five 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} parameterization, resulting in estimates $\hat{U}_\tau$ of U and $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$ of Σ.

Results in Fig. 6(a) show that thresholding on the original scale outperforms thresholding on the logarithmic scale under the Σ_{ltu} parameterization 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 nonsparse 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 Figs. 6(b) and 6(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 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 Fig. 6(d), for comparison to (a), the ratio of the maximum ℓ_1 row norm of the two matrices, defined for $A \in \mathsf{M}(p)$ as

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

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 Fig. K.4 in the Supplementary Material. 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 Fig. 7. Interestingly, the pattern of relative behaviour mirrors that of $\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$.

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 the

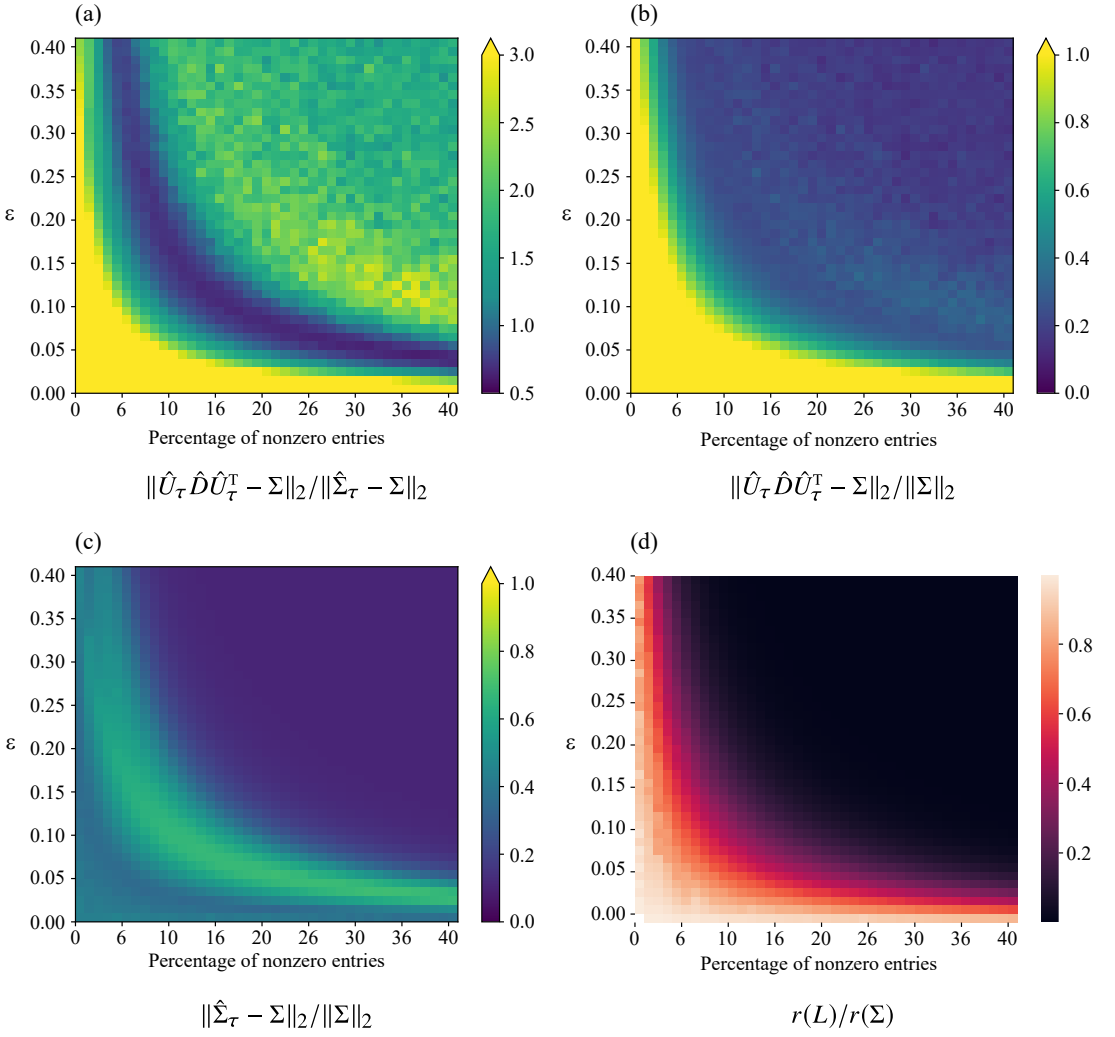

Fig. 6. (a),(b),(c) Relative ℓ_2 errors and (d) relative ℓ_1 row norms for different combinations of ϵ (y axis) and levels of sparsity of B, measured by the percentage of nonzero entries (x axis). Pixels are median values over 100 simulations with $n = 150$, $p = 100$ for different combinations of ϵ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

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

	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%)
$\hat{U}_\tau \hat{D} \hat{U}_\tau^{\mathrm{T}}$	65.5% (3.1%)	67.3% (3.6%)	69.7% (2.4%)

Supplementary Material. For leukemia patients ($p = 3571$), the results show a slight improvement in accuracy, from a median of 95.5% on the original scale to 97.7% for the 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.

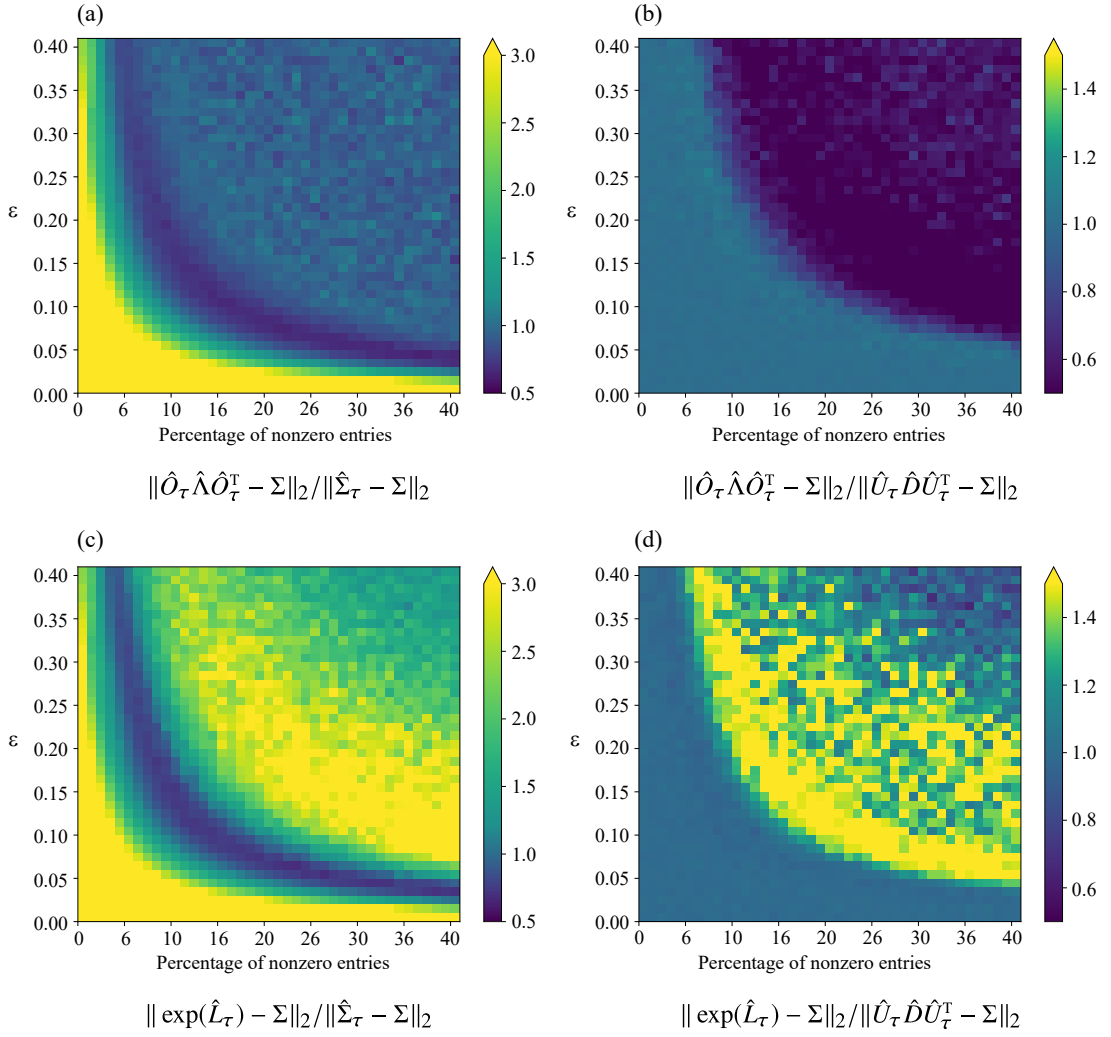

Fig. 7. Relative ℓ_2 errors for different combinations of ϵ (y axis) and levels of sparsity of B, measured by the percentage of nonzero entries (x axis). Each entry corresponds to the ratio of median errors over 100 Monte Carlo simulations with $n = 150$ and $p = 100$.

10. Closing discussion

The work has uncovered insights into the interpretation of sparsity on nonstandard 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} parameterization.

Acknowledgement

We are grateful to two 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 includes a formalized definition of reparameterization, simulation results for § 9.1 and proofs

References

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

Battey, H. S. (2017). Eigen structure of a new class of structured covariance and inverse covariance matrices. *Bernoulli* **23**, 3166–77.

Battey, H. S. (2019). On sparsity scales and covariance matrix transformations. *Biometrika* **106**, 605–17.

Battey, H. S. (2023). Inducement of population sparsity *Can. J. Statist*. **51**, 760–8.

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

Bickel, P. J. & Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–604.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *J. R. Statist. Soc. Suppl*. **5**, 171–6.

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

Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.* **8**, 204–18.

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

Draisma, J. & Zwiernik, P. (2017). Automorphism groups of Gaussian Bayesian networks. *Bernoulli* **23**, 1102–29.

Drton, M. & Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist*. **33**, 247–57.

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

Fan, J., Liao, Y. & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Statist. Soc. B* **75**, 603–80.

Grover, A. & Leskovec, J. (2016). node2vec: scalable feature learning for networks. In *Proc. 22nd ACM SIGKDD Int. Conf. Know. Disc. Data Mining*, pp. 855–64. New York: Association for Computing Machinery.

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

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

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

Skovgaard, L. T. (1984). A Riemannian geometry of the multivariate normal model. *Scand. J. Statist*. **11**, 211–23.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statist. Comp*. **17**, 395–416.

Wermuth, N. & Cox, D. R. (2004). Joint response graphs and separation induced by triangular systems. *J. R. Statist. Soc. B* **66**, 687–717.

Zwiernik, P. (2025). Entropic covariance models. *Ann. Statist.* **53**, 1371–405.

[*Received on* 16 *July* 2024. *Accepted on* 29 *September* 2025]