REGRESSION GRAPHS AND SPARSITY-INDUCING REPARAMETRIZATIONS

JAKUB RYBAK, HEATHER BATTEY, AND KARTHIK BHARATH

Summary. That parametrization and population-level sparsity are intrinsically 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 main purpose of this paper is to provide insight into structures on the physically-natural scale that induce and are induced by sparsity after reparametrization. In a sense the richest of the four structures initially uncovered turns out to be that of the joint-response graphs studied by Wermuth & Cox (2004), while the most restrictive is that induced by sparsity on the scale of the matrix logarithm, studied by Battey (2017). This points to a class of reparametrizations for the chain-graph models (Andersson et al., 2001), with undirected and directed acyclic graphs as special cases. While much of the paper is focused on exact zeros after reparametrization, an important insight is the interpretation of approximate zeros, which explains the 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 insights have a bearing on methodology, some aspects of which are discussed in the supplementary material where an estimator with high-dimensional statistical guarantees is presented.

Some key words: Chain graphs; 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 of a high-dimensional interest parameter; and to prevent accumulation of estimation error associated with a high-dimensional nuisance parameter. 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 barely touches on sample quantities; it is concerned with the more fundamental question of whether there are parametrizations enjoying a population-level sparsity not present 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.

That parametrization and sparsity are intrinsically linked was emphasized by Battey (2023), who pointed to an implicit invocation of the idea in the form of parameter orthogonalization (Cox & Reid, 1987). The latter entails a traversal of parametrization space by solving a system of partial differential equations. These are constructed in such a way that interpretation of the interest parameter is retained, and such that in the new parametrization, the nuisance parameter is orthogonal to the interest parameter in the relevant sense. In other words, the nuisance parameter is redefined in such a way that it induces sparsity on the relevant block of the Fisher information matrix, a population-level object, reducing its role. Any interpretation of the original nuisance parameter is sacrificed in favour of a more reliable inference for the interest parameter. Inducement of population-level sparsity does not seem to have received subsequent attention until the recent examples highlighted by Battey (2023), although it could be argued that the development of Gaussian graphical models (e.g. Lauritzen, 1996; Cox & Wermuth, 1996) is in this vein, together with the important work on graphical modelling for extremes by Engelke & Hitz (2020).

The motivating question for this paper is whether, for a given covariance matrix, not obviously sparse in any domain, a sparsity-inducing reparametrization can be deduced. Battey (2017) and Rybak & Battey (2021) provided a proof of concept for the idea. 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. In the case of Gaussian graphical models, the precision matrix is the interest parameter by virtue of the interpretability ascribed to its zeros. Thus, both aspects are of interest and are addressed here. A third type of situation, to which the present paper does not contribute, is when the covariance matrix is a nuisance parameter that has a known structure up to a low-dimensional parameter. This situation is common in some settings, for instance in the analysis of split-plot or Latin square designs with block effects treated as random.

The contribution of the work to be presented is manifold. A starting point is the identification of new parametrizations in which sparsity conveniently manifests in a vector space. For these new parametrizations we uncover the structure induced on physically natural scales through sparsity on the transformed scale, and the converse result of interest: that matrices encoding such structure are sparse after reparametrization. While much of the paper is focused on exact zeros after reparametrization, an important insight is the interpretation of approximate zeros, which explains the 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. A further insight, although not taken up until section 6 in order to improve readability, is the unification of old and new parametrizations via a class of matrix decompositions representing the chain graphs. This recovers the four fundamental parametrizations as special cases of those induced by the chain-graph models (Andersson et al., 2001), with undirected and directed acyclic graphs as special cases.

Because the population-level sparsity manifests in a vector space, imposition of sparsity on relevant sample quantities produces a covariance estimate that necessarily respects the positive definite cone constraint in much the same way that the canonical parametrization of the full exponential family evades complicated constraints that would otherwise arise. These benefits are transferred to any sensible methodology, and we discuss one approach with high-dimensional statistical guarantees in the supplementary material. Our view, echoing to some extent Cox (2001), is that the question of how to do the analysis is important but secondary to understanding the substantive implications of sparsity at a population level.

The paper has a strong group-theoretic and geometric underpinning, in itself enlightening but not encouraging a broad readership. We have placed most of this discussion in the supplementary material. Some readers may prefer the exposition of an earlier preprint (Rybak, Battey & Bharath, 2024).

2. NOTATION

The following subsets of the vector space $\mathsf{M}(p)$ of $p \times p$ real matrices are extensively referenced; most of them are matrix Lie groups with matrix multiplication as the group operation: the symmetric matrices $\mathsf{Sym}(p)$; the skew symmetric matrices $\mathsf{Sk}(p)$; the general linear group of nonsingular matrices $\mathsf{GL}(p)$; the orthogonal matrices $\mathsf{O}(p) \subset \mathsf{GL}(p)$; the rotation group of special orthogonal matrices $\mathsf{SO}(p) \subset \mathsf{O}(p)$ with determinant $+1$; the symmetric positive definite matrices $\mathsf{PD}(p) \subset \mathsf{GL}(p)$; the diagonal matrices $\mathsf{D}(p)$; the permutation matrices $\mathsf{P}(p) \subset \mathsf{O}(p)$; the lower triangular matrices with arbitrary diagonal entries $\mathsf{LT}(p)$; the lower triangular matrices with unit diagonal entries $\mathsf{LT}_{\mathsf{u}}(p) \subset \mathsf{LT}(p)$; the strictly lower triangular matrices $\mathsf{LT}_{\mathsf{s}}(p) \subset \mathsf{LT}(p)$. Versions of the upper triangular matrices $\mathsf{UT}(p)$ are defined analogously, and versions of $\mathsf{D}(p)$ and $\mathsf{LT}(p)$ with positive diagonal elements are differentiated using the subscript $+$.

The notation $\mathsf{V}(p)$ is used for a generic vector subspace of $\mathsf{M}(p)$ and $\mathsf{Cone}_p \subset \mathsf{Sym}(p)$ represents the interior of a convex cone within $\mathsf{Sym}(p)$ excluding the origin, an open subset of $\mathsf{Sym}(p)$. 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. The symbols $\oplus$ and $\otimes$ denote, respectively, the direct sum and product of two vector spaces; when the context is clear, $\otimes$ also represents the Kronecker product of matrices, while $A \oplus B$ denotes a block-diagonal matrix with blocks A and B. The vectors $e_1, \dots, e_d$ denote the canonical basis vectors for $\mathbb{R}^d$, where $e_i \in \mathbb{R}^d$ is a vector with 1 for its ith component and zero elsewhere. The index set $\{1, \dots, p\}$ is written $[p]$.

3. REPARAMETRIZATION

3.1. **Formulation.** The set Cone_p is, from one perspective, the natural parameter domain for parametrizing $\mathsf{PD}(p)$. The question we seek to address is whether there is another parameter domain that is perhaps less physically natural, but in which a population-level sparsity is present, encapsulating the relevant set of multivariate dependencies. In other words, we seek a reparametrization that induces sparsity on the covariance matrix Σ. The problem is addressed from the opposite direction, by first considering the parameter domains in which sparsity can be fruitfully represented, and then studying the form of multivariate dependencies that are implied by such sparsity. The deduced structures are found in section 4 to be both necessary and sufficient, so that the direction of interest is recovered through this route.

In a sense to be clarified, a construction based on the chain graph representations discussed in section 6 subsumes the dependence structures of Battey (2017) and Rybak & Battey (2021), plus those identified in section 4 in 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 models interpretation on the physically natural scale.

3.2. **Four parametrizations.** From an initial parametrization $\sigma \in \mathsf{Cone}_p$ of $\Sigma \in \mathsf{PD}(p)$, formalized in the supplementary material, we study four maps of the form $\sigma \mapsto \alpha(\sigma) \in \mathbb{R}^{p(p+1)/2}$ or $\sigma \mapsto (\alpha, d)(\sigma) \in \mathbb{R}^{p(p+1)/2}$ so that sparsity of α respects the positive definiteness constraint on Σ. These reparametrizations can equivalently be written

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

where $D(d) = \mathrm{diag}(d_1, \dots, d_p)$ and e^A denotes the matrix exponential of a square matrix A. For each of the four reparametrization maps, the parameter domain $\mathbb{R}^{p(p+1)/2}$ is identified with a different vector space. 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 indicate which of the matrix sets, $\mathsf{PD}(p)$, $\mathsf{SO}(p)$, $\mathsf{LT}_{+}(p)$ and $\mathsf{LT}_{\mathsf{u}}(p)$ respectively, parametrized in terms of α, are represented as the image of the exponential map from $\mathsf{Sym}(p)$, $\mathsf{Sk}(p)$, $\mathsf{LT}(p)$ and $\mathsf{LT}_{\mathsf{s}}(p)$ respectively. In each case $L(\alpha)$ depends on α through its expansion in the canonical basis

$$L(\alpha) = \alpha_1 B_1 + \cdots + \alpha_m B_m. \tag{3.1}$$

In an obvious notation, the canonical basis matrices corresponding to each parametrization are: $\mathcal{B}_{sym} := \{B_1, \ldots, B_{p(p+1)/2}\}$ consisting of $p(p-1)/2$ non-diagonal matrices of the form $e_j e_k^{\mathrm{T}} + e_k e_j^{\mathrm{T}}$ for $j < k$ and p diagonal matrices of the form $e_j e_j^{\mathrm{T}}$; $\mathcal{B}_{sk} := \{B_1, \ldots, B_{p(p-1)/2}\}$ consisting of skew symmetric matrices of the form $e_j e_k^{\mathrm{T}} - e_k e_j^{\mathrm{T}}$ for $j < k$; $\mathcal{B}_{lt} := \{B_1, \ldots, B_{p(p+1)/2}\}$, consisting of lower triangular matrices of the form $e_k e_j^{\mathrm{T}}$, $j \leq k$; and $\mathcal{B}_{ltu} := \{B_1, \ldots, B_{p(p-1)/2}\}$ consisting of strictly lower triangular matrices of the form $e_k e_j^{\mathrm{T}}$ with $j < k$. The canonical basis is part of the definition of the reparametrization maps (formalized in section B and C of the supplementary material).

We investigate in section 4 the structure induced on the original scale, i.e., in $\sigma \in \mathsf{Cone}_p$ or equivalently in Σ through sparsity constraints on α, and establish that the uncovered structure is both necessary and sufficient for sparsity of α. This allows us to say that the structure is both sparsity-induced and sparsity-inducing in the appropriate parametrization. Importantly, the matrix on the transformed scale can be considerably sparser than Σ, a conclusion with strong statistical implications. We revisit this point in section 7.

3.3. **Legitimacy of the four reparametrization maps.**

The matrix logarithm $\log : \mathsf{PD}(p) \to \mathsf{Sym}(p)$ of a covariance matrix was considered by Leonard & Hsu (1992) and Chiu et al. (1996) as a tool to incorporate dependence on covariates. In a precursor to the present work, Battey (2017) exploited the vector-space properties of $\mathsf{Sym}(p)$ and examined the structure induced on $\Sigma(\alpha)$ through sparsity of α in an expansion of $L = \log(\Sigma)$ in the canonical basis.

Starting from the natural parametrization $\sigma \mapsto \mathbf{vech}^{-1}(\sigma) = \Sigma$, where $\sigma \in \mathsf{Cone}_p$ and **vech** is the half-vectorization map which picks out the upper triangular part of the vectorization of Σ, 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 1). 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 (3.1) 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 of the supplementary material.

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 $d = \operatorname{diag}(D)$ can be easily incorporated and corresponds to further structure. The composition of three maps described in Figure 1 (top right) consists of a diffeomorphism $\phi_o : \mathsf{Cone}_p \to \mathbb{R}^{p(p+1)/2}$, a map b_{sk} from (α, d) to $D(d)$ and $L(\alpha)$ via (3.1) 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 of the supplementary material.

The situation is analogous for the two new reparametrization maps Σ_{lt} and Σ_{ltu}, depicted in the bottom row of Figure 1. 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 (supplementary material, Proposition C.4). The so-called LDU 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 $\mathcal{B}_{ltu}$.

$$\begin{tikzcd}
\alpha \arrow[d, "b_{sym}"] & & \sigma \arrow[ll, "\phi_{pd}"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L \in \mathsf{Sym}(p) \arrow[rr, "e^L"] & & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}
\qquad
\begin{tikzcd}
(\alpha, d) \arrow[d, shift right] \arrow[d, shift left, "b_{sk}"] & & \sigma \arrow[ll, "\phi_o"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L, D \in \mathsf{Sk}(p) \times \mathsf{D}(p) \arrow[rr, "e^L e^D (e^L)^{\mathrm{T}}"] & & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}$$

$$\begin{tikzcd}
\alpha \arrow[d, "b_{lt}"] & & \sigma \arrow[ll, "\phi_{lt}"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L \in \mathsf{LT}(p) \arrow[rr, "e^L (e^L)^{\mathrm{T}}"] & & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}
\qquad
\begin{tikzcd}
(\alpha, d) \arrow[d, shift right] \arrow[d, shift left, "b_{ltu}"] & & \sigma \arrow[ll, "\phi_{ltu}"'] \arrow[d, "\mathbf{vech}^{-1}"] \\
L, D \in \mathsf{LT}_{\mathsf{s}}(p) \times \mathsf{D}(p) \arrow[rr, "e^L e^D (e^L)^{\mathrm{T}}"] & & \mathsf{PD}(p) \ni \Sigma
\end{tikzcd}$$

FIGURE 1. Reparametrization maps for Σ_{pd} (top left), Σ_o (top right), Σ_{lt} (bottom left), and Σ_{ltu} (bottom right).

These four reparametrizations are the base cases from which a unified perspective emerges in section 6. This unification stems from the interpretation in a notional Gaussian model provided in section 5. The subsuming graphical structure is that the chain graphs (e.g. Andersson et al., 2001) which allow for directed and undirected edges between nodes.

4. STRUCTURE OF $\Sigma(\alpha)$ WHEN α IS SPARSE

Consider the matrices $L \in \mathsf{V}(p) \subset \mathsf{M}(p)$ of section 3.2, 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 $\|\alpha\|_0 = \sum_j \mathbb{I}\{\alpha_j \neq 0\} = s^* < p \ll m$. Corollaries 4.1–4.4 to be presented specify the structure induced on each of $\Sigma_{pd}(\alpha)$, $\Sigma_o(\alpha, d)$, $\Sigma_{lt}(\alpha)$ and $\Sigma_{ltu}(\alpha, d)$ through sparsity of α. The basis coefficients α and the structures expounded in Corollaries 4.1–4.4 are directly interpretable and are elucidated in section 5 from a graphical modelling perspective, assuming for some of the latter statements an underlying Gaussian model.

Corollary 4.1 for the map $\alpha \mapsto \Sigma_{pd}(\alpha)$ was derived by Battey (2017) using elementary arguments. Specifically, this entailed showing that for a matrix $L(\alpha) \in \mathsf{Sym}(p)$, represented in terms of the canonical basis $\mathcal{B}_{sym}$, a basis for $\mathsf{im}(L) := \{y \in \mathbb{R}^p : Lx = y, \; x \in \mathbb{R}^p\}$ is provided both by the eigenvectors of Σ corresponding to the non-unit eigenvalues, and by $\mathcal{L}(\alpha)$, the set of unique non-zero column vectors of the basis matrices picked out from $\mathcal{B}_{sym}$ by the non-zero elements of α. It follows that the rank of $L(\alpha)$, i.e. the dimension of $\mathsf{im}(L)$, is the equal to $d^* = |\mathcal{L}(\alpha)| < 2s^*$. The precise value of d^* depends on the configuration of α, but can be specified in expectation if the support of α is taken as a simple random sample of size s^* from $[p(p+1)/2]$. Since the column space and the row space of a symmetric matrix coincide, the non-trivial eigenvectors of $L(\alpha)$ have only d^* non-zero entries, which corresponds to a very specific structure on $\Sigma(\alpha)$, summarized in Corollary 4.1.

The analogous study of $(\alpha, d) \mapsto \Sigma_o(\alpha, d)$ by Rybak & Battey (2021) required considerable technical modification, using a representation of normal operators in terms of block-diagonal matrices, which simplifies calculations involving the matrix logarithm. The analysis to be presented for $\alpha \mapsto \Sigma_{lt}(\alpha)$ and $(\alpha, d) \mapsto \Sigma_{ltu}(\alpha, d)$ is more involved still, since triangular matrices are not necessarily normal.

Lemma 4.1 and Theorem 4.1 are general results applying in different ways to the four cases discussed in §3. They are specialized to these cases of interest in Corollaries 4.1–4.4 to expound the structure induced on $\alpha \mapsto \Sigma_\bullet(\alpha)$ or $(\alpha, d) \mapsto \Sigma_\bullet(\alpha, d)$ through sparsity of α. While the statements of Corollaries 4.1 and 4.2 are not new, new proofs in supplementary material D are presented in terms of the general formulation rather than the more elementary proofs provided by Battey (2017) and Rybak and Battey (2021).

For a matrix $M \in \mathsf{M}(p)$ possessing a matrix logarithm $M = e^L$, let $M = QJQ^{-1}$ be a real Jordan decomposition of M (e.g. Horn and Johnson, 2012, p. 202). 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 $|\mathcal{A}| = p - |\mathcal{A}^c|$.

Let d_r^* and d_c^* denote the number of non-zero rows and columns of L respectively. The number of indices for which the corresponding row or column of L is non-zero is denoted by d^*. These quantities are strongly related to s^* when α is sparse. A bound which is always valid is $\max\{d_r^*, d_c^*\} \leq 2s^*$, although $\max\{d_r^*, d_c^*\}$ can be considerably smaller than this, as it depends on the configuration of basis elements picked out by the sparse α. Indeed, it can arise that $\max\{d_r^*, d_c^*\} \ll p$ even when s^* exceeds p, provided that the configuration of non-zero elements of α produces zero rows or columns of L.

Lemma 4.1. *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, $|\mathcal{A}| \leq \max\{d_r^*, d_c^*\}$.*

For normal matrices in $\mathsf{M}(p)$, i.e. those satisfying $M^{\mathrm{T}}M = MM^{\mathrm{T}}$, $d_r^* = d_c^*$ so that Lemma 4.1 recovers Lemma 2.1 of Battey (2017) and Proposition 3.1 of Rybak and Battey (2021). Theorem 4.1 establishes general conditions for logarithmic sparsity of a matrix $M = e^L \in \mathsf{M}(p)$. In this and subsequent results, the dependence of d_r^* and d_c^* on s^* is implicit and suppressed in the notation.

Theorem 4.1. *Consider $M = e^L \in \mathsf{M}(p)$ where $L \in \mathsf{V}(p)$, a vector space with canonical basis $\mathcal{B}$ of dimension d. 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 has $p - d_r^*$ rows of the form e_j^{T} for some $j \in [p]$, all distinct, and $p - d_c^*$ columns of the form e_j. Of these, $p - d^*$ coincide after transposition. If M is normal, $d_r^* = d_c^* = d^*$.*

Figure 2 shows an example of a structure of M established by Theorem 4.1.

While arbitrary constraints on the value of s^* are avoided in Theorem 4.1 and in subsequent formal statements, 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. As discussed below, relatively large values of s^* also entail a reduction in the underlying graphical model in many cases, a point worth highlighting, but less relevant from the standpoint of inducing sparsity through reparametrization.

The importance of Theorem 4.1 lies in its implications for the four parametrizations of section 3.2. Specifically, for all four, the form of $\Sigma(\alpha)$ induced by a sparse α, established in Corollaries 4.1–4.4, follows immediately from Theorem 4.1. An example at the end of this section illustrates these results. For each of the parametrizations studied here, the induced structure for $\Sigma(\alpha)$ is necessary and sufficient for sparsity of α.

Corollary 4.1. *The image of the map $\alpha \mapsto \Sigma_{pd}(\alpha) = e^{L(\alpha)}$ is logarithmically sparse in the sense that $\|\alpha\|_0 = s^*$ in the basis representation for $L(\alpha)$ if and only if Σ is of the*

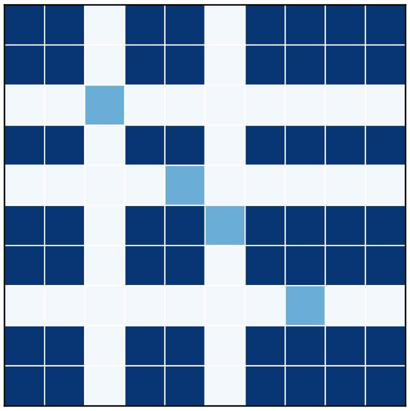


Figure 2. Example of a structure of M as established in Theorem 4.1 with $p = 10$, $d_r^* = 7$, $d_c^* = 8$ and $d^* = 9$. The entries that are zero by Theorem 4.1 are light blue, those equal to one are medium blue, and the remaining entries, whose values are unconstrained, are dark blue.

form $\Sigma = P\Sigma^{(0)}P^{\mathrm{T}}$*, where* $P \in \mathsf{P}(p)$ *is a permutation matrix and* $\Sigma^{(0)} = \Sigma_1^{(0)} \oplus I_{p-d^*}$ *with* $\Sigma_1^{(0)} \in \mathsf{PD}(d^*)$ *of maximal dimension, in the sense that it is not possible to find another permutation* $P \in \mathsf{P}(p)$ *such that the dimension of the identity block is larger than* $p - d^*$*.*

To return to the discussion of the permissible value of s^*, there are $p(p+1)/2$ basis elements for L. For α to induce a pattern of zeros in Σ of the type discussed in Corollary 4.1, 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.

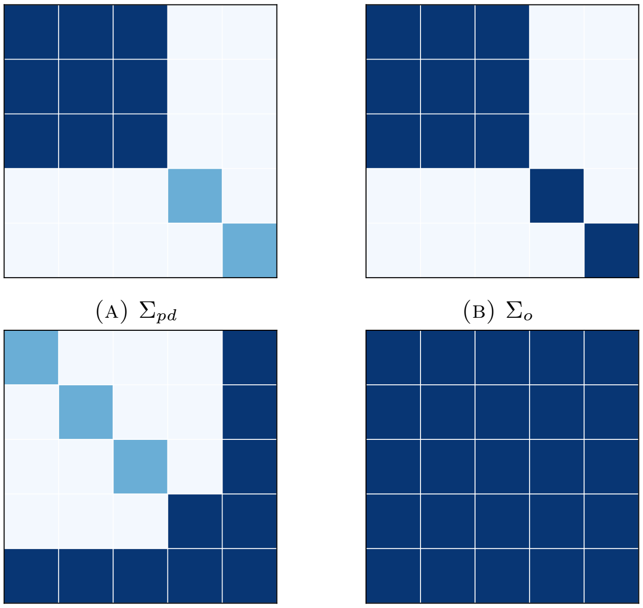
Corollary 4.2. *The image of the map* $(\alpha, d) \mapsto \Sigma_o(\alpha, d) = e^{L(\alpha)}e^{D}(e^{L(\alpha)})^{\mathrm{T}}$ *is logarithmically sparse in the sense that* $\|\alpha\|_0 = s^*$ *in the basis representation for* $L(\alpha)$ *if and only if* Σ *is of the form* $\Sigma = P\Sigma^{(0)}P^{\mathrm{T}}$*, where* $P \in \mathsf{P}(p)$ *is a permutation matrix and* $\Sigma^{(0)} = \Sigma_1^{(0)} \oplus D_{p-d^*}$*, where* $D_{p-d^*} \in \mathsf{D}(p-d^*)$ *and* $\Sigma_1^{(0)} \in \mathsf{PD}(d^*)$ *is of maximal dimension, in the sense that it is not possible to find another permutation* $P \in \mathsf{P}(p)$ *such that the dimension of the diagonal block is larger than* $p - d^*$*.*

Corollary 4.3. *The image of the map* $\alpha \mapsto \Sigma_{lt}(\alpha) = e^{L(\alpha)}(e^{L(\alpha)})^{\mathrm{T}}$ *is logarithmically sparse in the sense that* $\|\alpha\|_0 = s^*$ *in the basis representation for* $L(\alpha)$ *if and only if* Σ *is of the form* $\Sigma = VV^{\mathrm{T}}$ *, where* $V = I_p + \Theta$ *and* $\Theta \in \mathsf{LT}_+(p)$ *has* $p - d_r^*$ *zero rows and* $p - d_c^*$ *zero columns, of which* $p - d^*$ *coincide after transposition.*

Corollary 4.4. *The image of the map* $\alpha \mapsto \Sigma_{ltu}(\alpha) = e^{L(\alpha)}e^{D}(e^{L(\alpha)})^{\mathrm{T}}$ *is logarithmically sparse in the sense that* $\|\alpha\|_0 = s^*$ *in the basis representation for* $L(\alpha)$ *if and only if* Σ *is of the form* $\Sigma = U\Psi U^{\mathrm{T}}$*, where* $\Psi = e^{D} \in \mathsf{D}_+(p)$*,* $U = I_p + \Theta$ *and* $\Theta \in \mathsf{LT}_s(p)$ *has* $p - d_r^*$ *zero rows and* $p - d_c^*$ *zero columns, of which* $p - d^*$ *coincide after transposition.*

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

Figure 3 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



(A) Σ_{pd} (B) Σ_o

(C) Σ_{lt}^r (D) Σ_{lt}^c

FIGURE 3. Structure of $\Sigma(\alpha)$ induced by sparsity of α. Zero entries are in light blue, unit entries are in medium blue, and the unrestricted entries are in dark blue.

corresponding covariance matrices Σ_{lt}^r and Σ_{lt}^c should represent relationships of similar inherent structural complexity. The following lemma confirms this intuition. Specifically, Lemma 4.2 shows that although Σ_{lt}^c can be fully dense, the sparsity restriction on α induces a low-rank structure on a submatrix of Σ_{lt}^c.

Lemma 4.2. *Consider a random vector Y, $Y = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}}, Y_3^{\mathrm{T}})^{\mathrm{T}}$ with covariance matrix Σ. Let $|Y_i|$ denote the length of the subvector Y_i. Then, columns of Θ in Corollary 4.4 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 $|Y_1|$.

If the map $\alpha \mapsto \Sigma_{lt}(\alpha)$ is instead replaced by an essentially equivalent representation $\alpha \mapsto \Sigma_{ut}(\alpha)$ in terms of upper triangular matrices, the analogous structures $\Sigma_{ut}^r(\alpha)$ and $\Sigma_{ut}^c(\alpha)$ are the same as for $\Sigma_{lt}^c(\alpha)$ and $\Sigma_{lt}^r(\alpha)$ respectively.

There is a relationship between the two parametrizations $\alpha \mapsto \Sigma_{lt}(\alpha)$ and $(\alpha, d) \mapsto \Sigma_{ltu}(\alpha, d)$ if zeros in d are allowed. 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 $L_s \in \mathsf{LT}_{\mathsf{s}}(p)$ and using the properties of the matrix exponential,

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

which recovers the LTU decomposition 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}.

5. Statistical interpretation in a notional Gaussian model

5.1. **Introduction.** All results so far have held irrespective of any distributional assumptions on the outcomes. The purpose of this section is to provide insight into the interpretation of sparsity on the transformed scale and thereby the implicit assumptions involved in enforcing sparsity when it is only approximately present. To this end, we consider a real or notional Gaussian graphical model, to which an interpretation in terms of conditional independencies can be attached. In section 5.2 we show how an interpretation of α in the Σ_{ltu} parametrization emerges from the work of Cox & Wermuth (1993) and Wermuth & Cox (2004), and particularly the relevance of an approximate zero in α, which has a strong implication for the application of these ideas. In section 5.3 interpret the structure uncovered in Corollary 4.4 in terms of role of underlying random variables as e.g., source, sink or transition nodes within a real or notional Gaussian model. Importantly, the analysis of section 5.2 points to a generalization and unification of the ideas presented so far, which is developed completely in section 6. There, we explain how the parametrizations Σ_{pd}, Σ_{lt} and Σ_{ltu} all arise as a special case of a system of reparametrizations from a representation of the chain graphs due to Andersson et al. (2001).

5.2. **Interpretation of basis coefficients.** With $[p] = \{1, \ldots, p\}$, let $a \subset [p]$ and $b = [p]\backslash a$ be disjoint subsets of variable indices, and in a departure from previous notation aligning with graphical modelling convention, let I_{aa} denote the identity matrix of dimension $|a|$. As a consequence of a block-diagonalization identity for symmetric matrices (Cox & Wermuth, 1993; Wermuth & Cox, 2004),

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} = \begin{pmatrix} I_{aa} & 0 \\ \Sigma_{ba}\Sigma_{aa}^{-1} & I_{bb} \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & 0 \\ 0 & \Sigma_{bb.a} \end{pmatrix} \begin{pmatrix} I_{aa} & \Sigma_{aa}^{-1}\Sigma_{ab} \\ 0 & I_{bb} \end{pmatrix}. \tag{5.1}$$

The matrix identity (5.1) holds independently of any distributional assumptions on the underlying random variables. 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 (Terras, 1988, Chapter 4). For a statistical interpretation, let $Y = (Y_a^{\mathrm{T}}, Y_b^{\mathrm{T}})^{\mathrm{T}}$ be a mean-centred random vector with covariance matrix Σ, $\Pi_{b|a}$ 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 residual covariance matrix, i.e. $Y_b = \Pi_{b|a}Y_a + \varepsilon_b$ and $\Sigma_{bb.a} = \mathrm{var}(\varepsilon_b)$. More explicitly, on fixing a component $i \in b$ and $j \in a$, the corresponding component of $\Pi_{b|a}$ is

$$\beta_{i.ja} = \frac{\partial \mathbb{E}\{(Y_i \mid Y_{a\backslash j} = y_{a\backslash j})(y_j)\}}{\partial y_j}. \tag{5.2}$$

The conditional expectation indicates conditioning on observed values of $Y_{a\backslash j}$ with y_j as a free variable. Although the ordering in Corollary 4.4 is immaterial, it is notationally convenient to order the underlying random variables as $Y = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ and apply the block-diagonalization identity recursively with, say $a(1) = 1$, $b(1) = [p]\backslash a(1)$, $a(k) = k$, $b(k) = b(k-1)\backslash a(k)$. There results a block diagonalization in 1×1 blocks, recovering the LDU decomposition. In this representation, the iith entry of $\Psi = e^D$ from Corollary 4.4 is given by $\Sigma_{ii.[i-1]}$, a mean-squared error from the linear projection of Y_i on $Y_1, \ldots, Y_{i-1}$, and the ijth entry of the lower-triangular matrix U, for $j < i$, is given by $U_{ij} = \beta_{i.j[j]}$, that is, the coefficient of Y_j in a regression of Y_i on $Y_1, \ldots, Y_j$.

Further interpretation for the entries of U is obtained by considering an upper-triangular decomposition of the precision matrix. 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,

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

An alternative expression for the matrix of regression coefficients of Y_b in a regression of Y_b on Y_a is thus $\Pi_{b|a} = -(\Sigma^{bb})^{-1}\Sigma^{ba}$ (Wermuth & Cox, 2004), which corresponds to the 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, $\bar{U} \in \mathsf{LT_s}(p)$ 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,

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

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

To illustrate this, consider a set of three variables (Y_1, Y_2, Y_3). The marginal effect of Y_3 on Y_1 is related to the conditional effects through Cochran's recursion (Cochran, 1938),

$$\beta_{3.1} = \beta_{3.12} + \beta_{3.21}\beta_{2.1}, \tag{5.3}$$

which generalizes to an arbitrary number of variables in the manner indicated below. The right-hand side of equation (5.3) 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 set equal to the corresponding 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, namely $Y_1 \to Y_3$ and $Y_1 \to Y_2 \to Y_3$, which correspond to the first and second term in (5.3) respectively.

The lower-triangular matrices U and $\bar{U}$ have the form,

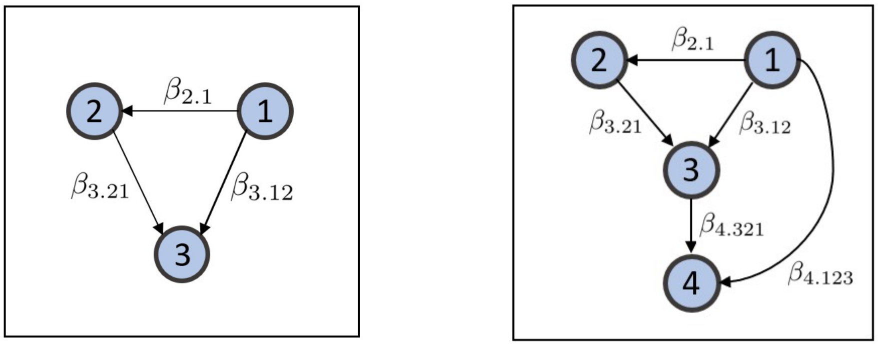
$$\bar{U} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{2.1} & 0 & 0 \\ \beta_{3.12} & \beta_{3.21} & 0 \end{pmatrix}, \quad U = I + \bar{U} + \bar{U}^2 = \begin{pmatrix} 1 & 0 & 0 \\ \beta_{2.1} & 1 & 0 \\ \beta_{3.12} + \beta_{3.21}\beta_{2.1} & \beta_{3.21} & 1 \end{pmatrix}. \tag{5.4}$$

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. When the joint distribution of (Y_1, Y_2, Y_3) is Gaussian, linear regression coefficients correspond to conditional dependencies, and the trivariate system can be represented by a directed acyclic graph, as shown in Figure 4(A).

For a system of four variables, (Y_1, Y_2, Y_3, Y_4) with a corresponding directed acyclic graph of the form shown in Figure 4(B), we have,

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \beta_{2.1} & 1 & 0 & 0 \\ \beta_{3.12} + \beta_{3.21}\beta_{2.1} & \beta_{3.21} & 1 & 0 \\ \beta_{4.123} + \beta_{4.321}\beta_{3.21}\beta_{2.1} + \beta_{4.321}\beta_{3.12} & \beta_{4.213} + \beta_{4.321}\beta_{3.21} & \beta_{4.321} & 1 \end{pmatrix}. \tag{5.5}$$

Again, U_{ij} for $j < i$ corresponds to the sum of effects of Y_j on Y_i along all directed paths connecting the two nodes, with edge weights given by regression coefficients, i.e. the marginal effect of Y_j on Y_i via Cochran's recursion.



(A) (B)

FIGURE 4. Directed acyclic graphs with edge weights given by regression coefficients β.

Using the properties of the matrix logarithm,

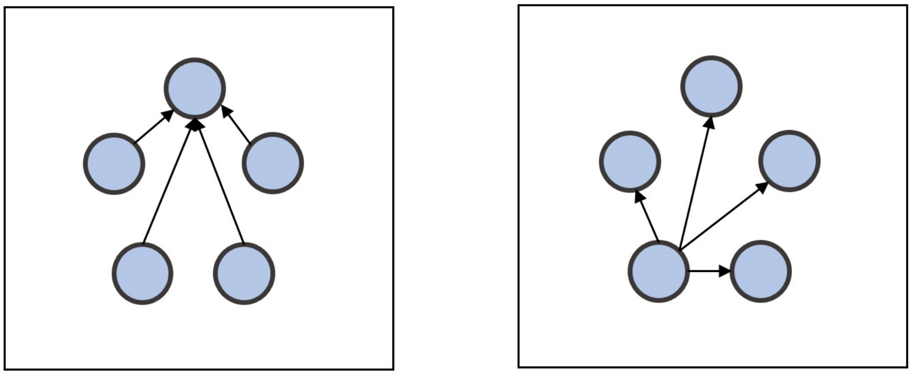
$$L = \log(U) = \log[(I - \bar{U})^{-1}] = -\log(I + \bar{U}) = \sum_{k=1}^{p-1} \frac{\bar{U}^k}{k}.$$

Thus, 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, a logarithmic transformation promotes approximate sparsity of the triangular matrix, the interepretation of a near-zero being that relatively direct paths have a small effect. This indicates the type of approximation that would be inherent in a statistical algorithm that sets small values of α to zero: in effect, 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. Since relatively short paths are likely to be more interpretable, logarithmic sparsity may have some advantages from a conventional statistical modelling point of view, where the goal is to isolate stable effects of core scientific importance, and where overconditioning leads to degeneracy (e.g. Bardorff-Nielsen and Cox, 1994, p. 29). The most compelling applications for these ideas are probably in the realm of genomics, however length-based weighted averages of paths between nodes have been proposed in other application areas in which shorter paths are deemed more important than longer ones (e.g. Estrada & Higham, 2010; Kloster & Gleich, 2014).

5.3. **Interpretation of sparsity structures.** Using the natural ordering in equation (5.2), the entries of U encapsulate dependencies between each pair of variables, Y_j, Y_i, $j < i$, conditional on other variables among $Y_1, \ldots, Y_{j-1}$, and marginalizing over the remaining variables $Y_{j+1}, \ldots, Y_{i-1}, Y_{i+1}, \ldots Y_p$.

In the terminology of e.g. Cox & Wermuth (1996), marginalization over a variable k indicated by $\not{k}$ in the depiction below, induces an edge between i and j if the marginalized variable is a transition node or a source node in a parent graph. No direction in the induced edge is implied if the marginalized variable is a source node, which is indicated by ----.

$$\begin{array}{cc} i \longleftarrow \not{k} \longrightarrow j, & i \longleftarrow \not{k} \longleftarrow j, \\ i \text{ ---- } j, & i \longleftarrow j. \end{array}$$



(A) $\Sigma_{lt}^{r}(\alpha)$ (B) $\Sigma_{lt}^{c}(\alpha)$

FIGURE 5. Directed acyclic graphs corresponding to an example depicted in Figure 3. Arrows indicate directed edges and nodes correspond to random variables.

By contrast, if i and j are separated by a sink node in a, then conditioning on such a variable, indicated by $\boxed{\circ}$ below, is edge inducing, with no direction implied.

$$i \longrightarrow \boxed{\circ} \longleftarrow j,$$
$$i \text{ — } j.$$

The entry U_{ij} for $j < i$ is zero if the effects of Y_j on Y_i cancel out. This happens, for example, when $\beta_{3.12} = -\beta_{3.21}\beta_{2.1}$ in equation (5.3). Alternatively, when no cancellations are present, $U_{ij} = 0$, $i > j$, implies an absence of a directed path from j to i in the corresponding directed acyclic graph. In addition, nodes $1, \dots, j-1$ cannot be sink nodes, and nodes $j+1, \dots, i-1, i+1, \dots, p$ cannot be transition or source nodes.

An assumption that Y is Gaussian of mean zero and covariance Σ allows the zeros in U to be interpreted as conditional independencies. If the ith row of $\Theta = U - I_p$ consists entirely of zeros, the dependence between i and any $j < i$ can be explained away by conditioning on the remaining variables $[j-1]$. This can happen for all $j < i$ only if $i \perp\!\!\!\perp j$ for all $j < i$, i.e., there are no direct edges between i and $j < i$, and the only connecting nodes, if any, are sink nodes $k > i$.

$$i \longrightarrow k \longleftarrow j, \qquad j < i < k.$$

Suppose that $U - I_p$ from Corollary 4.4 has a zero jth column. This implies that for every $i > j$, the only source or transition nodes connecting i and j are in the conditioning sets $[j-1]$ (otherwise dependence is induced through marginalization), and that there are no sink nodes among these conditioning variables (as conditioning on sink-nodes is edge-inducing). Consequently, the examples from Figure 3 can be represented by directed acyclic graphs presented in Figure 5.

6. SPARSE PARAMETRIZATIONS FOR CHAIN GRAPHS

The interpretation of basis coefficients discussed in section 5.2 points to two ways in which sparsity after reparametrization can induce graphical structure. Corollaries 4.1 and 4.2 induce undirected graphical structure, while Corollaries 4.3 and 4.4 relate to sparsity on the edges of a directed acyclic graph. So-called chain graphs have both directed and undirected components and therefore subsume the structures identified earlier as synonymous with sparsity after reparametrization. Corollary 6.1 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 the results of Corollaries 4.1, 4.3 and 4.4 as special cases, the first two being edge cases in which all connected components consist of undirected edges and directed edges respectively.

Consider a graph $G = (V, E)$ with possibly both directed and undirected edges. We assume that G contains no semi-directed cycles. Then G constitutes a chain graph (Drton & Eichler, 2006, p. 83). When two nodes $v, w \in V$ are connected by a path connected solely of undirected edges, we say that u and w are equivalent. Let $\mathcal{T}$ be a set of equivalence classes, called chain components, of this equivalence relation. Define a new graph $\mathcal{D} = (\mathcal{T}, \mathcal{E})$ with chain components as nodes. Since we assume that there are no semi-directed cycles in G, the graph of $\mathcal{D}$ is a directed acyclic graph. The graph G is thus decomposed into a directed acyclic graph $\mathcal{D}$, and a set of undirected graphs corresponding to chain components.

In Gaussian graphical models, chain graphs are usually characterized by the so-called alternative Markov property (Andersson et al., 2001). A normal distribution $N(0, \Sigma)$ satisfies the alternative Markov property if and only if

$$\Sigma^{-1} = (I - B)\Omega^{-1}(I - B^{\mathrm{T}}), \tag{6.1}$$

where Ω^{-1}_{uv} is zero if an undirected edge (u, v) is not in the graph, and $B_{uv} = 0$ 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 $B_{uv} = 0$, and vice-versa. Every directed acyclic graph can be represented by a triangular matrix, so we can always find a permutation matrix P such that decomposition (6.1) 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 permutation matrix 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 a convenient ordering of variables has been chosen such that B is triangular and Ω^{-1} is block-diagonal. The factorization (6.1) then represents the precision matrix as a product of block-diagonal and block-triangular matrices. The block-triangular matrix captures connections between groups, where groups are defined by chain components, while the block-diagonal matrix describes within-group connections.

Suppose that a random vector $X \sim N(0, \Sigma)$ is partitioned into c blocks, $X = (X_1, \ldots, X_c)$, where each block X_i constitutes a chain component, that is, the variables within X_i form a connected undirected graphical model. Let $p_i = |X_i|$ denote the dimension of the sub-vector X_i. Decomposition of the precision matrix (6.1) implies a decomposition of the covariance matrix,

$$\Sigma = T\Omega T^{\mathrm{T}}, \tag{6.2}$$

where (recalling the definition of $\oplus$ from section 2) $\Omega = \Omega_1 \oplus \Omega_2 \oplus \cdots \oplus \Omega_c$, $\Omega_i \in \mathsf{PD}(p_i)$, and $T = (I - B^{\mathrm{T}})^{-1} \in \mathsf{LT}_s(p)$, with diagonal blocks of the form $I_{p_1}, \ldots, I_{p_c}$, where subscripts indicate the dimensions. The factorization (6.2) can be obtained by successive block-triangularization of Σ.

The corresponding block-triangular transformation can be written

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

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 $\delta = (\delta_1^{\mathrm{T}} \ldots \delta_c^{\mathrm{T}})^{\mathrm{T}}$. $L(\alpha)$ is block-triangular with c diagonal blocks, each equal to a $p_i \times p_i$ identity matrix. Let p_δ denote the dimension of δ. Then, $\alpha \in \mathbb{R}^{p(p+1)/2 - p_\delta}$.

The parametrization $\Sigma_{bt}(\alpha, \delta)$ unifies and generalizes parameterizations introduced in Section 3. To see this, note that when $D(\delta)$ is diagonal we recover the parameterization $\Sigma_{ltu}(\alpha, \delta)$, whereas when $D(\delta)$ consists of a single block of dimension $p \times p$, we recover $\Sigma_{pd}(\delta)$.

To distinguish the sparsity structure induced on T and Ω, we use d_r^* and d_c^* to represent the number of non-zero rows and columns of L, and d^* to represent the number of unique indices of non-zero rows and columns of L. An analogous quantity for $D(\delta)$ is denoted by d_δ^*.

Corollary 6.1. *The image of the map* $(\alpha, \delta) \mapsto \Sigma_{bt}(\alpha, \delta) = e^{L(\alpha)} e^{D(\delta)} (e^{L(\alpha)})^{\mathrm{T}}$ *is logarithmically sparse in the sense that* $\|\alpha\|_0 = s_\alpha^*$ *and* $\|\delta\|_0 = s_\delta^*$ *in the basis representation of* $L(\alpha)$ *and* $D(\delta)$ *respectively if and only if* $\Sigma = T\Omega T^T$ *where* $\Omega = \Omega_1 \oplus \Omega_2 \oplus \cdots \oplus \Omega_c$*, for some* $c \leq p$*, and*

1. (*Sparsity of DAG*)*:* $T = I_p + A$ *and* $A \in \mathsf{LT_s}(p)$ *has* $p - d_r^*$ *zero rows and* $p - d_c^*$ *zero columns, of which* $p - d^*$ *coincide after transposition.*
2. (*Sparsity of chain components*)*:* Ω *is of the form* $\Omega = P\Omega^{(0)} P^{\mathrm{T}}$*, where* $P \in \mathsf{P}(p)$ *is a permutation matrix and* $\Omega^{(0)} = \Omega_1^{(0)} \oplus D_{p-d_\delta^*}$*, where* $D_{p-d_\delta^*} \in \mathsf{D}(p - d_\delta^*)$ *and* $\Omega_1^{(0)} \in \mathsf{PD}(d_\delta^*)$ *is block-diagonal and of maximal dimension, in the sense that it is not possible to find another permutation* $P \in \mathsf{P}(p)$ *such that the dimension of the diagonal block is larger than* $p - d_\delta^*$*.*

Note that $L(\alpha) = \log(T) = -\log(I - B^{\mathrm{T}}) = -(\log(I - B))^{\mathrm{T}}$. With the appropriate basis matrices, the basis coefficients of $\log(I - B)$ are equal to $-\alpha$. Thus, sparsities of $I - B$ and T on the transformed 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 6.1 can be obtained for the precision matrix by noticing that d_r^* and d_c^* are equal to the number of non-zero columns and rows of $-L(\alpha)^{\mathrm{T}} = \log(I - B)$ respectively. Part (1) of Corollary 6.1 thus describes structures arising when the sparsity patterns of T and $I - B$ coincide after transposition. For example, suppose that the ith column of $I - B$ is zero, that is, node i has no parents. Then, the ith row of $I - T$ is zero as well, since none of the variables $X_1, \dots, X_{i-1}$ has any effect on X_i. This is reflected on the transformed scale by a zero ith row of $L(\alpha)$.

The block triangularization in (6.2) represents the partial Iwasawa coordinates of the positive definite matrix Σ^{-1} arising from the Iwasawa decomposition of the general linear group $\mathsf{GL}(p)$ of invertible matrices (Terras, 1988). This follows from the identification $\mathsf{PD}(p) \cong \mathsf{GL}(p)/\mathsf{O}(p)$ seen via the fact that $X^{\mathrm{T}} X$ is positive definite for every $X \in \mathsf{GL}(p)$ (Section E in supplementary material). 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.

Interestingly, information geometry of the zero-mean multivariate Gaussian (Skovgaard, 1984) induced by 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.

7. SPARSITY COMPARISONS

7.1. **Exact zeros.** Section 4 studied the structure induced on Σ, in terms of exact zeros, through sparsity of α. An aspect not discussed in detail there is the observation of Battey (2017) and Rybak & Battey (2021) that α may be considerably sparser than its half-vectorization $\sigma = \mathbf{vech}(\Sigma(\alpha))$, both in terms of exact zeros and approximate zeros. This is most clearly illustrated through the parametrization $\alpha \mapsto \Sigma_{pd}(\alpha)$. As shown by Battey (2017), and as is implicit in Corollary 4.1, the number of non-unit eigenvalues of $\Sigma_{pd}(\alpha)$

when $\|\alpha\|_0 = s^* < p$ is $d^* < 2s^*$ and the number non-zero entries of any eigenvector ξ_j associated with a non-unit eigenvalue λ_j of Σ is also d^*. Thus, with $\mathcal{A} = \{j : \lambda_j \neq 1\}$, the (q,r)th entry of the matrix logarithm is $L_{qr} = \sum_{j\in\mathcal{A}} \log(\lambda_j)\xi_{qj}\xi_{rj}$. If $q \in \mathcal{A}$, then $\xi_{qj} \neq 0$ and $\log(\lambda_j) \neq 0$ for $j \in \mathcal{A}$. However $L_{qr} = 0$ is still feasible by cancellation in the sum even if $q, r \in \mathcal{A}$. Suppose that $L_{qr} = 0$. If $q \in \mathcal{A}^c$ or $r \in \mathcal{A}^c$ then $\Sigma_{qr} = 0$ by Corollary 4.1. However if $L_{qr} = 0$ and $q, r \in \mathcal{A}$, then $\Sigma_{qr} \neq 0$ by construction, illustrating the possibility that $L = \log(\Sigma)$ may be substantially sparser than Σ. A random sparse L is necessarily sparser in terms of exact zeros than e^L. However a random dense Σ with the structure specified by Corollary 4.1 almost surely has a matrix logarithm with the same degree of sparsity in terms of exact zeros as Σ. Notwithstanding, it is often the case that L is much sparser than Σ under more general notions of sparsity. A similar phenomenon applies in the other parametrizations. The relative sparsity levels in the four parametrizations are explored in §7.2.

7.2. **Approximate zeros.**

A more general notion of sparsity suitable for the four cases is q-sparsity. Defined for $\alpha \mapsto \Sigma_{pd}(\alpha)$ and $\alpha \mapsto \Sigma_{lt}(\alpha)$ as

$$s_q^* = s_q^*(\alpha) = \left(\sum_{j=1}^{p(p+1)/2} |\alpha_j|^q\right)^{1/q}$$

and for $(\alpha, d) \mapsto \Sigma_o(\alpha, d)$ and $(\alpha, d) \mapsto \Sigma_{ltu}(\alpha, d)$ as

$$s_q^* = s_q^*(\alpha, d) = \left(\sum_{j=1}^{p(p-1)/2} |\alpha_j|^q + \sum_{j=1}^{p} |d_j|^q\right)^{1/q}.$$

A comparable definition of q-sparsity for $\Sigma \in \mathsf{PD}(p)$ is

$$s_q = s_q(\sigma) = \left(\sum_{j=1}^{p(p+1)/2} |\sigma_j|^q\right)^{1/q}.$$

In the simulations to follow, sparsity after reparametrization is compared to sparsity on the original scale via the ratio, s_q^*/s_q.

We consider both directions of the discussion in §7.1 by simulation for $q > 0$. Let $\mathsf{V}(p)$ be the relevant vector space for each of the four parametrizations. In one direction, we generate a sparse $L \in \mathsf{V}(p)$ or $(L, D) \in \mathsf{V}(p) \otimes \mathsf{D}(p)$ and compare their q-sparsity level to that of the corresponding covariance matrix $\Sigma \in \mathsf{PD}(p)$. For the converse question, we generate a dense $\Sigma \in \mathsf{PD}(p)$ and compare its q-sparsity level to that of $L \in \mathsf{V}(p)$ or $(L, D) \in \mathsf{V}(p) \otimes \mathsf{D}(p)$ after reparametrization.

7.2.1. *Simulations for sparse $L \in V(p)$.*

Figure 6 was obtained by generating 250 realizations of a random, sparse $L \in \mathsf{V}(p)$ with $p = 20$ by taking the support of α to be random samples of size s^* from the index set $\{1, \ldots, p(p-1)/2\}$ for the two cases $L \in \mathsf{Sk}(p)$ and $L \in \mathsf{LT_s}(p)$ or from the index set $\{1, \ldots, p(p+1)/2\}$ for the two cases $L \in \mathsf{Sym}(p)$ and $L \in \mathsf{LT}(p)$. This was done for different values of s^* as indicated in Figure 1. The values of the nonzero basis coefficients were then drawn from a uniform distribution on $[-1, 1]$. The positive elements of the diagonal matrices in the decompositions $\Sigma_o = O\Lambda O^{\mathrm{T}}$ and $\Sigma_{ltu} = U\Psi U^{\mathrm{T}}$ are obtained from χ_4^2 and $\exp(1)$ distributions, as indicated.

7.2.2. *Simulations for dense $\Sigma \in PD(p)$.*

Rather than constructing Σ from a sparse $L \in \mathsf{V}(p)$, we now sample covariance matrices directly, and construct the corresponding tranformation on the logarithmic scale. A random dense Σ is obtained from the representation $\Sigma = O\Lambda O^T$ by generating O uniformly over $\mathsf{SO}(p)$ according to Haar measure and by generating the positive of the diagonal matrix Λ from a gamma distribution of shape τ and scale κ, written $\Gamma(\tau, \kappa)$. The results are reported in Figure 7.

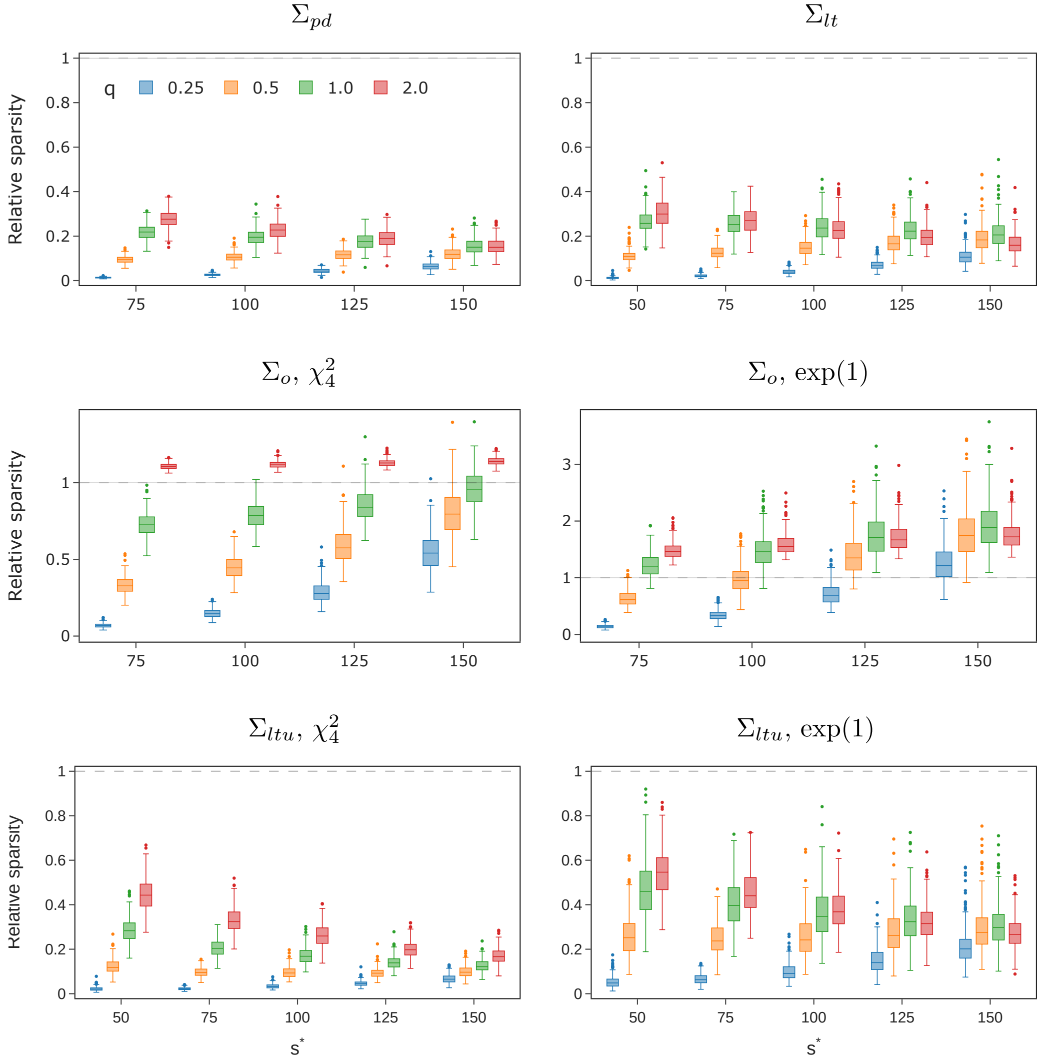
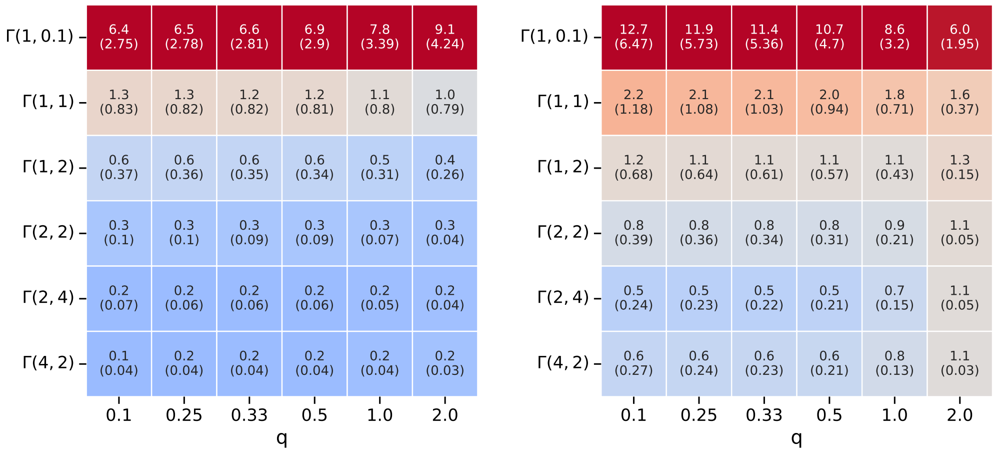


FIGURE 6. Simulations for random sparse $L \in \mathsf{V}(p)$. Simulated ratios s_q^*/s_q for different values of $s^* = \|\alpha\|_0$ and q.

8. METHODOLOGICAL IMPLICATIONS

The insights uncovered in this work transfer benefits to any reasonable methodology. The supplementary material offers one possibility, exploiting both the structure uncovered on the original scale and the sparsity on the transformed scale. Theoretical guarantees are provided for the resulting estimators, under a notional double-asymptotic regime in which the dimension $p = p(n)$ is allowed to grow with n at rate $\log p/n \to 0$. In particular, the formulation is such that approximate rather than exact zeros are allowed in α. We find in Proposition I.1 of the supplementary material that the resulting estimator $\tilde{\Sigma}$ converges in spectral norm at rate

$$\|\tilde{\Sigma} - \Sigma\|_2 = O_p\left(s(p)\, \left(n^{-1} \log p\right)^{\frac{(1-q)\kappa}{2(\kappa+1)}} \right),$$

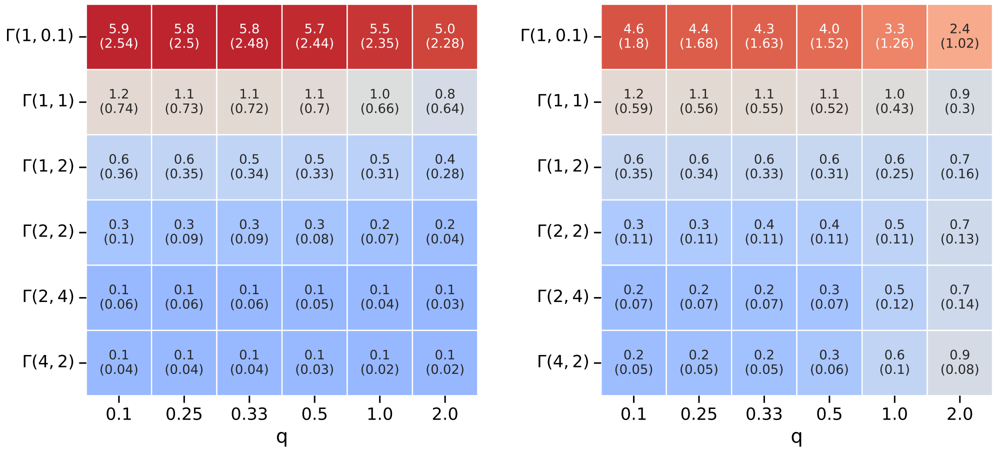


	0.1	0.25	0.33	0.5	1.0	2.0
Γ(1, 0.1)	6.4 (2.75)	6.5 (2.78)	6.6 (2.81)	6.9 (2.9)	7.8 (3.39)	9.1 (4.24)
Γ(1, 1)	1.3 (0.83)	1.3 (0.82)	1.2 (0.82)	1.2 (0.81)	1.1 (0.8)	1.0 (0.79)
Γ(1, 2)	0.6 (0.37)	0.6 (0.36)	0.6 (0.35)	0.6 (0.34)	0.5 (0.31)	0.4 (0.26)
Γ(2, 2)	0.3 (0.1)	0.3 (0.1)	0.3 (0.09)	0.3 (0.09)	0.3 (0.07)	0.3 (0.04)
Γ(2, 4)	0.2 (0.07)	0.2 (0.06)	0.2 (0.06)	0.2 (0.06)	0.2 (0.05)	0.2 (0.04)
Γ(4, 2)	0.1 (0.04)	0.2 (0.04)	0.2 (0.04)	0.2 (0.04)	0.2 (0.04)	0.2 (0.03)

(A) Σ_{pd}

	0.1	0.25	0.33	0.5	1.0	2.0
Γ(1, 0.1)	12.7 (6.47)	11.9 (5.73)	11.4 (5.36)	10.7 (4.7)	8.6 (3.2)	6.0 (1.95)
Γ(1, 1)	2.2 (1.18)	2.1 (1.08)	2.1 (1.03)	2.0 (0.94)	1.8 (0.71)	1.6 (0.37)
Γ(1, 2)	1.2 (0.68)	1.1 (0.64)	1.1 (0.61)	1.1 (0.57)	1.1 (0.43)	1.3 (0.15)
Γ(2, 2)	0.8 (0.39)	0.8 (0.36)	0.8 (0.34)	0.8 (0.31)	0.9 (0.21)	1.1 (0.05)
Γ(2, 4)	0.5 (0.24)	0.5 (0.23)	0.5 (0.22)	0.5 (0.21)	0.7 (0.15)	1.1 (0.05)
Γ(4, 2)	0.6 (0.27)	0.6 (0.24)	0.6 (0.23)	0.6 (0.21)	0.8 (0.13)	1.1 (0.03)

(B) Σ_{o}



	0.1	0.25	0.33	0.5	1.0	2.0
Γ(1, 0.1)	5.9 (2.54)	5.8 (2.5)	5.8 (2.48)	5.7 (2.44)	5.5 (2.35)	5.0 (2.28)
Γ(1, 1)	1.2 (0.74)	1.1 (0.73)	1.1 (0.72)	1.1 (0.7)	1.0 (0.66)	0.8 (0.64)
Γ(1, 2)	0.6 (0.36)	0.6 (0.35)	0.5 (0.34)	0.5 (0.33)	0.5 (0.31)	0.4 (0.28)
Γ(2, 2)	0.3 (0.1)	0.3 (0.09)	0.3 (0.09)	0.3 (0.08)	0.2 (0.07)	0.2 (0.04)
Γ(2, 4)	0.1 (0.06)	0.1 (0.06)	0.1 (0.06)	0.1 (0.05)	0.1 (0.04)	0.1 (0.03)
Γ(4, 2)	0.1 (0.04)	0.1 (0.04)	0.1 (0.04)	0.1 (0.03)	0.1 (0.02)	0.1 (0.02)

(C) Σ_{lt}

	0.1	0.25	0.33	0.5	1.0	2.0
Γ(1, 0.1)	4.6 (1.8)	4.4 (1.68)	4.3 (1.63)	4.0 (1.52)	3.3 (1.26)	2.4 (1.02)
Γ(1, 1)	1.2 (0.59)	1.1 (0.56)	1.1 (0.55)	1.1 (0.52)	1.0 (0.43)	0.9 (0.3)
Γ(1, 2)	0.6 (0.35)	0.6 (0.34)	0.6 (0.33)	0.6 (0.31)	0.6 (0.25)	0.7 (0.16)
Γ(2, 2)	0.3 (0.11)	0.3 (0.11)	0.4 (0.11)	0.4 (0.11)	0.5 (0.11)	0.7 (0.13)
Γ(2, 4)	0.2 (0.07)	0.2 (0.07)	0.2 (0.07)	0.3 (0.07)	0.5 (0.12)	0.7 (0.14)
Γ(4, 2)	0.2 (0.05)	0.2 (0.05)	0.2 (0.05)	0.3 (0.06)	0.6 (0.1)	0.9 (0.08)

(D) Σ_{ltu}

FIGURE 7. Simulations for random dense $\Sigma \in \mathsf{PD}(p)$. Simulated averages (standard deviations) of the ratio s_q^*/s_q for different values of q and different distributions of eigenvalues of Σ. Blue entries signify greater sparsity after reparametrization.

where $s(p)$ and q characterize the sparsity of the logarithmic transformations of T and Ω from (6.2) in a sense made explicit in the supplementary material. The parameter κ controls the rate of growth of the number of parents for each node as $n, p \to \infty$. The approach is based on thresholding (Bickel & Levina, 2008a,b) after logarithmic transformation of a suitable pilot estimator. Ideally a more direct approach might be found, perhaps along the lines of Zwiernik (2023) whose elegant formulation in terms of Bregman divergences covers linear constraints on a class of matrix functions of a covariance matrix, which includes the matrix logarithm, the matrix inverse, and certain matrix powers.

Acknowledgements

KB thanks Sergey Oblezin for helpful discussions on Lie group decompositions, and acknowledges support from grants EPSRC EP/V048104/1, EP/X022617/1, NSF 2015374 and NIH R37-CA21495. HB acknowledges support from an EPSRC research fellowship under grant number EP/T01864X/1.

References

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.

Cox, D. R. (2001). Discussion of 'Statistical modelling: the two cultures' by Leo Breiman. *Statist. Sci.*, 16, 216–218.

Chiu, T. Y. M., Leonard, T. and Tsui, K-W.(1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.*, 91, 198-210.

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.

Estrada, E. and Higham, D. J. (2014). Network properties revealed through matrix functions. *SIAM review*, 52, 696–714.

Kloster, K. and Gleich, D. F. (2014). Heat kernel based community detection. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1386–1395.

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

Leonard, T an Hsu, J. S. J.(1992). Bayesian inference for a covariance matrix. *Ann. Statist.*, 20, 1669-1696.

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

Rybak, J., Battey, H. S. and Bharath, K. (2024). Regression graphs and sparsity-inducing reparametrizations. *arXiv:2402.09112v1.*

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

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

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

Zwiernik, P. (2023). Entropic covariance models. *arXiv:2306.03590.*

Jakub Rybak, Heather Battey: Department of Mathematics, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK

Email address: `jakub.rybak18@imperial.ac.uk` `h.battey@imperial.ac.uk`

Karthik Bharath: School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK.

Email address: `Karthik.Bharath@nottingham.ac.uk`