# Group symmetry and covariance regularization

## Parikshit Shah and Venkat Chandrasekaran

*University of Wisconsin, Madison and California Institute of Technology*
*e-mail:* pshah@discovery.wisc.edu; venkatc@caltech.edu

**Abstract:** Statistical models that possess symmetry arise in diverse settings such as random fields associated to geophysical phenomena, exchangeable processes in Bayesian statistics, and cyclostationary processes in engineering. We formalize the notion of a symmetric model via group invariance. We propose projection onto a group fixed point subspace as a fundamental way of regularizing covariance matrices in the high-dimensional regime. In terms of parameters associated to the group we derive precise rates of convergence of the regularized covariance matrix and demonstrate that significant statistical gains may be expected in terms of the sample complexity. We further explore the consequences of symmetry in related model-selection problems such as the learning of sparse covariance and inverse covariance matrices. We also verify our results with simulations.

**AMS 2000 subject classifications:** Primary 62F12, 62H12.
**Keywords and phrases:** Group invariance, covariance selection, exchangeability, high dimensional asymptotics.

Received November 2011.

## Contents

## 1.  Introduction

An important feature of many modern data analysis problems is the small number of samples available relative to the dimension of the data. Such *high-dimensional* settings arise in a range of applications in bioinformatics, climate studies, and economics. A fundamental problem that arises in the high-dimensional regime is the poor behaviour of sample statistics such as empirical covariance matrices [8, 9]. Accordingly, a fruitful and active research agenda over the last few years has been the development of methods for high-dimensional statistical inference and modeling that take into account *structure* in the underlying model. Some examples of structural assumptions on statistical models include models with a few latent factors (leading to low-rank covariance matrices) [19], models specified by banded or sparse covariance matrices [8, 9], and Markov or graphical models [27, 29, 34].

The focus of this paper is on exploiting *group symmetries* in covariance matrices. Models in which the distribution of a collection of random variables is invariant under certain permutations of the variables have been explored in many diverse areas of statistical research [17, 24]. Symmetry assumptions have played a prominent role within the context of covariance matrices and multivariate Gaussians [1] and these assumptions are of interest in numerous applications (see [24] for a detailed list, and [33] for some historical comments on the subject).

We systematically investigate the statistical and computational benefits of exploiting symmetry in high-dimensional covariance estimation. As a very simple example consider the estimation of the variances of $p$ independent random variables from $n$ samples of the variables. Since the variables are independent, this problem is equivalent to estimating the variances of each variable separately from $n$ samples. Suppose we are additionally given that the variables are identically distributed – we now need to estimate just a single parameter from what are effectively $n \times p$ samples. This very elementary example demonstrates the potential for improvements in the sample complexity as well as computational complexity in models with symmetries. We generalize this basic insight to much more complicated settings in which the distribution underlying a collection of random variables may be symmetric with respect to an arbitrary subgroup of the symmetric group. More formally, we investigate multivariate Gaussians specified

by *invariant covariance matrices*:

$$\Sigma = \Pi_g \Sigma \Pi_g^T \qquad \forall \Pi_g \in \mathfrak{G},$$

where $\mathfrak{G}$ is some subgroup of the symmetric group, i.e., the group of all permutations. Associated to each subgroup $\mathfrak{G}$ of the symmetric group is the *fixed-point subspace* $\mathcal{W}_{\mathfrak{G}}$ of matrices that is invariant with respect to conjugation by each element of $\mathfrak{G}$. This subspace plays a fundamental role via Schur's lemma from representation theory in our analysis of the benefits in sample complexity and computational complexity of regularizing invariant covariance matrices.

The general advantages of symmetry exploitation are numerous

- Problem size: One advantage is that when symmetry is incorporated in the model, the problem size often reduces, and the new instance can have significantly fewer number of variables and constraints, which can lead to dramatic computational speed-ups. This is exploited for example in finite-element methods for partial differential equations [20].
- Better statistical properties: As we will see in this paper, exploiting symmetry also leads to statistical gains in terms of order-of-magnitude gains in the sample complexity of model selection.
- Numerical benefits: Another advantage is the removal of *degeneracies* in the problem that arises from the high multiplicity of eigenvalues that is typically associated to symmetric models. Such multiplicities are a common source of difficulties in numerical methods, and they can be properly handled by suitably exploiting symmetry in the problem. Symmetry-aware methods in general have better numerical conditioning, are more reliable and lead to faster and more accurate solutions.

This paper consists of three main technical contributions. First, we precisely quantify in terms of properties of $\mathfrak{G}$ the improvement in rates of convergence when the sample covariance matrix (obtained from $n$ samples) is projected onto $\mathcal{W}_{\mathfrak{G}}$. Our analysis holds for the spectral or operator norm, the Frobenius norm, and the $\ell_\infty$ norm (maximum absolute entry). Second, we study the implications of these improved rates of convergence by specializing recent results on covariance estimation of Bickel and Levina [8], and of Ravikumar et al. [34] to our setting with invariant covariance matrices. These results quantitatively demonstrate the benefits of taking symmetry structure into account in covariance estimation tasks. Finally, we discuss the computational benefits of incorporating symmetry in covariance regularization. Specifically, we describe how large-dimensional convex programs based on regularized maximum-likelihood can be simplified significantly to smaller-sized convex programs by incorporating prior information about symmetry.

Symmetry and its consequences in the context of statistics, and more broadly in computational mathematics, has been well studied. We mention some related work here. In the Gaussian setting, work has been done related to testing of symmetry in statistical models [32, 40, 42]. In the context of Markov random fields symmetry restrictions have been studied in [2, 6, 7, 25, 28, 41]. We also mention

the work of Højsgaard and Lauritzen [24], who study edge and vertex symmetries for covariance and inverse covariance (graphical models) and exploitation of symmetry for inference problems. They also cite interesting examples from the social sciences (such as distribution of test scores among students and results of psychological tests) where symmetric structure has been experimentally observed. More broadly, symmetry has played an important role in other computational problems such as optimization [16, 23, 26, 39], in the computation of kissing numbers and sphere-packing problems [4], in coding theory [36], in truss topology optimization [5], and in the design of rapidly mixing Markov chains [10].

In the remainder of Section 1 we explain a few specific applications where exploiting symmetry in statistical models is of interest. We also set up some of the notation that will be used throughout the paper. In Section 2 we introduce the necessary group theoretic preliminaries, the notion of $\mathfrak{G}$-invariant covariance matrices, and related statistical quantities. In Section 3 we state some of our main results that quantify the statistical gains when symmetry is exploited judiciously. In Section 4 we extend our results from Section 3 to selecting sparse covariance and inverse covariance models. In Section 5, we make some brief remarks regarding how computational benefits can be gained when selecting symmetric models. In Section 6, we present some simulation results that clarify the statistical gains one may expect when symmetry is exploited. In Section 7, we state some concluding remarks.

## 1.1. Applications

In this section we motivate the problem under consideration by briefly describing a few applications in which estimation of symmetric covariance structure arises naturally.

### 1.1.1. Random fields in physical phenomena

Many signals that capture the spatial variability of natural phenomena are described in the framework of random fields [30]. Random fields arise in a variety of engineering contexts, for example the temperature distribution of materials, the circulation of ocean fields in oceanography, or the transport of groundwater in hydrology [30]. A classical model is associated with the well-known Poisson's equation:

$$\nabla^2 \phi(x) = f(x) \qquad x \in \mathbb{R}^d \tag{1}$$

where $\nabla^2$ is the Laplace operator. In hydrology, this equation governs the steady state flow of a fluid in $\mathbb{R}^3$, whereas in electrostatics it governs the distribution of the potential in a variety of settings such as transmission lines [30]. Under the right technical conditions, if $f(x)$ is a white noise process, the stochastic process $\phi(x)$ in (1) is a Gauss-Markov random field [35]. While Poisson's equation (1) is a *sample-path* representation of the process $\phi(x)$, a particularly useful representation of this process is the covariance $R(x_1, x_2)$ associated to this random field.

Indeed the covariance $R(x_1, x_2)$ is the Green's function of the biharmonic equation, and its properties have been studied in detail by Moura and Goswami [30].

A useful observation that is pertinent in our context is that the Laplacian as well as the biharmonic equation are invariant under the action of the orthogonal group, i.e. they are isotropic. If in addition the forcing process $f(x)$ has isotropic stochastic properties (one such example being a white noise process) and the boundary conditions of the physical process are also isotropic, then the covariance $R(x_1, x_2)$ will also be isotropic.

When such random fields are encountered in practice, the usual engineering approach is to discretize the domain so as to make the problem computationally tractable. The restriction of the covariance $R(x_1, x_2)$ to the finite discretization of the domain results is a covariance matrix $\hat{R}$. Recognizing the symmetry in the problem, it is often useful to construct a *symmetry-preserving discretization* of the space [31]. Such a modeling paradigm naturally leads to a covariance matrix that is invariant with respect to a discrete group $\mathfrak{G}$ which is a finite subgroup of the orthogonal group. A motivating question in this paper then is:

"Given samples of the stochastic process $\phi(x)$ on this discretized domain, can one exploit the symmetry of the space to obtain a good statistical approximation of the covariance matrix $\hat{R}$?"

### 1.1.2. Bayesian models

A fundamental notion in Bayesian statistics is that of exchangeability. Indeed, in such settings one often deals with a collection of random variables $X_1, \ldots, X_n$ where it is not appropriate to assume that this collection is *i.i.d.*. It is far more natural to assume instead that the collection of random variables is *exchangeable* or *partially exchangeable*. We give two such examples:

1. **Urn models:** Consider an urn with $n_r$ red balls and $n_b$ blue balls. A quantity $T$ balls are randomly drawn *without replacement* from the urn. Let us define the random variable:

$$X_i = \left\{ \begin{array}{ll} 0 & \text{if the } i^{th} \text{ ball is red} \\ 1 & \text{if the } i^{th} \text{ ball is blue.} \end{array} \right.$$

Clearly the random variables $X_1, \ldots, X_T$ are *not i.i.d.* On the other hand, it is well-known that this sequence of random variables is exchangeable, i.e. for any permutation $\sigma$ on the indices $1, \ldots, T$ and for any set $A$

$$\mathbb{P}\left((X_1, \ldots, X_T) \in A\right) = \mathbb{P}\left((X_{\sigma(1)}, \ldots, X_{\sigma(T)}) \in A\right).$$

In the terminology developed in this paper, this distribution is invariant with respect to the group $\mathfrak{G}$ corresponding to the *symmetric group*. Such urn models are commonly used metaphors for problems related to computing statistical properties of a population of objects.

Suppose the number of red and blue balls in the urn is unknown, and one wants to estimate the probability of drawing two consecutive blue balls in

the first two draws empirically from multiple trials of $T$ draws each. It is an easy exercise to check that:

$$\mathbb{E}\left[X_1 X_2\right] = \mathbb{P}\left(X_1 = 1, X_2 = 1\right).$$

By exchangeability $\mathbb{E}\left[X_1 X_2\right] = \mathbb{E}\left[X_i X_j\right]$ for any $i, j \in \{1, \ldots, T\}$. Hence the computation of the required probability reduces to estimating the empirical covariance matrix. It is important to exploit the underlying symmetry in this computation.

2. **Clinical tests:** Suppose a clinical test is conducted on a set of $T$ patients wherein a drug is administered to the patients and a physiological response $x_1, x_2, \ldots, x_T$ is measured for each patient. The set of patients are divided into groups $T_1, \ldots, T_h$ of patients of similar physiological properties (say based on characteristics such as combinations of similar age, ethnicity, sex, and body size). It is often reasonable to assume that the distributions of the responses $X_i$ of patients within the same class are exchangeable (but not *i.i.d.*). Sources of non-independence include latent factors such as residual effects of a previous experiment, physical effects such as the temperature, etc. Thus, the distribution of $(X_1, \ldots, X_n)$ may be assumed to be invariant with respect to permutation of patient labels within the same class $T_i$. Specifically, if $\text{Sym}_k$ denotes the symmetric group of permutations on $k$ indices then the the distribution of $(X_1, \ldots, X_n)$ is invariant under the action of the group $\mathfrak{G} = \text{Sym}_{|T_1|} \times \ldots \times \text{Sym}_{|T_h|}$. In such clinical tests, it is of importance to estimate the covariance matrix of the responses from a limited number of trials.

### 1.1.3. Cyclostationary processes

An interesting special case of random phenomena with group symmetries are the class of well-studied stochastic processes known as *cyclostationary processes* [22]. These processes arise in random phenomena that obey periodic stochastic properties, for example the yearly variation of temperature in a particular location. Such processes are frequently encountered in engineering and arise in a variety of contexts such as communication systems (presence of a sinusoidal component), radar (rotating elements), and the study of vibrations in mechanical engineering. Moreover, a special case of such processes are the classical stationary processes, which are ubiquitous in engineering.

Formally, a discrete time stochastic process $\{X_t\}_{t \in \mathbb{Z}}$ is said to be cyclostationary with period $N$ if for any measurable set $A$

$$\mathbb{P}\left((X_{t_1}, \ldots X_{t_k}) \in A\right) = \mathbb{P}\left(X_{t_1+N}, \ldots X_{t_k+N}) \in A\right) \quad \text{for all } t_1, \ldots, t_k \in \mathbb{Z}.$$

Thus the joint probability distribution of such a process is invariant under shifts of $N$ time units. In the particular instance when $N = 1$ we obtain a *stationary* process. An important statistical property associated to this process is the *auto-correlation*

$$R(t, \tau) := \mathbb{E}\left[X_t X_\tau\right].$$

If we restrict attention to the first $kN$ time units for some integers $k$, it is clear that the restriction of the autocorrelation admits a finite representation as a matrix $R \in \mathbb{R}^{kN \times kN}$. In the stationary case (i.e. $N = 1$), from shift invariance it is also apparent that $R(t, \tau)$ depends *only* on $|t - \tau|$, so that the resulting covariance matrix is *Toeplitz*. A commonplace and particularly useful approximation for large Toeplitz matrices is to convert them to *circulant* matrices [38]. In the more general cyclostationary case (the period $N > 1$), the matrix $R$ is block Toeplitz and approximated via a *block circulant* one where each block is of size $N \times N$.

In group-theoretic terms, the covariance matrix $R$ is invariant under the action of the cyclic group of order $k$ (the group captures cyclic translations of the indices). This group is abelian, and its action on the index set is precisely the cyclic shift operator by $N$ units. As is well-known from the representation theory of the cyclic group, the *Fourier transform* (block) diagonalizes circulant matrices, and the resulting (block) diagonal matrix is called the spectrum of the autocorrelation. These observations are classical results and we mention them here to point out an interesting special case of our framework.

### 1.1.4. MAR processes

Many physical phenomena exhibit multi-scale behavior. In these models, there is a natural notion of scale, and the process exhibits interesting behavior at multiple scales. Examples of such processes abound in the signal processing literature [43], and include models in oceanography, surface reconstruction and texture discrimination in computer vision and image segmentation and denoising in image processing.

An important class of models that have been studied in this context are multi-scale autoregressive (MAR) processes [43]. These can be represented as state space models driven by white Gaussian noise, where the states are indexed by a tree and the state variables are related by affine dynamics operating from the root to the leaves of the tree.

For concreteness we will study MAR processes evolving on full binary trees. Consider a full binary tree $\mathfrak{T}_d = (V, E)$ of depth $d$ with $|V| = 2^{d+1} - 1$ nodes. Let $L \subset V$ denote the leaves of $\mathfrak{T}_d$ and $r \in V$ denote the root. For a node $v \in V$ that is not the root, let $\uparrow v$ denote the parent of $v$. Let $d(v)$ denote the depth of the vertex $v$ (i.e. the distance from the root). The state vector $x \in \mathbb{R}^{|V|}$ is indexed by the tree, so that $x_v$ denotes the state variable indexed by the node $v \in V$. The root node is distributed as $x_r \sim \mathcal{N}(0, 1)$ and the remaining states evolve according to the following recursion:

$$x_v = A_{d(v)} x_{\uparrow v} + w_v, \tag{2}$$

where $w_v \sim \mathcal{N}(0, 1)$ is *i.i.d.* Gaussian noise. As a consequence of this recursive relationship between the random variables, the joint distribution of the state $x$ is Gaussian and posseses *Markov* structure with respect to the tree $\mathfrak{T}_d$ [43].

In many processes of interest, only the states indexed by the leaf nodes are observable, the other nodes of the tree are latent. Let $x_L \in \mathbb{R}^{|L|}$ denote the random vector corresponding to the leaf nodes. Let $\Sigma \in \mathbb{R}^{|L| \times |L|}$ be its covariance matrix (indexed by $L$ also). In (2), the dynamics depend only on the scale (i.e. the dynamical evolution of $x_v$ from $x_{\uparrow v}$ depends only on the scale $d(v)$ and white noise), the statistical properties of $x_L$ (in particular $\Sigma$) will inherit the symmetries of the graph $\mathfrak{T}_d$.

Note that the tree $\mathfrak{T}_d$ has a non-trivial automorphism group, and is in fact related to the wreath product, see Example 2.1 for more details. An interesting question in the context of MAR processes is, given a few samples of $x$, how can one exploit the symmetries to reliably construct the covariance matrix $\Sigma$?

### 1.2. Notation

We describe some of the notation that will be adopted through the rest of this paper. We will commonly deal with a random vector $X \in \mathbb{R}^p$, where $p$ is the dimension of the space. Occasionally, we will encounter the complex vector space $\mathbb{C}^p$. We will be interested in inferring statistical properties of this random vector from *i.i.d.* samples $X_1, \ldots, X_n$ where $n$ is the number of samples. We will study the problem in a *high-dimensional* setting, i.e. where $p \gg n$ so that the dimension of the space is much larger than the number of samples.

We will mostly consider *normally distributed* random vectors $X$ with zero mean and covariance $\Sigma$, we will denote this by $\mathcal{N}(0, \Sigma)$. The inverse of the covariance is called the *concentration matrix* and we use the notation $\Theta := \Sigma^{-1}$. The probability of an event $A$ will be denoted by $\mathbb{P}(A)$ and the expectation of a random variable $Y$ is denoted by $\mathbb{E}[Y]$. The indicator function of an event $A$ is denoted by $\mathbb{1}(A)$. If $h(X)$ is a real valued function of a $\mathbb{R}^p$-valued random vector, we will use the notation that $h(X) = O_P(f(p, n))$ to mean that $h(X) \leq Cf(p, n)$ with probability exceeding $1 - \frac{1}{p^\tau}$ for some constants $C$ and $\tau > 0$.

Given a finite group $\mathfrak{G}$ which is a subgroup of the symmetric group, we consider the *group action* of $\mathfrak{G}$ on the index set $\{1, \ldots, p\}$. Specifically, an element $g \in \mathfrak{G}$ can be interpreted as a permutation, and $g$ acts on $\{1, \ldots, p\}$ by permuting these indices. To every $g \in \mathfrak{G}$ one can associate a permutation matrix $\Pi_g \in \mathbb{R}^p$.

We will denote the positive-definiteness (respectively positive semidefiniteness) of a symmetric matrix $M$ by $M \succ 0$ ($M \succeq 0$). The cone of symmetric positive definite matrices in $\mathbb{R}^{p \times p}$ will be denoted by $\mathcal{S}_{++}^p$. The maximum eigenvalue of a symmetric matrix $M$ is denoted by $\lambda_{\max}(M)$. Its *trace* (sum of the diagonal entries) will be denoted by $\text{Tr}(M)$. We will encounter different norms on the space of symmetric matrices:

1. Frobenius norm: Denoted by $\|M\|_F = \left(\text{Tr}(M^T M)\right)^{\frac{1}{2}}$.
2. Spectral norm: Denoted by $\|M\|$ is the maximum singular value of $M$.
3. Elementwise $\ell_\infty$ norm: Denoted by $\|M\|_\infty = \max_{i,j} |M_{ij}|$.
4. The operator $\infty$ norm: Denoted by $\|M\|_{\infty,\infty} = \max_i \sum_j |M_{ij}|$.
5. Off-diagonal $\ell_1$ norm: Denoted by $\|M\|_{1,\text{off}} = \sum_{i,j,i \neq j} |M_{ij}|$.

## 2. Symmetric models

### *2.1. Symmetry, groups, and group representations*

The central concepts that formalize the notion of symmetry mathematically are *groups* and *invariance*. Abstractly, a physical object or process (in our case a covariance matrix) is said to possess symmetry if it is invariant under the action of a group. For background on groups and group actions we refer the reader to any standard algebra textbooks, for example [3, 18, 37]. In this paper we will be dealing with finite groups $\mathfrak{G}$, i.e. groups whose cardinality $|\mathfrak{G}|$ is finite. A basic result in group theory, Cayley's theorem, establishes that every finite group is isomorphic to a subgroup of the symmetric group (i.e. the group of permutations with the group operation being permutation composition). Consequently, elements of finite groups $\mathfrak{G}$ can be conveniently represented via permutation matrices, and in our context these naturally induce an action of $\mathfrak{G}$ on matrices $M \in \mathbb{R}^{p \times p}$ via conjugation.

Specifically, letting $\Pi_g$ denote the permutation matrix corresponding to $g \in \mathfrak{G}$, a *group action* is a map $\varphi$:

$$\varphi : \mathfrak{G} \times R^{p \times p} \to \mathbb{R}^{p \times p}$$
$$(g, M) \mapsto \Pi_g M \Pi_g^T.$$

A matrix $M$ is said to be *invariant* with respect to the group if $M = \Pi_g M \Pi_g^T$ for all $g \in \mathfrak{G}$. The set

$$\mathcal{W}_\mathfrak{G} = \left\{ X \in \mathbb{R}^{p \times p} \mid \Pi_g^T X \Pi_g = X \ \forall \Pi_g \in \mathfrak{G} \right\}$$

is called the *fixed point subspace* of the group $\mathfrak{G}$ and it is the set of all matrices that are invariant with respect to the group. Informally, this is the set of matrices that possess "symmetry" with respect to the group. We let $\mathcal{P}_\mathfrak{G}(\cdot) : \mathbb{R}^{p \times p} \to \mathcal{W}_\mathfrak{G}$ denote the Euclidean projection operator that projects a matrix onto the fixed point subspace.

A fundamental tool in the study of invariance is the theory of group representations. We provide a brief review of concepts from group representation theory which are relevant for our discussion. Let $GL(\mathbb{C}^N)$ denote the general linear group of invertible matrices in $\mathbb{C}^{N \times N}$. A *group representation* is a group homomorphism

$$\rho : \mathfrak{G} \to GL(\mathbb{C}^N)$$
$$g \mapsto \rho(g).$$

The representation naturally induces an action of $\mathfrak{G}$ onto the vector space $\mathbb{C}^N$ via

$$\rho : \mathfrak{G} \times \mathbb{C}^N \to \mathbb{C}^N$$
$$(g, v) \mapsto \rho(g)v.$$

Thus, the representation naturally associates to each $g \in \mathfrak{G}$ a linear transformation $\rho(g)$ that acts on $\mathbb{C}^N$. Since a linear transformation is characterized by

its invariant subspaces, it is sufficient to study the invariant subspaces of the linear transformations $\rho(g)$ for $g \in \mathfrak{G}$. Indeed, if $\rho$ is a given representation and $W \subseteq \mathbb{C}^N$ is a subspace, we say that $W$ is $\mathfrak{G}$-*invariant* with respect to $\rho$ if $\rho(g)W \subseteq W$ for every $g \in \mathfrak{G}$. We say that a representation $\rho$ is *irreducible* if the only $\mathfrak{G}$-invariant subspaces are the trivial subspaces $0$ and $\mathbb{C}^N$. *Schur's lemma*, a classical result in group representation theory, establishes that finite groups have only finitely many irreducible representations, and provides the following orthogonal decomposition theorem of $\mathbb{C}^N$ in terms of the $\mathfrak{G}$-invariant subspaces.

**Lemma 2.1** (Schur's lemma [37])**.** *For a finite group $\mathfrak{G}$ there are only finitely many inequivalent irreducible representations (indexed by $\mathcal{I}$) $\vartheta_1, \ldots, \vartheta_{|\mathcal{I}|}$ of dimensions $s_1, \ldots, s_{|\mathcal{I}|}$. The dimensions $s_i$ divide the group order $|\mathfrak{G}|$, and satisfy $\sum_{i \in \mathcal{I}} s_i^2 = |\mathfrak{G}|$. Every linear representation of $\mathfrak{G}$ has a canonical decomposition*

$$\rho = m_1 \vartheta_1 \oplus \cdots \oplus m_{|\mathcal{I}|} \vartheta_{|\mathcal{I}|},$$

*where the $m_i$ are the multiplicities. Correspondingly, there is an* isotypic *decomposition of $\mathbb{C}^N$ into invariant subspaces $W_i$:*

$$\mathbb{C}^N = W_1 \oplus \cdots \oplus W_{|\mathcal{I}|}$$

*where each $W_i$ is again a direct sum of isomorphic copies $W_i = W_{i1} \oplus \cdots \oplus W_{i,m_i}$ with multiplicity $m_i$.* [1]

A basis of this decomposition (that depends only on the group) transforming with respect to the matrices $\vartheta(g)$ is called symmetry adapted, and can be explicitly computed algorithmically [20, 37]. This basis defines a change of coordinates by a unitary matrix $T \in \mathbb{C}^{N \times N}$. One of the main consequences of Schur's lemma is that the symmetry adapted basis *block diagonalizes the fixed point subspace*, i.e. every matrix in the fixed point subspace is commonly diagonalized by $T$. If $M \in \mathcal{W}_{\mathfrak{G}}$ is a matrix in the fixed point subspace, then changing coordinates with respect to $T$ decomposes $M$ into a block diagonal form as follows:

$$T^* M T = \begin{bmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_{|\mathcal{I}|} \end{bmatrix} \qquad M_i = \begin{bmatrix} B_i & & 0 \\ & \ddots & \\ 0 & & B_i \end{bmatrix}. \qquad (3)$$

In the above decomposition the diagonal blocks $M_i \in \mathbb{R}^{m_i s_i \times m_i s_i}$ can be further decomposed into $m_i$ repeated diagonal blocks $B_i \in \mathbb{R}^{s_i \times s_i}$ (recall that $s_i$ are the sizes of the irreducible representations and $m_i$ are the multiplicities). Thus, the symmetry restriction $M \in \mathcal{W}_{\mathfrak{G}}$ reduces the degrees of freedom in the problem of interest. This observation plays a central role in our paper.

**Example 2.1.** To fix ideas, we will consider the following three examples to illustrate our results throughout the paper.

---

[1] Schur's lemma, as stated classically, provides a decomposition over the complex field $\mathbb{C}^N$. However, a real version can be adapted in a straightforward manner [37, pp. 106-109],[23] and the irreducible real representations are called *absolutely irreducible*.

1. **Cyclic group:** The cyclic group of order $p$ is isomorhpic to $\mathbb{Z}/p\mathbb{Z}$ (the group operation is the addition operation $+$). Let $M \in \mathbb{R}^{p \times p}$ be a circulant matrix of the form:

$$M = \begin{bmatrix} M_1 & M_2 & \dots & M_p \\ M_p & M_1 & \dots & M_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ M_2 & M_3 & \dots & M_1 \end{bmatrix}. \tag{4}$$

   If $\rho(k)$ is a permutation matrix corresponding to cyclic shifts of size $k$ of the indices $\{1, \dots, p\}$, then it is clear that $\rho(k)M\rho(k)^T = M$. Indeed the fixed point subspace is precisely the set of circulant matrices and since the group is abelian, can be described as

$$\mathcal{W}_{\mathfrak{G}} = \left\{ M \in \mathbb{R}^{p \times p} \mid M\rho(1) = \rho(1)M \right\}.$$

   (Since 1 is the generator of the cyclic group, invariance with respect to 1 implies invariance with respect to the other elements).
   It is well known that if $T = \mathcal{F} \in \mathbb{C}^{p \times p}$ is the Fourier transform matrix, then $T^*MT$ is block diagonal. (The columns of $\mathcal{F}$ form the symmetry adapted basis). Here $|\mathcal{I}| = p$ (there are $p$ irreducible representations), $m_i = 1$ and $s_i = 1$ for all $i \in \mathcal{I}$. In our setting, covariance matrices of such structure arise from cyclostationary processes.

2. **Symmetric group:** Let us now consider $\mathfrak{G} = \mathrm{Sym}_p$, the full symmetric group of permutations on $p$ indices. Let $M \in \mathbb{R}^{p \times p}$ be have the following structure:

$$M = \begin{bmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{bmatrix}. \tag{5}$$

   It is straightforward to check that $\Pi_g M = M \Pi_g$ for every permutation matrix $\Pi_g$, matrices satisfying this property constitute the fixed point subspace. It is well known that via the *permutation representation* there is an orthogonal matrix [2] $T$ such that:

$$T^*MT = M = \left[ \begin{array}{c|ccc} c_1 & 0 & \dots & 0 \\ \hline 0 & c_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & c_2 \end{array} \right].$$

   Here $|\mathcal{I}| = 2, s_1 = 1, m_1 = 1, s_2 = 1, m_2 = p-1$. In our setting, covariance matrices with this structure arise when we have a completely exchangeable sequence of random variables.

---

[2] In fact any orthogonal matrix $T$ whose first column is a multiple of the all ones vector, with the remaining columns span the orthogonal subspace will achieve the desired diagonalization.
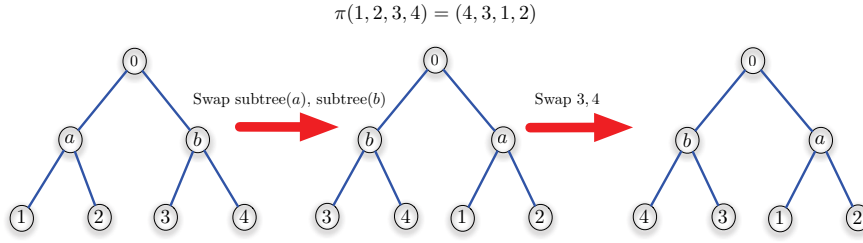
$$\pi(1,2,3,4) = (4,3,1,2)$$



FIG 1. *A permutation element $\pi \in \mathfrak{G}_2$ (where $\mathfrak{G}_2 = \mathbb{Z}_2 \wr \mathbb{Z}_2$) is generated recursively. First the subtrees rooted at a and b respectively are permuted. Then on the left subtree the exchange permutation is applied to the leaves, while the identity permutation is applied to the right subtree.*

3. **Wreath product of $\mathbb{Z}_2$:** Recall the notion of an MAR process described in Section 1.1.4. In the estimation of the covariance matrix of the leaf nodes $L$ of the process, it is evident that there are a number of symmetries that can be exploited. We note that there is a natural notion of *scale* in the tree $\mathfrak{T}_d$, with the root node at depth $0$ and the leaves at depth $d$. We begin by describing the automorphism group of $\mathfrak{T}_d$. A *valid* permutation at a node of $\mathfrak{T}_d$ is either the identity permutation or one that swaps the left and right subtrees rooted at that node. More generally, one obtains an element of the automorphism group of $\mathfrak{T}_d$ as follows: One first applies a valid permutation at the root node (depth $0$), followed by valid permutations at the nodes at depth $1$, and so on recursively (see Fig 1). Formally, the automorphism group of $\mathfrak{T}_d$ is equal to the $d$-fold iterated *wreath product*, $\wr^d \mathbb{Z}_2$, of the cyclic group of cardinality $2$ [21] (In this paper we only describe the case of wreath products of $\mathbb{Z}_2$ for the sake of simplicity. Our results can be extended to more general wreath products).

   The covariance $\Sigma \in \mathbb{R}^{2^d \times 2^d}$ indexed by the leaves $L$ inherits the symmetries of the tree $\mathfrak{T}_d$ as the coarse-to-fine dynamics (2) are uniform across each scale. In fact one can check that the automorphism group of $\mathfrak{T}_d$ is in one-to-one correspondence with the invariance group $\mathfrak{G}_d$ of $\Sigma$. A symmetry-adapted basis $T$ that diagonalizes $\Sigma$ is given by the Haar wavelet transformation [21]. One can check that $T^* \Sigma T$ has $d+1$ distinct eigenvalues, one corresponding to each scale from depth $0$ to depth $d$ (see Fig. 2). The eigenvalue associated to depth $k$ has multiplicity $2^k$ for $k = 0, \ldots, d-1$, and the eigenvalue associated to depth $d$ has multiplicity $1$.

4. **Product of symmetric groups:** Let $T_1, \ldots, T_k$ be finite index sets where $|T_i| = r_i$. We consider the product of symmetric groups $\mathrm{Sym}_{|T_1|} \times \cdots \times \mathrm{Sym}_{|T_k|}$ acting on the index set $T_1 \times \cdots \times T_k$. Consider a block $k \times k$ matrix of the form:

$$M = \begin{bmatrix} M_{11} & M_{12} & \ldots & M_{1k} \\ M_{12}^T & M_{22} & \ldots & M_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1k}^T & M_{2k}^T & \ldots & M_{kk} \end{bmatrix}, \tag{6}$$

where for each $j \in \{1,\ldots,k\}$, $M_{jj} \in \mathbb{R}^{r_j \times r_j}$ is of the form (5), and each $M_{ij}$ for $i \neq j$ is a constant multiple of the all ones matrix. Let $\sigma = (\sigma_1,\ldots,\sigma_k) \in \mathrm{Sym}_{|T_1|} \times \ldots \times \mathrm{Sym}_{|T_k|}$ be a permutation with this product structure. Let $\Pi_\sigma$ be its corresponding permutation matrix. Then it is clear that for a matrix of the form described above, $\Pi_\sigma M = M\Pi_\sigma$, and the set of matrices satisfying this relation constitute the corresponding fixed point subspace $\mathcal{W}_\mathfrak{G}$. Using standard representation theory results, it is possible to show that there is an orthogonal matrix $T$ such that

$$T^* M T = \begin{bmatrix} C & 0 & \ldots & 0 \\ 0 & c_1 I_{r_1-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & c_k I_{r_k-1} \end{bmatrix}.$$

where $C \in \mathbb{R}^{k \times k}$ and $I_r \in \mathbb{R}^{r \times r}$ is the identity matrix of size $r$. (In fact, upto a permutation of the columns, $T$ can be chosen to be block diagonal where each diagonal block is of the form described in the preceding example). In the above example, $|\mathcal{I}| = k + 1$, $s_1 = k$, $m_1 = 1$. For $i \in \{2,\ldots,k+1\}$, $s_i = 1$, $m_i = r_{i-1} - 1$. Covariance matrices with product symmetric group structure arise when one has a set of random variables that are *partially* exchangeable.

Covariance matrices of interest in this paper will be invariant with respect to a group $\mathfrak{G}$, and hence belong to the fixed point subspace $\mathcal{W}_\mathfrak{G}$. A fundamental operation in our paper will be the Euclidean projection of a given matrix $M$ onto $\mathcal{W}_\mathfrak{G}$. The next lemma relates projection to any important notion called *Reynolds averaging.* Reynolds averaging operates by generating the orbit of $M$ with respect to the group, i.e. the collection of matrices $\Pi_g M \Pi_g^T$, and averaging with respect to the orbit.

**Lemma 2.2.** *Let $\mathcal{P}_\mathfrak{G}(\cdot)$ be the Euclidean projection operator onto the fixed point subspace $\mathcal{W}_\mathfrak{G}$. Then*

$$\mathcal{P}_\mathfrak{G}(M) = \frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} \Pi_g M \Pi_g^T.$$

*Proof.* Since $\mathcal{W}_\mathfrak{G}$ is a subspace, it suffices to show that $\mathrm{Tr}((M - \mathcal{P}_\mathfrak{G}(M)) \times \mathcal{P}_\mathfrak{G}(M)) = 0$ (by the orthogonality principle and that fact that $\langle A, B \rangle = \mathrm{Tr}(AB)$ is the inner product that generates the Euclidean norm over the space of symmetric matrices). We obtain the proof via the following chain of equalities:

$$\mathrm{Tr}\left(\mathcal{P}_\mathfrak{G}(M)\mathcal{P}_\mathfrak{G}(M)\right) = \mathrm{Tr}\left(\frac{1}{|\mathfrak{G}|}\sum_{h \in \mathfrak{G}}\Pi_h M \Pi_h^T \frac{1}{|\mathfrak{G}|}\sum_{g \in \mathfrak{G}}\Pi_g M \Pi_g^T\right)$$

$$= \frac{1}{|\mathfrak{G}|^2}\sum_{h \in \mathfrak{G}}\mathrm{Tr}\left(\Pi_h M \Pi_h^T \sum_{g \in \mathfrak{G}}\Pi_g M \Pi_g^T\right)$$

$$= \frac{1}{|\mathfrak{G}|^2} \sum_{h \in \mathfrak{G}} \mathrm{Tr} \left( M \sum_{g \in \mathfrak{G}} \Pi_h^T \Pi_g M \Pi_g^T \Pi_h \right)$$

$$= \frac{1}{|\mathfrak{G}|} \sum_{h \in \mathfrak{G}} \mathrm{Tr} \left( M \mathcal{P}_{\mathfrak{G}}(M) \right)$$

$$= \mathrm{Tr} \left( M \mathcal{P}_{\mathfrak{G}}(M) \right).$$

$\square$

**Remark 2.1.** Note that the fixed point subspace, when suitably transformed, is a set of block diagonal matrices of the form (3) as a consequence of Schur's lemma. Projection of a matrix $M$ onto the fixed point subspace in these coordinates corresponds precisely to *zeroing* components of $M$ that are off the block diagonals, following by *averaging* the diagonal blocks corresponding to multiplicities of the same irreducible representation.

## 2.2. Group theoretic parameters of interest

Given a covariance matrix $\Sigma$ which is invariant with respect to the action of a group $\mathfrak{G}$, we introduce the following parameters that depend on $\Sigma$ and $\mathfrak{G}$. Our subsequent results characterizing the rates of convergence of the estimated covariance to the true covariance will be expressed in terms of these group theoretic parameters.

Given $\Sigma$, it will often be convenient to associate informally a weighted undirected complete graph to it, so that the entries of $\Sigma$ correspond to edge weights. We will view the rows and columns of $\Sigma$ as being indexed by a set $V$ so that $\Sigma \in \mathbb{R}^{|V| \times |V|}$. The group $\mathfrak{G}$ is then precisely the *automorphism group* of this weighted graph. Let $(i, j) \in V \times V$ (we will call this an "edge"). Define the *edge orbit* as

$$\mathcal{O}(i, j) = \{(g(i), g(j)) \mid g \in \mathfrak{G}\}.[3]$$

**Example 2.2.** Consider an MAR process given by (2), so that its corresponding covariance matrix $\Sigma$ indexed by the leaves $L$ is $\mathfrak{G}_d$-invariant. The tuples $(i, j), (k, l)$ lie in the same edge orbit if the lowest common ancestor of $i, j$ and that of $k, l$ are at the same depth in the tree $T_d$. Consequently, there are $d + 1$ distinct edge orbits. More precisely the orbits are given by

$$\mathrm{Orb}_k = \{(i, j) : i \text{ and } j \text{ have a common ancestor at depth } k\} \quad \text{for } k = 0, \ldots, d$$

(see Fig. 2).

We further define $\mathcal{V}(i, j)$ to be the set of nodes that appear in the edge orbit $\mathcal{O}(i, j)$, i.e.

$$\mathcal{V}(i, j) = \{k \in V \mid \exists l \in V, g \in \mathfrak{G} \text{ such that } (g(k), g(l)) \in \mathcal{O}(i, j)\}.$$

---

[3]The pairs $(g(i), g(j))$ are to be viewed as unordered. Also, each edge is enumerated only once so that if $(g(i), g(j)) = (i, j)$ then we treat them as the same edge in $\mathcal{O}(i, j)$.
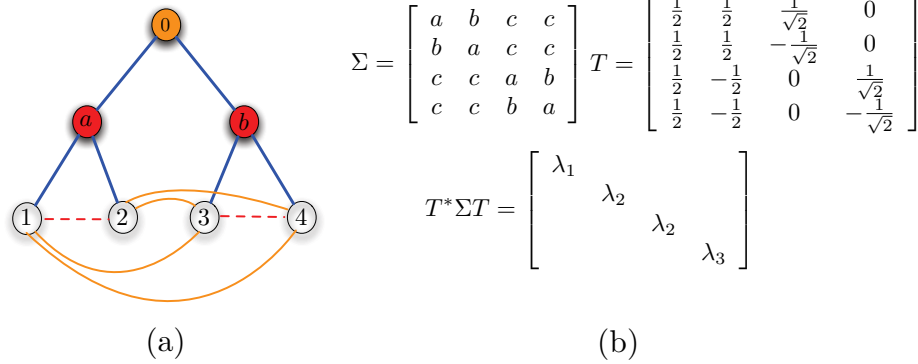
$$\Sigma = \begin{bmatrix} a & b & c & c \\ b & a & c & c \\ c & c & a & b \\ c & c & b & a \end{bmatrix} \quad T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$T^*\Sigma T = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \lambda_2 & \\ & & & \lambda_3 \end{bmatrix}$$

(a)                                                         (b)

FIG 2. *(a) The tree $\mathfrak{T}$ along with the edge orbits. There are three edge orbits: one correspond-*
*ing to the diagonal entries of $\Sigma$ (not shown), one corresponding to pairs $(i,j)$ with lowest*
*common ancestor at depth $1$ (shown in red) and one corresponding to pairs $(i,j)$ with lowest*
*common ancestor at depth $0$ (shown in orange). (b) The matrix $\Sigma$ is $\mathfrak{G}_2$-invariant, and can*
*be diagonalized by the Haar wavelet transformation $T$.*

We define $d_{ij}$ (called the degree of the orbit $\mathcal{O}(i,j)$) to be the maximum number of times any node appears in the edge orbit $\mathcal{O}(i,j)$, i.e.

$$d_{ij}(k) = \sum_{l \in \mathcal{V}(i,j)} \mathbb{1}\left(\exists l \in V \mid (k,l) \in \mathcal{O}(i,j)\right),$$

and $d_{ij} = \max_{k \in \mathcal{V}(i,j)} d_{ij}(k)$.

**Definition 2.1.** Let $M \in \mathcal{W}_{\mathfrak{G}}$ be a matrix in the fixed point subspace. We define the *orbit parameter* $\mathcal{O}$ to be the size of the smallest orbit (with respect to the action of $\mathfrak{G}$ on M), i.e.

$$\mathcal{O} = \min_{i,j} |\mathcal{O}(i,j)|$$

The *normalized orbit parameter* $\mathcal{O}_d$ is defined to be

$$\mathcal{O}_d = \min_{i,j} \frac{|\mathcal{O}(i,j)|}{d_{ij}}.$$

**Example 2.3.** For a covariance matrix of the form (4) shown in Example 2.1,

$$\mathcal{O}(1,2) = \{(1,2),(2,3),\ldots,(p-1,p),(p,1)\}$$
$$\mathcal{V}(1,2) = \{1,\ldots,p\}$$
$$d_{12} = 2 \qquad \text{(since each node appears twice in } \mathcal{O}(1,2)\text{)}.$$

For this example $\mathcal{O} = p$ and $\mathcal{O}_d = \frac{p}{2}$.

**Example 2.4.** For a covariance matrix of the form (5) with the action of the symmetric group of order $p$, it is easy to check that $\mathcal{O} = p$ (there are only two orbits, the first is $\mathcal{O}(1,1)$ which is of size $p$ and degree 1, and the other is $\mathcal{O}(1,2)$ of size $\frac{p(p-1)}{2}$ and degree $p - 1$. Hence $\mathcal{O} = p$ and $\mathcal{O}_d = \frac{p}{2}$.

**Example 2.5.** For a covariance matrix of the from (6) with the action of the product of symmetric groups, if the sizes $|T_i| = r_i$ for $i = 1, \ldots, k$ then it is clear that $\mathcal{O} = \min_i r_i$ and $\mathcal{O}_d = \min_i \frac{r_i}{2}$.

**Example 2.6.** For a covariance matrix $\Sigma$ associated to the MAR process (2) a straightforward counting argument shows that

$$|\mathrm{Orb}_k| = 2^{d-1} 2^{d-k-1} = \frac{p^2}{2^{k+2}} \text{ for } k = 0, \ldots, d-1$$

and $|\mathrm{Orb}_d| = p$ (corresponding to the diagonal entries of $\Sigma$). Hence $\mathcal{O} = p$.

Moreover in the orbit $\mathrm{Orb}_k$ each vertex appears $2^{d-k-1}$ times for $k = 0, \ldots, d-1$, hence the normalized orbit parameter is

$$\mathcal{O}_d = \frac{p}{2}.$$

In addition to these combinatorial parameters, we will also be interested in the following *representation-theoretic* parameters (which are also intrinsic to the group and the fixed point subspace $\mathcal{W}_{\mathfrak{G}}$). Given $\Sigma$ and its block diagonalization (3) as described by Schur's lemma, we will be interested in the parameters $s_i$ (the sizes of the blocks) and $m_i$ (the multiplicities) for $i \in \mathcal{I}$. (We note that given a fixed point subspace $\mathcal{W}_{\mathfrak{G}}$, multiplicities corresponding to many irreducible components may be zero, we call these the inactive components. We will only be interested in the active components, i.e. the ones with non-zero multiplicity. Thus the set of irreducible representations $\mathcal{I}$ mentioned henceforward will include only active representations).

### 2.3. Symmetric covariance models

Let $X \sim \mathcal{N}(0, \Sigma)$ be a normally distributed $\mathbb{R}^p$-valued random vector where the covariance $\Sigma$ is unknown. Given a small number of samples $X^{(1)}, \ldots, X^{(n)}$, we wish to learn properties of the model (for example, the covariance matrix $\Sigma$, the inverse covariance $\Theta := \Sigma^{-1}$, the sparsity pattern of $\Sigma$ and $\Theta$). We will be interested in the *high-dimensional setting* where $p$ is large and $n \ll p$, so that the number of samples is much smaller than the ambient dimension of the problem.

Throughout the paper we will make the two following assumptions regarding our model:

**Assumption 2.1.**

1. *We assume that $\Sigma$ is* invariant *with respect to a known group $\mathfrak{G}$ so that $\Sigma \in \mathcal{W}_{\mathfrak{G}}$.*

2. *We assume that the spectral norm of the true covariance matrix is bounded independent of the model size p, i.e.*

$$\|\Sigma\| \leq \psi$$

*for some constant $\psi$.*

**Remark 2.2.** A simple but important consequence of the $\mathfrak{G}$-invariance of $\Sigma$ (i.e. $\Pi_g \Sigma \Pi_g^T = \Sigma$ for all $g \in \mathfrak{G}$) is that $\Sigma_{ii} = \Sigma_{g(i),g(i)}$, and $\Sigma_{ij} = \Sigma_{g(i),g(j)}$ for all $g \in \mathfrak{G}$.

Given samples $X^{(1)}, \ldots X^{(n)}$ we define the following.

1. The *sample Reynolds average* of the $k^{th}$ sample is:

$$\hat{\Sigma}_{\mathcal{R}}^{(k)} := \mathcal{P}_{\mathfrak{G}} \left( X^{(k)} (X^{(k)})^T \right) = \frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} \Pi_g X^{(k)} X^{(k)^T} \Pi_g^T. \qquad (7)$$

2. The *empirical covariance matrix* is:

$$\Sigma^n := \frac{1}{n} \sum_{k=1}^{n} X^{(k)} X^{(k)^T}. \qquad (8)$$

3. The $\mathfrak{G}$-*empirical covariance matrix* is:

$$\hat{\Sigma} := \mathcal{P}_{\mathfrak{G}} \left( \Sigma^n \right). \qquad (9)$$

Note that the $(i, j)$ entry of the $\mathfrak{G}$-empirical covariance matrix may be written as:

$$\hat{\Sigma}_{ij} = \frac{1}{n|\mathfrak{G}|} \sum_{k=1}^{n} \sum_{g \in \mathfrak{G}} X_{g(i)}^{(k)} X_{g(j)}^{(k)}, \qquad (10)$$

where $X_i^{(k)}$ is the $i^{th}$ component of the sample $X^{(k)}$. We point out that (10) may be interpreted as averaging $X_i^{(k)} X_j^{(k)}$ with respect to the orbit $\mathcal{O}(i, j)$, followed by averaging over the samples $k = 1, \ldots, n$. In fact $\hat{\Sigma}$ is the *maximum likelihood estimate* of the population covariance matrix $\Sigma$. As we will argue subsequently, due to the double averaging, the estimate of the covariance matrix obtained via the $\mathfrak{G}$-empirical covariance matrix is a much better estimator as compared to the empirical covariance matrix. Indeed, quantification of this fact is one of the central themes of this paper.

**Remark 2.3.** Note that Lemma 2.2 ensures that if $M \succeq 0$, then $\mathcal{P}_{\mathfrak{G}}(M) \succeq 0$ (since positive definiteness is preserved under conjugation and addition). Hence the projection of the empirical covariance matrix $\mathcal{P}_{\mathfrak{G}}(\Sigma^n)$ is a meaningful estimate of $\Sigma$ in the sense that it is a valid covariance matrix.

*2.3.1. Sufficient statistics*

We make a few brief remarks about the *sufficient statistics* for symmetric Gaussian models. For an arbitrary (zero-mean) model, it is well-known that the sufficient statistics correspond to quadratics associated to the variance, i.e. the sufficient statistics are

$$\phi_{ij}(x) = x_i x_j \qquad \text{for } (i,j) \in V \times V.$$

For $\mathfrak{G}$-invariant models, the sufficient statistics are

$$\phi_e(x) = x_i x_j \qquad \text{for } e = (i,j) \in (V \times V)|\mathfrak{G},$$

where $(V \times V)|\mathfrak{G}$ denotes the *quotient* of $V \times V$ with respect to $\mathfrak{G}$ with respect to the equivalence relation $(i,j) \cong (k,l)$ if $g(i) = k$ and $g(j) = l$ for some $g \in \mathfrak{G}$. Put more simply, the sufficient statistics correspond to quadratics $\phi_{ij}$, and we have only one statistic per *orbit* $\mathcal{O}(i,j)$ (as opposed to one per pair $(i,j)$). The representative sufficient statistic from orbit $\mathcal{O}(i,j)$ may be picked to be:

$$\frac{1}{|\mathcal{O}(i,j)|} \sum_{(i,j) \in \mathcal{O}(i,j)} x_i x_j.$$

An alternate set of sufficient statistics can be obtained via Schur's lemma by observing that $Y = T^* X$ is an equivalent statistical model whose covariance is block diagonal. The sufficient statistics are $\theta_{ij}(y) = y_i y_j$ for $i,j$ belonging to the same irreducible component. Note that one only needs one set of statistics per irreducible component, and they may be expressed as above by averaging the corresponding sufficient statistics with respect to the multiplicities of each irreducible component.

## 3. Rates of convergence

In this section we will be interested in the statistical properties of the estimator $\hat{\Sigma} := \mathcal{P}_{\mathfrak{G}}(\Sigma^n)$. Note that given samples $X^{(1)}, \ldots X^{(n)}$, $\mathbb{E}(\hat{\Sigma}) = \Sigma$. In this section, we will be interested in establishing *rates of convergence* of $\hat{\Sigma}$ in different norms, i.e. we will provide upper bounds on $\mathbb{P}((\|\hat{\Sigma} - \Sigma\|_p \geq \delta)$, where $\| \cdot \|_p$ are the Frobenius, spectral and $\ell_\infty$ norms. The rates presented here hold for the covariance estimation of any sub-Gaussian random vector $X$. Though we present them in the context of normally distributed random vectors, identical analysis holds for sub-Gaussians more generally.

### *3.1. Spectral norm and Frobenius norm rates*

To obtain our rate results, we use the following standard result regarding spectral norm rates for an arbitrary normal model.

**Lemma 3.1.** *Let $X \sim \mathcal{N}(0, \Sigma^*)$ be a normal model with $\Sigma^* \in \mathbb{R}^{p \times p}$. Let $\psi = \|\Sigma^*\|$. Given $\delta > 0$ with $\delta \leq 8\psi$, let the number of samples satisfy $n \geq \frac{64p\psi^2}{\delta^2}$. Let $\Sigma^n$ be the empirical covariance matrix formed from the $n$ samples. Then we have*

$$\mathbb{P}\left(\|\Sigma^n - \Sigma^*\| \geq \delta\right) \leq 2 \exp\left(\frac{-n\delta^2}{128\psi^2}\right).$$

*Proof.* This is a standard result, a proof follows from [15, Theorem II.13]. $\quad\square$

**Theorem 3.1.** *Given an invariant covariance matrix $\Sigma$ with $\psi = \|\Sigma\|$ and a sample covariance $\Sigma^n$ formed from $n$ i.i.d. samples drawn from a Gaussian distribution according to covariance $\Sigma$, we have that*

$$\mathbb{P}[\|(\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n))\| \geq \delta] \leq |\mathcal{I}| \exp\left(-\frac{n \min_{i \in \mathcal{I}} m_i \delta^2}{128\psi^2}\right)$$

*for $n \geq \max_{i \in \mathcal{I}} \frac{64 s_i \psi^2}{m_i \delta^2}$ and for $\delta \leq 8\psi$.*

*Proof.* Due to the unitary invariance of the spectral norm, we may assume that $\Sigma$ is put in the suitable block-diagonal coordinates based on Schur's representation lemma. Letting $\mathcal{I}$ denote the indices of the set of irreducible representations, using Lemma 3.1 we have for each $i \in \mathcal{I}$ that

$$\mathbb{P}[\|(\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n))_i\| \geq \delta] \leq 2 \exp\left(-\frac{n m_i \delta^2}{128\psi^2}\right)$$

for $\delta \leq 8\psi$ and for $n m_i \geq \frac{64 s_i \psi^2}{\delta^2}$. Note that we get the extra factor of $m_i$ (the multiplicity of the representation indexed by $i$) since blocks corresponding to the same irreducible representation but different multiplicity are independent and identically distributed. Hence we average over the $m_i$ blocks of size $s_i$ (see Remark 2.1). Consequently we have by the union bound that

$$\mathbb{P}[\|(\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n))\| \geq \delta] \leq |\mathcal{I}| \exp\left(-\frac{n \min_{i \in \mathcal{I}} m_i \delta^2}{128\psi^2}\right)$$

for $n \geq \max_{i \in \mathcal{I}} \frac{64 s_i \psi^2}{m_i \delta^2}$. $\quad\square$

**Theorem 3.2.** *Given an invariant covariance matrix $\Sigma$ with $\psi = \|\Sigma\|_2$ and a sample covariance $\Sigma^n$ formed from $n$ i.i.d. samples drawn from a Gaussian distribution according to covariance $\Sigma$, we have that*

$$\mathbb{P}[\|(\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n))\|_F \geq \delta\sqrt{p}] \leq |\mathcal{I}| \exp\left(-\frac{n \min_{i \in \mathcal{I}} m_i \delta^2}{128\psi^2}\right)$$

*for $n \geq \max_{i \in \mathcal{I}} \frac{64 s_i \psi^2}{m_i \delta^2}$ and for $\delta \leq 8\psi$.*

*Proof.* The proof of this result follows directly from that of the previous theorem by noting that $\|\cdot\|_F \leq \sqrt{p}\|\cdot\|$ for $p \times p$ matrices. $\quad\square$

As a direct corollary of the preceding theorems, we get the following sample complexity results.

**Corollary 3.1.** *Assume that $\psi$ is a constant (independent of $p, n$). The sample complexity of obtaining $\|\Sigma - \hat{\Sigma}\| \leq \delta$ for $\delta \leq 8\psi$ with probability exceeding $1 - \frac{1}{p^c}$ for some constant $c$ is given by:*

$$n \geq C_1 \max \left\{ \max_{i \in \mathcal{I}} \frac{s_i}{m_i \delta^2}, \max_{i \in \mathcal{I}} \frac{\log p}{m_i \delta^2} \right\},$$

*for some constant $C_1$.*

*The sample complexity of obtaining $\|\Sigma - \hat{\Sigma}\|_F \leq \delta$ for $\delta \leq 8\psi$ with probability exceeding $1 - \frac{1}{p^c}$ for some constant $c$ is given by:*

$$n \geq C_2 \max \left\{ \max_{i \in \mathcal{I}} \frac{s_i p}{m_i \delta^2}, \max_{i \in \mathcal{I}} \frac{p \log p}{m_i \delta^2} \right\},$$

*for some constant $C_2$.*

**Remark 3.1.** It is instructive to compare the sample complexity described by Lemma 3.1 versus the rates obtained from Theorem 3.1 for the case of the covariance matrices invariant with respect to the cyclic, the symmetric, and the product of symmetric groups. Note that the spectral norm sample complexity obtained from Lemma 3.1 is $n \geq \frac{Cp}{\delta^2}$ for some constant $C$. As explained in Example 2.1, for the cyclic group case $s_i = m_i = 1$ for all $i$ so that the sample complexity is $n \geq \frac{C \log p}{\delta^2}$. For the symmetric group $s_1 = 1, m_1 = 1$ and $s_2 = 1, m_2 = p - 1$ so that the sample complexity becomes $n \geq \frac{C \log p}{\delta^2}$. For the case of product of $k$ symmetric groups, the sample complexity is $n \geq \frac{C \max\{k, \log p\}}{\delta^2}$. Finally, for the MAR example (invariance with respect to the wreath product of $\mathbb{Z}_2$), $s_i = 1$ for all $i$ and the smallest multiplicity is 1. Hence the sample complexity is $n \geq \frac{C \log p}{\delta^2}$.

Hence in all the cases, projection onto the fixed point subspace delivers a dramatic improvement in the sample complexity.

### 3.2. $\ell_\infty$ *norm rates*

In this section we examine the $\ell_\infty$ rates of convergence of the estimate

$$\hat{\Sigma} := \mathcal{P}_{\mathfrak{G}} \left( \Sigma^n \right).$$

In this section, as elsewhere in the paper, we will assume that our covariance model $\Sigma$ has bounded spectral norm, i.e. $\|\Sigma\|$ is bounded independent on $p$ and $n$. It is possible to make the dependence on this quantity more explicit at the expense of more notation in a straightforward manner, for simplicity and clarity we omit doing so.

**Theorem 3.3.** *Let $\mathcal{O}(i,j)$ be the orbit of the element $(i,j)$, let its degree be $d_{ij}$. Then there are constants $C_1, C_2$ such that for all $i, j$*

$$\mathbb{P}\left(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t\right) \leq \max\left\{\exp\left(-C_1 n |\mathcal{O}(i,j)| t^2\right), \exp\left(-C_2 \frac{n|\mathcal{O}(i,j)|}{d_{ij}} t\right)\right\}.$$

The proof is provided in the appendix.

Recall that $\mathcal{O} := \min_{i,j} |\mathcal{O}(i,j)|$ denotes the size of the smallest edge orbit, and $\mathcal{O}_d := \min_{i,j} \frac{|\mathcal{O}(i,j)|}{d_{ij}}$, and let $n_{\mathcal{O}}$ denote the number of *distinct* (i.e. inequivalent) edge orbits $\mathcal{O}(i,j)$. Note that $n_{\mathcal{O}} \leq p^2$ trivially.

**Corollary 3.2.** *We have,*

$$\mathbb{P}\left(\max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > t\right) \leq n_{\mathcal{O}} \max\left\{\exp\left(-C_1 n \mathcal{O} t^2\right), \exp\left(-C_2 n \mathcal{O}_d t\right)\right\}.$$

*Proof.* From Theorem 3.3, and by the definition of $\mathcal{O}$ and $\mathcal{O}_d$ we have for each $(i,j)$:

$$\mathbb{P}\left(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > t\right) \leq \max\left\{\exp\left(-C_1 n \mathcal{O} t^2\right), \exp\left(-C_2 n \mathcal{O}_d t\right)\right\}.$$

By taking union bound over the distinct orbits we obtain the required result. □

**Corollary 3.3.** *We have the following sample complexity results:*

$$|\Sigma_{ij} - \hat{\Sigma}_{ij}| \leq O_P\left(\max\left\{\sqrt{\frac{\log p}{n|\mathcal{O}(i,j)|}}, \frac{d_{ij}\log p}{n|\mathcal{O}(i,j)|}\right\}\right) \tag{11}$$

*For $|\Sigma_{ij} - \hat{\Sigma}_{ij}| \leq \delta$ with probability exceeding $1 - \frac{1}{p^c}$ for some constant $c$, we need*

$$n \geq C_3 \max\left\{\frac{\log p}{\delta^2 |\mathcal{O}(i,j)|}, \frac{d_{ij}\log p}{\delta|\mathcal{O}(i,j)|}\right\},$$

*for some constant $C_3$.*

*Proof.* This is a simple consequence of Theorem 3.3. □

**Corollary 3.4.** *We have the following sample complexity results for the $\ell_\infty$ norm:*

$$\max_{i,j} |\Sigma_{ij} - \hat{\Sigma}_{ij}| \leq O_P\left(\max\left\{\sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d}\right\}\right) \tag{12}$$

*To obtain an error bound $\max_{i,j} |\Sigma_{ij} - \hat{\Sigma}_{ij}| \leq \delta$ with probability exceeding $1 - \frac{1}{p^c}$ for some constant $c$, we need*

$$n \geq C_3 \max\left\{\frac{\log p}{\delta^2 \mathcal{O}}, \frac{\log p}{\delta \mathcal{O}_d}\right\},$$

*for some constant $C_3$.*

*Proof.* This is a simple consequence of Corollary 3.2. □

We remind the reader that the group-theoretic parameters $\mathcal{O}$ and $\mathcal{O}_d$ were introduced in Section 2.2. These parameters roughly capture the sizes of the orbits in the group. Since samples corresponding to the same orbit are statistically equivalent (though the information is not *i.i.d.*), we may expect to reuse this information to get statistical gains. Corollary 3.4 makes these gains explicit, indeed the sample complexity is smaller by a factor corresponding to the size of the orbit.

**Remark 3.2.** It is again instructive to contrast the standard $\ell_\infty$ rates versus the improved rates derived in this section for symmetric models. The standard $\ell_\infty$ rates require a sample complexity of $n \geq \frac{C \log p}{\delta^2}$ for $\delta = o(1)$. For the cyclic and symmetric group, our rates imply a sample complexity of $n \geq \frac{C \log p}{\delta^2 p}$ samples. (Similarly for the case of the product of symmetric groups our sample complexity is $n \geq \frac{C \log p}{\delta^2 \min_i r_i}$.) We point out that in the case of models invariant with respect to the symmetric group, the overall $\ell_\infty$ rate is governed by the rate of convergence of the diagonal entries of $\Sigma$ corresponding to the orbit $\mathcal{O}(1,1)$. We point out that the orbit $\mathcal{O}(1,2)$ is much larger, and the sample complexity for estimating $\Sigma_{12}$ is $n \geq C \max(\frac{\log p}{\delta^2 p(p-1)}, \frac{C \log p}{\delta p})$.

Finally for models invariant with respect to the wreath product of $\mathbb{Z}_2$ we have a sample complexity of $n \geq C(\frac{\log p}{\delta^2 p})$ for small $\delta$ (i.e. $0 < \delta < 1$).

**Remark 3.3.** We note that in the examples in the preceding remark related to the symmetric, cyclic and wreath product groups, if we fix $n = 1$ (i.e. one sample) and let $p \to \infty$ then the $\ell_\infty$ norm error tends to zero. In other words, for large $p$, a *single snapshot* of the process is enough to reproduce the covariance properties with good accuracy, a phenomenon qualitatively similar to stationary ergodic processes.

## 4. Sparse model selection

### 4.1. Sparse covariance models

Estimation of covariance matrices in a high dimensional setting is an important problem in statistical analysis since it is an essential prerequisite for data analysis procedures such as dimension reduction via PCA, classification, and testing independence relations. In the high dimensional setting (which arise for example from the discretization of Markov random fields), the dimension of the problem may be larger than the number of samples available. In this setting it is well-known that the empirical covariance matrix is a poor estimator for the covariance matrix [8]. There is a growing body of literature on understanding regularization procedures for covariance matrices that possess specific structure. For example, if the covariance matrix known to be sparse, Bickel and Levina [8] show that *thresholding* the empirical covariance matrix is both a natural and statistically meaningful means of regularization. Indeed they provide precise rates of convergence of the thresholded estimate to the true covariance matrix.

In this section we analyze sparse covariance matrices arising from Gaussian models with symmetries. We show that if the covariance matrix is known to be invariant under the action of $\mathfrak{G}$ and is sparse, a statistically meaningful way of regularizing the empirical covariance matrix is to first project it onto the fixed point subspace, followed by thresholding. Doing so provides a faster rate of convergence and reduces the sample complexity by a factor that depends on the group (more specifically on the size of the smallest orbit and the orbit degree).

More precisely, let $X \sim \mathcal{N}(0, \Sigma)$, and let $\Sigma$ be invariant under the action of the group $\mathfrak{G}$. Furthermore, we adopt a model of sparsity on $\Sigma$ that is somewhat more general (Bickel and Levina [8]) since it encompasses the usual notion of sparsity as well as power-law decay in the coefficients of $\Sigma$ as follows. Define the set of approximately sparse covariance matrices via a *uniformity class* as:

$$\mathcal{U}(q) = \left\{ \Sigma \mid \Sigma_{ii} \leq M, \sum_{j=1}^{p} |\Sigma_{ij}|^q \leq c_0(p), \text{ for all } i \right\} \tag{13}$$

for $0 < q < 1$. If $q = 0$ we have a set of sparse matrices as follows:

$$\mathcal{U}(0) = \left\{ \Sigma \mid \Sigma_{ii} \leq M, \sum_{j=1}^{p} \mathbb{1}(\Sigma_{ij} \neq 0) \leq c_0(p), \text{ for all } i \right\}. \tag{14}$$

Here we think of $c_0(p)$ as a parameter that controls the sparsity level of $\Sigma$, and is to be thought of as sublinear in $p$. Given samples $X_1, \ldots, X_n$ from this distribution, recall that $\Sigma^n$ is the empirical covariance matrix and the $\mathfrak{G}$-empirical covariance matrix is $\hat{\Sigma} := \mathcal{P}_{\mathfrak{G}}(\Sigma^n)$.

Recall that $\mathbb{1}(A)$ denote the indicator function for the event $A$ and define the thresholding operator (at threshold $s$) by

$$\mathcal{T}_s(M) = [M_{ij} \mathbb{1}(|M_{ij}| \geq s)], \tag{15}$$

so that the $(i, j)$ entry of $M$ is retained if it is larger than $s$ and is zero otherwise. We propose the following estimator for the covariance:

$$\Sigma_{\text{est}} = \mathcal{T}_t(\mathcal{P}_{\mathfrak{G}}(\Sigma^n))$$

for a suitable choice of threshold parameter $t$ to be specified in the subsequent theorem. As in [8], we note that the thresholding operator preserves symmetry and under mild assumptions on the eigenvalues, preserves positive-definiteness. Indeed, the next theorem provides a rate of convergence of our thresholded estimate to the true covariance in spectral norm. As a consequence, if the smallest eigenvalue of $\Sigma$ is bounded away from 0 (independent of $p$ and $n$), the theorem also establishes that with high probability the thresholded estimate is positive definite.

**Theorem 4.1.** *Suppose* $\Sigma \in \mathcal{U}(q)$ *satisfies Assumtion 2.1. Choose the threshold parameter*

$$t = M' \max\left\{ \sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d} \right\}$$

*for a constant $M'$. Then uniformly on $\mathcal{U}(q)$*

$$\|\mathcal{T}_t\left(\mathcal{P}_{\mathfrak{G}}\left(\Sigma^n\right)\right) - \Sigma\| = O_P\left(c_0(p)\max\left\{\sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d}\right\}^{1-q}\right).$$

**Remark 4.1.** Note that the sample complexity gains observed in the $\ell_\infty$ norms translate directly to gains for sparse covariance selection. Indeed, in the absence of symmetry the deviations $\delta = O_P\left(c_0(p)(\frac{\log p}{n})^{\frac{1-q}{2}}\right)$. In the case of models that are sparse and invariant with respect to the cyclic group, we see $\delta = O_P\left(c_0(p)(\frac{\log p}{np})^{\frac{1-q}{2}}\right)$. Moreover, for invariance with respect to the symmetric group, we have $\delta = O_P\left(c_0(p)(\frac{\log p}{np})^{1-q}\right)$. Hence we again observe sample complexity gains corresponding to the orbit parameters.

**Remark 4.2.** We note that as in [9], sample complexity results can be established in the Frobenius norm, and also for models with alternative forms of sparsity [8] arising from banding, tapering, etc. We omit a detailed discussion of the same in this paper.

### 4.2. Sparse concentration models

In the preceeding section we studied a means of regularizing structured covariance matrices which were sparse and symmetric with respect to $\mathfrak{G}$. In many applications, it is much more suitable to study sparsity in the concentration (i.e. inverse covariance) matrix. These are often best described using the terminology of *graphical models* [27].

Formally we assume that the inverse covariance of $\Theta^* := \Sigma^{-1}$ is sparse. We will associate a weighted undirected graph $\mathcal{G} = (V, E)$ with $p$ nodes and edge weights $\Theta^*_{ij}$ so that $\Theta^*_{ij} \neq 0$ if and only if $(i, j) \in E$. In the context of Gaussian graphical models, the edges in the graph have a particularly appealing interpretation which makes them suitable in the modeling of Gauss-Markov random fields. Indeed, the edges represent conditional independence relations (i.e. $X_i$ and $X_j$ are conditionally independent of a set of nodes $X_S$ for $S \subseteq [p]$ if and only if $S$ is a separator of $i$ and $j$ in the graph $\mathcal{G}$).

A fundamental problem in graphical models learning is that of inferring the structure of the graph $\mathcal{G}$ and the corresponding non-zero entries of $\Theta^*$ from measurements (see [34] and the references therein); this problem is called the *graphical model selection problem*. In the high-dimensional setup, one wishes to do so from a small number of measurements $n \ll p$. This has been shown [29, 34] to be possible provided the graph has additional structure, and is simple in the sense that it has low degree.

Ravikumar et al. [34] analyze a convex optimization based procedure to obtain consistent model selection of $\Theta^*$ in the high-dimensional regime. Specifically, if $\Sigma^n$ is the empirical covariance matrix, they propose selecting $\Theta^*$ as the solution to the following $\ell_1$ regularized log det problem:

$$\hat{\Theta} := \underset{\Theta \in \mathcal{S}^p_{++}}{\arg\min}\ \text{Tr}(\Sigma^n\Theta) - \log\det(\Theta) + \mu_n\|\Theta\|_{1,\text{off}}. \tag{16}$$

This approach has several appealing properties. Firstly the above optimization problem is *convex*, i.e. both the optimization objective and the constraint set are convex. Consequently, this formulation is amenable to efficient computation, since one can apply standard numerical procedures to efficiently solve for the global optimum of the above convex optimization problem.

Secondly, the above formulation is statistically meaningful. Indeed if $\mu_n = 0$ one recovers the standard maximum entropy formulation of the problem, and the $\ell_1$ penalty on the off-diagonal terms in $\Theta$ can be viewed as a regularization term which encourages sparsity. In [34], the authors prove that the solution to problem (16) has good statistical properties (under reasonable statistical assumptions). The model selection can be shown to be consistent with good rates of convergence (see [34, Theorem 1, Theorem 2]) when the regularizer $\mu_n$ is chosen judiciously. Importantly, these rates are governed by the *degree* of the graph $\mathcal{G}$ and only the logarithm of the ambient dimension $p$.

In this section, we will study in a similar setup the problem of Gaussian graphical model selection in which the graph $\mathcal{G}$ has low degree and also a non-trivial automorphism group. Recall that $\Sigma \in \mathfrak{G}$ implies that $\Theta^* \in \mathfrak{G}$, hence $\mathfrak{G}$ is the *automorphism group* of the graph $\mathcal{G}$. We will assume that the group $\mathfrak{G}$ is non-trivial and known, where as the graph $\mathcal{G}$ is unknown. We examine how knowledge of the structure of $\mathfrak{G}$ can be exploited for the model selection problem using i.i.d. samples $X^{(1)}, \ldots, X^{(n)}$ drawn from $\mathcal{N}(0, (\Theta^*)^{-1})$ to obtain statistical gains in terms of the sample complexity.

We propose exploiting the known symmetry structure by solving the modified log-determinant program with subspace constraints:

$$\hat{\Theta} := \underset{\Theta \in \mathcal{S}_{++}^p \cap \mathcal{W}_\mathfrak{G}}{\arg\min} \ \mathrm{Tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{1,\mathrm{off}}. \tag{17}$$

The following lemma shows that solving the above problem is equivalent to symmetrizing the empirical covariance matrix directly and then solving the standard formulation.

**Lemma 4.1.** *Let* $\hat{\Sigma} := \mathcal{P}_\mathfrak{G}(\Sigma^n)$. *The optimization problem*

$$\underset{\Theta \in \mathcal{S}_{++}^p}{\arg\min} \ \mathrm{Tr}(\hat{\Sigma}\Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{1,off} \tag{18}$$

*has the same solution as* (17).

*Proof.* Let $\Theta_0$ be the optimal solution of (17). Let $f_s(\Theta) = \mathrm{Tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{1,\mathrm{off}}$. Then by definition $f_s(\Theta_0) = f_s(\Pi_g \Theta \Pi_g^T)$ for all $g \in \mathfrak{G}$. Since $f_s(\Theta)$ is convex, by Jensen's inequality we have:

$$f_s\left(\frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} \Pi_g \Theta_0 \Pi_g^T\right) \leq f_s(\Theta_0).$$

However, $\frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} \Pi_g \Theta_0 \Pi_g^T$ is feasible for the problem (17), hence

$$f_s \left( \frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} \Pi_g \Theta_0 \Pi_g^T \right) = f_s(\Theta_0).$$

Next, note that

$$\mathrm{Tr} \left( \Sigma^n \frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} \Pi_g \Theta_0 \Pi_g^T \right) = \mathrm{Tr} \left( \hat{\Sigma} \Theta_0 \right).$$

Hence the problem (17) has the same optimal value as:

$$\underset{\Theta \in \mathcal{S}_{++}^p \cap \mathcal{W}_{\mathfrak{G}}}{\arg\min} \ \mathrm{Tr}(\hat{\Sigma}\Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{1,\mathrm{off}}. \tag{19}$$

(Both the optimization problems are strictly convex, and hence have the same unique solution [34, Appendix A]).

Now we note that $\mathrm{Tr}(\hat{\Sigma}\Theta) = \mathrm{Tr}(\Pi_g \hat{\Sigma} \Pi_g^T \Theta) = \mathrm{Tr}(\hat{\Sigma} \Pi_g \Theta \Pi_g^T)$ for all $g \in \mathfrak{G}$. Moreover $\log\det(\cdot)$ is unitarily invariant. Lastly, note that $\Pi_g \Theta \Pi_g^T$ for any permutation matrix $\Pi_g$ has the effect of permuting the rows and columns of $\Theta$, which does not affect the $\|\cdot\|_{1,\mathrm{off}}$ norm. Hence the objective function of (18) is invariant with respect to $\mathfrak{G}$ and is convex. Since the constraint set of (18) (the positive definite cone $\mathcal{S}_{++}^p$) is also invariant with respect to $\mathfrak{G}$ and convex, by a standard convexity argument it is possible to show that its solution must belong to the fixed point subspace. Thus the subspace constraint in (19) is redundant, and may be dropped. $\qquad\square$

Before we proceed to state the statistical properties of the optimal solution to (17), we must make some standard statistical assumptions. We denote $\Sigma$ and $\Theta^*$ to be the covariance and concentration matrix respectively for the true model. We denote by $S$ the set of indices $(i,j)$ (including the diagonal entries) that are non-zero in $\Theta^*$, and denote by $S^c$ the complement of this set. For any two index sets $T, T'$, and define the matrix $M_{T,T'}$ to be the submatrix of $M$ with rows and columns indexed by $T, T'$ respectively. Finally we define $\Gamma^* := \Theta^{*-1} \otimes \Theta^{*-1}$, where $\otimes$ denotes the Kronecker product.

We define the following quantities:

$$\kappa_{\Sigma^*} = \|\Sigma\|_{\infty,\infty} \qquad \kappa_{\Gamma^*} = \| (\Gamma_{SS}^*)^{-1} \|_{\infty,\infty}. \tag{20}$$

For simplicity, we will assume that that these quantities are *constants* in our asymptotic setting, i.e. they are bounded independent of $p$ and $n$ (although it is straightforward to make the dependence more explicit). Finally we make the following *incoherence* assumption which is also standard for such statistical models [34, Assumption 1]:

**Assumption 4.1.** *There exists some $\alpha \in (0,1]$ such that*

$$\max_{e \in S^c} \|\Gamma_{eS}^* (\Gamma_{SS}^*)^{-1}\|_1 \le 1 - \alpha.$$

We define the *degree* of the graphical model to be

$$d_{\mathcal{G}} := \max_{i=1,\ldots,p} \left\{ \sum_{j=1}^{p} \mathbb{1}_{\left\{ \Theta_{ij}^* \neq 0 \right\}} \right\}.$$

Note that this precisely corresponds to the degree of the graph $\mathcal{G}$.

**Theorem 4.2.** *Consider a symmetric Gaussian graphical model satisfying Assumption 2.1 and Assumption 4.1. Let $\hat{\Theta}$ be the unique solution to the problem log-determinant program* (17) *with regularization parameter*

$$\lambda_n = M_1 \max \left\{ \sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d} \right\}$$

*for some constant $M_1$. Suppose the sample size is bounded by $n > \max(n_1, n_2)$,*

$$n_1 = M_2 \frac{d_{\mathcal{G}}^2 \log p}{\mathcal{O}} \qquad n_2 = M_3 \frac{d_{\mathcal{G}} \log p}{\mathcal{O}_d}$$

*for some constants $M_2$ and $M_3$. Then with probability exceeding $1 - \frac{1}{p^\tau}$ for some constant $\tau > 0$ we have:*

1. *The estimate $\hat{\Theta}$ satisfies:*

$$\|\hat{\Theta} - \Theta^*\|_\infty \le M_0 \max \left\{ \sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d} \right\},$$

   *for some constant $M_0$.*
2. *It specifies an edge set $E(\hat{\Theta})$ that is a subset of the true edge set $E(\Theta^*)$, and includes all edges $(i,j)$ with $|\Theta_{ij}^*| > M_4 \max\{\sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d}\}$ for some constant $M_4$.*

*Proof.* Is a direct consequence of Corollary 3.3 and [34, Theorem 1]. $\square$

**Remark 4.3.** We note that the constants $M_0, M_1, M_2, M_3, M_4$ depend on the parameters $\|\Sigma\|, \alpha, \kappa_{\Sigma^*}, \kappa_{\Gamma^*}, \tau$ in a way that can be made explicit, as has been done in [34].

A more refined version of the second part of the preceding theorem that is analogous to [34, Theorem 2], and that establishes model selection consistency can also be established using our $\ell_\infty$ bounds, though we omit the statement of a precise result here.

**Remark 4.4.** Note that sample complexity gains in the $\ell_\infty$ norm rates are inherited for sparse inverse covariance selection. Indeed, the standard sample complexity requires $n \ge C d_{\mathcal{G}}^2 \log p$ samples to get an accuracy of

$$\|\hat{\Theta} - \Theta^*\|_\infty \le M \sqrt{\frac{\log p}{n}},$$

we have significant gains for models with symmetry. For the cyclic group, we need $n \geq C d_{\mathcal{G}}^2 \frac{\log p}{p}$ to obtain

$$\|\hat{\Theta} - \Theta^*\|_\infty \leq M\sqrt{\frac{\log p}{np}}.$$

For the symmetric group also we need $n \geq C d_{\mathcal{G}} \frac{\log p}{p}$ to obtain

$$\|\hat{\Theta} - \Theta^*\|_\infty \leq M\sqrt{\frac{\log p}{np}}.$$

### 4.2.1. Extension to latent variable graphical model selection

The results presented in [34] have been extended in [13] to deal with the setting in which one may not observe samples of all the relevant variables. In contrast to graphical model selection in which one seeks a sparse approximation to a concentration matrix, [13] proposed a convex optimization method to approximate a concentration matrix as the sum of a sparse matrix and a low-rank matrix. The low-rank component captures the effect of marginalization over a few latent variables, and the sparse component specifies the conditional graphical model among the observed variables conditioned on the latent variables. The main result in [13] is that consistent model selection in the latent-variable setting (i.e., decomposition into sparse and low-rank components with the components having the same support/rank as the true underlying components) via tractable convex optimization is possible under suitable identifiability conditions even when the number of latent variables and the number of samples of the observed variables are on the same order as the number of observed variables. The sample complexity in these consistency results can be improved by exploiting symmetry in the underlying model in a similar manner to our presentation above. To consider just one simple example, if the marginal covariance matrix of the observed variables is invariant with respect to the cyclic group, then consistent latent-variable model selection using the convex programming estimator of [13] and by exploiting symmetry is possible (under suitable identifiability conditions) when the number of latent variables is on the same order as the number of observed variables, but with the number of samples of the observed variables growing only logarithmically in the number of observed variables. This improvement in the rates follows directly from our analysis in Section 3.

## 5. Computational aspects

In the preceding sections we have argued that knowledge of symmetry in a statistical model can lead to significant statistical gains in terms of sample complexity. To do so, one must exploit knowledge regarding symmetry judiciously by regularizing the model appropriately. We will now investigate the computational efficiency associated to symmetry regularization.

### 5.1. Projection onto the fixed-point subspace

A fundamental computational step in exploiting symmetry in statistical models is to project the empirical covariance matrix $\Sigma^n$ onto the fixed point subspace $\mathcal{W}_{\mathfrak{G}}$, i.e. computing

$$\hat{\Sigma} = \mathcal{P}_{\mathfrak{G}}(\Sigma^n).$$

As argued in Section 3, this estimator is a far more accurate estimator of the covariance as compared to the empirical covariance matrix. Moreover, in the selection of sparse covariance models with symmetry, our estimator is of the form:

$$\Sigma_{\text{est}} = \mathcal{T}_t\left(\mathcal{P}_{\mathfrak{G}}(\Sigma^n)\right)$$

where $t$ is an appropriate threshold level as described in Section 4.1. In the selection of sparse graphical models as described in Section 4.2, one needs to solve problem (18), which requires the computation of $\hat{\Sigma} = \mathcal{P}_{\mathfrak{G}}(\Sigma^n)$.

Conceptually the projection can be computed in numerous ways. If the group is of small order, it is possible to compute the projection using Lemma 2.2. If the group has a small number of generators (for example the cyclic group) $\mathfrak{G}_{\text{gen}} \subseteq \mathfrak{G}$, then the fixed point subspace can be expressed as

$$\mathcal{W}_{\mathfrak{G}} = \left\{M \mid \rho(g)M = M\rho(g) \ \forall g \in \mathfrak{G}_{\text{gen}}\right\},$$

and thus projections can be computed by solving the constrained least squares problem

$$\mathcal{P}_{\mathfrak{G}}(\Sigma^n) = \underset{M \,:\, \rho(g)M=M\rho(g) \ \forall g \in \mathfrak{G}_{\text{gen}}}{\arg\min} \|M - \Sigma^n\|_F^2.$$

One need only specify one matrix equality constraint corresponding to each generator of the group (the constraints corresponding to other group elements are implied by the generators), so that for groups such as abelian groups this assumes an especially simple form.

Lastly, if the diagonalizing matrix $T$ corresponding to the irreducible representations from Lemma 2.1 are available, the projections are especially easy to compute. To compute $\mathcal{P}_{\mathcal{W}_{\mathfrak{G}}}(\Sigma^n)$ one can simply compute $\Phi := T^*\Sigma^n T$. One can truncate the off-diagonal blocks of $\Omega$ and average with respect to the multiplicities as described by Remark 2.1 to obtain $\Phi'$. Finally, one obtains $\mathcal{P}_{\mathcal{W}_{\mathfrak{G}}}(\Sigma^n) = T\Phi'T^*$.

### 5.2. Solution of the $\ell_1$ regularized $\log\det$ problem

Given a Gaussian graphical model which is symmetric with respect to the group $\mathfrak{G}$, we examined in Section 4.2 how one can reliably recover the model by solving the optimization problem (17) (or equivalently (18)). This optimization problem is convex and expressible as a $\log\det$ conic program. An interesting consequence of symmetry reduction techniques that are now well-understood in optimization [23, 39] is that by exploiting the group invariance of the objective function

and the constraint set one can reduce the size of (18). Let $\Gamma := T^*\Theta T$, and $S := T^*\hat{\Sigma}T$. Then (18) reduces to:

$$\min_{\Gamma} \quad \mathrm{Tr}\,(S\Gamma) - \log\det(\Gamma) + \mu_n\|T\Gamma T^*\|_{1,\mathrm{off}}$$

$$\text{subject to:} \quad \Gamma \succ 0.$$

In the resulting conic program in the transformed domain, the optimization variable $\Gamma$ is block diagonal, so that there is a reduction in the problem size. Furthermore, many iterative algorithms require linear algebraic operations (such as matrix inversion). These steps can be sped up for block diagonal matrices, and the complexity is governed by the size of the largest irreducible component as given by Schur's lemma.

In a high dimensional setup where the problem size $p$ is large, it is often prohibitively expensive to implement an interior point method to solve the problem due to the computational burden of the Newton step [11]. Moreover, the above formulation is inefficient due to the introduction of additional variables $S_{ij}$. For these reasons, it is far more desirable to implement an *accelerated first order method* that uses the special structure of the objective function. One such method described by d'Aspremont et al. [14, pp. 4-5], provides an efficient iterative algorithm with low memory requirements. A careful examination of their algorithm reveals that the iterations are $\mathfrak{G}$-invariant, i.e. each iterate of the algorithm lies on the fixed-point subspace. Moreover, the computational bottlenecks in the implementation of the algorithm correspond to inversion and eigenvalue decompositions of matrices that belong to the fixed point subspace. Here again, the unitary transform $T$ and the associated block-diagonal structure is particularly useful and will provide significant computational speedups.

## 6. Experimental results

In this section we report results obtained from a set of simulations that illustrate the efficacy of accounting for symmetry via the methods proposed in this paper.

In the first simulation experiment we generate a covariance matrix $\Sigma$ of size $p = 50$ that is circulant and hence invariant with respect to the cyclic group. The first row of the circulant matrix is randomly generated (and then modified to ensure the matrix is positive definite). Figure 3 illustrates the rates of convergence that are observed in different matrix norms as the number of samples are increased. (For each sampling level, 20 trials are conducted and the average error is reported). It is apparent that the estimate formed by projection of $\Sigma^n$ onto the fixed-point subspace far outperforms the estimate formed by $\Sigma^n$.

In the next simulation, we generate a *sparse* circulant covariance matrix $\Sigma$ of size $p = 50$ (so that it is again invariant with respect to the cyclic group). The entries as well as the sparsity pattern of the first row of the circulant matrix are random. The sparsity pattern is such that twenty percent of the matrix has non-zero entries. Figure 4 compares the performance obtained by different means of regularization. The first is symmetry-oblivious and uses the
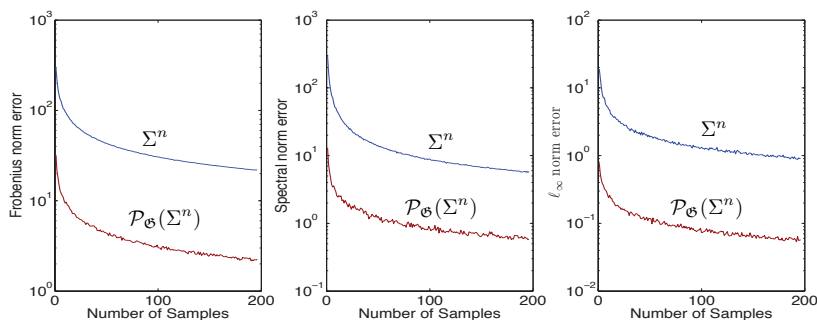
FIG 3. *A comparison of the rates of convergence for a covariance matrix that is invariant with respect to the cyclic group. The $\mathfrak{G}$-empirical covariance matrix is a better estimate as compared to the empirical covariance matrix.*
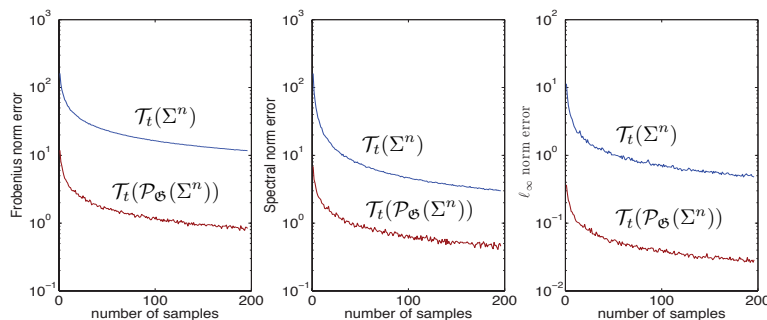


FIG 4. *A comparison of the rates of convergence for a sparse covariance matrix that is invariant with respect to the cyclic group. Projection followed by thesholding provides a better estimate.*

methodology proposed by Bickel and Levina [8], that thresholds the matrix at an appropriate level (that is proportional to $\sqrt{\frac{\log p}{n}}$). The second methodology is symmetry aware, and projects the empirical covariance matrix to the fixed-point subspace, followed by thresholding (proportional to $\sqrt{\frac{\log p}{pn}}$ as proposed in Section 4.1). We compare the empirically observed error in different norms.

In the last experiment, we consider the problem of recovering a sparse inverse covariance matrx (i.e. a graphical model) of size $p = 50$. The graph corresponding to the graphical model is a cycle. The entries $\Theta^*_{ii} = 1$ for all $i$. The edge weights of the cycle are all set to $\Theta^*_{ij} = -0.25$, the remaining entries of $\Theta^*$ are zero. We perform an experiment pertaining to *model selection consistency*, i.e. recovering the structure of the graph from a small number of measurements. We compare two methodologies. The first is symmetry-oblivious, and uses the
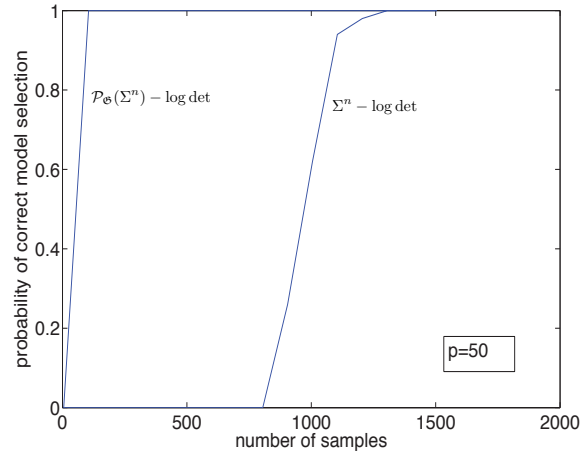
FIG 5. *A comparison of the model selection consistency obtained by* $\mathcal{P}_{\mathfrak{G}}(\Sigma^n)$ *based* $\log\det$ *formulation versus a* $\Sigma^n$ *based* $\log\det$ *formulation.*

$\ell_1$ regularized $\log\det$ formulation proposed in [34] using only the empirical co-variance matrix $\Sigma^n$ in the objective function. The second uses the formulation (18) proposed in Section 4.2, i.e. we first project $\Sigma^n$ onto the fixed-point sub-space corresponding to the cyclic group, and then solve the $\ell_1$ regularized $\log\det$ problem. The empirically observed probabilities of correct model selection us-ing these two methods are shown in Fig. 5. This plot is generated by varing the number of samples from $n = 5$ to $n = 1505$ at increments of 100 samples. For each sample level, we perfom 50 experiments and report the average probability of success. The regularization parameter for the symmetry oblivious method is chosen proportional to $\sqrt{\frac{\log p}{n}}$, as proposed in [34], and for the symmetry aware method it is chosen to be proportional to $\sqrt{\frac{\log p}{np}}$, as proposed in Section 4.2. Both the symmetry oblivious, as well as the symmetry aware formulations re-quired the solution of $\ell_1$ regularized $\log\det$ problems. These were solved using the solver COVSEL [14].

As seen in these experiments, for models containting symmetry the symmetry aware methodology proposed in this paper significantly outperforms symmetry oblivious methods.

## 7. Conclusion

In this paper we have examined Gaussian models with symmetries in the high-dimensional regime. We have suggested a natural means of estimating a $\mathfrak{G}$-invariant model from a small number of samples. The key idea is the formation

of the projection of the empirical covariance matrix onto the fixed point subspace, which enjoys good statistical properties. As one of the main results of this paper, we have made precise the statistical gains enjoyed by this estimate by establishing explicit sample complexity bounds for the Frobenius, spectral, and $\ell_\infty$ norms. These sample complexity bounds depend naturally on certain group theoretic parameters that characterize the extent to which the group theoretic structure allows sample reuse. We have further used our sample complexity bounds to analyze sparse covariance models, and sparse graphical models that enjoy symmetry. Finally we provide experimental evidence that verifies our results via computer simulations.

## Acknowledgement

## Appendix A: $\ell_\infty$ rates

Before we present the formal proof, we remind the reader of some basic machinery related to pre-Gaussian random variables.

**Definition A.1.** A random variable $Y \in \mathbb{R}$ is called *pre-Gaussian* [12] with parameters $\tau, \Lambda$ if

$$\mathbb{E} \exp(\lambda Y) \leq \exp\left(\frac{1}{2}\tau^2\lambda^2\right),$$

for all $\lambda \in (-\Lambda, \Lambda)$.

A useful fact about the pre-Gaussian variables (see [12, Chapter 1]) is that if $Y$ is pre-Gaussian with parameters $\tau, \Lambda$, the random variable $cY$ is pre-Gaussian with parameters $|c|\tau, \frac{\Lambda}{|c|}$. The following lemma characterizes the tail behavior of pre-Gaussian random variables.

**Lemma A.1.** *Suppose $Y$ is a pre-Gaussian random variable with parameters $\tau, \Lambda$. Then*

$$\mathbb{P}(|Y| > t) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\tau^2}\right) & \text{if } 0 < t < \Lambda\tau^2 \\ 2\exp\left(-\frac{\Lambda t}{2}\right) & \text{if } \Lambda\tau^2 < t. \end{cases}$$

*Proof.* See [12, Chapter 1]. $\qquad\square$

The canonical examples of pre-Gaussian random variables are $\chi^2$ distributed random variables. Indeed, let $D \in \mathbb{R}^{p \times p}$ be a diagonal positive definite matrix and let $X \sim \mathcal{N}(0, D)$ be a $\mathbb{R}^p$-valued Gaussian random variable. Then the random variable $X^T X - \mathrm{Tr}(D)$ is pre-Gaussian (see [12, Lemma 5.1, Example 3.1]) with parameters

$$\tau^2 = \mathrm{Tr}(D) \qquad \Lambda = \frac{1}{2\max_i D_{ii}}. \tag{21}$$

**Lemma A.2.** *Let $Y \in \mathbb{R}^n \sim \mathcal{N}(0, \Sigma)$. Let $Q = Q^T \succ 0$ and let $\lambda_{\max}(\cdot)$ denote the maximum eigenvalue. Then*

$$\mathbb{E} \exp\left(\lambda(Y^T Q Y - \mathrm{Tr}(Q\Sigma))\right) \leq \exp\left(\frac{1}{2}\lambda^2 \mathrm{Tr}(Q\Sigma)\right) \ \forall \lambda \in [-\Lambda, \Lambda],$$

*where $\Lambda = \frac{1}{2\lambda_{\max}((\Sigma^{\frac{1}{2}})^T Q \Sigma^{\frac{1}{2}})}$.*

*Proof.* For $Y \sim \mathcal{N}(0, \Sigma)$ let us define $X = Q^{\frac{1}{2}} Y$. Then $X \sim \mathcal{N}(0, Q^{\frac{1}{2}} \Sigma Q^{\frac{1}{2}})$ and $Y^T Q Y - \mathrm{Tr}(Q\Sigma) = X^T X - \mathrm{Tr}(Q\Sigma)$. Consider $R = U^T X$ where $U$ is unitary and has the property that $U^T Q^{\frac{1}{2}} \Sigma Q^{\frac{1}{2}} U = D$ and $D$ is diagonal. Applying (21) to $R^T R - \mathrm{Tr}(Q\Sigma) = Y^T Q Y - \mathrm{Tr}(Q\Sigma)$, and using the fact that $\mathrm{Tr}(Q^{\frac{1}{2}} \Sigma Q^{\frac{1}{2}}) = \mathrm{Tr}(Q\Sigma)$, and $\lambda_{\max}(Q^{\frac{1}{2}} \Sigma Q^{\frac{1}{2}}) = \lambda_{\max}(\Sigma^{\frac{1}{2}} Q \Sigma^{\frac{1}{2}})$, we have the desired result. $\square$

In the subsequent analysis, it will be more convenient to deal with normalized random variables, which we define as:

$$\bar{X}_{g(i)}^{(k)} := \frac{X_{g(i)}^{(k)}}{\sqrt{\Sigma_{ii}}}.$$

Note that $\mathbb{E}\left(\bar{X}_{g(i)}^{(k)}\right)^2 = 1$. We also define:

$$\rho_{ij} := \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} \Sigma_{jj}}},$$

and note that by symmetry of the covariance matrix $\Sigma_{ii} = \Sigma_{g(i),g(i)}$ and hence $\rho_{ij} = \rho_{g(i),g(i)}$. Define

$$Y_{ij}^{(k)} = \sum_{(g(i),g(j)) \in \mathcal{O}(i,j)} \bar{X}_{g(i)}^{(k)} \bar{X}_{g(j)}^{(k)}. \tag{22}$$

We observe that

$$Y_{ij}^{(k)} - |\mathcal{O}(i,j)| \rho_{ij} = \frac{1}{4}\left(U_{ij}^{(k)} - V_{ij}^{(k)}\right) \qquad \text{where,}$$
$$U_{ij}^{(k)} = \sum_{(g(i),g(j)) \in \mathcal{O}(i,j)} (\bar{X}_{g(i)}^{(k)} + \bar{X}_{g(j)}^{(k)})^2 - 2|\mathcal{O}(i,j)|(1 + \rho_{ij}) \tag{23}$$
$$V_{ij}^{(k)} = \sum_{(g(i),g(j)) \in \mathcal{O}(i,j)} (\bar{X}_{g(i)}^{(k)} - \bar{X}_{g(j)}^{(k)})^2 - 2|\mathcal{O}(i,j)|(1 - \rho_{ij}).$$

For the covariance matrix of the normalized random variables occuring in the orbit of $(i,j)$ define $\sigma_{ij} = \|\Sigma_{\mathcal{V}(i,j)}\|$, the spectral norm restricted to the variables occuring in the orbit. Note that $\sigma_{ij} \leq \|\Sigma\|$ and is hence bounded by a constant. Also recall that $d_{ij}$ is the degree of the orbit $\mathcal{O}(i,j)$.

**Lemma A.3.** *The following inequalities hold:*

$$\mathbb{E}\exp(\lambda U_{ij}^{(k)}) \leq \exp\left(|\mathcal{O}(i,j)|\lambda^2(1 + \rho_{ij})\right) \text{ for all } \lambda \in \left[-\frac{1}{4d_{ij}\sigma_{ij}}, \frac{1}{4d_{ij}\sigma_{ij}}\right]$$

$$\mathbb{E}\exp(\lambda V_{ij}^{(k)}) \leq \exp\left(|\mathcal{O}(i,j)|\lambda^2(1 - \rho_{ij})\right) \text{ for all } \lambda \in \left[-\frac{1}{4d_{ij}\sigma_{ij}}, \frac{1}{4d_{ij}\sigma_{ij}}\right].$$

*Proof.* Fix a particular sample $k$, a pair $(i,j)$, and consider the corresponding $\mathcal{O}(i,j)$ and $\mathcal{V}(i,j)$. Let $Z \in \mathbb{R}^{|\mathcal{V}(i,j)|}$ be the vector of all the distinct nodes that occur on the edge orbit $\mathcal{O}(i,j)$. Let $W^+$ be the vector of all the terms $X_{g(i)}^{(k)} + X_{g(j)}^{(k)}$, and $W^-$ be the vector of all the terms $X_{g(i)}^{(k)} - X_{g(j)}^{(k)}$. Then there are matrices $A^+$ and $A^-$ (node-edge incidence matrices) such that

$$W^+ = A^+ Z \qquad W^- = A^- Z.$$

Note that $U_{ij}^{(k)} = (W^+)^T W^+ - 2|\mathcal{O}(i,j)|(1 + \rho_{ij})$ and $V_{ij}^{(k)} = (W^-)^T W^- - 2|\mathcal{O}(i,j)|(1 - \rho_{ij})$. Let $\Sigma_Z$ be the submatrix of the covariance matrix associated to $Z$. Then we note that we can write

$$(W^+)^T W^+ = Z^T (A^+)^T A^+ Z$$

where $Z \in \mathbb{R}^{|\mathcal{V}(i,j)|} \sim \mathcal{N}(0, \Sigma_Z)$. Note that by definition of $A^+$ and $Z$, $\mathrm{Tr}((A^+)^T \times A^+ \Sigma_Z) = 2\mathcal{O}(i,j)(1 + \rho_{ij})$ (similar equality holds for $A^-$). To be able to invoke Lemma A.2 we need to bound two quantities:

$$\lambda_{\max}\left((\Sigma_Z^{\frac{1}{2}})^T (A^+)^T A^+ \Sigma_Z^{\frac{1}{2}}\right) \text{ and } \lambda_{\max}\left((\Sigma_Z^{\frac{1}{2}})^T (A^-)^T A^- \Sigma_Z^{\frac{1}{2}}\right).$$

Note that

$$\lambda_{\max}\left((\Sigma_Z^{\frac{1}{2}})^T (A^+)^T A^+ \Sigma_Z^{\frac{1}{2}}\right) \le \|\left(A^+\right)^T A^+\|\|\Sigma_Z\|$$
$$\le \sigma_{ij}\|\left(A^+\right)^T A^+\|.$$

By the definition of $A^+$, we have

$$\|\left(A^+\right)^T A^+\| = \max_{x \ne 0} \frac{\sum_{(g(i),g(j)) \in \mathcal{O}(i,j)} (x(i) + x(j))^2}{\sum_{k \in \mathcal{V}(i,j)} x_k^2}$$
$$\le \max_{x \ne 0} \frac{2\sum_{(g(i),g(j)) \in \mathcal{O}(i,j)} \left(x(i)^2 + x(j)^2\right)}{\sum_{k \in \mathcal{V}(i,j)} x_k^2}$$
$$\le \max_{x \ne 0} \frac{2d_{ij} \sum_{k \in \mathcal{V}(i,j)} x_k^2}{\sum_{k \in \mathcal{V}(i,j)} x_k^2}$$
$$\le 2d_{ij}.$$

Hence $\lambda_{\max}((\Sigma_Z^{\frac{1}{2}})^T (A^+)^T A^+ \Sigma_Z^{\frac{1}{2}}) \le 2d_{ij}\sigma_{ij}$. (Similar inequalities hold for $A^-$). Now using Lemma A.2, the result follows. $\qquad\square$

**Lemma A.4.** $\dfrac{\hat{\Sigma}_{\mathcal{R},ij}^{(k)}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} = \dfrac{1}{|\mathcal{O}(i,j)|} Y_{ij}^{(k)}.$

*Proof.*

$$\hat{\Sigma}_{\mathcal{R},ij} = \frac{1}{|\mathfrak{G}|} \sum_{g \in \mathfrak{G}} X_{g(i)} X_{g(j)}$$
$$= \frac{1}{|\mathfrak{G}|} \sum_{(g(i),g(j)) \in \mathcal{O}(i,j)} n_{ij} X_{g(i)} X_{g(j)},$$

where $n_{ij}$ is the number of times edge $(i, j)$ (and hence all other edges in the orbit) are repeated in the Reynolds average. Now note that the size of the orbit divides the size of the group, and in fact $\frac{|\mathfrak{G}|}{|\mathcal{O}(i,j)|} = n_{ij}$ (in fact both these are equal to the cardinality of the stabilizer of the edge $(i, j)$). Dividing throughout by $\sqrt{\Sigma_{ii}\Sigma_{jj}}$, we get the desired result. $\qquad\square$

Let

$$U_{ij} = \frac{1}{4n|\mathcal{O}(i,j)|} \sum_{k=1}^{n} U_{ij}^{(k)} \qquad V_{ij} = \frac{1}{4n|\mathcal{O}(i,j)|} \sum_{k=1}^{n} V_{ij}^{(k)} \qquad Y_{ij} = \frac{1}{4n|\mathcal{O}(i,j)|} \sum_{k=1}^{n} Y_{ij}^{(k)}. \tag{24}$$

**Lemma A.5.**

$$\mathbb{P}\left(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > \sqrt{\Sigma_{ii}\Sigma_{jj}}t\right) \leq \mathbb{P}\left(|U_{ij}| > \frac{t}{2}\right) + \mathbb{P}\left(|V_{ij}| > \frac{t}{2}\right).$$

*Proof.* We have

$$\begin{aligned}
\frac{\hat{\Sigma}_{ij} - \Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} &= \frac{1}{n|\mathcal{O}(i,j)|} \sum_{k=1}^{n} Y_{ij}^{(k)} - \rho_{ij} \\
&= \frac{1}{n} \sum_{k=1}^{n} \frac{1}{|\mathcal{O}(i,j)|} \left(Y_{ij}^{(k)} - |\mathcal{O}(i,j)|\rho_{ij}\right) \\
&= \frac{1}{n} \sum_{k=1}^{n} \frac{1}{4|\mathcal{O}(i,j)|} \left(U_{ij}^{(k)} - V_{ij}^{(k)}\right) \qquad \text{(using (23))} \\
&= U_{ij} - V_{ij} \qquad \text{(by (24)).}
\end{aligned}$$

Now

$$\left\{|\hat{\Sigma}_{ij} - \Sigma_{ij}| > \sqrt{\Sigma_{ii}\Sigma_{jj}}t\right\} = \{|U_{ij} - V_{ij}| > t\} \subseteq \left\{|U_{ij}| > \frac{t}{2}\right\} \cup \left\{|V_{ij}| > \frac{t}{2}\right\},$$

from which the conclusion follows. $\qquad\square$

*Proof of Theorem 3.3.* Note that $\mathbb{P}\left(|U_{ij}| > \frac{t}{2}\right) \leq 2\mathbb{P}\left(U_{ij} > \frac{t}{2}\right)$. Now $U_{ij} = \frac{1}{4n|\mathcal{O}(i,j)|} \sum_{k=1}^{n} U_{ij}^{(k)}$, so that

$$\mathbb{E}\exp(\lambda U_{ij}) = \mathbb{E}\exp\left(\frac{1}{4n|\mathcal{O}(i,j)|} \sum_{k=1}^{n} U_{ij}^{(k)}\right).$$

From Lemma A.3,

$$\mathbb{E}\exp(\lambda U_{ij}^{(k)}) \leq \exp\left(|\mathcal{O}(i,j)|\lambda^2(1 + \rho_{ij})\right) \text{ for all } \lambda \in \left[-\frac{1}{4d_{ij}\sigma_{ij}}, \frac{1}{4d_{ij}\sigma_{ij}}\right].$$

By rescaling, we have

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{1}{4n|\mathcal{O}(i,j)|}\lambda U_{ij}^{(k)}\right) \\
&\leq \exp\left(\frac{(1 + \rho_{ij})}{16|\mathcal{O}(i,j)|n^2}\lambda^2\right) \text{ for all } \lambda \in \left[-\frac{n|\mathcal{O}(i,j)|}{d_{ij}\sigma_{ij}}, \frac{n|\mathcal{O}(i,j)|}{d_{ij}\sigma_{ij}}\right].
\end{aligned}$$

Since the random variables $U_{ij}^{(k)}$ are independent and identically distributed (they correspond to different samples) for $k = 1, \ldots, n$, we have that

$$\mathbb{E} \exp\left(\lambda U_{ij}\right) \leq \exp\left(\frac{(1+\rho_{ij})}{16|\mathcal{O}(i,j)|n}\lambda^2\right) \text{ for all } \lambda \in \left[-\frac{n|\mathcal{O}(i,j)|}{d_{ij}\sigma_{ij}}, \frac{n|\mathcal{O}(i,j)|}{d_{ij}\sigma_{ij}}\right].$$

Thus $U_{ij}$ is pre-Gaussian with parameters $\tau^2 = \frac{(1+\rho_{ij})}{16|\mathcal{O}(i,j)|n}$ and $\Lambda = \frac{n|\mathcal{O}(i,j)|}{d_{ij}\sigma_{ij}}$. From Lemma A.1 we have

$$\mathbb{P}\left(|U_{ij}| > \frac{t}{2}\right) \leq \begin{cases} 2\exp\left(-\frac{2t^2 n|\mathcal{O}(i,j)|}{1+\rho_{ij}}\right), & \text{for } 0 \leq t \leq \frac{1+\rho_{ij}}{16d_{ij}\sigma_{ij}} \\ 2\exp\left(-\frac{n|\mathcal{O}(i,j)|t}{2d_{ij}\sigma_{ij}}\right), & \text{for } t > \frac{1+\rho_{ij}}{16d_{ij}\sigma_{ij}}. \end{cases}$$

Another way of rewriting the above is:

$$\mathbb{P}\left(|U_{ij}| > \frac{t}{2}\right) \leq \max\left\{2\exp\left(-\frac{2t^2 n|\mathcal{O}(i,j)|}{1+\rho_{ij}}\right), 2\exp\left(-\frac{n|\mathcal{O}(i,j)|t}{2d_{ij}\sigma_{ij}}\right)\right\}.$$

By a similar argument,

$$\mathbb{P}\left(|V_{ij}| > \frac{t}{2}\right) \leq \max\left\{2\exp\left(-\frac{2t^2 n|\mathcal{O}(i,j)|}{1-\rho_{ij}}\right), 2\exp\left(-\frac{n|\mathcal{O}(i,j)|t}{2d_{ij}\sigma_{ij}}\right)\right\}.$$

Let $C_1 = \min_{i,j}\{\frac{2}{1+\rho_{ij}}, \frac{2}{1-\rho_{ij}}\}$, and $C_2 = \min_{i,j}\{\frac{2}{\sigma_{ij}}, \frac{2}{\sigma_{ij}}\}$. (Note that all these quantities are constants since the spectral norm $\|\Sigma\|$ is bounded by a constant by assumption). Using Lemma A.5, we get the following.

$$\mathbb{P}\left(|\hat{\Sigma}_{ij} - \Sigma_{ij}| > \sqrt{\Sigma_{ii}\Sigma_{jj}}t\right)$$
$$\leq \max\left\{\exp\left(-C_1 n|\mathcal{O}(i,j)|t^2\right), \exp\left(-\frac{C_2 n|\mathcal{O}(i,j)|t}{d_{ij}}\right)\right\}.$$

Noting that $\sqrt{\Sigma_{ii}\Sigma_{jj}}$ is bounded by a constant, we replace $t$ by $\frac{t}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$, and alter the constants in the exponent suitably to obtain the required result. $\square$

## Appendix B: Proof of Theorem 4.1

We define the following function that captures the $\ell_\infty$ norm sample complexity

$$\delta(n, p) := \max\left\{\sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d}\right\}. \tag{25}$$

It will be convenient to work with the notation $\hat{\Sigma} = \mathcal{P}_{\mathfrak{G}}(\Sigma^n)$.

*Proof.* The proof is almost identical to the proof of [8, Theorem 1]. We present it below for the sake of completeness. Note that from Corollary 3.2,

$$\max_{i,j}|\hat{\Sigma}_{ij} - \Sigma_{ij}| = O_P\left(\delta(n, p)\right).$$

Now we note that

$$\|\mathcal{T}_t(\hat{\Sigma}) - \Sigma\| \leq \|\mathcal{T}_t(\Sigma) - \Sigma\| + \|\mathcal{T}_t(\hat{\Sigma}) - \mathcal{T}_t(\Sigma)\|,$$

and we will focus on bounding the different terms.

Recall that for a symmetric matrix $M$, $\|M\| \leq \max_{j=1,\ldots,p} \sum_{i=1}^{p} |M_{ij}|$. (This is because $\|M\| \leq (\|M\|_{1,1}\|M\|_{\infty,\infty})^{\frac{1}{2}} = \|M\|_{1,1}$). As a consequence, the first term is bounded by

$$\max_i \sum_{j=1}^{p} |\Sigma_{ij}| \mathbb{1}(|\Sigma_{ij}| \leq t) \leq t^{1-q} c_0(p).$$

The second term can be bounded as

$$\|\mathcal{T}_t(\hat{\Sigma}) - \mathcal{T}_t(\Sigma)\| \leq \max_i \sum_{j=1}^{p} |\hat{\Sigma}_{ij}| \mathbb{1}(|\hat{\Sigma}_{ij}| \geq t, |\Sigma_{ij}| < t)$$

$$+ \max_i \sum_{j=1}^{p} |\Sigma_{ij}| \mathbb{1}(|\hat{\Sigma}_{ij}| < t, |\Sigma_{ij}| \geq t)$$

$$+ \max_i \sum_{j=1}^{p} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \mathbb{1}(|\hat{\Sigma}_{ij}| \geq t, |\Sigma_{ij}| \geq t)$$

$$= \mathrm{I} + \mathrm{II} + \mathrm{III}.$$

Now note that

$$\mathrm{III} \leq \max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \max_i \sum_{j=1}^{p} |\hat{\Sigma}_{ij}|^q t^{-q} = O_P\left(c_0(p) t^{-q} \delta(n,p)\right).$$

For term II we note that

$$\mathrm{II} \leq \max_i \sum_{j=1}^{p} (|\hat{\Sigma}_{ij} - \Sigma_{ij}| + |\hat{\Sigma}_{ij}|) \mathbb{1}(|\hat{\Sigma}_{ij}| < t, |\Sigma_{ij}| \geq t)$$

$$\leq \max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \sum_{j=1}^{p} \mathbb{1}(|\Sigma_{ij}| \geq t) + t \max_i \sum_{j=1}^{p} \mathbb{1}(|\Sigma_{ij}| \geq t)$$

$$= O_P\left(c_0(p) t^{-q} \delta(n,p) + c_0(p) t^{1-q}\right).$$

To bound I, we pick $\gamma \in (0,1)$ and note that

$$\mathrm{I} \leq \max_i \sum_{j=1}^{p} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \mathbb{1}(|\hat{\Sigma}_{ij}| \geq t, |\Sigma_{ij}| < t) + \max_i \sum_{j=1}^{p} |\Sigma_{ij}| \mathbb{1}(|\Sigma_{ij}| < t)$$

$$\leq \max_i \sum_{j=1}^{p} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \mathbb{1}(|\hat{\Sigma}_{ij}| \geq t, |\Sigma_{ij}| \leq \gamma t)$$

$$+ \max_i \sum_{j=1}^{p} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \mathbb{1}(|\hat{\Sigma}_{ij}| \geq t, \gamma t |\Sigma_{ij}| \leq t) + t^{1-q} c_0(p)$$

$$\leq \max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \left( \sum_{j=1}^{p} \mathbb{1}(|\hat{\Sigma_{ij}} - \Sigma_{ij}| > (1-\gamma)t) + c_0(p)(\gamma t)^{-q} \right) + t^{1-q}c_0(p)$$

$$\leq \max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| c_0(p)(\gamma t)^{-q} + t^{1-q}c_0(p) \text{ (with high probability)}$$

$$= O_P \left( c_0(p)(\gamma t)^{-q} \delta(n,p) + t^{1-q}c_0(p) \right).$$

The second to last inequality follows from the fact that:

$$\mathbb{P} \left( \max_i \sum_{j=1}^{p} \mathbb{1}(|\hat{\Sigma_{ij}} - \Sigma_{ij}| > (1-\gamma)t) > 0 \right) = \mathbb{P} \left( \max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| > (1-\gamma)t \right)$$

$$\leq \frac{1}{p^c}$$

for $t$ as chosen in the statement of the theorem. Combining all these inequalities we get the required result. □

## References

[1] S. Andersson. Invariant normal models. *Annals of Statistics*, 3(1):132–154, 1975. MR0362703

[2] S. Andersson and J. Madsen. Symmetry and lattice conditional independence in a multivariate normal distribution. *Annals of Statistics*, 26, 1998. MR1626059

[3] M. Artin. *Algebra*. Prentice Hall, 1991. MR1129886

[4] C. Bachoc and F. Vallentin. New upper bounds for kissing numbers from semidefinite programming. *Journal of the American Mathematical Society*, 21:909–924, 2008. MR2393433

[5] Y. Bai, E. de Klerk, D. V. Pasechnik, and R. Sotirov. Exploiting group symmetry in truss topology optimization. *Optimization and Engineering*, 10:331–349, 2009. MR2539339

[6] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society: Series B*, 36:192–236, 1974. MR0373208

[7] J. Besag and P. A. Moran. On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, 62:555–562, 1975. MR0391451

[8] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008. MR2485008

[9] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008. MR2387969

[10] S. Boyd, P. Diaconis, P. A. Parrilo, and L. Xiao. Fastest mixing Markov chain on graphs with symmetries. *SIAM Journal on Optimization*, 20:792, 2009.

[11] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. MR2061575

[12] V. V. BULDYGIN AND Y. V. KOZACHENKO. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, 2000. MR1743716

[13] V. CHANDRASEKARAN, P. A. PARRILO, AND A. S. WILLSKY. Latent variable graphical model selection via convex optimization. *Preprint available on* www.arxiv.org/abs/1008.1290v1, 2010.

[14] A. D'ASPREMONT, O. BANERJEE, AND L. EL GHAOUI. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30:56–66, 2007. MR2399568

[15] K. R. DAVIDSON AND S.J. SZAREK. Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*, I:317–366, 2001. MR1863696

[16] E. DE KLERK, D. V. PASECHNIK, AND A. SCHRIJVER. Reduction of symmetric semidefinite programs using the regular ∗-representation. *Mathematical Programming*, 109:613–624, 2007. MR2296566

[17] P. DIACONIS. *Group Representations in Probability and Statistics*, volume 11 of *Lecture notes-monograph series*. Institute of Mathematical Statistics, 1988. MR0964069

[18] D. DUMMIT AND R. FOOTE. *Abstract Algebra*. Wiley, 2004. MR2286236

[19] J. FAN, Y. FAN, AND J. LV. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):43, 2007. MR2472991

[20] A. FÄSSLER AND E. STIEFEL. *Group theoretical methods and their applications*. Birkhäuser, 1992. MR1158662

[21] R. FOOTE, G. MIRCHANDANI, D. ROCKMORE, D. HEALY, AND T. OLSON. A wreath product group approach to signal and image processing: Part I - multiresolution analysis. *IEEE Transactions on Signal Processing*, 48(1):102–132, 2000. MR1736279

[22] W. A. GARDNER, A. NAPOLITANO, AND L. PAURA. Cyclostationarity: Half a century of research. *Signal Processing*, 86(4):639–697, 2006.

[23] K. GATERMANN AND P. A. PARRILO. Symmetry groups, semidefinite programs, and sums of squares. *Journal of Pure and Applied Algebra*, 192(1-3):95–128, 2004. MR2067190

[24] S. HØJSGAARD AND S. LAURITZEN. Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, 70(5):1005–1027, 2008. MR2530327

[25] B. HYLLEBERG, B. JENSEN, AND E. ØRNBØL. Graphical symmetry models. *Master's thesis, University of Aalborg*, 1993.

[26] Y. KANNO, M. OHSAKI, K. MUROTA, AND N. KATOH. Group symmetry in interior point methods for semidefinite programs. *Optimization and Engineering*, 2:293–320, 2001. MR1980276

[27] S. LAURITZEN. *Graphical Models*. Clarendon Press, Oxford, 1996. MR1419991

[28] J. MADSEN. Invariant normal models with recursive graphical Markov structure. *Annals of Statistics*, 28:1150–1178, 2000. MR1810923

[29] N. MEINSHAUSEN AND P. BÜHLMANN. High dimensional graphs and variable selection with the LASSO. *Annals of Statistics*, 34:1436–1462, 2006. MR2278363

[30] J.M.F. MOURA AND S. GOSWAMI. Gauss-Markov random fields (GMrf) with continuous indices. *Information Theory, IEEE Transactions on*, 43(5):1560–1573, Sep 1997. MR1476787

[31] H. Z. MUNTHE-KAAS. On group Fourier analysis and symmetry preserving discretizations of PDEs. *Journal of Physics A: Mathematical and General*, 39(19):5563–5584, 2006. MR2220776

[32] I. OLKIN AND S. J. PRESS. Testing and estimation for a circular stationary model. *Annals of Mathematical Statistics*, 40:1358–1373, 1969. MR0245139

[33] M. PERLMAN. A review of multivariate analysis, comment: Group symmetry covariance models. *Statistical Science*, 2:421–425, 1987.

[34] P. RAVIKUMAR, M. J. WAINWRIGHT, G. RASKUTTI, AND B. YU. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 2011. MR2836766

[35] H. RUE AND L. HELD. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005. MR2130347

[36] A. SCHRIJVER. New code upper bounds from the Terwilliger algebra and semidefinite programming. *IEEE Transactions on Information Theory*, 51:2859–2866, 2005. MR2236252

[37] J. P. SERRE. *Linear Representations of Finite Groups*. Springer, 1977. MR0450380

[38] T. STROHMER. Four short stories about Toeplitz matrix calculations. *Linear Algebra and its Applications*, 343-344(0):321–344, 2002. MR1878948

[39] F. VALLENTIN. Symmetry in semidefinite programs. *Preprint available on* www.arxiv.org/abs/0706.4233v3, 2008. MR2460523

[40] D. F. VOTAW. Testing compound symmetry in a normal multivariate distribution. *Annals of Mathematical Statistics*, 19:447–473, 1948. MR0027999

[41] P. WHITTLE. On stationary processes in the plane. *Biometrika*, 41:439–449, 1954. MR0067450

[42] S. S. WILKS. Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Annals of Mathematical Statistics*, 17:257–281, 1946. MR0017498

[43] A. WILLSKY. Multiresolution markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, 2002.