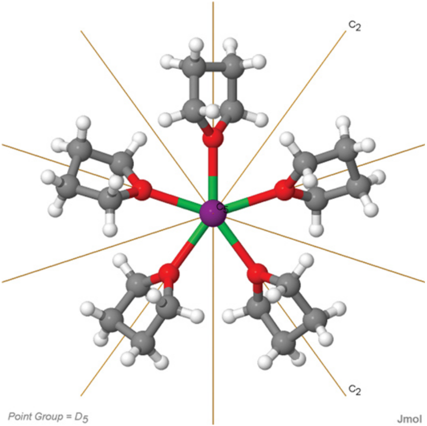
Group Symmetry and Covariance Regularization

Parikshit Shah
University of Wisconsin

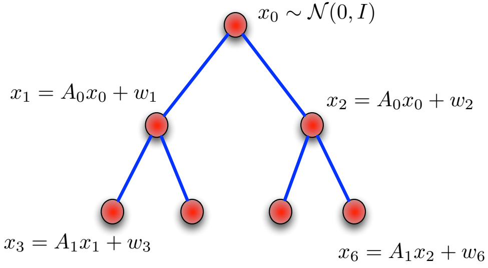
Joint work with Venkat Chandrasekaran

Motivation

- Symmetry is common in science and engineering.
- Symmetry in statistical models.
- How to exploit known group structure?
- Message: Symmetry-aware methods provide huge statistical and computational gains.



Applications: MAR processes



Important class of stochastic models for multi-scale processes, e.g. oceanography, computer vision.

- What is the covariance among the leaf nodes?
- Symmetries: automorphism group of T_d.
- Formally: Σ invariant under action of: $\mathbb{Z}_2$ wr $\mathbb{Z}_2 \ldots$ wr $\mathbb{Z}_2$.
- Can we exploit symmetries? Haar wavelet transform ...

Applications: Random Fields

- Physical phenomena: oceanography, hydrology, electromagnetics
- Poisson's equation (stochastic input):

$$\bigtriangledown^2 \phi(x) = f(x).$$

- Green's function: covariance process $R(x_1, x_2)$.
- Symmetry: Laplacian, boundary conditions, $R(x_1, x_2)$.

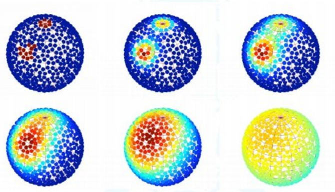


Image courtesy: H. Z. Munthe Kaas, "On Group Fourier Analysis and Symmetry Preserving Discretizations of PDEs"

- Symmetry-preserving discretization.

Other Applications

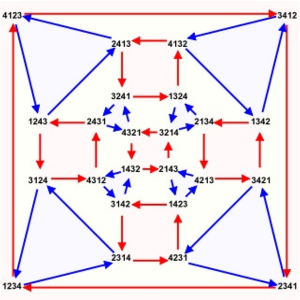
- Partial exchangeability: Clinical Tests
  1. N patients, T groups of similar characteristics
  2. $X_1, \ldots, X_N$ physiological responses
  3. Patients within same group exchangeable (but not i.i.d.)
- Cyclostationarity: periodic phenomena such as vibrations, sinusoidal components ...

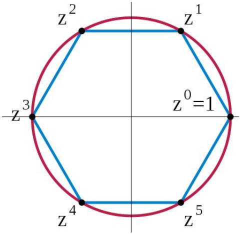
We model symmetry of covariance Σ via $\mathfrak{G}$-invariance.

Problem statement: Given $\mathfrak{G}$ infer information about Σ.

Group Theory: Basics

- Finite group $\mathfrak{G} = (G, \circ)$
  1. G - collection of permutations on $[p]$
  2. $\circ$ - composition
  3. Closure under composition
- Examples
  1. Symmetric group: $\mathcal{S}_p$.
  2. Cyclic group: $\mathbb{Z}/p\mathbb{Z}$.
  3. Cartesian products: $\mathfrak{G}_1 \times \mathfrak{G}_2$.
  4. Other products: semi-direct, wreath.





Group Theory: Group Action

Let $\mathfrak{G}$ be a finite group (of permutation matrices), and $\mathbb{R}_+^{p\times p}$ be PSD matrices. A group action is a map

$$\begin{aligned} \mathcal{A} : & G \times \mathbb{R}_+^{p\times p} \to \mathbb{R}_+^{p\times p} \\ & \left(\Pi_g, \Sigma\right) \mapsto \Pi_g \Sigma \Pi_g^T. \end{aligned}$$

- $\mathfrak{G}$ "acts on" matrices by permuting indices.

Definition

Σ is $\mathfrak{G}$-invariant if

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

- Formalizes notion of a symmetric model.
- Fixed point subspace: $W_{\mathfrak{G}} = \left\{\Sigma : \ \Pi_g \Sigma \Pi_g^T = \Sigma \quad \forall \Pi_g \in \mathfrak{G}\right\}$.

Fixed Point Subspace Projection

- Statistical model $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathbb{R}^{p \times p}$.
- Symmetry: $\Sigma \in W_{\mathfrak{G}}$.
- Model Selection: Given i.i.d. samples $X_1, \ldots, X_n$ recover Σ.

$$\Sigma^n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

- High-D regime ($n \ll p$), Σ^n a poor estimate.
- $\mathfrak{G}$-empirical covariance:

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

- Main contribution: statistical analysis of this estimator.

Fixed Point Subspace Projection

- Statistical model $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathbb{R}^{p \times p}$.
- Symmetry: $\Sigma \in W_{\mathfrak{G}}$.
- Model Selection: Given i.i.d. samples $X_1, \ldots, X_n$ recover Σ.

$$\Sigma^n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

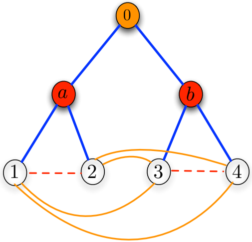
- High-D regime ($n \ll p$), Σ^n a poor estimate.
- $\mathfrak{G}$-empirical covariance:

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

- Main contribution: statistical analysis of this estimator.

Fixed Point Projection: An Example

- MAR Process invariant w.r.t. $\mathbb{Z}_2$ wr $\mathbb{Z}_2 \ldots$ wr $\mathbb{Z}_2$.



(a)

$$\Sigma = \begin{bmatrix} a & b & c & c \\ b & a & c & c \\ c & c & a & b \\ c & c & b & a \end{bmatrix} \quad T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

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

(b)

- How to compute fixed-point subspace projection?
- Use Haar wavelet transform T:

$$\mathcal{P}_{\mathfrak{G}}(\Sigma^n) = T\mathcal{D}\left(T^*\Sigma^n T\right)T^*.$$

Statistical gains: Convergence in spectral norm

- $\|\Sigma - \Sigma^n\| \leq \delta$ w.h.p. if $n = O\left(\frac{p}{\delta^2}\right)$.
- However, $\|\Sigma - \mathcal{P}_{\mathfrak{G}}\left(\Sigma^n\right)\| \leq \delta$ w.h.p. if $n = O\left(\frac{\log p}{\delta^2}\right)$ for $\mathfrak{G} = \mathrm{cyclic}, \mathrm{symmetric}$.
- Proof: Fourier transform diagonalizes circulant matrices.

- How do we generalize?

Group Theory: Representation

- $\mathfrak{G}$-invariant matrices can be simultaneously block diagonalized.

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

$\mathcal{I}$: (active) irreducible representations
s_i: dimension of B_i
m_i: multiplicity of B_i

- **Theorem:** $\|\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n)\| \leq \delta$ w.h.p. provided

$$n = \mathcal{O}\left(\max\left\{\max_{i \in \mathcal{I}} \frac{s_i}{m_i \delta^2}, \max_{i \in \mathcal{I}} \frac{\log p}{m_i \delta^2}\right\}\right).$$

Group Theory: Representation

- $\mathfrak{G}$-invariant matrices can be simultaneously block diagonalized.

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

$\mathcal{I}$: (active) irreducible representations
s_i: dimension of B_i
m_i: multiplicity of B_i

- **Theorem**: $\|\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n)\| \leq \delta$ w.h.p. provided

$$n = \mathcal{O}\left(\max\left\{\max_{i \in \mathcal{I}} \frac{s_i}{m_i \delta^2}, \max_{i \in \mathcal{I}} \frac{\log p}{m_i \delta^2}\right\}\right).$$

Statistical gains: Convergence in ℓ_∞ norm

- $\|\Sigma - \Sigma^n\|_{\ell_\infty} \leq O\left(\sqrt{\frac{\log p}{n}}\right)$.
- $\|\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n)\|_{\ell_\infty} \leq O\left(\sqrt{\frac{\log p}{pn}}\right)$ for $\mathfrak{G} = \text{cyclic}$.
- Proof idea: Reynolds averaging

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

  $\Rightarrow$ Average over edge orbits.
- For cyclic group edge orbits are of size p.

- How do we generalize?

Edge Orbit Parameters

- Combinatorial parameters:

  The edge orbit of (i,j) is $\mathcal{O}(i,j) := \{(g(i), g(j)) \mid g \in \mathfrak{G}\}$.

  The degree d_{ij} is the max. number of times any variable appears in $\mathcal{O}(i,j)$.

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

- **Theorem**: We have w.h.p. that

$$\|\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n)\|_{\ell_\infty} \leq \mathcal{O}\left(\max\left\{\sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d}\right\}\right).$$

- Delicate issues: non-i.i.d. averaging, sample reuse.

Edge Orbit Parameters

- Combinatorial parameters:

  The edge orbit of (i,j) is $\mathcal{O}(i,j) := \{(g(i), g(j)) \mid g \in \mathfrak{G}\}$.

  The degree d_{ij} is the max. number of times any variable appears in $\mathcal{O}(i,j)$.

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

- **Theorem**: We have w.h.p. that

$$\|\Sigma - \mathcal{P}_{\mathfrak{G}}(\Sigma^n)\|_{\ell_\infty} \leq \mathcal{O}\left(\max\left\{\sqrt{\frac{\log p}{n\mathcal{O}}}, \frac{\log p}{n\mathcal{O}_d}\right\}\right).$$

- Delicate issues: non-i.i.d. averaging, sample reuse.

Application: Covariance Estimation

- Covariance Estimation: Bickel-Levina thresholding

$$\hat{\Sigma} := \text{threshold}_t\left(\Sigma^n\right).$$

If Σ has at most d nonzeros per row/column,

$$\|\Sigma - \hat{\Sigma}\| < \sqrt{\frac{d^2 \log p}{n}} \text{w.h.p.}$$

- Symmetry-aware thresholding: Consider $\mathfrak{G} = \text{cyclic}$

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

If Σ has at most d nonzeros per row/column,

$$\|\Sigma - \hat{\Sigma}_{\mathfrak{G}}\| < \sqrt{\frac{d^2 \log p}{pn}} \text{w.h.p.}$$

- Rates in previous slides give results for general groups.

Application: Covariance Estimation

- Covariance Estimation: Bickel-Levina thresholding

$$\hat{\Sigma} := \mathrm{threshold}_t\left(\Sigma^n\right).$$

If Σ has at most d nonzeros per row/column,

$$\|\Sigma - \hat{\Sigma}\| < \sqrt{\frac{d^2 \log p}{n}} \text{w.h.p.}$$

- Symmetry-aware thresholding: Consider $\mathfrak{G} = \mathrm{cyclic}$

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

If Σ has at most d nonzeros per row/column,

$$\|\Sigma - \hat{\Sigma}_{\mathfrak{G}}\| < \sqrt{\frac{d^2 \log p}{pn}} \text{w.h.p.}$$

- Rates in previous slides give results for general groups.

Application: Gaussian Graphical Model Selection

- Zeros of Σ^{-1} encode conditional independence relations.
- ℓ_1-regularized log-l'hood [Yuan and Lin, Ravikumar et al.]:

$$\hat{\Theta} := \arg\min_{\Theta \in \mathcal{S}^p_{++}} \mathrm{tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}.$$

$\hat{\Theta}$, Σ^{-1} have same zero pattern w.h.p. if $n = O\left(d^2 \log p\right)$, where d is degree of graph.

- If Σ is $\mathfrak{G}$-invariant for $\mathfrak{G} = \mathrm{cyclic}$:

$$\hat{\Theta}_{\mathfrak{G}} := \arg\min_{\Theta \in \mathcal{S}^p_{++} \cap W_{\mathfrak{G}}} \mathrm{tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}.$$

$\hat{\Theta}_{\mathfrak{G}}$, Σ^{-1} have same zero pattern w.h.p. if $n = O\left(\frac{d^2 \log p}{p}\right)$.

- Again, rates in previous slides $\Rightarrow$ scaling in general groups.

Application: Gaussian Graphical Model Selection

- Zeros of Σ^{-1} encode conditional independence relations.
- ℓ_1-regularized log-l'hood [Yuan and Lin, Ravikumar et al.]:

$$\hat{\Theta} := \arg\min_{\Theta \in \mathcal{S}^p_{++}} \mathrm{tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}.$$

  $\hat{\Theta}$, Σ^{-1} have same zero pattern w.h.p. if $n = O\left(d^2 \log p\right)$, where d is degree of graph.

- If Σ is $\mathfrak{G}$-invariant for $\mathfrak{G} = \mathrm{cyclic}$:

$$\hat{\Theta}_{\mathfrak{G}} := \arg\min_{\Theta \in \mathcal{S}^p_{++} \cap W_{\mathfrak{G}}} \mathrm{tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}.$$

  $\hat{\Theta}_{\mathfrak{G}}$, Σ^{-1} have same zero pattern w.h.p. if $n = O\left(\frac{d^2 \log p}{p}\right)$.

- Again, rates in previous slides $\Rightarrow$ scaling in general groups.

Computational Gains

- When T known $\mathcal{P}_{\mathcal{G}}(\cdot)$ efficiently computable.
- Exploiting symmetries in convex optimization:

  If objective and constraint functions $\mathfrak{G}$-invariant, then solution in fixed-point subspace.
  $\Rightarrow$ reduction in problem size.
  $\Rightarrow$ improved numerical conditioning.
- For example

$$\underset{\Theta \in \mathcal{S}^p_{++} \cap W_{\mathfrak{G}}}{\arg\min} \quad \mathrm{tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}$$

Computational Gains

- When T known $\mathcal{P}_{\mathcal{G}}(\cdot)$ efficiently computable.
- Exploiting symmetries in convex optimization:

  If objective and constraint functions $\mathfrak{G}$-invariant, then solution in fixed-point subspace.
  $\Rightarrow$ reduction in problem size.
  $\Rightarrow$ improved numerical conditioning.
- For example

$$\underset{\Theta \in \mathcal{S}^p_{++} \cap W_{\mathfrak{G}}}{\arg\min} \quad \mathrm{tr}(\Sigma^n \Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}$$

Computational Gains

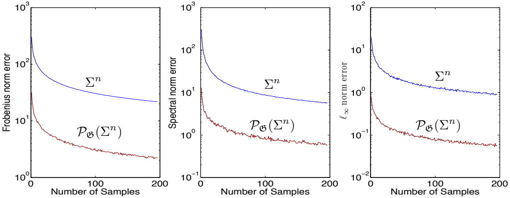
- When T known $\mathcal{P}_{\mathcal{G}}(\cdot)$ efficiently computable.
- Exploiting symmetries in convex optimization:

  If objective and constraint functions $\mathfrak{G}$-invariant, then solution in fixed-point subspace.
  $\Rightarrow$ reduction in problem size.
  $\Rightarrow$ improved numerical conditioning.
- For example

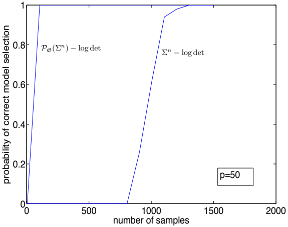
$$\underset{\Theta \in \mathcal{S}_{++}^{p} \cap W_{\mathfrak{G}}}{\arg\min} \; \mathrm{tr}(\Sigma^{n}\Theta) - \log\det(\Theta) + \mu_n \|\Theta\|_{\ell_1}$$

Experiments

Gaussian model invariant with respect to cyclic group, $p = 50$.



Inverse covariance corresponding to a cycle graph, invariant with respect to cyclic group, $p = 50$.



Conclusion

- Statistical models with symmetries.
- Fixed-point projection as means of regularization.
- Improved rates for several model selection and estimation tasks.
- Computational benefits.
- Current efforts: approximately symmetric models.

http://arxiv.org/abs/1111.7061