Identifying and Estimating Dynamical Covariance Matrices with Hierarchical Rank Structure

Robin Armstrong[1], Samuel Otto[2], and Anil Damle[3]

26th Conference of the International Linear Algebra Society
June 24th, 2025

[1] Cornell University, Center for Applied Mathematics
[2] Cornell University, Sibley School of Mechanical and Aerospace Engineering
[3] Cornell University, Department of Computer Science

Covariance Matrices in Dynamical Systems Modeling

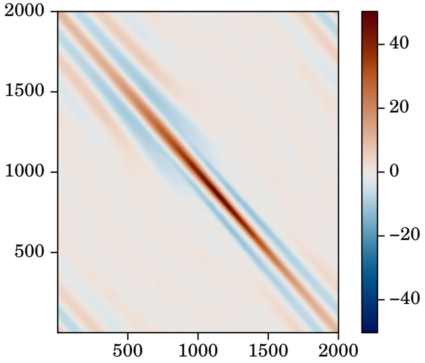
- Consider a dynamical system in n dimensions:

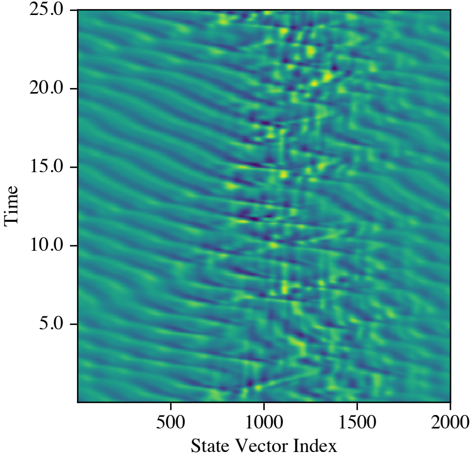
$$\mathbf{x} \in \mathbb{R}^n, \qquad \frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t).$$

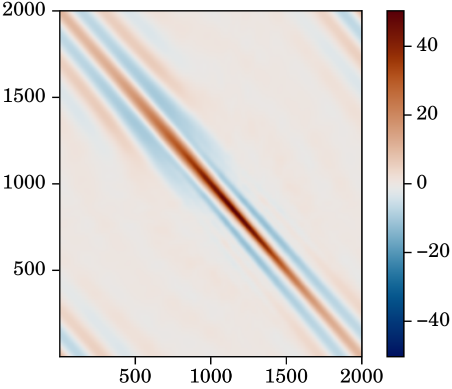
- **Goal:** to estimate the covariance matrix of a measure associated with the dynamics. For example:
  - Model reduction: covariance of the *invariant measure* reveal the appropriate subspace for POD.
  - Data assimilation: covariance of a *prior measure* represents uncertainty about the state, and spreads information from observed variables to unobserved ones.
- We begin with an *ensemble* or *sample* covariance estimate:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{s-1} \sum_{i=1}^{s} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}},$$

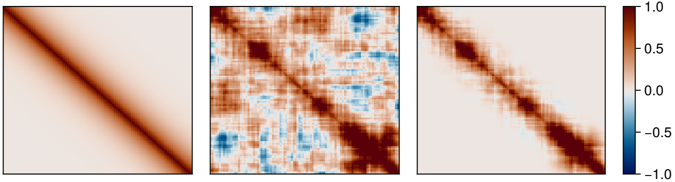
where $\mathbf{x}_1, \ldots, \mathbf{x}_s$ are an ensemble of model states.




25.0
20.0
15.0
10.0
5.0
Time
500
1000
1500
2000
State Vector Index


2000
1500
1000
500
500
1000
1500
2000
40
20
0
−20
−40

Sampling Errors and Localization



- Sample size may be *much* smaller than phase space dimension; $s \ll n$.
- In this case, $\widehat{\boldsymbol{\Sigma}}$ will be unacceptably noisy.
- **Localization** [4] regularizes $\widehat{\boldsymbol{\Sigma}}$ by attenuating “unphysical” correlations:

$$\boldsymbol{\Sigma} = \mathbf{L} \circ \widehat{\boldsymbol{\Sigma}},$$

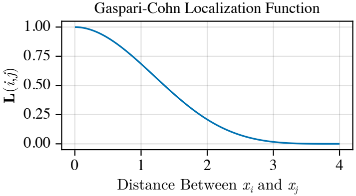
where $\circ$ = element-wise product, $\mathbf{L}$ = a symmetric localizing matrix in $[0, 1]^{n \times n}$.

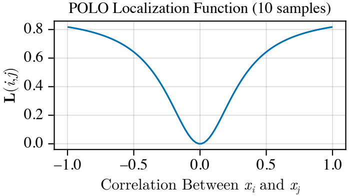
Localization Functions

- **Distance-based localization:** correlations are attenuated based on spatial separation between variables.
  - Admits a data-sparse matrix representation.
  - Encodes *a priori* physical structure.
  - Careful tuning needed for localization radius.
- **Prior optimal localization (POLO):** [6] small empirical correlations are attenuated more than large ones.

$$\mathcal{L}(\rho) = \frac{(s-1)\rho^2}{1+s\rho^2}.$$

  - No tuning needed.
  - Adapts to sample size.
  - **Optimal for multivariate Gaussian samples** [6].
  - **Expensive to construct for high-dimensional systems**.

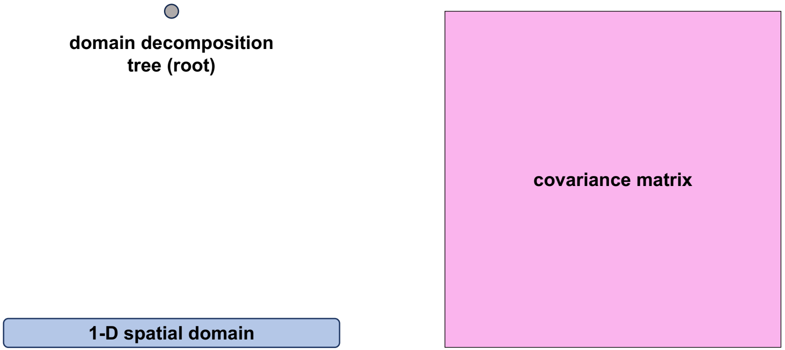




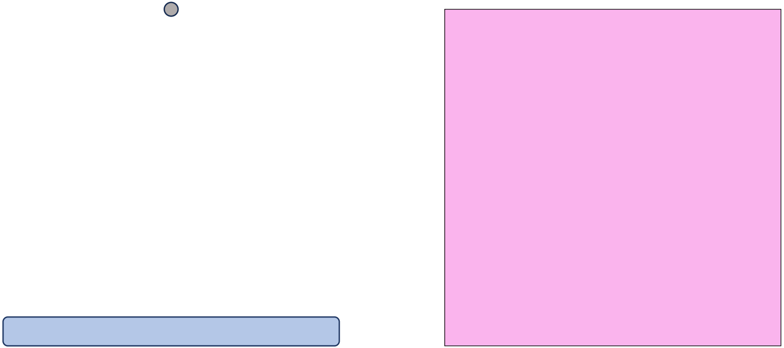
Our Contributions

- **The problem we address:** POLO localization does not produce a data-sparse matrix.
  - ...at least, not without $\mathcal{O}(n^2)$ work to create sparsity!
- **Our solution:** represent the POLO estimator as a hirarchically rank-structured matrix [3].
  - Spatial domain is recursively partitioned into nested subdomains.
  - At each level, admissible blocks (correlations between sufficiently far-apart subdomains) are compressed to low-rank form.
- **The challenge we face:** we must create a hierarchically rank structured representation by only looking at our samples.
  - We are not allowed to form the dense-matrix POLO estimator first!

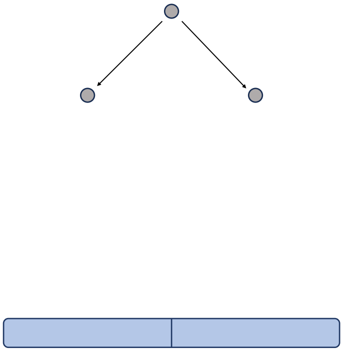
Compression Using Hierarchical Rank Structure

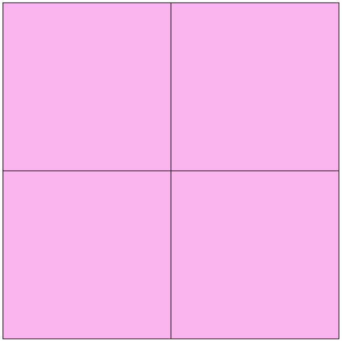


Compression Using Hierarchical Rank Structure

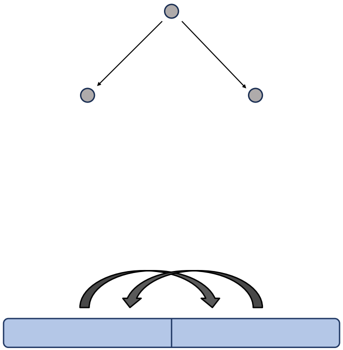


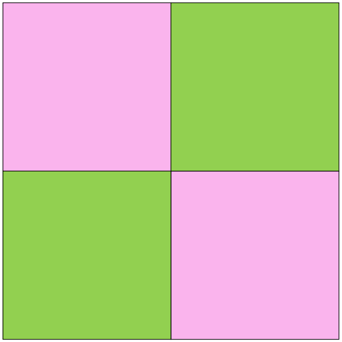
Compression Using Hierarchical Rank Structure



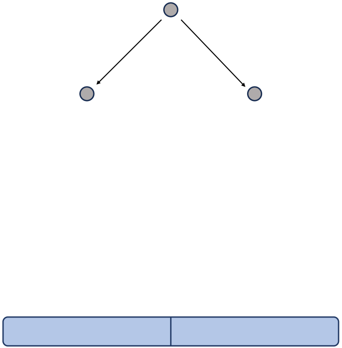


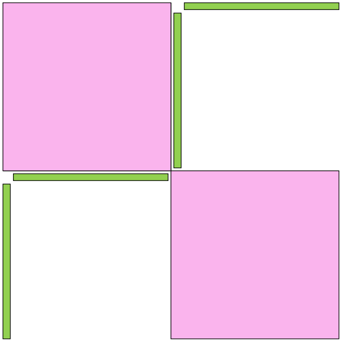
Compression Using Hierarchical Rank Structure



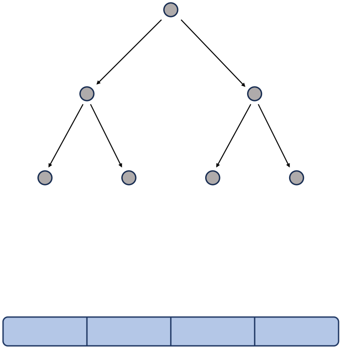


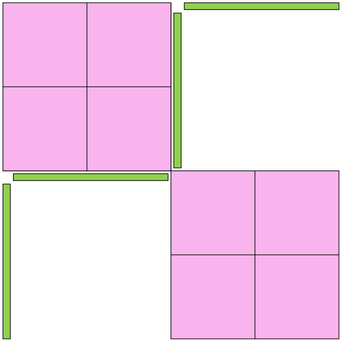
Compression Using Hierarchical Rank Structure



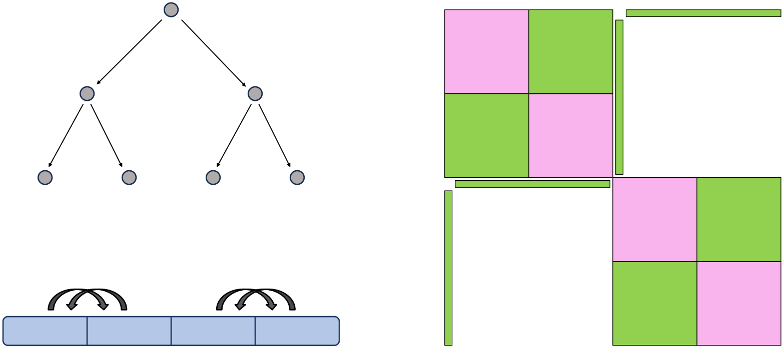


Compression Using Hierarchical Rank Structure

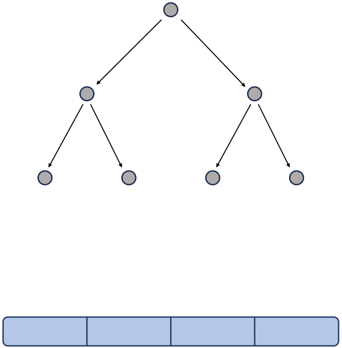


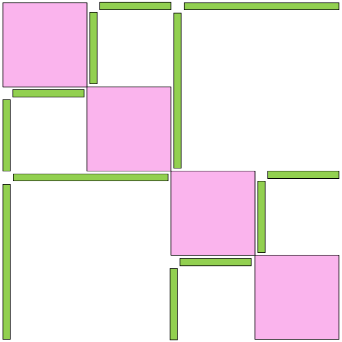


Compression Using Hierarchical Rank Structure

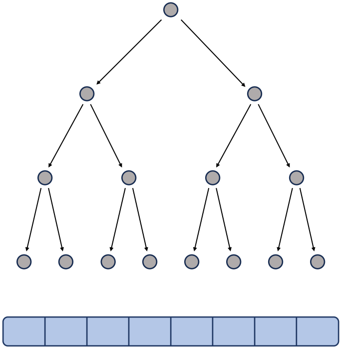


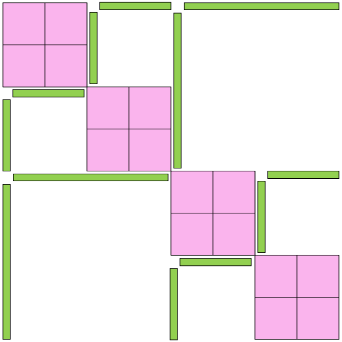
Compression Using Hierarchical Rank Structure



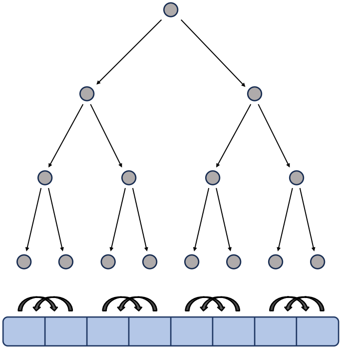


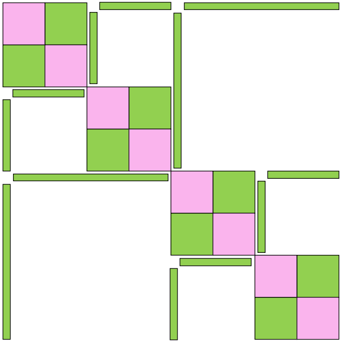
Compression Using Hierarchical Rank Structure



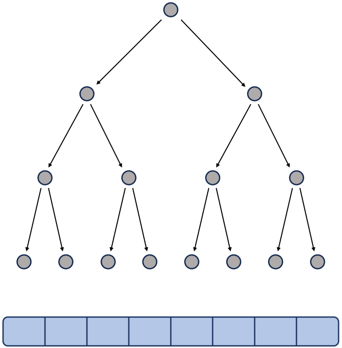


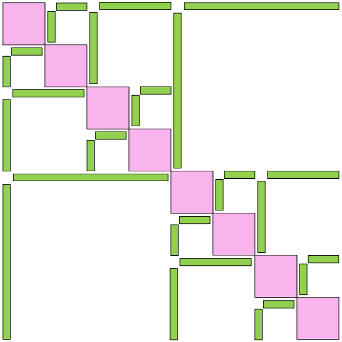
Compression Using Hierarchical Rank Structure





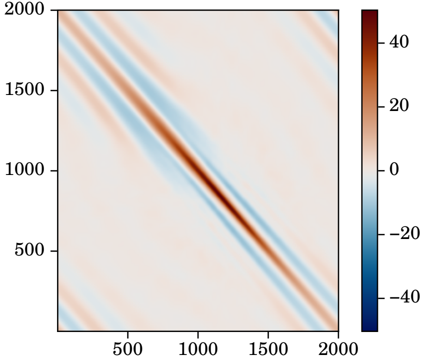
Compression Using Hierarchical Rank Structure



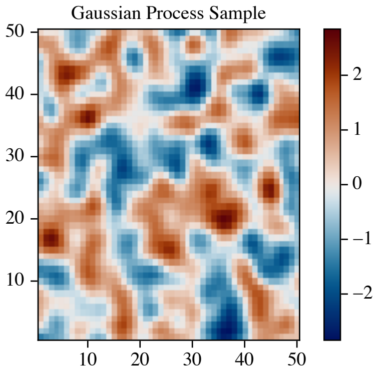


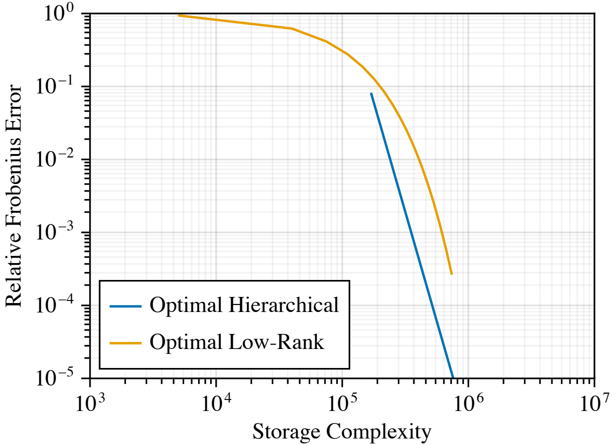
Is Hierarchical Rank Structure Appropriate Here?

- **Informally:** hierarchical rank structure is appropriate *when long-distance correlations vary more smoothly than short-distance correlations.*
- **More Formally:** kernels that are *asymptotically smooth* [3] (correlations and their derivatives decay algebraically away from the diagonal) have hierarchical rank structure.
- Many details to consider, such as:
  - What is the procedure for recursively partitioning space?
  - What is the criterion for determining which blocks are admissible (can be compressed)?
- **Examples of hierarchical rank structure:**
  - Greens functions associated with elliptic operators.
  - **Diffusion-based covariance models used in numerical weather prediction** [8].



Hierarchical Rank Structure vs Low Rank Structure





Estimating Admissible Blocks: Overview

- **Notation:** $\widehat{\boldsymbol{\Sigma}}$ = the unlocalized covariance matrix, $(\mathcal{X}, \mathcal{Y})$ = an admissible subdomain pair, and

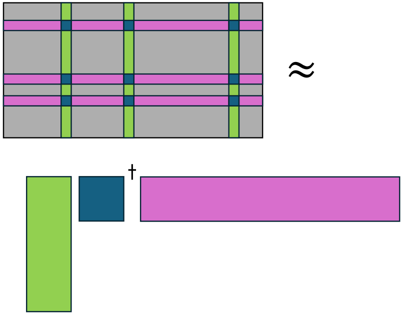
$$\mathcal{L}(\rho) = \frac{(s-1)\rho^2}{1 + s\rho^2} = \text{POLO localizer for } s \text{ samples.}$$

- **Goal:** form an approximate low-rank factorization of a localized admissible block,

$$\boldsymbol{\Sigma}(\mathcal{X}, \mathcal{Y}) = \mathcal{L}(\widehat{\boldsymbol{\Sigma}}(\mathcal{X}, \mathcal{Y})) \approx \mathbf{L}_{\mathcal{X}} \mathbf{L}_{\mathcal{Y}}^{\mathrm{T}},$$

using as few entry evaluations in $\boldsymbol{\Sigma}(\mathcal{X}, \mathcal{Y})$ as possible.

- **Approach:** a generalized Nyström approximation [7]. The challenge is to select good skeleton rows/columns without looking "too hard" at $\boldsymbol{\Sigma}(\mathcal{X}, \mathcal{Y})$.



Column/Row Selection via Proxy Points

- We want to choose a skeleton column set $\mathcal{S} \subseteq \mathcal{Y}$, $|\mathcal{S}| = r \ll |\mathcal{Y}|$, such that

$$\sigma_{\min}(\mathbf{\Sigma}(\mathcal{X}, \mathcal{S})) \approx \max_{\mathcal{T} \subseteq \mathcal{Y}, |\mathcal{T}|=r} \sigma_{\min}(\mathbf{\Sigma}(\mathcal{X}, \mathcal{T})).$$

- **Column-pivoted QR (CPQR)** factorization provides a good solution,

$$\begin{bmatrix} \mathbf{\Sigma}(\mathcal{X}, \mathcal{S}) & \mathbf{\Sigma}(\mathcal{X}, \mathcal{Y} \setminus \mathcal{S}) \end{bmatrix} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_2 \end{bmatrix},$$

  but requires evaluating all the entries in $\mathbf{\Sigma}(\mathcal{X}, \mathcal{Y})$.
- **Proxy point method:** select a set of **proxy points** [9] $\mathcal{P} \subseteq \mathcal{X}$ with $|\mathcal{P}| \ll |\mathcal{X}|$. Pivot on $\mathbf{\Sigma}(\mathcal{P}, :)$ instead:

$$\begin{bmatrix} \mathbf{\Sigma}(\mathcal{P}, \mathcal{S}) & \mathbf{\Sigma}(\mathcal{P}, \mathcal{Y} \setminus \mathcal{S}) \end{bmatrix} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_2 \end{bmatrix}.$$

- How to choose proxy points without looking at $\mathbf{\Sigma}(\mathcal{X}, \mathcal{Y})$?

Proxy Point Selection

- Proxy points should correspond to rows that capture the rank structure of Row $\mathbf{\Sigma}(\mathcal{X}, \mathcal{Y})$.
- **Use CPQR on rows of $\widehat{\mathbf{\Sigma}}(\mathcal{X}, \mathcal{Y})$ (unlocalized).**
- If $\mathbf{Z}_{\mathcal{X}}$, $\mathbf{Z}_{\mathcal{Y}}$ are centered, normalized samples restricted to $\mathcal{X}$ and $\mathcal{Y}$, then

$$\widehat{\mathbf{\Sigma}}(\mathcal{X}, \mathcal{Y}) = \mathbf{Z}_{\mathcal{X}} \mathbf{Z}_{\mathcal{Y}}^{\mathrm{T}} = (n_{\mathcal{X}} \times s) \cdot (s \times n_{\mathcal{Y}}),$$

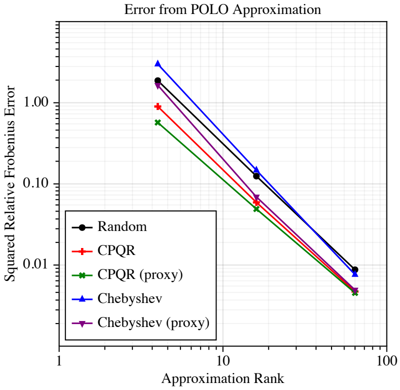
where $s \ll \min\{n_{\mathcal{X}}, n_{\mathcal{Y}}\}$.

Lemma

CPQR factorization on $\mathbf{A}_1 = \mathbf{Z}_{\mathcal{Y}} \mathbf{Z}_{\mathcal{X}}^{\mathrm{T}}(:, I)$ yields an identical column permutation as CPQR factorization on

$$\mathbf{A}_2 = (\mathbf{Z}_{\mathcal{Y}}^{\mathrm{T}} \mathbf{Z}_{\mathcal{Y}})^{1/2} \mathbf{Z}_{\mathcal{X}}^{\mathrm{T}}(:, I).$$

Note that $\mathbf{A}_1$ is $n_{\mathcal{Y}} \times n_{\mathcal{X}}$, while $\mathbf{A}_2$ is only $s \times n_{\mathcal{X}}$.



Full Estimation Procedure

Algorithm 1 Estimating Low-Rank Factors Of An Admissible Block

1: **input:** $\mathbf{Z}_{\mathcal{X}} \in \mathbb{R}^{n_{\mathcal{X}} \times s}$, $\mathbf{Z}_{\mathcal{Y}} \in \mathbb{R}^{n_{\mathcal{Y}} \times s}$ (centered and normalized sample vectors on $\mathcal{X}$ and $\mathcal{Y}$)
2: **input:** k, $p \geq 1$ (admissible block rank and skeletonization rank)

3: $\mathcal{P}_1 \leftarrow \texttt{CPQRColumnSelect}((\mathbf{Z}_{\mathcal{Y}}^{\mathrm{T}}\mathbf{Z}_{\mathcal{Y}})^{1/2}\mathbf{Z}_{\mathcal{X}}^{\mathrm{T}}, p)$ *# selecting p skeleton columns*
4: $\mathcal{S}_{\mathcal{Y}} \leftarrow \texttt{CPQRColumnSelect}(\mathbf{\Sigma}(\mathcal{P}_1, :), p)$
5: $\mathcal{P}_2 \leftarrow \texttt{CPQRColumnSelect}((\mathbf{Z}_{\mathcal{X}}^{\mathrm{T}}\mathbf{Z}_{\mathcal{X}})^{1/2}\mathbf{Z}_{\mathcal{Y}}^{\mathrm{T}}, p)$ *# selecting p skeleton rows*
6: $\mathcal{S}_{\mathcal{X}} \leftarrow \texttt{CPQRColumnSelect}(\mathbf{\Sigma}(:, \mathcal{P}_2)^{\mathrm{T}}, p)$
7: $\mathbf{L}_1\mathbf{L}_2^{\mathrm{T}} \leftarrow \mathbf{\Sigma}(:, \mathcal{S}_{\mathcal{Y}})\mathbf{\Sigma}(\mathcal{S}_{\mathcal{X}}, \mathcal{S}_{\mathcal{Y}})^{\dagger}\mathbf{\Sigma}(\mathcal{S}_{\mathcal{X}}, :)$ *# rank-p generalized Nyström approximation*

8: $\mathbf{Q}_1, \mathbf{R}_1 \leftarrow \texttt{ThinQR}(\mathbf{L}_1)$, $\mathbf{Q}_2, \mathbf{R}_2 \leftarrow \texttt{ThinQR}(\mathbf{L}_2)$ *# optimal reduction to rank k*
9: $\mathbf{U}, \mathbf{\Gamma}, \mathbf{V} \leftarrow \texttt{SVD}(\mathbf{R}_1\mathbf{R}_1^{\mathrm{T}})$
10: $\mathbf{B}_1 \leftarrow \mathbf{Q}_1\mathbf{U}(:, 1:r)\mathbf{\Gamma}(1:r, 1:r)^{1/2}$
11: $\mathbf{B}_2 \leftarrow \mathbf{Q}_2\mathbf{V}(:, 1:r)\mathbf{\Gamma}(1:r, 1:r)^{1/2}$

12: **return** $\mathbf{B}_1, \mathbf{B}_2$

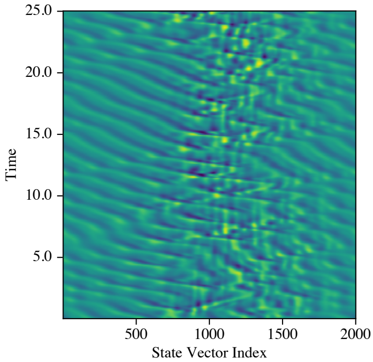
Test Case 1: "Storm Track" Dynamics

- A modification of "model II" from Lorenz [5]. Like the Lorenz '96 model, but:
  - admits waves much larger than the grid spacing, and
  - has a "stable" region of strong damping and a "chaotic" region of weak damping.

  Based off a system in [1].
- **Domain:** 2000 grid points in 1D with periodic boundary conditions.
- **Partition:** recursive bisection until domain has at most 10 grid points.
- **Admissibility criterion:**

$$\min\{\ell(\mathcal{X}), \ell(\mathcal{Y})\} \leq d(\mathcal{X}, \mathcal{Y}),$$

where $\ell(\cdot)$ = domain length, and $d(\mathcal{X}, \mathcal{Y}) = \inf_{x\in\mathcal{X}, y\in\mathcal{Y}} |x - y|$.



Test Case 2: 2D Gaussian Process

- **“Ground-truth” covariance**:

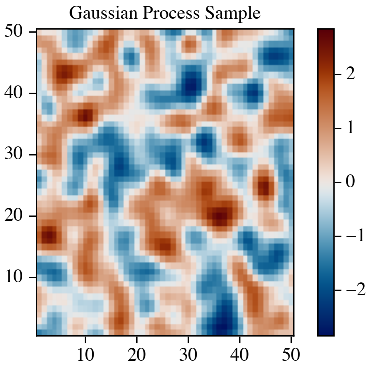
$$\boldsymbol{\Sigma}_0((x_i, y_i), (x_j, y_j)) = \exp\left(-\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{2\sigma^2}\right),$$

where $\sigma = 0.05$ and (x_i, y_i) are nodes of a 50×50 uniform grid on $[0, 1] \times [0, 1]$.

- Samples in $\mathbb{R}^{2500}$ drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.
- **Partition:** recursive bisection of rectangles along longer axis, until max. sidelength is below 0.2.
- **Admissibility criterion:**

$$\min\{\ell(\mathcal{X}), \ell(\mathcal{Y})\} \leq d(\mathcal{X}, \mathcal{Y}),$$

where $\ell(\cdot) =$ max sidelength of rectangle, and $d(\mathcal{X}, \mathcal{Y}) = \inf_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} \|\mathbf{x} - \mathbf{y}\|_2$.



Test Case 3: 2D Quasigeostrophic Turbulence

- **Quasigeostrophic flow** approximates the motion of a rotating fluid where Coriolis and pressure-gradient forces are nearly in balance [2].
- Commonly used as a simplified model of atmospheric flow.
- Covariance has *mild spatial inhomogeneity.*
- **Domain:** 128×128 grid on a 2D square with periodic boundary conditions.
- **Partition:** bisecting rectangles until longest side spans at most 10 gridpoints.
- **Admissibility Criterion:** same as before.
- Simulated with code from `https://github.com/jswhit/sqgturb`.

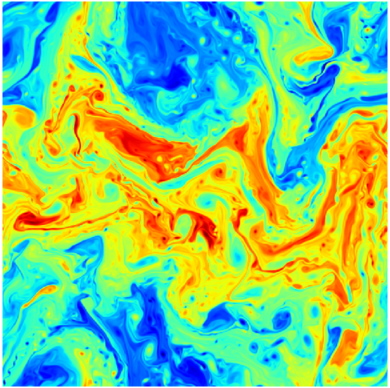
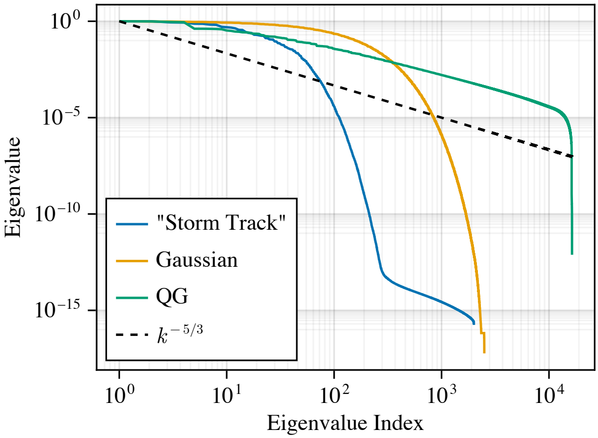
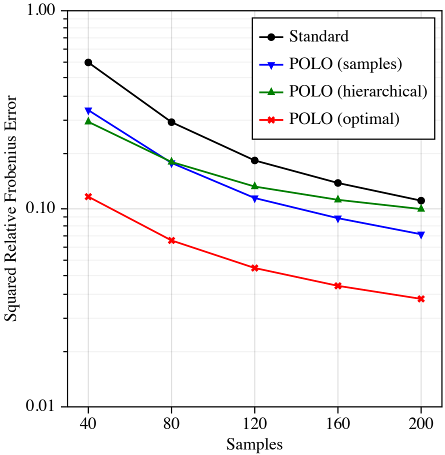


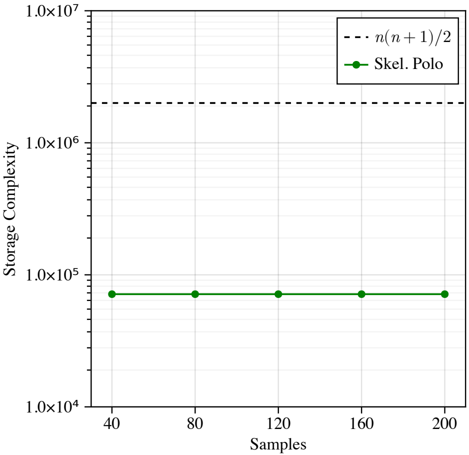
Figure: from `https://github.com/jswhit/sqgturb`.

Problem Difficulty

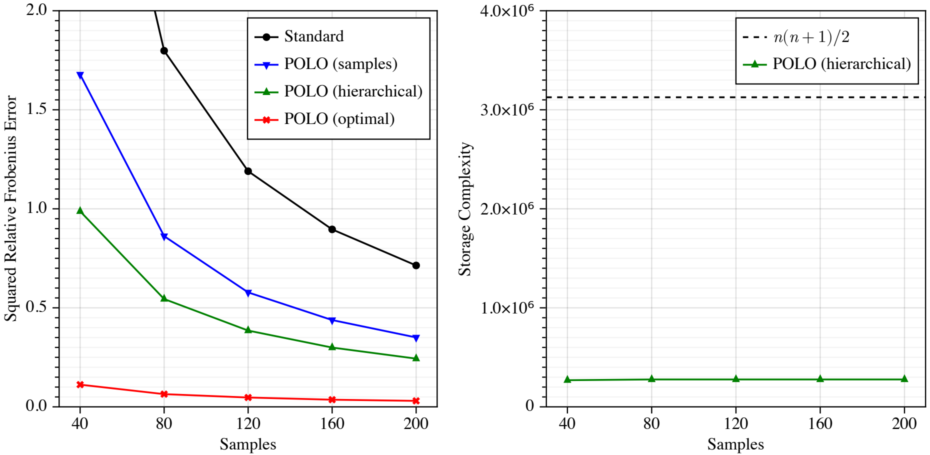


Results: "Storm Track" Dynamics

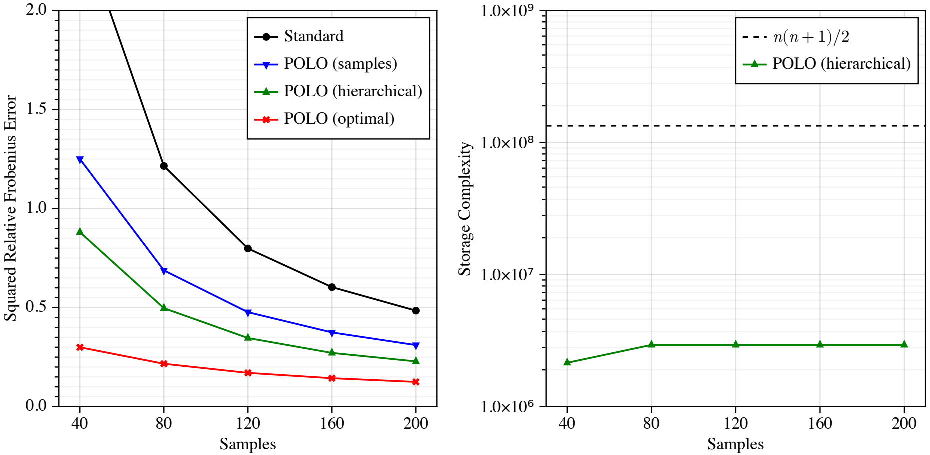




Results: Gaussian Process



Results: Quasigeostrophic Turbulence



Conclusions

- Localization is an essential aspect of covariance matrix estimation when very few samples are available.
- The POLO localizer has optimality properties and requires no tuning, but expensive to construct and store.
- **We have developed:** *a data-sparse, efficiently constructable covariance estimator that corresponds to a hierarchically rank-structured approximation of POLO localization.*
- **Next steps:**
  1. More complex test cases, including 3D spatial domains.
  2. **Positive-definite estimators**,which POLO itself is not.
  3. Testing performance in model reduction and data assimilation problems.

References I

Craig H. Bishop, Jeffrey S. Whitaker, and Lili Lei.
Gain form of the ensemble transform Kalman filter and its relevance to satellite data assimilation with model space ensemble covariance localization.
Monthly Weather Review, 145(11):4575 – 4592, 2017.

Roger Daley.
Atmospheric Data Analysis.
Cambridge Atmosphere and Space Science Series. Cambridge University Press, 1991.

Wolfgang Hackbush.
Hierarchical Matrices: Algorithms and Applications.
Springer, 2015.

Thomas M. Hamill, Jeffrey S. Whitaker, and Chris Snyder.
Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter.
Monthly Weather Review, 129(11), 2001.

References II

Edward N. Lorenz.
Designing chaotic models.
Journal of the Atmospheric Sciences, 62(5):1574 – 1587, 2005.

David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar, and Daniel Hodyss.
High-dimensional covariance estimation from a small number of samples.
Journal of Advances in Modeling Earth Systems, 16(9):e2024MS004417, 2024. e2024MS004417 2024MS004417.

Yatian Wang, Hua Xiang, Chi Zhang, and Songling Zhang.
A generalized Nystrom method with column sketching for low-rank approximation of nonsymmetric matrices, 2024.

Anthony Weaver and Philippe Courtier.
Correlation modelling on the sphere using a generalized diffusion equation.
Quarterly Journal of the Royal Meteorological Society, 127(575):1815–1846, 2001.

References III

Xin Xing and Edmond Chow.
Interpolative decomposition via proxy points for kernel matrices.
SIAM Journal on Matrix Analysis and Applications, 41(1):221–243, 2020.