Sum of Kronecker Products Representation and Its Cholesky Factorization for Spatial Covariance Matrices from Large Grids

Jian Cao

King Abdullah University of Science and Technology

Joint work with: Prof. Marc G. Genton, Prof. David Keyes, and Prof. George Turkiyyah

Mar 1, 2021

Motivation

Several compressed representations for matrices (among others):

- Sparse matrices (record only non-zero coefficients)
- Hierarchical matrices ($\mathcal{H}$)
- Hierarchical semiseparable matrices (HSS)

Both $\mathcal{H}$-matrices (Hackbusch, 2015) and HSS matrices (Chandrasekaran et al., 2005) can be interpreted as utilizing the row/column-wise similarity

Can we utilize the **block-wise similarity**? If so, it would be the sum-of-Kronecker-product (SKP) representation

For example, the covariance matrix from a stationary kernel on a grid is block-Toeplitz with Toeplitz blocks if indexed along the column/row

Construction of the SKP Representation

The nearest Kronecker product problem, $\text{argmin}_{\mathbf{U},\mathbf{V}} \|\mathbf{\Sigma} - \mathbf{U} \otimes \mathbf{V}\|_F$, was discussed in Van Loan and Pitsianis (1993). An algorithm for finding $\mathbf{U}$ and $\mathbf{V}$ is through the singular value decomposition (SVD) of the rearranged $\mathbf{\Sigma}$, denoted by $\tilde{\mathbf{\Sigma}}$

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \Rightarrow \tilde{\mathbf{\Sigma}} = \begin{bmatrix} \text{vec}(\mathbf{\Sigma}_{11})^\top \\ \text{vec}(\mathbf{\Sigma}_{21})^\top \\ \text{vec}(\mathbf{\Sigma}_{12})^\top \\ \text{vec}(\mathbf{\Sigma}_{22})^\top \end{bmatrix}$$

The singular vectors corresponding to the largest singular value of $\tilde{\mathbf{\Sigma}}$ are matricized into $\mathbf{U}$ and $\mathbf{V}$

Construction of the SKP Representation Continued

The SKP representation, $\sum_{i=1}^{k} \mathbf{U}_i \otimes \mathbf{V}_i$, is a natural extension if the singular vectors corresponding to the top k singular values of $\tilde{\mathbf{\Sigma}}$ are matricized

However, the construction with SVD has a complexity of $O(n^3)$, where n is the dimension of $\mathbf{\Sigma}$. Potential improvements include:

- Randomized SVD $\rightarrow O(n^2)$ if $k \ll n$
- Adaptive-cross-approximation (ACA) $\rightarrow O(n)$ if $k \ll n$

Here, the more greedy algorithm is chosen to maximize efficiency. However, randomized SVD would typically be the minor computation cost in the whole procedure

Construction of the SKP Representation Continued

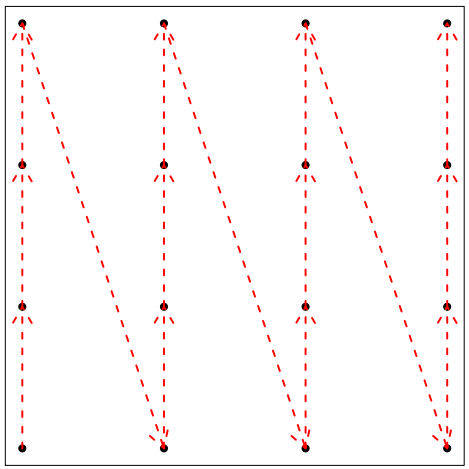
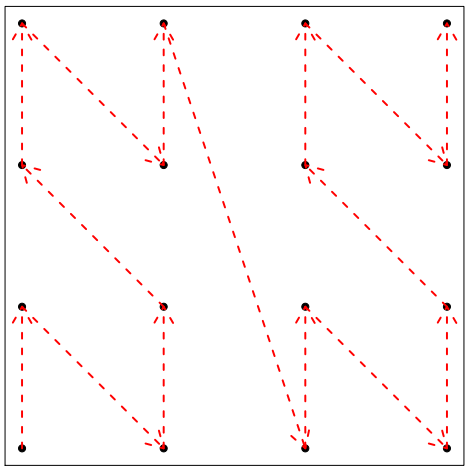
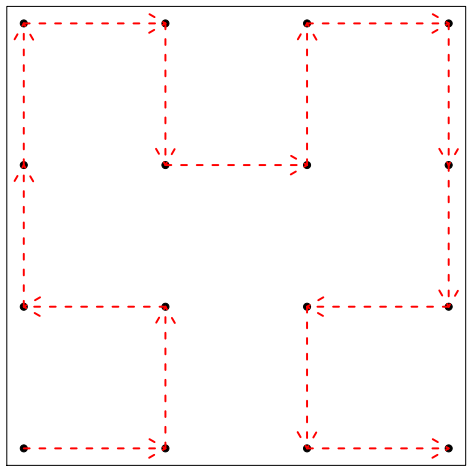
We use covariance matrices from four $s_1 \times s_2$ regular grids in the unit square and an exponential kernel, $\exp(-\|\mathbf{h}\|/\beta), \beta = 0.3$ to compare the SVD-based decomposition and the ACA-based decomposition. The dimensions of $\{\mathbf{V}_i\}$ are $s_2 \times s_2$

Table: Construction of the SKP representations for covariance matrices using SVD and ACA. Error is measured by $\|\mathbf{\Sigma} - \sum_{i=1}^{k} \mathbf{U}_i \otimes \mathbf{V}_i\|_F / \|\mathbf{\Sigma}\|_F$ and time is measured in seconds

(s_1, s_2)	(64, 64)	(128, 128)	(128, 256)	(256, 256)
Method	SVD | ACA	SVD | ACA	SVD | ACA	SVD | ACA
Error	5.5e−6 | 2.8e−5	1.4e−6 | 8.8e−6	*NA* | 1.4e−5	*NA* | 4.1e−6
k	9 | 9	11 | 11	*NA* | 11	*NA* | 13
Time	4.4e+1 | 1.4e−2	2.7e+3 | 3.7e−2	*NA* | 6.0e−2	*NA* | 2.7e−1

Indexing for the SKP Representation

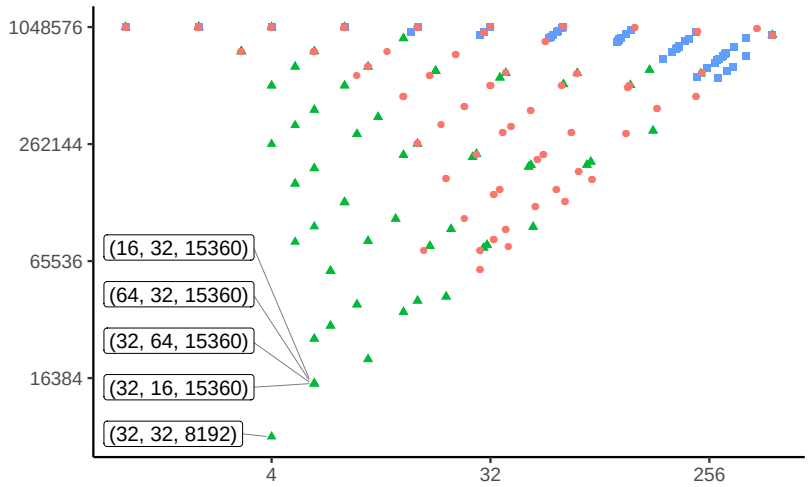
Three ordering methods are experimented to find the one most aligned with the SKP representation, namely the *y*-major order, the Morton's order, and the Hilbert curve order, from the left to the right

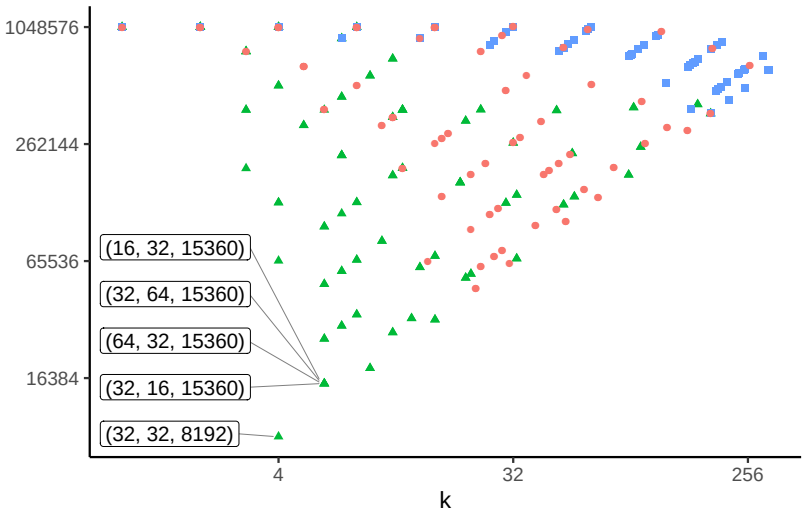




Unlike hierarchical matrices and HSS matrices, **preserving locality is not strictly necessary**

Indexing for the SKP Representation Continued

The storage costs for both stationary, $\exp(-\|\mathbf{h}\|/\beta_0)$, and non-stationary, $\beta^{1/2}(\mathbf{s}_i)\beta^{1/2}(\mathbf{s}_j)\left\{\frac{\beta^2(\mathbf{s}_i)+\beta^2(\mathbf{s}_j)}{2}\right\}^{-1/2}\exp\left(-\sqrt{\frac{2\mathbf{h}^2}{\beta^2(\mathbf{s}_i)+\beta^2(\mathbf{s}_j)}}\right)$, kernels are compared under these three orders. The geometry used is a 32×32 grid in the unit square. Here, $\{\mathbf{s}_i\}_{i=1}^n$ denotes the spatial locations

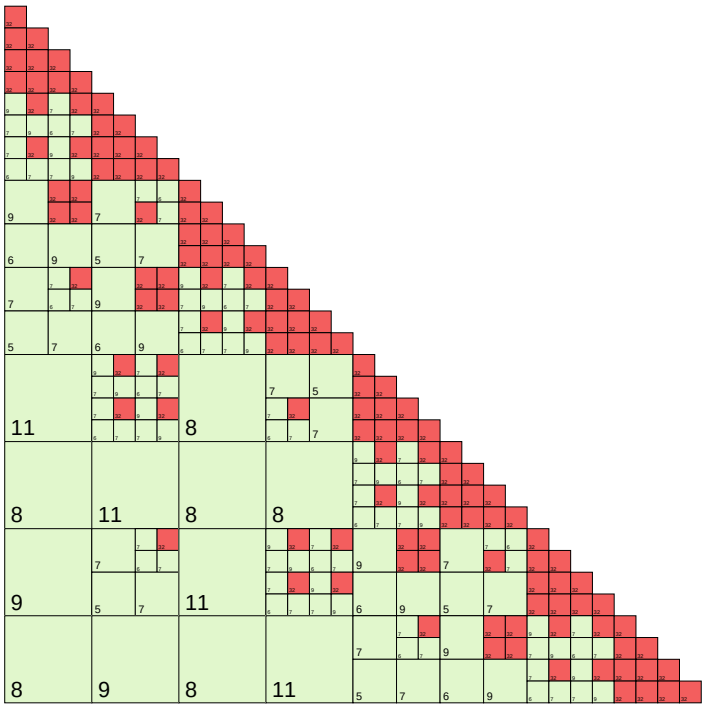


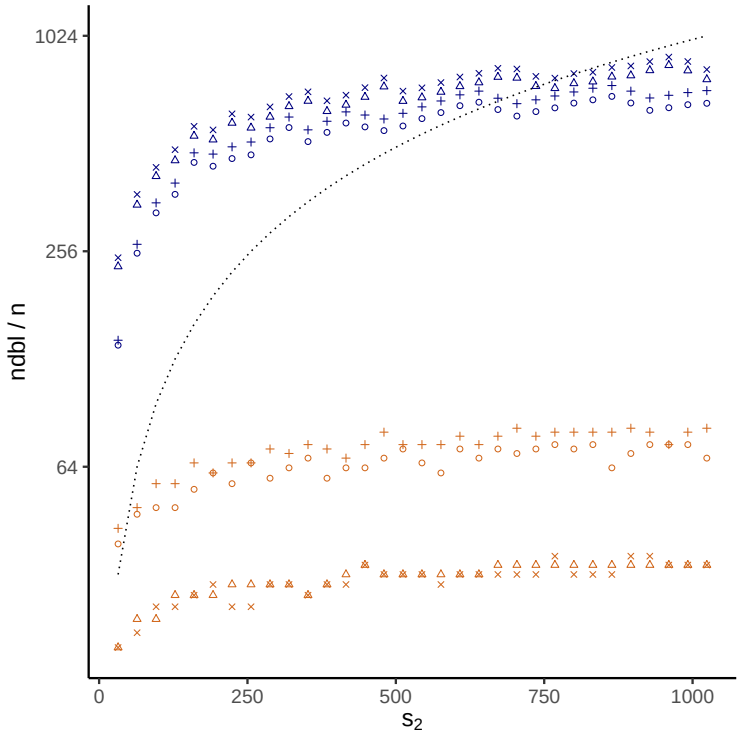


The y-axis is the number of float numbers used, denoted by *ndbl*. Letting $(m_2, n_2) = \dim(\mathbf{V}_i)$, the five points in each plot with the smallest *ndbl* is annotated by $(m_2, n_2, ndbl)$. **y-major order** is the most aligned

Storage Cost Comparison – Regular Grid

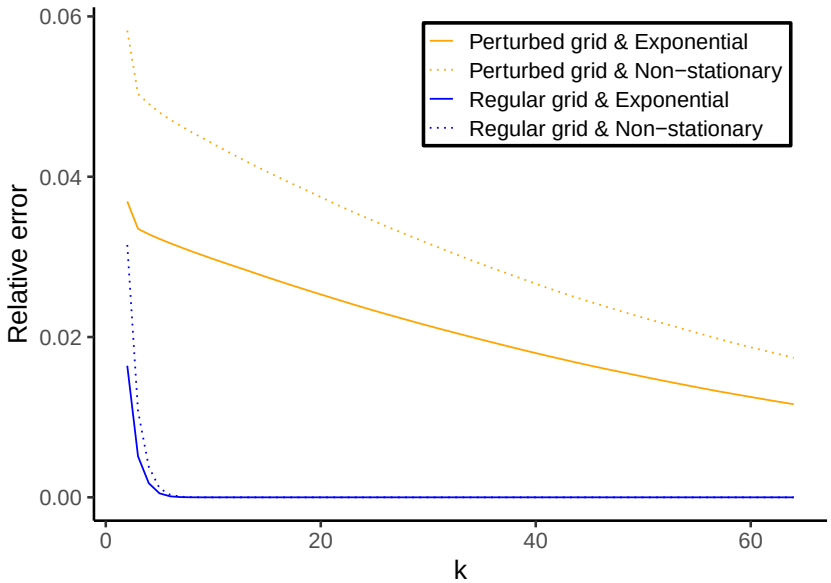
Use $\mathcal{H}$-matrix (illustrated on the left) and SKP to approximate to the same accuracy, 10^{-5}, under 4 scenarios, i.e., rectangular/square grid and stationary/non-stationary kernel. n is proportional to s_2. The storage costs of SKP (top 4 dotted curves) are **significantly smaller** than $\mathcal{H}$-matrices (bottom 4 dotted curves)

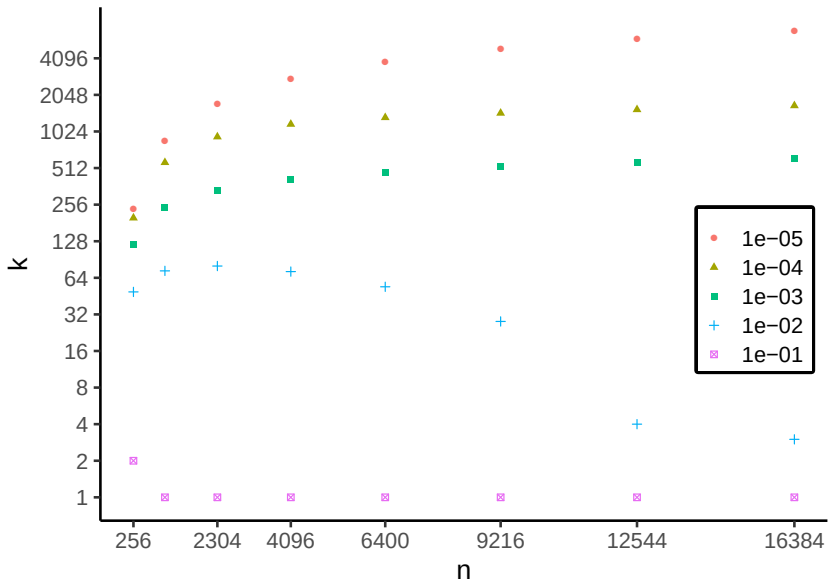




Storage Cost Comparison – Perturbed Grid

However, the approximation error **decreases slowly** with k on a 32×32 perturbed grid and the k needed to reach a high accuracy **increases fast** with n





The SKP is **not suitable** for irregular geometries

Cholesky Factorization under SKP

Overarching idea: block Cholesky factorization (Akbudak et al., 2017) that uses a **coordinate system** and **discovers new base blocks** dynamically

Three SKP representations:

$$\mathbf{\Sigma} \approx \sum_{i=1}^{k^{\mathbf{\Sigma}}} \mathbf{U}_i^{\mathbf{\Sigma}} \otimes \mathbf{V}_i^{\mathbf{\Sigma}}, \quad \mathbf{L} \approx \sum_{i=1}^{k^{\mathbf{L}}} \mathbf{U}_i^{\mathbf{L}} \otimes \mathbf{V}_i^{\mathbf{L}}, \quad \mathbf{D} \approx \sum_{i=1}^{k^{\mathbf{D}}} \mathbf{U}_i^{\mathbf{D}} \otimes \mathbf{V}_i^{\mathbf{D}}$$

where $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$ and $\mathbf{D} = \mathbf{L}^{-1}$. Given $\sum_{i=1}^{k^{\mathbf{\Sigma}}} \mathbf{U}_i^{\mathbf{\Sigma}} \otimes \mathbf{V}_i^{\mathbf{\Sigma}}$, the goal is to compute $\sum_{i=1}^{k^{\mathbf{L}}} \mathbf{U}_i^{\mathbf{L}} \otimes \mathbf{V}_i^{\mathbf{L}}$, where $\sum_{i=1}^{k^{\mathbf{D}}} \mathbf{U}_i^{\mathbf{D}} \otimes \mathbf{V}_i^{\mathbf{D}}$ is a middle product

$\{\mathbf{U}_i\}$ are interpreted as the **coordinates** while $\{\mathbf{V}_i\}$ are interpreted as the **base blocks**

Cholesky Factorization under SKP Continued

Two sets of matrix multiplications, relative to which the coordinates of the base blocks are defined:

$$\mathcal{S} = \{\mathbf{V}_{l_1}^{\mathbf{\Sigma}}\mathbf{V}_{l_2}^{\mathbf{D}\top}, l_1 = 1, \ldots, k^{\mathbf{\Sigma}}, l_2 = 1, \ldots, k^{\mathbf{D}}\}$$
$$\mathcal{H} = \{\mathbf{V}_{l_1}^{\mathbf{L}}\mathbf{V}_{l_2}^{\mathbf{L}\top}, l_1 = 1, \ldots, k^{\mathbf{L}}, l_2 = 1, \ldots, k^{\mathbf{L}}\}$$

Therefore, the off-diagonal part of $\mathbf{L}$ can be projected onto $\mathcal{S}$ and the Schur complement can be projected onto $\mathcal{H}$ because the Cholesky factorization mainly involves the following two parts:

- $\mathbf{\Sigma}_{ij} \leftarrow \mathbf{\Sigma}_{ij} - \mathbf{L}_{i,1:j-1}\mathbf{L}_{j,1:j-1}^{\top}$
- $\mathbf{L}_{ij} \leftarrow \mathbf{\Sigma}_{ij}\mathbf{D}_{jj}^{\top}$, for $i > j$

Total complexity: $O(k_{\max}^2 n^{3/2} + k_{\max}^4 n + n^2)$, $k_{\max} = \max(k^{\mathbf{\Sigma}}, k^{\mathbf{L}}, k^{\mathbf{D}})$

Cholesky Factorization under SKP Continued

However, the Cholesky factorization algorithm has the issue of finding the optimal set of base blocks $\{\mathbf{V}_i^{\mathbf{L}}\}_{i=1}^{k^{\mathbf{L}}}$ because:

1. It operates on block columns ***sequentially*** from left to right
2. Unknown blocks of $\mathbf{L}$ are initially represented with ***coordinates*** relative to $\mathcal{S}$, and then they are ***projected*** onto $\{\mathbf{V}_i^{\mathbf{L}}\}$
3. The residuals are added to $\{\mathbf{V}_i^{\mathbf{L}}\}$ if greater than the threshold

Therefore, new bases are computed ***conditional*** on the old bases instead of all bases being computed simultaneously. The former generally has a larger $k^{\mathbf{L}}$ than the latter, for which we fix $k_{\max} \geq k^{\mathbf{L}}$ to control the computation cost

Cholesky Factorization under SKP Continued

Cholesky factorization in 1M dimensions for the exponential kernel finishes within 600 seconds but fails for the non-stationary kernel because $k_{\max}$ is fixed and new bases cannot be continuously added

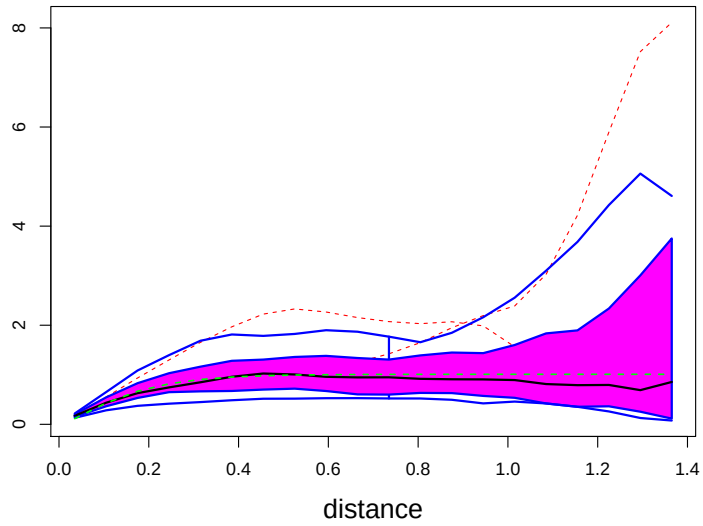
Table: Cholesky factorization of the exponential and non-stationary kernel matrices. Here, n spatial variables are located on a regular grid in the unit square

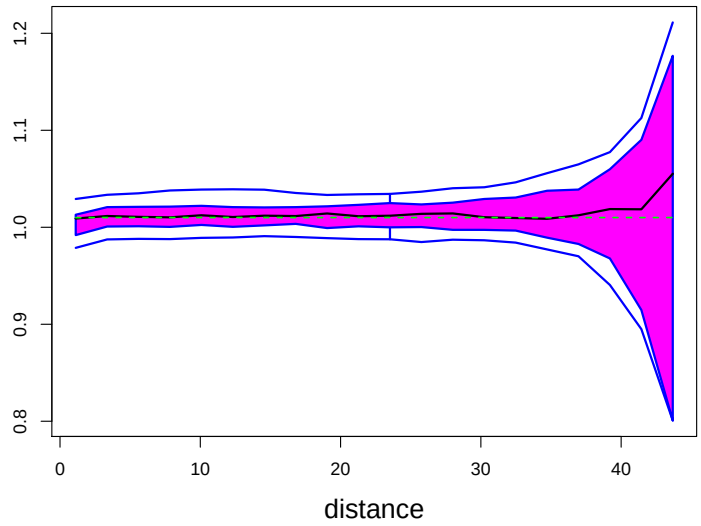
	Err $n = 2^{10}$	Err $n = 2^{12}$	Err $n = 2^{14}$	Time $n = 2^{20}$
Exponential	1.6%	4.1%	3.7%	596s
Non-stationary	1.3%	1.8%	8.0%	N.A.

It is a challenge how to **adjust existing base blocks** when new base blocks are introduced

Simulation of Large GRF

Simulate GRF in 1M dimensions under the Whittle kernel. Draw the functional boxplots of the empirical semivariogram based on 100 replicates





The semivariogram from the simulated data is well aligned with its theoretical value. The embedding methods (Gneiting et al., 2006) have a lower complexity but only work for limited kernels which do not include the Whittle kernel

Summary

Contributions:

- Proposed a novel SKP representation that is more efficient for kernel matrices from a regular grid than other compressed representations
- Concluded that the y-major order is the most aligned with the SKP representation
- Developed a Cholesky factorization algorithm using an efficient relative coordinate system and a base-block-discovery scheme

Limitations:

- The SKP representation cannot approximate matrices from irregular geometries efficiently
- The base blocks discovered in the Cholesky factorization are not optimal
- The Cholesky factorization has a complexity of $O(n^2)$

Bibliography

Akbudak, K., Ltaief, H., Mikhalev, A., and Keyes, D. (2017), "Tile low rank cholesky factorization for climate/weather modeling applications on manycore architectures," in *International Supercomputing Conference*, Springer, pp. 22–40.

Cao, J., Genton, M. G., Keyes, D. E., and Turkiyyah, G. M. (2021), "Sum of Kronecker products representation and its Cholesky factorization for spatial covariance matrices from large grids," *Computational Statistics & Data Analysis*, 107165.

Chandrasekaran, S., Gu, M., and Lyons, W. (2005), "A fast adaptive solver for hierarchically semiseparable representations," *Calcolo*, 42, 171–185.

Gneiting, T., Ševčíková, H., Percival, D. B., Schlather, M., and Jiang, Y. (2006), "Fast and exact simulation of large Gaussian lattice systems in 2D: exploring the limits," *Journal of Computational and Graphical Statistics*, 15, 483–501.

Hackbusch, W. (2015), *Hierarchical Matrices: Algorithms and Analysis*, vol. 49, Springer.

Van Loan, C. F. and Pitsianis, N. (1993), "Approximation with Kronecker products," in *Linear Algebra for Large Scale and Real-time Applications*, Springer, pp. 293–314.