Linear-time factorization of covariance matrices

Kenneth L. Ho (Stanford)

Joint work with Lexing Ying

SIAM CSE 2015

Introduction

- Covariance matrices are central to Gaussian-process–based statistical modeling

Introduction

- Covariance matrices are central to Gaussian-process–based statistical modeling
- Many common covariance functions are long-ranged

Exponential (λ large):	$C(r;\lambda) = \exp\left(-\frac{r}{\lambda}\right)$
Matérn (ν small or λ large):	$C(r;\nu,\lambda) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}r}{\lambda}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{\lambda}\right)$
Rational quadratic:	$C(r;\alpha,\lambda) = \left(1+\frac{r^2}{2\alpha\lambda^2}\right)^{-\alpha}$

- Costs of computing with dense covariance matrices:

$y = Ax$	$O(N^2)$
$x = (A+\sigma^2 I)^{-1} b$	$O(N^3)$
$A = BB^{\mathsf{T}}$	$O(N^3)$
$\Delta = \log\det A$	$O(N^3)$

Introduction

- Covariance matrices are central to Gaussian-process–based statistical modeling
- Many common covariance functions are long-ranged

Exponential (λ large):	$C(r;\lambda) = \exp\left(-\frac{r}{\lambda}\right)$
Matérn (ν small or λ large):	$C(r;\nu,\lambda) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}r}{\lambda}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{\lambda}\right)$
Rational quadratic:	$C(r;\alpha,\lambda) = \left(1+\frac{r^2}{2\alpha\lambda^2}\right)^{-\alpha}$

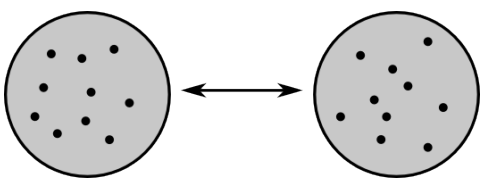
- Costs of computing with dense covariance matrices:

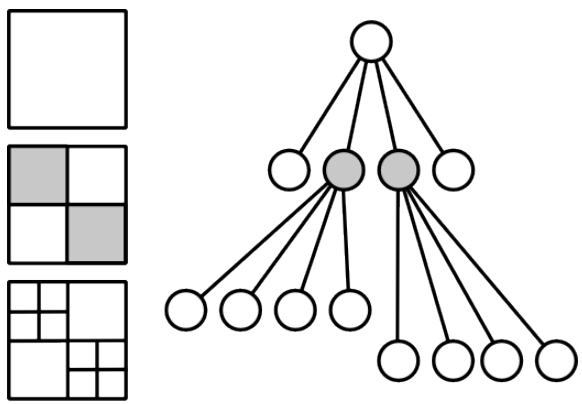
$y = Ax$	$O(N^2)$	$\to O(N)$
$x = (A+\sigma^2 I)^{-1} b$	$O(N^3)$	$\to O(N)$
$A = BB^{\mathsf{T}}$	$O(N^3)$	$\to O(N)$
$\Delta = \log\det A$	$O(N^3)$	$\to O(N)$

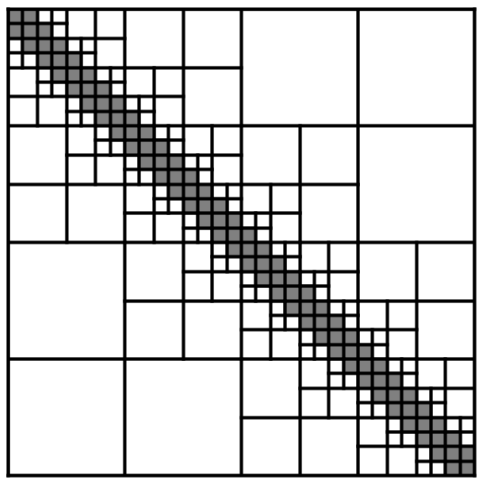
- Goal: enable large-scale calculations by accelerating to **linear** complexity

Main observation

- Covariance matrix is dense but **structured**
- Smooth far field $\Longrightarrow$ low-rank off-diagonal blocks
- Decompose and compress hierarchically
- Similar in flavor to fast multipole methods and treecodes







- See also MS241 (linear-complexity dense linear algebra) on Tuesday

[Ambikasaran/Foreman-Mackey/Greengard/Hogg/O'Neil 2014, Ambikasaran/Li/Kitanidis/Darve 2013, Ambikasaran/O'Neil 2014, Anitescu/Chen/Wang 2012, Chen 2014, Chen/Wang/Anitescu 2014, Saibaba/Kitanidis 2012]

Overview

Problem setting:

- Matrix can be low-rank but best if rank is not too small
  - Otherwise just use low-rank techniques (random sampling)
- Low geometric dimensionality: think time or space
- Fixed-domain asymptotics ($N \to \infty$ with λ fixed)

Results:

- Generalized Cholesky decomposition by **recursive skeletonization**
  - Originally developed for solving integral equations/PDEs
- Optimal $O(N)$ complexity with small constants
- Kernel-independent: depends weakly on specific covariance function
- Interpretation as adaptive model reduction

[Gillman/Young/Martinsson 2012, Ho/Greengard 2012, Ho/Ying 2013, Martinsson/Rokhlin 2005]

Overview

Problem setting:

- Matrix can be low-rank but best if rank is not too small
  - Otherwise just use low-rank techniques (random sampling)
- Low geometric dimensionality: think time or space
- Fixed-domain asymptotics ($N \to \infty$ with λ fixed)

Results:

- Generalized Cholesky decomposition by **recursive skeletonization**
  - Originally developed for solving integral equations/PDEs
- Optimal $O(N)$ complexity with small constants
- Kernel-independent: depends weakly on specific covariance function
- Interpretation as adaptive model reduction

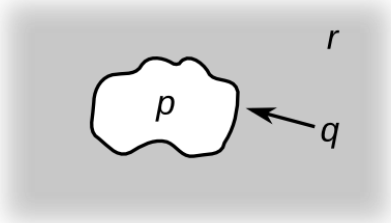
Tools: sparse elimination, interpolative decomposition, skeletonization

[Gillman/Young/Martinsson 2012, Ho/Greengard 2012, Ho/Ying 2013, Martinsson/Rokhlin 2005]

Sparse elimination

Let

$$A = \begin{bmatrix} A_{pp} & A_{qp}^{\mathsf{T}} & \\ A_{qp} & A_{qq} & A_{rq}^{\mathsf{T}} \\ & A_{rq} & A_{rr} \end{bmatrix}$$



be a "sparse" SPD matrix. Then define

$$S_p = \begin{bmatrix} I & -A_{pp}^{-1} A_{qp}^{\mathsf{T}} & \\ & I & \\ & & I \end{bmatrix} \quad \Longrightarrow \quad S_p^{\mathsf{T}} A S_p = \begin{bmatrix} A_{pp} & & \\ & * & A_{rq}^{\mathsf{T}} \\ & A_{rq} & A_{rr} \end{bmatrix}.$$

- Classical tool in numerical PDEs
- DOFs p have been eliminated
- Interactions involving r are unchanged

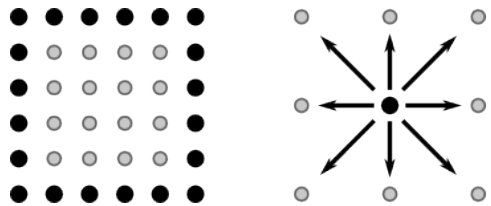
Interpolative decomposition

If $A_{:,q}$ has numerical rank k, then there exist

- skeleton ($\hat{q}$) and redundant ($\check{q}$) columns partitioning $q = \hat{q} \cup \check{q}$ with $|\hat{q}| = k$
- an interpolation matrix T_q

such that

$$A_{:,\check{q}} \approx A_{:,\hat{q}} T_q.$$



- Essentially a pivoted QR written slightly differently
- Rank-revealing to any specified precison $\epsilon > 0$ (controllable error)
- Fast randomized algorithms are available

[Cheng/Gimbutas/Martinsson/Rokhlin 2005, Gu/Eisenstat 1996, Halko/Martinsson/Tropp 2011]

Skeletonization

- Efficient elimination of redundant DOFs
- Let $A = \begin{bmatrix} A_{pp} & A_{qp}^{\mathsf{T}} \\ A_{qp} & A_{qq} \end{bmatrix}$ with A_{qp} low-rank
- Apply ID to A_{qp}: $A_{q\check{p}} \approx A_{q\hat{p}} T_p$
- Reorder $A = \begin{bmatrix} A_{\check{p}\check{p}} & A_{\hat{p}\check{p}}^{\mathsf{T}} & A_{q\check{p}}^{\mathsf{T}} \\ A_{\hat{p}\check{p}} & A_{\hat{p}\hat{p}} & A_{q\hat{p}}^{\mathsf{T}} \\ A_{q\check{p}} & A_{q\hat{p}} & A_{qq} \end{bmatrix}$, define $Q_p = \begin{bmatrix} I & & \\ -T_p & I & \\ & & I \end{bmatrix}$
- Sparsify via ID: $Q_p^* A Q_p \approx \begin{bmatrix} * & * & \\ * & A_{\hat{p}\hat{p}} & A_{q\hat{p}}^{\mathsf{T}} \\ & A_{q\hat{p}} & A_{qq} \end{bmatrix} \xrightarrow{\text{elim}} \begin{bmatrix} * & & \\ & * & A_{q\hat{p}}^{\mathsf{T}} \\ & A_{q\hat{p}} & A_{qq} \end{bmatrix}$
- Reduces to a subsystem involving skeletons only

[Ho/Ying 2013, Xia/Xi/Gu 2012]

Algorithm

Build tree.
for each level $\ell = 0, 1, 2, \ldots, L$ from finest to coarsest **do**
    Let C_ℓ be the set of all cells on level ℓ.
    for each cell $c \in C_\ell$ **do**
        Skeletonize remaining DOFs in c.
    end for
end for

[Ho/Ying 2013]

Algorithm

Build tree.
for each level $\ell = 0, 1, 2, \ldots, L$ from finest to coarsest **do**
    Let C_ℓ be the set of all cells on level ℓ.
    for each cell $c \in C_\ell$ **do**
        Skeletonize remaining DOFs in c.
    end for
end for

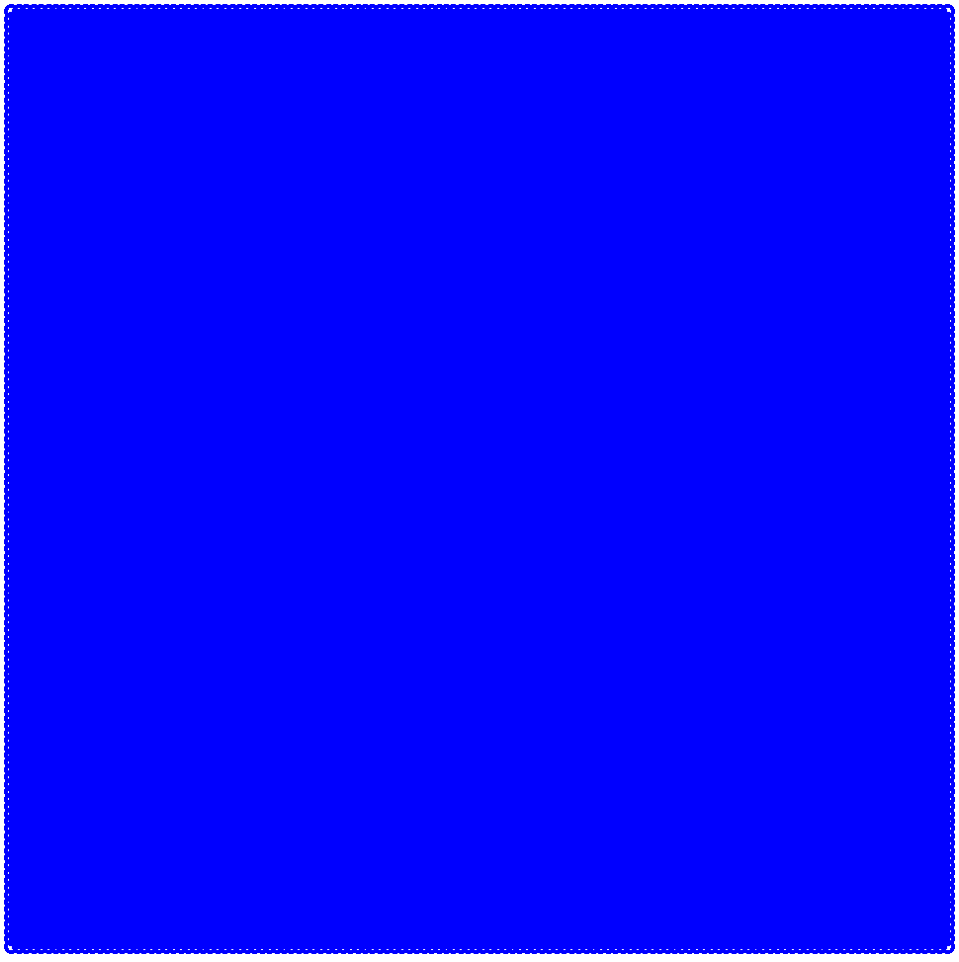
Example. Matérn ($\nu = 3/2$) in the unit square (2D Gaussian random field):

$$C(r; \lambda) = \left(1 + \frac{\sqrt{3}r}{\lambda}\right) \exp\left(-\frac{\sqrt{3}r}{\lambda}\right), \quad \lambda = \frac{1}{4}.$$

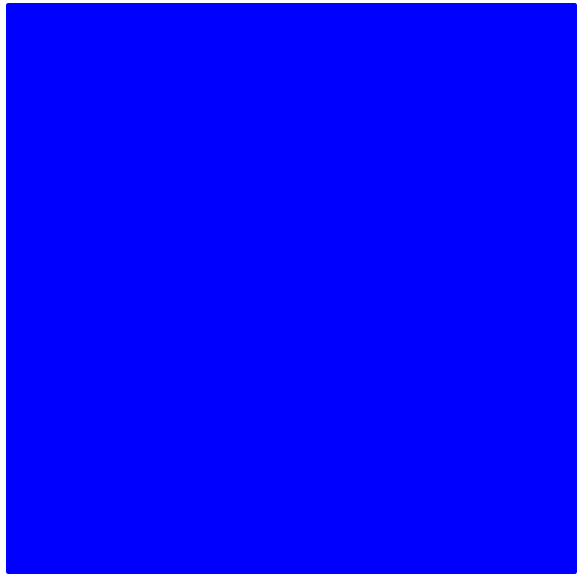
Approximate to relative precision $\epsilon = 10^{-6}$: $N = 16384 \to 543$.

[Ho/Ying 2013]

Level 0

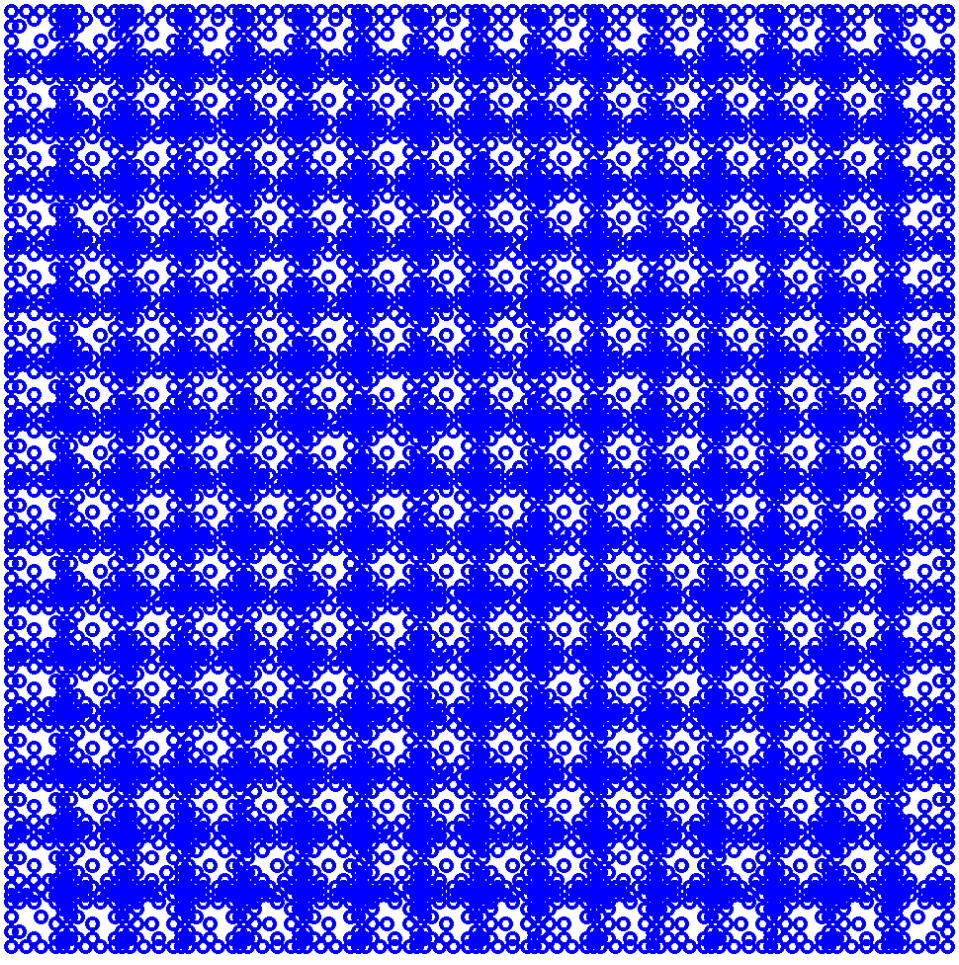


domain

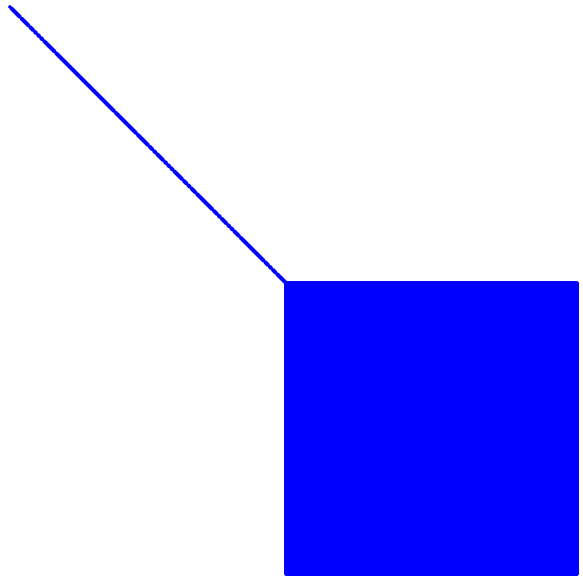


matrix

Level 1

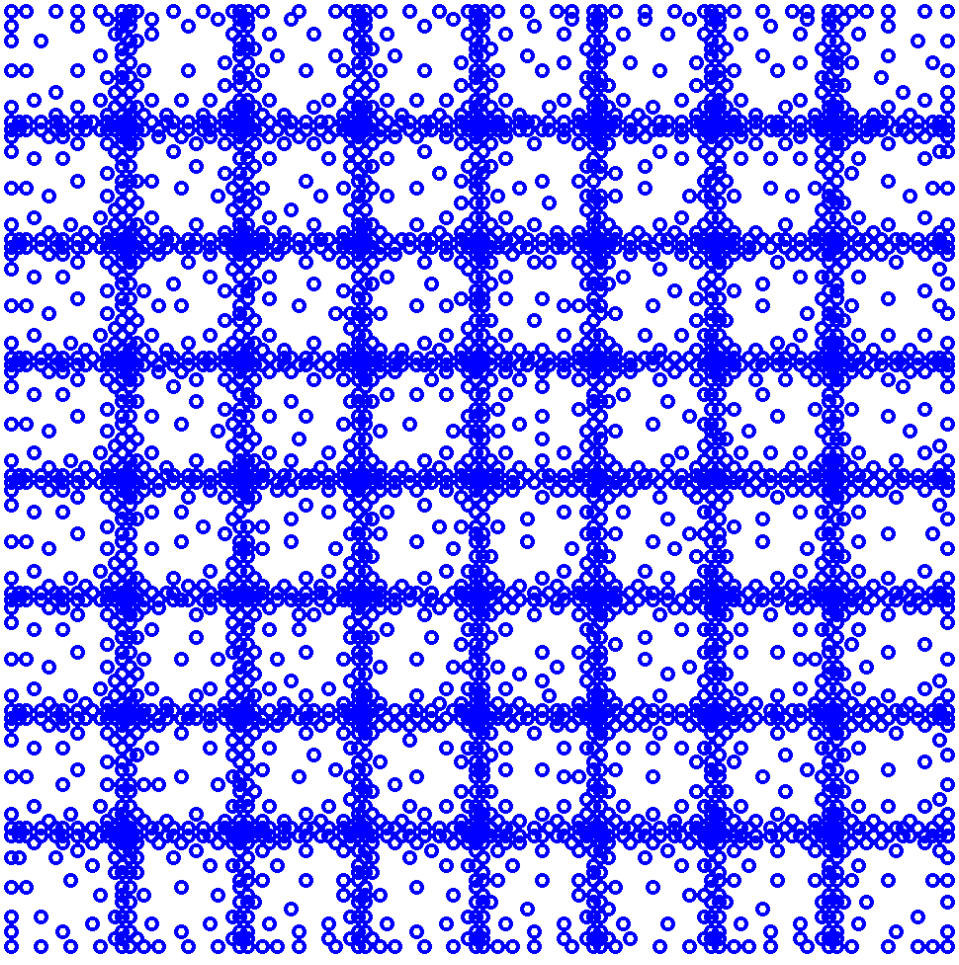


domain

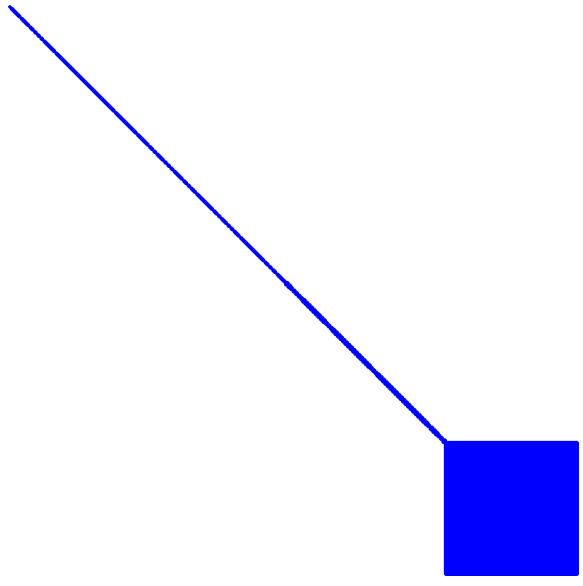


matrix

Level 2

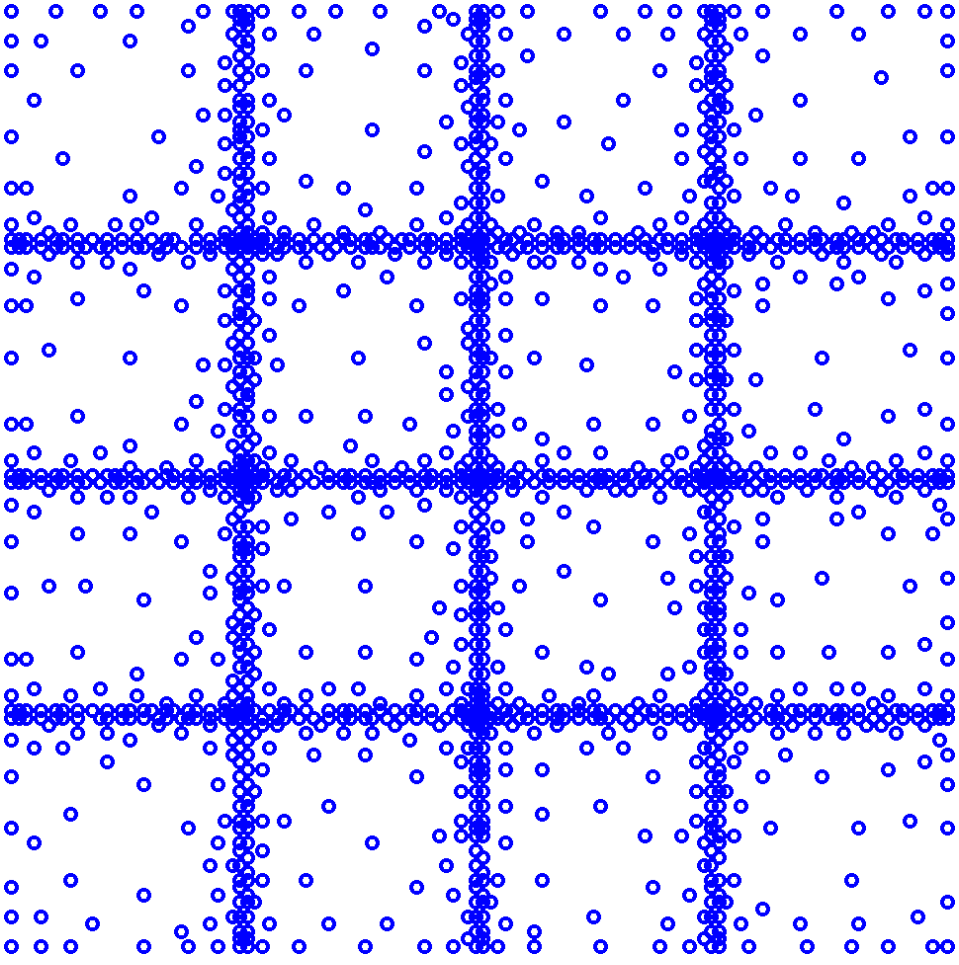


domain

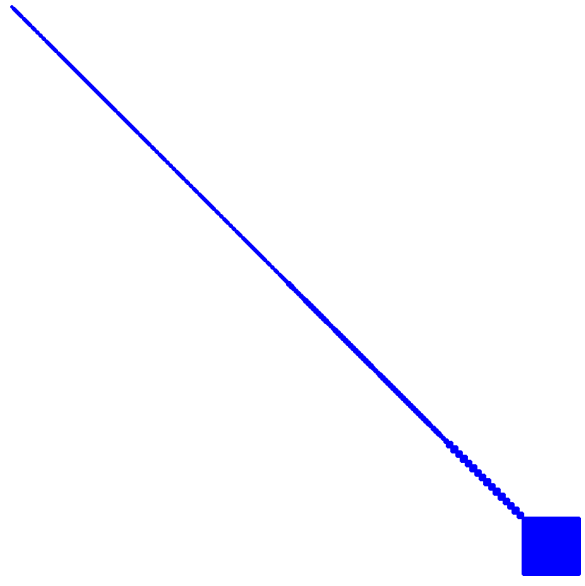


matrix

Level 3

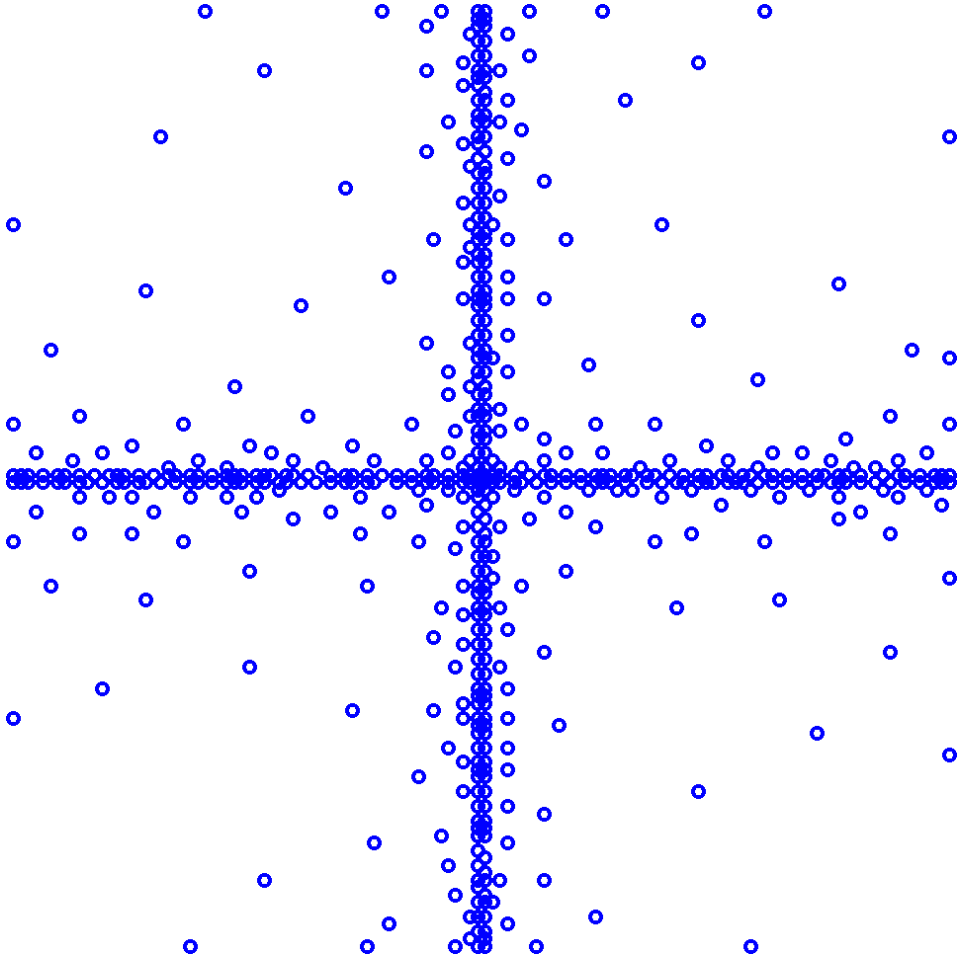


domain

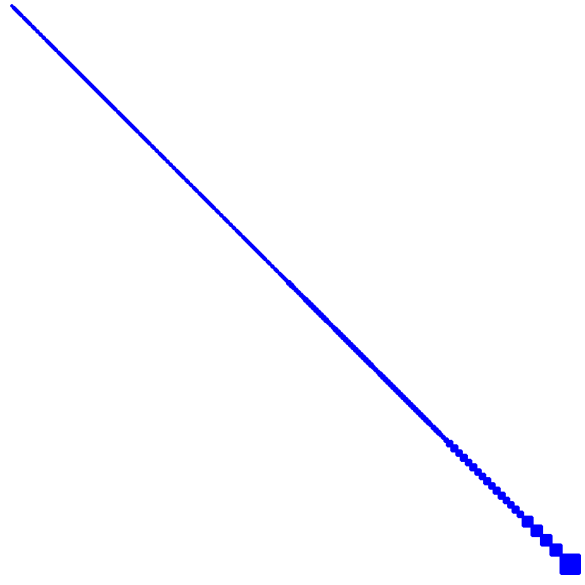


matrix

Level 4



domain



matrix

Properties

- Skeletonization operators:

$$U_\ell = \prod_{c \in C_\ell} Q_c S_c, \qquad Q_p = \begin{bmatrix} I & & \\ * & I & \\ & & I \end{bmatrix}, \quad S_p = \begin{bmatrix} I & * & \\ & I & \\ & & I \end{bmatrix},$$

- Symmetric block diagonalization:

$$D \approx U_{L-1}^{\mathsf{T}} \cdots U_0^{\mathsf{T}} A U_0 \cdots U_{L-1}$$

- Generalized Cholesky/LDL^{T} decomposition (SPD if $\epsilon\kappa(A) < 1$):

$$A \approx U_0^{-\mathsf{T}} \cdots U_{L-1}^{-\mathsf{T}} D U_{L-1}^{-1} \cdots U_0^{-1}$$
$$A^{-1} \approx U_0 \cdots U_{L-1} D^{-1} U_L^{\mathsf{T}} \cdots U_0^{\mathsf{T}}$$

- Fast multiplication/inversion, square root, $\det A = \det D$
- All operations are **very cheap** once the factorization has been constructed
- Skeletons: reduced order model at each length scale

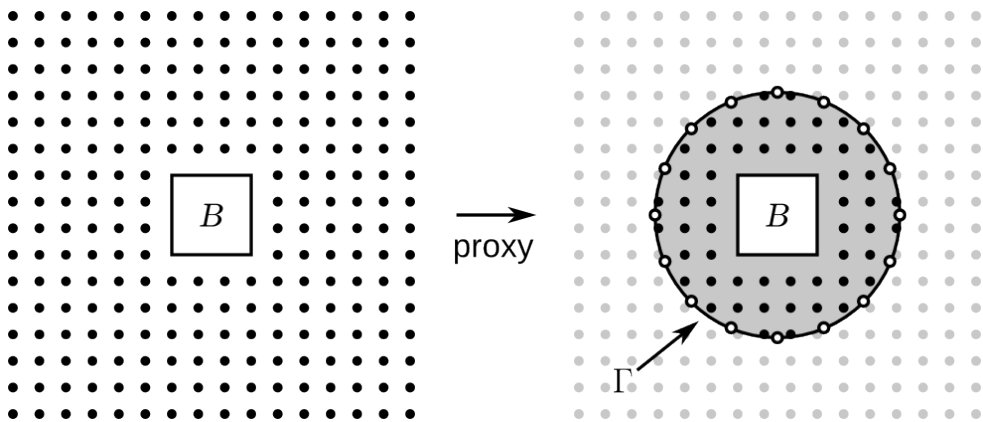
[Ho/Ying 2013]

Accelerated compression

- Main cost of algorithm is computing IDs (of $A_{p^C,p}$)
- Naive compression is global $\implies$ total cost of at least $O(N^2)$
- Observation: if $W_{:,q} = XY_{:,q}$ and $Y_{:,\check{q}} = Y_{:,\hat{q}}T_q$, then

$$W_{:,\check{q}} = XY_{:,\check{q}} = XY_{:,\hat{q}}T_q = W_{:,\hat{q}}T_q$$

- Can replace tall-and-skinny ID of W by short-and-skinny ID of Y
- How to find Y? Use analyticity, sampling, etc.



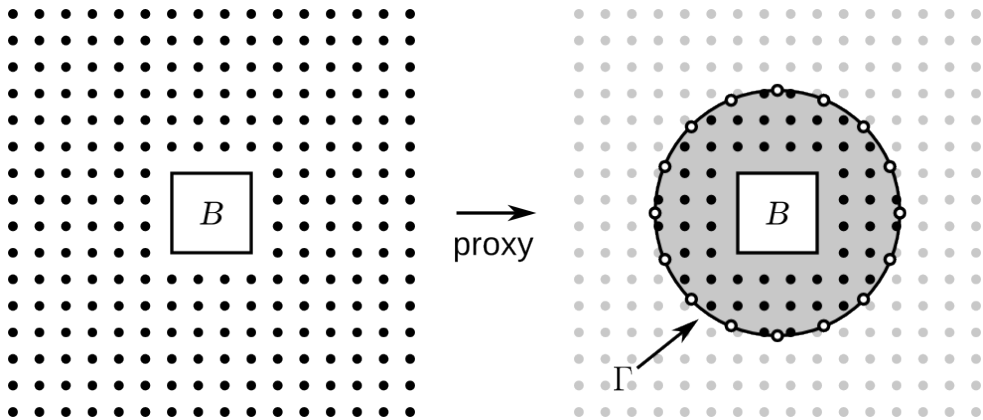
[Ho/Ying 2013]

Accelerated compression

- Main cost of algorithm is computing IDs (of $A_{p^C,p}$)
- Naive compression is global $\implies$ total cost of at least $O(N^2)$
- Observation: if $W_{:,q} = XY_{:,q}$ and $Y_{:,\check{q}} = Y_{:,\hat{q}} T_q$, then

$$W_{:,\check{q}} = XY_{:,\check{q}} = XY_{:,\hat{q}} T_q = W_{:,\hat{q}} T_q$$

- Can replace tall-and-skinny ID of W by short-and-skinny ID of Y
- How to find Y? Use analyticity, sampling, etc.



Results in this talk:

- Keep all near-field interactions
- Sample far field on a few concentric rings of radii 1, 2, 4, 8, etc.

[Ho/Ying 2013]

Theorem

If the off-diagonal block rank is bounded, then constructing the approximate factorization requires $O(N)$ operations.

[Ho/Ying 2013]

Theorem

If the off-diagonal block rank is bounded, then constructing the approximate factorization requires $O(N)$ operations.

- For fixed-domain asymptotics, interaction length scale is independent of N
- Therefore, number of "distinct" interactions is bounded
- Rank is bounded $\implies$ **linear** complexity
- Note: constant has the form $O(2^d)$

[Ho/Ying 2013]

Theorem

If the off-diagonal block rank is bounded, then constructing the approximate factorization requires $O(N)$ operations.

- For fixed-domain asymptotics, interaction length scale is independent of N
- Therefore, number of "distinct" interactions is bounded
- Rank is bounded $\implies$ **linear** complexity
- Note: constant has the form $O(2^d)$

What about increasing-domain asymptotics?

- Number of interactions grows as $1/\lambda \sim N^{1/d}$
- Cost becomes $O(N^{3(1-1/d)})$
- Must do additional work to recover linear complexity
  - Example: hierarchical interpolative factorization

[Ho/Ying 2013]

Numerical benchmarks in MATLAB

Matérn ($\nu = 3/2$, $\lambda = 1/8$) with nugget effect of $\sigma^2 = 0.01$

- Point distributions: unit line (1D) or square (2D)

d	ϵ	N	$\lvert s_L \rvert$	t_f (s)	$t_{a/s}$ (s)	t_d (s)	e_a	e_d
1D	10^{-08}	262144	4	1.8e+1	4.8e−1	9.8e−2	1.1e−08	1.2e−9
		524288	4	3.5e+1	1.1e+0	2.0e−1	4.3e−07	1.8e−7
		1048576	4	7.0e+1	2.0e+0	3.9e−1	4.7e−07	1.1e−7
1D	10^{-12}	262144	4	1.8e+1	4.7e−1	9.9e−2	2.1e−13	—
		524288	4	3.5e+1	9.3e−1	2.0e−1	2.8e−13	—
		1048576	4	7.0e+1	1.9e+0	4.0e−1	3.0e−13	—
2D	10^{-06}	256^2	214	7.4e+0	1.2e−1	1.8e−2	5.8e−07	2.6e−6
		512^2	219	2.8e+1	4.1e−1	7.3e−2	1.8e−06	4.1e−6
		1024^2	220	1.1e+2	1.6e+0	2.9e−1	1.7e−06	8.0e−6
2D	10^{-09}	256^2	1081	3.2e+1	2.1e−1	1.8e−2	5.4e−10	—
		512^2	1227	6.7e+1	5.9e−1	7.4e−2	1.1e−09	—
		1024^2	1301	1.7e+2	1.9e+0	3.0e−1	4.0e−09	—

- Can be heavily accelerated by a more careful implementation

Example: Gaussian process regression

- **Unknown** function $f(x)$ on $[0, 1]$
- Prior: zero mean, Matérn covariance $C(x, x')$ with $\nu = 3/2$ and $\lambda = 1/8$
- Measurements $y_1 = f(x_1) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, at N uniform random points
- Estimate values of $y_2 = f(x_2)$ at N equispaced points:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right), \qquad \begin{aligned} A_{11} &= C(x_1, x_1) + \sigma^2 I \\ A_{21} &= C(x_2, x_1) \\ A_{22} &= C(x_2, x_2) \end{aligned}$$

$$\implies y_2 \mid y_1 \sim \mathcal{N}(\mu_{\text{post}}, A_{\text{post}}), \qquad \begin{aligned} \mu_{\text{post}} &= A_{21} A_{11}^{-1} y_1 \\ A_{\text{post}} &= A_{22} - A_{21} A_{11}^{-1} A_{12} \end{aligned}$$

- $N \sim 10^6$, $\sigma^2 = 0.01$: 273 s to compute μ_{post} to precision 10^{-5}
- Generate conditional samples via $y_2 = \mu_{\text{post}} + z_2 - A_{21} A_{11}^{-1} z_1$, where $z \sim \mathcal{N}(0, A)$
- Estimate posterior variances to precision 10^{-2} by sampling: $\sim$ 30 min

Summary

- Efficient factorization of covariance matrices
  - Apply, solve, square root, determinant, etc.
  - Extends to general structured matrices with low-rank off-diagonal blocks
- **Linear** complexity under fixed-domain asymptotics
  - Can extend to increasing-domain asymptotics with some work
- Applications: Gaussian processes, maximum likelihood estimation, etc.
  - *There is no* $O(N^3)$ *bottleneck!*
- Key idea: sparsification and elimination (skeletonization) via the ID
- Naturally parallelizable: independent for-loops at each level
- However, effective only in low geometric dimensions
  - High-dimensional setting will require new ideas
- Extensions: posterior variances by selected inversion, online data assimilation

[Ho/Ying 2013, Minden/Damle/Ho/Ying 2014]

References

- K.L. Ho, L. Greengard. A fast direct solver for structured linear systems by recursive skeletonization. SIAM J. Sci. Comput. 34 (5): A2507–A2532, 2012.
- K.L. Ho, L. Ying. Hierarchical interpolative factorization for elliptic operators: differential equations. Preprint, arXiv:1307.2895 [math.NA], 2013. To appear, Comm. Pure Appl. Math.
- K.L. Ho, L. Ying. Hierarchical interpolative factorization for elliptic operators: integral equations. Preprint, arXiv:1307.2666 [math.NA], 2013. To appear, Comm. Pure Appl. Math.
- V. Minden, A. Damle, K.L. Ho, L. Ying. A technique for updating hierarchical factorizations of integral operators. Preprint, arXiv:1411.5706 [math.NA], 2014.