Compression, inversion and sparse approximate PCA of dense kernel matrices in near linear computational complexity

Florian Schäfer

ICERM 2017
Compression, inversion and approximate PCA of dense kernel matrices in near linear computational complexity

Florian Schäfer, T.J. Sullivan, Houman Owhadi

1. A numerical experiment
2. Disintegration of measure and Gaussian elimination
3. Near-linear complexity algorithms using the theory of Gamblets
4. Further numerical results
A numerical experiment

\[ \{ x_i \}_{i \in I} \subset [0, 1]^2, \text{ with } \# I = N = 16641 \]
A numerical experiment

- \( \{x_i\}_{i \in I} \subset [0, 1]^2 \), with \( \#I = N = 16641 \)
- Define \( K(r) \) as Matérn kernel with smoothness parameter \( \nu = 1 \) and lengthscale \( l = 0.4 \).
A numerical experiment

- \( \{x_i\}_{i \in I} \subset [0, 1]^2 \), with \( \# I = N = 16641 \)
- Define \( K(r) \) as Matérn kernel with smoothness parameter \( \nu = 1 \) and lengthscale \( l = 0.4 \).
- \( \Gamma_{i,j} := K(\|x_i - x_j\|) \).

\[ \{x_i\}_{i \in I} \subset [0, 1]^2, \text{ with } \# I = N = 16641 \]

Define \( K(r) \) as Matérn kernel with smoothness parameter \( \nu = 1 \) and lengthscale \( l = 0.4 \).

\( \Gamma_{i,j} := K(\|x_i - x_j\|) \).
A numerical experiment

- $\Gamma$, interpreted as covariance matrix, describes a Gaussian field with second order smoothness.
A numerical experiment

- $\Gamma$, interpreted as covariance matrix, describes a Gaussian field with second order smoothness.
- Alternatively, $K$ can be seen as the Green’s function of a fourth order elliptic PDE, on the whole space.
A numerical experiment

- \( \Gamma \), interpreted as covariance matrix, describes a Gaussian field with fourth order smoothness.
- Alternatively, \( K \) can be seen as the Green's function of a fourth order elliptic PDE, on the whole space.
- Matrices of this kind appear in both statistics and scientific computing.
A numerical experiment

- $\Gamma$, interpreted as covariance matrix, describes a Gaussian field with fourth order smoothness.
- Alternatively, $K$ can be seen as the Green’s function of a fourth order elliptic PDE, on the whole space.
- Matrices of this kind appear in both statistics and scientific computing.
- We need to apply the Matrix and its inverse, and compute its determinant.
A numerical experiment

- $\Gamma$, interpreted as covariance matrix, describes a Gaussian field with fourth order smoothness.
- Alternatively, $K$ can be seen as the Green’s function of a fourth order elliptic PDE, on the whole space.
- Matrices of this kind appear in both statistics and scientific computing.
- We need to apply the Matrix and its inverse, and compute its determinant.
- $\Gamma$ is dense, and hence has $N^2$ storage cost. Direct inversion via Gaussian elimination has $O(N^3)$ complexity in time.
A numerical experiment

Can we be more efficient?
A numerical experiment

- Can we be more efficient?
- Many existing methods: Quadrature formulas, subsampling, randomised approximations, low rank approximations, fast multipole methods, hierarchical matrices, wavelet methods, inducing points, covariance tapering . . . .
A numerical experiment

- Can we be more efficient?

- Many existing methods: Quadrature formulas, subsampling, randomised approximations, low rank approximations, fast multipole methods, hierarchical matrices, wavelet methods, inducing points, covariance tapering . . .

- We provide a simple algorithm, with rigorous error bounds and near-linear complexity.
Even writing down the matrix has $N^2$ complexity.
A numerical experiment

- Even writing down the matrix has $N^2$ complexity.
- Therefore, we subsample $\Gamma$:

$$\tilde{\Gamma}_{i,j} := \begin{cases} 
\Gamma_{i,j}, & \text{for } (i,j) \in S_2 \\
0, & \text{else}
\end{cases}$$
A numerical experiment

- Even writing down the matrix has $N^2$ complexity.
- Therefore, we subsample $\Gamma$:

$$\tilde{\Gamma}_{i,j} := \begin{cases} 
\Gamma_{i,j}, & \text{for } (i,j) \in S_2 \\
0, & \text{else}
\end{cases}$$

$$\#S_2 = 5528749 = 0.0189N^2.$$ We have thrown away all but 2 percent of the entries, without even touching them!
A numerical experiment

- Even writing down the matrix has $N^2$ complexity.
- Therefore, we subsample $\Gamma$:
  \[
  \tilde{\Gamma}_{i,j} := \begin{cases} 
  \Gamma_{i,j}, & \text{for } (i,j) \in S_2 \\
  0, & \text{else}
  \end{cases}
  \]
- $\#S_2 = 5528749 = 0.0189N^2$. We have thrown away all but 2 percent of the entries, without even touching them!
- We will see later: $S_2$ does not depend on the entries of $\Gamma$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example.png}
\end{figure}
We have compressed $\Gamma$ to 2 percent of its original size.
A numerical experiment

- We have compressed $\Gamma$ to 2 percent of its original size.
- How much information have we retained?
A numerical experiment

- We have compressed $\Gamma$ to 2 percent of its original size.
- How much information have we retained?
- Consider relative error in operator norm:
A numerical experiment

- We have compressed $\Gamma$ to 2 percent of its original size.
- How much information have we retained?
- Consider relative error in operator norm:

$$\frac{\| \Gamma - \tilde{\Gamma} \|}{\| \Gamma \|} = 0.9662$$
A numerical experiment

- We have compressed $\Gamma$ to 2 percent of its original size.
- How much information have we retained?
- Consider relative error in operator norm:

$$\frac{\|\Gamma - \tilde{\Gamma}\|}{\|\Gamma\|} = 0.9662$$

- $\tilde{\Gamma}$ is a very bad approximation of $\Gamma$. 
A numerical experiment

Can we obtain a better approximation of $\Gamma$, using only $\tilde{\Gamma}$?
A numerical experiment

- Can we obtain a better approximation of $\Gamma$, using only $\tilde{\Gamma}$?
- Consider $L = \text{ICHOL}(\tilde{\Gamma})$, the incomplete Cholesky factorisation of $\tilde{\Gamma}$, ignoring all fill-in.
Can we obtain a better approximation of $\Gamma$, using only $\tilde{\Gamma}$?

Consider $L = \text{ICHOL}(\tilde{\Gamma})$, the incomplete Cholesky factorisation of $\tilde{\Gamma}$, ignoring all fill-in.

$$\frac{\|\Gamma - LL^T\|}{\|\Gamma\|} = 3.0676 \times 10^{-04}$$
A numerical experiment

- Can we obtain a better approximation of $\Gamma$, using only $\tilde{\Gamma}$?
- Consider $L = \text{ICHOL}\left(\tilde{\Gamma}\right)$, the incomplete Cholesky factorisation of $\tilde{\Gamma}$, ignoring all fill-in.

\[
\frac{\left\| \Gamma - LL^T \right\|}{\left\| \Gamma \right\|} = 3.0676 \times 10^{-04}
\]

\[
K_{\text{approx}}
\]

\[
K_{\text{true}}
\]
A numerical experiment

Decompose $\{x_i\}_{i \in I}$ into a nested hierarchy as:

$$\{x_i\}_{i \in I^{(1)}} \subset \{x_i\}_{i \in I^{(2)}} \subset \{x_i\}_{i \in I^{(3)}} \subset \cdots \subset \{x_i\}_{i \in I^{(q)}} = \{x_i\}_{i \in I} \quad (1.1)$$
A numerical experiment

- We define $J^{(k)} := I^{(k)} \setminus I^{(k-1)}$ and define the sparsity pattern:

$$S_2 := \left\{ (i, j) \in I \times I \mid i \in J^{(k)}, j \in J^{(l)}, \text{dist} (x_i, x_j) \leq 2 \cdot 2^{-\min(k,l)} \right\}.$$
A numerical experiment

- We define $J^{(k)} := I^{(k)} \setminus I^{(k-1)}$ and define the sparsity pattern:

$$S_2 := \left\{ (i, j) \in I \times I \left| i \in J^{(k)}, j \in J^{(l)}, \text{dist} (x_i, x_j) \leq 2 \ast 2^{\min(k,l)} \right. \right\}.$$  

- We order the elements of $I$ from coarse to fine, that is from $J^{(1)}$ to $J^{(q)}$. 

![Graphs showing sparsity patterns](image)
A numerical experiment

- $L$ provides a good approximation of $\Gamma$ at only 2 percent of the storage cost.
A numerical experiment

- $L$ provides a good approximation of $\Gamma$ at only 2 percent of the storage cost.
- Can be computed in near linear complexity in time and space.
A numerical experiment

- $L$ provides a good approximation of $\Gamma$ at only 2 percent of the storage cost.
- Can be computed in near linear complexity in time and space.
- Allows for approximate evaluation of $\Gamma$, $\Gamma^{-1}$, and $\det(\Gamma)$ in near-linear time.
A numerical experiment

- $L$ provides a good approximation of $\Gamma$ at only 2 percent of the storage cost.
- Can be computed in near linear complexity in time and space.
- Allows for approximate evaluation of $\Gamma$, $\Gamma^{-1}$, and $\det(\Gamma)$ in near-linear time.
- Allows for sampling of $X \sim N(0, \Gamma)$ in near-linear time.
In this work, we
In this work, we prove that this phenomenon holds whenever the covariance function $K$ is the Green’s function of an elliptic boundary value problem.
In this work, we prove that this phaenomenon holds whenever the covariance function $K$ is the Green’s function of an elliptic boundary value problem.

prove that it leads to an algorithm with computational complexity of $O \left( N \log^2 (N) \left( \log (1/\epsilon) + \log^2 (N) \right)^{4d+1} \right)$ in time and $O\left( N \log(N) \log^d (N \frac{1}{\epsilon}) \right)$ in space for an approximation error of $\epsilon$. 
In this work, we prove that this phenomenon holds whenever the covariance function $K$ is the Green’s function of an elliptic boundary value problem.

prove that it leads to an algorithm with computational complexity of $O\left( N \log^2 (N) \left( \log (1/\epsilon) + \log^2 (N) \right)^{4d+1} \right)$ in time and $O\left( N \log(N) \log^d (N^{1/\epsilon}) \right)$ in space.

show that even though the Matérn family is not covered rigorously by our theoretical results, we get good approximation results, in particular in the interior of the domain.
A numerical experiment

- In this work, we
- prove that this phenomenon holds whenever the covariance function $K$ is the Green’s function of an elliptic boundary value problem.
- prove that it leads to an algorithm with computational complexity of $O\left(N \log^2(N) \left( \log(1/\epsilon) + \log^2(N) \right)^{4d+1} \right)$ in time and $O\left(N \log(N) \log^d\left(N \frac{1}{\epsilon}\right) \right)$ in space.
- show that even though the Matérn family is not covered rigorously by our theoretical results, we get good approximation results, in particular in the interior of the domain.
- show that as a byproduct of our algorithm we obtain a sparse approximate PCA with near optimal approximation property.
Let $X$ be a centered Gaussian vector with covariance $\Theta$. 
Let $X$ be a centered Gaussian vector with covariance $\Theta$.

Assume we want to compute $\mathbb{E} [f (X)]$ for some function $f$. 
Let \( X \) be a centered Gaussian vector with covariance \( \Theta \).

Assume we want to compute \( \mathbb{E}[f(X)] \) for some function \( f \).

Use Monte Carlo, but for \( \Theta \) large, each sample is expensive.
Let \( X \) be a centered Gaussian vector with covariance \( \Theta \).

Assume we want to compute \( \mathbb{E} [f(X)] \) for some function \( f \).

Use Monte Carlo, but for \( \Theta \) large, each sample is expensive.

Idea: use disintegration of measure:

\[
\mathbb{E} [f(X)] = \mathbb{E} \left[ \mathbb{E} [f(X) \mid Y] (Y) \right].
\]
Let $X$ be a centered Gaussian vector with covariance $\Theta$.

Assume we want to compute $\mathbb{E}[f(X)]$ for some function $f$.

Use Monte Carlo, but for $\Theta$ large, each sample is expensive.

Idea: use disintegration of measure:

$$\mathbb{E}[f(X)] = \mathbb{E}[\mathbb{E}[f(X)|Y](Y)].$$

Choose $Y$, such that $Y$ and $\mathbb{E}[f(X)|Y]$ can be sampled cheaply.
Consider $X \in \mathbb{R}^N$, $\{x_i\}_{1 \leq i \leq N} \subset [0, 1]$ and $\Theta_{i,j} := \exp \left( -|x_i - x_j| \right)$. 
Consider $X \in \mathbb{R}^N$, $\{x_i\}_{1 \leq i \leq N} \subset [0, 1]$ and $\Theta_{i,j} := \exp \left( -|x_i - x_j| \right)$.

Corresponds to a prior on the space $H^1(0, 1)$ of mean square differentiable functions.
Consider $X \in \mathbb{R}^N$, $\{x_i\}_{1 \leq i \leq N} \subset [0, 1]$ and $\Theta_{i,j} := \exp(-|x_i - x_j|)$.

Corresponds to a prior on the space $H^1(0, 1)$ of mean square differentiable functions.

Assume $x_i$ are ordered in increasing order and $x_{[N/2]} \approx 1/2$. 
Consider $X \in \mathbb{R}^N$, $\{x_i\}_{1 \leq i \leq N} \subset [0, 1]$ and $\Theta_{i,j} := \exp \left( -|x_i - x_j| \right)$. Corresponds to a prior on the space $H^1 (0, 1)$ of mean square differentiable functions.

Assume $x_i$ are ordered in increasing order and $x_{\lfloor N/2 \rfloor} \approx 1/2$.

We then have, for $i < \lfloor N/2 \rfloor < j$:

$$\text{Cov} \left[ X_i, X_j \mid X_{\lfloor N/2 \rfloor} \right] \approx 0.$$
Consider $X \in \mathbb{R}^N$, $\{x_i\}_{1 \leq i \leq N} \subset [0, 1]$ and $\Theta_{i,j} := \exp \left( -|x_i - x_j| \right)$.

Corresponds to a prior on the space $H^1(0, 1)$ of mean square differentiable functions.

Assume $x_i$ are ordered in increasing order and $x_{\lfloor N/2 \rfloor} \approx 1/2$

We then have, for $i < \lfloor N/2 \rfloor < j$ : $\text{Cov} \left[ X_i, X_j | X_{\lfloor N/2 \rfloor} \right] \approx 0.$
Consider $X \in \mathbb{R}^N$, $\{x_i\}_{1 \leq i \leq N} \subset [0, 1]$ and $\Theta_{i,j} := \exp \left( -|x_i - x_j| \right)$.

Corresponds to a prior on the space $H^1(0, 1)$ of mean square differentiable functions.

Assume $x_i$ are ordered in increasing order and $x_{\lfloor N/2 \rfloor} \approx 1/2$.

We then have, for $i < \lfloor N/2 \rfloor < j$:

$$\text{Cov} \left[ X_i, X_j | X_{\lfloor N/2 \rfloor} \right] \approx 0.$$
For two observation sites $x_i, x_j$, the covariance conditional on the observation sites inbetween is small.
For two observation sites $x_i, x_j$, the covariance conditional on the observation sites inbetween is small.

Disintegration of Gaussian Measures and the Screening Effect

- For two observation sites $x_i, x_j$, the covariance conditional on the observation sites inbetween is small.
- Let us take $Y = X_{\lfloor N/2 \rfloor}$. Then $Y$ is cheap to sample, and the covariance matrix of $X|Y$ has only $2 \left(\frac{N}{2}\right)^2$ nonnegligible entries.
For two observation sites $x_i, x_j$, the covariance conditional on the observation sites inbetween is small.


Let us take $Y = X_{\lfloor N/2 \rfloor}$. Then $Y$ is cheap to sample, and the covariance matrix of $X|Y$ has only $2 \left(\frac{N}{2}\right)^2$ nonneglegible entries.

When using Cholesky decomposition, this yields a factor 4 improvement of computational speed.
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- Look at a single step of Block Cholesky decomposition:
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- Look at a single step of Block Cholesky decomposition:
- This corresponds to:

\[
\begin{pmatrix}
\Theta_{11} & \Theta_{12} \\
\Theta_{21} & \Theta_{22}
\end{pmatrix}
= 
\begin{pmatrix}
\text{Id} & 0 \\
\Theta_{21}\Theta_{11}^{-1} & \text{Id}
\end{pmatrix}
\begin{pmatrix}
\Theta_{11} & 0 \\
0 & \Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}
\end{pmatrix}
\begin{pmatrix}
\text{Id} & \Theta_{11}^{-1}\Theta_{12} \\
0 & \text{Id}
\end{pmatrix}
\]
Look at a single step of Block Cholesky decomposition:

This corresponds to:

\[
\begin{pmatrix}
\Theta_{11} & \Theta_{12} \\
\Theta_{21} & \Theta_{22}
\end{pmatrix}
= \begin{pmatrix}
\text{Id} & 0 \\
\Theta_{21}\Theta_{11}^{-1} & \text{Id}
\end{pmatrix}
\begin{pmatrix}
\Theta_{11} & 0 \\
0 & \Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}
\end{pmatrix}
\begin{pmatrix}
\text{Id} & \Theta_{11}^{-1}\Theta_{12} \\
0 & \text{Id}
\end{pmatrix}
\]

Note, that for \( \left( \Theta_{21}\Theta_{11}^{-1} \right) b = \mathbb{E} \left[ X_2 | X_1 = b \right] \), and
\[
\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12} = \text{Cov} \left[ X_2 | X_1 \right].
\]
Look at a single step of Block Cholesky decomposition:

This corresponds to:

\[
\begin{pmatrix}
\Theta_{11} & \Theta_{12} \\
\Theta_{21} & \Theta_{22}
\end{pmatrix}
= \begin{pmatrix}
\text{Id} & 0 \\
\Theta_{21}\Theta_{11}^{-1} & \text{Id}
\end{pmatrix}
\begin{pmatrix}
\Theta_{11} & 0 \\
0 & \Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}
\end{pmatrix}
\begin{pmatrix}
\text{Id} & \Theta_{11}^{-1}\Theta_{12} \\
0 & \text{Id}
\end{pmatrix}
\]

Note, that for \(\Theta_{21}\Theta_{11}^{-1}\) \(b = \mathbb{E}[X_2|X_1 = b]\), and \(\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12} = \text{Cov}[X_2|X_1]\).

(Block-)Cholesky decomposition is computationally equivalent to the disintegration of Gaussian measures.
Look at a single step of Block Cholesky decomposition:
This corresponds to:

\[
\begin{pmatrix}
\Theta_{11} & \Theta_{12} \\
\Theta_{21} & \Theta_{22}
\end{pmatrix}
= \begin{pmatrix}
\text{Id} & 0 \\
\Theta_{21}\Theta_{11}^{-1} & \text{Id}
\end{pmatrix}
\begin{pmatrix}
\Theta_{11} & 0 \\
0 & \Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12}
\end{pmatrix}
\begin{pmatrix}
\text{Id} & \Theta_{11}^{-1}\Theta_{12} \\
0 & \text{Id}
\end{pmatrix}
\]

Note, that for \( \begin{pmatrix} \Theta_{21}\Theta_{11}^{-1} \end{pmatrix} b = \mathbb{E} [X_2 | X_1 = b] \), and
\[
\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12} = \text{Cov} [X_2 | X_1].
\]

(Block-)Cholesky decomposition is computationally equivalent to the disintegration of Gaussian measures.

Follows immediately from well known formulas, but rarely used in the literature. One Example: Bickson (2008).
This suggests to choose a *bisective* elimination ordering:
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- This suggests to choose a *bisective* elimination ordering:

- Let's start computing the Cholesky decomposition
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- This suggests to choose a *bisective* elimination ordering:

- Let's start computing the Cholesky decomposition
- We observe a *fade-out* of entries!
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

What about higher dimensional examples?
Sparse Factorisation of Dense Matrices: \textit{fade-out} instead of \textit{fill-in}

- How about higher dimensional examples?
- In 2d, use quadsection:
We know that the result of the factorisation is sparse, but can we compute it efficiently?
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- We know that the result of the factorisation is sparse, but can we compute it efficiently?
- Key observation: The entry \((i, j)\) is used for the first time with the \(\min (i, j)\)-th pivot.
We know that the result of the factorisation is sparse, but can we compute it efficiently?

Key observation: The entry \((i, j)\) is used for the first time with the \(\min (i, j)\)-th pivot.

If we know they will be negligible until we use them, we don’t have to update them, nor know them in the first place.
We know that the result of the factorisation is sparse, but can we compute it efficiently?

Key observation: The entry \((i, j)\) is used for the first time with the \(\min (i, j)\)-th pivot.

If we know they will be negligible until we use them, we don’t have to update them, nor know them in the first place.
Bisective/Quadsective ordering is the reverse of nested dissection.

Indeed, for $P$ the order-reversing permutation matrix, we have:

$$(\Theta)^{-1} = \left(LL^T\right)^{-1} = L^{-T}L^{-1}$$

$$\implies P(\Theta)^{-1}P = P L^{-T} P P L^{-1} P = \left(PL^{-T} P\right) \left(PL^{-T} P\right)^T,$$

But we have $L^{-1} = L^T (\Theta)^{-1}$.

For a sparse elimination ordering of $\Theta$, the reverse ordering leads to sparse factorisation of $(\Theta)^{-1}$
Sparse Factorisation of Dense Matrices: \textit{fade-out} instead of \textit{fill-in}

- We obtain a very simple algorithm:
We obtain a very simple algorithm:

- Given a positive definite matrix $\Theta$ and a Graph $G$, such that $\Theta^{-1}$ is sparse according to $G$. 
We obtain a very simple algorithm:

- Given a positive definite matrix $\Theta$ and a Graph $G$, such that $\Theta^{-1}$ is sparse according to $G$.
- Obtain inverse nested dissection ordering for $G$. 
We obtain a very simple algorithm:

- Given a positive definite matrix $\Theta$ and a Graph $G$, such that $\Theta^{-1}$ is sparse according to $G$.
- Obtain inverse nested dissection ordering for $G$.
- Set entries $(i,j)$ that are separated after pivot number $\min(i,j)$ to zero.
Sparse Factorisation of Dense Matrices: \textit{fade-out} instead of \textit{fill-in}

- We obtain a very simple algorithm:
- Given a positive definite matrix $\Theta$ and a Graph $G$, such that $\Theta^{-1}$ is sparse according to $G$.
- Obtain inverse nested dissection ordering for $G$.
- Set entries $(i, j)$ that are separated after pivot number $\min(i, j)$ to zero.
- Compute incomplete Cholesky factorisation.
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

Remaining problems with our approach:
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- Remaining problems with our approach:
  - Nested dissection does not lead to near-linear complexity algorithms
Sparse Factorisation of Dense Matrices: *fade-out* instead of *fill-in*

- Remaining problems with our approach:
  - Nested dissection does not lead to near-linear complexity algorithms
  - Precision matrix will not be exactly sparse. How is it localised?
Remaining problems with our approach:

- Nested dissection does not lead to near-linear complexity algorithms
- Precision matrix will not be exactly sparse. How is it localised?
- The answer can be found in the recent literature on numerical homogenisation:
“Gamlet” bases have been introduced as part of the game theoretical approach to numerical PDE (Owhadi (2017), Owhadi and Scovel (2017)).
“Gamblet” bases have been introduced as part of the game theoretical approach to numerical PDE (Owhadi (2017), Owhadi and Scovel (2017)).

Assume our covariance matrix is

$$\Theta_{i,j} = \int_{[0,1]^2} \phi_i^{(q)}(x) G(x, y) \phi_j^{(q)}(y) \, dx \, dy$$

For $$\phi_i^{(q)} := 1_{[(i-1)h^q, ih^q]}$$ and $$G$$ the Green’s function of a second order elliptic PDE.
“Gamblet” bases have been introduced as part of the game theoretical approach to numerical PDE (Owhadi (2017), Owhadi and Scovel (2017)).

Assume our covariance matrix is

$$\Theta_{i,j} = \int_{[0,1]^2} \phi_i^{(q)}(x) G(x, y) \phi_j^{(q)}(y) \, dx \, dy$$

For $\phi_i^{(q)} := 1_{[(i-1)h^q, ih^q]}$ and $G$ the Green’s function of a second order elliptic PDE.

Corresponds to $X_i(\omega) = \int_0^1 \phi_i^{(q)}(x) u(x, \omega) \, dx$, with $u(\omega)$ solution to elliptic SPDE with Gaussian forcing.
Similiar to our case, only with $\mathbb{1}_{[(i-1)h^q,ih^q]}$ instead of dirac mesure.
Similar to our case, only with $\mathbb{1}_{[(i-1)h^q, ih^q]}$ instead of dirac mesure.

For $\phi_i^{(k)} := \mathbb{1}_{[(i-1)h^k, ih^k]}$, Owhadi and Scovel (2017) shows:
Similar to our case, only with $I_{[(i-1)h^q, ih^q]}$ instead of dirac mesure.

For $\phi_i^{(k)} := I_{[(i-1)h^k, ih^k]}$, Owhadi and Scovel (2017) shows:

$\psi_i^{(k)} := \mathbb{E} \left[ u \mid \int_0^1 u(x) \phi_j^{(k)}(x) \, dx = \delta_{i,j} \right]$ is exponentially localised, on a scale of $h^k$: 
Similiar to our case, only with $\mathbb{1}_{[(i-1)h^q,ih^q]}$ instead of dirac mesure.

For $\phi_i^{(k)} := \mathbb{1}_{[(i-1)h^k,ih^k]}$, Owhadi and Scovel (2017) shows:

$$\psi_i^{(k)} := \mathbb{E} \left[u \middle| \int_0^1 u(x) \phi_j^{(k)}(x) \, dx = \delta_{i,j} \right]$$ is exponentially localised, on a scale of $h^k$:

Main idea: Estimate on exponential decay of a conditional expectation implies exponential decay of a Cholesky factors.
Transform to multiresolution basis to obtain block matrix:

\[
(\Gamma_{k,l})_{i,j} = \int_{[0,1]^2} \phi_i^{(k)}(x) G(x, y) \phi_j^{(l)}(y) \, dx \, dy
\]
Sparse factorisation of dense matrices using gamblets

- Transform to multiresolution basis to obtain block matrix:

\[
(\Gamma_{k,l})_{i,j} = \int_{[0,1]^2} \phi_i^{(k),\chi}(x) G(x, y) \phi_j^{(l),\chi}(y) \, dx \, dy
\]

- Where the \( \{ \phi_j^{(k),\chi} \}_{j \in J(k)} \) are chosen as Haar basis functions.
Then the results of Owhadi (2017) and Owhadi and Scovel (2017) imply that:
Then the results of Owhadi (2017) and Owhadi and Scovel (2017) imply that:

\[ \chi_i^{(k)} := \mathbb{E} \left[ u \mid \int_0^1 u(x) \phi_j^{(l)}, \chi(x) \, dx = \delta_{i,j} \delta_{k,l}, \forall l \leq k \right] \] is exponentially localised, on a scale of \( h^k \):

\[ \left| \chi_i^{(k)} \left( x - x_i^{(k)} \right) \right| \leq C \exp \left( -\frac{\gamma}{h^k} \| x - x_i^{(k)} \| \right). \]
Then the results of Owhadi (2017) and Owhadi and Scovel (2017) imply that:

$$\chi_i^{(k)} := \mathbb{E} \left[ u \| \int_0^1 u(x) \phi_j^{(l)}(x) \chi(x) \, dx = \delta_{i,j} \delta_{k,l}, \forall l \leq k \right]$$

is exponentially localised, on a scale of $h^k$:

$$\left| \chi_i^{(k)}(x - x_i^{(k)}) \right| \leq C \exp \left( -\frac{\gamma}{h^k} \| x - x_i^{(k)} \| \right).$$

Furthermore, the stiffness matrices decay exponentially on each level:

$$B_{i,j}^{(k)} := \int_0^1 \chi_i^{(k)}(x) G^{-1} \chi_j^{(k)}(x) \, dx \leq \exp \left( -\gamma \| x_i - x_j \| \right)$$
Then the results of Owhadi (2017) and Owhadi and Scovel (2017) imply that:

\[ \chi^{(k)}_i := \mathbb{E} \left[ u \int_0^1 u(x) \phi_j^{(l)} \psi_k^{(m)} (x) \, dx = \delta_{i,j} \delta_{k,l}, \forall l \leq k \right] \]

is exponentially localised, on a scale of \( h^k \):

\[ \left| \chi^{(k)}_i \left( x - x_i^{(k)} \right) \right| \leq C \exp \left( -\frac{\gamma}{h^k} \| x - x_i^{(k)} \| \right). \]

Furthermore, the stiffness matrices decay exponentially on each level:

\[ B^{(k)}_{i,j} := \int_0^1 \chi^{(k)}_i (x) G^{-1} \chi^{(k)}_j (x) \, dx \leq \exp \left( -\gamma \| x_i - x_j \| \right) \]

Finally, we have for a constant \( \kappa \):

\[ \text{cond} \left( B^{(k)} \right) \leq \kappa, \forall k \]
The above properties will allow us to show localisation of the (block) Cholesky factors:
The above properties will allow us to show localisation of the (block) Cholesky factors:

Consider the two-scale case:

\[
\begin{pmatrix}
\Gamma_{11} & \Gamma_{12} \\
\Gamma_{21} & \Gamma_{22}
\end{pmatrix} = \begin{pmatrix}
\text{Id} & 0 \\
\Gamma_{21} & \text{Id}
\end{pmatrix} \begin{pmatrix}
\Theta_{11} & 0 \\
0 & \Gamma_{22} - \Gamma_{21} \Gamma_{11}^{-1} \Gamma_{12}
\end{pmatrix} \begin{pmatrix}
\text{Id} & \Gamma_{11}^{-1} \Gamma_{12} \\
0 & \text{Id}
\end{pmatrix}
\]
The above properties will allow us to show localisation of the (block) Cholesky factors:

Consider the two-scale case:

\[
\begin{pmatrix}
\Gamma_{11} & \Gamma_{12} \\
\Gamma_{21} & \Gamma_{22}
\end{pmatrix}
= \begin{pmatrix}
\text{Id} & 0 \\
\Gamma_{21} \Gamma_{11}^{-1} & \text{Id}
\end{pmatrix}
\begin{pmatrix}
\Gamma_{11} & 0 \\
0 & \Gamma_{22} - \Gamma_{21} \Gamma_{11}^{-1} \Gamma_{12}
\end{pmatrix}
\begin{pmatrix}
\text{Id} & \Gamma_{11}^{-1} \Gamma_{12} \\
0 & \text{Id}
\end{pmatrix}
\]

\[
\left( \Gamma_{21} \Gamma_{11}^{-1} \right)_{i,j} = \mathbb{E} \left[ \int u_{\phi_i}^{(2)} \chi \, dx \bigg| \int u_{\phi_m}^{(1)} \chi \, dx = \delta_{j,m} \right] = \int \phi_i^{(2)} \chi \chi_j^{(1)} \, dx
\]

\[
\Gamma_{22} - \Gamma_{21} \Gamma_{11}^{-1} \Gamma_{12} = \text{Cov} \left[ \int u_{\phi_i}^{(2)} \chi \, dx \bigg| \int u_{\phi_i}^{(1)} \chi \, dx \right] = \left( B^{(2)} \right)^{-1}
\]
\[
\left( \Gamma_{21} \Gamma_{11}^{-1} \right)_{i,j} = \int \phi_i^{(2)}(x) \chi_j^{(1)}(x) \, dx \leq C \exp \left( -\frac{\gamma}{h} \| x_i^{(2)} - x_j^{(1)} \| \right)
\]
\[
\left( \Gamma_{21} \Gamma_{11}^{-1} \right)_{i,j} = \int \phi_i^{(2)} \chi_j^{(1)} \, dx \leq C \exp \left( -\frac{\gamma}{h} \left\| x_i^{(2)} - x_j^{(1)} \right\| \right)
\]

\[ (\Gamma_2 \Gamma_1^{-1})_{i,j} = \int \phi_i^{(2)}(x) \chi_j^{(1)}(x) \, dx \leq C \exp \left( -\frac{\gamma}{h} \| x_i^{(2)} - x_j^{(1)} \| \right) \]


Therefore: \[ \left( (B^{(2)})^{-1} \right)_{i,j} \leq C \exp \left( -\frac{\gamma}{h^2} \| x_i^{(2)} - x_j^{(2)} \| \right) \]
\[
\left( \Gamma_{21} \Gamma^{-1}_{11} \right)_{i,j} = \int \phi^{(2)}_i \chi_j^{(1)} \, dx \leq C \exp \left( -\frac{\gamma}{h} \left\| x^{(2)}_i - x^{(1)}_j \right\| \right)
\]


Therefore: \[
\left( (B^{(2)})^{-1} \right)_{i,j} \leq C \exp \left( -\frac{\gamma}{h^2} \left\| x^2_i - x^{(2)}_j \right\| \right).
\]

Argument can be extended to multiple scales. Results in exponentially decaying (block-)Cholesky factors.
Sparse factorisation of dense matrices using gamblets

\[
\left( \Gamma_{21} \Gamma_{11}^{-1} \right)_{i,j} = \int \phi_i^{(2)} \chi_j^{(1)} \, dx \leq C \exp \left( -\frac{\gamma}{h} \left\| x_i^{(2)} - x_j^{(1)} \right\| \right)
\]


Therefore:

\[
\left( (B^{(2)})^{-1} \right)_{i,j} \leq C \exp \left( -\frac{\gamma}{h^2} \left\| x_i^{2} - x_j^{(2)} \right\| \right).
\]

Argument can be extended to multiple scales. Results in exponentially decaying (block-)Cholesky factors.

These factors can be approximated in time complexity by (block-)Cholesky decomposition in computational complexity of

\[
\mathcal{O} \left( N \log^2 (N) \left( \log (1/\epsilon) + \log^2 (N) \right)^{4d+1} \right)
\]

in time and

\[
\mathcal{O} \left( N \log(N) \log^d \left( N \frac{1}{\epsilon} \right) \right)
\]

in space for an approximation error of \( \epsilon \).
How about $\phi_i^{(q)} = \delta_{x_i(q)}$, i.e. pointwise sampling?
How about $\phi_i^{(q)} = \delta_{x_i^{(q)}}$, i.e. pointwise sampling?

In Owhadi and Scovel (2017), analogue results for pointwise samples are obtained using averaging:
We are left with a simple algorithm:
Sparse factorisation of dense matrices using gamblets

- We are left with a simple algorithm:
- Let $\Gamma$ be $\Theta$ expressed in multiresolution basis.
We are left with a simple algorithm:

- Let $\Gamma$ be $\Theta$ expressed in multiresolution basis.
- Throw away all entries outside of $S_\rho$, defined as

$$S_\rho := \left\{(i, j) \in I \times I \middle| i \in J^{(k)}, j \in J^{(l)}, \text{dist} \left( x_i^{(k)}, x_j^{(l)} \right) \leq \rho \ast h_{\text{min}(k,l)} \right\}.$$
We are left with a simple algorithm:

- Let $\Gamma$ be $\Theta$ expressed in multiresolution basis.
- Throw away all entries outside of $S_\rho$, defined as

$$S_\rho := \left\{ (i, j) \in I \times I \mid i \in J^{(k)}, j \in J^{(l)}, \text{dist} (x_i^{(k)}, x_j^{(l)}) \leq \rho \ast h_{\text{min}}^{(k,l)} \right\}.$$

- Compute incomplete (block-)Cholesky decomposition of $\Gamma$ restricted to $S_\rho$. 
We are left with a simple algorithm:

Let $\Gamma$ be $\Theta$ expressed in multiresolution basis.

Throw away all entries outside of $S_{\rho}$, defined as

$$S_{\rho} := \left\{ (i, j) \in I \times I \mid i \in J^{(k)}, j \in J^{(l)}, \text{dist} \left( x_i^{(k)}, x_j^{(l)} \right) \leq \rho \times h_{\min(k,l)} \right\}.$$ 

Compute incomplete (block-)Cholesky decomposition of $\Gamma$ restricted to $S_{\rho}$.

Factorisation can be done in $\mathcal{O}(N \text{poly}(\rho \log(N)))$, error decays exponentially with $\rho$. 
We are left with two closely related problems:
We are left with two closely related problems:

The multiresolution basis, in order to satisfy the conditions of the proof of bounded condition numbers given in Owhadi and Scovel (2017) needs to satisfy the vanishing moment condition:

\[
\int_{\tau_i^{(k)}} p_{\phi_i}^{(k)} \chi \, dx = 0, \quad \forall p \in \mathcal{P}_{s-1} \left( \tau_i^{(k)} \right),
\]

for a \( \tau_i^{(k)} \) of diameter \( \approx h^k \) and 2s the order of the elliptic operator.
We are left with two closely related problems:

The multiresolution basis, in order to satisfy the conditions of the proof of bounded condition numbers given in Owhadi and Scovel (2017) needs to satisfy the vanishing moment condition:

\[
\int_{\tau_i^{(k)}} p \phi_i^{(k)} p, \chi \, dx = 0, \forall p \in \mathcal{P}_{s-1} \left( \tau_i^{(k)} \right),
\]

for a \( \tau_i^{(k)} \) of diameter \( \approx h^k \) and \( 2s \) the order of the elliptic operator.

Therefore, the multiresolution basis depends on the operator.
We are left with two closely related problems:

The multiresolution basis, in order to satisfy the conditions of the proof of bounded condition numbers given in Owhadi and Scovel (2017) needs to satisfy the vanishing moment condition:

\[ \int_{\tau_i^{(k)}} p \phi_{i,\chi}^{(k)} dx = 0, \forall p \in P_{s-1} \left( \tau_i^{(k)} \right), \]

for a \( \tau_i^{(k)} \) of diameter \( \approx h^k \) and \( 2s \) the order of the elliptic operator.

Therefore, the multiresolution basis depends on the operator.

Also, averaging over large regions required for coarse basis functions. Leads to \( \mathcal{O} \left( N^2 \right) \) complexity of basis transform.
Can we get rid of vanishing moment condition?
Can we get rid of vanishing moment condition?

Conditions in Owhadi and Scovel (2017) are (roughly speaking):

\[
\frac{1}{C}H^k \leq \lambda_{\text{min}} \left( \Theta |_{\Phi^{(k)}} \right) \\
\lambda_{\text{max}} \left( \Theta |_{\perp \Phi^{(k-1)}} \right) \leq CH^{k-1}.
\]
Can we get rid of vanishing moment condition?

Conditions in Owhadi and Scovel (2017) are (roughly speaking):

\[
\frac{1}{C} H^k \leq \lambda_{\min}(\Theta|_{\phi(k)}) \\
\lambda_{\max}(\Theta|_{\perp \phi(k-1)}) \leq CH^{k-1}.
\]

Moving to finer scales, the discrete space contains more and more oscillatory functions (small eigenvalues).
Can we get rid of vanishing moment condition?

Conditions in Owhadi and Scovel (2017) are (roughly speaking):

\[
\frac{1}{C} H^k \leq \lambda_{\text{min}} (\Theta|_{\Phi(k)}) \\
\lambda_{\text{max}} (\Theta|_{\perp \Phi(k-1)}) \leq CH^{k-1}.
\]

Moving to finer scales, the discrete space contains more and more oscillatory functions (small eigenvalues).

But its in the orthogonal complement, of a larger space, low modes are “projected out”.
Can we get rid of vanishing moment condition?

Conditions in Owhadi and Scovel (2017) are (roughly speaking):

\[
\frac{1}{C} H^k \leq \lambda_{\text{min}} (\Theta|_{\Phi^{(k)}}) \\
\lambda_{\text{max}} (\Theta|_{\perp \Phi^{(k-1)}}) \leq CH^{k-1}.
\]

Moving to finer scales, the discrete space contains more and more oscillatory functions (small eigenvalues).

But its in the orthogonal complement, of a larger space, low modes are "projected out".

Balance of these effects leads to bounded condition numbers.
- Gamblets are more robust!
Gamblets are more robust!

Can replace the conditions with (roughly speaking):

$$\frac{1}{C} H^k \leq \lambda_{\text{min}} \left( \Theta \big|_{\Phi(k)} \right)$$

$$\max_{\phi \in \Phi^k, \|\phi\| = 1} \min_{\varphi \in \Phi^{k-1}, \|\varphi\| \leq C} (\phi - \varphi)^T \Theta (\phi - \varphi) \leq CH^{k-1}.$$
Gamblets are more robust!

Can replace the conditions with (roughly speaking):

$$\frac{1}{C} H^k \leq \lambda_{\min} \left( \Theta |\Phi^{(k)} \right)$$

$$\max_{\phi \in \Phi^k, \|\phi\|=1} \min_{\phi \in \Phi^{k-1}: \|\phi\| \leq C} (\phi - \varphi)^T \Theta (\phi - \varphi) \leq C H^{k-1}.$$

The gamblets find the optimal orthogonalisation themselves!
We can use subsampling as an aggregation scheme!

\[ \phi_i^{(1)} = \delta(x - x_i^{(1)}) \]

\[ \phi_j^{(2)} = \delta(x - x_j^{(2)}) \]

\[ \phi_i^{(3)} = \delta(x - x_i^{(3)}) \]

\[ \Omega \]

\[ \pi_{i,1}^{(1,2)} \]

\[ \pi_{i,2}^{(2,3)} \]

\[ W_{i,2}^{(2)} \]
Our algorithm now consists of three steps:

1. Reorder the variables hierarchically
2. Obtain the entries in $S_2$ (or more generally $S_\rho$), set other entries to zero.
3. Compute the incomplete Cholesky decomposition
Our algorithm now consists of three steps:

1. Reorder the variables hierarchically
2. Obtain the entries in $S_2$ (or more generally $S_\rho$), set other entries to zero.
3. Compute the incomplete Cholesky decomposition

At this point, for theoretical guarantees we need to replace step three with an incomplete Block factorisation. All numerical evidence indicates that this is not necessary.
As observed in Owhadi 2017, Hou and Zhang 2017, gamblets provide a near-optimal sparse PCA. We obtain a PCA with the same approximation property, by keeping only the first $k$ columns of $L$. 
As observed in Owhadi 2017, Hou and Zhang 2017, gamblets provide a near-optimal sparse PCA. We obtain a PCA with the same approximation property, by keeping only the first $k$ columns of $L$.

By reversing the elimination ordering, we obtain a near linear complexity Cholesky factorisation of the sparse/exponentially decaying inverse of $\Theta$. 
Problems at the boundary

Figure: $\nu = 1, \ l = 0.4$

$log_{10}(L_{:,j})$
Problems at the boundary

$K_{N,j}$

$\|x_N - x_j\|$
Decay of approximation error

$\log_{10}(E_{rel})$ vs $\rho$

- $\nu = 1, l = 0.2$
- $\nu = 1, l = 0.4$
- $\nu = 2, l = 0.2$
Figure: Near optimal sparse PCA: First panel: $\nu = 1$, $l = 0.2$, $\delta_x = 0.2$ and $\rho = 6$. Second panel: $\nu = 2$, $l = 0.2$ and $\delta_x = 0.2$ and $\rho = 8$. 

Perturbation of the Mesh

\[
\delta_x \parallel \Gamma^\rho - \Gamma \parallel \parallel \Gamma \parallel \parallel \Gamma^\rho - \Gamma \parallel_{\text{Fro}} \parallel \Gamma \parallel_{\text{Fro}}^2
\]

<table>
<thead>
<tr>
<th>$\delta_x$</th>
<th>$\parallel \Gamma^\rho - \Gamma \parallel$</th>
<th>$\parallel \Gamma^\rho - \Gamma \parallel / \parallel \Gamma \parallel$</th>
<th>$\parallel \Gamma^\rho - \Gamma \parallel_{\text{Fro}}$</th>
<th>$\parallel \Gamma^\rho - \Gamma \parallel_{\text{Fro}} / \parallel \Gamma \parallel_{\text{Fro}}$</th>
<th>#S</th>
<th>#S/!N^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>4.336e-03</td>
<td>1.560e-06</td>
<td>1.669e-02</td>
<td>1.026e-06</td>
<td>2.125e+07</td>
<td>7.675e-02</td>
</tr>
<tr>
<td>0.4</td>
<td>4.495e-03</td>
<td>1.617e-06</td>
<td>1.706e-02</td>
<td>1.063e-06</td>
<td>2.128e+07</td>
<td>7.683e-02</td>
</tr>
<tr>
<td>2.0</td>
<td>4.551e-03</td>
<td>1.638e-06</td>
<td>1.820e-02</td>
<td>1.077e-06</td>
<td>2.127e+07</td>
<td>7.682e-02</td>
</tr>
<tr>
<td>4.0</td>
<td>8.158e-03</td>
<td>2.940e-06</td>
<td>2.976e-02</td>
<td>1.933e-06</td>
<td>2.119e+07</td>
<td>7.652e-02</td>
</tr>
</tbody>
</table>

**Table:** Compression and accuracy for $q = 7$, $l = 0.2$, $\rho = 5$, $\nu = 1$ and different values of $\delta_x$. 
Data on low dimensional manifold

$$\delta_z \frac{\|\Gamma^\rho - \Gamma\|}{\|\Gamma\|} \frac{\|\Gamma^\rho - \Gamma\|_{\text{Fro}}}{\|\Gamma\|_{\text{Fro}}} \#S \#S/N^2$$

<table>
<thead>
<tr>
<th>$\delta_z$</th>
<th>$|\Gamma^\rho - \Gamma|$</th>
<th>$|\Gamma^\rho - \Gamma|/|\Gamma|$</th>
<th>$|\Gamma^\rho - \Gamma|_{\text{Fro}}$</th>
<th>$|\Gamma^\rho - \Gamma|<em>{\text{Fro}}/|\Gamma|</em>{\text{Fro}}$</th>
<th>#S</th>
<th>#S/N^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>5.049e-03</td>
<td>1.560e-06</td>
<td>1.885e-02</td>
<td>1.026e-06</td>
<td>2.126e+07</td>
<td>7.677e-02</td>
</tr>
<tr>
<td>0.1</td>
<td>6.341e-02</td>
<td>1.648e-06</td>
<td>1.232e-01</td>
<td>1.077e-06</td>
<td>2.083e+07</td>
<td>7.521e-02</td>
</tr>
<tr>
<td>0.2</td>
<td>1.204e-01</td>
<td>1.749e-06</td>
<td>2.203e-01</td>
<td>1.126e-06</td>
<td>1.976e+07</td>
<td>7.137e-02</td>
</tr>
<tr>
<td>0.4</td>
<td>1.954e-01</td>
<td>3.550e-06</td>
<td>5.098e-01</td>
<td>2.197e-06</td>
<td>1.722e+07</td>
<td>6.218e-02</td>
</tr>
</tbody>
</table>

Table: Compression and accuracy for $q = 7$, $l = 0.2$, $\rho = 5$, $\nu = 1$, $\delta_x = 2$ and different values of $\delta_z$. 
Table: Compression and accuracy for $q = 7$, $l = 0.2$, $\rho = 6$, $\delta_x = 0.2$ and different values of $\nu$.