DARTMOUTH
Department of Mathematics

OBLIQUE PROJECTIONS AND LOW-RANK STRUCTURE IN INVERSE PROBLEMS

Graduate Student Seminar

22nd September 2023

Jonathan Lindbloom

OUTLINE

OBLIQUE PROJECTIONS

Orthogonal projection operator

Let $\mathcal{X} \subset \mathbb{R}^n$ be a subspace. Then the orthogonal projection operator $\boldsymbol{P}_{\mathcal{X}}(\cdot)$ is the linear operator satisfying

1. $\forall \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{P}_{\mathcal{X}}(\boldsymbol{x}) = \boldsymbol{x}$
2. $\forall \boldsymbol{x} \in \mathcal{X}^{\perp},\ \boldsymbol{P}_{\mathcal{X}}(\boldsymbol{x}) = \mathbf{0}$

Orthogonal projection operator

Let $\mathcal{X} \subset \mathbb{R}^n$ be a subspace. Then the orthogonal projection operator $\boldsymbol{P}_{\mathcal{X}}(\cdot)$ is the linear operator satisfying

1. $\forall \boldsymbol{x} \in \mathcal{X}, \ \boldsymbol{P}_{\mathcal{X}}(\boldsymbol{x}) = \boldsymbol{x}$
2. $\forall \boldsymbol{x} \in \mathcal{X}^{\perp}, \ \boldsymbol{P}_{\mathcal{X}}(\boldsymbol{x}) = \boldsymbol{0}$

Decomposition of vectors

We can decompose any $\boldsymbol{u} \in \mathbb{R}^n$ uniquely as

$$\boldsymbol{u} = \boldsymbol{x} + \boldsymbol{x}_{\perp}$$

where $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{x}_{\perp} \in \mathcal{X}^{\perp}$.

Optimization representation

The orthogonal projector $\boldsymbol{P}_{\mathcal{X}}(\cdot)$ can be expressed as

$$\boldsymbol{P}_{\mathcal{X}}(\boldsymbol{x}) = \arg\min_{\hat{\boldsymbol{x}} \in \mathcal{X}} \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2$$

for any $\boldsymbol{x} \in \mathbb{R}^n$.

Matrix representation

The orthogonal projector $\boldsymbol{P}_{\mathcal{X}}$ can be represented by the matrix

$$\boldsymbol{P}_{\mathcal{X}} = \boldsymbol{X}\boldsymbol{X}^{\dagger}$$

for any matrix $\boldsymbol{X}$ such that $\mathcal{X} = \operatorname{range}(\boldsymbol{X})$. If we furthermore require that the columns of $\boldsymbol{X}$ are linearly independent, then this specializes to

$$\boldsymbol{P}_{\mathcal{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

Oblique projection operator

Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ be subspaces that intersect trivially. Then the projection onto $\mathcal{X}$ along $\mathcal{Y}$ is the linear operator $\boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\cdot)$ satisfying

1. $\forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}) = \boldsymbol{x}$
2. $\forall \boldsymbol{y} \in \mathcal{Y}, \quad \boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{y}) = \boldsymbol{0}$
3. $\forall \boldsymbol{z} \in \mathbb{R}^n, \quad \boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{z}) \in \mathcal{X}$

Oblique projection operator

Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ be subspaces that intersect trivially. Then the projection onto $\mathcal{X}$ along $\mathcal{Y}$ is the linear operator $\boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\cdot)$ satisfying

1. $\forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}) = \boldsymbol{x}$
2. $\forall \boldsymbol{y} \in \mathcal{Y}, \quad \boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{y}) = \boldsymbol{0}$
3. $\forall \boldsymbol{z} \in \mathbb{R}^n, \quad \boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{z}) \in \mathcal{X}$

Decomposition of vectors

We can decompose any $\boldsymbol{u} \in \mathbb{R}^n$ uniquely as

$$\boldsymbol{u} = \boldsymbol{x} + \boldsymbol{y} + \boldsymbol{z}$$

where $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$, $\boldsymbol{z} \in (\mathcal{X} \cup \mathcal{Y})^{\perp}$.

Optimization representation

The oblique projector $\boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\cdot)$ can be expressed as

$$\boldsymbol{E}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{z}) = \boldsymbol{X}\left(\underset{\boldsymbol{w} \text{ s.t. } \boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{z})=\boldsymbol{0}}{\arg\min} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{z}\|_2\right)$$

for any matrix $\boldsymbol{X}$ such that $\mathcal{X} = \mathrm{range}(\boldsymbol{X})$ and any matrix $\boldsymbol{Y}$ such that $\mathcal{Y}^\perp = \mathrm{range}(\boldsymbol{Y})$.

Matrix representation

The oblique projector $\boldsymbol{E}_{\mathcal{X},\mathcal{Y}}$ can be represented by the matrix

$$\boldsymbol{E}_{\mathcal{X},\mathcal{Y}} = \boldsymbol{X}\left(\boldsymbol{Y}^T\boldsymbol{X}\right)^{\dagger}\boldsymbol{Y}^T$$

for any matrix $\boldsymbol{X}$ such that $\mathcal{X} = \mathrm{range}(\boldsymbol{X})$ and any matrix $\boldsymbol{Y}$ such that $\mathcal{Y}^{\perp} = \mathrm{range}(\boldsymbol{Y})$.

Some identities

$$\boldsymbol{E}_{\mathcal{X},\mathcal{Y}} + \boldsymbol{E}_{\mathcal{Y},\mathcal{X}} = \boldsymbol{P}_{\mathcal{X}\cup\mathcal{Y}}$$
$$\boldsymbol{E}_{\mathcal{X},\mathcal{Y}} + \boldsymbol{E}_{\mathcal{Y},\mathcal{X}} + \boldsymbol{P}_{(\mathcal{X}\cup\mathcal{Y})^\perp} = \boldsymbol{I}$$

$\boldsymbol{A}$-orthogonality and oblique complement

Let $\boldsymbol{x} \perp_{\boldsymbol{A}} \boldsymbol{y}$ denote

$$\boldsymbol{x} \perp_{\boldsymbol{A}} \boldsymbol{y} \Leftrightarrow \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{y} = 0.$$

If $\mathcal{X} \subset \mathbb{R}^n$ is a subspace, then we say that

$$\mathcal{X}^{\perp_{\boldsymbol{A}}} = \{\boldsymbol{y} \in \mathbb{R}^n : \forall \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{x} \perp_{\boldsymbol{A}} \boldsymbol{y}\}$$

is its oblique complement w.r.t. $\boldsymbol{A}$. For the oblique projector $\boldsymbol{E}_{\mathcal{X},\mathcal{X}^{\perp_{\boldsymbol{A}}}}$, we just write $\boldsymbol{E}_{\mathcal{X}}$.

Matrix representation and splitting

The oblique projector $\boldsymbol{E}_{\mathcal{X}} = \boldsymbol{E}_{\mathcal{X},\mathcal{X}^{\perp_{\boldsymbol{A}}}}$ can be expressed as

$$\boldsymbol{E}_{\mathcal{X}} = \boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^{\dagger}\boldsymbol{A},$$

for any matrix satisfying $\mathcal{X} = \mathrm{range}(\boldsymbol{X})$. Also, we can split any vector $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\boldsymbol{x} = \boldsymbol{E}_{\mathcal{X}}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{E}_{\mathcal{X}})\boldsymbol{x}$$

which satisfies

$$\boldsymbol{E}_{\mathcal{X}}\boldsymbol{x} \perp_{\boldsymbol{A}} (\boldsymbol{I} - \boldsymbol{E}_{\mathcal{X}})\boldsymbol{x}.$$

Why is

$$\boldsymbol{E}_{\mathcal{X}}\boldsymbol{x} \perp_{\boldsymbol{A}} (\boldsymbol{I} - \boldsymbol{E}_{\mathcal{X}})\boldsymbol{x}?$$

Let $\boldsymbol{E}_{\mathcal{X}} = \boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^{\dagger}\boldsymbol{A}$. Then we can show $\boldsymbol{A}$-orthogonality by showing that

$$\left\langle \boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^{\dagger}\boldsymbol{A}\boldsymbol{x}, \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^{\dagger}\boldsymbol{A}\right)\boldsymbol{x} \right\rangle_{\boldsymbol{A}^T\boldsymbol{A}} = 0.$$

Expanding, we see that

$$
\begin{aligned}
\langle \ldots, \ldots \rangle_{\boldsymbol{A}^T\boldsymbol{A}} &= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T((\boldsymbol{A}\boldsymbol{X})^\dagger)^T\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} \\
&= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T((\boldsymbol{A}\boldsymbol{X})^\dagger)^T(\boldsymbol{A}\boldsymbol{X})^T(\boldsymbol{A}\boldsymbol{X})(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} \\
&= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T((\boldsymbol{A}\boldsymbol{X})^\dagger)^T(\boldsymbol{A}\boldsymbol{X})^T(\boldsymbol{A}\boldsymbol{X})(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} \\
&= \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{X})(\boldsymbol{A}\boldsymbol{X})^\dagger\boldsymbol{A}\boldsymbol{x} \\
&= 0
\end{aligned}
$$

since $\forall \boldsymbol{B} \in \mathbb{R}^{m \times n}$,

$$(\boldsymbol{B}^\dagger)^T\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{B}.$$

Oblique pseudoinverse

Let $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ with $p \leq n$ such that $\mathcal{X} = \mathrm{range}(\boldsymbol{X})$. Then we define the oblique pseudoinverse as $\boldsymbol{X}_{\mathcal{Y}}^{\dagger} \in \mathbb{R}^{n\times p}$ where

$$\boldsymbol{X}_{\mathcal{Y}}^{\dagger} = \boldsymbol{E}_{\mathcal{Y},\ker(\boldsymbol{X})}\boldsymbol{X}^{\dagger}.$$

Oblique pseudoinverse

Let $\boldsymbol{X} \in \mathbb{R}^{p \times n}$ with $p \leq n$ such that $\mathcal{X} = \mathrm{range}(\boldsymbol{X})$. Then we define the oblique pseudoinverse as $\boldsymbol{X}_{\mathcal{Y}}^{\dagger} \in \mathbb{R}^{n \times p}$ where

$$\boldsymbol{X}_{\mathcal{Y}}^{\dagger} = \boldsymbol{E}_{\mathcal{Y}, \ker(\boldsymbol{X})} \boldsymbol{X}^{\dagger}.$$

If $\mathcal{Y} = \ker(\boldsymbol{X})^{\perp}$, then $\boldsymbol{X}_{\mathcal{Y}}^{\dagger} = \boldsymbol{X}^{\dagger}$ (just the Moore-Penrose inverse).

Properties of oblique pseudoinverse

1. $\boldsymbol{X}\boldsymbol{X}_{\mathcal{Y}}^{\dagger} = \boldsymbol{P}_{\mathcal{X}}$
2. $\boldsymbol{X}_{\mathcal{Y}}^{\dagger}\boldsymbol{X} = \boldsymbol{E}_{\mathcal{Y},\ker(\boldsymbol{X})}$
3. $\boldsymbol{X}^{\dagger} = \boldsymbol{P}_{\mathrm{range}(\boldsymbol{X}^T)}\boldsymbol{X}_{\mathcal{Y}}^{\dagger}$
4. If $\mathcal{Y} = \mathrm{range}(\boldsymbol{Y})$, then $\boldsymbol{X}_{\mathcal{Y}}^{\dagger} = \boldsymbol{Y}(\boldsymbol{X}\boldsymbol{Y})^{\dagger}$.

REGULARIZED LEAST-SQUARES

Motivation: for general, regularized least-squares problems of the form

$$\boldsymbol{x}^\star = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{F}\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}\|_2^2,$$

with $\boldsymbol{F} \in \mathbb{R}^{m\times n}$, $\boldsymbol{R} \in \mathbb{R}^{k\times n}$, $\ker(\boldsymbol{F}) \cap \ker(\boldsymbol{R}) = \{\boldsymbol{0}\}$, we often would like to convert this using a change-of-variables to solving a problem of the form

$$\boldsymbol{z}^\star = \operatorname*{arg\,min}_{\boldsymbol{z}\in\mathbb{R}^k} \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2$$

for some $\boldsymbol{A}$ to be determined, and some relation between $\boldsymbol{z}$ and $\boldsymbol{x}$ to be determined.

Why would we like to convert to standard form? The solution we desire is given explicitly by

$$\boldsymbol{x} = \left(\boldsymbol{F}^T\boldsymbol{F} + \boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\boldsymbol{F}^T\boldsymbol{y}.$$

For high-dimensional problems, we must employ iterative methods such as the Conjugate Gradient method to apply the inverse to a vector. The efficiency of this method depends highly on the condition number of $\boldsymbol{Q} = \boldsymbol{F}^T\boldsymbol{F} + \boldsymbol{R}^T\boldsymbol{R}$. The (heuristic) observation is that for typical choices of $\boldsymbol{F}$ and $\boldsymbol{R}$, making a change-of-variables and dealing instead with $\tilde{\boldsymbol{Q}} = \boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{I}$ gives a matrix with better conditioning and thus easier/quicker to apply the needed inverse.

Strongest assumption: if we assume that $\boldsymbol{R}^{-1}$ exists, then with the change-of-variables $\boldsymbol{z} = \boldsymbol{R}\boldsymbol{x} \Leftrightarrow \boldsymbol{x} = \boldsymbol{R}^{-1}\boldsymbol{z}$ we obtain the solution by solving

$$\boldsymbol{z}^\star = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\arg\min}\ \|\boldsymbol{F}\boldsymbol{R}^{-1}\boldsymbol{z} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2$$

and recovering $\boldsymbol{x}^\star = \boldsymbol{R}^{-1}\boldsymbol{z}^\star$.

A slightly weaker assumption: if we assume that $\ker(\boldsymbol{R}) = \{\mathbf{0}\}$ ($\boldsymbol{R}$ has linearly independent columns), then $\boldsymbol{R}^T\boldsymbol{R}$ is invertible and a matrix square root such as the Cholesky factor $\boldsymbol{L}$ in $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{L}\boldsymbol{L}^T$ exists and can be computed. With the change-of-variables $\boldsymbol{z} = \boldsymbol{L}^T\boldsymbol{x} \Leftrightarrow \boldsymbol{x} = \boldsymbol{L}^{-T}\boldsymbol{z}$, we obtain the solution by solving

$$\boldsymbol{z}^\star = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\arg\min}\ \|\boldsymbol{F}\boldsymbol{L}^{-T}\boldsymbol{z} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2$$

and recovering $\boldsymbol{x}^\star = \boldsymbol{L}^{-T}\boldsymbol{z}^\star$.

But what to do when $\boldsymbol{R}$ not invertible and has a nontrivial kernel?

Oblique projections to the rescue!

Oblique projections to the rescue!

$$\boldsymbol{x}^\star = \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathbb{R}^n} \; \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}\|_2^2$$

Oblique projections to the rescue!

$$\boldsymbol{x}^\star = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \ \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}\|_2^2$$

Consider the splitting $\mathbb{R}^n = \ker(\boldsymbol{R}) \cup \ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}$, and for the solution $\boldsymbol{x}^\star = \boldsymbol{x}_1 + \boldsymbol{x}_2$.

Oblique projections to the rescue!

$$\boldsymbol{x}^\star = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \;\|\boldsymbol{F}\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}\|_2^2$$

Consider the splitting $\mathbb{R}^n = \ker(\boldsymbol{R}) \cup \ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}$, and for the solution $\boldsymbol{x}^\star = \boldsymbol{x}_1 + \boldsymbol{x}_2$. Then, inserting the splitting, we arrive at two separate problems

$$\operatorname*{arg\,min}_{\boldsymbol{x}_1\in\ker(\boldsymbol{R})} \;\|\boldsymbol{F}\boldsymbol{x}_1-\boldsymbol{y}\|_2^2, \qquad \operatorname*{arg\,min}_{\boldsymbol{x}_2\in\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}} \;\|\boldsymbol{F}\boldsymbol{x}_2-\boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}_2\|_2^2.$$

For the second problem, we need the oblique projector $\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}$. This is given by

$$\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}} = \boldsymbol{R}^{\dagger}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}} \boldsymbol{R},$$

for any $\boldsymbol{W}$ such that $\operatorname{span}(\boldsymbol{W}) = \ker(\boldsymbol{R})$. The oblique pseudoinverse can be expressed as

$$\boldsymbol{R}^{\dagger}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}} = \left(\boldsymbol{I} - \boldsymbol{W}(\boldsymbol{F}\boldsymbol{W})^{\dagger}\boldsymbol{F}\right)\boldsymbol{R}^{\dagger}.$$

We can also show that

$$\boldsymbol{R}\boldsymbol{R}^{\dagger}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{R} = \boldsymbol{R}$$

So we see that

$$\underset{\boldsymbol{x}_2 \in \ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{x}_2 - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}_2\|_2^2$$

So we see that

$$\underset{\boldsymbol{x}_2 \in \ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}{\arg\min} \quad \|\boldsymbol{F}\boldsymbol{x}_2 - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}_2\|_2^2$$

is the same as solving

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{x}\|_2^2$$

So we see that

$$\underset{\boldsymbol{x}_2 \in \ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}{\arg\min} \quad \|\boldsymbol{F}\boldsymbol{x}_2 - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}_2\|_2^2$$

is the same as solving

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{x}\|_2^2$$

which is the same as

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{R}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{R}^{\#}\boldsymbol{R}\boldsymbol{x}\|_2^2$$

So we see that

$$\underset{\boldsymbol{x}_2 \in \ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}{\arg\min} \quad \|\boldsymbol{F}\boldsymbol{x}_2 - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}_2\|_2^2$$

is the same as solving

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{E}_{\ker(\boldsymbol{R})^{\perp_{\boldsymbol{F}}}}\boldsymbol{x}\|_2^2$$

which is the same as

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{R}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{R}^{\#}\boldsymbol{R}\boldsymbol{x}\|_2^2$$

which is the same as

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\arg\min} \; \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{R}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}\|_2^2.$$

This final problem can be written as

$$\underset{\boldsymbol{z} \in \mathrm{range}(\boldsymbol{R})}{\arg\min} \quad \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{z} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2.$$

What have we accomplished?

This final problem can be written as

$$\underset{\boldsymbol{z}\in\operatorname{range}(\boldsymbol{R})}{\arg\min} \ \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{z}-\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2.$$

What have we accomplished? It turns out, we can show that the solution to this problem is the same as the solution to the unconstrained problem

$$\underset{\boldsymbol{z}\in\mathbb{R}^k}{\arg\min} \ \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{z}-\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2.$$

SUMMARY

We have shown that the solution $\boldsymbol{x}^\star$ to

$$\boldsymbol{x}^\star = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\ \|\boldsymbol{F}\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \|\boldsymbol{R}\boldsymbol{x}\|_2^2$$

can be written as

$$\boldsymbol{x}^\star = \boldsymbol{R}^{\#}\boldsymbol{z}^\star + \boldsymbol{W}(\boldsymbol{F}\boldsymbol{W})^{\dagger}\boldsymbol{y}$$

where

$$\begin{aligned}\boldsymbol{z}^\star &= \underset{\boldsymbol{z}\in\mathbb{R}^k}{\arg\min}\ \|\boldsymbol{F}\boldsymbol{R}^{\#}\boldsymbol{z}-\boldsymbol{y}\|_2^2 + \|\boldsymbol{z}\|_2^2 \\ &= \left((\boldsymbol{R}^{\#})^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{R}^{\#} + \boldsymbol{I}\right)^{-1}(\boldsymbol{F}\boldsymbol{R}^{\#})^T\boldsymbol{y}.\end{aligned}$$

LOW RANK STRUCTURE

MATRIX DETERMINANT LEMMA

Matrix Determinant Lemma (part 1)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Then

$$\det\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\right)^{-1} = \left(1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}\right)\det(\boldsymbol{A})$$

for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$.

MATRIX DETERMINANT LEMMA

Matrix Determinant Lemma (part 2)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be invertible, and let $\boldsymbol{U} \in \mathbb{R}^{n\times k}$, $\boldsymbol{V} \in \mathbb{R}^{n\times k}$. Then

$$\det\left(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{V}^T\right) = \det\left(\boldsymbol{I}_k + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U}\right)\det(\boldsymbol{A}).$$

SM IDENTITY

Sherman-Morrison Identity

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be invertible and let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. Then $\boldsymbol{B} := \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T$ is invertible iff $1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$, in which case

$$\boldsymbol{B}^{-1} = \left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

SMW IDENTITY

Sherman-Morrison-Woodbury Identity

We have

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$$

when all of these products and inverses make sense.

GAUSSIAN SAMPLING

Skipping many of these details, in statistical inverse problems we often find ourselves in the situation that we would like to sample from the Gaussian

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{F}^T \boldsymbol{F} + \boldsymbol{Q}, \quad \boldsymbol{\mu} = \boldsymbol{Q}^{-1} \boldsymbol{F}^T \boldsymbol{y}.$$

GAUSSIAN SAMPLING

Skipping many of these details, in statistical inverse problems we often find ourselves in the situation that we would like to sample from the Gaussian

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{F}^T \boldsymbol{F} + \boldsymbol{Q}, \quad \boldsymbol{\mu} = \boldsymbol{Q}^{-1} \boldsymbol{F}^T \boldsymbol{y}.$$

Since $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{A}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, we know that if we could find a square root factorization $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$ then we could draw a sample from this Gaussian.

GAUSSIAN SAMPLING

Skipping many of these details, in statistical inverse problems we often find ourselves in the situation that we would like to sample from the Gaussian

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{F}^T \boldsymbol{F} + \boldsymbol{Q}, \quad \boldsymbol{\mu} = \boldsymbol{Q}^{-1} \boldsymbol{F}^T \boldsymbol{y}.$$

Since $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{A}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, we know that if we could find a square root factorization $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$ then we could draw a sample from this Gaussian.

But the problem is that even though computing a square root factorization of $\boldsymbol{Q}$ may be feasible, computing a square root factorization of $\boldsymbol{F}^T \boldsymbol{F} + \boldsymbol{Q}$ may not be feasible.

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \left(\boldsymbol{F}^T\boldsymbol{F} + \boldsymbol{Q}\right)^{-1} \\
&= \left(\boldsymbol{F}^T\boldsymbol{F} + \boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{Q}^{\frac{1}{2}}\right)^{-1} \\
&= \left(\boldsymbol{Q}^{\frac{1}{2}}\left(\boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{Q}^{-\frac{1}{2}} + \boldsymbol{I}_n\right)\boldsymbol{Q}^{\frac{1}{2}}\right)^{-1} \\
&= \boldsymbol{Q}^{-\frac{1}{2}}\left(\boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{Q}^{-\frac{1}{2}} + \boldsymbol{I}_n\right)^{-1}\boldsymbol{Q}^{-\frac{1}{2}}.
\end{aligned}
$$

If the posterior covariance is close to a low-rank update of the prior covariance, then

$$\boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{Q}^{-\frac{1}{2}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T$$

is a good approximation.

Then, by the SMW identity we have

$$
\begin{aligned}
\left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T + \boldsymbol{I}_n\right)^{-1} &= \boldsymbol{I}_n - \boldsymbol{V}\left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}^T\boldsymbol{V}\right)^{-1}\boldsymbol{V}^T \\
&= \boldsymbol{I}_n - \boldsymbol{V}\left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{I}\right)^{-1}\boldsymbol{V}^T \\
&= \boldsymbol{I}_n - \boldsymbol{V}\left(\operatorname{diag}\left(\frac{\lambda_i + 1}{\lambda_i}\right)\right)^{-1}\boldsymbol{V}^T \\
&= \boldsymbol{I}_n - \boldsymbol{V}\operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + 1}\right)\boldsymbol{V}^T \\
&= \boldsymbol{I}_n - \boldsymbol{V}_r\operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + 1}\right)\boldsymbol{V}_r^T - \sum_{i=r+1}^{n}\operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + 1}\right)\boldsymbol{v}_i\boldsymbol{v}_i^T \\
&\approx \boldsymbol{I}_n - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T
\end{aligned}
$$

where $\boldsymbol{D}_r := \operatorname{diag}(\frac{\lambda_i}{\lambda_i+1}) \in \mathbb{R}^{r\times r}$.

The final expression for the covariance is

$$\boldsymbol{\Sigma} \approx \boldsymbol{Q}^{-\frac{1}{2}} \left(\boldsymbol{I} - \boldsymbol{V}_r \boldsymbol{D}_r \boldsymbol{V}_r^T \right) \boldsymbol{Q}^{-\frac{1}{2}}.$$

It turns out that this approximation also provides us with an expression for a square root of the covariance:

$$\left(\boldsymbol{I}_n - \boldsymbol{V}_r \boldsymbol{D}_r \boldsymbol{V}_r^T \right)^{1/2} = \boldsymbol{I}_n - \boldsymbol{V}_r \left[\boldsymbol{I}_n \pm (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}} \right] \boldsymbol{V}_r^T.$$

$$
\begin{aligned}
&\left(\boldsymbol{I}_n - \boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T\right)\left(\boldsymbol{I}_n - \boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T\right)^T \\
&= \boldsymbol{I}_n - 2\,\boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T + \boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T\boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T \\
&= \boldsymbol{I}_n - 2\,\boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T + \boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T \\
&= \boldsymbol{I}_n - 2\,\boldsymbol{V}_r\left[\boldsymbol{I}_n + (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T + \boldsymbol{V}_r\left[\boldsymbol{I}_n + 2\,(\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}} + \boldsymbol{I}_n - \boldsymbol{D}_r\right]\boldsymbol{V}_r^T \\
&= \boldsymbol{I}_n + \boldsymbol{V}_r\left[-2\boldsymbol{I}_n - 2\,(\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}}\right]\boldsymbol{V}_r^T + \boldsymbol{V}_r\left[2\boldsymbol{I}_n + 2\,(\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}} - \boldsymbol{D}_r\right]\boldsymbol{V}_r^T \\
&= \boldsymbol{I}_n - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T
\end{aligned}
$$

So a square root of the covariance is

$$\boldsymbol{\Sigma}^{\frac{1}{2}} \approx \boldsymbol{Q}^{-\frac{1}{2}} \left(\boldsymbol{I}_n - \boldsymbol{V}_r \left[\boldsymbol{I}_n \pm (\boldsymbol{I}_n - \boldsymbol{D}_r)^{\frac{1}{2}} \right] \boldsymbol{V}_r^T \right).$$

This has the nice benefit of letting us take advantage of a square root factorization of $\boldsymbol{Q}$, which may be much cheaper to compute than for $\boldsymbol{\Sigma}$.

REFERENCES

Per Christian Hansen
"Discrete Inverse Problems: Insight and Algorithms"
Society for Industrial and Applied Mathematics, 2010

Håvard Rue, Leonhard Held
"Gaussian Markov Random Fields"
CRC Press, 2005

Alessio Spantini
"On the low-dimensional structure of Bayesian inference"
Massachusetts Institute of Technology, 2017