Bayesian Optimal Experiment Design via Column Subset Selection

Srinivas Eswar, Amit N. Subrahmanya, Vishwas Rao, Arvind K. Saibaba

Abstract

Inverse problems involve the process of calculating parameters of a mathematical model from observational data [3]. Often these problems are ill-posed and a Bayesian approach is used to produce a posterior distribution for the unobservable parameters. A key question is "how best to acquire data" in such a setting. We consider the case of Bayesian linear inverse problems where there are m candidate sensor locations, and we need to pick the k "best" ones.

Consider the measurement equation

$$\mathbf{d} = \mathbf{F}\mathbf{m} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{d} \in \mathbb{R}^m$ is the data, $\mathbf{F} \in \mathbb{R}^{m\times n}$ is the mathematical model, and $\mathbf{m} \in \mathbb{R}^n$ is the parameter to be reconstructed. The observations are assumed to be perturbed with additive uncorrelated Gaussian noise, i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{noise}})$. We assume that $m < n$, which makes the problem underdetermined. If we assume our prior to also be Gaussian, $\mathbf{m} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{pr}}, \boldsymbol{\Gamma}_{\text{pr}})$, the posterior will also be a Gaussian with covariance $\boldsymbol{\Gamma}_{\text{post}} = (\mathbf{F}^{\mathsf{T}}\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{F} + \boldsymbol{\Gamma}_{\text{pr}}^{-1})^{-1}$ and mean $\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Gamma}_{\text{post}}(\mathbf{F}^{\mathsf{T}}\boldsymbol{\Gamma}_{\text{noise}}^{-1}\mathbf{d} + \boldsymbol{\Gamma}_{\text{pr}}^{-1}\boldsymbol{\mu}_{\text{pr}})$.

The rows of $\mathbf{F}$ correspond to the m different candidate sensor locations and we would like to select only k locations to collect data. To determine the optimal sensor placements, we solve the following combinatorial optimization problem

$$\min_{W\subset\{1,\cdots,m\}} \phi(W), \quad \text{subject to } |W| \le k. \tag{2}$$

Here $\phi(W)$ is a set-valued function which determines the quality of the sensor placement. In this work we focus on the A-optimality criterion, which minimizes average posterior variance, and D-optimality, which measures the information gain from the prior to the posterior. These criteria amounts to measuring the trace and log-determinant of the posterior covariance matrices respectively. For the current problem, these criteria take the form

$$\phi_A(W) = \mathsf{trace}\left(\boldsymbol{\Gamma}_{\text{pr}}^{1/2}\left(\mathbf{I} + \mathbf{C}\mathbf{C}^{\mathsf{T}}\right)^{-1}\boldsymbol{\Gamma}_{\text{pr}}^{1/2}\right) \quad \text{and} \quad \phi_D(W) = -\mathsf{logdet}\left(\mathbf{I} + \mathbf{C}\mathbf{C}^{\mathsf{T}}\right), \tag{3}$$

where $\mathbf{C} = \mathbf{A}(:, W)$ are the columns of an appropriately formed matrix indexed by W. Here $\mathbf{A} := \boldsymbol{\Gamma}_{\text{pr}}^{1/2}\mathbf{F}^{\mathsf{T}}\boldsymbol{\Gamma}_{\text{noise}}^{-1/2} \in \mathbb{R}^{n\times m}$ is the prior-preconditioned forward operator and selecting k columns is akin to selecting sensors. Note that we use $\phi(W)$ and $\phi(\mathbf{C})$ interchangeably.

Assuming the following partitioned SVD of $\mathbf{A}$ with $1 \le k \le m$,

$$\mathbf{A} = \begin{bmatrix}\mathbf{U}_k & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{V}_k & \mathbf{V}_\perp\end{bmatrix}^{\mathsf{T}}.$$

Now our structural bounds are for column selection of the form $\mathbf{A}\boldsymbol{\Pi} = \begin{bmatrix}\mathbf{A}\boldsymbol{\Pi}_1 & \mathbf{A}\boldsymbol{\Pi}_2\end{bmatrix} = \begin{bmatrix}\mathbf{C} & \mathbf{T}\end{bmatrix}$ with an identical permutation of the truncated right singular vectors $\mathbf{V}_k^{\mathsf{T}}\boldsymbol{\Pi} = \begin{bmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\end{bmatrix}$.

Theorem 1 *[1] Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $k \le \operatorname{rank}(\mathbf{A})$. Then for any permutation $\boldsymbol{\Pi}$ such that $\operatorname{rank}(\mathbf{V}_{11}) = k$ and $\mathbf{A}\boldsymbol{\Pi} = \begin{bmatrix}\mathbf{C} & \mathbf{T}\end{bmatrix}$ we have,*

$$\frac{\sigma_i(\mathbf{A})}{\left\|\mathbf{V}_{11}^{-1}\right\|_2} \le \sigma_i(\mathbf{C}) \le \sigma_i(\mathbf{A}), \quad 1 \le i \le k.$$

The bounds on individual singular values of $\mathbf{C}$ are key to obtaining bounds and algorithms for the different OED objectives. Let $\mathbf{C}_D^{\text{opt}}$ denote the optimal selection for the D-optimality criteria (respectively $\mathbf{C}_A^{\text{opt}}$ for A-optimality). Then utilizing Theorem 1, we can see that

$$\begin{aligned} \phi_D(\mathbf{A}) \leq \phi_D(\mathbf{\Sigma}_k) \leq \phi_D(\mathbf{C}_D^{\text{opt}}) \leq \phi_D(\mathbf{C}) \leq \phi_D\left(\mathbf{\Sigma}_k / \left\|\mathbf{V}_{11}^{-1}\right\|_2\right) \text{ and} \\ \frac{t\left(\mathbf{\Sigma}_k\right) + (n-k)}{\left\|\mathbf{\Gamma}_{\text{pr}}^{-1}\right\|_2} \leq \phi_A(\mathbf{C}_A^{\text{opt}}) \leq \phi_A\left(\mathbf{C}\right) \leq \left\|\mathbf{\Gamma}_{\text{pr}}\right\|_2 \left(t\left(\mathbf{\Sigma}_k / \left\|\mathbf{V}_{11}^{-1}\right\|_2\right) + (n-k)\right), \end{aligned} \tag{4}$$

where $t(\mathbf{X}) = \sum_{i=1}^{\text{rank}(\mathbf{X})} \frac{1}{1+\sigma_j^2(\mathbf{X})}$. Not surprisingly, the performance of the selected columns depend on the top-k singular values of $\mathbf{A}$. If the discarded singular values, $\mathbf{\Sigma}_\perp$, are not negligible, we cannot expect $\mathbf{C}^{\text{opt}}$ to be close to $\mathbf{A}$ in either criterion. Note that the error bounds for the D-optimality case is much cleaner than A-optimality due to the absence of the prior term which factors out as a constant because of the logdet objective. Another point of concern is the presence of the terms with n for A-optimality, which in principle can be extremely large. This term arises due to the ill-posed nature of the inverse problem and corresponds to the singular values of 1 in $\mathbf{I}_n + \mathbf{C}\mathbf{C}^\mathsf{T}$. These values multiply out for D-optimality but are harder to remove in the A-optimality case prompting the development of relative bounds.

Equation (4) clearly identifies the factor $\left\|\mathbf{V}_{11}^{-1}\right\|_2$ to optimize for in an OED algorithm. Also since $\mathbf{V}_{11}$ is an invertible submatrix of $\mathbf{V}_k$, we have $\left\|\mathbf{V}_{11}^{-1}\right\|_2 \geq 1$. We wish to make this value as close to 1 as possible by finding a set of k well-conditioned columns of $\mathbf{V}_k^\mathsf{T}$. This is exactly the Golub-Klema-Stewart approach for subset selection [4], which we further accelerate using randomized approaches. Inspired by rank-revealing factorizations [2] and exchange algorithms for OED [5], we also investigate column-swapping based methods on model inverse problems.

The explicit connection to column subset selection gives us many avenues for future work. Is it possible to extend our techniques to the correlated noise or to nonlinear problems? Can we reduce the gap to $\phi(\mathbf{\Sigma}_k)$ by combining sensor information in a sensible manner? What if our optimization criteria is some user specified goal?

References

[1] Eswar, S., Rao, V. & Saibaba, A. Bayesian D-Optimal Experimental Designs via Column Subset Selection. *ArXiv Preprint ArXiv:2402.16000.* (2024)

[2] Gu, M. & Eisenstat, S. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal On Scientific Computing.* **17**, 848-869 (1996)

[3] Hansen, P. Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. (SIAM,1998)

[4] Golub, G., Klema, V. & Stewart, G. Rank degeneracy and least squares problems. (Stanford University,1976)

[5] Fedorov, V. Theory of optimal experiments. (Academic Press,1972)