Linear system identification using augmented Krylov subspaces

Fabio Matti[1] Martin Skovgaard Andersen[2] Tianshi Chen[3] Daniel Kressner[1]

[1]ANCHP, École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, 2800, Denmark
[3]School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, 518172, China

We present a method for modeling noisy outputs generated by one-dimensional linear time-invariant systems. To do so, we need to evaluate a maximum likelihood objective, composed of the quadratic form $\boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y}$ and the log-determinant $\mathrm{Tr}(\log(\lambda \boldsymbol{I} + \boldsymbol{A}))$, for a positive semi-definite $\boldsymbol{A}$ and $\lambda > 0$. Rather than approximating the two terms separately, we approximate them in an augmented Krylov subspace.

Problem setting

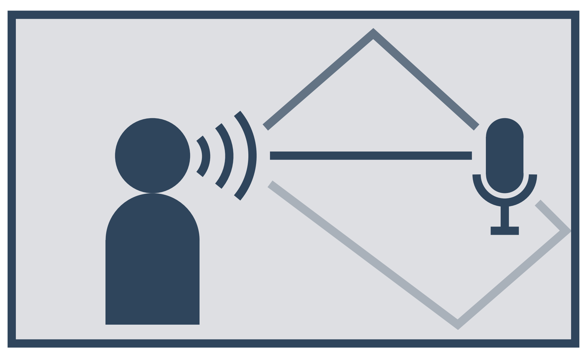
Assume output $\boldsymbol{y} \in \mathbb{R}^m$ is related to input $\boldsymbol{u} \in \mathbb{R}^m$ through

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{e}.$$

- $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ Toeplitz with entries $\Phi_{ij} = u_{i-j}$ if $i > j$ else $\Phi_{ij} = 0$
- $\boldsymbol{\theta} \in \mathbb{R}^n$ unknown impulse response
- $\boldsymbol{e} \in \mathbb{R}^m$ additive noise

Example: Speech dereverberation

Speech recorded by a microphone in a room, where it reverberates off the walls.



Bayesian approach

Assume independent Gaussian priors

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, \nu \boldsymbol{K}(\beta)) \quad \text{and} \quad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

with $\nu, \sigma > 0$, and a kernel matrix $\boldsymbol{K}(\beta) \in \mathbb{R}^{n \times n}$.

Consider kernels given as Cholesky factorizations $\boldsymbol{K}(\beta) = \boldsymbol{L}(\beta)\boldsymbol{L}(\beta)^\top$, where $\boldsymbol{L}(\beta)$ and $\boldsymbol{L}(\beta)^\top$ can be applied to vectors in $\mathcal{O}(n)$ operations.

Goal: Find β, ν, and σ^2 which best fit the output.

Setting $\lambda = \nu/\sigma^2$ and, for now, putting aside the parameter β of the kernel ($\boldsymbol{K} = \boldsymbol{K}(\beta)$), we get the marginal maximum likelihood estimate

$$\boxed{\psi(\lambda) = \log\Big(\underbrace{\boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y}}_{\text{(I)}}\Big) + \underbrace{\mathrm{Tr}\log(\lambda \boldsymbol{I} + \boldsymbol{A})}_{\text{(II)}}}$$

with $\boldsymbol{A} := \boldsymbol{\Phi}\boldsymbol{K}\boldsymbol{\Phi}^\top \in \mathbb{R}^{m \times m}$.

Existing: Direct algorithm [1]

Compute SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$ and evaluate

$$\text{(I)} = \frac{\|\boldsymbol{y}\|_2^2 - \|\tilde{\boldsymbol{y}}\|_2^2}{\lambda} + \sum_{i=1}^{n} \frac{\tilde{y}_i^2}{s_i + \lambda},$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{U}^\top \boldsymbol{y}$, and

$$\text{(II)} = (m - n)\log(\lambda) + \sum_{i=1}^{n} \log(s_i + \lambda).$$

Existing: Indirect algorithm [1]

Use Sherman–Woodbury–Morrison to rewrite

$$\text{(I)} = \lambda^{-1}\boldsymbol{y}^\top(\boldsymbol{I} - \tilde{\boldsymbol{\Phi}}(\lambda \boldsymbol{I} + \tilde{\boldsymbol{\Phi}}^\top\tilde{\boldsymbol{\Phi}})^{-1}\tilde{\boldsymbol{\Phi}}^\top)\boldsymbol{y},$$

and solve the involved regularized least-squares problem with preconditioned LSQR algorithm.

Compute (II) with Girard–Hutchinson estimator and truncated Mercator series of logarithm.

Krylov methods for (I) & (II)

Construct (block) Krylov subspace

$$\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{Z}) = \mathrm{span}\{\boldsymbol{Z}, \boldsymbol{A}\boldsymbol{Z}, \dots, \boldsymbol{A}^{k-1}\boldsymbol{Z}\},$$

for particular vectors or matrices $\boldsymbol{Z}$. Use Lanczos to compute orthonormal basis $\boldsymbol{W}_k$ of $\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{Z})$, along with $\boldsymbol{T}_k = \boldsymbol{W}_k^\top \boldsymbol{A}\boldsymbol{W}_k$.

(I) As typical, use $\boldsymbol{Z} = \boldsymbol{y}$ and approximate

$$\boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y} \approx \boldsymbol{y}^\top \boldsymbol{W}_k(\lambda \boldsymbol{I} + \boldsymbol{T}_k)^{-1}\boldsymbol{W}_k^\top \boldsymbol{y} = \|\boldsymbol{y}\|_2^2(\lambda \boldsymbol{I} + \boldsymbol{T}_k)_{11}^{-1}.$$

(II) Can use $\boldsymbol{Z} = \boldsymbol{\Omega}$ for some Gaussian random matrix $\boldsymbol{\Omega} \in \mathbb{R}^{m \times n_\Omega}$ and approximate

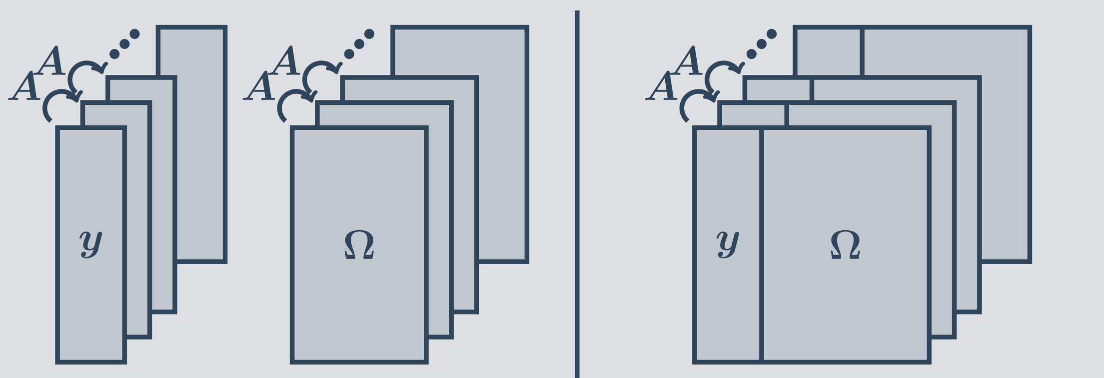
$$\mathrm{Tr}(\log(\lambda \boldsymbol{I} + \boldsymbol{A})) \approx \mathrm{Tr}(\log(\lambda \boldsymbol{I} + \boldsymbol{W}_k \boldsymbol{T}_k \boldsymbol{W}_k^\top)).$$

New: Augmented Krylov subspace

By combining $\boldsymbol{Z} = [\boldsymbol{y}, \boldsymbol{\Omega}] \in \mathbb{R}^{m \times (1+n_\Omega)}$ we construct augmented Krylov subspace

$$\mathcal{K}_k(\boldsymbol{A}, [\boldsymbol{y}, \boldsymbol{\Omega}]) = \mathcal{K}_k(\boldsymbol{A}, \boldsymbol{y}) + \mathcal{K}_k(\boldsymbol{A}, \boldsymbol{\Omega}).$$

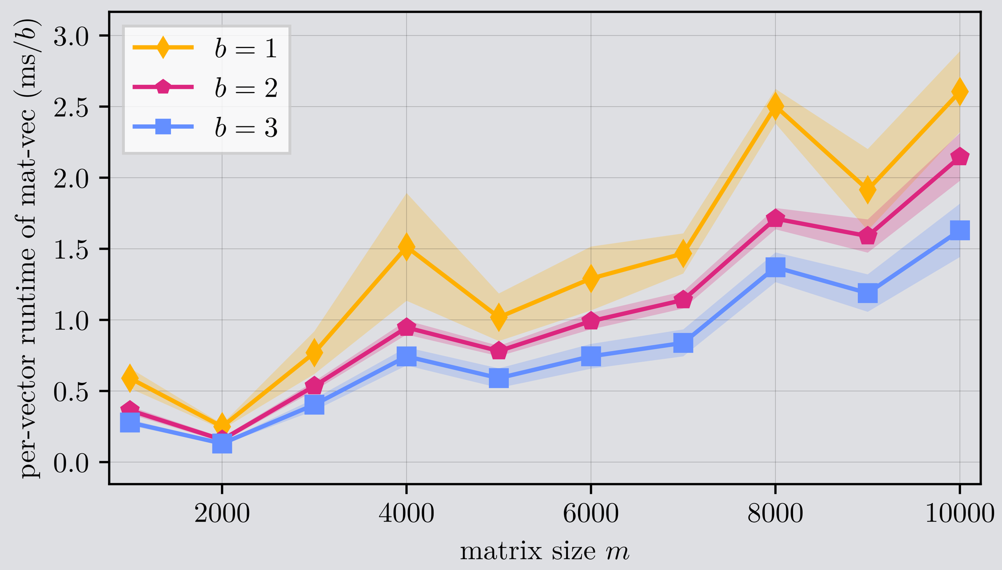
Use block Lanczos to compute orthonormal basis $\widehat{\boldsymbol{W}}_k$ of $\mathcal{K}_k(\boldsymbol{A}, [\boldsymbol{y}, \boldsymbol{\Omega}])$ and $\widehat{\boldsymbol{T}}_k = \widehat{\boldsymbol{W}}_k^\top \boldsymbol{A}\widehat{\boldsymbol{W}}_k$.



separate Krylov subspaces | augmented Krylov subspace

Advantages of blocked mat-vecs

Forming the augmented Krylov subspace can be advantageous, since multiplying the matrix $\boldsymbol{A}$ repeatedly with blocked b vectors is faster than multiplying it with the b vectors individually.



Augmentation does not harm

The following two theorems show that augmenting the estimators can only improve the accuracy.

Theorem: Augmented approx. of (I) [2]

It holds that

$$0 \leq \boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y} - \|\boldsymbol{y}\|_2^2(\lambda \boldsymbol{I} + \widehat{\boldsymbol{T}}_k)_{11}^{-1} \quad \text{(augm.)}$$
$$\leq \boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y} - \|\boldsymbol{y}\|_2^2(\lambda \boldsymbol{I} + \boldsymbol{T}_k)_{11}^{-1} \quad \text{(stand.)}$$

Theorem: Augmented approx. of (II) [2]

It holds that

$$\mathrm{Tr}(\log(\lambda \boldsymbol{I} + \boldsymbol{A}))$$
$$\geq \mathrm{Tr}(\log(\lambda \boldsymbol{I} + \widehat{\boldsymbol{W}}_k \widehat{\boldsymbol{T}}_k \widehat{\boldsymbol{W}}_k^\top)) \quad \text{(augmented)}$$
$$\geq \mathrm{Tr}(\log(\lambda \boldsymbol{I} + \boldsymbol{W}_k \boldsymbol{T}_k \boldsymbol{W}_k^\top)) \quad \text{(standard)}$$

provided that $k(n_\Omega + 1) \leq m$.

Implicit preconditioning

Additionally, there is an even stronger result we can show for the approximation of (I).

Theorem: Implicit preconditioning [3, 2]

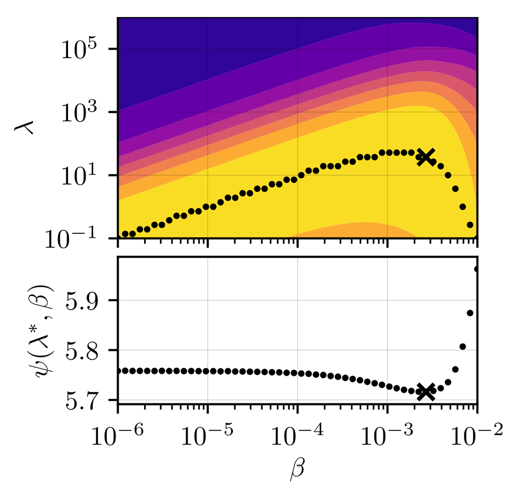
With the notation introduced above, it holds that

$$0 \leq \frac{\boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y} - \|\boldsymbol{y}\|_2^2(\lambda \boldsymbol{I} + \widehat{\boldsymbol{T}}_k)_{11}^{-1}}{\boldsymbol{y}^\top(\lambda \boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{y}} \leq 4\left(\frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1}\right)^{2(k-s)},$$

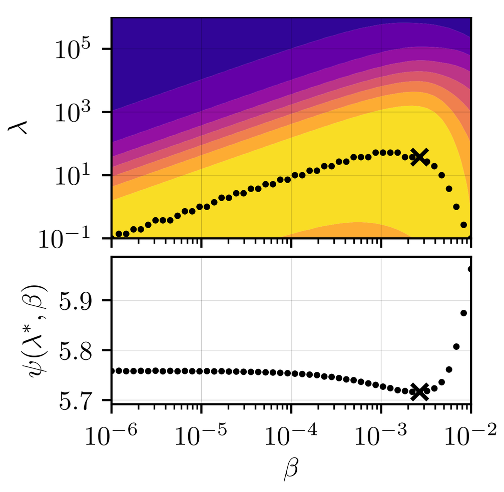
where $\tilde{\kappa}$ is the condition number of the "preconditioned" matrix $\boldsymbol{P}^{-1/2}(\lambda \boldsymbol{I} + \boldsymbol{A})\boldsymbol{P}^{-1/2}$ for any preconditioner $\boldsymbol{P} = (\boldsymbol{I} + \boldsymbol{X})^{-1}$ with $\mathrm{range}(\boldsymbol{X}) \subseteq \mathcal{K}_{s+1}(\boldsymbol{A}, \boldsymbol{\Omega})$ for some $s \leq k$.

Augmentation has an effect like preconditioning.

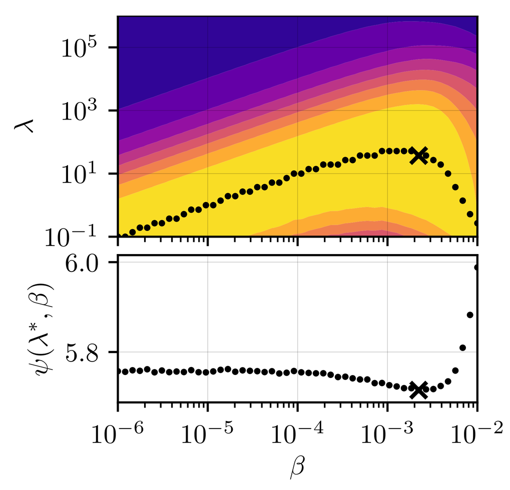
Numerical experiments



(a) Direct algorithm.



(b) Indirect algorithm.



(c) Krylov-augmented.

	runtime (s)
Direct algorithm	86.96
Indirect algorithm	937.68
Krylov-augmented	11.91

(d) Runtime comparison.

[1] L. Chen, T. Chen, and M. S. Andersen, "Fast kernel-based regularized system identification using Bayesian optimization," *IEEE Trans. Autom. Control*, vol. 71, no. 4, 2026.
[2] F. Matti, M. S. Andersen, T. Chen, and D. Kressner, "Kernel-based linear system identification using augmented krylov subspaces," *arXiv*, 2026.
[3] T. Chen, C. Huber, E. Lin, and H. Zaid, "Preconditioning without a preconditioner using randomized block Krylov subspace methods," *ETNA*, pp. 63–92, 2026.