A Comparative Study of Invariance-Aware Loss Functions for Deep Learning-based Gridless Direction-of-Arrival Estimation

Kuan-Lin Chen and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego

IEEE ICASSP 2025

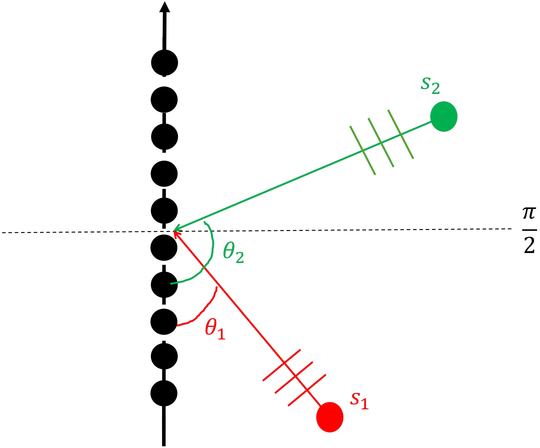
Code is available at https://github.com/kjason/SubspaceRepresentationLearning.

Outline

1. Direction-of-arrival (DoA) estimation
   - Background
   - Optimization-based approaches
2. DNN-based covariance matrix reconstruction
   - Modeling
   - Frobenius norm
   - Scale invariance
3. Geodesic distances
   - Covariance matrix reconstruction
   - Subspace reconstruction
4. Numerical results

Outline

The direction-of-arrival (DoA) estimation problem



A uniform linear array (ULA)

Assumptions

- Under the standard assumptions, the snapshot $\mathbf{y}(t) \in \mathbb{C}^m$ at time $t \in [T]$ can be modeled as

$$\mathbf{y}(t) = \sum_{i=1}^{k} s_i(t)\mathbf{a}(\theta_i) + \mathbf{n}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{n}(t), \quad \mathbf{n}(t) \sim \mathcal{CN}(\mathbf{0}, \eta\mathbf{I}_m) \quad (1)$$

  where $\mathbf{a}(\theta) : [0, \pi] \to \mathbb{C}^m$ is the array manifold of the m-element uniform linear array (ULA) whose i-th element is given by

$$[\mathbf{a}(\theta)]_i = e^{j2\pi\left(i-1-\frac{(m-1)}{2}\right)\frac{d}{\lambda}\cos\theta}, i \in [m] \quad (2)$$

  and $\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{a}(\theta_1) & \mathbf{a}(\theta_2) & \cdots & \mathbf{a}(\theta_k) \end{bmatrix}$. The k signals have equal powers.
- Given $\{\mathbf{y}(t)\}_{t=1}^{T}$ and $k \in [m-1]$, how to find $\theta_1, \theta_2, \cdots, \theta_k$?

Sparse linear arrays (SLAs)

- Let $n \leq m$ and $\mathcal{S} = \{e_1, e_2, \cdots, e_n\} \subset [m]$. Consider minimum redundancy arrays (MRAs) or nested arrays.[a]
- The snapshot received on this physical array is

$$\mathbf{y}_{\mathcal{S}}(t) = \mathbf{\Gamma}\mathbf{y}(t). \tag{3}$$

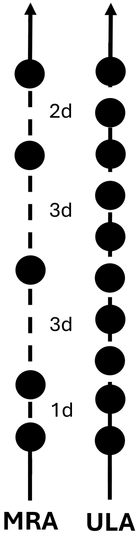
where $\mathbf{\Gamma} \in \mathbb{R}^{n \times m}$ is a row selection matrix given by

$$[\mathbf{\Gamma}]_{ij} = \begin{cases} 1, & \text{if } e_i = j, \\ 0, & \text{otherwise,} \end{cases}, i \in [n], j \in [m]. \tag{4}$$

- Let $\mathbf{R}_0$ be the SCM of the ULA. The noiseless SCM of the SLA/MRA is

$$\mathbf{R}_{\mathcal{S}} = \mathbf{\Gamma}\mathbf{R}_0\mathbf{\Gamma}^{\mathsf{T}}. \tag{5}$$

- Define $\hat{\mathbf{R}}_{\mathcal{S}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_{\mathcal{S}}(t)\mathbf{y}_{\mathcal{S}}^{\mathsf{H}}(t)$.



[a] Pal, Piya, and Palghat P. Vaidyanathan. "Nested arrays: A novel approach to array processing with enhanced degrees of freedom." *IEEE Transactions on Signal Processing* 58, no. 8 (2010).

Outline

1. Direction-of-arrival (DoA) estimation
   - Background
   - Optimization-based approaches
2. DNN-based covariance matrix reconstruction
   - Modeling
   - Frobenius norm
   - Scale invariance
3. Geodesic distances
   - Covariance matrix reconstruction
   - Subspace reconstruction
4. Numerical results

The maximum likelihood problem

- $\mathbf{R}_0 + \eta \mathbf{I}_m$ is positive semidefinite and possibly Toeplitz.
- $\mathbf{y}_{\mathcal{S}}(t) \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\mathcal{S}} + \eta \mathbf{I}_n)$.
- One can formulate the following constrained optimization problem according to the maximum likelihood principle:

$$\begin{aligned} \min_{\mathbf{v} \in \mathbb{C}^m} \quad & \log\det\left(\mathbf{\Gamma}\,\mathrm{Toep}(\mathbf{v})\mathbf{\Gamma}^{\mathsf{T}}\right) + \mathrm{tr}\left(\left(\mathbf{\Gamma}\,\mathrm{Toep}(\mathbf{v})\mathbf{\Gamma}^{\mathsf{T}}\right)^{-1} \hat{\mathbf{R}}_{\mathcal{S}}\right) \\ \text{subject to} \quad & \mathrm{Toep}(\mathbf{v}) \succeq 0. \end{aligned} \tag{6}$$

Convex relaxation and majorization-minimization:

- SPA (Yang et al., 2014), Wasserstein distance minimization (Wang et al., 2019), StructCovMLE (Pote and Rao, 2023), etc

The sparse and parametric approach (SPA)

Based on the covariance fitting criterion (Stoica et al., 2010)[1], Yang et al. (2014)[2] formulated the SPA which involves the following problem:

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{H}^n,\mathbf{v}\in\mathbb{C}^m} \quad & \operatorname{tr}(\mathbf{X}) + \operatorname{tr}\left(\hat{\mathbf{R}}_{\mathcal{S}}^{-1}\boldsymbol{\Gamma}\operatorname{Toep}(\mathbf{v})\boldsymbol{\Gamma}^{\mathsf{T}}\right) \\ \text{subject to} \quad & \begin{bmatrix} \mathbf{X} & \hat{\mathbf{R}}_{\mathcal{S}}^{\frac{1}{2}} & \\ \hat{\mathbf{R}}_{\mathcal{S}}^{\frac{1}{2}} & \boldsymbol{\Gamma}\operatorname{Toep}(\mathbf{v})\boldsymbol{\Gamma}^{\mathsf{T}} & \\ & & \operatorname{Toep}(\mathbf{v}) \end{bmatrix} \succeq 0. \end{aligned} \tag{7}$$

[1] Stoica, Petre, Prabhu Babu, and Jian Li. "New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data." *IEEE Transactions on Signal Processing* 59, no. 1 (2010).

[2] Yang, Zai, Lihua Xie, and Cishen Zhang. "A discretization-free sparse and parametric approach for linear array signal processing." *IEEE Transactions on Signal Processing* 62, no. 19 (2014).

Outline

Reconstruction models

Training a covariance matrix reconstruction model can be formulated as minimizing the empirical risk

$$\min_{W} \quad \frac{1}{L} \sum_{l=1}^{L} d\left(g \circ f_W\left(\hat{\mathbf{R}}_{\mathcal{S}}^{(l)}\right), h\left(\mathbf{R}^{(l)}\right)\right). \tag{8}$$

- $f_W : \mathbb{C}^{n \times n} \to \mathbb{C}^{m \times m}$ is a DNN model with parameters W.
- $\{\hat{\mathbf{R}}_{\mathcal{S}}^{(l)}, \mathbf{R}^{(l)}\}_{l=1}^{L}$ is a dataset of sample covariance matrices at the SLA and noiseless covariance matrices at the corresponding ULA.
- h is a function that extracts the learning target.
- g is a transformation that ensures some properties of a valid covariance matrix. For example, picking the function

  $$g(\mathbf{E}) = \mathbf{E}\mathbf{E}^{\mathrm{H}} + \delta \mathbf{I} \tag{9}$$

  for some $\delta \geq 0$ enforces the predicted matrix being always positive semidefinite (or positive definite).
- $d : \mathbb{C}^{m \times m} \times \mathbb{C}^{m \times m} \to [0, \infty)$ is a loss function of choice.

Outline

How to design the loss function?

Question 1

How to design d in (8)?

Frobenius norm:

$$d_{\mathrm{Fro}}\left(\hat{\mathbf{R}}, \mathbf{R}\right) = \left\|\hat{\mathbf{R}} - \mathbf{R}\right\|_F. \tag{10}$$

- $\alpha\mathbf{R}$ for any $\alpha \in \mathbb{R} \setminus \{0\}$ leads to identical signal and noise subspaces.
- $d_{\mathrm{Fro}}(\alpha\mathbf{R}, \mathbf{R}) \to \infty$ as $\alpha \to \infty$ or $\alpha \to -\infty$ for any positive definite matrix $\mathbf{R}$.

Outline

A scale-invariant loss function

To avoid such a penalization and allow a larger solution space, we propose the following *scale-invariant reconstruction loss*:

$$d_{\mathsf{SI}}\left(\hat{\mathbf{R}}, \mathbf{R}\right) = -\log\left(\frac{\left\|\alpha^* \mathbf{R}\right\|_F}{\epsilon + \left\|\alpha^* \mathbf{R} - \hat{\mathbf{R}}\right\|_F}\right) \tag{11}$$

where $\epsilon \geq 0$ is a constant and

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}} \left\|\alpha \mathbf{R} - \hat{\mathbf{R}}\right\|_F . \tag{12}$$

- The scale-invariant reconstruction loss d_{SI} is invariant to scaling of the matrices in the following sense:

$$d_{\mathsf{SI}}\left(\gamma \mathbf{R}, \mathbf{R}\right) \to -\infty \quad \text{as} \quad \epsilon \to 0 \tag{13}$$

  for every $\gamma \neq 0$.
- Approximately, this property allows d_{SI} to expand the solution space from a point to a line in $\mathbb{C}^{m \times m}$.
- If $\epsilon > 0$, in general we have $d_{\mathsf{SI}}\left(\gamma \mathbf{R}_1, \mathbf{R}_1\right) \neq d_{\mathsf{SI}}\left(\gamma \mathbf{R}_2, \mathbf{R}_2\right)$ for $\mathbf{R}_1 \neq \mathbf{R}_2$.

Signal and noise subspaces

Denoting $\mathbf{E}_s$ as a matrix whose columns are signal eigenvectors of $\mathbf{R}$, the scale-invariant reconstruction loss can be applied to the signal subspace matrix $h(\mathbf{R}) = \mathbf{E}_s \mathbf{E}_s^{\mathsf{H}}$ as follows

$$-\log\left(\frac{\left\|\alpha^* \mathbf{E}_s \mathbf{E}_s^{\mathsf{H}}\right\|_F}{\epsilon + \left\|\alpha^* \mathbf{E}_s \mathbf{E}_s^{\mathsf{H}} - g \circ f_W\left(\hat{\mathbf{R}}_{\mathcal{S}}\right)\right\|_F}\right) \tag{14}$$

where

$$\alpha^* = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}} \left\|\alpha \mathbf{E}_s \mathbf{E}_s^{\mathsf{H}} - g \circ f_W\left(\hat{\mathbf{R}}_{\mathcal{S}}\right)\right\|_F. \tag{15}$$

The same formulation as in (14) and (15) can be applied to the noise subspace, where $h(\mathbf{R}) = \mathbf{E}_n \mathbf{E}_n^{\mathsf{H}}$ and $\mathbf{E}_n$ denotes the noise subspace.

Outline

The affine invariant distance

The affine invariant distance (Barthelme and Utschick, 2021)

$$d_{\mathrm{Aff}}\left(\hat{\mathbf{R}}, \mathbf{R}\right) = \left\| \log\left(\mathbf{R}^{-\frac{1}{2}} \hat{\mathbf{R}} \mathbf{R}^{-\frac{1}{2}}\right) \right\|_F \tag{16}$$

measures the length of the shortest curve between two positive definite matrices (Bhatia, 2007).

Proposition 1

For every m-by-m Hermitian matrix such that $\mathbf{R} \succ 0$ *and for every* $\alpha > 0$,

$$d_{\mathrm{Aff}}\left(\alpha \mathbf{R}, \mathbf{R}\right) = \sqrt{m} |\log \alpha|. \tag{17}$$

- A logarithmic growth in terms of scaling, much slower than the linear rate of the Frobenius norm.
- Despite d_{Aff} being not scale-invariant, its increased distance is invariant to the underlying matrix and only depends on the scaling factor, unlike the Frobenius norm, which depends on the matrix.
- For $\mathbf{R}_1 \neq \mathbf{R}_2$, we have $d_{\mathrm{Aff}}\left(\alpha \mathbf{R}_1, \mathbf{R}_1\right) = d_{\mathrm{Aff}}\left(\alpha \mathbf{R}_2, \mathbf{R}_2\right)$, ensuring the same penality for perfect fittings.

Outline

1. Direction-of-arrival (DoA) estimation
   - Background
   - Optimization-based approaches
2. DNN-based covariance matrix reconstruction
   - Modeling
   - Frobenius norm
   - Scale invariance
3. Geodesic distances
   - Covariance matrix reconstruction
   - Subspace reconstruction
4. Numerical results

Subspace representation learning (Chen and Rao, 2025)

- Let $\mathcal{D} = \left\{ \hat{\mathbf{R}}_{\mathcal{S}}^{(l)}, \mathcal{U}^{(l)} \right\}_{l=1}^{L}$ be a dataset. Construct

$$f_W : \mathbb{C}^{n \times n} \times [m-1] \to \bigcup_{k=1}^{m-1} \mathrm{Gr}(k, m) \tag{18}$$

  where $\mathrm{Gr}(k, m)$ is the Grassmann manifold or *Grassmannian* such that

$$f_{W^*} \left(\hat{\mathbf{R}}_{\mathcal{S}}, k \right) \approx \mathcal{U} \tag{19}$$

  where $\mathcal{U}$ is the corresponding signal or noise subspace.
- Solve

$$\min_{W} \quad \frac{1}{L} \sum_{l=1}^{L} d_{k=k^{(l)}} \left(f_W \left(\hat{\mathbf{R}}_{\mathcal{S}}^{(l)}, k^{(l)} \right), \mathcal{U}^{(l)} \right) \tag{20}$$

  where $d_k : \mathrm{Gr}(k, m) \times \mathrm{Gr}(k, m) \to [0, \infty)$ is some distance.

The geodesic distance

- Let $\mathcal{U}, \tilde{\mathcal{U}} \in \mathrm{Gr}(k, m)$ and $\mathbf{U} \in \mathbb{C}^{m \times k}$ and $\tilde{\mathbf{U}} \in \mathbb{C}^{m \times k}$ be matrices whose columns form unitary bases of $\mathcal{U}$ and $\tilde{\mathcal{U}}$, respectively.

- The principal angles $\boldsymbol{\phi}_k = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_k \end{bmatrix}^{\mathsf{T}}$ between $\mathcal{U}$ and $\tilde{\mathcal{U}}$ can be calculated by

$$\phi_i \left(\mathcal{U}, \tilde{\mathcal{U}} \right) = \cos^{-1} \left(\sigma_i \left(\mathbf{U}^{\mathsf{H}} \tilde{\mathbf{U}} \right) \right) \tag{21}$$

  for $i \in [k]$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ are the singular values of $\mathbf{U}^{\mathsf{H}} \tilde{\mathbf{U}}$.

- The *geodesic distance*

$$d_{\mathrm{Gr}-k} \left(\mathcal{U}_1, \mathcal{U}_2 \right) = \sqrt{\sum_{i=1}^{k} \phi_i^2 \left(\mathcal{U}_1, \mathcal{U}_2 \right)} \tag{22}$$

  defines the length of the shortest curve between two points on the Grassmannian $\mathrm{Gr}(k, m)$.

Outline

Experimental setup

- $n = 4$, $m = 7$, $\mathcal{S} = \{1, 2, 5, 7\}$ if not explicitly specified.
- $T = 50$ if not explicitly specified. $\text{SNR} = 10 \log_{10} \left(\frac{\frac{1}{k} \sum_{i=1}^{k} p_i}{\eta} \right)$.
- $p_1 = p_2 = \cdots = p_k$.
- $k \in [m-1]$.
- For any $k \in [m-1]$, the DoAs $\theta_1, \theta_2, \cdots, \theta_k$ are selected at random in the range $\left[\frac{1}{6}\pi, \frac{5}{6}\pi\right]$ with a minimum separation constraint $\min_{i \neq j} |\theta_i - \theta_j| \geq \frac{1}{45}\pi$.

Evaluation and methods

Evaluation:

- The mean squared error (MSE)

$$\frac{1}{L_{\text{test}}} \sum_{l=1}^{L_{\text{test}}} \frac{1}{k} \min_{\mathbf{\Pi} \in \mathcal{P}_k} \left\| \mathbf{\Pi} \hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l \right\|_2^2 \tag{23}$$

  for a source number $k \in [m-1]$ where $L_{\text{test}} = 10^4$ is the total number of random trials.

Methods:

- We denote the proposed loss functions in (11) and (14) as "SI-Cov" and "SI-Sig," respectively.
- Optimization-based approaches
  - Direct augmentation (DA) (Pillai et al., 1985)
  - SPA (Yang et al., 2014)
  - Wasserstein distance based approach (WDA) (Wang et al., 2019)
- DNN-based covariance matrix reconstruction
  - Cov and Cov-Aff (Barthelme and Utschick, 2021)
  - Cov-T (Wu et al., 2022)
- DNN-based subspace reconstruction
  - Subspace representation learning (Subspace) (Chen and Rao, 2025)

The DNN architecture, dataset, and training procedure

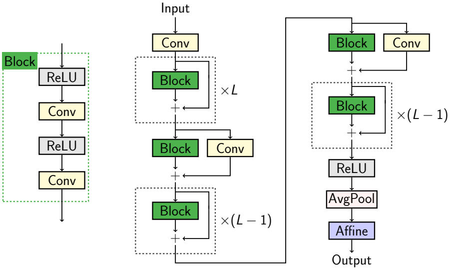


Figure 1: An illustration of a 3-stage L-block ResNet model (He et al., 2016).

- Wide ResNet 16-8 (Zagoruyko and Komodakis, 2016) (11M parameters).
- For each $k \in [m-1]$, there are 2×10^6 and 6×10^5 random data points for training and validation, respectively. $\min_{i \neq j} |\theta_i - \theta_j| \geq \frac{1}{60}\pi$.
- Train 50 epochs with the SGD algorithm and one cycle learning rate scheduler (Smith and Topin, 2019). The batch size is 4096.
- The learning rates for Cov, SI-Cov, SI-Sig, Cov-Aff, and Subspace are 0.01, 0.05, 0.2, 0.005, and 0.1, respectively, found by grid search.

MSE vs. SNR

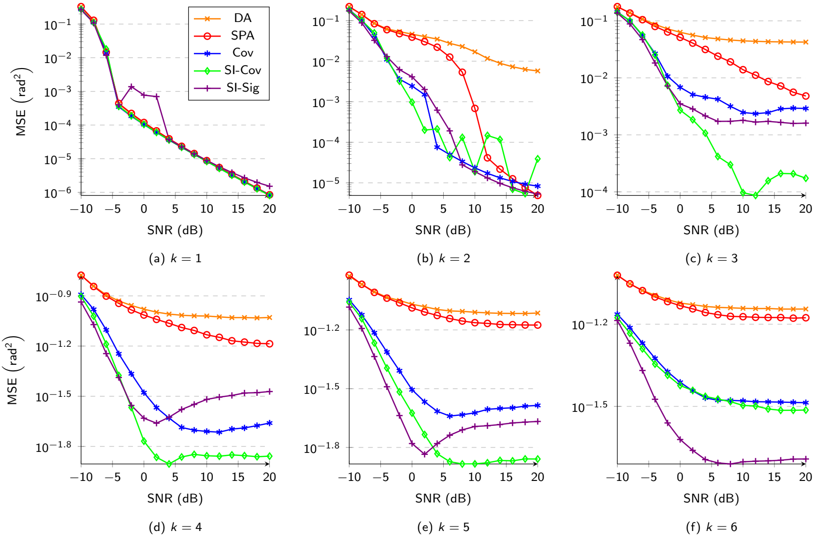
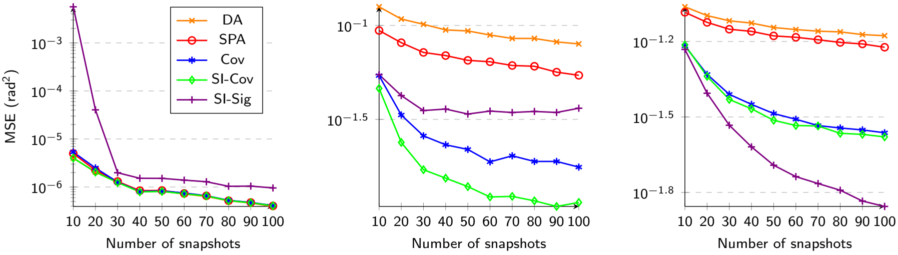


Figure 2: MSE vs. SNR. The proposed scale-invariant covariance matrix reconstruction approach (SI-Cov) outperforms DA, SPA, and Cov when $k > 2$.

MSE vs. Number of snapshots



(a) $k = 1$ (b) $k = 4$ (c) $k = 6$

Figure 3: MSE vs. number of snapshots. Despite these models are trained with their corresponding loss functions at $T = 50$ snapshots, they are able to perform well on a wide range of snapshots.

- The proposed SI-Cov outperforms Cov, showing the advantage of using the scale-invariant strategy.

On greater degrees of invariance

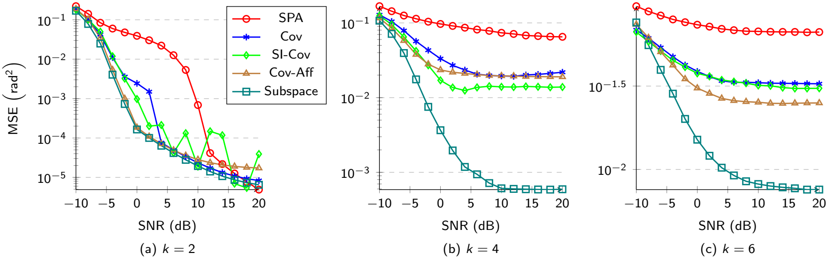


Figure 4: The subspace loss outperforms all the other loss functions designed for covariance matrix reconstruction.

- In general, increasing the degrees of invariance leads to a better optimization landscape and yields better performance.
- SI-Cov and Cov-Aff have mixed results.

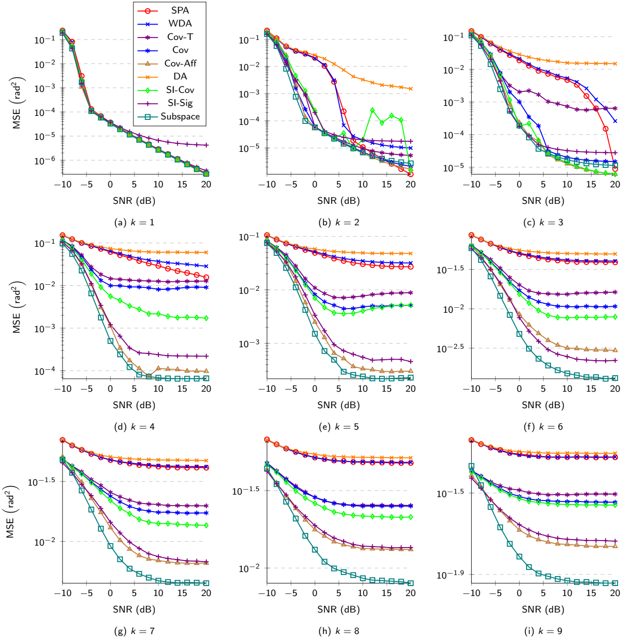


Figure 5: MSE vs. SNR. $n = 5$.

Summary

- A new family of scale-invariant loss functions is proposed for gridless DoA estimation using SLAs.
- We study several loss functions and analyze how invariance properties of a loss can play an important role in shaping the optimization landscape of a DNN model.
- The scale-invariant losses outperform the Frobenius norm that does not have an invariance property.
- The subspace loss is better than the scale-invariant losses and the affine invariant distance.
- These observations provide evidence that greater invariance enhances a DNN's solution space, improving performance in gridless DoA estimation.

References

Barthelme, A. and Utschick, W. (2021). DoA estimation using neural network-based covariance matrix reconstruction. *IEEE Signal Processing Letters*, 28:783–787.

Bhatia, R. (2007). *Positive definite matrices*. Princeton series in applied mathematics. Princeton University Press.

Chen, K.-L. and Rao, B. D. (2025). Subspace representation learning for sparse linear arrays to localize more sources than sensors: A deep learning methodology. *IEEE Transactions on Signal Processing*, 73:1293–1308.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE.

Pillai, S. U., Bar-Ness, Y., and Haber, F. (1985). A new approach to array geometry for improved spatial spectrum estimation. *Proceedings of the IEEE*, 73(10):1522–1524.

Pote, R. R. and Rao, B. D. (2023). Maximum likelihood-based gridless DoA estimation using structured covariance matrix recovery and SBL with grid refinement. *IEEE Transactions on Signal Processing*, 71:802–815.

Smith, L. N. and Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.

Stoica, P., Babu, P., and Li, J. (2010). New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data. *IEEE Transactions on Signal Processing*, 59(1):35–47.

Wang, M., Zhang, Z., and Nehorai, A. (2019). Grid-less DOA estimation using sparse linear arrays based on Wasserstein distance. *IEEE Signal Processing Letters*, 26(6):838–842.

Wu, X., Yang, X., Jia, X., and Tian, F. (2022). A gridless DOA estimation method based on convolutional neural network with Toeplitz prior. *IEEE Signal Processing Letters*, 29:1247–1251.

Yang, Z., Xie, L., and Zhang, C. (2014). A discretization-free sparse and parametric approach for linear array signal processing. *IEEE Transactions on Signal Processing*, 62(19):4959–4973.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press.