Cramèr-Rao Bound for Estimation After Model Selection and its Application to Sparse Vector Estimation

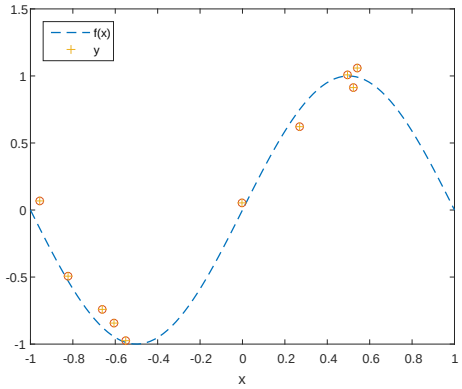
Elad Meir[1,2] and Tirza Cherlow-Routtenberg[1]

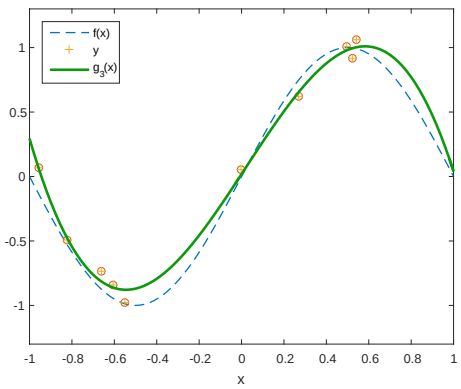
[1] School of Electrical and Computer Engineering
Ben-Gurion University of the Negev, Israel.
[2] TechSee Augmented Vision Ltd.

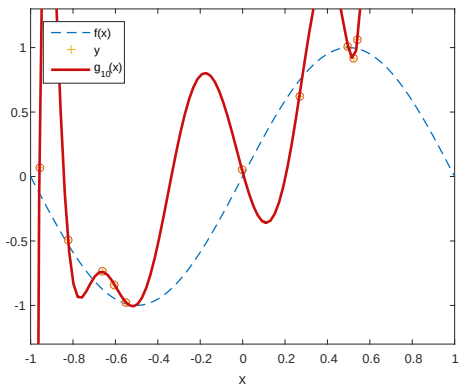
Overview

Motivation

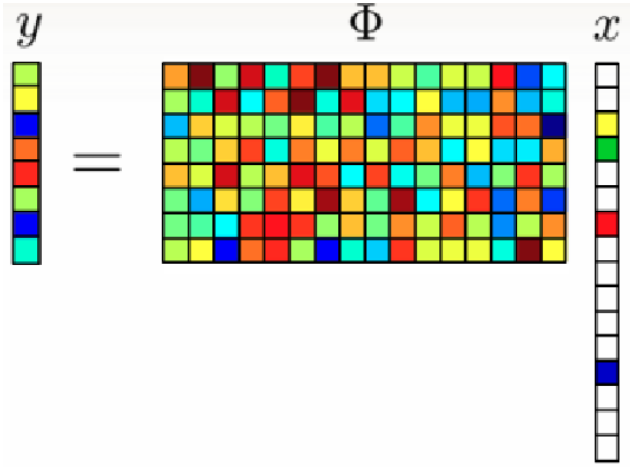
Coefficients estimation of polynomial regression







Sparse Signal processing



Estimation after model selection

The usual practice:

- select a model from a pool of candidate models, based on the data (e.g. AIC, MDL)
- estimate and analyze selected model, disregarding the selection process (maximum likelihood, least squares)

Applications:
multivariate data analysis, machine learning, graph analysis, ...

Non-Bayesian estimation after model selection

- **Goal:** include model selection process in the analysis and estimation
- **Previous works on the selection effect:**
  - effects on standard errors [Pöetscher 1991, Efron 2014]
  - confidence intervals [Benjamini and Yakutieli 2005, Kabaila and Leeb 2006]
  - Cramèr-Rao bound for signals under unknown model order (SMS) [Sando, Mitra and Stoica 2002].
  - Cramèr-Rao bound for estimation after *parameter selection* [Routtenberg and Tong 2016, Harel and Routtenberg 2019] - the model is known, selection of parameter of interest
- **Our approach:**
  - formulation of non-Bayesian post model selection estimation
  - coherence
  - (selective) unbiasedness
  - (selective) CRB

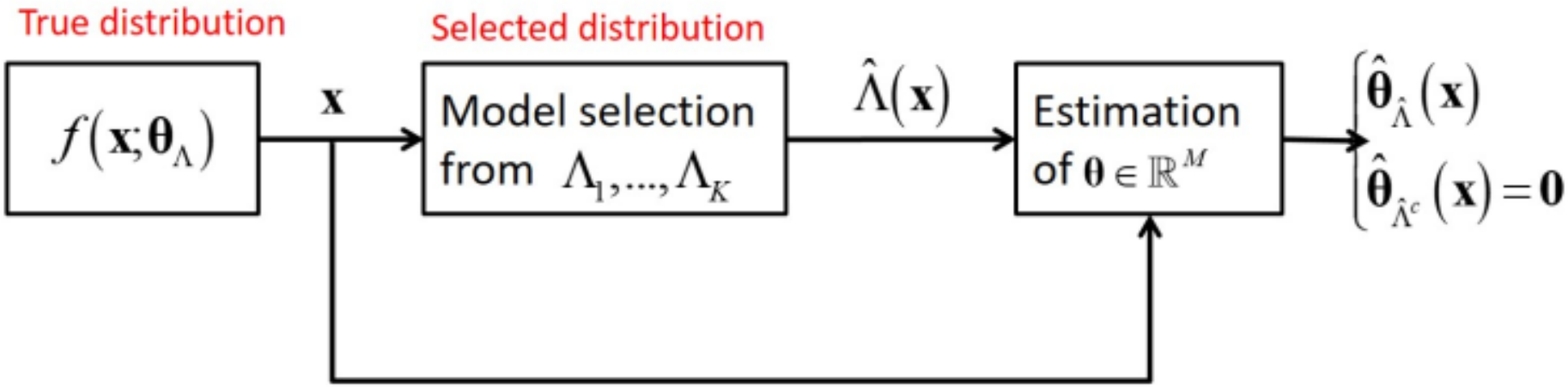
Outline

Problem formulation: estimation after model selection

- $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]^T \in \mathbb{R}^M$ - unknown deterministic parameter vector
- $\mathbf{x} \in \Omega_{\mathbf{x}}$ - random observation vector, which is truly distributed by
- $f(\mathbf{x}; \boldsymbol{\theta}_\Lambda)$ - true probability density function
- Λ and Λ^c - the true support of $\boldsymbol{\theta}$ and its complement, according to the true model. s.t. $\boldsymbol{\theta}_\Lambda \in \mathbb{R}^{|\Lambda|}$, and $\boldsymbol{\theta}_{\Lambda^c} = \mathbf{0}$
- $\{\Lambda_k\}_k$ - is the set of all candidate supports, thus
- $\{f(\mathbf{x}; \boldsymbol{\theta}_{\Lambda_k})\}_k$ - set of pdfs, parameterized by $\boldsymbol{\theta}$ with the suitable support from $\{\Lambda_k\}_k$
- $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \to \mathbb{R}^M$ - estimator of $\boldsymbol{\theta}$ based on $\mathbf{x}$
- $\hat{\Lambda}$ - the selection rule
- $\pi_k(\boldsymbol{\theta}_\Lambda) \triangleq \Pr(\hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}_\Lambda)$ - the probability of selecting the kth model

Estimation after model selection:

The measurement vector, $\mathbf{x}$, is generated by the p.d.f., $f(\mathbf{x};\boldsymbol{\theta}_\Lambda)$, then a model is selected by a pre-determined selection rule - $\hat{\Lambda}$. In the second stage, the unknown parameter, $\boldsymbol{\theta}$ is estimated based on the observation vector and the selected model.



Coherence estimation

Theorem

An estimator $\hat{\boldsymbol{\theta}}$ is said to be a coherent estimator with respect to the selection rule $\hat{\Lambda}$, if $\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} = \mathbf{0}$

Structure:

- a model is selected according to a *predetermined* data-driven selection rule, $\hat{\Lambda}$
- the selected parameters, $\hat{\boldsymbol{\theta}}_{\hat{\Lambda}}$, are estimated
- the deselected parameters are set to be zero, i.e. $\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} = \mathbf{0}$

Zero Padded (ZP) vectors and $\mathbf{D}_k(\Lambda)$ matrix

Definition

For an arbitrary vector, $\mathbf{a} \in \mathbb{R}^M$, and any candidate support set, Λ_k, $k = 1, \ldots, K$, the vector $\mathbf{a}^{\text{ZP}}_{\Lambda_k}$, is a zero-padded, M-length vector, whose non-padded elements correspond to the elements of $\mathbf{a}_{\Lambda_k}$.

Definition

For any $k = 1, \ldots, K$, the $\mathbf{D}_k(\Lambda)$ is a $M \times |\Lambda|$ matrix with the elements

$$[\mathbf{D}_k(\Lambda)]_{m,l} \triangleq \begin{cases} 1, & m \in \Lambda_k,\ m = [\Lambda]_l \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

$\forall l = 1, \ldots, M$, where $[\Lambda]_l$ denotes the lth element of the true support set (not to be confused with the lth candidate support set, Λ_l).

Outline

cost function - motivation

- The Cramér-Rao bound (CRB) provides a lower bound on the mean squared error (MSE) of any mean-unbiased estimator and is used as a benchmark in non-Bayesian estimation
- The conventional CRB does not take into account the selection process, thus, it is inappropriate for estimation after model selection
- We wish to derive a CRB-type bound that takes into account the selection process

- The Cramér-Rao bound (CRB) provides a lower bound on the mean squared error (MSE) of any mean-unbiased estimator

The term mean square error (MSE) refers to the squared mean difference between the estimated and true values. Although this cost function is widely used in non-Bayesian estimation, it does not include the model selection procedure therefore neglects the selection process

Novel cost function

Selected-square-error (SSE) matrix, and the corresponding mean SSE (MSSE)

Proposition (selected-square-error matrix)

$$\mathbf{C}(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \boldsymbol{\theta}) \triangleq \left(\hat{\boldsymbol{\theta}}^{ZP}_{\hat{\Lambda}} - \boldsymbol{\theta}^{ZP}_{\hat{\Lambda}}\right)\left(\hat{\boldsymbol{\theta}}^{ZP}_{\hat{\Lambda}} - \boldsymbol{\theta}^{ZP}_{\hat{\Lambda}}\right)^T \tag{2}$$

Proposition (mean-selected-squared-error)

$$\mathrm{E}_{\boldsymbol{\theta}_\Lambda}\left[\mathbf{C}(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \boldsymbol{\theta})\right] = \sum_{k=1}^{K} \pi_k(\boldsymbol{\theta}_\Lambda) \mathrm{E}_{\boldsymbol{\theta}_\Lambda}\left[(\hat{\boldsymbol{\theta}}^{ZP}_{\Lambda_k} - \boldsymbol{\theta}^{ZP}_{\Lambda_k})(\hat{\boldsymbol{\theta}}^{ZP}_{\Lambda_k} - \boldsymbol{\theta}^{ZP}_{\Lambda_k})^T | \hat{\Lambda} = \Lambda_k\right] \tag{3}$$

A scalar version is available in the article

- The Cramér-Rao bound (CRB) provides a lower bound on the mean squared error (MSE) of any mean-unbiased estimator

Mean-unbiasedness (*unbiasedness*) considers the estimated, and expected values, but does not include the model selection procedure.
Moreover, most post selection estimator tend to bias, making the CRB unsuitable for our use case.

Proposition (Selective Unbiasdness)

An estimator, $\hat{\boldsymbol{\theta}}$, is an unbiased estimator for the problem of estimating the true parameter vector, $\boldsymbol{\theta}_\Lambda$, in the Lehmann sense w.r.t. the SSE matrix defined earlier, and the selection rule, $\hat{\Lambda}$, iff

$$\mathbf{b}_k(\boldsymbol{\theta}, \Lambda) \stackrel{\triangle}{=} \mathrm{E}_{\boldsymbol{\theta}_\Lambda}\left[\hat{\boldsymbol{\theta}}^{ZP}_{\Lambda_k} - \boldsymbol{\theta}^{ZP}_{\Lambda_k} | \hat{\Lambda} = \Lambda_k\right] = \mathbf{0}, \forall \boldsymbol{\theta}^{ZP}_{\Lambda_k} \in \mathbb{R}^{|\Lambda|}, \tag{4}$$

for all $k = 1, \ldots, K$, such that $\pi_k(\boldsymbol{\theta}_\Lambda) \neq 0$.

proof appears in the article, Appendix A

In practice, coherent post model selection estimators tend to be biased. Thus (4) is not zero

Proposition (Selective bias)

$$\mathbf{b}_k(\boldsymbol{\theta},\Lambda) \stackrel{\triangle}{=} \mathrm{E}_{\boldsymbol{\theta}_\Lambda}\left[\hat{\boldsymbol{\theta}}^{ZP}_{\Lambda_k} - \boldsymbol{\theta}^{ZP}_{\Lambda_k} | \hat{\Lambda} = \Lambda_k\right] \neq \mathbf{0}, \forall k = 1,\ldots,K, \tag{5}$$

while its derivative w.r.t. the true parameter is defined as

$$\mathbf{G}_k(\boldsymbol{\theta},\Lambda) \stackrel{\triangle}{=} \nabla_{\boldsymbol{\theta}_\Lambda}\mathbf{b}_k(\boldsymbol{\theta}_\Lambda),\ \forall k = 1,\ldots,K. \tag{6}$$

The selection bias affects the relation between the MSE and the MSSE of the estimator, as we can see in

Proposition (MSE-MSSE coherent relations)

For any coherent estimator, the MSE satisfies

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}},\boldsymbol{\theta},\Lambda)=\mathrm{E}_{\boldsymbol{\theta}_\Lambda}\left[\mathbf{C}(\hat{\boldsymbol{\theta}},\hat{\Lambda},\boldsymbol{\theta})\right]+\sum_{k=1}^{K}\pi_k(\boldsymbol{\theta}_\Lambda)\boldsymbol{\theta}^{ZP}_{\Lambda_k^c}(\boldsymbol{\theta}^{ZP}_{\Lambda_k^c})^T$$
$$-\sum_{k=1}^{K}\pi_k(\boldsymbol{\theta}_\Lambda)\left(\boldsymbol{\theta}^{ZP}_{\Lambda_k^c}\mathbf{b}_k^T(\boldsymbol{\theta},\Lambda)+\mathbf{b}_k(\boldsymbol{\theta},\Lambda)(\boldsymbol{\theta}^{ZP}_{\Lambda_k^c})^T\right). \qquad (7)$$

proof appears in the article, Appendix B

Proposition (Selective CRB on the MSSE)

Under regularity conditions, the MSSE of any coherent and selective biased *estimator, with the selective bias from* (4)*, is bounded by*

$$\mathrm{E}_{\boldsymbol{\theta}_\Lambda}[\mathbf{C}(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \boldsymbol{\theta}_\Lambda)] \succeq \mathbf{B}_{sCRB}(\boldsymbol{\theta}_\Lambda), \tag{8}$$

where the biased selective CRB is defined as

$$\mathbf{B}_{sCRB}(\boldsymbol{\theta}_\Lambda) \stackrel{\triangle}{=} \sum_{k=1}^{K} \pi_k(\boldsymbol{\theta}_\Lambda) \left[(\mathbf{D}_{\Lambda_k}^{ZP} + \mathbf{G}_{\Lambda_k}^{ZP}) \mathbf{J}_k^{-1} (\mathbf{D}_{\Lambda_k}^{ZP} + \mathbf{G}_{\Lambda_k}^{ZP})^T \right] \tag{9}$$

selective CRB

The selective CRB matrix bound

Theorem (sCRB on the MSE)

For any coherent estimator, the MSE satisfies

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_\Lambda) \succeq \mathbf{B}_{sCRB}(\boldsymbol{\theta}_\Lambda) + \sum_{k=1}^{K} \pi_k(\boldsymbol{\theta}_\Lambda)\boldsymbol{\theta}^{ZP}_{\Lambda^c_k}\left(\boldsymbol{\theta}^{ZP}_{\Lambda^c_k}\right)^T - \sum_{k=1}^{K} \pi_k(\boldsymbol{\theta}_\Lambda)\left(\boldsymbol{\theta}^{ZP}_{\Lambda^c_k}\mathbf{b}_k^T(\boldsymbol{\theta},\Lambda) + \mathbf{b}_k(\boldsymbol{\theta},\Lambda)(\boldsymbol{\theta}^{ZP}_{\Lambda^c_k})^T\right).$$

Proof (Appendix C), and the marginal version appear in the article (Section C)

selective unbiased CRB

The selective CRB matrix bound

Theorem (sCRB on the MSE)

For any coherent and selective unbiased estimator, the MSE satisfies

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_\Lambda) \succeq \mathbf{B}_{sCRB}(\boldsymbol{\theta}_\Lambda) + \sum_{k=1}^{K} \pi_k(\boldsymbol{\theta}_\Lambda) \boldsymbol{\theta}^{ZP}_{\Lambda_k^c} \left(\boldsymbol{\theta}^{ZP}_{\Lambda_k^c}\right)^T .$$

Outline

General Linear Model

The General Linear Model (GLM) is given by

- its observations $\forall k = 1, \ldots, K$,

$$\mathbf{x} = \mathbf{H}_k \boldsymbol{\theta}_{\Lambda_k} + \mathbf{w}, \tag{10}$$

- the matrices $\mathbf{H}_k \in \mathbb{R}^{N \times |\Lambda_k|}$, $k = 1, \ldots, K$, are assumed to be known full column rank matrices
- $\boldsymbol{\theta} \in \mathbb{R}^M$ is a deterministic unknown vector
- $\mathbf{w}$ is a zero-mean i.i.d. Gaussian random vector, with known variance
- the coherent Maximum Selected Likelihood (MSL) estimator, is given for $k = 1, \ldots, K$, by

$$\hat{\boldsymbol{\theta}}_{\Lambda_k}^{\mathrm{ML}|k} = \left(\mathbf{H}_k^T \mathbf{H}_k\right)^{-1} \mathbf{H}_k^T \mathbf{x}, \tag{11}$$

- coherecy dictates $\hat{\boldsymbol{\theta}}_{\Lambda_k^c} = \mathbf{0}$

Vs. SNR

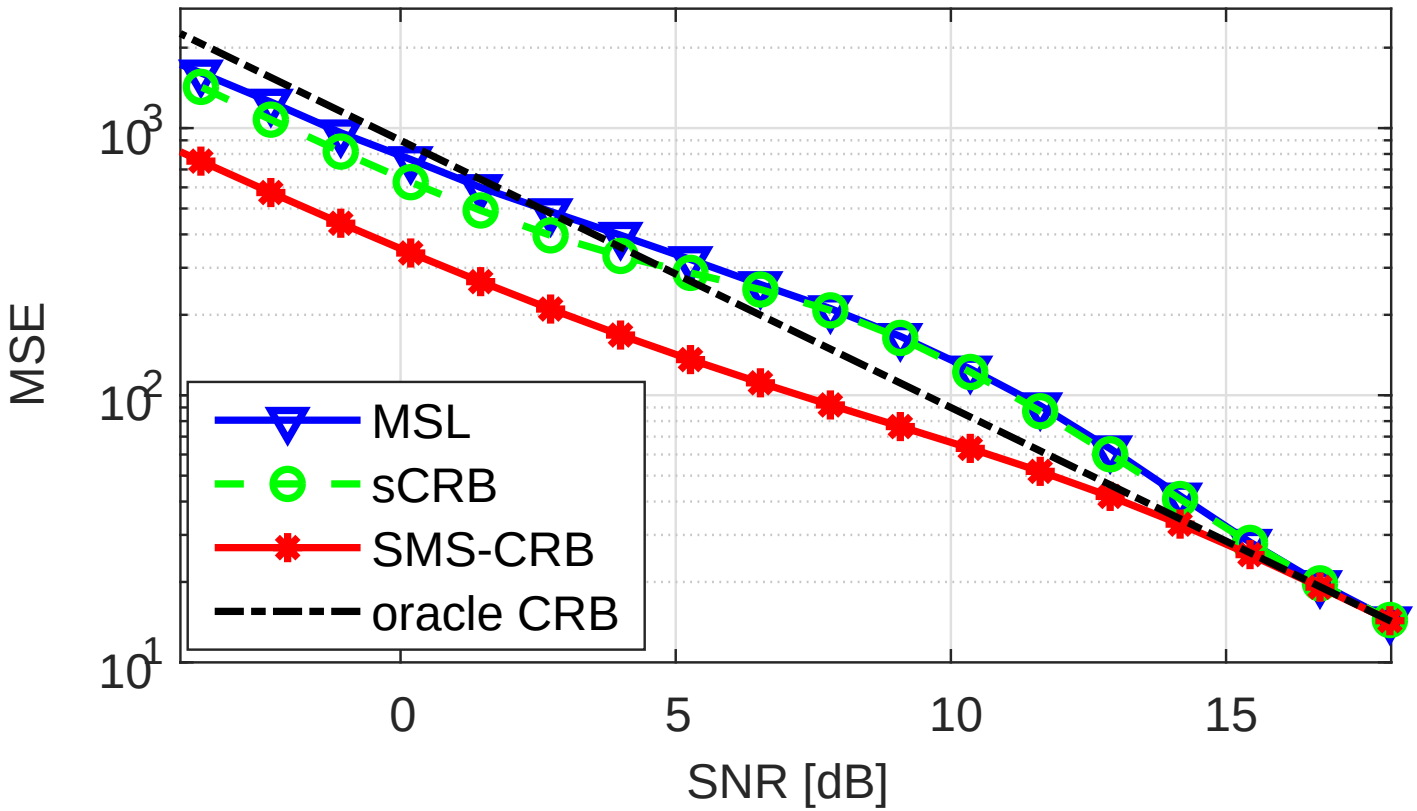


Figure: General linear model with AIC selection rule: MSE, sCRB, SMS-CRB, oracle CRB, vs. $\mathrm{SNR} \triangleq 10\log 10\frac{||\mathbf{H}\boldsymbol{\theta}||^2}{N\sigma^2}$, with varying σ, $N = 1500$ samples, $\boldsymbol{\theta} = [4, -3]^T$, $\mathbf{h}_1 = [1, \ldots, 1]^T$, and the values of $\mathbf{h}_2$ are randomly drawn from uniform distribution in interval $[0, 10]$.

Vs. π_k

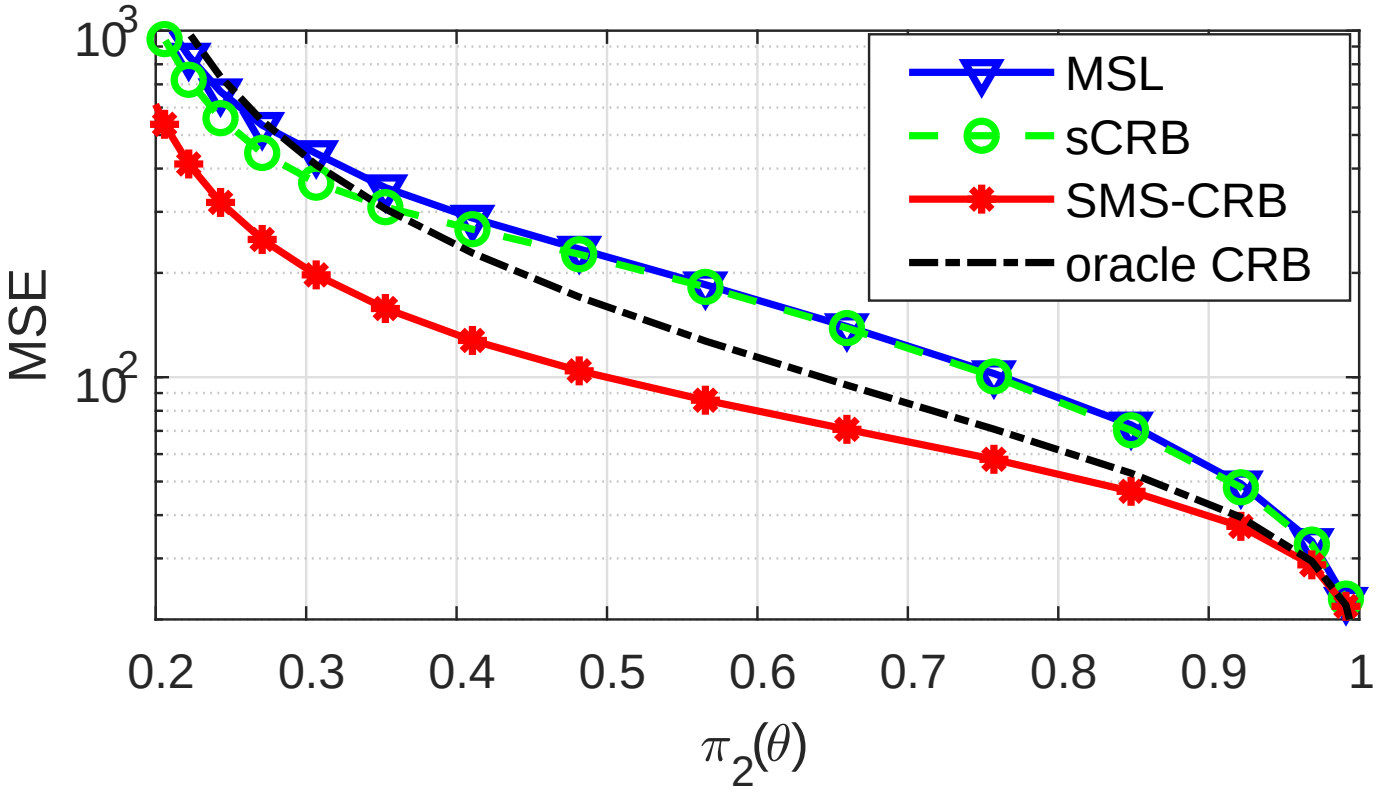


Figure: General linear model with AIC selection rule: MSE, sCRB, SMS-CRB, oracle CRB, vs. $\mathrm{SNR} \triangleq 10\log 10\frac{||\mathbf{H}\boldsymbol{\theta}||^2}{N\sigma^2}$, with varying σ, $N = 1500$ samples, $\boldsymbol{\theta} = [4, -3]^T$, $\mathbf{h}_1 = [1, \ldots, 1]^T$, and the values of $\mathbf{h}_2$ are randomly drawn from uniform distribution in interval $[0, 10]$.

Vs. GIC functions

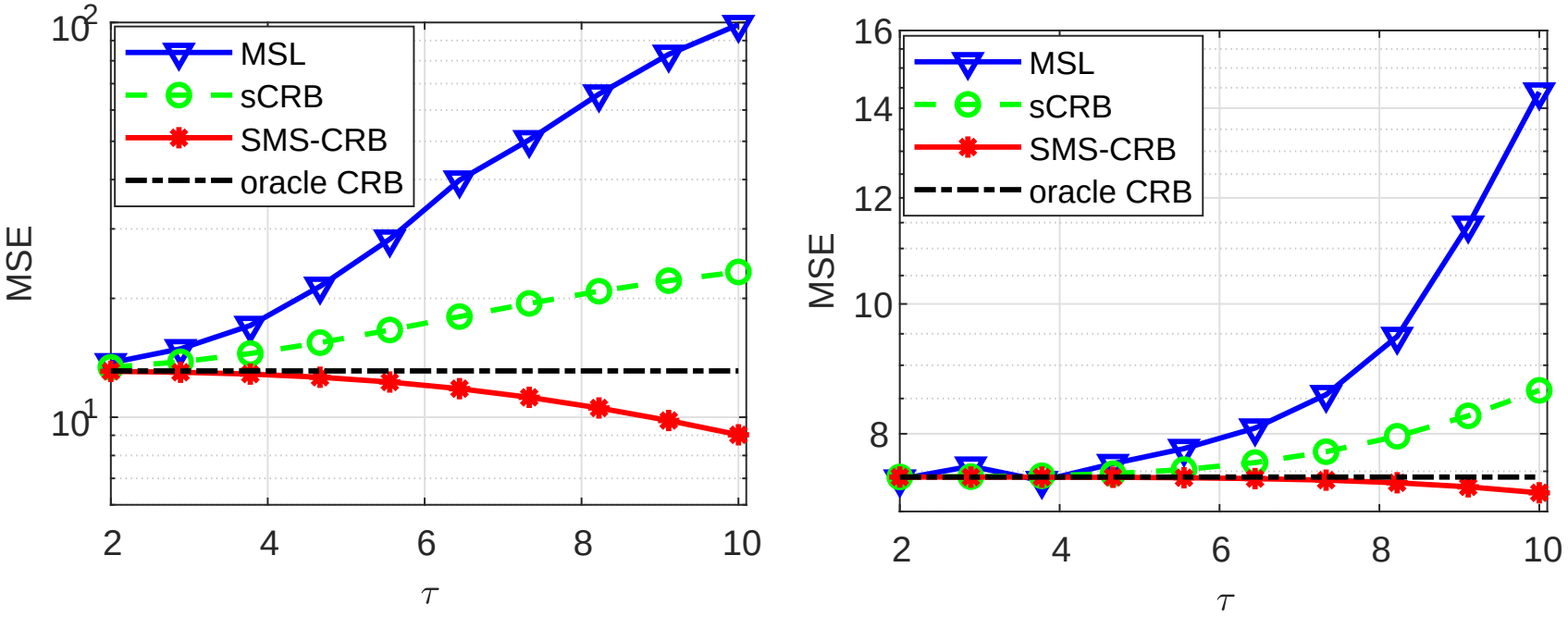


Figure: General linear model with GIC selection rule: The MSE of the MSL estimator compared to selective CRB and the SMS-CRB versus different values of the parameter $\tau(N, |\Lambda_k|)$, $N = 150$, with SNR$= -3.5$dB (left) and 0dB (right).

OST

For sparse vector estimation, we inspect the MSE of

$$\hat{\boldsymbol{\theta}}_{\hat{\Lambda}}^{\text{ML—OST}} = \left(\mathbf{A}_{\hat{\Lambda}}^T \mathbf{A}_{\hat{\Lambda}}\right)^{-1} \mathbf{A}_{\hat{\Lambda}}^T \mathbf{x}, \tag{12}$$

where the selection function is OST

$$m \in \hat{\Lambda} \text{ if } |\mathbf{a}_m^T \mathbf{x}| > c > 0, \ \forall m = 1, \ldots, M$$

and, of course, coherency dictates $\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c}^{\text{ML—OST}} = \mathbf{0}$.

1st example

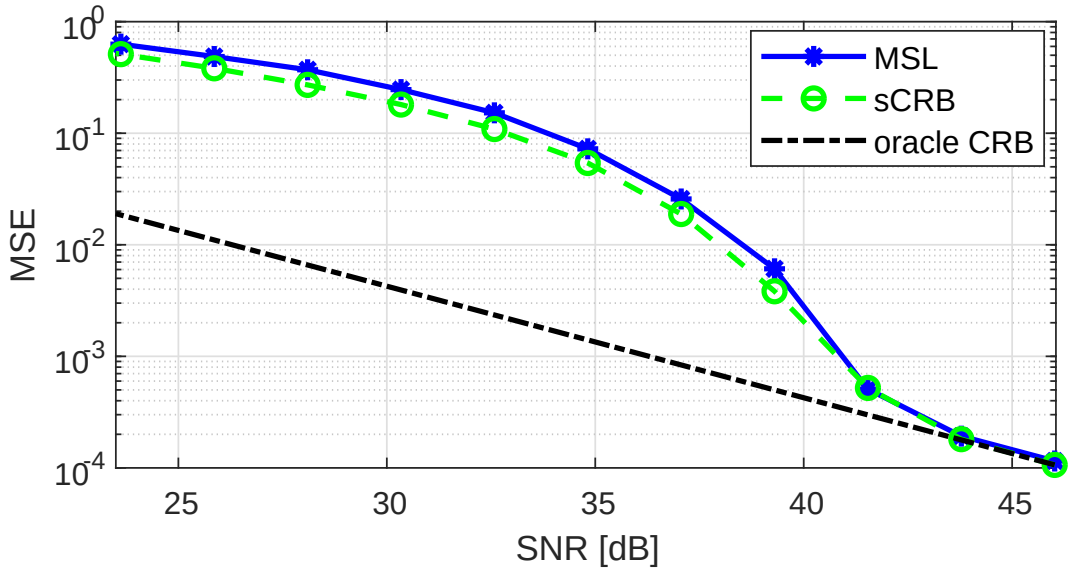


Figure: $\boldsymbol{\theta}_\Lambda = \mathbf{1}$. The MSE of the MSL estimator, sCRB, and oracle CRB under OST rule, with $c = 0.95$, versus SNR. Random 7×14 dictionary, $\mathbf{A}$, $\mathbf{a}_m^T \mathbf{a}_m = 1$, mutual coherence is $\mu = 0.5673$. $|\Lambda| = 3$

2nd example

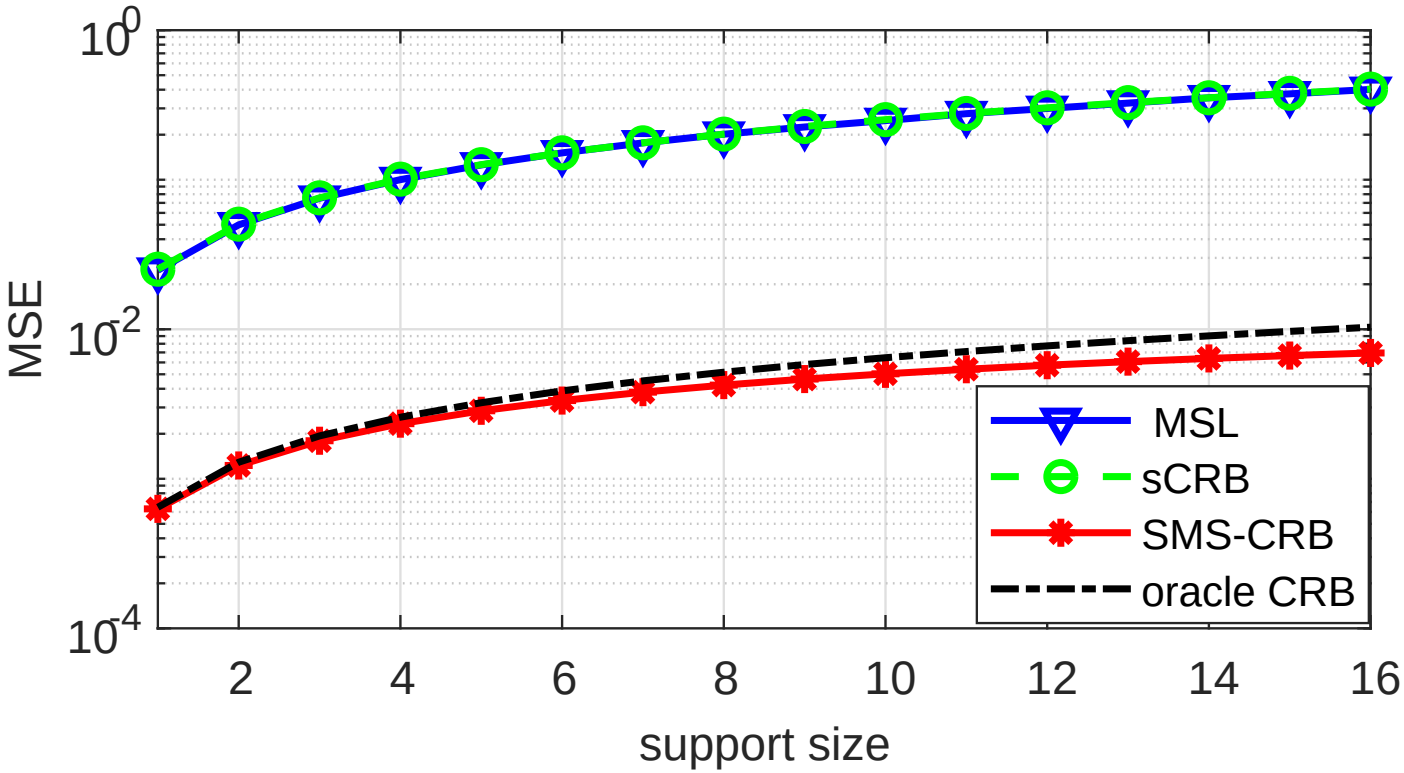


Figure: $\boldsymbol{\theta}_\Lambda = \mathbf{1}$. The MSE of the MSL estimator, sCRB, and oracle CRB under OST rule, with $c = 0.95$. Hadamard 16×16 dictionary, $\mathbf{A}$, $\mathbf{a}_m^T \mathbf{a}_m = 1$, $\sigma = 0.1594$, $1 \leq |\Lambda| \leq 16$

Biased sCRB with identity dictionary, $\mathbf{A} = \mathbf{I}$, with $L = M$.

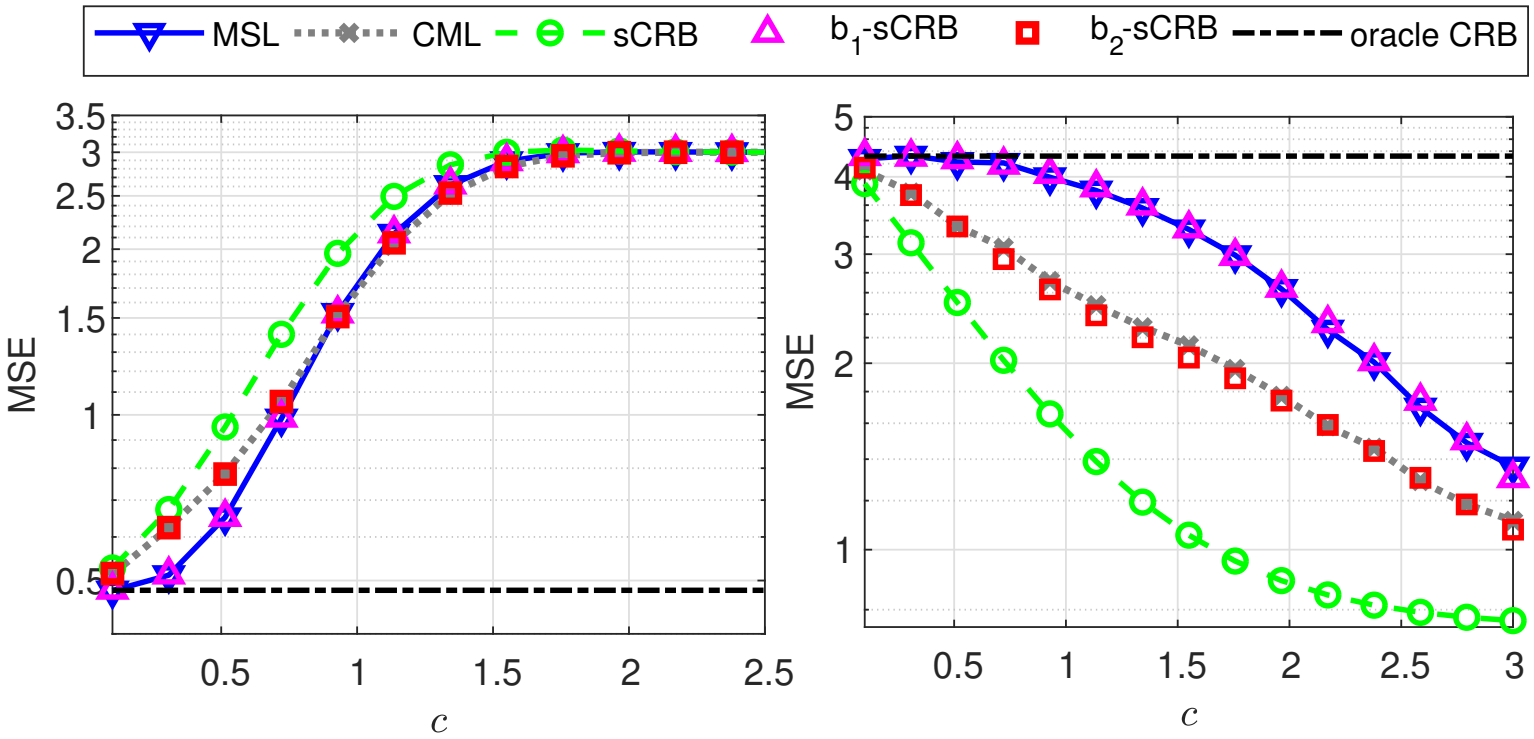


Figure: Sparse vector estimation: The MSE of the MSL and of the post-selection conditional ML (CML) estimators, where the likelihood is selected by the OST rule, the selective CRB, the biased selective CRB with the MSL bias, b_1-sCRB, the biased selective CRB with the CML bias, b_2-sCRB, and the oracle CRB, versus the threshold, c, for 1) $\theta_m = 1, \sigma = 0.4$ (left); 2) $\theta_m = 0.5, \sigma = 1.2$ (right).

Conclusion

- formulation of coherent estimation after model selection
- a measure of performance - selected-square-error is considered
- a Lehmann-sense selective unbiasedness definition is introduced
- an appropriate selective Cramér Rao Bound is derived, with its biased version and for sparse vector estimation
- simulations show that the selective CRB is valid, and tighter than the SMS-CRB, and predicts the threshold phenomenon

Outline

Our recent publications:

- N. Harel and T. Routtenberg, "Post-Parameter-Selection Maximum-Likelihood Estimation", IEEE Workshop on Statistical Signal Processing (SSP 2021).
- E. Meir and T. Routtenberg, "Cramér-Rao Bound for Estimation After Model Selection with Applications to Sparse Vector Estimation", IEEE Transactions on Signal Processing", vol. 69, pp. 2284-2301, 2021.
- N. Harel and T. Routtenberg, "Bayesian Estimation After Model Selection", IEEE Signal Processing Letters, vol. 28, pp. 175-179, 2021.
- T. Weiss, T. Routtenberg, and H. Messer, "Total Performance Evaluation of Intensity Estimation after Detection, Signal Processing, vol. 183, pp. 1-8, 2021.
- N. Harel and T. Routtenberg, "Low-Complexity Methods for Estimation After Parameter Selection, IEEE Transactions on Signal Processing, vol. 68, pp. 1152-1167, 2020.
- E. Meir and T. Routtenberg, "Selective Cramr-Rao Bound for Estimation After Model Selection", IEEE Workshop on Statistical Signal Processing (SSP 2018).
- S. Cohen, T. Routtenberg, and L. Tong, "Non-Bayesian Parameter Estimation of the Probability of the Missing Mass", arXiv:2101.04329.


Fin