Sparse Sensing for Statistical Inference

Model-driven and data-driven paradigms

Geert Leus, Sundeep Chepuri, and Georg Kail

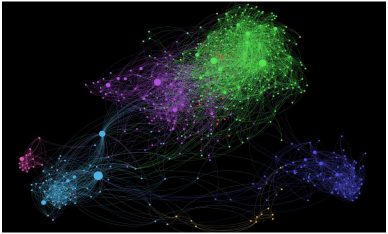
ITA 2016, 04 Feb 2016

Delft University of Technology
TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Power networks, grid analytics

Health informatics

Radio astronomy (e.g., SKA)

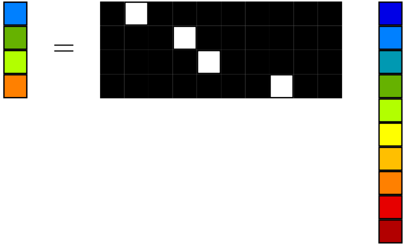


Internet, social media

Design sparse sensing functions

What is sparse sensing?

$\mathbf{y} \in \mathbb{R}^{d}$ $\quad \boldsymbol{\Phi}(\mathbf{w}) = \overbrace{\mathrm{diag_r}(\mathbf{w})}^{\{0,1\}}$ $\quad \mathbf{x} \in \mathbb{R}^{D}$



What is sparse sensing?

Design $\mathbf{w} \in \{0,1\}^D$ to select the most "informative" d ($\ll D$) samples for

- data acquisition (e.g., offline sensor selection)
- data selection (e.g., outlier rejection, censoring)
- Or, combination of above —hybrid

State of the art

- Compressive sensing: sparse signal recovery

  [Donoho-2006], [Candès-Wakin-2008]

- Sensor selection: model-driven

  - convex optimization: design $\{0,1\}^M$ selection vector

    [Joshi-Boyd-2009], [Chepuri-Leus-2015]

  - greedy methods and heuristics: submodularity

    [Krause-Singh-Guestrin-08], [Ranieri-Chebira-Vetterli-14]

- Censoring (or data selection) and outlier rejection: data-driven

  [Rago-Willett-Shalom-96], [Msechu-Giannakis-12]

Model-driven, data-driven, or hybrid?

Linear regression setup— model-driven

- Observations follow

$$x_m = \mathbf{a}_m^T \boldsymbol{\theta} + n_m,\ m = 1, 2, \ldots, D$$

  - $\boldsymbol{\theta} \in \mathbb{R}^p$ unknown parameter
  - n_m i.i.d. zero-mean unit-variance Gaussian noise

Problem statement

Given $\{\mathbf{a}_m\}$ and noise pdf, i.e., only the data model, design $\mathbf{w}$ to select the best subset of d ($\ll D$) sensors

- Best subset of d ($\ll D$) sensors are obtained by optimizing a scalar function (trace, max. eigenvalue, log det) of the CRB matrix of $\boldsymbol{\theta}$:

$$f(\mathbf{w}) = g\left\{\left(\sum_{m=1}^{D} w_m \mathbf{a}_m \mathbf{a}_m^{\mathsf{T}}\right)^{-1}\right\}$$

Optimization problem

- Sparse sensing function can be designed by solving

$$\min_{\mathbf{w}\in\mathcal{W}} \; f(\mathbf{w})$$

$\mathcal{W} = \{\mathbf{w} \in \{0,1\}^D \mid \|\mathbf{w}\|_0 = d\}.$

- The above problem is convex on $\mathbf{w}$ if the set $\mathcal{W}$ is relaxed to $\mathcal{W}_{\mathrm{c}} = \{\mathbf{w} \in [0,1]^D \mid \|\mathbf{w}\|_1 = d\}.$
- Monotone submodular cost functions can be optimized using near-optimal greedy methods.

\+ Sampler needs to be designed offline only once, and samplers are obtained by optimizing ensemble performance.

\- Design might suffer from any possible outliers or model mismatch.

Model-driven design for other inference tasks

- Can be generalized to nonlinear models by optimizing scalar functions of the Cramér-Rao bound matrix.

  S.P. Chepuri and G. Leus. Sparsity-Promoting Sensor Selection for Non-linear Measurement Models. *IEEE Trans. on Signal Processing, Volume 63, Issue 3, pp. 684-698, February 2015*.

- Samplers for nonlinear models with correlated errors can be obtained by solving a convex program.

  S.P. Chepuri and G. Leus. Sparse Sensing for Estimation with Correlated Observations. In Proc. of Asilomar 2015, Pacific Grove, CA, Nov. 2015.

- Samplers can be designed for detection problems by optimizing Kullback-Leibler or Bhattacharyya distance measures.

  S.P. Chepuri and G. Leus. *Sparse Sensing for Distributed Detection*. *Trans. on Signal Processing*, October 2015.

Linear regression setup— data-driven

- The output data $\{x_m\}_{m=1}^{D}$ is possibly contaminated with up to o outliers.
- We know the (uncontaminated) data model

$$\bar{x}_m = \mathbf{a}_m^T \boldsymbol{\theta} + n_m,\ m = 1, 2, \ldots, D$$

- $\boldsymbol{\theta} \in \mathbb{R}^p$ unknown parameter
- n_m i.i.d. zero-mean unit-variance Gaussian noise

Problem statement

Given $\{x_m\}$, $\{\mathbf{a}_m\}$, and noise pdf:

(a) design $\mathbf{w}$ to censor less-informative samples and reject outliers

(b) estimate $\boldsymbol{\theta}$ that performs using the uncensored data

Data-driven design

- Data samples with smaller residuals are informative; more generally the one's with large likelihood
- Sensing function is obtained by solving

$$\min_{\mathbf{w}\in\mathcal{W},\boldsymbol{\theta}} \sum_{m=1}^{D} w_m \left(x_m - \mathbf{a}_m^T\boldsymbol{\theta}\right)^2 \Leftrightarrow \min_{\mathbf{w}\,\in\,\mathcal{W}} r(\mathbf{w})$$

$$r(\mathbf{w}) = \mathbf{x}_w^\mathsf{T}\left(\mathbf{I} - \mathbf{A}_w\left(\mathbf{A}_w^\mathsf{T}\mathbf{A}_w\right)^{-1}\mathbf{A}_w^\mathsf{T}\right)\mathbf{x}_w.$$
$$\mathbf{A}_w = \mathrm{diag}_r(\mathbf{w})\mathbf{A};\, \mathbf{x}_w = \mathrm{diag}_r(\mathbf{w})\mathbf{x}.$$

- Also known as least trimmed squares.
- This problem is non-convex in general
  - Can be convexified for linear Gaussian case
  - Markov chain Monte Carlo methods (e.g., Metropolis-Hastings sampling)

Generalization of data-driven design

- For linear Gaussian models, the cost function amounts to sparsity based outlier rejection methods

  [Fuchs-1999], [Rousseeuw-Leroy-2005], [Giannakis et al.-2011]

- Can be generalized to nonlinear models by optimizing likelihood function parameterized with $\boldsymbol{\theta}$ and $\mathbf{w}$.

  G. Kail, S.P. Chepuri, and G. Leus. Robust Censoring Using Metropolis-Hastings Sampling. *IEEE Journ. of Selec. Topics in Signal Processing*, Nov. 2015.

+ Designs are robust to outliers.

- Sampler needs to be designed for each data realization, and samplers are obtained by optimizing instantaneous measure (performance could be bad).

Hybrid Model-data-driven design

- Data-driven samplers are robust to outliers, but don't take the resulting inference performance (i.e., MSE) into account

- Model-driven samplers are MSE optimal, but are not robust to outliers.

- Hybrid model-data-driven designs allow to combine (and trade) the above two advantages.

+ Designs are robust to outliers, and take into account the inference performance.

- Sampler needs to be designed for each data realization.

Optimization problem

- Hybrid model-data-driven sensing scheme jointly optimizes the likelihood function (i.e., residual) and the MSE:

$$\min_{\mathbf{w} \in \mathcal{W}} \; r(\mathbf{w}) + \lambda f(\mathbf{w})$$

$\mathcal{W} = \{\mathbf{w} \in \{0,1\}^D \mid \|\mathbf{w}\|_0 = d\}$.

- $\lambda \to 0(\infty)$ results in the related data (model)-driven scheme

- This problem is non-convex in general
  - Can be convexified for linear Gaussian case
  - Markov chain Monte Carlo methods (e.g., Metropolis-Hastings sampling)

Convex optimization based solver

- Hybrid model-data-driven sensing design is equivalent to

$$
\begin{aligned}
\min_{\mathbf{w} \in \mathcal{W}, t_1, t_2} \quad & t_1 + \lambda t_2 \\
\text{s.t.} \quad & r(\mathbf{w}) \leq t_1, \\
& f(\mathbf{w}) \leq t_2.
\end{aligned}
$$

- Using Schur complement and $\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} = \mathrm{diag}(\mathbf{w})$

$$
\begin{aligned}
\min_{\mathbf{w} \in \mathcal{W}, t_1, t_2} \quad & t_1 + \lambda t_2 \\
\text{s.t.} \quad & \begin{bmatrix} \mathbf{A}^{\mathsf{T}}\mathrm{diag}(\mathbf{w})\mathbf{A} & \mathbf{A}^{\mathsf{T}}\mathrm{diag}(\mathbf{w})\mathbf{x} \\ \mathbf{x}^{\mathsf{T}}\mathrm{diag}(\mathbf{w})\mathbf{A} & t_1 - \mathbf{x}^{\mathsf{T}}\mathrm{diag}(\mathbf{w})\mathbf{x} \end{bmatrix} \succeq \mathbf{0}, \\
& f(\mathbf{w}) \leq t_2.
\end{aligned}
$$

- The above problem is convex after relaxing $\mathcal{W}$ to $\mathcal{W}_c$.

Metropolis-Hastings sampler

- Powerful concept for generating samples from complex distributions:
  1. Each iteration generates a proposal $\widetilde{\mathbf{w}}$ from some proposal distribution $q\big(\widetilde{\mathbf{w}}\big|\mathbf{w}^{(j-1)}\big)$.
  2. The new sample $\mathbf{w}^{(j)}$ is obtained as

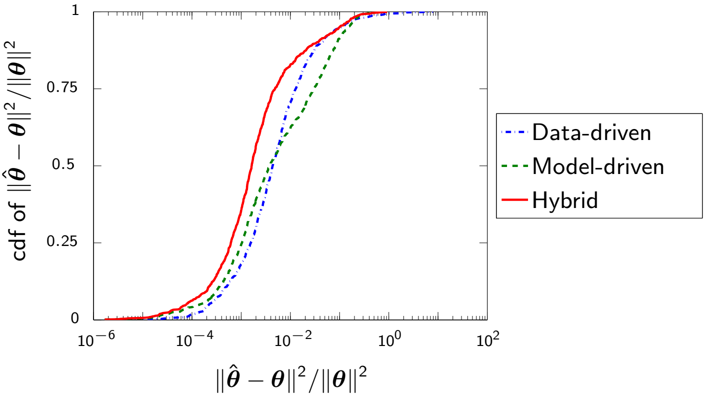
$$\mathbf{w}^{(j)} = \begin{cases} \widetilde{\mathbf{w}} & \text{with probability } \alpha_j \\ \mathbf{w}^{(j-1)} & \text{with probability } 1-\alpha_j \end{cases}$$

$$\alpha_j = \min\left\{ \frac{p_{\mathrm{t}}\big(\widetilde{\mathbf{w}}\big)\, q\big(\mathbf{w}^{(j-1)}\big|\widetilde{\mathbf{w}}\big)}{p_{\mathrm{t}}\big(\mathbf{w}^{(j-1)}\big)\, q\big(\widetilde{\mathbf{w}}\big|\mathbf{w}^{(j-1)}\big)}, 1 \right\}$$

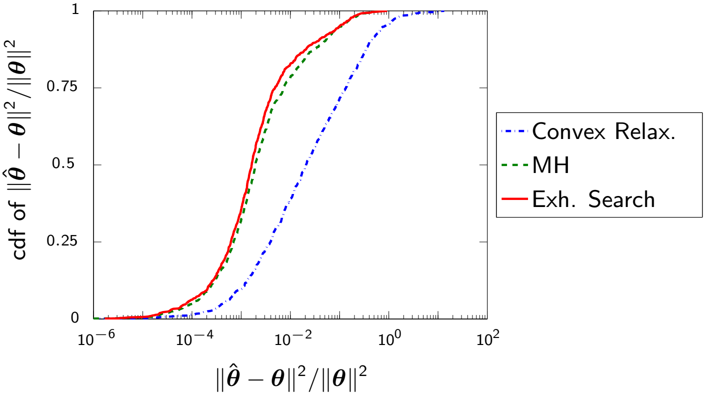
$$p_{\mathrm{t}}(\mathbf{w}) \propto \exp\left(-\frac{1}{2\,\sigma^2}\left(r(\mathbf{w}) + \lambda f(\mathbf{w}) \right) \right)$$

- Doesn't involve convex relaxations ($\mathcal{W}$ to $\mathcal{W}_c$).

Numerical experiments



- Best performance of the hybrid scheme is not necessarily in between the best performances of the data-driven and model-driven schemes.



- Error distribution achieved with the MH method almost coincides with that of exhaustive search.

Conclusions and future directions

- Model-driven, data-driven, hybrid sparse sensing
  - for basic inference problems
  - respective strengths and weaknesses

- Future directions
  - Correlated observations, clustering, and classification
  - Greedy algorithms (submodular)

Thank You!!