Data modeling using nuclear norm heuristic

Ivan Markovsky

University of Southampton

Behavioral approach to data modeling

- data set $\mathscr{D} \subset \mathscr{U} \quad \xrightarrow{\text{data modeling problem}} \quad$ model $\mathscr{B} \in \mathscr{M}$
  - set of all possible observations $\mathscr{U}$
  - model class $\mathscr{M}$
- basic criteria in any data modeling problem are:
  - “simple” model and
  - “good” fit of the data by the model

  contradicting objectives
- core issue in data modeling complexity–accuracy trade-off

- in the classical setting, models are viewed as equations and a model class is a parameterized equation
- in the behavioral setting, models are subsets of $\mathscr{U}$ and equations are used as representations of models
- allows us to define equivalence of model representations
- establish links among data modeling methods
- model complexity and misfit (lack of fit) b/w data and model have appealing geometrical definitions

Complexity–accuracy trade-off

- a linear model $\mathscr{B}$ is a subspace of $\mathscr{U}$
- a complexity measure of $\mathscr{B}$ is its dimension — $\dim(\mathscr{B})$
- misfit — distance from $\mathscr{D}$ to $\mathscr{B}$

$$M(\mathscr{D},\mathscr{B}) := \operatorname{dist}(\mathscr{D},\mathscr{B}) := \min_{\widehat{\mathscr{D}}\subset\mathscr{B}} \|\mathscr{D}-\widehat{\mathscr{D}}\|_{\mathscr{U}}.$$

- data modeling problem: given $\mathscr{D}\subset\mathscr{U}$ and $\|\cdot\|_{\mathscr{U}}$

$$\text{minimize} \quad \text{over all linear models } \mathscr{B} \quad \begin{bmatrix}\dim(\mathscr{B})\\ M(\mathscr{D},\mathscr{B})\end{bmatrix} \qquad \text{(DM)}$$

- a bi-objective optimization problem

Low-rank approximation and rank minimization

- the data set $\mathscr{D}$ can be parameterized by a real vector $p \in \mathbb{R}^{n_p}$ via a map $\mathscr{S} : \mathbb{R}^{n_p} \to \mathbb{R}^{m \times n}$
- $\mathscr{S}$ depends on the application but is often affine
- in static linear modeling problems, $\mathscr{S}(p)$ is unstructured
- in dynamic LTI modeling problems, $\mathscr{S}(p)$ is block-Hankel
- Fact
$$\dim(\mathscr{B}) \geq \operatorname{rank}\big(\mathscr{S}(p)\big) \tag{$*$}$$

Low-rank approximation and rank minimization

- $\|\mathscr{D} - \widehat{\mathscr{D}}\|_{\mathscr{U}} = \|p - \widehat{p}\| = \|p\|$

- weighted 1-, 2-, and ∞-(semi)norms:

$$\|\widetilde{p}\|_{w,1} := \|w \odot \widetilde{p}\|_1 := \textstyle\sum_{i=1}^{n_p} |w_i \widetilde{p}_i|$$

$$\|\widetilde{p}\|_{w,2} := \|w \odot \widetilde{p}\|_2 := \sqrt{\textstyle\sum_{i=1}^{n_p} (w_i \widetilde{p})^2}$$

$$\|\widetilde{p}\|_{w,\infty} := \|w \odot \widetilde{p}\|_\infty := \max_{i=1,\dots,n_p} |w_i \widetilde{p}_i|$$

- w — nonnegative vector, specifying the weights
- $\odot$ — element-wise product

Low-rank approximation and rank minimization

- (DM) becomes a matrix approximation problem:

$$\text{minimize} \quad \text{over } \widehat{p} \quad \begin{bmatrix} \operatorname{rank}\big(\mathscr{S}(\widehat{p})\big) \\ \|p-\widehat{p}\| \end{bmatrix} \tag{DM'}$$

- two possible ways to scalarizations:

1. Misfit minimization with a bound r on the model complexity

$$\text{minimize} \quad \text{over } \widehat{p} \quad \|p-\widehat{p}\| \quad \text{subject to} \quad \operatorname{rank}\big(\mathscr{S}(\widehat{p})\big) \leq r \tag{LRA}$$

2. Model complexity minimization with a bound e on the misfit

$$\text{minimize} \quad \text{over } \widehat{p} \quad \operatorname{rank}\big(\mathscr{S}(\widehat{p})\big) \quad \text{subject to} \quad \|p-\widehat{p}\| \leq e \tag{RM}$$

- (LRA) — low-rank approximation problem
- (RM) — rank minimization problem
- method for solving (RM) can solve (LRA) (using bisection) and vice verse
- varying $r, e \in [0, \infty)$ the solutions of (LRA) and (RM) sweep the trade-off curve (Pareto optimal solutions of (DM))

Nuclear norm heuristics

- with a few exceptions (LRA) and (RM) are non-convex optimization problems
- The main classes of heuristics for solving them are:
  - local optimization methods
  - subspace-type methods, and
  - convex relaxations
- the nuclear norm can be used as a surrogate for the rank
- generalization of the ℓ_1-norm heuristic for sparse vector approximation

Nuclear norm heuristics

- leads to a semidefinite optimization problem
- existing algorithms with provable convergence properties and readily available high quality software packages
- additional advantage is flexibility: affine inequality constraints in the data modeling problem still leads to semidefinite optimization problems
- disadvantage: the number of optimization variables depends quadratically on the number of data points

Nuclear norm heuristics for SLRA

- nuclear norm: $\|M\|_* =$ sum of the singular values of M
- regularized nuclear norm minimization

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{p} \quad \|\mathscr{S}(\widehat{p})\|_* + \gamma \|p - \widehat{p}\| \\ \text{subject to} & G\widehat{p} \leq h \end{array}$$

- using the fact

$$\|M\|_* < \mu \quad \Longleftrightarrow \quad \frac{1}{2}\big(\operatorname{trace}(U) + \operatorname{trace}(V)\big) < \mu \quad \text{and} \quad \begin{bmatrix} U & M^\top \\ M & V \end{bmatrix} \succeq 0$$

we obtain an equivalent SDP problem

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{p}, U, V, \nu \quad \frac{1}{2}\big(\operatorname{trace}(U) + \operatorname{trace}(V)\big) + \gamma\nu \\ \text{subject to} & \begin{bmatrix} U & \mathscr{S}(\widehat{p})^\top \\ \mathscr{S}(\widehat{p}) & V \end{bmatrix} \succeq 0, \quad \|p - \widehat{p}\| < \nu, \quad G\widehat{p} \leq h \end{array}$$

Nuclear norm heuristics for SLRA

- convex relaxation of (LRA)

$$\text{minimize} \quad \text{over } \widehat{p} \quad \|p - \widehat{p}\| \quad \text{subject to} \quad \|\mathscr{S}(\widehat{p})\|_* \leq \mu \tag{RLRA}$$

- motivation: approx. with appropriately chosen bound on the nuclear norm tends to give solutions $\mathscr{S}(\widehat{p})$ of low rank

- (RLRA) can also be written in the equivalent form

$$\text{minimize} \quad \text{over } \widehat{p} \quad \|\mathscr{S}(\widehat{p})\|_* + \gamma \|p - \widehat{p}\| \tag{RLRA'}$$

  γ — regularization parameter related to μ in (RLRA)

- this is a regularized nuclear norm minimization problem

Conclusions

- regularized nuclear norm min. is a general and flexible tool
- can be used as a relaxation for low-rank approximation problems with the following desirable features:
  - arbitrary affine structure
  - any weighted 2-norm or even a weighted semi-norm
  - affine inequality constraints
  - regularization
- issues:
  - effectiveness in comparison with other heuristics
  - currently applicable to small sample sizes problems only

Questions?