Low-rank approximation: a tool for data modeling

Ivan Markovsky

University of Southampton

Outline

Exact line fitting

$$\text{the points } w_i = (x_i, y_i),\ i = 1, \dots, N \text{ lie on a line} \qquad (*)$$

$$\Updownarrow$$

$$\text{there is } (a,b,c) \neq 0, \text{ such that} \quad ax_i + by_i + c = 0, \text{ for } i = 1, \dots, N$$

$$\Updownarrow$$

$$\text{there is } (a,b,c) \neq 0, \text{ such that} \quad \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_N \\ y_1 & \cdots & y_N \\ 1 & \cdots & 1 \end{bmatrix} = 0$$

$$\Updownarrow$$

$$\operatorname{rank} \left(\begin{bmatrix} x_1 & \cdots & x_N \\ y_1 & \cdots & y_N \\ 1 & \cdots & 1 \end{bmatrix} \right) \leq 2 \qquad (**)$$

- restatement of problem $(*)$ as an equivalent problem $(**)$
- however, $(**)$ is a standard problem in linear algebra
- the solution generalizes to
  1. multivariable data (points in $\mathbb{R}^q$) fitted by an affine set
  2. time-series fitting by linear time-invariant dynamical models
  3. data fitting by nonlinear models

Exact conic section fitting

the points $w_i = (x_i, y_i)$, $i = 1, \dots, N$ lie on a conic section

$$\Updownarrow$$

there are $A = A^\top$, b, c, at least one of them nonzero, such that

$$w_i^\top A w_i + b^\top w_i + c = 0, \text{ for } i = 1, \dots, N$$

$$\Updownarrow$$

there is $(a_{11}, a_{12}, a_{22}, b_1, b_2, c) \neq 0$, such that

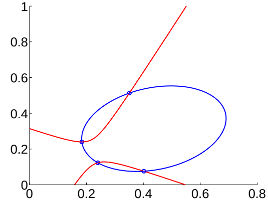
$$\begin{bmatrix} a_{11} & 2a_{12} & b_1 & a_{22} & b_2 & c \end{bmatrix} \begin{bmatrix} x_1^2 & \cdots & x_N^2 \\ x_1 y_1 & \cdots & x_N y_N \\ x_1 & \cdots & x_N \\ y_1^2 & \cdots & y_N^2 \\ y_1 & \cdots & y_N \\ 1 & \cdots & 1 \end{bmatrix} = 0$$

the points $w_i = (x_i, y_i)$, $i = 1, \ldots, N$ lie on a conic section

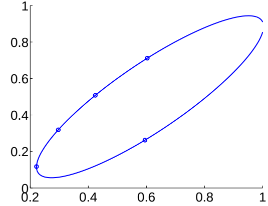
$$\Updownarrow$$

$$\operatorname{rank}\left(\begin{bmatrix} x_1^2 & \cdots & x_N^2 \\ x_1 y_1 & \cdots & x_N y_N \\ x_1 & \cdots & x_N \\ y_1^2 & \cdots & y_N^2 \\ y_1 & \cdots & y_N \\ 1 & \cdots & 1 \end{bmatrix}\right) \leq 5$$

- $N < 5 \rightsquigarrow$ nonunique fit



- $N = 5$ (different points) $\rightsquigarrow$ unique fit



- $N > 5 \rightsquigarrow$ generically no conic section fits the data exactly

Exact fitting by linear homogeneous recurrence relations with constant coefficients

the sequence $w = (w_1, \dots, w_T)$ is generated by linear recurrence relations with lag $\leq \ell$

$$\Updownarrow$$

there is $a = (a_0, a_1, \dots, a_\ell) \neq 0$, such that

$$a_0 w_i + a_1 w_{i+1} + \cdots + a_\ell w_{i+\ell} = 0, \text{ for } i = 1, \dots, T - \ell$$

$$\Updownarrow$$

there is $a = (a_0, a_1, \dots, a_\ell) \neq 0$, such that

$$a^\top \begin{bmatrix} w_1 & w_2 & \cdots & w_{T-\ell} \\ w_2 & w_3 & \cdots & w_{T-\ell+1} \\ \vdots & \vdots & & \vdots \\ w_{\ell+1} & w_{\ell+2} & \cdots & w_T \end{bmatrix} = a^\top \mathscr{H}_\ell(w) = 0$$

the sequence $w = (w_1, \dots, w_T)$ is a linear recursion with lag $\leq \ell$

$$\Updownarrow$$

$$\operatorname{rank}\left(\begin{bmatrix} w_1 & w_2 & \cdots & w_{T-\ell} \\ w_2 & w_3 & \cdots & w_{T-\ell+1} \\ \vdots & \vdots & & \vdots \\ w_{\ell+1} & w_{\ell+2} & \cdots & w_T \end{bmatrix}\right) \leq \ell$$

- $T \leq 2\ell \rightsquigarrow$ there is exact fit (independent of w)
- $T > 2\ell \rightsquigarrow$ generically there is no exact fit

Existence of greatest common divisor

$$p(z) := p_0 + p_1 z + \cdots + p_m z^m \quad \text{and} \quad q(z) := q_0 + q_1 z + \cdots + q_n z^n$$

have a GCD of degree $\geq \ell$

$$\Updownarrow$$

$$\cdots$$

$$\Updownarrow$$

$$\text{rank}\left(\begin{bmatrix} p_0 & & & q_0 & & \\ \vdots & \ddots & & \vdots & \ddots & \\ p_m & & p_0 & q_n & & q_0 \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_m & & & q_n \end{bmatrix} \right) \leq m + n - \ell$$

Data, model, and model class

	line fitting	conic section fitting	linear recurrence with lag $\leq \ell$	GCD
data	points (in $\mathbb{R}^2$)	points (in $\mathbb{R}^2$)	sequence	pair of polynomials
model	line (in $\mathbb{R}^2$)	conic section	autonomous LTI system	polynomials with nontrivial GCD
model class	{ lines (in $\mathbb{R}^2$) }	{ conic sections }	class of LTI systems	?

Continue the sequences

$$(\quad 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ ?\ )$$

$$(\ -5,\ 5,\ 0,\ 5,\ 5,\ 10,\ ?\ )$$

$$(\quad 1,\ 0,\ -1,\ -1,\ 0,\ 1,\ ?\ )$$

An algorithm for continuation of a sequence

Input: $w = (w_1, \dots, w_T)$

1: $\ell := 1$

2: **while** rank$\big(\mathscr{H}_\ell(w)\big) = \ell + 1$ **do**

3: $\quad \ell := \ell + 1$

4: **end while**

5: compute nonzero vector a in the left null space of $\mathscr{H}_\ell(w)$

Output: $w_{T+1} = -\dfrac{1}{a_\ell}(a_0 w_{T-\ell+1} + a_1 w_{T-\ell+2} + \cdots + a_{\ell-1} w_T)$

Outline

Abstract setting for data modeling

- data space $\mathscr{U}$

  examples: $\mathbb{R}^q$, $(\mathbb{R}^q)^T$, $\mathbb{R}[z] \times \mathbb{R}[z]$, $\{\text{true}, \text{false}\}$

- data $\mathscr{D} = \{\mathscr{D}_1, \ldots, \mathscr{D}_N\} \subset \mathscr{U}$

  $\mathscr{D}_i \in \mathscr{U}$ — observation, relalization, or outcome

- model $\mathscr{B} \subset \mathscr{U}$

  an exclusion rule, declares what outcomes are possible

- model class $\mathscr{M} \subset 2^{\mathscr{U}}$

Exact vs approximate models

- $\mathscr{B}$ is an exact model for $\mathscr{D}$ if $\mathscr{D} \subset \mathscr{B}$

  otherwise $\mathscr{B}$ is an approximate model for $\mathscr{D}$

- $\mathscr{B} = \mathscr{U}$ is a (trivial) exact model for any $\mathscr{D} \subset \mathscr{U}$

  $\rightsquigarrow$ we want nontrivial model

  $\rightsquigarrow$ notion of model complexity

- any model is approximate model for any data set

  $\rightsquigarrow$ we need to quantify the approximation accuracy

  $\rightsquigarrow$ notion of model accuracy (w.r.t. the data)

Summary

- data set $\mathscr{D} \subset \mathscr{U}$ $\xrightarrow{\text{data modeling problem}}$ 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

Notes

- in the classical setting, models are viewed as equations
  and a model class is a parameterized equation
- in our setting, models are subsets of the data space $\mathscr{U}$
  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

Model complexity

- the “smaller” a model is the more powerful/useful it is
- the “bigger” a model is the more complex it is
- we prefer simple models over complex ones
- exact modeling problem:

  find the least complex model that fits the data exactly

  (*cf.*, the example on 13)

Linear model complexity

- a linear model $\mathscr{B}$ is a subspace of $\mathscr{U}$ ($\mathscr{U}$ is a vector space)
- the complexity of $\mathscr{B}$ is defined as its dimension
- in the linear case

$$\mathscr{D} \subset \mathscr{B} \qquad \Longrightarrow \qquad \text{span}(\mathscr{D}) \subset \mathscr{B}$$

  and the rank of the data matrix is $\leq \dim(\mathscr{B})$
- $\text{span}(\mathscr{D})$ — the smallest linear model, consistent with $\mathscr{D}$

Model accuracy

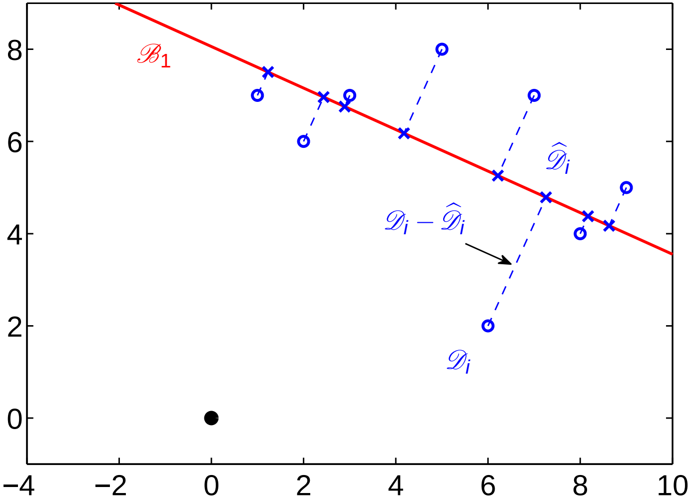
- let $\mathscr{U}$ be a normed vector space with norm $\|\cdot\|$
- the distance between the data $\mathscr{D}$ and a model $\mathscr{B}$

$$\operatorname{dist}(\mathscr{D},\mathscr{B}) := \min_{\widehat{\mathscr{D}}\subset\mathscr{B}} \|\mathscr{D}-\widehat{\mathscr{D}}\| \tag{1}$$

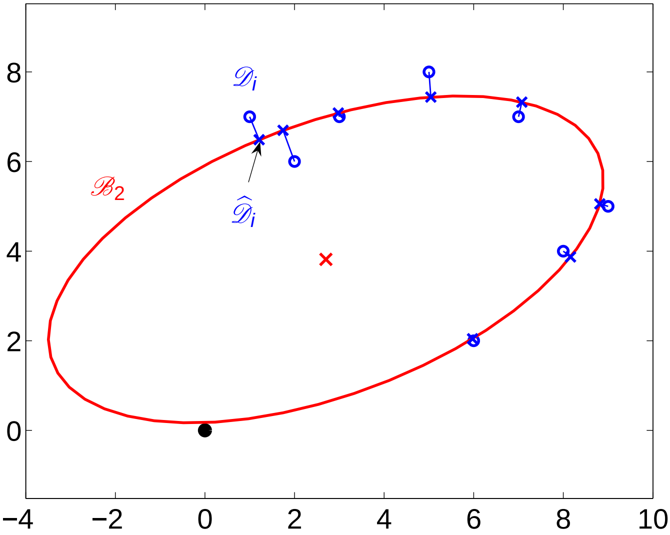
  measures the lack of fit (misfit) between $\mathscr{D}$ and $\mathscr{B}$

- (1) is the projection of the data on the model

Example: $\mathscr{U} = \mathbb{R}^2$, $\mathscr{B}$ linear, Euclidean norm



Example: $\mathscr{U} = \mathbb{R}^2$, $\mathscr{B}$ quadratic, Euclidean norm



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

The data matrix $\mathscr{S}(p)$

- 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

  ($\mathscr{S}$ is affine in case of linear models)
- in static linear modeling problems, $\mathscr{S}(p)$ is unstructured
- in dynamic LTI modeling problems, $\mathscr{S}(p)$ is block-Hankel
- fact

$$\dim(\mathscr{B}) \geq \operatorname{rank}\big(\mathscr{S}(p)\big) \qquad (*)$$

The approximation criterion

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

- 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
- in the stochastic setting of errors-in-variables modeling, $\|\cdot\|$ corresponds to the distribution of the measurement noise

Low-rank approximation and rank minimization

- (DM) becomes a matrix approximation problem:

$$\text{minimize} \quad \text{over } \widehat{p} \quad \begin{bmatrix} \text{rank}\left(\mathscr{S}(\widehat{p})\right) \\ \|p - \widehat{p}\| \end{bmatrix} \qquad \text{(DM')}$$

- two possible scalarizations:

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

$$\text{minimize} \quad \text{over } \widehat{p} \quad \|p - \widehat{p}\| \quad \text{subject to} \quad \text{rank}\left(\mathscr{S}(\widehat{p})\right) \leq r \qquad \text{(LRA)}$$

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

$$\text{minimize} \quad \text{over } \widehat{p} \quad \text{rank}\left(\mathscr{S}(\widehat{p})\right) \quad \text{subject to} \quad \|p - \widehat{p}\| \leq e \qquad \text{(RM)}$$

- (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))

- r is discrete and “small”
  e is continuous and generally unknown

- in applications, an upper bound for r is often specified

Example: approximate line fitting in $\mathbb{R}^2$

$$\text{minimize} \quad \text{over } \mathscr{B} \in \{\,\text{lines}\,\} \quad \text{dist}(\mathscr{D}, \mathscr{B})$$

$$\Updownarrow$$

$$\text{minimize} \quad \text{over } \widehat{x}_i, \widehat{y}_i, i = 1, \dots, N \quad \sum_{i=1}^{N} \left\| \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \widehat{x}_i \\ \widehat{y}_i \end{bmatrix} \right\|_2^2$$

$$\text{subject to} \quad \text{rank}\left(\begin{bmatrix} \widehat{x}_1 & \cdots & \widehat{x}_N \\ \widehat{y}_1 & \cdots & \widehat{y}_N \\ 1 & \cdots & 1 \end{bmatrix} \right) \leq 2$$

can be solved globally using the singular value decomposition of the data matrix

Example: approximate conic section fitting in $\mathbb{R}^2$

$$\text{minimize} \quad \text{over } \mathscr{B} \in \{\text{conic sections}\} \quad \text{dist}(\mathscr{D}, \mathscr{B})$$

$$\Updownarrow$$

$$\text{minimize} \quad \text{over } \widehat{x}_i, \widehat{y}_i, i = 1, \dots, N \quad \sum_{i=1}^{N} \left\| \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \widehat{x}_i \\ \widehat{y}_i \end{bmatrix} \right\|_2^2$$

$$\text{subject to} \quad \text{rank} \left(\begin{bmatrix} \widehat{x}_1^2 & \cdots & \widehat{x}_N^2 \\ \widehat{x}_1 \widehat{y}_1 & \cdots & \widehat{x}_N \widehat{y}_N \\ \widehat{x}_1 & \cdots & \widehat{x}_N \\ \widehat{y}_1^2 & \cdots & \widehat{y}_N^2 \\ \widehat{y}_1 & \cdots & \widehat{y}_N \\ 1 & \cdots & 1 \end{bmatrix} \right) \leq 5$$

Outline

Examples

A setting for data modeling

Solution methods

Algorithms

- with a few exceptions (LRA) and (RM) are non-convex optimization problems
- all general methods are heuristics
- main classes of methods for solving (LRA) and (RM) are:
    - global optimization
    - local optimizations
    - convex relaxations
        - subspace methods and
        - methods based on nuclear norm heuristics

Unstructured low-rank approximation

$$\widehat{D}^* := \arg\min_{\widehat{D}} \|D - \widehat{D}\|_{\mathrm{F}} \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq r$$

Theorem (closed form solution)

Let $D = U\Sigma V^\top$ be the SVD of D and define

$$U =: \begin{array}{cc} \begin{array}{cc} r & n-r \end{array} & \\ \begin{bmatrix} U_1 & U_2 \end{bmatrix} & m \end{array}, \quad \Sigma =: \begin{array}{cc} \begin{array}{cc} r & n-r \end{array} & \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} & \begin{array}{c} r \\ n-r \end{array} \end{array} \quad \text{and} \quad V =: \begin{array}{cc} \begin{array}{cc} r & n-r \end{array} & \\ \begin{bmatrix} V_1 & V_2 \end{bmatrix} & m \end{array}$$

An optimal low-rank approximation solution is

$$\widehat{D}^* = U_1 \Sigma_1 V_1^\top, \qquad (\widehat{\mathscr{B}}^* = \ker(U_2^\top) = \operatorname{colspan}(U_1)).$$

It is unique if and only if $\sigma_r \neq \sigma_{r+1}$.

Structured low-rank approximation

No closed form solution is known for the general SLRA problem

$$\widehat{p}^* := \arg\min_{\widehat{p}} \|p - \widehat{p}\| \quad \text{subject to} \quad \operatorname{rank}\left(\mathscr{S}(\widehat{p})\right) \leq r.$$

NP-hard, consider solution methods based on local optimization

Representing the constraint in a kernel form, the problem is

$$\min_{R,\ RR^\top = I_{m-r}} \left(\min_{\widehat{p}} \|p - \widehat{p}\| \quad \text{subject to} \quad R\mathscr{S}(\widehat{p}) = 0 \right)$$

Note: Double minimization with bilinear equality constraint.

There is a matrix $G(R)$, such that $R\mathscr{S}(\widehat{p}) = 0 \iff G(R)\widehat{p} = 0$.

Variable projection vs. alternating projections

Two ways to approach the double minimization:

- Variable projections (VARPRO):
  solve the inner minimization analytically

$$\min_{R,\ RR^\top = I_{m-r}} \operatorname{vec}^\top\big(R\mathscr{S}(\widehat{p})\big)\Big(G(R)G^\top(R)\Big)^{-1}\operatorname{vec}\big(R\mathscr{S}(\widehat{p})\big)$$

$\rightsquigarrow$ a nonlinear least squares problem for R only.

- Alternating projections (AP):
  alternate between solving two least squares problems

VARPRO is globally convergent with a super linear conv. rate.

AP is globally convergent with a linear convergence rate.

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
- in my experience, the nuclear norm heuristics is less effective than alternative heuristics

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

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

we obtain an equivalent SDP problem

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

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

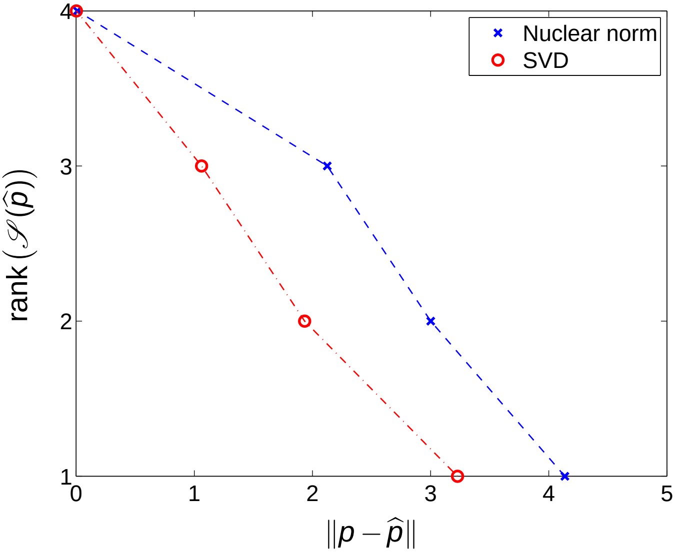
- (RLRA) can also be written in the equivalent form

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

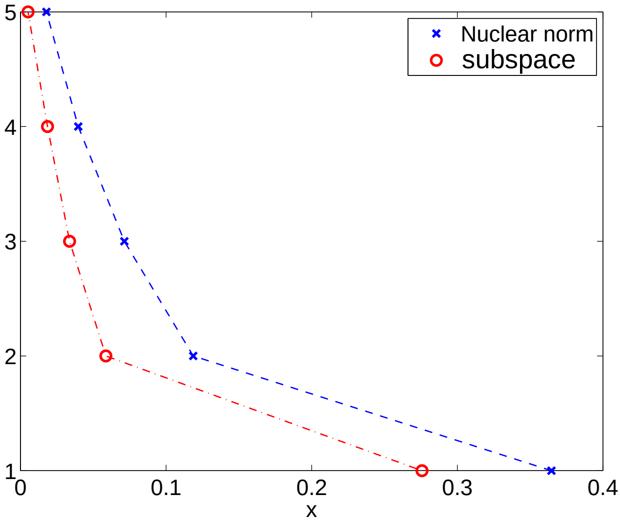
  γ — regularization parameter related to μ in (RLRA)

- this is a regularized nuclear norm minimization problem

Unstructured problem's trade-off curve



Hankel structured problem's trade-off curves



Conclusions

- common pattern in data modeling

  data is exact for a model of bounded complexity

  $$\Updownarrow$$

  matrix constructed from the data is rank deficient

- exact modeling $\approx$ rank computation

- approximate modeling is a biobjective opt. problem

  accuracy vs complexity trade-off

- computationally approx. modeling leads to SLRA and RM

- 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?