Affine data modeling by low-rank approximation

Ivan Markovsky

University of Southampton

Data fitting example

Problem: given data points

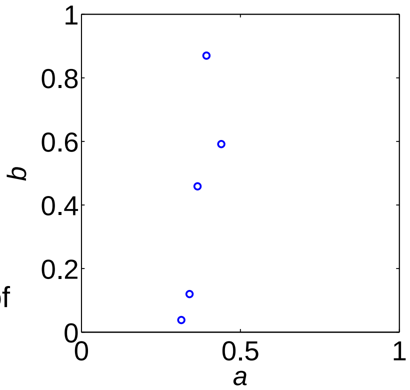
$$d_i = (a_i, b_i), \qquad i = 1, \ldots, N$$

find a fitting line

$$\mathscr{A}_{p,c} = \{ p\ell + c \mid \ell \in \mathbb{R} \}$$

that minimizes the sum-of-squares of the orthogonal distances

$$\operatorname{dist}(\mathscr{A}_{p,c}, d_i) = \min_{\widehat{d}_i \in \mathscr{A}_{p,c}} \| d_i - \widehat{d}_i \|_2$$



A “two-stage solution”

Problem: $\quad$ minimize $\quad$ over p and c $\quad \sum_{i=1}^{N} \big(\text{dist}(\mathscr{A}_{p,c}, d_i)\big)^2$

Heruistic solution method:

1. Compute the data mean

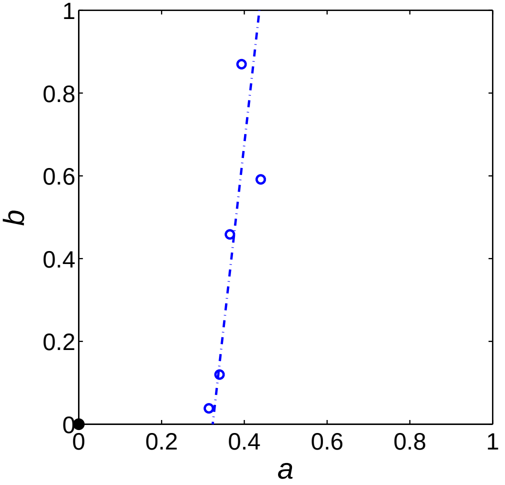
$$\widehat{c} := (d_1 + \cdots + d_N)/N$$

2. Solve the problem

$$\text{minimize} \quad \text{over } p \quad \sum_{i=1}^{N} \big(\text{dist}(\mathscr{A}_{p,\widehat{c}}, d_i)\big)^2$$

Step 2 is a standard problem, which can be solved by the singular value decomposition of the data matrix

$$D := \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}$$



Empirical observation: the two-stage solution $\mathscr{A}_{\widehat{p},\widehat{c}}$ coincides with the optimal solution $\mathscr{A}_{p^*,c^*}$, however, (p^*, c^*) is not unique.

General problem: low-rank approximation

Consider $q \times N$ data matrix and affine space of dimension r.

$$\begin{array}{lll} \text{minimize} & \text{over } \widehat{D} \text{ and } c & \|D - c\mathbf{1}^\top - \widehat{D}\|_\mathrm{F} \\ \text{subject to} & \operatorname{rank}(\widehat{D}) \leq r & \end{array} \qquad (*)$$

Notation:

- $\|D\|_\mathrm{F} := \sqrt{\sum_{i=1}^N \|d_i\|^2}$ — Frobenius norm
- $\mathscr{M}(D) := (d_1 + \cdots + d_N)/N \in \mathbb{R}^q$ — mean
- $\mathscr{C}(D) := D - \mathscr{M}(D)\mathbf{1}^\top$ — centering

Main results

Theorem (Optimality of the two-stage procedure)

A solution to $(*)$ *is the mean of* D, $c^* = \mathscr{M}(D)$, *and an optimal in a Frobenius norm rank-*r *approximation* $\widehat{D}^*$ *of the centered data matrix* $\mathscr{C}(D)$.

Theorem (Nonuniqueness)

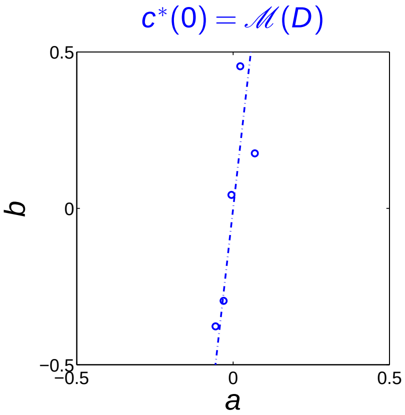
Let

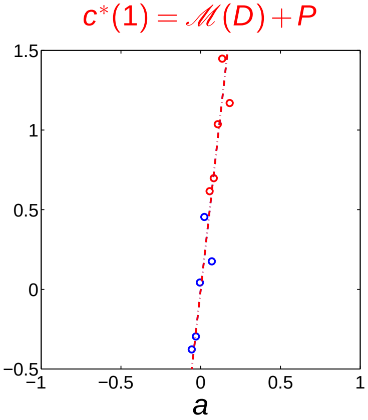
$$\widehat{D} = PL, \qquad \text{where} \quad P \in \mathbb{R}^{q \times r} \text{ and } L \in \mathbb{R}^{r \times N}$$

be a rank revealing factorization of an optimal in a Frobenius norm rank-r *approximation of the centered data matrix* $\mathscr{C}(D)$. *The solutions of* $(*)$ *are of the form*

$$\begin{aligned} c^*(z) &= \mathscr{M}(D) + Pz \\ \widehat{D}^*(z) &= P(L - z\mathbf{1}^\top) \end{aligned} \qquad \text{for} \quad z \in \mathbb{R}^r.$$

Geometry of the nonuniqueness





Conclusions and future work

- In a reprospect the results are trivial, however, as far as I know they are not available in the literature.
- Slight modifications of problem $(*)$ make the two stage procedure suboptimal.
- Modification #1: use a weighted norm

  $\|D\|_\Sigma := \|\Sigma \odot D\|_\mathrm{F}$, where $\Sigma \in \mathbb{R}_+^{q \times N}$ and $\odot$ is Hadamard prod.

- Modification #2: impose constraints on $\widehat{D}$, *e.g.*,
  - nonnegativity
  - Hankel structure.

Thank you