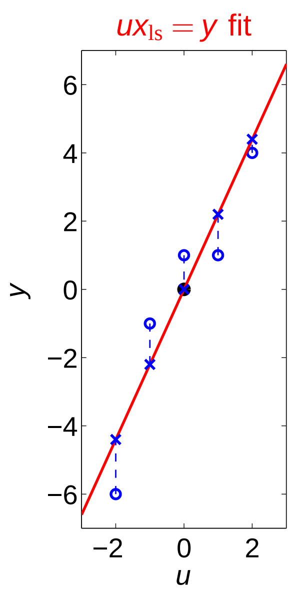
Low-rank approximation: Applications and algorithms

Ivan Markovsky

School of Electronics and Computer Science
University of Southampton

Outline

What is a model?



Classic problem: Fit the points

$$d_1 = \begin{bmatrix} -2 \\ -6 \end{bmatrix},\ d_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix},\ \ldots,\ d_5 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

by a line passing through the origin.

Classic solution: Define $d_i =: \operatorname{col}(u_i, y_i)$ and solve the least squares problem

$$\operatorname{col}(u_1, \ldots, u_5)x = \operatorname{col}(y_1, \ldots, y_5).$$

The model is the line

$$\mathscr{B} := \{\, d = \operatorname{col}(u, y) \mid ux_{\text{ls}} = y \,\}$$

and not the equation $ux_{\text{ls}} = y$.

Model representations

In general,

a linear static model is a subspace $\mathscr{B}$ of the data space $\mathbb{R}^q$

Representations of a linear static model $\mathscr{B} \subseteq \mathbb{R}^q$:

kernel	$\mathscr{B} = \ker(R)$	$:= \{\, d \mid Rd = 0 \,\}$
image	$\mathscr{B} = \operatorname{image}(P)$	$:= \{\, d = Pv \mid \text{for all } v \,\}$
input/output	$\mathscr{B} = \mathscr{B}_{\text{i/o}}(X)$	$:= \{\, d = \operatorname{col}(u, y) \mid Xu = y \,\}$

Links among model representations

input/output kernel image

$$\underset{\text{input/output}}{ux=y} \quad\iff\quad \underset{\text{kernel}}{\underbrace{\begin{bmatrix}x & -1\end{bmatrix}}_{R}\begin{bmatrix}u\\ y\end{bmatrix}=0} \quad\iff\quad \underset{\text{image}}{\begin{bmatrix}u\\ y\end{bmatrix}=\underbrace{\begin{bmatrix}1\\ x\end{bmatrix}}_{P}u}$$

Therefore, $\mathscr{B}=\{\,d=\operatorname{col}(u,y)\mid ux=y\,\}=\ker(R)=\operatorname{image}(P)$

In general:

$$\begin{tikzcd}
\mathscr{B}=\ker(R) \arrow[rr, leftrightarrow, "RP=0"] \arrow[dr, shift left, "X^\top=-R_o^{-1}R_i"] & & \mathscr{B}=\operatorname{image}(P) \arrow[dl, shift right, "X^\top=P_oP_i^{-1}"'] \\
& \mathscr{B}=\mathscr{B}_{i/o}(X) \arrow[ul, shift left, "R=[X^\top\ -I]"] \arrow[ur, shift right, "P^\top=[I\ X]"'] &
\end{tikzcd}$$

Model complexity

The dimension of $\mathscr{B}$, $\dim(\mathscr{B})$ is a measure of $\mathscr{B}$'s "complexity".

In the example,

$$\dim(\mathscr{B})=1=\operatorname{rank}(P)=2-\operatorname{rank}(R)$$

- $\mathscr{B}=\{0\}$ has $\dim(\mathscr{B})=0$ and is the least complex model,
- $\mathscr{B}=$ "line passing through the origin" has $\dim(\mathscr{B})=1$, and
- $\mathscr{B}=\mathbb{R}^2$ has $\dim(\mathscr{B})=2$ and is the most complex model.

Exact modelling

Consider a given data set

$$\mathscr{D}=\{\,d_1,\dots,d_N\,\}\subset\mathbb{R}^q$$

A model $\mathscr{B}\subseteq\mathbb{R}^q$ is exact for the data $\mathscr{D}$ if $\mathscr{D}\subseteq\mathscr{B}$.

Exact data modelling problem:

Find a least complex model $\mathscr{B}$ in a given set of models $\mathscr{M}$ that fits the data $\mathscr{D}$ exactly or assert that such a model doesn't exist.

Notes

- The most basic (and simple) data modelling problem.
- The exact case should be considered before (more complicated) approximate and stochastic cases.
- The solution of the exact problem is useful in the solution of approximate and stochastic cases.
- A core question in all sciences. For example,
  - Kepler's laws define a model that fit exactly planetary trajectories,
  - Newtons laws of dynamics define a model that fit exactly the trajectory of any moving body

Exact linear static model

In the case, when the model class

$$\mathscr{M} = \text{set of all linear static models}$$

the solution of the exact modelling problem is simple.

- A solution always exists and is unique.
- It is given by $\mathscr{B} = \operatorname{image}(D)$, where

$$D := \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix} \in \mathbb{R}^{q \times N}$$

- The complexity of $\mathscr{B}$ is equal to the rank of D

Generically, $\operatorname{rank}(D) = q$, so $\mathscr{B}$ is trivial model (fits everything).

Approximate modelling

In the case, when the model class

$$\mathscr{M} = \text{set of all linear static models of bounded complexity}$$

a solution to the exact modelling problem may not exist.

Approximate modelling problem:

Find a smallest (in specified sense) perturbation of the data $\mathscr{D}$ that renders exact modelling of the perturbed data solvable.

Low-rank approximation (LRA)

Let $\widehat{D} \in \mathbb{R}^{q \times N}$ be the perturbed data. We want

1. $\widehat{D}$ to be as close as possible to D, *e.g.*, $\min \|D - \widehat{D}\|$
2. $\widehat{\mathscr{B}}$ to be an exact model for $\widehat{D}$, *i.e.*, $\widehat{D} \in \widehat{\mathscr{B}}$
3. $\widehat{\mathscr{B}}$ to has complexity bounded by $r < q$, *i.e.*, $\dim(\widehat{\mathscr{B}}) \leq r$

$$\widehat{D} \in \widehat{\mathscr{B}} \quad \text{and} \quad \dim(\widehat{\mathscr{B}}) \leq r \quad \Longrightarrow \quad \operatorname{rank}(\widehat{D}) \leq r$$

Approximate modelling problem: Given $D \in \mathbb{R}^{q \times N}$, r, and $\|\cdot\|$,

$$\text{minimize} \quad \text{over } \widehat{D} \quad \|D - \widehat{D}\| \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq r$$

Relation to regression problems

The classical approach for data fitting is regression ($AX \approx B$). Regression correspond to LRA with input/output representation.

$$AX = B \quad \Longrightarrow \quad \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} X \\ -I \end{bmatrix} = 0 \quad \Longrightarrow \quad \operatorname{rank}(\begin{bmatrix} A & B \end{bmatrix}) \leq \operatorname{coldim}(X)$$

$\Longrightarrow$ indeed, regression is a way to achieve low-rank approximation.

However, $AX = B$ does not imply $\operatorname{rank}(\begin{bmatrix} A & B \end{bmatrix}) \leq \operatorname{coldim}(X)$.

$\rightsquigarrow$ existence of ill-posed regression problems.

Ill-posedness/conditioning is a consequence of imposing input/output structure on the model that is not corroborated by the data.

Rank minimization (RM)

Approximate modeling is a trade-off between:

- fitting accuracy ($\|D - \widehat{D}\|$) and
- model complexity ($\operatorname{rank}(\widehat{D})$)

Two possible scalarizations of the bi-objective optimization are:

LRA: maximize accuracy under a constraint on complexity

RM: minimize complexity under a constraint ($\mathscr{C}$) on accuracy

$$\text{minimize} \quad \text{over } \widehat{D} \quad \operatorname{rank}(\widehat{D}) \quad \text{subject to} \quad \widehat{D} \in \mathscr{C}$$

RM (as well as LRA) is NP-hard, however, there are effective heuristics for RM, *e.g.*, with $\widehat{D} = \operatorname{diag}(\widehat{d})$, $\operatorname{rank}(\widehat{D}) = \operatorname{card}(\widehat{d})$,

$$\ell_1 \text{ heuristic:} \quad \min_{\widehat{d}} \|\widehat{d}\|_1 \quad \text{subject to} \quad \operatorname{diag}(\widehat{d}) \in \mathscr{C}$$

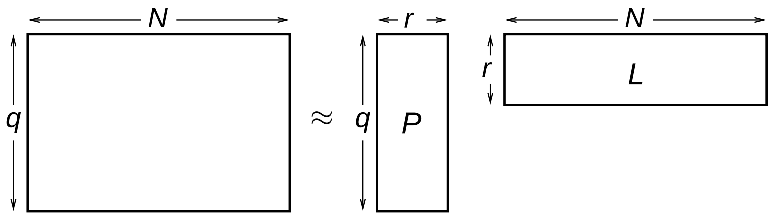
Generalizations

- Cost functions
  - weighted norms $\quad \|\Delta\| = \operatorname{vec}^\top(\Delta) W \operatorname{vec}(\Delta)$
  - information criteria $\quad \|\Delta\| \leftrightarrow \log\det(\Delta)$
- Constraints on $\widehat{D}$
  - structured, *e.g.*, Hankel, Sylvester, sparse
  - nonnegative
  - exact elements
- Data D
  - tensors
  - categorical data
  - missing data

Applications

- System theory
  1. Approximate realization
  2. Model reduction
  3. System identification
- Signal processing
  4. Linear prediction
  5. FIR modeling
  6. Harmonic retrieval
  7. Array processing
  8. Image deblurring
- Computer algebra
  9. Approximate GCD
- Machine learning
  10. Data compression
  11. Natural language proc.
  12. Psychometrics
  13. Recommender systems
- Computer vision
  14. Structure from motion
- Chemometrics
  15. Multivariate calibration

Data compression



qN elements vs $(q+N)r$ elements

For $r \ll \max(q, N)$, $qN \gg (q+N)r \quad \Longrightarrow \quad$ data compression

Natural language processing

Consider N documents, involving q terms and r concepts.

d_{ij} — frequency of ith term in jth document

ℓ_{kj} — relevance of kth concepts to jth document

p_{ik} — frequency of ith term in a document of kth concept only

$$d_j = \begin{bmatrix} d_{1j} \\ \vdots \\ d_{qj} \end{bmatrix} \in \mathbb{R}^q, \qquad p_k = \begin{bmatrix} p_{1k} \\ \vdots \\ p_{qk} \end{bmatrix} \in \mathbb{R}^q, \qquad \ell_k = \begin{bmatrix} \ell_{1j} \\ \vdots \\ \ell_{rN} \end{bmatrix} \in \mathbb{R}^r$$

Latent semantic analysis model:

$$\underbrace{\begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}}_{D} = \underbrace{\begin{bmatrix} p_1 & \cdots & p_r \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} \ell_1 & \cdots & \ell_N \end{bmatrix}}_{L}$$

Exact latent semantic analysis

Assuming

- fewer concepts than terms or documents,
- independent concepts, *i.e.*, $p_1, \ldots, p_r$ linearly independent,
- independent relevance vectors $\ell_1, \ldots, \ell_r$

$\operatorname{rank}(D) =$ the number of concepts related to the documents.

In a rank revealing factorization $D = PL$,

- P indicate relevance of the concepts to the documents
- L indicate the term frequencies related to the concepts

Approximate latent semantic analysis

The LSA model does not hold exactly because

- the notion of (small number of) concepts is an idealization
- linearity assumption $D = PL$ is not likely to hold in practice

LRA is used to find a few concepts explaining the data approx.

Document classification:

similarity of documents is evaluated in the concepts space

Synonym discovery:

terms are clustered in the concepts space

Documents search by keywords:

translate first the keywords to a vector in the concepts space and then finding a cluster of documents nearby this vector

Psychometrics

The data D consists of test scores of a group of people

Psychometrics tries to explain the data as a result of a few underlying abilities.

d_{ij} — score in ith test of jth person

ℓ_{kj} — amount of kth ability in jth person

p_{ik} — score in ith test of a person with kth ability only

Factor analysis model:

$$\underbrace{\begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}}_{D} = \underbrace{\begin{bmatrix} p_1 & \cdots & p_r \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} \ell_1 & \cdots & \ell_N \end{bmatrix}}_{L}$$

$\Longrightarrow \quad \operatorname{rank}(D) = \#$ of abilities relevant for the tests.

verbal, quantitative, and analytical ability, are believed to be most important in explaining one's academic performance.

Outline

Basic low-rank approximation problem

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

Theorem (closed form solution)

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

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

An optimal low-rank approximation solution is

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

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

The basic LRA problem is an exception: other approx. criteria and extra constraints lead to NP-hard problems.

Double minimization nature of LRA:

$$\min_{\widehat{\mathscr{B}} \in \mathscr{M}} \left(\min_{\widehat{D}} \|D - \widehat{D}\|_{\mathrm{F}} \quad \text{subject to} \quad \widehat{D} \in \widehat{\mathscr{B}} \right) \qquad (*)$$

If $\widehat{\mathscr{B}}$ is linear, the inner minimization has analytic solution.
It gives the distance of the data to the model $\widehat{\mathscr{B}}$.

Using an image representation of the model:

$$\min_{P\in\mathbb{R}^{q\times r}} \left(\min_{L\in\mathbb{R}^{r\times N}} \|D - PL\|_{\mathrm{F}} \right) = \min_{L\in\mathbb{R}^{r\times N}} \left(\min_{P\in\mathbb{R}^{q\times r}} \|D - PL\|_{\mathrm{F}} \right) \qquad (**)$$

For fixed P the problem is linear in L and vice verse.

Variable projection vs. alternating projections

Two ways to approach the double minimization:

- Variable projections (VARPRO):
  solve the inner minimization of $(*)$ analytically
  $\rightsquigarrow$ nonlinear least squares problem for the model parameters
- Alternating projections (AP):
  Alternate between the least squares problems, resulting from $(**)$ with fixed P and L, respectively

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

AP is globally convergent with a linear convergence rate.

Summary

- Linear static models $=$ subspaces. Can be represented as image or kernel of matrix, or graph of map (input/output)
- Exact modeling is not practical but is conceptually useful.
- Approximate modeling is a bi-objective optimization: accuracy vs complexity trade-off.
- LRA — approximate modeling with complexity bound. Regression is special case when input/output is used.
- Most LRA problems have no analytic solution. Two basic solution approaches: VARPRO and AP

Thank you