Chapter 1
Nonlinearly structured low-rank approximation

Ivan Markovsky and Konstantin Usevich

Abstract
Polynomially structured low-rank approximation problems occur in

- algebraic curve fitting, *e.g.*, conic section fitting,
- subspace clustering (generalized principal component analysis), and
- nonlinear and parameter-varying system identification.

The maximum likelihood estimation principle applied to these nonlinear models leads to nonconvex optimization problems and yields inconsistent estimators in the errors-in-variables (measurement errors) setting. We propose a computationally cheap and statistically consistent estimator based on a bias correction procedure, called adjusted least-squares estimation. The method is successfully used for conic section fitting and was recently generalized to algebraic curve fitting. The contribution of this book's chapter is the application of the polynomially structured low-rank approximation problem and, in particular, the adjusted least-squares method to subspace clustering, nonlinear and parameter-varying system identification. The classical in system identification input-output notion of a dynamical model is replaced by the behavioral definition of a model as a set, represented by implicit nonlinear difference equations.

Key words: structured low-rank approximation, conic section fitting, subspace clustering, nonlinear system identification.

Ivan Markovsky and Konstantin Usevich
Department ELEC, Vrije Universiteit Brussel, Pleinlaan 2, Building K, B-1050 Brussels, Belgium,
e-mail: Ivan.Markovsky@vub.ac.be, e-mail: Konstantin.Usevich@vub.ac.be

1.1 Introduction

Data modeling, missing data estimation, and dimensionality reduction problems are closely related to the problem of approximating a given matrix by another matrix of reduced rank. Apart from the approximation criterion and the desired rank, the low-rank approximation problem involves additional constraints that represent prior knowledge about the to-be-estimated "true" data generating system. Common examples are non-negativity and structure (*e.g.*, Hankel, Toeplitz, and Sylvester) of the approximation matrix.

The reduced rank of the approximation matrix corresponds to the reduction of dimensionality as well as to the reduction of the model complexity in data modeling. In linear time-invariant system identification, for example, the rank of the data matrix is related to the order of the model. By the Eckart-Young-Mirsky theorem [Eckart and Young(1936)], unstructured optimal in spectral and Frobenius norm reduced rank approximation is obtained from the truncated singular value decomposition of the matrix. With a few exceptions, this result has not been generalized to structured approximation problems and weighted approximation criteria. For structured weighted approximation problems convex relaxations as well as local optimization methods have been developed, see [Markovsky(2012)].

In this book's chapter, we consider the low-rank approximation problem with the constraint that the rank deficient matrix is polynomially structured. Formally, the *polynomially structured low-rank approximation problem* is defined as follows.

Given a data matrix D, an approximation criterion $\|D-\widehat{D}\|$, a polynomial mapping $\Phi : \widehat{D} \mapsto \widehat{D}_{\text{ext}}$, and an upper bound on the rank r,

$$\begin{aligned} &\text{minimize} && \text{over } \widehat{D} \quad \|D-\widehat{D}\| \\ &\text{subject to} && \operatorname{rank}\big(\Phi(\widehat{D})\big) \leq r. \end{aligned} \qquad \text{(PSLRA)}$$

The polynomially structured low-rank approximation problem (PSLRA) has applications in

- curve fitting [Markovsky(2012), Chapter 6],
- manifold learning [Ma and Fu(2011), Zhang and Zha(2005)],
- subspace clustering [Vidal *et al.*(2005)], and
- nonlinear system identification [Vandersteen(1997), Vajk and Hetthéssy(2003)].

The simplest special case of nonlinear curve fitting is conic section fitting, which leads to low-rank approximation with quadratic structure constraint, see Sections 1.2.1 and 1.3. More involved is the application to subspace clustering, which is low-rank approximation with Veronese structure of the approximation and an additional (factorizability) condition on the kernel.

As an optimization problem, (PSLRA) is nonconvex. Contrary to affine structured low-rank approximation problems (see [Markovsky(2008), Markovsky(2014)] for an overview of recent results on this problem), (PSLRA) does not allow the approximation matrix $\widehat{D}$ to be eliminated analytically via the variable projections method [Golub and Pereyra(2003)]. Therefore, the number of optimization variables is of the order of magnitude of the number of data points. This makes the use of local optimization methods infeasible for medium to large scale polynomially structured low-rank approximation problems.

Closely related to low-rank approximation is the principal component analysis method [Jolliffe(2002), Jackson(2003)]. Principal component analysis gives a stochastic interpretation of the deterministic low-rank approximation. Vice verse, low-rank approximation is a deterministic optimization problem resulting from the principal component analysis method. Nonlinearly structured low-rank approximation problems are considered in the principal component analysis context under the names of principal curves [Hastie and Stuetzle(1989)] and kernel principal component analysis [Schölkopf *et al.*(1999)], [Bishop(2006), Chapter 12]. The kernel principal component analysis method is unstructured low-rank approximation of the matrix $\Phi(\mathscr{D})$, *i.e.*, it does not impose the polynomial structure of the approximating matrix.

We adopt the errors-in-variables stochastic model, *i.e.*, the given data is obtained from true data that satisfies a true data generating model plus additive noise, see [Cheng and Schneeweiss(1998)]. The noise is assumed to be zero mean, independent, Gaussian identically distributed with a covariance matrix that is known up to a scaling factor. The solution of the polynomially structured low-rank approximation problem (PSLRA) is a maximum likelihood estimator in the errors-in-variable setting. It is well known, see, *e.g.*, [Kukush and Zwanzig(1996)], that the maximum likelihood estimator is inconsistent in nonlinear errors-invariables estimation problems.

The method proposed in this book's chapter is a generalization of the adjusted least squares method of [Kukush *et al.*(2004), Markovsky *et al.*(2004)] developed for ellipsoid fitting. The adjustment procedure is motivated from the idea of correcting for the bias of the unstructured low-rank approximation method. The bias correction is explicitly given in terms of the noise variance and a procedure for the estimation of the noise variance is proposed. A generalization of the adjusted least squares method to algebraic curve fitting is described in [Markovsky(2012), Chapter 6]. In this contribution, we show that polynomially structured low-rank approximation problems appear naturally in subspace clustering and nonlinear system identification, so that the adjusted least squares algorithm is a promising estimation method also in these application areas.

1.1.1 Outline

In Section 1.2, we start with an overview of the application of polynomially structured low-rank approximation to conic section fitting, subspace clustering, and nonlinear system identification. The data is assumed exact and in the original problem is reduced to rank deficiency of a polynomially structured matrix depending on the

data. The left kernel of the rank deficient matrix contains the parameters of the exact fitting model. Section 1.3 deals with the conic section fitting problem in the presence of noise. Two popular methods in computer vision—algebraic and geometric conic section fitting—are related to structured low-rank approximation. Section 1.4 generalizes the results of Section 1.3 to higher order curves. The algebraic and geometric fitting methods, however, are inconsistent in the errors-in-variables setting. This motivates the development of the bias correction procedure in Section 1.5. Section 1.7 outlines current and future work.

1.2 Applications

1.2.1 Conic section fitting

A conic section is a set of points defined by a second order algebraic equation:

$$\mathscr{B}(S,u,v) = \{\, d \in \mathbb{R}^2 \mid d^\top S d + u^\top d + v = 0 \,\}.$$

The symmetric matrix S, the vector u, and the scalar v are parameters of the conic section. At least one of them is assumed to be nonzero, so that the trivial case $\mathscr{B}(0,0,0) = \mathbb{R}^2$ is excluded. The class of conic sections include lines, union of two lines, hyperbolas, parabolas, and ellipses.

The *conic section fitting problem* is informally defined as follows:

Given a set of points

$$\mathscr{D} = \{\, d_1, \ldots, d_N \,\} \subset \mathbb{R}^2,$$

find a conic section $\widehat{\mathscr{B}} = \mathscr{B}(S,u,v)$, such that data $\mathscr{D}$ is "well" approximated by the fitting curve $\widehat{\mathscr{B}}$.

The approximation criterion is specified by a distance measure $\operatorname{dist}(\mathscr{D},\widehat{\mathscr{B}})$—the smaller the distance, the better the fit. Two different distance measures—the so called algebraic and geometric distance measures—and the corresponding approximation problems—algebraic and geometric conic section fitting problems—are considered in Section 1.3.

Next, we consider the *exact conic section fitting problem:*

Find $\widehat{\mathscr{B}} = \mathscr{B}(S,u,v)$, such that data $\mathscr{D}$ is fitted exactly by the curve $\widehat{\mathscr{B}}$, *i.e.*,

$$\mathscr{D} \subset \widehat{\mathscr{B}}. \qquad \text{(ExactFit)}$$

By definition, the points $d_i = (a_i, b_i)$, $i = 1, \ldots, N$ lie on a (nontrivial) conic section if there is a symmetric matrix S, a vector u, and a scalar v, at least one of them nonzero, such that

$$d_i^\top S d_i + u^\top d_i + v = 0, \qquad \text{for} \quad i = 1, \dots, N.$$

Equivalently, with $S = \left[\begin{smallmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{smallmatrix}\right]$ and $u = \left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$, the exact fitting condition (ExactFit) is that there is a nonzero vector

$$\theta := \begin{bmatrix} s_{11} & s_{12} & u_1 & s_{22} & u_2 & v \end{bmatrix},$$

such that

$$\underbrace{\begin{bmatrix} s_{11} & s_{12} & u_1 & s_{22} & u_2 & v \end{bmatrix}}_{\theta} \underbrace{\begin{bmatrix} a_1^2 & \cdots & a_N^2 \\ 2a_1b_1 & \cdots & 2a_Nb_N \\ a_1 & \cdots & a_N \\ b_1^2 & \cdots & b_N^2 \\ b_1 & \cdots & b_N \\ 1 & \cdots & 1 \end{bmatrix}}_{\Phi(\mathscr{D})} = 0. \tag{$*$}$$

Since the matrix $\Phi(\mathscr{D})$ constructed from the data has six rows and a left kernel of dimension at least one ($\theta \neq 0$), the exact fitting condition (ExactFit) is furthermore equivalent to the condition that $\Phi(\mathscr{D})$ is rank deficient:

$$\operatorname{rank}\left(\Phi(\mathscr{D})\right) \leq 5.$$

If an exact conic section fit of data $\mathscr{D}$ is nonunique (see Figure 1.1 for an example with four data points), the left kernel of $\Phi(\mathscr{D})$ has dimension higher than one. Moreover, all exact conic section fits of data $\mathscr{D}$ are parameterized by the vectors in the left kernel of $\Phi(\mathscr{D})$.

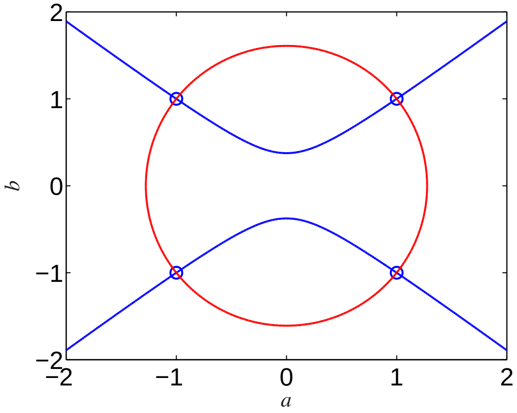

Fig. 1.1 Example of a nonunique solution: there are infinitely many conic sections fitting the four data points (circles) exactly. Two of them are shown in the figure.

Summary:

- The exact conic section fitting problem is a rank test problem for a matrix $\Phi(\mathscr{D})$ that depends quadratically on the data $\mathscr{D}$.
- All exact models

$$\mathscr{B}(\theta) := \mathscr{B}\left(\begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}, \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, v \right) \tag{$\mathscr{B}(\theta)$}$$

  are obtained from the left kernel of $\Phi(\mathscr{D})$ via ($*$).
- The approximate conic section fitting problem is a quadratically structured low-rank approximation problem. The full rank matrix $\Phi(\mathscr{D})$ is approximated by a matrix $\Phi(\widehat{\mathscr{D}})$ with the same structure and rank at most five. The conic section approximation of the data is the exact model for $\widehat{\mathscr{D}}$. This problem is further treated in Section 1.3.

1.2.2 Subspace clustering

In the conic section fitting problem, the set of candidate models (the model class) is the set of conic sections. In this section, the data

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

is fitted by a model $\mathscr{B} \subset \mathbb{R}^q$ that is the union of n-subspaces $\mathscr{B}_1, \dots, \mathscr{B}_n$ with bounded dimensions

$$\dim(\mathscr{B}_1) \leq r_1, \dots, \dim(\mathscr{B}_n) \leq r_n.$$

The union of subspaces model admits a representation

$$\mathscr{B}(R^1, \dots, R^n) = \{\, d \in \mathbb{R}^q \mid (R^1 d) \cdots (R^n d) = 0 \,\},$$

where $R^1 \in \mathbb{R}^{(q-r_1)\times q}, \dots, R^n \in \mathbb{R}^{(q-r_n)\times q}$ are parameters of the model. At least one of the R^i's is assumed to be nonzero in order to avoid the trivial model $\mathscr{B}(0, \dots, 0) = \mathbb{R}^q$. Note that in the case $q = 2$ and $n = 2$, with $r_1 = r_2 = 1$, the union of two lines model $\mathscr{B}(R^1, R^2)$ is a special conic section $\mathscr{B}(S, u, v)$, with

$$S = (R^1)^\top R^2 + (R^2)^\top R^1, \quad u = 0, \quad \text{and} \quad v = 0.$$

Fitting a set of points $\mathscr{D}$ in $\mathbb{R}^q$ by a union of lines model $\mathscr{B}(R^1, \dots, R^n)$ is a type of a clustering problem. Indeed, the data $\mathscr{D}$ is clustered into the r subspaces:

$$\mathscr{B}_i = \mathscr{B}(R^i) = \{\, d \in \mathbb{R}^q \mid R^i d = 0 \,\} \quad \text{for } i = 1, \dots, n.$$

The problem of fitting the model $\mathscr{B}(R^1, \dots, R^n)$ to the data $\mathscr{D}$ is the subspace clustering of [Vidal *et al.*(2005)], also called the generalized principal component analysis problem.

Next, we consider a simplified version of the subspace clustering problem when $q = 2$ and $r = 2$ and the data is fitted exactly.

Given a data set $\mathscr{D}$, find $\widehat{\mathscr{B}} = \mathscr{B}(R^1, R^2)$, such that $\mathscr{D}$ is fitted exactly by $\widehat{\mathscr{B}}$, *i.e.*, (ExactFit) holds.

The data points $d_i \in \mathbb{R}^2$, $i = 1, \dots, N$ lie on a union of two lines if and only if there are vectors R^1 and R^2, at least one of which is nonzero, such that

$$(R^1 d_i)(R^2 d_i) = 0, \qquad \text{for} \quad i = 1, \dots, N.$$

This condition can be written in a matrix form as

$$\underbrace{\begin{bmatrix} R_1^1 R_1^2 & R_1^1 R_2^2 + R_2^1 R_1^2 & R_2^1 R_2^2 \end{bmatrix}}_{\theta} \underbrace{\begin{bmatrix} a_1^2 & \cdots & a_N^2 \\ a_1 b_1 & \cdots & a_N b_N \\ b_1^2 & \cdots & b_N^2 \end{bmatrix}}_{\Phi(\mathscr{D})} = 0. \qquad (**)$$

We showed that if (ExactFit) holds,

$$\operatorname{rank}(\Phi(\mathscr{D})) \leq 2.$$

In subspace clustering, the rank constraint is only a *necessary* condition for exact data fitting. In addition, a basis vector θ of the left kernel of $\Phi(\mathscr{D})$ should have the structure

$$\begin{aligned} \theta_1 &= 1 \\ \theta_2 &= \alpha + \beta \\ \theta_3 &= \alpha\beta, \end{aligned} \qquad (***)$$

for some α and β. This is a polynomial factorization condition that makes possible to map the estimated parameter θ to the the model parameters R^1, R^2 by solving the equations:

$$\begin{aligned} \theta_1 &= R_1^1 R_1^2 \\ \theta_2 &= R_1^1 R_2^2 + R_2^1 R_1^2 \\ \theta_3 &= R_2^1 R_2^2. \end{aligned} \qquad \text{(FACTORIZE)}$$

Applied on the data in the example of Figure 1.1, the kernel computation of the matrix $\Phi(\mathscr{D})$, followed by the solution of (FACTORIZE) yields the exact fit shown in Figure 1.2. Note that the obtained model $\mathscr{B}(R^1, R^2)$ is a particular conic section fitting exactly the data.

Summary:

- A necessary condition for exact subspace clustering is rank deficiency of a matrix $\Phi(\mathscr{D})$ that depends quadratically on the data $\mathscr{D}$ with an additional factorizability condition.
- All exact union of subspaces models are obtained from the left kernel of $\Phi(\mathscr{D})$ by solving a system of nonlinear equations. In the special case of union of two lines, the matrix $\Phi(\mathscr{D})$ is given in $(**)$, the factorization condition is $(***)$, and the system of equations in (FACTORIZE).
- The approximate subspace clustering problem is a quadratically structured low-rank approximation problem with a factorizability constraint. Currently, there are no specialized methods developed for solving this problem. The approach used instead is to solve the structured low-rank approximation problem without the factorizability constraint and then solve the factorization problem approximately.

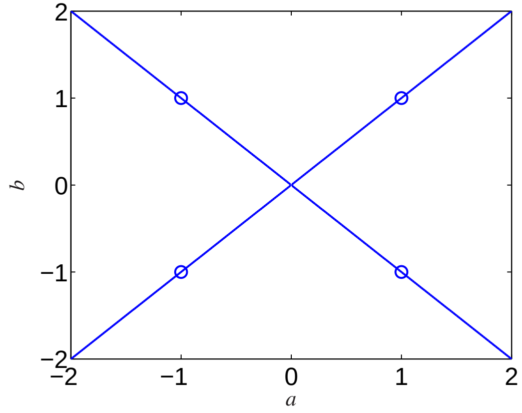

Fig. 1.2 Example of subspace clustering: fitting the data (circles) by a union of two lines.

1.2.3 Nonlinear system identification

The conic section and subspace clustering applications, reviewed in the previous sections, have the following main features:

1. *multivariable data:* the relation among several observed variables $d_1, \dots, d_q$ is modeled, and
2. *nonlinear model:* the modeled relation among the observed variables is nonlinear. In nonlinear system identification, an additional feature is:
3. *dynamical model:* the data w is a time series[1]

$$w = \big(w(1), \dots, w(T)\big), \qquad \text{where} \quad w(t) \in \mathbb{R}^q$$

and the modeled relation involves the variables at different moments of time.

Let σ be the backwards shift operator

[1] We use the notation d for data in problems involving static models and w for data in problems involving dynamical models.

$$(\sigma w)(t) = w(t+1).$$

A finite-dimensional nonlinear multivariable dynamical model $\mathscr{B}$ is defined by a relation R among the variables of w and a finite number ℓ of their shifts $\sigma w, \dots, \sigma^\ell w$, *i.e.*,

$$\mathscr{B}(R) = \{\, w \mid R(w, \sigma w, \dots, \sigma^\ell w) = 0 \,\}. \tag{KER}$$

We refer to (KER) as a *kernel representation* of the system $\mathscr{B} = \mathscr{B}(R)$.

Following the behavioral setting in systems and control (see, *e.g.*, the three-part paper [Willems(1987)] and the book [Polderman and Willems(1998)]), in (KER) we make no separation of the variables into inputs and outputs. This contrasts with the classical definition [Sontag(1990)] of a (nonlinear) dynamical system is a signal processor, accepting one variable u as an input and producing another variable as an output y:

$$u \xrightarrow{\text{system}} y.$$

In discrete-time, the relation between u and y can be defined by a difference equation

$$y = f(u, \sigma u, \dots, \sigma^\ell u, \sigma y, \dots, \sigma^\ell y). \tag{I/O DE}$$

The corresponding dynamical system is

$$\mathscr{B} = \{\, w \mid \text{(I/O DE) holds} \,\}. \tag{I/O}$$

(I/O DE) is refered to as an input-output representation of the system.

An input-output representation (I/O) of a dynamical system $\mathscr{B}$ is a special case of a kernel representation (KER) (take $w = (u,y)$ and $R = y - f(u,y)$). However, not every kernel representation can be representation in an input-output form, *i.e.*, the kernel representation is more general. The importance of this fact is evident already in the static case: conic section and union of two linear models, have no input-output representations $a = f(b)$ or $b = f(a)$ because they are not which are not graphs of functions.

Example 1 (First order SISO model with quadratic nonlinearity). A first order, quadratic, single input single output dynamical system has $q = 2$ variables, *e.g.*, $w_1 = u$ is an input and $w_2 = y$ is an output. Such a system has a kernel representation

$$R(u, y, \sigma u, \sigma y) = \sum_{i+j+k+l=2} R_{ijkl} u^i y^j (\sigma u)^k (\sigma y)^l. \tag{SISO KER}$$

Defining

- the vector of model parameters

$$\theta = \begin{bmatrix} R_{2000} & R_{1100} & R_{1010} & R_{1001} & R_{0200} & R_{0110} & R_{0101} & R_{0020} & R_{0011} & R_{0002} \end{bmatrix}$$

  and
- the corresponding vector of monomials in u, y, σu and σy

$$\phi(w) = \operatorname{col}\big(u^2, uy, u\sigma u, u\sigma y, y^2, y\sigma u, y\sigma y, (\sigma u)^2, \sigma u \sigma y, (\sigma y)^2\big),$$

we see that the kernel representation is linear in the parameters

$$R(u, y, \sigma u, \sigma y) = \theta^\top \phi(w).$$

Example 2 (Wiener-Hammerstein model). A Wiener-Hammerstein model is a block-oriented nonlinear system, where a static nonlinearity f_u is followed by a linear time-invariant system and another static nonlinearity f_y

$$u \xrightarrow{f_u} u' \xrightarrow{\text{LTI}} y' \xrightarrow{f_y} y.$$

Assuming that the function f_y is invertible, the Wiener-Hammerstein model can be rewritten as a kernel representation

$$R(\sigma) f(w) = 0,$$

where

$$f = \begin{bmatrix} f_u \\ f_y^{-1} \end{bmatrix} \quad \text{and} \quad R(\sigma) f(w) = R_0 f(w) + R_1 f(\sigma w) + \cdots + R_\ell f(\sigma^\ell w) = 0.$$

Therefore, the Wiener-Hammerstein model becomes a special case of the nonlinear kernel representation (KER).

Consider, first, the *exact nonlinear system identification problem:*

> Given a time-series w, find a model $\widehat{\mathscr{B}} = \mathscr{B}(R)$ that fits the data exactly, *i.e.*, $w \in \widehat{\mathscr{B}}$.

For a finite time series a nonunique exact fitting model always exists. Of interest is, however, to find the "simplest" in some sense exact model. This leads us to the notion of complexity of a nonlinear system $\widehat{\mathscr{B}} = \mathscr{B}(R)$.

Definition 1 (Polynomial dynamical model's complexity). The complexity of the model $\widehat{\mathscr{B}} = \mathscr{B}(R)$ is the integer triple $(\mathtt{m}, \ell, \mathtt{d}) \in \mathbb{N}^3$

1. $\mathtt{m}$ — number of inputs (independent variables),
2. ℓ — maximum lag, and
3. $\mathtt{d}$ — degree of R.

Example 1 defines a class of models with complexity bounded by $(1,1,2)$. Then, the exact system identification problem becomes a parameter estimation problem:

$$\theta \underbrace{\big[\phi\big(x(1)\big) \cdots \phi\big(x(T-\ell)\big)\big]}_{\Phi(w)} = 0,$$

where

$$x(t) := \operatorname{col}\big(w(t), w(t+1), \dots, w(t+\ell)\big).$$

Summary:

- Exact nonlinear system identification is equivalent to a rank test for a matrix $\Phi(w)$ that depends polynomially on the data w. The matrix $\Phi(w)$ has in addition Hankel structure due to the repeated elements $w(t+1),\dots,w(t+\ell)$ in the columns $\phi(x(t))$ and $\phi(x(t+1))$ of $\Phi(w)$.
- The model parameters are obtained from the left kernel of $\Phi(\mathscr{D})$.
- The approximate nonlinear system identification problem is a polynomially structured low-rank approximation problem.

1.3 Conic section fitting in the errors-in-variables setting

The conic section fitting problem is extensively studied in the computer vision literature, see, *e.g.*, [Bookstein(1979), Gander *et al.*(1994), Kanatani(1994), Fitzgibbon *et al.*(1999)]. The so called "algebraic fitting" methods minimize the equation error and lead to unstructured low-rank approximation. The "geometric fitting" methods minimize the sum of squares of the orthogonal distances from the data points to the fitting curve. These methods lead to polynomially structured low-rank approximation.

As estimators in the errors-in-variables setting, both the algebraic and geometric fitting methods are biased. In [Kukush *et al.*(2004), Markovsky *et al.*(2004)], a bias correction procedure called adjusted least squares method is proposed. The adjusted least squares method is cheap to compute and gives good fits in the geometric sense.

Apart from estimation error, in computer vision, important are the properties of invariance to translation, rotation, and scaling, and the boundedness of the fitting curve (*e.g.*, ellipsoidal fit rather than hyperbolic or parabolic). The invariance properties of the adjusted least squares method are studied in [Shklyar *et al.*(2007)].

1.3.1 Problem formulation

Given a set of points

$$\mathscr{D} = \{d_1,\dots,d_N\} \subset \mathbb{R}^2,$$

the conic section fitting problem aims to find a conic section $\mathscr{B}$, which fits the data $\mathscr{D}$ as well as possible in the sense of minimizing a specified distance measure $\operatorname{dist}(\mathscr{D},\mathscr{B})$ between the data $\mathscr{D}$ and the model $\mathscr{B}$. A natural choice of the fitting criterion is the sum of squares

$$\operatorname{dist}(\mathscr{D},\mathscr{B}) = \sqrt{\sum_{i=1}^{N} \operatorname{dist}^2(d_i,\mathscr{B})}$$

of the orthogonal distances

$$\operatorname{dist}(d_i,\mathscr{B}) := \sqrt{\min_{\widehat{d}_i\in\mathscr{B}} \|d_i-\widehat{d}_i\|^2} \qquad \text{(dist)}$$

from the data points d_i to the curve $\mathscr{B}$.

Let $\mathscr{P}_2$ be the set of conic sections. (dist) leads to the following problem

$$\text{minimize} \quad \text{over } \widehat{\mathscr{B}} \in \mathscr{P}_2 \quad \operatorname{dist}\big(\mathscr{D},\widehat{\mathscr{B}}\big). \qquad \text{(CSF)}$$

Using the representation $\mathscr{B}(\theta)$ of a conic section (see ($\mathscr{B}(\theta)$) on page 6), we obtain the following parameter optimization problem

$$\text{minimize} \quad \text{over } \theta \neq 0 \quad \operatorname{dist}\big(\mathscr{D},\mathscr{B}(\theta)\big).$$

1.3.2 Equivalence to low-rank approximation

In Section 1.2.1, we showed that exact conic section fitting is equivalent to rank deficiency of a structured matrix $\Phi(\mathscr{D})$.

Proposition 1. *The data $\mathscr{D}$ is fitted exactly by a conic section $\mathscr{B} \in \mathscr{P}_2$ if and only if the "extended data matrix"*

$$\Phi(\mathscr{D}) := \big[\phi(d_1)\ \cdots\ \phi(d_N)\big], \qquad \textit{where } \phi(\left[\begin{smallmatrix}a\\b\end{smallmatrix}\right]) =: \big[a^2\ ab\ a\ b^2\ b\ 1\big]^\top$$

has rank less than or equal to 5*,* i.e.*,*

$$\mathscr{D} \subset \mathscr{B} \in \mathscr{P}_2 \quad \Longleftrightarrow \quad \operatorname{rank}\big(\Phi(\mathscr{D})\big) \leq 5.$$

Let

$$D := \big[d_1\ \cdots\ d_N\big] \quad \text{and} \quad \widehat{D} := \big[\widehat{d}_1\ \cdots\ \widehat{d}_N\big]$$

be the data matrix and the approximating matrices, respectively. By Proposition 1, the conic section fitting problem is a quadratically structured low-rank approximation problem

$$\text{minimize} \quad \text{over } \widehat{D} \in \mathbb{R}^{2\times N} \quad \|D-\widehat{D}\|_{\mathrm{F}} \quad \text{subject to} \quad \operatorname{rank}\big(\Phi(\widehat{D})\big) \leq 5.$$

Problem (CSF) defines what is called in the computer vision literature a geometric fitting methods. Geometric fitting is intuitively appealing, however, it leads to hard nonconvex optimization problems. In addition, geometric fitting methods are biased in the errors-in-variables setup, see Section 1.5.

1.3.3 Algebraic fitting method

The algebraic method for conic section fitting is defined by the optimization problem

$$\text{minimize} \quad \text{over } \theta \neq 0 \quad \sqrt{\sum_{i=1}^{N} \|d_i^\top S(\theta) d_i + u^\top(\theta) d_i + v(\theta)\|_2^2}.$$

It has no simple geometrical interpretation, however, it has the advantage of being readily commutable, as shown in the next proposition.

Proposition 2. *Algebraic fitting is equivalent to unstructured low-rank approximation of the extended data matrix $\Phi(\mathscr{D})$.*

The algebraic fitting method coincides with the kernel principal component analysis with the feature map ϕ. Both the geometric and the algebraic fitting methods, however, yield inconsistent estimators. In addition, the geometric fitting method is a hard nonconvex optimization problem. These deficiencies of the methods are corrected by the bias correction procedure, described in Section 1.5, which is computationally cheap and yields a consistent estimator under certain specified assumptions.

1.4 From conic sections to algebraic curves

In Section 1.3, we have seen that conic section fitting in the geometric sense leads to a quadratically structured low-rank approximation problem and in the algebraic sense to unstructured low-rank approximation. This section generalizes these results to algebraic algebraic hypersurfaces (one row) or algebraic varieties (several rows) [Cox *et al.*(2004)]. The corresponding computational problem is polynomially structured low-rank approximation.

Consider a static nonlinear model

$$\mathscr{B} = \ker(R) := \{\, d \in \mathbb{R}^q \mid R(d) = 0 \,\}$$

defined by a multivariable polynomial

$$R_\Theta(d) = \sum_{k=1}^{q_{\text{ext}}} \Theta_k \phi_k(d) = \Theta \phi(d), \qquad (R_\Theta)$$

where Θ is an $\mathtt{p} \times q_{\text{ext}}$ parameter matrix and

$$\phi(d) := \begin{bmatrix} \phi_1(d) & \cdots & \phi_{q_{\text{ext}}}(d) \end{bmatrix}^\top$$

is a vector of a priori chosen monomials $\phi_k(d)$.[2]

In what follows, we assume that the monomials are ordered in $\phi(d)$ in decreasing degree according to the lexicographic ordering (with alphabet the indexes of d). For example, a full parameterization of a second order curve ($\mathtt{d} = 2$) in two variables ($q = 2$) is

[2] The choice of the monomials is related to the model class selection in system identification.

$$q_{\text{ext}} = 6 \qquad \text{and} \qquad \begin{aligned} \phi^\top(x,y) &= \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 & \phi_5 & \phi_6 \end{bmatrix} \\ &= \begin{bmatrix} x^2 & xy & x & y^2 & y & 1 \end{bmatrix} \end{aligned}$$

In general,

$$\phi_k(d) = d_{1\cdot}^{\mathtt{d}_{k1}} \cdots d_{q\cdot}^{\mathtt{d}_{kq}}, \qquad \text{for} \quad k = 1, \dots, q_{\text{ext}}, \qquad (\phi_k)$$

where

- $d_{1\cdot}, \dots, d_{q\cdot} \in \mathbb{R}$ are the elements of $d \in \mathbb{R}^q$, and
- $\mathtt{d}_{ki} \in \mathbb{Z}_+$, is the degree of the ith element of d in the kth monomial ϕ_k.

The matrix formed from the degrees $\mathtt{d}_{ki}$

$$\mathtt{D} = \begin{bmatrix} \mathtt{d}_{ki} \end{bmatrix} \in \mathbb{R}^{q_{\text{ext}} \times q}$$

uniquely defines the vector of monomials ϕ. The matrix of degrees $\mathtt{D}$ depends only on the number of variables q and the degree $\mathtt{d}$. For example, with $q = 2$ and $\mathtt{d} = 2$,

$$\mathtt{D}^\top = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 1 & 0 \end{bmatrix}.$$

Minimality of the kernel representation is equivalent to the condition that the parameter Θ is full row rank. The nonuniqueness of R_Θ corresponds to a nonuniqueness of Θ. The parameters Θ and $Q\Theta$, where Q is a nonsingular matrix, define the same model. Therefore, without loss of generality, we can assume that the representation is minimal and normalize it, so that

$$\Theta \Theta^\top = I_{\mathtt{p}}.$$

Note that a $\mathtt{p} \times q_{\text{ext}}$ full row rank matrix Θ defines via (R_Θ) a polynomial matrix R_Θ, which in turn defines a kernel representation of an algebraic hyperserface $\mathscr{B}_\Theta$ of dimension $\mathtt{m}$ and degree $\mathtt{d}$ (the model). This model class is denoted by $\mathscr{P}^q_{\mathtt{m},\mathtt{d}}$. Thus, Θ defines a function

$$\mathscr{B}_\Theta : \mathbb{R}^{\mathtt{p} \times q_{\text{ext}}} \to \mathscr{P}^q_{\mathtt{m},\mathtt{d}}.$$

Vice verse, a model $\mathscr{B}$ in $\mathscr{P}^q_{\mathtt{m},\mathtt{d}}$ corresponds to a (nonunique) $\mathtt{p} \times q_{\text{ext}}$ full row rank matrix Θ, such that $\mathscr{B} = \mathscr{B}_\Theta$. For a given q, there are mappings

$$\mathtt{d} \to q_{\text{ext}} \qquad \text{and} \qquad \mathtt{m} \to \mathtt{p},$$

defined by

$$q_{\text{ext}} := \binom{q + \mathtt{d}}{\mathtt{d}} = \frac{(q + \mathtt{d})!}{\mathtt{d}!\, q!}$$

and $\mathtt{p} = q - \mathtt{m}$, respectively.

Proposition 3 (Algebraic fit $\iff$ unstructured low-rank approximation). *The algebraic fitting problem for the model class of affine varieties with bounded complexity $\mathscr{P}^q_{\mathtt{m},\mathtt{d}}$*

$$\begin{aligned} &\text{minimize} \quad \text{over } \Theta \in \mathbb{R}^{\mathtt{p}\times q_{ext}} \quad \sqrt{\sum_{j=1}^{N} \left\| R_{\Theta}(d_j) \right\|_{\mathrm{F}}^{2}} \\ &\textit{subject to} \quad \Theta\Theta^{\top} = I_{\mathtt{p}} \end{aligned} \tag{AM$'_{\Theta}$}$$

is equivalent to the unstructured low-rank approximation problem

$$\begin{aligned} &\text{minimize} \quad \text{over } \widehat{\Phi} \in \mathbb{R}^{q\times \mathtt{p}} \quad \|\Phi_{\mathtt{d}}(\mathscr{D}) - \widehat{\Phi}\|_{\mathrm{F}} \\ &\textit{subject to} \quad \operatorname{rank}(\widehat{\Phi}) \leq q_{ext} - \mathtt{p}. \end{aligned} \tag{LRA}$$

Proposition 4 (Geometric fit $\iff$ polynomial structured low rank approx). *The geometric fitting problem for the model class of affine varieties with bounded complexity $\mathscr{P}^{q}_{\mathtt{m},\mathtt{d}}$*

$$\text{minimize} \quad \text{over } \mathscr{B} \in \mathscr{P}^{q}_{\mathtt{m},\mathtt{d}} \quad \operatorname{dist}(\mathscr{D}, \mathscr{B}) \tag{AM}$$

is equivalent to the polynomially structured low-rank approximation problem

$$\begin{aligned} &\text{minimize} \quad \text{over } \widehat{D} \in \mathbb{R}^{q\times N} \quad \|D - \widehat{D}\|_{\mathrm{F}} \\ &\textit{subject to} \quad \operatorname{rank}\big(\Phi_{\mathtt{d}}(\widehat{D})\big) \leq q_{ext} - \mathtt{p}. \end{aligned} \tag{PSLRA}$$

Corollary 1. *The algebraic fitting problem (AM$'_{\Theta}$) is a relaxation of the geometric fitting problem (AM), obtained by removing the structure constraint of the approximating matrix $\Phi_{\mathtt{d}}(\widehat{D})$.*

1.5 Bias correction method for (PSLRA)

Assume that the data $\mathscr{D}$ is generated according to the errors-in-variables model

$$d_j = \overline{d}_j + \widetilde{d}_j, \qquad \text{where} \quad \overline{d}_j \in \overline{\mathscr{B}} \in \mathscr{P}^{q}_{\mathtt{m},q} \qquad \text{and} \quad \operatorname{vec}\left(\begin{bmatrix} \widetilde{d}_1 & \cdots & \widetilde{d}_N \end{bmatrix}\right) \sim \mathrm{N}(0, \sigma^2 I_{qN}). \tag{EIV}$$

Here $\overline{\mathscr{B}}$ is the to-be-estimated true model. The estimate $\widehat{\mathscr{B}}$ obtained by the algebraic fitting method (AM$'_{\Theta}$) is biased, *i.e.*, $\mathbf{E}(\widehat{\mathscr{B}}) \neq \overline{\mathscr{B}}$. In this section, we derive a bias correction procedure. The correction depends on the noise variance σ^2, however, the noise variance can be estimated from the data $\mathscr{D}$ together with the model parameter $\widehat{\Theta}$. The resulting bias corrected estimate $\widehat{\mathscr{B}}_{\mathrm{c}}$ is invariant to rigid transformations. Simulation results show that $\widehat{\mathscr{B}}_{\mathrm{c}}$ has smaller orthogonal distance to the data than alternative direct methods.

Define the matrices

$$\Psi := \Phi_{\mathtt{d}}(\mathscr{D})\Phi_{\mathtt{d}}^{\top}(\mathscr{D}) \qquad \text{and} \qquad \overline{\Psi} := \Phi_{\mathtt{d}}(\overline{\mathscr{D}})\Phi_{\mathtt{d}}^{\top}(\overline{\mathscr{D}})$$

The algebraic fitting method computes the rows of parameter estimate $\widehat{\Theta}$ as eigenvectors related to the $\mathtt{p}$ smallest eigenvalues of Ψ. We construct a “corrected” matrix Ψ_{c}, such that

$$\mathbf{E}(\Psi_{\mathrm{c}}) = \overline{\Psi}. \tag{$*$}$$

This property ensures that the corrected estimate $\widehat{\Theta}_{\mathrm{c}}$, obtained from the eigenvectors related to the $\mathtt{p}$ smallest eigenvalues of Ψ_{c}, is a consistent estimator in the errors-in-variables setting (EIV), *i.e.*, the estimator $\widehat{\theta}$ converges to the true parameter value $\overline{\theta}$ as the sample size N goes to infinity.

The key tool to achieve bias correction is the sequence of the Hermite polynomials, defined by the recursion

$$h_0(x) = 1, \quad h_1(x) = x, \quad \text{and} \quad h_k(x) = x h_{k-1}(x) - (k-2) h_{k-2}(x), \quad \text{for } k = 2, 3, \ldots$$

(See Table 1.1 for explicit expressions of $h_2, \ldots, h_{10}$.) The Hermite polynomials have the deconvolution property

Table 1.1 Explicit expressions of the Hermite polynomials $h_2, \ldots, h_{10}$.

$h_2(x)$	$= x^2 - 1$
$h_3(x)$	$= x^3 - 3x$
$h_4(x)$	$= x^4 - 6x^2 + 3$
$h_5(x)$	$= x^5 - 10x^3 + 15x$
$h_6(x)$	$= x^6 - 15x^4 + 45x^2 - 15$
$h_7(x)$	$= x^7 - 21x^5 + 105x^3 - 105x$
$h_8(x)$	$= x^8 - 28x^6 + 210x^4 - 420x^2 + 105$
$h_9(x)$	$= x^9 - 36x^7 + 378x^5 - 1260x^3 + 945x$
$h_{10}(x)$	$= x^{10} - 45x^8 + 630x^6 - 3150x^4 + 4725x^2 - 945$

$$\mathbf{E}\left(h_k(\overline{x} + \widetilde{x})\right) = \overline{x}^{k}, \qquad \text{where} \quad \widetilde{x} \sim \mathrm{N}(0,1). \tag{$**$}$$

We have,

$$\Psi = \sum_{\ell=1}^{N} \phi(d_\ell)\phi^{\top}(d_\ell) = \sum_{\ell=1}^{N} \left[\phi_i(d_\ell)\phi_j(d_\ell)\right]_{i,j=1}^{q,q},$$

and, from (ϕ_k), the (i,j)th element of Ψ is

$$\psi_{ij} = \sum_{\ell=1}^{N} d_{1\ell}^{\mathtt{d}_{i1}+\mathtt{d}_{j1}} \cdots d_{q\ell}^{\mathtt{d}_{iq}+\mathtt{d}_{jq}} = \sum_{\ell=1}^{N} \prod_{k=1}^{q} (\overline{d}_{k\ell} + \widetilde{d}_{k\ell})^{\mathtt{d}_{iq}+\mathtt{d}_{jq}}.$$

By the data generating assumption (EIV), $\widetilde{d}_{k\ell}$ are independent, zero mean, normally distributed. Then, using the deconvolution property ($**$) of the Hermite polynomials, we have that

$$\psi_{\mathrm{c},ij} := \sum_{\ell=1}^{N} \prod_{k=1}^{q} h_{\mathtt{d}_{ik}+\mathtt{d}_{jk}}(d_{k\ell}) \tag{ψ_{ij}}$$

has the unbiasedness property $(*)$, *i.e.*,

$$\mathbf{E}(\psi_{\mathrm{c},ij}) = \sum_{\ell=1}^{N} \prod_{k=1}^{q} \overline{d}_{k\ell}^{\mathtt{d}_{ik}+\mathtt{d}_{jk}} =: \overline{\psi}_{ij}.$$

The elements $\psi_{\mathrm{c},ij}$ of the corrected matrix are even polynomials of σ of degree less than or equal to

$$\mathtt{d}_{\psi} = \left\lceil \frac{q\mathtt{d}+1}{2} \right\rceil.$$

The following code constructs a $1 \times (\mathtt{d}_{\psi}+1)$ vector of the coefficients of $\psi_{\mathrm{c},ij}$ as a polynomial of σ^2. Note that the product of Hermite polynomials in (ψ_{ij}) is a convolution of their coefficients [Markovsky(2012), Chapter 6].

The corrected matrix

$$\Psi_{\mathrm{c}}(\sigma^2) = \overline{\Psi}_{\mathrm{c}} + \sigma^2 \Psi_{\mathrm{c},1} + \cdots + \sigma^{2\mathtt{d}_{\psi}} \Psi_{\mathrm{c},\mathtt{d}_{\psi}}$$

is then obtained by computing its elements in the lower triangular part

The rows of the parameter $\widehat{\Theta}$ form a basis for the $\mathtt{p}$-dimensional (approximate) null space of $\Psi_{\mathrm{c}}(\sigma^2)$

$$\Theta \Psi_{\mathrm{c}}(\sigma^2) = 0.$$

Computing simultaneously σ and Θ is a *polynomial eigenvalue problem:* the noise variance estimate is the minimum eigenvalue and the parameter estimate is a corresponding eigenvector.

1.6 Numerical example

In this section we illustrate the application of the adjusted least squares method on a problem in nonlinear system identification. Consider the first order single-input single-output system with second order nonlinearity

$$\overline{\mathscr{B}} = \{\, w = (u,y) \mid \sigma y = (u+1)y \,\}.$$

A kernel representation of the system is

$$R(w) = \underbrace{\begin{bmatrix} 1 & 1 & -1 \end{bmatrix}}_{\overline{\theta}} \underbrace{\begin{bmatrix} w_1(t)w_2(t) \\ w_2(t) \\ w_2(t+1) \end{bmatrix}}_{\phi\left(w(t),w(t+1)\right)} = 0.$$

The data is generated in the errors-in-variables setting $w = \overline{w} + \widetilde{w}$, where $\overline{w}$ is a trajectory of the system $\overline{\mathscr{B}}$ with input

$$\overline{u} = \frac{1}{2}(-1,-1,-1,-1,-1,1,1,1,1,1)$$

and initial conditions $\overline{u}(0) = 0$ and $\overline{y}(0) = 1$. The disturbance $\widetilde{w}$ is a zero mean white Gaussian noise. Its standard deviation is varied from zero to a value that corresponds to the signal-to-noise ratio 13.

In order to estimate the model parameters, we approximate the extended data matrix

$$\Phi(w_1) = \begin{bmatrix} \phi\left(w(1),w(2)\right) & \phi\left(w(2),w(3)\right) & \cdots & \phi\left(w(9),w(10)\right) \end{bmatrix} \in \mathbb{R}^{3\times 9}$$

by a matrix of rank 2. One method for doing this is unstructured low-rank approximation, computed via the singular value decomposition. Another method is the adjusted least squares method described in the paper. We compare the methods by Monte Carlo simulation with $K = 500$ noise realizations. The performance criterion is the average parameter error

$$e = \frac{1}{K} \sum_{k=1}^{K} \|\overline{\theta} - \widehat{\theta}^k\|_2, \qquad (e)$$

where $\widehat{\theta}^k$ is the estimate in the kth noise realization. The results are shown in Figure 1.3.

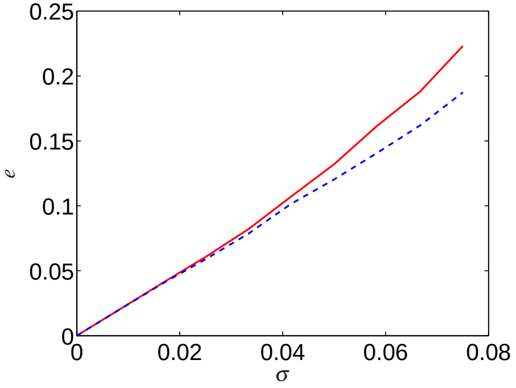

Fig. 1.3 Comparison of unstructured low-rank approximation (solid line) and bias corrected approximation algorithm (dashed line) in terms of the parameter error (e).

1.7 Summary

The polynomially structured low-rank approximation problem (PSLRA) studied in this book's chapter is a generic problem with many applications in machine leaning, computer vision, and system identification. It is, however, a hard nonconvex optimization problem, for which there are currently only heuristic methods. A commonly used heuristic is to ignore the polynomial matrix structure and solve a corresponding unstructured low-rank approximation problem. This approach is known in the machine learning literature as the kernel principal component analysis method. We improved the kernel principal component analysis from a statistical estimation point of view by developing a bias correction procedure, called adjusted least squares. The main assumption is that the data is generated in the errors-in-variables setting and the noise is zero mean independent and Gaussian distributed. The noise variance is estimated from the data. The main computational step is solving a polynomial eigenvalue problem.

Applications of the polynomially structured low-rank approximation problem in conic section fitting, subspace clustering, and nonlinear system identification were presented. Other applications in computer vision are:

- camera calibration,
- motion analysis,
- image matching,
- pose estimation, and
- surface reconstruction.

More generally the adjusted least squares method can be applied on any application where kernel principal component analysis is used, replacing the biased kernel principal component analysis by the consistent adjusted least squares algorithm. This can lead to a significant performance improvement in large sample size and low signal-to-noise cases.

There are links between the adjusted least squares method, the method of [Chojnacki *et al.*(2004), Matei and Meer(2006)] for heteroscedastic errors-in-variables estimators, and the nonlinear dimension reduction method of [Zhang and Zha(2005)]. Current and future work aims at a formal consistency proof of the adjusted least squares estimator for general polynomially structured low-rank models with errors in the variables, invariance of the estimator to translation, rotation and scaling, boundedness of the estimated model, and test on benchmark nonlinear system identification problems.

Acknowledgements

Funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement number 258581

"Structured low-rank approximation: Theory, algorithms, and applications" is gratefully acknowledged.

References

Bishop(2006). Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.

Bookstein(1979). Bookstein, F. L. (1979). Fitting conic sections to scattered data. *Computer Graphics Image Proc.*, **9**, 59–71.

Cheng and Schneeweiss(1998). Cheng, C. and Schneeweiss, H. (1998). Polynomial regression with errors in the variables. *J. R. Stat. Soc. B*, **60**(1), 189–199.

Chojnacki *et al.*(2004). Chojnacki, W., Brooks, M., Hengel, A. V. D., and Gawley, D. (2004). From FNS to HEIV: a link between two vision parameter estimation methods. *IEEE Trans. Pattern Anal. Machine Intelligence*, **26**, 264—268.

Cox *et al.*(2004). Cox, D., Little, J., and O'Shea, D. (2004). *Ideals, varieties, and algorithms*. Springer.

Eckart and Young(1936). Eckart, G. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211–218.

Fitzgibbon *et al.*(1999). Fitzgibbon, A., Pilu, M., and Fisher, R. (1999). Direct least-squares fitting of ellipses. *IEEE Trans. Pattern Anal. Machine Intelligence*, **21**(5), 476–480.

Gander *et al.*(1994). Gander, W., Golub, G., and Strebel, R. (1994). Fitting of circles and ellipses: Least squares solution. *BIT*, **34**, 558–578.

Golub and Pereyra(2003). Golub, G. and Pereyra, V. (2003). Separable nonlinear least squares: the variable projection method and its applications. *Institute of Physics, Inverse Problems*, **19**, 1–26.

Hastie and Stuetzle(1989). Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. American Statistical Association*, **84**, 502–516.

Jackson(2003). Jackson, J. (2003). *A User's Guide to Principal Components*. Wiley.

Jolliffe(2002). Jolliffe, I. (2002). *Principal component analysis*. Springer-Verlag.

Kanatani(1994). Kanatani, K. (1994). Statistical bias of conic fitting and renormalization. *IEEE Trans. Pattern Anal. Machine Intelligence*, **16**(3), 320–326.

Kukush and Zwanzig(1996). Kukush, A. and Zwanzig, S. (1996). On inconsistency of the least squares estimator in nonlinear functional error-in-variables models. Preprint N96-12, Institut für Mathematische Stochastik, Universität Hamburg.

Kukush *et al.*(2004). Kukush, A., Markovsky, I., and Van Huffel, S. (2004). Consistent estimation in an implicit quadratic measurement error model. *Comput. Statist. Data Anal.*, **47**(1), 123–147.

Ma and Fu(2011). Ma, Y. and Fu, Y. (2011). *Manifold Learning Theory and Applications*. CRC Press.

Markovsky(2008). Markovsky, I. (2008). Structured low-rank approximation and its applications. *Automatica*, **44**(4), 891–909.

Markovsky(2012). Markovsky, I. (2012). *Low Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Springer.

Markovsky(2014). Markovsky, I. (2014). Recent progress on variable projection methods for structured low-rank approximation. *Signal Processing*, **96PB**, 406–419.

Markovsky *et al.*(2004). Markovsky, I., Kukush, A., and Van Huffel, S. (2004). Consistent least squares fitting of ellipsoids. *Numerische Mathematik*, **98**(1), 177–194.

Matei and Meer(2006). Matei, B. and Meer, P. (2006). Estimation of nonlinear errors-in-variables models for computer vision applications. *IEEE Trans. Pattern Anal. Machine Intelligence*, **28**(10), 1537–1552.

Polderman and Willems(1998). Polderman, J. and Willems, J. C. (1998). *Introduction to mathematical systems theory*. Springer-Verlag, New York.

Schölkopf *et al.*(1999). Schölkopf, B., Smola, A., and Müller, K. (1999). *Kernel principal component analysis.*, pages 327–352. MIT Press, Cambridge, MA.

Shklyar *et al.*(2007). Shklyar, S., Kukush, A., Markovsky, I., and Van Huffel, S. (2007). On the conic section fitting problem. *Journal of Multivariate Analysis*, **98**, 588–624.

Sontag(1990). Sontag, E. D. (1990). *Mathematical control theory: Deterministic finite dimensional systems*. Springer-Verlag.

Vajk and Hetthéssy(2003). Vajk, I. and Hetthéssy, J. (2003). Identification of nonlinear errors-in-variables models. *Automatica*, **39**(12), 2099–2107.

Vandersteen(1997). Vandersteen, G. (1997). *Identification of Linear and Nonlinear Systems in an Errors-in-Variables Least Square*. Ph.D. thesis, Vrije Universiteit Brussel, http://wwwtw.vub.ac.be/elec/Papers on web/Papers/GerdVandersteen/Phd.pdf.

Vidal *et al.*(2005). Vidal, R., Ma, Y., and Sastry, S. (2005). Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Analysis and Machine Intelligence*, **27**(12), 1945–1959.

Willems(1987). Willems, J. C. (1986, 1987). From time series to linear system—Part I. Finite dimensional linear time invariant systems, Part II. Exact modelling, Part III. Approximate modelling. *Automatica*, **22, 23**, 561–580, 675–694, 87–115.

Zhang(1997). Zhang, Z. (1997). Parameter estimation techniques: A tutorial with application to conic fitting. *Image Vision Comp. J.*, **15**(1), 59–76.

Zhang and Zha(2005). Zhang, Z. and Zha, H. (2005). Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. on Scientific Computing*, **26**, 313–338.