

ALGORITHMS FOR THE WEIGHTED
ORTHOGONAL PROCRUSTES PROBLEM AND
OTHER LEAST SQUARES PROBLEMS

Thomas Viklands

PH.D. THESIS, 2006
UMINF-06.10

DEPARTMENT OF COMPUTING SCIENCE
UMEÅ UNIVERSITY
SE-901 87 UMEÅ, SWEDEN



*Akademisk avhandling som med tillstånd av rektor vid Umeå Universitet fram-
läggs till offentlig granskning fredagen den 7 april klockan 10.15 i sal MA121,
MIT-huset, för avläggande av filosofie doktorsexamen.*

Abstract

In this thesis, we present algorithms for local and global minimization of some *Procrustes* type problems. Typically, these problems are about rotating and scaling a known set of data to fit another set with applications related to determination of rigid body movements, factor analysis and multidimensional scaling. The known sets of data are usually represented as matrices, and the rotation to be determined is commonly a matrix Q with orthonormal columns.

The algorithms presented use Newton and Gauss-Newton search directions with optimal step lengths, which in most cases result in a fast computation of a solution.

Some of these problems are known to have several minima, e.g., the weighted orthogonal Procrustes problem (WOPP). A study on the maximal amount of minima has been done for this problem. Theoretical results and empirical observations gives strong indications that there are not more than 2^n minimizers, where n is the number of columns in Q . A global optimization method to compute all 2^n minima is presented.

Also considered in this thesis is a cubically convergent iteration method for solving nonlinear equations. The iteration method presented uses second order information (derivatives) when computing a search direction. Normally this is a computational heavy task, but if the second order derivatives are constant, which is the case for quadratic equations, a performance gain can be obtained. This is confirmed by a small numerical study.

Finally, regularization of ill-posed nonlinear least squares problems is considered. The quite well known L-curve for linear least squares problems is put in context for nonlinear problems.

Preface

This thesis consists of the following six papers.

- I.** P. Å. Wedin and T. Viklands. Algorithms for 3-dimensional Weighted Orthogonal Procrustes Problems. Technical Report UMINF-06.06, Department of Computing Science, Umeå University, Umeå, Sweden, 2006.
- II.** T. Viklands and P. Å. Wedin. Algorithms for Linear Least Squares Problems on the Stiefel manifold. Technical Report UMINF-06.07, Department of Computing Science, Umeå University, Umeå, Sweden, 2006.
- III.** T. Viklands. On the Number of Minima to Weighted Orthogonal Procrustes Problems. Technical Report UMINF-06.08, Department of Computing Science, Umeå University, Umeå, Sweden, 2006. Submitted for publication in *BIT*.
- IV.** T. Viklands. On Global Minimization of Weighted Orthogonal Procrustes Problems. Technical Report UMINF-06.09, Department of Computing Science, Umeå University, Umeå, Sweden, 2006. Submitted for publication in *BIT*.
- V.** T. Viklands. A Cubic Convergent Iteration Method. Technical Report UMINF-05.10, Department of Computing Science, Umeå University, Umeå, Sweden, 2005.
- VI.** T. Viklands and M. Gulliksson. Optimization Tools for Solving Nonlinear Ill-posed Problems. Fast solution of discretized optimization problems (Berlin, 2000), 255–264, Internat. Ser. Numer. Math., 138, Birkhäuser, Basel, 2001.

In Chapter 1 an introduction to the optimization problems considered is presented along with an overview of all papers. The papers are referred to by their roman numbers.

Acknowledgements

I wish to express my sincere thanks and gratitude to my supervisor Prof. Per-Åke Wedin for his guidance, interest, engagement and for his wide and solid knowledge in mathematics and numerics.

I am also special grateful to Bo Kågström, Frank Drewes and Lennart Edblom for valuable reviews of my papers, and for providing me with financial support.

Thank you Gunilla Wikström and Jerry Eriksson for your help and support. Special thanks also to Mårten Gulliksson for your assistance and cooperation during my first years as a research student, and for several pleasant conference journeys.

I would also like to express my gratitude to Niclas Börlin, Erik Elmroth, Robert Granat, Pedher and Stefan Johansson, Lena Kallin and all the others at the department for their assistance in and discussions regarding optimization, linear algebra, thesis and article writing, administration, numerics etcetera.

Financial support has partly been provided by the Swedish Foundation for Strategic Research under the frame program grant A3 02:128.

Umeå, February 14, 2006

Thomas Viklands

Contents

1	Introduction and overview	1
1.1	Procrustes problems	4
1.1.1	Rigid body movements	4
1.1.2	Psychometrics	5
1.1.3	The OPP and WOPP	7
1.1.4	Canonical form of a WOPP	8
1.2	Linear Least squares problems on the Stiefel manifold	9
2	Algorithms for least squares problems on the Stiefel manifold	11
2.1	The 3-dimensional WOPP, Paper I	11
2.2	Linear least squares problems on the Stiefel manifold, Paper II	11
3	Global minimization of a WOPP	13
3.1	The number of minima to a WOPP, Paper III	14
3.2	Computing all minimizers, Paper IV	15
4	Quadratic equations	17
5	Ill-posed problems	19
5.1	The L-curve for nonlinear problems, Paper VI	20
6	Software	21
6.1	WOPP software	21
6.2	L-curve toolbox	21
7	Research biography and reflections	23
	References	26
	Paper I	31
	Paper II	57
	Paper III	83

Paper IV	109
Paper V	143
Paper VI	159
Notations	175

Chapter 1

Introduction and overview

The main part of this thesis is about an optimization problem known as the *weighted orthogonal Procrustes problem* (WOPP), which we define as:

Definition 1.0.1 With $Q \in \mathbb{R}^{m \times n}$ where $n \leq m$, let A , X and B be known real matrices of compatible dimensions with $\text{rank}(A) = m$ and $\text{rank}(X) = n$. Let $\|\cdot\|_F$ denote the Frobenius matrix norm. The optimization problem

$$\min_Q \|AQX - B\|_F^2, \text{ subject to } Q^T Q = I_n, \quad (1.0.1)$$

is called a *weighted orthogonal Procrustes problem*.

The Frobenius matrix norm can be regarded as the Euclidean norm for matrices. For vectors, the Euclidean norm is commonly known as the 2-norm, $\|\cdot\|_2$. With a vector $y \in \mathbb{R}^k$, its Euclidean length is

$$\|y\|_2 = \sqrt{\sum_{i=1}^k y_i^2}.$$

For a matrix $Y \in \mathbb{R}^{m \times n}$, the Frobenius norm $\|Y\|_F$ is

$$\|Y\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n y_{i,j}^2}.$$

The WOPP is a linear least squares problem defined on a *Stiefel manifold*. A Stiefel manifold [30], commonly denoted $\mathcal{V}_{m,n}$, is the set of all matrices $Q \in \mathbb{R}^{m \times n}$, having orthonormal columns. $\mathcal{V}_{m,n}$ is also referred to as the Stiefel manifold of orthogonal n -frames in \mathbb{R}^m ,

$$\mathcal{V}_{m,n} = \{Q \in \mathbb{R}^{m \times n} : Q^T Q = I_n\}.$$

A set of nonzero vectors $\{q_1, \dots, q_n\}$ in \mathbb{R}^m is said to be *orthogonal* if $q_i^T q_j = 0$ when $i \neq j$. If additionally $q_i^T q_i = 1$ (normalized), the set is said to be orthonormal [12].

The definition of an *orthogonal matrix* is well known. A square matrix $Q \in \mathbb{R}^{m \times m}$ is said to be orthogonal if $Q^T Q = I_m$, hence $Q Q^T = I_m$ and $Q^T = Q^{-1}$. This may seem a bit ambiguous since the columns of Q form not just an orthogonal basis but also an orthonormal basis. When speaking of an *orthonormal matrix* $Q = [q_1, \dots, q_n]$, we mean that the columns of Q form an orthonormal basis, i.e., $Q^T Q = I_n$. If $n = m$ then Q is orthogonal (a square orthonormal matrix).

Trough out this thesis, we assume that $n \leq m$. Evidently $Q^T Q \neq I_n$ if $n > m$.

The WOPP can be regarded as a generalization of the *orthogonal Procrustes problem* (OPP).

Definition 1.0.2 With $Q \in \mathbb{R}^{m \times n}$ where $n \leq m$, let X with $\text{rank}(X) = n$ and B be known real matrices of correct dimensions. We call the optimization problem

$$\min_Q \|QX - B\|_F^2, \text{ subject to } Q^T Q = I_n, \quad (1.0.2)$$

an *orthogonal Procrustes problem*.

The OPP has an analytical solution, that can be derived by using the *singular value decomposition* (SVD) of XB^T . The WOPP on the other hand, typically needs to be solved by iterative optimization algorithms.

To derive the solution to (1.0.2), use the property that for a matrix $Y \in \mathbb{R}^{m \times n}$, $\|Y\|_F^2 = \text{tr}(Y^T Y)$ where $\text{tr}(\cdot)$ is the *matrix-trace*,

$$\text{tr}(\tilde{Y}) = \sum_{i=1}^n \tilde{y}_{i,i}, \quad \tilde{Y} = Y^T Y.$$

We can then write

$$\|QX - B\|_F^2 = \text{tr}((QX - B)^T (QX - B)) =$$

$$= \text{tr}(X^T Q^T Q X) - 2\text{tr}(Q X B^T) + \text{tr}(B^T B) = \|X\|_F^2 - 2\text{tr}(Q X B^T) + \|B\|_F^2,$$

since $Q^T Q = I_n$. Solving (1.0.2) is then done by maximizing $\text{tr}(Q X B^T)$. To do so, let $U \Sigma V^T = X B^T$ be a SVD, then

$$\text{tr}(Q X B^T) = \text{tr}(Q U \Sigma V^T) = \text{tr}(V^T Q U \Sigma) = \text{tr}(\Sigma Z) = \sum_{i=1}^n \sigma_{i,i} z_{i,i} \quad (1.0.3)$$

where $Z = V^T Q U$. Since $Z \in \mathbb{R}^{m \times n}$ is orthonormal, any $z_{i,j} \leq 1$ for all $i = 1, \dots, m$ and $j = 1, \dots, n$. Hence, the sum in (1.0.3) is maximized if $Z = I_{m,n}$, and the solution \hat{Q} to (1.0.2) is given by $\hat{Q} = V I_{m,n} U^T$. The OPP is well studied, and is mentioned in introductory textbooks as, e.g., [12].

As a simple and small example of a WOPP, consider the following: Find the minimum distance between the ellipse

$$x = \alpha_1 \cos \phi, \quad y = \alpha_2 \sin \phi$$

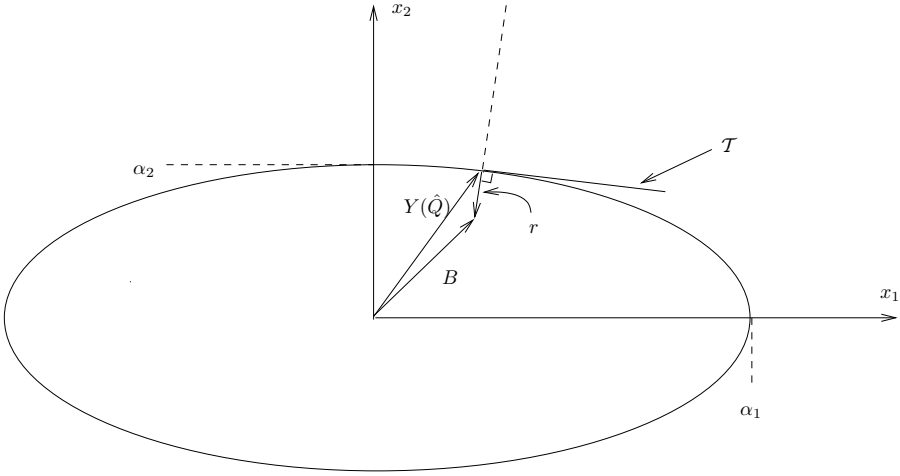


Figure 1: $Y(Q)$ is here an ellipse with semi major and semi minor axes α_1 and α_2 , respectively. The minimum distance occurs at $Y(\hat{Q})$ where the residual $r = B - Y(\hat{Q})$ is orthogonal to the tangent T of $Y(Q)$ at \hat{Q} .

and the point B according to Figure 1.

With $Q = [\cos \phi, \sin \phi]^T$ we can express the ellipse as the vector valued function

$$Y(Q) = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix} = A Q.$$

For vectors, the Frobenius norm is the same as the 2-norm. We can write the optimization problem as a WOPP (with $X = 1$)

$$\min_Q \|A Q - B\|_2^2, \text{ subject to } Q^T Q = 1. \quad (1.0.4)$$

Since $n = 1$ in this case, there are no orthogonality constraints. The task is to find the orthonormal matrix \hat{Q} (a normalized vector, $Q^T Q = 1$), that minimizes the distance between the ellipse $Y(Q)$ and the point B . A solution to (1.0.4) can be computed by using iterative methods as, e.g., Newton's method. Though, for this simple case of a WOPP, a solution can also be computed by solving a fourth degree polynomial, see Paper I. There can also be two different minimizers for (1.0.4), see Section 3.1.

Note that if $\alpha_1 = \alpha_2 = \chi$, $Y(Q)$ describes a circle with radii χ . Then by taking $X = \chi$, (1.0.4) can be written as an OPP

$$\min_Q \|Q X - B\|_2^2, \text{ subject to } Q^T Q = 1. \quad (1.0.5)$$

The solution \hat{Q} to (1.0.5) is unique, and is easily computed by taking $\hat{Q} = B/\|B\|_2$.

Though these examples are very simple, they illustrate the difficulty of computing a solution to a WOPP compared to an OPP.

1.1 Procrustes problems

There are several types of optimization problems involving Stiefel manifolds. One class of problems are different types of Procrustes problems, arising in a wide area of applications. Commonly, the term *Procrustes analysis* or *Procrustes rotation* is used instead of Procrustes problems. For example, orthogonal Procrustes analysis and weighted Procrustes rotation to address the OPP and WOPP, respectively.

The name Procrustes comes from the Greek mythology. Procrustes (the Stretcher) was a robber and torturer that had an iron bed in which he desired to put his victims. To make them fit the bed he cut off their limbs or alternatively stretched them out. In the end, karma came to Procrustes as he was fitted in his own bed by Theseus.

1.1.1 Rigid body movements

The ellipse problem discussed is to the least very simple, due to its low dimension ($m = 2$ and $n = 1$). To give an example of a WOPP of larger dimension, we consider the problem of determining a rigid body movement.

Consider a rigid body with n landmarks x_1, \dots, x_n in \mathbb{R}^3 that is subject to a translation $t \in \mathbb{R}^3$, and a rotation $M \in \mathbb{R}^{3 \times 3}$, taking the landmarks into the positions c_1, \dots, c_n . In rigid body applications, commonly M (for Motion) is used to represent an orthogonal matrix.

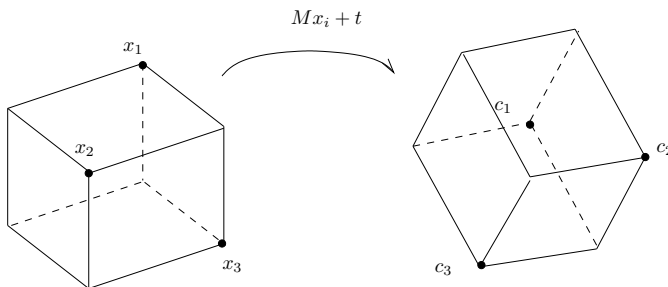


Figure 2: A rigid body with three landmarks undergoing a rotation and translation.

The motion of the rigid body can be written as

$$Mx_i + t = b_i$$

where M , with $\det(M) = 1$, describes the rotations around the three axes. Given x_1, \dots, x_n and c_1, \dots, c_n , the rotation M can be computed by solving an

orthogonal Procrustes problem (OPP) as follows [29]. Let $X = [x_1 - \bar{x}, \dots, x_n - \bar{x}]$ and $C = [c_1 - \bar{c}, \dots, c_n - \bar{c}]$ where \bar{x} and \bar{c} are the mean value vectors of x_i and c_i , $i = 1, \dots, n$, then M is given by solving

$$\min_M \|MX - C\|_F^2, \quad \text{subject to } M^T M = I_3, \det(M) = 1. \quad (1.1.1)$$

The solution is given by using the SVD of XC^T , and if XC^T is nonsingular the solution is unique.

Suppose now that the accuracies of the landmarks x_i and c_i , $i = 1, \dots, n$ are different, dependent on the coordinate axes (in \mathbb{R}^3). Let us say, that along the third axis (z -axis), we have a noticeable lower accuracy than for the first and second axes (x - and y -axes). It is then preferred to give these z -axis coordinates a lesser impact when computing a solution. To do this we can *weight* the OPP by using a weighting matrix A . For instance, let

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \alpha \end{bmatrix},$$

where $0 < \alpha < 1$ is a suitable chosen scalar. The weighted residual is then $A(MX - C)$ and we get a WOPP on the form

$$\min_M \|A(MX - C)\|_F^2, \quad \text{subject to } M^T M = I_3, \det(M) = 1. \quad (1.1.2)$$

Observe that by taking $B = AC$, (1.1.2) is on the form stated in Definition 1.0.1. Commonly iterative optimization algorithms are used to compute a solution to (1.1.2). Moreover, a WOPP can have several minima, which leads to the problem of deciding if a computed solution is the "best" one.

It can also be desired to weight the OPP from right. Assume that the accuracy when measuring some, specific, landmarks is worse than the average. Then by constructing a diagonal matrix W , we can give these landmarks a low weight as $(MX - C)W$. The right-weighted OPP then becomes

$$\min_M \|(MX - C)W\|_F^2, \quad \text{subject to } M^T M = I_3, \det(M) = 1.$$

By taking $X := XW$ and $B = CW$, we see that solving this problem is done by solving an OPP.

1.1.2 Psychometrics

The OPP originates from factor analysis in psychometrics in the 1950s and 1960s, e.g., [16, 18]. The task is to determine an orthogonal matrix $Q \in \mathbb{R}^{m \times m}$, that rotates a factor (data) matrix A , to fit some hypothesis matrix B . Typically in psychometrics, the points to be rotated are ordered row wise in A , not column wise in X as with the rigid body movement example shown above. We denote the rotation of A as $Y(Q) = AQ$. Hence, given A and B , we wish to find Q such that $Y(Q) \approx B$.

When using the Euclidean distance to measure the distance between the rotation $Y(Q)$ and B , the optimal orthogonal matrix Q is given by solving

$$\min_Q \|AQ - B\|_F^2, \text{ subject to } Q^T Q = I_m. \quad (1.1.3)$$

Since $n = m$, resulting in that Q is an orthogonal matrix, (1.1.3) is an OPP with a solution that can be derived by using the SVD of $B^T A$.

A more common formulation of the OPP, in psychometrics, is with the usage of the matrix-trace,

$$\min_Q \text{tr}(AQ - B)^T (AQ - B), \text{ subject to } Q^T Q = I_m.$$

Extensions of (1.1.3) were later on considered. The case when Q is an orthonormal matrix, i.e., $Q \in \mathbb{R}^{m \times n}$ with $n \leq m$ was considered in, e.g., [5]. Given A and B , it is desired to find Q such that $Y(Q) \approx B$ but now with $Q^T Q = I_n$. In [5], to measure the similarity of the two matrices Y and B , the degree of collinearity of the rows y_i^T and b_i^T in Y and B respectively is used. Hence, the solution \hat{Q} is computed from

$$\max \sum_i y_i^T b_i = \max \text{tr}(YB^T) = \max \text{tr}(AQB^T), \text{ subject to } Q^T Q = I_n, \quad (1.1.4)$$

by using the SVD $U\Sigma V^T = B^T A$, yielding $\hat{Q} = V I_{m,n} U^T$. This solution is computed in a similar manner as for an OPP. \hat{Q} is not necessary the same as the solution given when the Euclidean distance is used, to measure distances between points in Y and B . Then we get

$$\min_Q \|AQ - B\|_F^2, \text{ subject to } Q^T Q = I_n. \quad (1.1.5)$$

By using the matrix-trace, we write the objective function in (1.1.5) as

$$\begin{aligned} \|AQ - B\|_F^2 &= \text{tr}(Q^T A^T A Q) - 2\text{tr}(AQB^T) + B^T B = \\ &= \text{tr}(AQQ^T A^T) - 2\text{tr}(AQB^T) + B^T B. \end{aligned}$$

If $n = m$, then $QQ^T = I_m$ and (1.1.5) becomes an OPP whose solution is computed by maximizing $\text{tr}(AQB^T)$, just as in (1.1.4). The differences occur when $n < m$, since then $QQ^T \neq I_n$, and the term $\text{tr}(AQQ^T A^T)$ becomes dependent on Q . To illustrate this, we can look at the example in [5]. There the hypothetical factor matrices to be matched are,

$$A = \begin{bmatrix} 0.76 & 0.32 & 0.5 \\ 0.5 & 0.5 & -0.4 \\ 0.52 & -0.36 & 0.5 \\ 0.5 & -0.5 & -0.4 \end{bmatrix}, \quad B = \begin{bmatrix} 0.7 & 0.1 \\ 0.8 & 0 \\ 0.1 & 0.7 \\ 0 & 0.8 \end{bmatrix}.$$

The solution to 1.1.4 is

$$\hat{Q} = \begin{bmatrix} 0.7444 & 0.6620 \\ 0.6651 & -0.7466 \\ 0.0582 & 0.0657 \end{bmatrix}, \quad Y(\hat{Q}) = \begin{bmatrix} 0.8077 & 0.2970 \\ 0.6815 & -0.0686 \\ 0.1768 & 0.6459 \\ 0.0164 & 0.6780 \end{bmatrix},$$

while the solution to (1.1.3) is

$$\bar{Q} = \begin{bmatrix} 0.7385 & 0.6570 \\ 0.6656 & -0.7462 \\ -0.1073 & -0.1076 \end{bmatrix}, \quad Y(\bar{Q}) = \begin{bmatrix} 0.7206 & 0.2067 \\ 0.7450 & -0.0016 \\ 0.0907 & 0.5565 \\ 0.0794 & 0.7446 \end{bmatrix}.$$

The discrepancies here, for the two different solutions \hat{Q} and \bar{Q} , are $\|\hat{Q} - \bar{Q}\|_F = 0.2398$, $\|Y(\hat{Q}) - B\|_F = 0.3052$ and $\|Y(\bar{Q}) - B\|_F = 0.2119$.

It can also be desirable to weight either the columns or the rows in the residual $AQ - B$, as described above in the rigid body movement example. When weighting of the columns in $AQ - B$, we get

$$\min_Q \text{tr}(AQ - B)^T \tilde{W}^2 (AQ - B), \quad \text{subject to } Q^T Q = I_n, \quad (1.1.6)$$

and the weighting of the rows in $AQ - B$ is

$$\min_Q \text{tr}(AQ - B) \bar{W}^2 (AQ - B)^T, \quad \text{subject to } Q^T Q = I_n, \quad (1.1.7)$$

where \tilde{W} and \bar{W} are known diagonal weighting matrices, [20, 22]. If $n = m$, then (1.1.6) becomes an OPP, in a similar way to the right-weighted OPP in Section 1.1.1. Equation (1.1.7) can be written as a WOPP according to Definition 1.0.1, by taking $B := B\bar{W}$ and $X = \bar{W}$.

1.1.3 The OPP and WOPP

Areas where the WOPP (and OPP) arise are in applications related to, e.g., rigid body movement and psychometrics as mentioned, factor analysis [15, 23], multivariate analysis and multidimensional scaling [6, 13], global positioning system [2]. Typically, it is about computing a matrix with orthonormal columns, when it is desired to match one set of data to another.

As mentioned earlier, the solution to a WOPP can not be computed as easily as for an OPP. Additionally, a WOPP can have several local minima. Hence a solution, computed by some iterative method, is not necessarily a global optimum. The formulation (1.0.1) has sometimes been referred to as the *Penrose regression problem*¹.

¹It is not clear from where this originates. It seems as the first time the term "Penrose regression" was used was in an older version (technical report) of [4] from 1997. Lars Eldén informed me that Penrose studied the best approximation of the matrix equation $AXC = B$ where A , X , C and B are any general matrices [26].

If Q is an orthogonal matrix, the requirement of a specific determinant sign is sometimes requested. Typically, this is the case in applications related to rigid body movements, when $\det(Q) = 1$ is desired.

If $m = n$, i.e., Q is orthogonal, we call the optimization problem(s) *balanced* and if $n < m$ we call them *unbalanced*. Consider (1.0.1) with $X = I_n$,

$$\min_Q \|AQ - B\|_F^2, \text{ subject to } Q^T Q = I_n. \quad (1.1.8)$$

As explained in Section 1.1.2, if (1.1.8) is balanced, then it is an OPP with a solution that may be derived by using the SVD of $B^T A$. If (1.1.8) is unbalanced, we get a WOPP type problem. The name *projection Procrustes problem* has also been used to address unbalanced case of (1.1.8), e.g., [14, 15]. The formulation (1.1.8) is commonly used to address the OPP, but then assuming that Q is orthogonal ($n = m$).

Results from earlier work on iterative algorithms for solving problems similar to (1.0.1) and the unbalanced case of (1.1.8) are reported in [3, 4, 8, 9, 11, 20, 24, 28]. In [3, 4], the method to compute a solution to (1.0.1) is based on solving an ordinary differential equation. In [8], geodesics of the Stiefel manifold are used for conjugated gradient and Newton's method to solve the unbalanced case of (1.1.8). In [24], an iterative, coordinate wise type, algorithm that uses plane rotations is presented for solving (1.0.1). The work in [11] is not presented with a connection to Procrustes problems and Stiefel manifolds as the other articles. But, one of the problems considered can be regarded as (1.1.8) with $n = 1$. [11] contains an algorithm to solve this optimization problem, based on a Lagrangian formulation.

To get an analytical expression of a solution to the WOPP, the optimization problem is sometimes modified. As in [22], the constraint $Q^T \bar{W}^2 Q = I_n$ is imposed (instead of $Q^T Q = I_n$) when studying (1.1.7). This results in that an analytic solution can be derived, in a similar procedure as when dealing with an OPP.

1.1.4 Canonical form of a WOPP

For a WOPP, the matrices A and X can always be considered as (redefined as) diagonal matrices

$$A = \text{diag}(\alpha_1, \dots, \alpha_m), \quad \alpha_i \geq \alpha_{i+1} > 0, \quad i = 1, \dots, m-1, \quad (1.1.9)$$

and

$$X = \text{diag}(\chi_1, \dots, \chi_n), \quad \chi_j \geq \chi_{j+1} > 0, \quad j = 1, \dots, n-1, \quad (1.1.10)$$

respectively. We call this the *canonical form* of a WOPP. The diagonal elements α_i and χ_i are merely the corresponding singular values of the "original" matrices A and X , e.g., see Paper III. Evidently it is preferred to make this diagonalization. Not only does it make the analysis of the problem easier, it also reduces the computational cost for algorithms used to solve the problem. From now on, we assume that A and X are diagonal matrices.

1.2 Linear Least squares problems on the Stiefel manifold

In this thesis, we also consider optimization problems on the form,

$$\min_Q \|f(Q) - b\|_2^2, \text{ subject to } Q \in \mathcal{V}_{m,n}, \quad (1.2.1)$$

where $f(Q) \in \mathbb{R}^k$ is a vector valued function of Q and $b \in \mathbb{R}^k$ a known vector. We restrict ourselves to the cases when $f(Q)$ is linear in Q .

Another way of writing (1.2.1) is

$$\min_Q \|f(Q) - b\|_2^2 \quad (1.2.2)$$

$$\text{subject to } q_i^T q_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{otherwise.} \end{cases} \quad (1.2.3)$$

There are two types of constraints for this problem. The orthogonality constraint that $q_i \perp q_j$ whenever $i \neq j$ and the normalizing constraint, $\|q_i\| = 1$ for all i .

The algorithms presented in Paper **I** and Paper **II** are based on the formulation according to (1.2.1). Those can be used to compute a solution to a WOPP. To write a WOPP on the form given in (1.2.1), we make use of the *Kronecker product* \otimes and the *vec-operator*. The (i, j) block of the Kronecker product $X \otimes A$ of two matrices X and A is $x_{i,j}A$. $\text{vec}(Q)$ is a stacking of the columns in $Q = [q_1, \dots, q_n] \in \mathbb{R}^{m \times n}$ into a vector

$$\text{vec}(Q) = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} \in \mathbb{R}^{mn}.$$

Let $Y(Q) = AQX$, then by using the *vec-operator* on Y and B we get

$$f(Q) = \text{vec}(Y) = [X^T \otimes A] \text{vec}(Q) \quad b = \text{vec}(B).$$

The function $Y(Q) \in \mathbb{R}^{m \times n}$ is now embedded in \mathbb{R}^{mn} .

Generally, any matrix function $Y(Q)$ that is linear in Q , can be expressed as a vector valued function $f(Q) = \text{vec}(Y(Q)) = F \text{vec}(Q)$ where $F \in \mathbb{R}^{k \times mn}$.

Chapter 2

Algorithms for least squares problems on the Stiefel manifold

2.1 The 3-dimensional WOPP, Paper I

In Paper I, we present an algorithm to solve the 3-dimensional WOPP (1.1.2). As a parametrization of M , the *Cayley transform* $C(S)$ of a skew-symmetric matrix $S = -S^T \in \mathbb{R}^{3 \times 3}$ is used,

$$C(S) = (I + S)(I - S)^{-1}.$$

The algorithm uses Newton or Gauss-Newton search directions and due to the geometry of the problem, optimal step lengths can be computed very simply. The weighted case of (1.1.1) has also been specially studied by others [1].

A poster of this work was presented at the *First SIAM-EMS Conference "AMCW" 2001*, Berlin, September 2-6, 2001.

2.2 Linear least squares problems on the Stiefel manifold, Paper II

In Paper II, we consider the least squares problem

$$\min_Q \frac{1}{2} \|f(Q) - b\|_2^2, \quad \text{subject to } Q \in \mathcal{V}_{m,n}, \quad (2.2.1)$$

where $f(Q) \in \mathbb{R}^k$ can be written as $f(Q) = F \text{vec}(Q)$ with $F \in \mathbb{R}^{k \times mn}$ and $\text{rank}(F) = \min(k, mn)$. There are some requirements on the matrix F , though. Suppose Q is parameterized with p parameters, then if $k < p$ the optimization problem is under-determined, in fact the Jacobian of $f(Q)$ will not have full

column rank. Even if $k \geq p$ it is not guaranteed that (2.2.1) is not under-determined. We illustrate this with a small example. Take $Q = [q_1, q_2, q_3] \in \mathbb{R}^{3 \times 3}$ then $p = 3$ is needed to represent Q . But with $F = [I_3, Z, Z] \in \mathbb{R}^{3 \times 9}$ where $Z \in \mathbb{R}^{3 \times 3}$ is a zero matrix, we see that $f(Q)$ is then independent of q_2 and q_3 due to multiplication with zeros, i.e.,

$$f(Q) = \begin{bmatrix} I_3 & Z & Z \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = q_1 + Zq_2 + Zq_3 = q_1.$$

Hence it is necessary that F corresponds to a sufficiently amount of data such that (2.2.1) is well-posed. The algorithm in Paper **II** can be used to some extent to solve some undetermined problems, but it is not developed to handle them. The best way to deal with an under-determined (rank-deficient, ill-posed) problem is to do a reformulation, if possible. As for the example given, it would be better to reformulate the problem with $Q \in \mathbb{R}^{3 \times 1}$ and $F = \text{diag}(1, 1, 1) \in \mathbb{R}^{3 \times 3}$.

The algorithm in Paper **II** started out as a generalization of the algorithm in Paper **I**. Hence the Cayley transform was used to parameterize $Q \in \mathbb{R}^{m \times n}$. Since the Cayley transform only works for orthogonal matrices, i.e., when $m = n$, a slight modification of it was needed to comply with the unbalanced cases when $n < m$. Also this algorithm uses Newton or Gauss-Newton methods to get a descent direction. Optimal step lengths could not be computed as easily as in Paper **I**. However, later on it was found out that by using the *matrix exponential* of a skew-symmetric matrix S , $\exp(S)$, instead of the Cayley transform to parameterize $Q \in \mathbb{R}^{m \times n}$, optimal step lengths could be computed rather simply. The choice of $\exp(S)$ results in a similar algorithm as when using the Cayley transform.

Parts of this work was presented at the *18th International Symposium on Mathematical Programming* (ISMP), Copenhagen, August 18-22, 2003.

Chapter 3

Global minimization of a WOPP

As mentioned, a WOPP can have several minima. The task of finding the "best" minimizer is a *global optimization* problem. In order to know if a computed minimizer is a global minimizer, a *sufficient condition* for global optimum is desired. That is, a condition that is true for a global minimizer but false for any local minimum. Deriving such a condition is not an easy task. In [9], a *necessary condition* for global optimum is presented. If this condition fails for a computed solution \hat{Q} , then \hat{Q} is a local minimum. If the necessary condition is true for \hat{Q} , then \hat{Q} can either be a local or global minimum.

Another way to classify if a minimizer \hat{Q} is a global optimum, is to compute all minima to the problem and then check which of those minima that yields the least objective function value. In order to do so it is useful to know how many minima the problem might have. The studies done in Paper **III** and Paper **IV**, presented below, leads to the following conjecture.

Conjecture 3.0.1 *The weighted orthogonal Procrustes problem*

$$\min_Q \|AQX - B\|_F^2, \text{ subject to } Q^T Q = I_n,$$

has at most 2^n unconnected minima.

By unconnected minima, we mean that the minima are distinct. For some special cases, there can be a continuum of minimizers. This can be illustrated by considering the ellipse example shown earlier. Let $A = I_2$ and let $B = 0$ (the origin), then any $Q \in \mathcal{V}_{2,1}$ is a minimum.

3.1 The number of minima to a WOPP, Paper III

Paper **III** contains a study on the number of minima to a WOPP. As a simple example of a WOPP with more than one minimizer, consider the ellipse problem mentioned earlier. As seen in Figure 1 a local minimum occurs in the fourth quadrant. What determines if the optimization problem has one or two minima is the flatness of ellipse along with the magnitude and direction of b . For a very flat ellipse where $\alpha_1 \gg \alpha_2$ it is more likely that a local minimum can exist as opposed to an "almost circular" ellipse with $\alpha_1 \approx \alpha_2$. If $\alpha_1 = \alpha_2$ then $f(Q)$ is a circle and only one minimizer $\hat{Q} = b/\|b\|$ exist if $b \neq 0$. Consider now the

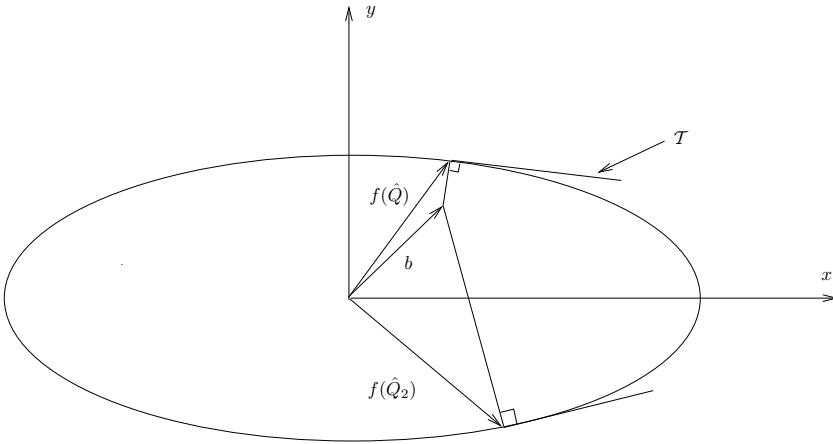


Figure 1: Two minima are found where the tangent of $f(Q)$ is orthogonal to the distance vector from b to the ellipse (the residual $r = b - f(Q)$). The minimum in the first quadrant is global.

case when $b = 0$. If $\alpha_1 = \alpha_2$ any $Q \in \mathcal{V}_{2,1}$ is a minimizer, i.e., a continuum of solutions arises. Roughly speaking we can say that we consider a continuum of solutions as one minimizer. However, if $\alpha_1 > \alpha_2$ the WOPP always has two distinct, *unconnected*, minima at $\hat{Q} = [0, \pm 1]^T$. This reasoning applies to any WOPP with $Q \in \mathbb{R}^{m \times 1}$, which we call the *ellipsoid cases* since the surface of $f(Q) = AQX$ is a hyper ellipsoid in \mathbb{R}^m . In this instance there can at most be two unconnected minima to the WOPP for any $b \in \mathbb{R}^m$, see Paper **III**.

The global minimum \hat{Q} to an ellipsoid case must fulfill the condition

$$\text{sign}(f_i(\hat{Q})) = \text{sign}(b_i) \quad \forall \quad i = 1, \dots, m.$$

That is, $f(\hat{Q})$ must lie in the same area as b , that is divided by the coordinate planes. For example consider $Q \in \mathbb{R}^2$, then $f(\hat{Q})$ and b must lie in the same quadrant and for $Q \in \mathbb{R}^3$ they must be in the same octant and so on.

For a WOPP of general dimension the geometry becomes more complicated than in the ellipsoid cases. Nevertheless, the surface of $f(Q)$ still have elliptic properties. For instance we can move from one point $f(\hat{Q})$ to another point $f(\tilde{Q})$ by following ellipses on the surface of $f(Q)$. When studying the maximum number of minima to a WOPP, using $b = 0$ ($B = 0$) is a naturally first approach. Also it is preferred to consider that $\alpha_i > \alpha_{i+1}$ and $\chi_j > \chi_{j+1}$ for all $i = 1, 2, \dots, m-1$ and $j = 1, 2, \dots, n-1$. Compare this to the task of determining the number of eigenvectors to a matrix $E \in \mathbb{R}^{m \times m}$. If we consider, e.g., the identity matrix $E = I \in \mathbb{R}^{m \times m}$, any $x \in \mathbb{R}^m$ with $\|x\| = 1$ is an eigenvector. But if E is a diagonal matrix according to $E = \text{diag}(\epsilon_1, \dots, \epsilon_m)$ where $\epsilon_i \neq \epsilon_j$ for all $i \neq j$, then the number of eigenvectors of E are finite.

However, in Paper **III** it is shown that a WOPP with $Q \in \mathbb{R}^{m \times n}$ has 2^n minima when $B = 0$. Empirical studies indicate that this is a valid upper bound for the maximal number of minima to a WOPP.

3.2 Computing all minimizers, Paper IV

Consider a very flat ellipse with $\alpha_1 \gg \alpha_2$ according to Figure 2, with \hat{b} lying in the first quadrant. The solution \hat{Q} is then also in the first quadrant. Now let $b = \hat{b} + \delta b$ be a perturbed measurement, as shown. Due to the flatness of the ellipse, the global minimum now is in the fourth quadrant. In this case the local minimum is a better approximation of the "correct" solution \hat{Q} , than the global minimum. This can also occur for higher dimensional problems. Hence computing all, or some, solutions to a WOPP could be to prefer.

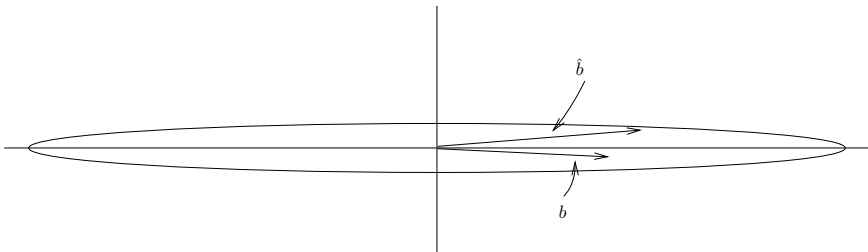


Figure 2:

In Paper **IV**, an algorithm to compute all minima to a WOPP is presented. To explain how this algorithm works we can again take the ellipsoid cases. Assume that the local minimum \hat{Q}_2 in Figure 1 is computed. To get a \tilde{Q} that is in the vicinity of the global minimum \hat{Q} we can use the normal N of $f(Q)$ at $Q = \hat{Q}_2$. The residual $r = b - f(\hat{Q}_2)$ coincides with the normal direction. Consider the function $f(\hat{Q}) + N\gamma$ where γ is a scalar and N the normal at $f(\hat{Q})$. Computing the intersection of $f(\hat{Q}) + N\gamma$ and the ellipse yields a Q_2 in the vicinity of \hat{Q} . Now Q_2 is a good initial value for an iterative method to compute

\hat{Q} . This method is roughly the same for any ellipsoid case since the normal is uniquely defined.

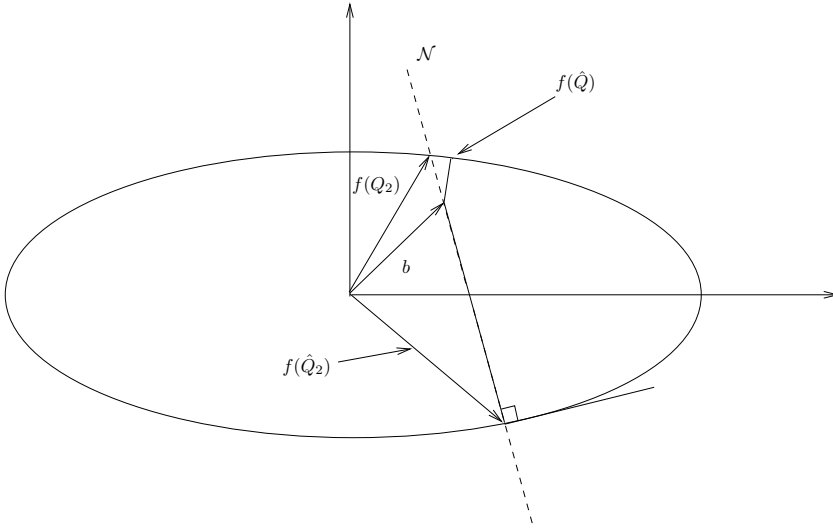


Figure 3: The normal plane \mathcal{N} (dashed line) at $f(\hat{Q}_2)$ intersects the surface of $f(Q)$ at $f(Q_2)$, in the vicinity of $f(\hat{Q})$.

From special studies of $Q \in \mathbb{R}^{2 \times 2}$ and $Q \in \mathbb{R}^{3 \times 2}$ it seems as if this normal plane method is a good method to use for computing all minimizers. Since the surface of $f(Q)$ is "built up by ellipses", intuitively this method should be viable for a WOPP of a general dimension. In these cases, computing the intersections of the normal plane and the surface of $f(Q)$ is done by computing all solutions to a set of quadratic equations. These equations can be formulated as a *continuous algebraic Riccati equation* (CARE), a well known quadratic matrix equation, [21, 27]. This equation has 2^n roots, the same number as the estimated maximal number of minima to a WOPP. Empirical studies shows that this method manages to compute all, or at least several, minima to the WOPPs considered.

A presentation of this work was held at the *18th International Symposium on Mathematical Programming* (ISMP), Copenhagen, August 18-22, 2003.

Chapter 4

Quadratic equations

In Paper **V**, an iteration method for solving $F(x) = 0$ where $F(x) \in \mathbb{R}^m$ and $x \in \mathbb{R}^m$ is presented. It exhibits cubic convergence by using second order information (derivatives) in each step. Other methods using second order information are the *Chebyshev method* [19], and *Halley's method* [25]. The implementation of these in \mathbb{R}^m for $m > 1$ from a "practical linear algebraic" point of view, does not seem easy. As an example, the Chebyshev method (in Banach spaces) is often presented as in [19],

$$x_{k+1} = x_k - (I + \frac{1}{2}F'(x_k)^{-1}F''(x_k)F'(x_k)^{-1}F(x_k))F'(x_k)^{-1}F(x_k).$$

How is the second order derivative $F''(x)$ (usually represented as a tensor) multiplied with the inverse of the Jacobian $F'(x_k)^{-1}$? Presentations of these methods seem too abstract. It is suspected that mathematicians are not interested in how the multiplication is performed, they are satisfied with the knowledge of that it is possible. In [7], a straightforward and practical presentation of Halley's method in \mathbb{R}^m is given. Paper **V** also contains the writers interpretation of Halley's method in several variables.

However, to use higher order information usually is computationally heavy. But for a quadratic problem this is not always the case. The second order derivatives of quadratic equations are constant, hence the computational cost will not grow that much. A small experimental study of this is done in Paper **V**, indicating an increased efficiency for quadratic problems with less than 15 parameters.

Finally, to prove cubic convergence for the method presented in Paper **V** when $m > 1$, a rather cumbersome and messy tensor arithmetic was invented and used. The paper [31] contains a short informal note about this tensor arithmetic.

Chapter 5

Ill-posed problems

An inverse problem is the task of, e.g., determining some parameters x in a model (function) $f(x)$ by using some observed data b where $b \approx f(x)$. Typically we can write a solution \hat{x} as $\hat{x} = f^{-1}(b)$, if the problem is well-posed. The definition of well-posedness was set up by Hadamard in the beginning of the 20th century as [10, 17],

- a) For all admissible data, a solution exist.
- b) For all admissible data, the solution is unique.
- c) The solution depends continuously on the data.

If any of the above properties does not hold, the problem is said to be ill-posed. In connection to c), are *ill-conditioned* problems. A problem is called ill-conditioned if a small perturbation in b yields a large perturbation of the solution \hat{x} .

Consider a nonlinear optimization problems as, e.g.,

$$\min_x \|f(x) - b\|_2^2, \quad (5.0.1)$$

where $f(x) \in \mathbb{R}^m$ is a nonlinear function of the parameters $x \in \mathbb{R}^n$ to be determined, and $b \in \mathbb{R}^m$ corresponds to some input data (measurements). Here (5.0.1) is ill-conditioned in the sense that any small perturbation Δb in b may result in a large perturbation of the solution \hat{x} of (5.0.1).

A commonly used approach for solving these types of optimization problems is *Tikhonov regularization*

$$\min_x \|f(x) - b\|_2^2 + \lambda \|L(x - x_c)\|_2^2, \quad (5.0.2)$$

where $\lambda > 0$ is called the *regularization parameter*, x_c the center of regularization and L a known weighting matrix. For simplicity, consider L as the identity matrix, then the regularized problem (5.0.2) corresponds to determining a $\hat{\lambda}$ and a solution $\hat{x}(\hat{\lambda})$ such that $\|\hat{x} - x_c\|_2$ is not large and $\|f(\hat{x}) - b\|_2$ is somewhat

optimal. How should this be done ? Given some x_c and a priori knowledge about the magnitude of the noise level, e.g., $\|\Delta b\|_2 \leq \delta_b$, then it would be an idea to formulate (5.0.2) as

$$\min_x \|x - x_c\|_2^2, \text{ subject to } \|f(x) - b\|_2^2 = \delta_b. \quad (5.0.3)$$

This is known as the *discrepancy principle*. We could also assume that for a given x_c , a solution \hat{x} should fulfill $\|\hat{x} - x_c\|_2 \leq \delta_x$ where δ_x is known. Then a solution could be computed by solving

$$\min_x \|f(x) - b\|_2^2, \text{ subject to } \|x - x_c\|_2^2 \leq \delta_x. \quad (5.0.4)$$

5.1 The L-curve for nonlinear problems, Paper VI

How should (5.0.1) or (5.0.2) be solved without having any a priori knowledge ? A quite popular method for linear least squares problems is the L-curve method [17]. The L-curve is given by plotting the solution norm $\|x(\lambda) - x_c\|$ as a function of the residual $\|f(x(\lambda)) - b\|$, in a log-log scale, for different values of λ . The idea is to pick a solution "in the corner" of the L-curve, since it is there that the solution norm starts to grow drastically.

In Paper VI a rather specialized investigation of the L-curve method in connection to nonlinear problems is done. Nonlinear problems can be extremely different from each other as opposed to linear problems. Hence the L-curves often vary in shape and similarities to the L-curve for linear problems can be uncommon.

Additionally, the center of regularization x_c plays a bigger role for nonlinear problems. Without any additional information, it can be hard to motivate that a solution \hat{x} is better than x_c itself just because \hat{x} yields a smaller residual norm.

Chapter 6

Software

6.1 WOPP software

The algorithms presented in Paper **I**, Paper **II** and Paper **IV** have been implemented in MATLAB. Though the algorithm in Paper **II** also manages the 3-dimensional case described in Paper **I**, a special routine for algorithm in Paper **I** is available. Also included are routines using the Cayley transform parametrization, instead of the matrix exponential function.

Details regarding the software can be found at

<http://www.cs.umu.se/~viklands/WOPP/index.html>.

6.2 L-curve toolbox

Initially our goal was to develop a toolbox for Tikhonov regularization of nonlinear optimization problems. In a similar manner as Hansen [17] did for linear problems. Due to the difficulties that arise with nonlinearity (convergence aspects, how to choose and update regularization parameters, time consuming computations, etc.), this turned out to be a difficult task. The black box type algorithm presented in Paper **IV** can and probably will fail, if the nonlinear function $f(x)$ is replaced by some other "general" function (coming from a different application). Therefore no toolbox software has been made public, yet.

Chapter 7

Research biography and reflections

My first years of research (1999-2001) were focused on algorithms for solving ill-posed nonlinear optimization problems, by using Tikhonov regularization (5.0.2) and the L-curve. Since no specific, real, application was considered, this gave rise to some problems. We mostly regarded $f(x) \in \mathbb{R}^m$ as an ill-conditioned function of the parameters $x \in \mathbb{R}^n$, and the weighting matrix L as the identity matrix, i.e.,

$$\min_x \|f(x) - b\|_2^2 + \lambda \|x - x_c\|_2^2. \quad (7.0.1)$$

Here we get something like a parameter estimation problem, which I now think is a bad approach. I figure it would be better to consider problems (applications) where x is a function instead, e.g., $x(t)$ where $t \in \mathbb{R}$. Then applying a smoothness condition on $x(t)$ by using a weighting matrix L .

Also, how should the center of regularization x_c be treated? Let us consider two choices of how to treat x_c .

1. x_c is a fix (known) point, that should not be changed as, e.g., a priori information.
2. x_c is just an arbitrary point (initial value) "in the vicinity" of the solution. Roughly speaking, it's not that important.

Our approach was to consider x_c as in item 2, and then as I see it, a small dilemma turns up. By using the L-curve method, suppose we compute a solution \hat{x} with some not-so-important x_c . Is \hat{x} really a better solution than x_c itself, just because it gives a smaller residual norm? Seems hard to state that, unless some additional information is provided.

If we now state that \hat{x} is a good solution to the problem, then by taking $x_c = \hat{x}$ and solving the problem again with the L-curve, we get a new solution \bar{x} . The new solution \bar{x} is not necessarily the same as the first computed solution \hat{x} . This does not feel right. In my opinion, having computed a solution \hat{x} by some

method with an arbitrary x_c , and then using $x_c = \hat{x}$ and solving the problem again should result in that the computed solution still is \hat{x} (if we disregard problems arising due to finite arithmetic). This is where the L-curve method fails, but the discrepancy principle does not.

However, I do not think the L-curve (or other heuristic methods) to solve nonlinear ill-posed problems are a waste, but I do believe a specific real application is needed. The test problems considered as, e.g., NMR spectroscopy and heat transfer equations, were purely artificial. Doing research with a "general" nonlinear function $f(x)$, i.e., having a black box approach, seemed rather hopeless at that time. Even though the area of ill-posed problems is more popular and huge compared to the area of WOPP, I think it was a good decision that I left it. In my opinion, research on deriving good or reasonable solutions to an ill-posed problem is more connected to statistics than developing optimization algorithms.

The first time I came in contact with a WOPP was during a course in optimization related to rigid body movement, held by my supervisor Per-Åke Wedin in the year 2000. Having only some basic knowledge about the common OPP, I became interested in the fact that a WOPP can have several minima while an OPP has an unique minimizer (if BX^T is nonsingular). Slowly my research became less focused on ill-posed problems, and instead turned towards the WOPP.

When we developed the algorithms in described Paper **I** and Paper **II**, I made a lot numerical tests. It was noted that there were some pattern between the different minima to a WOPP. For instance, if \hat{Q}_1 and \hat{Q}_2 are two minima, they can look quite similar apart from some + or – signs on different elements. By studying some special and low-dimensional cases, I found out that using the normal plane, to compute additional minima, seemed as a good method. The result was the normal plane algorithm presented in Paper **IV**.

Before the connection to the CARE was discovered in Paper **IV**, a heuristic method was used to compute all solutions to the quadratic equations (computation of the normal plane intersections). It was early noted that there always seemed to be 2^n solutions. This was later seen to be the exact number of solutions, when the CARE formulation was used. However, to compute all 2^n solutions *Newton's method* with random initial values was used until all solutions were found. To speed this up, a new iteration method was considered Paper **V**. In connection to this work, I became interested in tensor representations of higher order derivatives. But it felt as if I was too far away from what I was supposed to be working with. Hence only a small note about the subject was done [31], and I turned back to the WOPP.

By empirical studies I noted that the maximal number of minima to a WOPP seemed to be given by the formula 2^n . Also if B was small, more minima likely occurred. This resulted in a special study of the case when $B = 0$ in Paper **III**. Initially when studying the WOPP with $B = 0$, the Cayley transform parametrization was used. This resulted in rather long and badly arranged equations. In [9] Eldén and Park uses the Lagrangian formulation of the WOPP, which inspired me. By using that formulation, the equations become more

foreseeable, but the proofs in Paper **III** can presumably be done different and simpler.

The WOPP is not a "well used", so to speak, optimization problem as the OPP. Very little has been published about its real world applications. Is it due to that since the OPP is very easy to solve, people do not want to complicate the problem by turning it into a WOPP, even though a weighting or such is desirable ? I do not know. However, I hope that some of this work can be useful for present and future persons, wishing to solve and do research about the WOPP and least squares problems defined on Stiefel manifolds.

References

- [1] P. G. Batchelor and J. M. Fitzpatrick. A Study of the Anisotropically Weighted Procrustes Problem. *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '00)*, page 212, 2000.
- [2] T. Bell. Global Positioning System-Based Attitude Determination and the Orthogonal Procrustes Problem. *Journal of Guidance, Control, and Dynamics*, 26(5):820–822, 2003.
- [3] M. T. Chu and N. T. Trendafilov. On a Differential Equation Approach to the Weighted Orthogonal Procrustes Problem. *Statistics and Computing*, 8(2):125–133, 1998.
- [4] M. T. Chu and N. T. Trendafilov. The Orthogonally Constrained Regression Revisted. *J. Comput. Graph. Stat.*, 10:746–771, 2001.
- [5] N. Cliff. Orthogonal rotation to congruence. *Psychometrika*, 31(1):33–42, 1966.
- [6] T. F. Cox and M. A. A. Cox. *Multidimensional scaling*. Chapman & Hall, 1994.
- [7] A. A. M. Cuyt and L. B. Rall. Computational Implementation of the Multivariate Halley Method for Solving Nonlinear Systems of Equations. *ACM Transactions on Mathematical Software*, 11(1):20–36, 1985.
- [8] A. Edelman, T. A. Arias, and S. T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.
- [9] L. Eldén and H. Park. A Procrustes problem on the Stiefel manifold. *Numer. Math.*, 82(4):599–619, 1999.
- [10] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.
- [11] W. Gander. Least Squares with a Quadratic Constraint. *Numer. Math.*, 36:291–307, 1981.

- [12] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1989.
- [13] J. C. Gower. Multivariate Analysis: Ordination, Multidimensional Scaling and Allied Topics. *Handbook of Applicable Mathematics*, VI:Statistics(B), 1984.
- [14] J. C. Gower. Orthogonal and projection procrustes analysis. In W. J. Krzanowski, editor, *Recent Advances in Descriptive multivariate analysis*. Oxford University Press, Oxford, 1995.
- [15] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*. Oxford University Press, 2004.
- [16] B. F. Green. The orthogonal approximation of an oblique simple structure in factor analysis. *Psychometrika*, 17:429–440, 1952.
- [17] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, 1998.
- [18] J. R. Hurley and R. B. Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioural Science*, 6:258–262, 1962.
- [19] M. A. Hernández J. M. Gutiérrez. New Recurrence Relations for Chebyshev method. *Appl. Math. Lett.*, 10(2):63–65, 1997.
- [20] M. A. Koschat and D. F. Swayne. A Weighed Procrustes Criterion. *Psychometrika*, 56(2):229–239, 1991.
- [21] P. Lancaster and L. Rodman. *The Algebraic Riccati Equation*. Oxford University Press, 1995.
- [22] R. W. Lissitz, P. H. Schönemann, and J. C. Lingo. A solution to the weighted procrustes problem in which the transformation is in agreement with the loss function. *Psychometrika*, 41(4):547–550, 1976.
- [23] W. Meredith. On weighted procrustes and hyperplane fitting in factor analytic rotation. *Psychometrika*, 42(4):491–522, 1977.
- [24] A. Mooijaart and J. J. F. Commandeur. A General Solution of the Weighed Orthonormal Procrustes Problem. *Psychometrika*, 55(4):657–663, 1990.
- [25] Ortega and Rheinholdt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, 1970.
- [26] R. Penrose. A Generalized Inverse for Matrices. *Proc. Cambridge Philos. Soc.*, 51:406–413, 1955.
- [27] J. E. Potter. Matrix Quadratic Solutions. *J. SIAM Appl. Math.*, 14(3):496–501, 1966.

- [28] T. Rapcsak. On Minimization on Stiefel manifolds. *European J. Oper. Res.*, 143(2):365–376, 2002.
- [29] I. Söderkvist and Per-Åke Wedin. On Condition Numbers and Algorithms for Determining a Rigid Body Movement. *BIT*, 34:424–436, 1994.
- [30] E. Stiefel. Richtungsfelder und Fernparallelismus in n -dimensionalen Mannigfaltigkeiten. *Commentarii Math. Helvetici*, 8:305–353, 1935-1936.
- [31] T. Viklands. A Note on Representations of Derivative Tensors of Vector Valued Functions. Technical Report UMINF-05.14, Department of Computing Science, Umeå University, Umeå, Sweden, 2005.