Total Frame Potential and its Applications in Data Clustering

Dissertation

zur Erlangung des akademischen Grades

eines Doktors der Naturwissenschaften

Der Fakultät für Mathematik

der Technischen Universität Dortmund

vorgelegt von

Tobias Springer

im Jahr 2013

Vorsitzender der Prüfungskommission:	Prof. Dr. Rainer Brück
Erstgutachter:	Prof. Dr. Joachim Stöckler
Zweitgutachterin:	Prof. Dr. Katja Ickstadt
Dritter Prüfer:	Prof. Dr. Christoph Buchheim
Wissenschaftlicher Mitarbeiter:	Dr. Thorsten Camps

Datum des Prüfungskolloquiums: 26.11.2013

Abstract

For the statistical analysis of microarray gene expression data, the clustering of short time series is an important objective in order to identify subsets of genes sharing a temporal expression pattern. An established method, the Short Time Series Expression Miner (STEM) by Ernst et al. ([Erns 05]), assigns time series data to the closest of suitably selected prototypes followed by the selection of significant clusters and eventual grouping. This algorithm identifies each time series by a corresponding vector in $\mathbb{R}^d$ which contains the data expressions at $d \in \mathbb{N}$ not necessarily equidistantly distributed points in time. In order to qualify for the term "short" time series, the number d is supposed to be small, e.g. $d \leq 12$.

For the clustering of normalized d-dimensional data $Y = \{y_j\}_{j=1,\ldots,N}$ we propose to minimize the Penalized Frame Potential

$$\mathrm{F}_\alpha(\Theta, Y) \quad = \quad \mathrm{TFP}(\Theta) \, - \alpha \sum_{\ell=1}^{m} \max_{j=1,\ldots,N} \langle y_j, \theta_\ell \rangle \tag{1}$$

on the m-fold unit sphere for the regularization parameter $\alpha \geq 0$. The functional contains the "Total Frame Potential" (TFP) whose minimizers are exactly the Finite Unit Norm Tight Frames (FUNTFs), see Benedetto and Fickus ([Bene 03]), and includes a data-driven component for the selection of prototypes. We show that the solution of the corresponding constrained optimization problem is naturally connected to the spherical Dirichlet cells

$$D_j \quad = \quad \left\{ v \in \mathbb{R}^d \, : \, \|v\|_2 = 1, \, y_j = \operatorname*{arg\,max}_{1 \leq k \leq N} \langle \, y_k, v \, \rangle \right\}$$

of the given normalized data. Furthermore, the minimizers of F_α are, given that $\alpha > 0$, in the interior of the Dirichlet cells and the objective function F_α is differentiable in the minimum with the extremal condition

$$4TT^*T + 2T\Lambda \quad = \quad \alpha Y_s$$

where T, $Y_s \in \mathbb{R}^d$ have normalized columns and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ contains the Lagrange multipliers from a corresponding constrained minimization problem.

The general problem is closely related to the search for point configurations on the unit sphere like in Tammes' ([Tamm 30]) or Thomson's Problem ([Thom 04]). Moreover, the minimization

of (1) (subject to the constraint that the solution is normalized) contains connections to problems in matrix completion (see e.g. Candès and Tao in [Cand 10] or Mazumder, Hastie and Tibshirani in [Mazu 10]).

The idea of using the frame potential in combination with a data-dependent term for optimization was originally proposed by Benedetto, Czaja and Ehler ([Bene 10]) for finding sparse coefficient representations. First results of our proposed method were published by Springer, Ickstadt and Stöckler ([Spri 11]).

The thesis presents the motivation of our approach by introducing the STEM algorithm for data clustering and outlining the connection to a proposal in [Bene 10]. We give an overview over the development in the theory of Finite Unit Norm Tight Frames. Moreover, we analyze the features of the Penalized Frame Potential and illustrate relations to other well-known optimization problems in the theory of Compressive Sensing. Finally, we present numerical results on the implementation of the functional by application on real and simulated data.

Acknowledgements

First and foremost, I would like to give a few words to the people who have contributed to this thesis in different ways. I am indebted to Prof. Dr. Joachim Stöckler who was an encouraging and motivating advisor. The numerous meetings were a great source of inspiration and almost always led to new insights. I am also grateful to Prof. Dr. Katja Ickstadt for supporting the ideas of implementing new methods and for agreeing to be a reviewer for this thesis.

Another important factor was the team at the *Lehrstuhl für Approximationstheorie* – especially PD Dr. Maria Charina, Tobias Kloos and Dr. Katrin Siemko – who created a very pleasant working atmosphere during the last four years. Included is our secretary Christine Mecke for the morning coffee and the help on administrative issues.

Since my family has always been helpful during the last years, I also thank my parents Susanne and Detlef Springer who always encouraged me in various ways and Alina Stöteknuel for being constantly supportive in all aspects. Finally, the help of my good friends Hendrik Blom and Arne Hauner in preparation for my thesis defense should not go unmentioned.

Contents

Chapter 1

Introduction

In a variety of fields, such as biology, economy or social sciences, time series are necessary to express characteristic features of underlying processes over time. For example, in the analysis of microarray gene expression data, the clustering of time series is an important objective in order to identify subsets of genes sharing a temporal expression pattern (see Figure 1.1). According to Ernst et al. ([Erns 05]), more than 80% of the time series in the Stanford Microarray Database consist of the values measured at eight time points or less. That leads to a large number of data in a low-dimensional space ([Spri 11]).

Since most methods for analyzing long time series are not well-suited or not even applicable for short time series, different approaches and algorithms have to be developed. Many established methods for the analysis of short time series consider the behavior of biological data only in the phase of the modeling of cluster prototypes. In [Spri 11], Springer, Ickstadt and Stöckler proposed a new method based on the minimization of the non-convex functional

$$\mathrm{F}_\alpha(\Theta, Y) \;=\; \frac{d}{m^2}\mathrm{TFP}(\Theta) + \alpha \left(m + 1 - \sum_{\ell=1}^{m} \max_{j=1,\dots,N} \langle y_j, \theta_\ell \rangle \right), \tag{1.1}$$

which also takes the actual (normalized) data $Y = \{y_j\}_{j=1,\dots,N}$ into account. It combines the "Total Frame Potential" (TFP) from [Bene 03] with a data-dependent penalty term. This technique of obtaining a tradeoff between regularization and minimizing cost imposed by a loss function is common in Statistics and Machine Learning Theory.

In this thesis, we analyze this functional on a mathematical basis, including an introduction

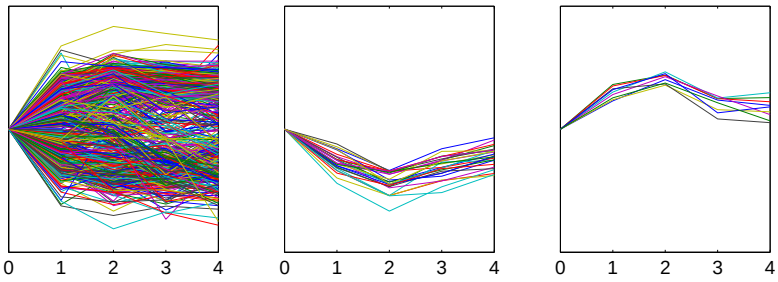


Figure 1.1: Sample data (left) and two groups of included short time series sharing similar expression patterns (middle and right)

into the necessary framework, and discuss the position of our method in the family of cluster algorithms as well as the relation of the inherent optimization to other problems in learning theory. Our central Theorem (Theorem 4.8) shows that for a positive regularization parameter α the minimizing family $\theta_1, \ldots, \theta_m$ of vectors on the unit sphere cannot be located on the spherical boundaries of the data-generated Dirichlet cells

$$D_j \quad = \quad \{v \in \mathbb{R}^d \,:\, \|v\|_2 = 1,\, \langle y_j, v\rangle = \max_{k=1,\ldots,N} \langle y_k, v\rangle\}\,.$$

Then it follows immediately that for each θ_ℓ there exists a unique $y_{s(\ell)}$ such that

$$\max_{j=1,\ldots,N} \langle y_j, \theta_\ell\rangle \quad = \quad \langle y_{s(\ell)}, \theta_\ell\rangle$$

holds in (1.1). This feature is the basis for a proposal of a group of related minimization problems which lead to further features of minimizers of the stated functional.

The outline is as follows. In Chapter 2, we introduce the basic theory of frames including the recent development on finite frames. We cite major results by Benedetto and Fickus ([Bene 03]) and Goyal et al. ([Goya 98]). The Total Frame Potential from [Bene 03] is considered from a linear algebra perspective and as an optimization problem with quadratic constraints using Lagrange multipliers. As will be shown, the objective function can be formulated in the eigenvalues of a Gramian matrix leading to a polynomial problem of total degree four. Furthermore, we extend existing results on the minimizers of the TFP to all extrema and show that every local maximum is also global (Theorem 2.14).

Chapter 3 gives a short overview on cluster algorithms in general. The focus lies on the so-called STEM algorithm by Ernst et al. ([Erns 05]), which contains connections to optimization

on unit spheres. Furthermore, a brief discussion on an inherent relation to classical problems by Tammes ([Tamm 30]) and Thomson ([Thom 04]) arising in biology and physics, respectively, is included. The basic idea is to generalize an approach by Benedetto, Czaja and Ehler from [Bene 10] in order to motivate the construction of the Penalized Frame Potential as a data-dependent version of the TFP whose minimizers serve as cluster centers (prototypes).

In Chapter 4 we analyze the behavior of the Penalized Frame Potential and extract simple characteristic features. Moreover, we characterize the minimizers in terms of Dirichlet cells of a certain subfamily of the underlying data on a unit sphere. This leads us to introduce mild relaxations of the given optimization problem in Chapter 5. We consider the minimization problem from the perspective of nonlinear optimization using the primal and their Lagrangian dual problems. For example, the optimization problem

$$(\mathrm{P2}*) \quad \begin{cases} \min\limits_{T\in\mathbb{R}^{d\times m}} & \|T^*T\|_F^2 + \alpha \|T - Y_s\|_F^2 \\ \text{s.t.} & \operatorname{trace}(T^*T) = m\,. \end{cases}$$

constitutes a mild relaxation where the primal objective function and the corresponding dual are equal in their respective optimal values, i.e. (P2$*$) does not possess a duality gap. In this context, tools from matrix analysis such as the Wielandt-Hoffman-Theorem for singular values will be introduced. We also discuss the relation to other optimization problems in the field of Compressive Sensing and formulate a heuristic method based on the relaxations for computing minimizers of the PFP.

Chapter 6 evaluates the performance of the proposed method compared to standard cluster algorithms such as STEM ([Erns 05]), DIB-C ([Kim 07]) and the well-known k-means algorithm. For the evaluation, simulated and real data from biological experiments will be used. Necessary tools for the evaluation such as permutation-based significance testing or the Adjusted Rand Index are introduced. We further present an example showing the applicability of our PFP-based algorithm in feature recognition in multispectral data. Finally, Chapter 7 serves as a brief overview on open problems which will be dealt with in future work.

Chapter 2

Frames

Frames were first introduced in 1952 by Duffin and Schaeffer in their work on non-harmonic Fourier series ([Duff 52]). Later, during the rise of wavelets and the corresponding applications in Signal Processing Theory, they drew attention due to ground-breaking works like the ones by Daubechies ([Daub 92]), Chui ([Chui 92]) or Hernández and Weiss ([Hern 96]).

The reason for the increased interest in frames in signal processing is mainly based on their ability in extracting and stressing characteristic features from signals compared to using standard orthonormal decompositions, e.g. wavelet bases. In contrast to bases, frames can be linearly dependent. The inherent redundancy leads to decompositions that are more stable against errors by corrupted or missing coefficients. A summary on the developments in frame theory and an overview on certain special cases can be found in the articles by Kovačević and Chebira ([Kova 07a, Kova 07b]).

Finite Unit Norm Tight Frames (FUNTFs) started attracting interest in the end of the 1990's and the beginning of the following decade due to publications by Goyal et al. ([Goya 98, Goya 01]) or Benedetto and Fickus ([Bene 03]). Goyal et al. ([Goya 98]) proved that randomly distributing m points independently and identically with a uniform distribution on the unit sphere asymptotically leads to FUNTFs as $m \to \infty$. In 2003, Benedetto and Fickus ([Bene 03]) characterized the class of FUNTFs as vectors in $\mathbb{K}^d$ which are exactly the minimizers of the (Total) Frame Potential, a functional that we introduce in Section 2.1. Minimization of the frame potential corresponds to finding configurations of m unit norm vectors which are in

equilibrium under the underlying (frame) force. The article initiated a fast development in this area whereas one has to admit that the theory basically rests on simple linear algebra due to the finite dimensionality. As we will show in the following chapters, many results on finite frames can be re-formulated using the singular value decomposition which simplifies the proofs as well.

Another reason for the increased consideration of FUNTFs was, for example, the optimality of analysis and synthesis of data in terms of a general quantization model ([Goya 01]). Shortly after the article by Benedetto and Fickus, Casazza generalized the frame potential approach by introducing the weighted frame potential distributing m vectors in $\mathbb{K}^d$ on arbitrary centered spheres with radii $r_1, \ldots, r_m$ ([Casa 04]). Together, Casazza and Fickus extended the frame potential concept even further to fusion frames ([Casa 09]).

In the theory of Compressed Sensing, where one is often interested in finding spanning systems in which the given data has a sparse coefficient representation, FUNTFs have also been studied ([Dono 06]). Ehler ([Ehle 12a]), Ehler and Okoudjou ([Ehle 12b]) created probabilistic versions of the frame potential, and Ehler and Galanis ([Ehle 11a]) showed their applicability in directional statistics. An exhaustive view on the recent development in the theory of finite frames is given by Casazza and Kutyniok in [Casa 13].

Later on, in Section 3.3, we adapt a functional proposed by Benedetto et al. in [Bene 10], by generating a weighted mean of the frame potential and a data-fitting term. This already lead us to introduce the Penalized Frame Potential in [Spri 11] for the selection of cluster prototypes, which we will analyze and discuss for both theoretical and practical purposes in this thesis. The primal objective will consist of the clustering of real-valued data vectors projected onto the unit sphere. This modeling justifies the concentration on FUNTFs which will be regarded primarily throughout this thesis after a general introduction into frame theory.

Definition 2.1. *Let $\mathcal{H}$ be a Hilbert space and I an index set. A family of vectors $\Theta = \{\theta_k\}_{k \in I}$ in $\mathcal{H}$ constitutes a frame, if constants $0 < A \leq B$ exist, such that for all $y \in \mathcal{H}$ the frame condition*

$$A \|y\|^2 \leq \sum_{k \in I} |\langle y, \theta_k \rangle|^2 \leq B \|y\|^2 \tag{2.1}$$

holds. A and B denote the frame bounds.

In the case of equal frame bounds $A = B$, the family $\{\theta_k\}_{k\in I}$ is called tight. Duffin and Schaeffer ([Duff 52]) defined frames for the Hilbert space $\mathcal{H} = L_2([0,1])$. In wavelet theory and signal processing, most results are formulated for the space of square-integrable functions over the real line, i.e. $\mathcal{H} = L_2(\mathbb{R})$.

In general, a family $\{\theta_k\}_{k\in I}$ forms by definition a Bessel sequence, if there exists a Bessel bound $B > 0$, such that the upper bound condition in (2.1) holds. It is easy to see that the corresponding operator

$$\begin{aligned} T^* \;:\; \mathcal{H} &\to \ell_2(I) \\ y &\mapsto (\langle y, \theta_k\rangle)_{k\in I} \end{aligned}$$

is bounded with $\|T^*\| \leq \sqrt{B}$. In functional analysis, T^* is often denoted as Bessel operator whereas the wavelet community commonly uses the terms analysis or decomposition operator. The adjoint operator T is called synthesis or reconstruction operator and given by

$$\begin{aligned} T \;:\; \ell_2(I) &\to \mathcal{H} \\ (c_k)_{k\in I} &\mapsto \sum_{k\in I} c_k\theta_k\,. \end{aligned}$$

If the lower bound condition in (2.1) also applies, i.e. $\{\theta_k\}_{k\in I}$ being a frame, the composition $S = TT^* : \mathcal{H} \to \mathcal{H}$ defines the frame operator. Furthermore, S is self-adjoint, positive, invertible and the inverse S^{-1} becomes itself a frame operator with bounds $0 < B^{-1} \leq A^{-1}$. The corresponding family $\{\tilde{\theta}_k\}_{k\in I}$ with $\tilde{\theta}_k = S^{-1}\theta_k$ for all $k \in I$ defines the canonical dual frame satisfying the identities

$$y = \sum_{k\in I} \langle y, \theta_k\rangle\ \tilde{\theta}_k = \sum_{k\in I} \langle y, \tilde{\theta}_k\rangle\ \theta_k \tag{2.2}$$

with unconditional convergence of both series for all $y \in \mathcal{H}$ ([Chri 08], Theorem 5.1.7). Note that one is often interested in finding other dual frames with certain features that are generally not satisfied by the canonical dual. For example, if $\mathcal{H} = L^2(\mathbb{R})$, compactness of the support is a common objective.

In the case of $A = B$, i.e. $\{\theta_k\}_{k\in I}$ constituting a tight frame, we have $S = A \cdot Id$ where Id denotes the identity on $\mathcal{H}$. Hence, (2.2) reduces to

$$y = A^{-1}\sum_{k\in I} \langle y, \theta_k\rangle\ \theta_k \qquad \forall y \in \mathcal{H} \tag{2.3}$$

and the frame condition (2.1) becomes the Parseval-type identity

$$A\|y\|^2 \;=\; \sum_{k\in I} |\langle y, \theta_k\rangle|^2 \qquad \forall y \in \mathcal{H}\,.$$

If $A = 1$, the frame $\{\theta_k\}_{k\in I}$ is also often referred to as a Parseval frame. By Casazza and Kovačević ([Casa 03]), the following theorem on Parseval frames is known as Naimark's theorem in operator theory and was first published by Akhiezer and Glazman in [Akhi 66]. Later on, the theorem was rediscovered and reformulated in the frame theoretical framework by Han and Larson in [Han 00a].

Theorem 2.2 (Naimark [Akhi 66], Han and Larson [Han 00a])**.** *The family $\{\theta_k\}_{k\in I}$ constitutes a Parseval frame of the Hilbert space $\mathcal{H}$ if and only if there exists a Hilbert space $\mathcal{H}_0 \supseteq \mathcal{H}$ with orthonormal basis $\{\varphi_k\}_{k\in I}$ such that the orthogonal projection $P : \mathcal{H}_0 \to \mathcal{H}$ satisfies $P\varphi_k = \theta_k$ for all $k \in I$.*

Note that if the elements of a Parseval frame are unit vectors, i.e. $\|\theta_k\| = 1$ for $k = 1, \ldots, m$, the family is an ONB of $\mathcal{H}$ and vice versa.

2.1 Finite Frames

Throughout the following chapters we will use the finite-dimensional Hilbert spaces $\mathcal{H} = \mathbb{K}^d$ ($\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$) and the index set $I = \{1, \ldots, m\}$ where $d, m \in \mathbb{N}$. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm induced by the inner product $\langle x, y\rangle = y^* x$ where y^* is the transposed complex conjugate of $y \in \mathbb{K}^d$. It is a well-known fact (and easy to verify) that the finite family $\Theta = \{\theta_k\}_{k=1,\ldots,m}$ is a frame of $\mathcal{H}$ if and only if it spans $\mathbb{K}^d$. Note that we use $\{\,\}$-braces both for sets and for families of vectors like in [Bene 03]. Families are allowed to contain multiplicities of single elements whereas sets are not. However, the meaning will become clear from the context.

With column vectors $\theta_k \in \mathcal{S}^{d-1}$, where

$$\mathcal{S}^{d-1} \;=\; \{v \in \mathbb{K}^d \mid \|v\| = 1\} \tag{2.4}$$

denotes the unit sphere in $\mathbb{K}^d$, the matrix

$$T \;=\; [\theta_1, \ldots, \theta_m] \in \mathbb{K}^{d\times m} \tag{2.5}$$

defines the Frame Matrix

$$S = TT^* = \sum_{k=1}^{m} \theta_k \theta_k^* \quad \in \mathbb{K}^{d\times d}$$

and the Gramian Matrix

$$G \;=\; T^*T \;=\; (\langle\, \theta_k, \theta_\ell \,\rangle)_{k,\ell=1\,\ldots,m} \quad \in \mathbb{K}^{m\times m}\,.$$

Obviously, since $\theta_k \in \mathcal{S}^{d-1}$, the diagonal entries of G satisfy $g_{k,k} = 1$. In the following, let $(\mathcal{S}^{d-1})^m = \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1}$ denote the m-fold Cartesian product of the unit sphere.

Example 2.3. *The simplest FUNTFs in* $\mathbb{R}^2$ *are given by the real and imaginary parts of the* m^{th} *complex roots of unity, e.g. for* $m = 3$ *we get the frame*

$$\Theta \;=\; \left\{(1,\,0)^T,\quad (-1/2,\,\sqrt{3}/2)^T,\quad (-1/2,\,-\sqrt{3}/2)^T\right\}$$

with corresponding Frame Matrix

$$S = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} = 3/2\, I_2$$

where I_2 *stands for the* 2×2 *identity matrix. Furthermore, the Gramian matrix is given by*

$$G = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix}.$$

Remark 2.4. (1) Note that the real and imaginary parts of the roots of unity form Grassmannian frames in $\mathbb{R}^2$: FUNTFs are called equiangular, if $|\,\langle\theta_k, \theta_\ell\rangle\,| = c$ for all $1 \le k < \ell \le m$ and some constant $c > 0$, i.e. the non-diagonal entries of the Gramian G are equal in absolute value. In general, the maximal frame correlation defined by Strohmer and Heath in [Stro 03]

$$\mathcal{M}(\Theta) \;=\; \max_{1\le k<\ell\le m} |\langle\theta_k, \theta_\ell\rangle| \tag{2.6}$$

satisfies the lower bound condition

$$\mathcal{M}(\Theta) \;\ge\; \sqrt{\frac{m-d}{d(m-1)}} \tag{2.7}$$

for all families $\Theta \in (\mathcal{S}^{d-1})^m$. Grassmannian frames are defined as the minimizers of (2.6). The right-hand side of (2.7) is a Welch bound ([Welc 74]). It constitutes a sharp bound since equality holds if and only if Θ is equiangular and tight ([Stro 03]). If furthermore all elements θ_k are normalized, Θ is an optimal Grassmannian frame. The problem of finding or constructing equiangular frames is closely related to arranging m linear subspaces of dimension $n < d$ in $\mathbb{R}^d$ such that the angles between the normal vectors are as large as possible, a problem which has been addressed by Conway et al. in [Conw 96] as a minimization problem in the Grassmannian space

$$\mathcal{G}(d,n) \quad = \quad \left\{ U \preceq \mathbb{R}^d \mid \dim(U) = n \right\}.$$

(2) The special class of harmonic frames is generated by taking $d \leq m$ rows of a discrete Fourier transform matrix M of size $m \times m$ and letting $\theta_1, \ldots, \theta_m \in \mathbb{K}^d$ denote the columns of that matrix. It is easy to see that $\{\theta_k\}_{k=1,\ldots,m}$ constitute an equal-norm Parseval frame for $\mathbb{K}^d$ with $\|\theta_k\| = \sqrt{\frac{d}{m}}$ for $k = 1, \ldots, m$ and normalization by $\sqrt{\frac{m}{d}}$ leads to a FUNTF. An example for a real-valued version of such a matrix of size 3×3 was constructed by Zimmermann ([Zimm 01]):

$$M \quad = \quad \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & \cos(\frac{2\pi}{3}) & \cos(\frac{4\pi}{3}) \\ 0 & \sin(\frac{2\pi}{3}) & \sin(\frac{4\pi}{3}) \end{bmatrix}$$

is already normalized appropriately and taking the last $d = 2$ rows implies that the frame in Example 2.3 also is harmonic. Note that in the case $\mathbb{K} = \mathbb{R}$ the choice of the d rows is not arbitrary.

Hochwald et al. ([Hoch 00]) propose the usage of harmonic tight frames in antenna array design which, interestingly enough, is again closely related to packings in Grassmannian spaces. The article also states that the construction of harmonic tight frames has been used earlier by Balan or Daubechies without publication.

(3) Multiplication of the frame elements in Example 2.3 by $\sqrt{A^{-1}} = \sqrt{2/3}$ leads to the Parseval frame

$$\tilde{\Theta} \quad = \quad \left\{ (\sqrt{2/3},\, 0)^T, \quad (-1/\sqrt{6},\, 1/\sqrt{2})^T, \quad (-1/\sqrt{6},\, -1/\sqrt{2})^T \right\}.$$

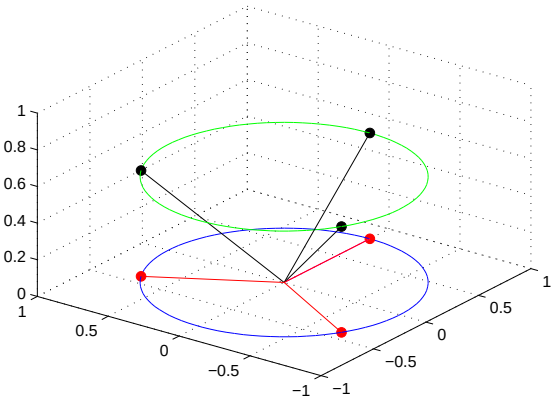


Figure 2.1: Example for Naimark's Theorem from Remark 2.4.3 with a Parseval frame for $\mathbb{R}^2$ (red) being the orthogonal projection of an orthonormal basis in $\mathbb{R}^3$ (black)

Identifying $\mathcal{H} = \mathbb{R}^2$ with the (x, y)-plane in $\mathcal{H}_0 = \mathbb{R}^3$ and letting $P : \mathcal{H}_0 \to \mathcal{H}$ the orthogonal projection, the family $\{\varphi_1, \varphi_2, \varphi_3\}$ with

$$\varphi_1 = \begin{pmatrix} \sqrt{2/3} \\ 0 \\ 1/\sqrt{3} \end{pmatrix}, \quad \varphi_2 = \begin{pmatrix} -1/\sqrt{6} \\ 1/\sqrt{2} \\ 1/\sqrt{3} \end{pmatrix}, \quad \varphi_3 = \begin{pmatrix} -1/\sqrt{6} \\ -1/\sqrt{2} \\ 1/\sqrt{3} \end{pmatrix}$$

is an orthonormal basis of $\mathbb{R}^3$ and satisfies $P\varphi_k = \theta_k$ for $k = 1, 2, 3$ in the sense of Naimark's Theorem (Theorem 2.2). Figure 2.1 presents the ONB consisting of $\varphi_1, \varphi_2, \varphi_3 \in \mathbb{R}^3$ and the corresponding vectors $\theta_1, \theta_2, \theta_3 \in \mathbb{R}^2$ which constitute a Parseval frame. △

Example 2.3 underlines the following important property of FUNTFs:

Lemma 2.5 ([Goya 98])**.** *Let $\Theta = \{\theta_k\}_{k=1,\dots,m} \in (\mathcal{S}^{d-1})^m$ be a family of $m \geq d$ unit vectors. Θ is an A-FUNTF, if and only if $S = AI_d \in \mathbb{K}^{d\times d}$ and $A = m/d$.*

The value m/d is often referred to as the measure of redundancy of the frame. Whereas all orthonormal bases $\Theta \in$ in $\mathbb{K}^d$ have $A = 1$, an increase of A shows the additional computational cost as well as redundancy in the information contained in representation (2.3).

Probably the most important characterization of FUNTFs for given dimension d and cardinality m was developed by Benedetto and Fickus in [Bene 03]. The idea is to define a repelling force between the frame elements leading to an equilibrium.

Definition 2.6 ([Bene 03]). *For a family* $\Theta = \{\theta_k\}_{k=1,\dots,m} \in (\mathcal{S}^{d-1})^m$*, the (Total) Frame Potential is defined as the mapping* $\mathrm{TFP} : (\mathcal{S}^{d-1})^m \to \mathbb{R}$*,*

$$\mathrm{TFP}\,(\Theta) \;=\; \sum_{k,\ell=1}^{m} |\langle \theta_k, \theta_\ell \rangle|^2 \;.$$

Using the above notations, the TFP can be calculated by

$$\mathrm{TFP}\,(\Theta) = \|G\|_F^2 = \mathrm{trace}\left((T^*T)^2\right) = \|S\|_F^2$$

with $\|.\|_F$ denoting the Frobenius norm on $\mathbb{K}^{m\times m}$. Moreover, if $T = U\Sigma V^*$ denotes the singular value decomposition (SVD) of T, where $U \in U(d)$, $V \in U(m)$ are unitary matrices and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_{\min\{d,m\}}) \in \mathbb{R}^{d\times m}$ with singular values $\sigma_j \geq 0$ for all j, the frame potential reads as

$$\mathrm{TFP}(\Theta) = \|\Sigma^T\Sigma\|_F^2 = \sum_{j=1}^{\min\{d,m\}} \sigma_j^4 \,.$$

Furthermore, the constraint that the trace of the Gramian G equals the cardinality of the family Θ can also be formulated in terms of the SVD by

$$m = \sum_{k=1}^{m} \|\theta_k\|^2 = \|T\|_F^2 = \|\Sigma\|_F^2 = \sum_{j=1}^{\min\{d,m\}} \sigma_j^2 \,.$$

Hence, using this relaxation, the TFP can be considered as a quartic polynomial under quadratic constraints. It is easy to see that the TFP under the given constraint is minimized by $\sigma_1 = \ldots = \sigma_d = \sqrt{m/d}$ if $m \geq d$ and $\sigma_1 = \ldots = \sigma_m = 1$ otherwise.

According to the following theorem, FUNTFs can be regarded as generalizations of the orthonormal sequences in $\mathbb{K}^d$. It also connects the frame potential to the FUNTFs and orthonormal sequences, respectively.

Theorem 2.7 (Theorem 7.1 in [Bene 03]).

(i) Every local minimizer of TFP *is also a global minimizer.*

(ii) If $m \leq d$*, then*

$$\min_{\Theta\in(\mathcal{S}^{d-1})^m} \mathrm{TFP}\,(\Theta) \;=\; m \tag{2.8}$$

and the minimizers are the orthonormal m*-sets in* $\mathbb{K}^d$*, i.e.* $\langle\, \theta_k, \theta_l \,\rangle = \delta_{k,\ell}$ *for all* $k, \ell = 1, \dots, m$*.*

(iii) If $m > d$, then

$$\min_{\Theta \in (\mathcal{S}^{d-1})^m} \mathrm{TFP}(\Theta) \;=\; m^2/d \tag{2.9}$$

and the minimizing families with m elements are the (m/d)-FUNTFs in $\mathbb{K}^d$.

Note that for the FUNTF Θ in example 2.3 we have

$$\mathrm{TFP}(\Theta) \;=\; 9/2 \tag{2.10}$$

and the singular values satisfy $\sigma_1 = \sigma_2 = \sqrt{3/2}$.

A method for generating FUNTFs $\{\theta_k\}_{k=1,\dots,m}$ in $\mathbb{R}^d$ for $m \geq d$ can be derived from Casazza's and Leon's algorithm in [Casa 02a] and [Casa 02b]. The main idea here is to construct the orthogonal matrix $V \in O(m)$ such that for $V^* = [v_1, \dots, v_m]$ it holds that $\|\tilde{v}_1\| = \dots = \|\tilde{v}_m\| = \sqrt{d/m}$, where $\tilde{v}_k = (v_{k,1}, \dots, v_{k,d})^T$ for $k = 1, \dots, m$ are generated by omitting the last $m - d$ rows of V^*. Using $U \in O(d)$ and $\Sigma = \mathrm{diag}(\sqrt{m/d}, \dots, \sqrt{m/d}) \in \mathbb{R}^{d \times m}$ it follows that $T = U\Sigma V^*$ has unit vectors as columns and satisfies the constraints for the singular values for minima of the frame potential.

Remark 2.8. **(1)** In [Casa 04], Casazza et al. gave an alternative proof for Theorem 2.7, by formulating a generalization for the eigenspaces of the frame operator S. We will adapt this concept in the following section for a characterization of all critical points of the frame potential.

(2) Goyal et al. define in [Goya 01] the equivalence relation

$$T_1 \sim T_2 \;\;\Leftrightarrow\;\; \exists U \in U(d),\, \Delta = \mathrm{diag}(\delta_1, \dots, \delta_d), \delta_k = \pm 1 : T_2 = \Delta U T_1\,,$$

where $U(d)$ is the group of unitary matrices in $\mathbb{C}^{d \times d}$. If $m = d + 1$, all FUNTFs are in the same equivalence class. For example, the frame from example 2.3 is a representative of the class containing all other FUNTFs with three elements. $\triangle$

2.2 Critical Points of the Frame Potential

In the following, Θ denotes the family of unit norm vectors $\theta_1, \dots, \theta_m$ and we concentrate on the case $\mathbb{K} = \mathbb{R}$ since it is quite illuminating. According to [Bene 03] and [Absi 08], we apply

the following definition:

Definition 2.9. *The finite family* $\Theta = \{\theta_k\}_{k=1,\dots,m} \in (\mathcal{S}^{d-1})^m$ *is called* (TFP-)*critical, if all* θ_k *for* $k = 1, \dots, m$ *are eigenvectors of the corresponding frame matrix* $S = TT^*$*. In addition, if* Θ *is critical, we also call* $T = [\theta_1, \dots, \theta_m]$ *critical. In the case, that a critical* Θ *is neither a local minimizer nor a local maximizer of* TFP*,* Θ *is called saddle point.*

By Theorem 2.7, local minima of TFP are global. Furthermore,

$$\min_{\Theta} \mathrm{TFP}(\Theta) \quad = \quad m \cdot \max\{1, m/d\}\,.$$

Applying the classical Lagrange approach for constrained minimization of the Frame Potential on the m-fold unit sphere in $\mathbb{R}^d$ gives the Lagrange function

$$\begin{aligned} \mathcal{L}(\Theta, \lambda) &= \mathrm{TFP}(\Theta) + \sum_{j=1}^{m} \lambda_j \left(\|\theta_j\|^2 - 1\right) \\ &= \sum_{k,\ell=1}^{m} |\langle \theta_k, \theta_\ell \rangle|^2 + \sum_{j=1}^{m} \lambda_j \left(\|\theta_j\|^2 - 1\right) \end{aligned}$$

where $\|\cdot\|$ is the euclidean norm. The m equality constraints

$$g_j(\Theta) \quad = \quad \|\theta_j\|^2 - 1, \quad j = 1, \dots, m,$$

describe the non-convex feasible set

$$(\mathcal{S}^{d-1})^m = \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1} = \{\Theta \mid g_j(\Theta) = 0,\, j = 1, \dots, m\}\,.$$

Since the Jacobian matrix of the mapping $F : \mathbb{R}^{dm} \to \mathbb{R}^m$, $F(\Theta) = (g_1(\Theta), \dots, g_m(\Theta))^T$ is

$$DF(\Theta) \quad = \quad \begin{bmatrix} 2\theta_1^T & & & \\ & 2\theta_2^T & & \\ & & \ddots & \\ & & & 2\theta_m^T \end{bmatrix} \in \mathbb{R}^{m \times dm}\,,$$

the full rank condition $\mathrm{rank}(DF)(\Theta) = m$ is satisfied for all $\Theta \in (\mathcal{S}^{d-1})^m$, which is necessary for the application of the Lagrange approach to the problem at hand, see e.g. [Rock 93].

For the derivative of the TFP let $j \in \{1, \dots, m\}$ and $\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_m \in \mathcal{S}^{d-1}$. Define the functions $h_j : \mathbb{R}^d \to \mathbb{R}$ by

$$h_j(\theta) \quad = \quad \sum_{k,\ell \neq j} \langle \theta_k, \theta_\ell \rangle^2 + 2 \sum_{k \neq j} \langle \theta_k, \theta \rangle^2 + \langle \theta, \theta \rangle^2\,.$$

Then, h_j is a quartic polynomial in the components of θ and the total derivative in θ is

$$\nabla h_j(\theta) \;=\; 4\sum_{k\neq j}\langle\theta_k,\theta\rangle\,\theta_k^T + 4\,\langle\theta,\theta\rangle\,\theta^T$$

which, with $\theta=\theta_j$, leads to

$$\nabla h_j(\theta_j) \;=\; 4\sum_{k=1}^{m}\langle\theta_k,\theta_j\rangle\,\theta_k^T\,,$$

or, equivalently,

$$\nabla h_j(\theta_j)^T \;=\; 4S\theta_j\,.$$

Therefore it holds that

$$\nabla\mathrm{TFP}(\Theta)^T \;=\; 4\,ST\,. \tag{2.11}$$

Analogously, the total derivative of $\mathcal{L}$ in Θ and λ leads to the system of extremal conditions

$$\begin{pmatrix} 4S\theta_1+2\lambda_1\theta_1 \\ \vdots \\ 4S\theta_m+2\lambda_m\theta_m \\ \|\theta_1\|^2-1 \\ \vdots \\ \|\theta_m\|^2-1 \end{pmatrix} \;\stackrel{!}{=}\; 0\in\mathbb{R}^{(d+1)m}\,,$$

or, with $\Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_m)$ denoting the diagonal matrix containing the Lagrange multipliers,

$$\begin{aligned} 4ST+2T\Lambda \;&=\; 0\in\mathbb{R}^{d\times m}, \\ \|\theta_j\|^2 \;&=\; 1,\quad j=1,\ldots,m\,. \end{aligned} \tag{2.12}$$

Note that (2.12) is equivalent to $S\theta_k=-\frac{\lambda_k}{2}\theta_k$, $k=1,\ldots,m$. This implies that the Lagrange multipliers $\lambda_1,\ldots,\lambda_m$ satisfy an eigenvalue equation and by denoting the spectrum of S by $\mathrm{spec}(S)$ we have $\{-\lambda_1/2,\ldots,-\lambda_m/2\}\subseteq\mathrm{spec}(S)$. Hence, $(-\frac{\lambda_k}{2},\theta_k)$ are eigenpairs of S and only critical Θ are candidates for extrema of TFP.

Since every FUNTF Θ satisfies $S=\frac{m}{d}I_d$, it follows from $\mathrm{Eig}(S,m/d)=\mathbb{R}^d$ that Θ is critical. However, Benedetto and Fickus show in [Bene 03] that critical Θ exist which do not constitute a FUNTF:

Example 2.10 ([Bene 03]). *Let* $N = \{\nu_1, \ldots, \nu_5] \in \mathbb{R}^{4\times 5}$ *with*

$$\nu_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \nu_2 = \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \\ 0 \\ 0 \end{pmatrix}, \nu_3 = \begin{pmatrix} -1/2 \\ -\sqrt{3}/2 \\ 0 \\ 0 \end{pmatrix}, \nu_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \nu_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Then the frame matrix S_N *is given by*

$$S_N = \begin{bmatrix} 3/2 & 0 & 0 & 0 \\ 0 & 3/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and it is easy to see that $\nu_1, \nu_2, \nu_3 \in \mathrm{Eig}(S, 3/2)$ *and* $\nu_4, \nu_5 \in \mathrm{Eig}(S, 1)$.

Theorem 2.11 ([Bene 03]). *A finite sequence of unit vectors* $\Theta = \{\theta_k\}_{k=1,\ldots,m}$ *is critical if and only if the sequence may be partitioned into a collection of mutually orthogonal vectors, each of which is a FUNTF for its span. Furthermore, the partition may be chosen explicitly to be* $\{E_\mu\}$ *where* $E_\mu = \{\theta_k : S\theta_k = \mu\theta_k\}$. *Also, the frame constant of* E_μ *is* μ, *and the spans of the* $\{E_\mu\}$ *are precisely the non-trivial eigenspaces of* S.

Consider $\mu_1 > \mu_2 > \ldots > \mu_s \geq 0$ as the pairwise distinct eigenvalues of S. Analogously to the proof of Theorem 7.4 in [Bene 03] we define for $j = 1, \ldots, s$ the index sets

$$I_j = \{k \in \{1, \ldots, m\} : S\theta_k = \mu_j\theta_k\},$$

which build a partition of $\{1, \ldots, m\}$ (with $I_s = \emptyset$, if S is not regular, i.e. $\mu_s = 0$). By Theorem 2.11, the families $\{\theta_k\}_{k\in I_j}$ build frames of the eigenspaces if Θ is critical. Due to the symmetry of S and the orthogonality of the eigenspaces in that case, the map TFP can be decomposed into the restrictions on its eigenspaces:

$$\mathrm{TFP}(\Theta) = \sum_{k,\ell=1}^{m} |\langle \theta_k, \theta_\ell \rangle|^2 = \sum_{j=1}^{s} \sum_{k,\ell\in I_j} |\langle \theta_k, \theta_\ell \rangle|^2 =: \sum_{j=1}^{s} \mathrm{TFP}_j(\Theta).$$

Using the notations $m_j = |I_j|$ and $d_j = \dim \mathrm{Eig}(S, \mu_j)$, Theorem 2.11 immediately leads to the following conclusion.

Corollary 2.12. *If Θ is critical, the restrictions on the eigenspaces of S satisfy*

$$\mathrm{TFP}_j(\Theta) = \sum_{k,\ell \in I_j} |\langle\, \theta_k, \theta_\ell \,\rangle|^2 = \mu_j m_j \,, \quad j = 1, \ldots, s \,.$$

Furthermore, $\mu_j = 1$ if $\{\theta_k\}_{k\in I_j}$ is ONB of $\mathrm{Eig}(S, \mu_j)$ and $\mu_j = \frac{m_j}{d_j}$ if the family is a frame.

Proof. Theorem 2.11 shows that $\mathrm{TFP}_j(\Theta) = m_j = d_j$ if $\{\theta_k\}_{k\in I_j}$ is an ONB of $\mathrm{Eig}(S, \mu_j)$ and $\mathrm{TFP}_j(\Theta) = \frac{m_j^2}{d_j}$ if it is a FUNTF. Then for $v \in \mathrm{Eig}(S, \mu_j)$ it follows that

$$\mu_j v = Sv = \sum_{k=1}^{m} \langle v, \theta_k \rangle \; \theta_k = \sum_{k\in I_j} \langle v, \theta_k \rangle \; \theta_k$$

which is v in the case of an ONB or $\frac{m_j}{d_j}\, v$ otherwise. □

The matrix $S = TT^*$ with $T = [\theta_1, \ldots, \theta_m]$ has only non-negative eigenvalues due to positive semi-definiteness. Since it holds that

$$\begin{aligned} \|S\theta_k\|^2 &= \|\sum_{\ell=1}^{m} \langle\, \theta_k, \theta_\ell \,\rangle \theta_\ell\|^2 \\ &= \|\theta_k + \sum_{\ell\neq k} \langle\, \theta_k, \theta_\ell \,\rangle \theta_\ell\|^2 \\ &= \|\theta_k\|^2 + 2\,\mathrm{Re}\,\langle\, \theta_k, \sum_{\ell\neq k} \langle\, \theta_k, \theta_\ell \,\rangle \theta_\ell \,\rangle + \|\sum_{\ell\neq k} \langle\, \theta_k, \theta_\ell \,\rangle \theta_\ell\|^2 \\ &= 1 + 2\sum_{\ell\neq k} |\langle\, \theta_k, \theta_\ell \,\rangle|^2 + \|\sum_{\ell\neq k} \langle\, \theta_k, \theta_\ell \,\rangle \theta_\ell\|^2 \geq 1 \,, \end{aligned} \tag{2.13}$$

the (normalized) column vectors of critical T are eigenvectors of S with eigenvalues greater or equal 1. Moreover, equality holds if and only if $\theta_k \perp \mathrm{span}\{\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_m\}$.

Lemma 2.13. *If Θ is critical, then $\mathrm{spec}(S) \subset (\{0\} \cup [1, m])$ and the Lagrange multipliers satisfy $-2m \leq \lambda_k \leq -2$, $k = 1, \ldots, m$.*

Proof. For the verification of the upper bound consider the singular value decomposition $T = U\Sigma V^*$. As seen before

$$m \;=\; \sum_{j=1}^{d} \sigma_j^2 \,,$$

with σ_j^2 being the eigenvalues of S. If an eigenpair (μ, v) with $v \notin \mathrm{span}\{\theta_1, \ldots, \theta_m\}$ exists, it follows that $\mu = 0$ since $\mathbb{R}^d = \mathrm{span}\{\theta_1, \ldots, \theta_m\} \oplus \ker(S)$ is an orthogonal sum. Together

with (2.13) it can be concluded that spec(S) is in $\{0\} \cup [1, m]$. Finally, the extremal condition $S\theta_k = -\frac{\lambda_k}{2}\theta_k$ gives $1 \leq -\frac{\lambda_k}{2} \leq m$ for all $k = 1, \ldots, m$ which completes the proof. □

In [Bene 03], Benedetto and Fickus consider only minima of the frame potential for the characterization of the FUNTFs. From

$$\mathrm{TFP}(\Theta) = \sum_{k,\ell=1}^{m} |\langle\, \theta_k, \theta_\ell \,\rangle|^2 \leq \sum_{k,\ell=1}^{m} \|\theta_k\|^2 \|\theta_\ell\|^2 = m^2\,,$$

we also get that Θ is a global maximum of the function TFP if and only if $c_1, \ldots, c_m \in \mathbb{C}$ exist with $|c_k| = 1$ and $c_1\theta_1 = \ldots = c_m\theta_m$. In that case the entries of the Gramian matrix satisfy $|g_{k,\ell}| = 1$. Thus, in $\mathbb{R}^d$, these are exactly the subsets of the unit sphere consisting of antipodal vectors. The following theorem states that also every local maximum of TFP is global.

Theorem 2.14. *Let $m, d \in \mathbb{N}$ and $\mu_1, \ldots, \mu_s \in \{0\} \cup [1, m]$ with $\mu_1 > \mu_2 > \ldots > \mu_s \geq 0$ denoting the pairwise distinct eigenvalues of S. The critical family $\Theta = \{\theta_k\}_{k=1,\ldots,m}$ is a saddle point of* TFP*, if and only if one of the following holds:*

(i) $\mu_2 \geq 1$,

(ii) $\mu_2 = 0$ *and the multiplicity* d_1 *of* μ_1 *satisfies* $1 < d_1 < \min\{d, m\}$.

Proof. A local and global minimum can be ruled out by Theorem 2.7, since these do only have the $(\min\{d, m\})$-fold eigenvalue $\mu_1 = \max\{1, m/d\}$. Hence, it suffices to show that no local maximum can exist under the assumptions.

By Corollary 2.12, the restrictions on the eigenspaces take their global minima in Θ:

$$\mathrm{TFP}_j(\Theta) = \sum_{k,\ell \in I_j} |\langle\, \theta_k, \theta_\ell \,\rangle|^2 = \mu_j |I_j|\,, \quad j = 1, \ldots, s.$$

If $\mu_2 \geq 1$, it holds that $\mu_1 > 1$ and the family $\{\theta_k\}_{k\in I_1}$ therefore is a FUNTF of $\mathrm{Eig}(S, \mu_1)$ by Theorem 2.11. Thus, a small perturbation on θ_{k_0}, $k_0 \in I_1$, in $\mathrm{Eig}(S, \mu_1)$, such that the FUNTF-condition is not satisfied, enlarges the function value of TFP_1 and the function value of $\mathrm{TFP} = \sum_{j=1}^{s} \mathrm{TFP}_j$ increases in the corresponding direction. If $\mu_2 = 0 \in \mathrm{spec}(S)$, almost the same argument can be used. In that case, $\mathbb{R}^d = \mathrm{Eig}(S, \mu_1) \oplus \mathrm{Ker}(S)$ is an orthogonal sum where $\dim \mathrm{Ker}(S) > 0$. Since Θ is critical, the family $\{\theta_k\}_{k\in I_1}$ builds a FUNTF or

orthonormal system of $\mathrm{Eig}(S, \mu_1)$. If $\dim \mathrm{Eig}(S, \mu_1) > 1$, perturbations on θ_{k_0} in $\mathrm{Eig}(S, \mu_1)$ revoke the minimality condition which, again, is equivalent to the existence of an ascend direction in Θ.

The only case which is open is $\mu_2 = 0$ and $d_1 = 1$. In that case it holds that $\mathrm{TFP}(\Theta) = m^2$, which corresponds to a global maximum. The equivalence follows directly from the fact, that no other cases are possible. □

The only case which has not been regarded in the proof is $\mu_2 = 0$ and $\dim \mathrm{Eig}(S, \mu_1) = 1$ which is only possible if $\mathrm{rank}(T) = \dim \mathrm{span}\{(\theta_1, \ldots, \theta_m)\} = 1$. In that case Θ is critical with the single positive eigenvalue $\mu_1 = m$ of S.

Corollary 2.15. *Let Θ be a critical family. Then Θ is a global maximum, if and only if $\mu_2 = 0$ and $\mu_1 = m$ is an eigenvalue of S with multiplicity 1.*

If two distinct eigenvalues $\mu_1 > \mu_2 > 1$ of S exist, directions of increase or decrease of the TFP can be constructed directly with the method of Benedetto and Fickus in the proof of Theorem 7.4 in [Bene 03]. The vectors θ_k, $k \in I_2$, are a frame of $\mathrm{Eig}(S, \mu_2)$. Due to the linear dependence, there exist $\beta_k \in \mathbb{C}$, $k \in I_2$, satisfying $\sum_{k \in I_2} \overline{\beta_k} \theta_k = 0$. Without any restriction, β_k can be chosen such that $|\beta_k|^2 < 1/2$. Let $\varepsilon > 0$, (μ_1, θ) eigenpair with normalized θ and $\tilde{\Theta} = \{\tilde{\theta}_k\}_{k=1,\ldots,m}$ with

$$\tilde{\theta}_k \quad = \quad \begin{cases} \sqrt{1 - \varepsilon^2 |\beta_k|^2}\, \theta_k + \varepsilon \beta_k \theta, & k \in I_2 \\ \theta_k, & k \notin I_2. \end{cases}$$

Then

$$\mathrm{TFP}(\tilde{\Theta}) \quad = \quad \mathrm{TFP}(\Theta) + 2(\mu_1 - \mu_2)\varepsilon^2 \left(\sum_{k \in I_2} |\beta_k|^2 \right) + R(\varepsilon)\varepsilon^4 \,,$$

where $R(\varepsilon)$ is bounded in magnitude and therefore $\mathrm{TFP}(\tilde{\Theta}) > \mathrm{TFP}(\Theta)$.

Thus, the restriction $\mu_2 > 1$ guarantees the existence of the linear coefficients β_k as described. For $\mu_2 = 1$ the construction from the proof of Theorem 2.14 can be used. On the other hand, we receive a decrease in function value, if instead of the elements in $\mathrm{Eig}(S, \mu_2)$ the elements of the spanning frame of $\mathrm{Eig}(S, \mu_1)$ are altered by the construction by Benedetto and Fickus.

Example 2.16. *Let $N = (\nu_1, \dots, \nu_5)$ be defined as in Example 2.10. Then*

$$S = \begin{bmatrix} 3/2 & 0 & 0 & 0 \\ 0 & 3/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathrm{TFP}(N) = \|S\|_F^2 = 13/2.$$

The vectors ν_1, ν_2, ν_3 constitute a FUNTF of $\mathrm{Eig}(S, 3/2)$ and ν_4, ν_5 are ONB of $\mathrm{Eig}(S, 1)$. Hence, N is critical and the partition of the index set $\{1, \dots, 5\}$ according to the proof of Theorem 2.14 is $I_1 = \{1, 2, 3\}$ and $I_2 = \{4, 5\}$. For $\beta_1 = \beta_2 = \beta_3 =: \beta \in \mathbb{R}$ we get $\sum_{k \in I_1} \beta \nu_k = 0$. Let $\gamma := \varepsilon|\beta|$ and define according to the construction $\hat{\nu}_4 = \nu_4$, $\hat{\nu}_5 = \nu_5$ and

$$\begin{aligned}
\hat{\nu}_1 &= \sqrt{1-\gamma^2} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \gamma \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{1-\gamma^2} \\ 0 \\ \gamma \\ 0 \end{pmatrix}, \\
\hat{\nu}_2 &= \frac{1}{2} \begin{pmatrix} -\sqrt{1-\gamma^2} \\ \sqrt{3-3\gamma^2} \\ 2\gamma \\ 0 \end{pmatrix}, \\
\hat{\nu}_3 &= \frac{1}{2} \begin{pmatrix} -\sqrt{1-\gamma^2} \\ -\sqrt{3-3\gamma^2} \\ 2\gamma \\ 0 \end{pmatrix}.
\end{aligned}$$

Then the new frame matrix $\hat{S} = \hat{T}\hat{T}^*$ is

$$\hat{S} = \mathrm{diag}\left(\frac{3}{2}(1-\gamma^2), \frac{3}{2}(1-\gamma^2), 3\gamma^2+1, 1\right)$$

and therefore

$$\begin{aligned}
\mathrm{TFP}(\hat{N}) = \|\hat{S}\|_F^2 &= \tfrac{9}{2}(1-\gamma^2)^2 + (3\gamma^2+1)^2 + 1 \\
&= \tfrac{13}{2} + \tfrac{27}{2}\gamma^4 - 3\gamma^2 \\
&= \mathrm{TFP}(N) + g(\gamma)
\end{aligned}$$

with $g(\gamma) := \frac{27}{2}\gamma^4 - 3\gamma^2 < 0$ for $0 < \gamma < \sqrt{2}/3$ and g strictly decreasing on $[0, 1/3]$. For $\gamma = 1/3$ we have $\hat{S} = \mathrm{diag}(\frac{4}{3}, \frac{4}{3}, \frac{4}{3}, 1)$. Furthermore, $\{\hat{\nu}_1, \dots, \hat{\nu}_4\}$ constitutes a FUNTF of $\mathrm{Eig}(S, 4/3)$ and $\hat{\nu_5}$ is ONB of $\mathrm{Eig}(S, 1)$.

2.3 Spectrum and Non-Tightness

In this last section on the introduction of Frames, we define the Non-Tightness as a means in order to measure "how far away" from a FUNTF a family of vectors in $\mathcal{S}^{d-1}$ is.

Definition 2.17. *Let $m \geq d$, $\Theta = \{\theta_k\}_{k=1,\dots,m} \subset \mathcal{S}^{d-1}$ be a finite family of (not necessarily pairwise distinct) normalized vectors, $T = [\theta_1, \dots, \theta_m] \in \mathbb{C}^{d\times m}$ and $S = TT^*$. Then the mapping* $\mathrm{NT} : \mathcal{S}^{d-1} \times \dots \times \mathcal{S}^{d-1} \to \mathbb{R}$,

$$\mathrm{NT}(\Theta) \;=\; \left\| S - \frac{m}{d}\, I_d \right\|_F^2$$

defines the Non-Tightness of the family Θ.

Obviously, the Non-Tightness is closely connected to the frame potential which satisfies $\mathrm{TFP}(\Theta) = \|S\|_F^2$. However, as we will see in Chapter 4, the Non-Tightness works as a helpful tool in the analysis of the asymptotic behavior of certain functionals used for the data clustering approach, which we introduce in Chapter 3.

It is easy to see from (2.1) that the eigenvalues $\mu_j = \sigma_j^2$ of an arbitrary frame operator are located in the interval $[A, B]$ where $0 < A \leq B < \infty$, which already implies the regularity of S. As stated in Lemma 2.5, if the columns of T constitute a FUNTF, then $A = B = m/d$. Let $T = U\Sigma V^*$ be again the singular value decomposition of T with $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_d) \in \mathbb{R}^{d\times m}$ and unitary matrices $U \in U(d)$ and $V \in U(m)$. By definition, NT is zero if and only if Θ is a FUNTF. Furthermore, by the symmetry of S, NT can be written as $\mathrm{TFP} - m^2/d$ since

$$\begin{aligned}
\mathrm{NT}(\Theta) &= \mathrm{trace}\left((S - \frac{m}{d}\, I_d)^2\right) \\
&= \|S\|_F^2 - \frac{2m}{d}\,\mathrm{trace}\,(TT^*) + \frac{m^2}{d^2}\,\mathrm{trace}\,(I_d) \\
&= \mathrm{TFP}(\Theta) - \frac{m^2}{d}\,.
\end{aligned}$$

The following proposition shows that the Non-Tightness has a natural interpretation as the variance of the spectrum of the frame matrix S.

Proposition 2.18. *Let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_d \geq 0$ denote the eigenvalues of the frame matrix S. Then $\bar{\mu} = \frac{m}{d}$ is the mean of the eigenvalues and the sample variance $\hat{\sigma}_\mu$ satisfies*

$$(d-1)\hat{\sigma}_\mu \;=\; \mathrm{NT}(\Theta)\,. \tag{2.14}$$

Proof. $\bar{\mu} = \frac{m}{d}$ follows directly from the fact that

$$m = \sum_{j=1}^{d} \sigma_j^2 = \sum_{j=1}^{d} \mu_j \, .$$

Then, due to the unitary invariance of the Frobenius norm,

$$\begin{aligned} \mathrm{NT}(\Theta) &= \left\| S - \frac{m}{d} \, I_d \right\|_F^2 \\ &= \left\| \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2) - \frac{m}{d} \, I_d \right\|_F^2 \\ &= \sum_{j=1}^{d} (\mu_j - \bar{\mu})^2 \end{aligned}$$

which establishes (2.14). □

Chapter 3

Cluster Algorithms for Short Time Series

Clustering real data describes the attempt to identify groups whose members are similar in terms of a predefined measure. This rather old problem affects many fields in science and a vast number of approaches exist in the literature. Nowadays, the clustering problem is commonly dealt with in Data Mining and Machine Learning Theory. A classification of different cluster algorithms into five major categories can be found, e.g., in the book of Han and Kamber ([Han 00b]). Based on this classification, various examples for partitioning (e.g. k-means, k-medoids, fuzzy c-means), hierarchical (agglomerative/ bottom-up, divisive/ top-down), density-based (DBSCAN, OPTICS), grid-based (STING) and model-based (AutoClass, ART) methods are described in Liao's exhaustive survey ([Liao 05]), where the problem of choosing appropriate distance or dissimilarity measures for the clustering process is also addressed.

The major part of the following chapter has already been published in [Spri 11]. However, the analysis and the evaluation including numerical results, which follow in Chapters 4, 5 and 6, have not been included in the mentioned contribution.

Classical methods for the analysis of long time series are based on principal component analysis and discrete wavelet transforms ([Hast 01], [Qu 03]). A feature-based incremental clustering

method is described by Vlachos et al. ([Vlac 03]) where the multilevel resolution capability of Haar wavelets is combined with a coarse-to-fine selection of centers for the k-means clustering algorithm. More recently, dimension reduction methods by kernel eigenmap methods like Laplacian ([Belk 03], [Ehle 11b]) and Schroedinger eigenmaps ([Czaj 13]) project the data into lower-dimensional subspaces with as little distortion as possible. In contrast, in most cases of microarray analysis, the lengths of the time series are rather small. For example, according to Ernst et al. ([Erns 05]), more than 80% of the time series in the Stanford Microarray Database (`http://smd.princeton.edu`, accessed November 28, 2013) consist of the values measured at eight time points or less. That leads to a large number of data in a low-dimensional space. Since most methods for analyzing long time series are not well-suited or even applicable for short time series, different approaches and algorithms have to be developed.

In some applications, data-dependent frames are computed in order to generate sparse coefficient representations of the given data. The advantage rests on the fact that frames as redundant spanning systems – sometimes also denoted as overcomplete dictionaries – allow for different choices of these representations. In the clustering problem for short time series we deal with the problem of determining cluster prototypes. These prototypes are in a sense also supposed to represent the expansion of the data, which is why we propose a frame theoretic approach in this thesis.

Prototype selection is typical in partitioning methods. Approaches for the application on data from short time series include the Short Time Series Expression Miner (STEM) developed by Ernst et al. ([Erns 05]) and Difference-Based Clustering (DIB-C) by Kim and Kim ([Kim 07]). In the following, we primarily concentrate on the STEM algorithm.

The first step of the STEM approach consists in constructing model profiles (cluster specific prototypes of time series) $p_1, \ldots, p_m$ such that

$$\min_{j \neq k} \quad \mathrm{dis}(p_j, p_k)$$

is maximized, where dis is an appropriate dissimilarity measure.

Section 3.1 gives a geometric interpretation of the dissimilarity measure in the STEM algorithm, justifying the application of frame theoretic tools. Afterwards, in Section 3.2, we show that the main problem in the STEM algorithm has a natural connection to some classical prob-

lems in distributing points on unit spheres ([Thom 04, Tamm 30]). Finally, we introduce the Penalized Frame Potential in Section 3.3, a data-driven functional based on a frame theoretic approach introduced by Benedetto et al. in [Bene 10] for sparse coefficient representations.

3.1 On the Short Time Series Expression Miner

The Short Time Series Expression Miner (STEM) by Ernst et al. ([Erns 05]) provides an algorithm for clustering short time series. Suppose, we are given N discrete time series $(s_{j,t})_{t=0,\ldots,d} = (s_{j,0}, \ldots, s_{j,d})$ of length $(d+1)$ where $N \gg d$ and d small (usually less than twelve). Firstly, a log-normalization transforms each time series $(s_{j,t})$ into points

$$(x_{j,t})_{t=0,\ldots,d} \quad = \quad (\log(s_{j,t}/s_{j,0}))_{t=0,\ldots,d} \in \mathbb{R}^{d+1}. \tag{3.1}$$

Note that $x_{j,0} = 0$. This is only a preliminary step. It presupposes positive data $s_{j,t}$ and is a useful normalization of raw data of an exponential growth or decay phase, e.g. measurements from short-time biological processes. In the following, we assume that all time series are in log-normalized representation.

The first step of the STEM algorithm defines a large set $P = \{p_1, \ldots, p_M\}$ of synthetic model profiles (interpreted as cluster-specific prototype time series), with $p_\ell \in \mathbb{R}^{d+1}$ and large $M \in \mathbb{N}$. For specifying P, the user defines a control parameter $c \in \mathbb{N}$ which is supposed to represent approximately the maximum change in values between successive time points of the time series. As the log-transformed data satisfies $x_{j,0} = 0$, each model profile $p_\ell \in P$ is required to satisfy $p_{\ell,0} = 0$. Moreover, from time t to $t+1$, the model profiles can only increase or decrease by an integer less than or equal to c, i.e.

$$p_{\ell,t+1} - p_{\ell,t} \quad \in \quad \{0, \pm 1, \ldots, \pm c\}\,.$$

As the set P of all possible profiles is large ($|P| = (2c+1)^d$ for time series of length $d+1$), the next step of the algorithm reduces P to a subset R of m profiles where m should be chosen moderately. Small m leads to the representation of separate clusters by the same prototype whereas large m causes higher computational cost.

Ernst et al. propose to solve

$$R^* \quad = \quad \underset{R \subset P, |R|=m}{\arg\max} \quad \min_{p,q \in R, p \neq q} \operatorname{dis}(p,q)\,, \tag{3.2}$$

where dis is an appropriate dissimilarity measure. In other words, R^* is chosen such that the minimal inherent dissimilarity in the profile set is maximized.

In the following we use

$$\operatorname{dis}(p,q) \quad = \quad 1 - \rho(p,q)$$

with Bravais-Pearson correlation coefficient ρ; see (3.5) for a mathematical definition. Note that non-integer values for the control parameter c in (3.2) could also be considered for generating the set P. However, scaling the profiles by positive constants does not bias the dissimilarity measure dis, i.e. $\operatorname{dis}(\alpha p, q) = \operatorname{dis}(p,q)$ for $\alpha > 0$ and profiles p, q.

Since the optimization problem of finding R^* is NP-hard, a greedy algorithm is proposed by Ernst et al. in order to find an approximate solution $\tilde{R}$ for the problem in (3.2). It starts with $\tilde{R}$ containing only the model profile that decreases by c at each time point, i.e. $p_1 = (0, -c, -2c, \ldots, -dc)$. After that, in each of the remaining $m-1$ steps, the profile p that meets

$$p \quad = \quad \underset{q \in P \setminus \tilde{R}}{\arg\max} \quad \min_{r \in \tilde{R}} \operatorname{dis}(q,r) \tag{3.3}$$

is added to $\tilde{R}$. Obviously, for the chosen dissimilarity in (3.3), due to perfect negative correlation, the second profile will always be $p_2 = (0, c, 2c, \ldots, dc)$.

Note that greedy techniques for maximizing intercluster distances often lead to satisfying prototype configurations with reasonable calculational effort, see, e.g., the algorithm proposed by Batchelor and Wilkins in [Batc 69].

The second step of the STEM algorithm consists in assigning each observed time series $(x_{j,t})_{t=0,\ldots,d}$ to the closest prototype in terms of the given dissimilarity measure. For more details, we refer to Ernst et al. ([Erns 05]). Since we focus on proposing an alternative method for finding prototypes, we will solely concentrate on the first step here.

As already seen in [Spri 11], we present our interpretation of the dissimilarity measure and

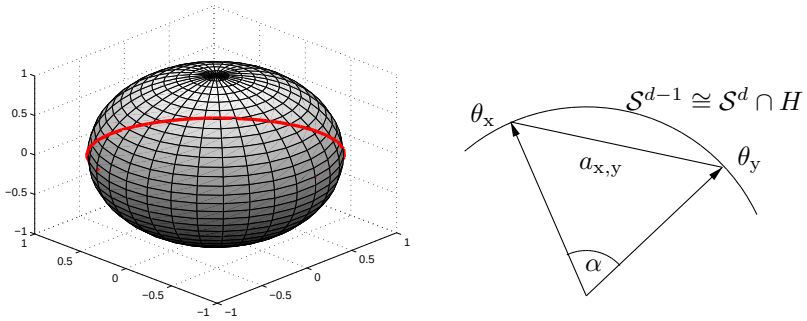


Figure 3.1: Intersection of sphere $\mathcal{S}^2$ and plane H (left), Euclidean distance $a_{\mathrm{x,y}} = \|\theta_{\mathrm{x}} - \theta_{\mathrm{y}}\|_2$ (right)

the optimization in (3.2) from a geometric perspective involving the (real) unit sphere

$$\mathcal{S}^{d-1} = \left\{ v \in \mathbb{R}^d \mid \|v\|_2 = 1 \right\} \subset \mathbb{R}^d. \tag{3.4}$$

Let $\mathrm{x} = (x_0, \ldots, x_d)$, $\mathrm{y} = (y_0, \ldots, y_d) \in \mathbb{R}^{d+1}$ be two vectors representing the values of two real-valued time series in log-ratio normalization and let $\bar{x} = (d+1)^{-1} \sum_{k=0}^{d} x_k$ and $\bar{y}$ be the mean values of x and y, respectively. Remember that the log-transformation causes $x_0 = y_0 = 0$.

According to the definition of the proposed dissimilarity measure in (3.3) we get

$$\mathrm{dis}(\mathrm{x},\mathrm{y}) \quad = \quad 1 - \rho(\mathrm{x},\mathrm{y}) \quad = \quad 1 - \frac{\sum_{k=0}^{d}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=0}^{d}(x_k - \bar{x})^2}\sqrt{\sum_{k=0}^{d}(y_k - \bar{y})^2}}\,. \tag{3.5}$$

Furthermore, we define in $\mathbb{R}^{d+1}$ the (d-dimensional) hyperplane

$$H = \left\{ v \in \mathbb{R}^{d+1} \,\middle|\, \sum_{k=0}^{d} v_k = 0 \right\} \tag{3.6}$$

containing all vectors in $\mathbb{R}^{d+1}$ having coordinate sums equal to zero. Then the orthogonal projection of $\mathbb{R}^{d+1}$ onto the linear subspace H is given by

$$\begin{aligned} \mathcal{P}_H \quad : \quad \mathbb{R}^{d+1} \quad &\to \quad H \\ \mathrm{x} \quad &\mapsto \quad \mathcal{P}_H(\mathrm{x}) = (x_0 - \bar{x}, \ldots, x_d - \bar{x})^T\,. \end{aligned}$$

Moreover, the point $\theta_{\mathrm{x}} = \mathcal{P}_H(\mathrm{x}) / \|\mathcal{P}_H(\mathrm{x})\|_2$ is in the intersection $H \cap \mathcal{S}^d$. Using the standard

inner product $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^{d+1}$, we observe that the correlation coefficient ρ can be expressed as

$$\begin{aligned} \rho(\mathrm{x}, \mathrm{y}) &= \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_k (x_k - \bar{x})^2} \sqrt{\sum_k (y_k - \bar{y})^2}} \\ &= \frac{\langle \mathcal{P}_H(\mathrm{x}), \mathcal{P}_H(\mathrm{y}) \rangle}{\|\mathcal{P}_H(\mathrm{x})\|_2 \, \|\mathcal{P}_H(\mathrm{y})\|_2} \\ &= \langle\, \theta_\mathrm{x}, \theta_\mathrm{y} \,\rangle \ . \end{aligned}$$

It is well known that $H \cap \mathcal{S}^d$ is isomorphic to the $(d-1)$-dimensional unit sphere $\mathcal{S}^{d-1} \subset \mathbb{R}^d$. Using this geometric identification, we denote the corresponding points on $\mathcal{S}^{d-1}$ again by θ_x and θ_y. Note that, for $(d+1) = 3$, $H \cap \mathcal{S}^d$ is a Riemannian circle (see Figure 3.1).

Application of the cosine formula leads to

$$\mathrm{dis}(\mathrm{x}, \mathrm{y}) \ = \ 1 - \langle\, \theta_\mathrm{x}, \theta_\mathrm{y} \,\rangle \quad = \quad 1 - \cos\alpha \ = \ a_{\mathrm{x},\mathrm{y}}^2 / 2$$

with α and $a_{\mathrm{x},\mathrm{y}}$ as in Figure 3.1.

Hence, the dissimilarity of the two time series can be interpreted as half the squared Euclidean distance between their projections onto the unit sphere $\mathcal{S}^{d-1}$. For example, two perfectly negatively correlated time series $\mathrm{x}, \mathrm{y} \in \mathbb{R}^{d+1}$ have $\mathrm{dis}(\mathrm{x}, \mathrm{y}) = 2$, their projections θ_x, θ_y are antipodal on $\mathcal{S}^{d-1}$ with $a_{\mathrm{x},\mathrm{y}} = 2$.

3.2 A Note on Tammes' Problem

We return to the interpretation using the set of profiles R and the Euclidean distance $a_{p,q} = \sqrt{2\,\mathrm{dis}(p,q)} = \|\theta_p - \theta_q\|_2$. For $d = 3$ (i.e., with observed time series at 4 time points), the problem of finding the m most distinct profiles

$$R^* \quad = \quad \underset{R \subset P, \# R = m}{\arg\max} \quad \min_{p,q \in R, p \neq q} a_{p,q} \tag{3.7}$$

is related to Tammes' problem ([Tamm 30]) of finding

$$\Theta_m \quad = \quad \underset{\{\theta_1, \ldots, \theta_m\} \subset \mathcal{S}^2}{\arg\max} \quad \min_{1 \leq j < k \leq m} \|\theta_j - \theta_k\|_2 \ , \tag{3.8}$$

which was proposed in order to simulate the distribution of pores on pollen grains. Note that the main difference consists in relaxing the restriction in (3.2).

A similar and even older problem has been considered by Thomson in [Thom 04] and is given by application of a certain heterogeneity measure instead of the Euclidean distance:

$$\Phi_m = \underset{\{\phi_1,\ldots,\phi_m\}\subset\mathcal{S}^2}{\arg\min} \mathcal{E}(\phi_1,\ldots,\phi_m) \tag{3.9}$$

where

$$\mathcal{E}(\phi_1,\ldots,\phi_m) = \sum_{1\leq j<k\leq m} \|\phi_j-\phi_k\|_2^{-1} \tag{3.10}$$

denotes the associated energy on the sphere. For example, the Fekete points ([Saff 97]) solve Thomson's problem; for a more exhaustive presentation of related problems and their solutions we refer to Conway and Sloane ([Conw 93]).

3.3 Motivation of the Penalized Frame Potential

In [Spri 11], we proposed a new method for the choice of the set of cluster prototypes $R = \{p_\ell \mid 1 \leq \ell \leq m\}$, viz. the first step of the STEM algorithm. More precisely, a replacement of the optimization criterion in problem (3.2) by a new criterion is applied, where the new one is based on the frame potential and a data-dependent penalty term. We assume that no further dimension reduction is applicable without changing the geometry of the data, therefore we only deal with the case $m > d$.

First of all, we recall, that the given time series $(s_{j,t})_{t=0,\ldots,d}$, $j = 1,\ldots,N$, are log-transformed and projected onto $\mathcal{S}^{d-1} \subset \mathbb{R}^d$. For the sake of consistency with the terminology used in other related work, we denote the matrix containing these projections by $Y \in \mathbb{R}^{d\times N}$ with normalized column vectors $y_j \in \mathcal{S}^{d-1}$.

Benedetto, Czaja and Ehler proposed in [Bene 10] to solve the minimization problem

$$\min_{\Theta\in(\mathcal{S}^{d-1})^m} \mathrm{TFP}(\Theta) + \mathrm{P}(\Theta, Y) \tag{3.11}$$

for finding sparse coefficient representations for retinal data. $\mathrm{P}(\cdot, Y)$ in (3.11) stands for a data-dependent penalty term, its choice depends on the given problem. The argument of choosing a criterion based on the frame potential is, that its minimization leads to point configurations on $\mathcal{S}^{d-1}$ which are "as orthogonal to each other as possible" ([Bene 03]). So the first part of

the criterion is responsible for the coverage of all dimensions of our data space. The remaining objective for choosing good candidates for cluster prototypes is to define a penalty term that guarantees approximation of the data.

In our case, we use the penalty term

$$\mathrm{P}(\Theta, Y) \;=\; m + 1 - \sum_{k=1}^{m} \max_{1\le j\le N} \langle y_j, \theta_k\rangle$$

and minimize the α-weighted (normalized) optimality criterion

$$\mathrm{F}_\alpha(\Theta, Y) \;=\; \frac{d}{m^2}\cdot \mathrm{TFP}(\Theta) + \alpha \cdot \mathrm{P}(\Theta, Y), \tag{3.12}$$

with $\alpha \in [0, \infty)$. Our objective is to find the minimizing families $\Theta = \{\theta_k\}_{k=1,\dots,m} \subset \mathcal{S}^{d-1}$ of the optimization problem

$$\min_{\Theta\in(\mathcal{S}^{d-1})^m} \quad \mathrm{F}_\alpha(\Theta, Y).$$

Both components are normalized such that their respective minimal values are 1. Thus, both expressions $\frac{d}{m^2}\cdot \mathrm{TFP}(\Theta)$ and $\mathrm{P}(\Theta, Y)$ are real numbers greater or equal to 1. Indeed, the data-independent (first) part of the criterion (3.12) takes its minimum value, if and only if Θ is a FUNTF. Hence, the minimizers in (3.12) for $\alpha = 0$ are exactly the FUNTFs, which agrees with the characterization in Chapter 2.

On the other hand, minimizing P enforces maximal correlation of each prototype θ_k with at least one data vector y_j. Therefore,

$$\mathrm{P}(\Theta, Y) \;\ge\; 1\,,$$

and equality holds if $\Theta \subset Y$. In the extreme case $\alpha \to \infty$, the minimizer Θ_α of (3.3) corresponds to a subset of Y (a feature which will be further investigated in Chapter 4). For the case "$\alpha = \infty$" even repetitions of the same unit vector $\theta_1 = \theta_2 = \cdots = \theta_m = y_i \in \mathcal{S}^{d-1}$ give a minimal solution. However, this "collapse" of the prototypes θ_j to only one point is prohibited by the influence of the data-independent frame potential TFP in (3.12).

Remark 3.1. (1) Using polarization and given that $\theta_k, y_j \in \mathcal{S}^{d-1}$ for all j and all k, the penalty term P can be rewritten as

$$m + 1 - \sum_{\ell} \max_{j} \langle\, y_j, \theta_\ell \,\rangle \;=\; 1 + \frac{1}{2}\sum_{\ell} \min_{j} \|y_j - \theta_\ell\|^2\,. \tag{3.13}$$

Hence, minimizing P is equivalent to minimizing the Euclidean distances between the data and the prototypes which stresses the objective of approximating (single) data points.

(2) The proposal by Benedetto et al. ([Bene 10]) uses a different penalty term than the one in [Spri 11]. Their original objective was to find sparse coefficient representations of retinal data. Using

$$p(\theta_k) \;=\; \sum_{j=1}^{N} |\langle y_j, \theta_k \rangle| \,, \quad k = 1, \ldots, m,$$

and the total separation TS defined as

$$\mathrm{TS}(\Theta, Y) \;=\; \min_{1 \le k < \ell \le m} |p(\theta_k) - p(\theta_\ell)| \;,$$

the applied penalty term is

$$\mathrm{P}_1(\Theta, Y) \;=\; \mathrm{TS}(\Theta, Y)^{-1} \,.$$

Alternatively, the proposed method also contains the term

$$\mathrm{P}_2(\Theta, Y) \;=\; \sum_{k=1}^{m} \sum_{j=1}^{N} |\langle y_j, \theta_k \rangle| \;, \tag{3.14}$$

which corresponds to the minimization of the ℓ_1-norm of the coefficient vectors of the data in the representation of the frame Θ. Candès and Tao showed in their renowned contribution considering data recovery from corrupted measurements ([Cand 05]), that under "suitable conditions" ([Bene 10]) ℓ_1-minimization becomes equivalent to finding sparse representations. Even if these conditions are hard to satisfy in practice, the ℓ_1-approach is a common means in compressive sensing.

However, P_1 measures the distance to the whole data set. On the other hand, minimizing P_2 leads to configurations θ_ℓ that are as orthogonal to the data as possible. Both approaches are unrewarding in specifying cluster prototypes. △

Chapter 4

Analysis of the Penalized Frame Potential

In [Spri 11], the applicability of the Penalized Frame Potential in data clustering was pointed out on the basis of experimental data. However, an extensive analysis of the minimizers has not been included. Since the minimizers for $\alpha = 0$ are exactly the FUNTFs, the following chapter addresses the problem of characterizing the minimizers for a positive parameter $\alpha > 0$.

The Sections are arranged in the following way: Section 4.1 deals with the case $\alpha \to \infty$. The core statement is that the minimizer converges towards the most frame-like subfamily of the data. The applied techniques are adapted from the analysis of similar problems. The main Theorem in Section 4.2 shows that each vector θ_ℓ in the minimizing family Θ corresponds to exactly one vector $y_{s(\ell)}$ in the data Y. This feature enables us to reformulate the functional by replacing the maximization in the data-dependent part and builds the basis for different simplifications which are introduced in Chapter 5. For the understanding of the minimizers, the notion of spherical Dirichlet cells will prove helpful, the necessary definitions will be given. Finally, Section 4.3 shortly characterizes the maximizers of the Penalized Frame Potential.

4.1 Asymptotic Behavior

In the following let $\alpha \in (0, \infty)$ and $Y = \{y_j \in \mathcal{S}^{d-1} \mid 1 \leq j \leq N\}$ be a family of normalized data vectors. Furthermore, let $Y^m = Y \times \ldots \times Y$ denote the m-fold Cartesian product of the data. As mentioned, the technique that will be applied in the following for the analysis of the behavior of the Penalized Frame Potential for $\alpha \to \infty$ is common in approximation theory (see [Boor 01]).

Firstly, define the function

$$\begin{aligned} \Phi : \mathbb{R} &\to \mathbb{R} \\ \alpha &\mapsto \min_{\Theta \in (\mathcal{S}^{d-1})^m} \mathrm{F}_\alpha(\Theta, Y) \end{aligned} \tag{4.1}$$

with $\mathrm{F}_\alpha(\cdot, Y) : \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1} \to \mathbb{R}$ defined as in (3.12):

$$\mathrm{F}_\alpha(\Theta, Y) = \frac{d}{m^2} \mathrm{TFP}(\Theta) + \alpha \mathrm{P}(\Theta, Y) \,.$$

In the following, we use the abbreviating representation

$$\Theta_\alpha = \operatorname*{arg\,min}_{\Theta \in (\mathcal{S}^{d-1})^m} \mathrm{F}_\alpha(\Theta, Y) \,, \tag{4.2}$$

which allows Φ from (4.1) to be expressed by $\Phi(\alpha) = (d/m^2)\mathrm{TFP}(\Theta_\alpha) + \alpha \mathrm{P}(\Theta_\alpha, Y)$. Again, as before in Chapter 3, we focus on the data-approximating penalty term

$$\begin{aligned} \mathrm{P}(\cdot, Y) : \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1} &\to \mathbb{R} \\ \Theta &\mapsto \mathrm{P}(\Theta, Y) = (m+1) - \sum_{\ell=1}^{m} \max_{1 \leq j \leq N} \langle\, y_j, \theta_\ell \,\rangle \,. \end{aligned}$$

According to Theorem 2.7 and the Cauchy-Schwarz inequality, $1 \leq \dfrac{d}{m^2} \mathrm{TFP}(\Theta) \leq d$ with

$$\frac{d}{m^2} \mathrm{TFP}(\Theta) = \begin{cases} 1 \Leftrightarrow & \Theta \text{ FUNTF} \\ d \Leftrightarrow & \sigma_1\theta_1 = \ldots = \sigma_m\theta_m, \quad |\sigma_1| = \ldots = |\sigma_m| = 1 \,. \end{cases}$$

Furthermore, $1 \leq \mathrm{P}(\Theta, Y) \leq 2m + 1$ where equality for the lower bound holds if and only if Θ is a subfamily of Y. The upper bound applies only in the case $y_1 = \ldots = y_N = -\theta_\ell$, $1 \leq \ell \leq m$. Hence, $\Phi(\alpha) = 1 + \alpha$ is obviously only possible, if a FUNTF is contained in the data Y.

Using the set

$$\mathcal{S}_0 \;=\; \{\Theta \in (\mathcal{S}^{d-1})^m \mid \mathrm{TFP}(\Theta) = m^2/d \vee \Theta \in Y^m\} \subset (\mathcal{S}^{d-1})^m\,,$$

simple calculations yield

$$\begin{aligned} g_1(\alpha) := 1+\alpha \;&=\; \min_{\Theta\in(\mathcal{S}^{d-1})^m} \frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \alpha \min_{\Psi\in(\mathcal{S}^{d-1})^m} \mathrm{P}(\Psi, Y) \\ &\leq\; \min_{\Theta\in(\mathcal{S}^{d-1})^m} \frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \alpha\,\mathrm{P}(\Theta, Y) \qquad\qquad = \;\Phi(\alpha) \\ &\leq\; \min_{\Theta\in\mathcal{S}_0} \frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \alpha\,\mathrm{P}(\Theta, Y) \\ &\leq\; \min\{1+\alpha(2m+1),\, d+\alpha\}. \end{aligned}$$

Since $d+\alpha < 1+\alpha(2m+1)$ for $\alpha > \frac{d-1}{2m}$ (which is less than $\frac{1}{2}$ in the general case $m > d$), we consider the linear function $g_d(\alpha) := d+\alpha$ as a first upper bound which leads to

$$g_1(\alpha) \leq \Phi(\alpha) \leq g_d(\alpha) \qquad \forall \alpha \geq 0\,. \tag{4.3}$$

Let

$$\tilde{\Psi} \;=\; \operatorname*{arg\,min}_{\Psi\in Y^m} \mathrm{TFP}(\Psi) \tag{4.4}$$

denote the most "frame-like" subfamily in terms of the frame potential of exactly m (not necessarily pairwise distinct) data vectors, then the following theorem improves this upper bound:

Theorem 4.1. *Let Y be a family of vectors $y_1, \ldots, y_N \in \mathbb{R}^d$ satisfying $\|y_j\| = 1$ for $j = 1, \ldots, N$ and let Φ, $\tilde{\Psi}$ and Θ_α as defined above. Then*

$$\lim_{\alpha\to\infty} \mathrm{TFP}(\Theta_\alpha) \;=\; \mathrm{TFP}(\tilde{\Psi})$$

and

$$\lim_{\alpha\to\infty} \mathrm{P}(\Theta_\alpha, Y) \;=\; 1\,,$$

with $\mathrm{TFP}(\Theta_\alpha)$ monotonously increasing and $\mathrm{P}(\Theta_\alpha, Y)$ monotonously decreasing.

In other words, Φ converges asymptotically towards the linear function

$$\begin{aligned} g^* : \mathbb{R} \;&\to\; \mathbb{R} \\ \alpha \;&\mapsto\; g^*(\alpha) = \frac{d}{m^2}\,\mathrm{TFP}(\tilde{\Psi}) + \alpha\,. \end{aligned} \tag{4.5}$$

The proof will be given in separate steps and results from the following lemmata.

Lemma 4.2. *For $\varepsilon > 0$ let $\Psi, \Theta \in (\mathcal{S}^{d-1})^m$ be two families with $\|\psi_\ell - \theta_\ell\| < \varepsilon$ for $\ell = 1, \dots, m$. Then*

$$|\mathrm{TFP}(\Psi) - \mathrm{TFP}(\Theta)| < 4m(m-1)\varepsilon(\varepsilon+1)\,.$$

Proof. For $k, \ell = 1, \dots, m$, $k \neq \ell$, let $c_{k,\ell} := \langle\, \theta_k - \psi_k, \theta_\ell \,\rangle + \langle\, \psi_k, \theta_\ell - \psi_\ell \,\rangle$. Then $|c_{k,\ell}| < 2\varepsilon$ and

$$\begin{aligned}\Big|\,|\langle\, \psi_k, \psi_\ell \,\rangle|^2 - |\langle\, \theta_k, \theta_\ell \,\rangle|^2\Big| &= \Big|\,|\langle\, \psi_k, \psi_\ell \,\rangle|^2 - |c_{k,\ell} + \langle\, \psi_k, \psi_\ell \,\rangle|^2\Big| \\ &= \big|2c_{k,\ell}\,\langle\, \psi_k, \psi_\ell \,\rangle + c_{k,\ell}^2\big| \\ &< 4\varepsilon + 4\varepsilon^2\,,\end{aligned}$$

which leads to

$$\begin{aligned}|\mathrm{TFP}(\Psi) - \mathrm{TFP}(\Theta)| &\leq \sum_{k,\ell=1}^{m} \Big|\,|\langle\, \psi_k, \psi_\ell \,\rangle|^2 - |\langle\, \theta_k, \theta_\ell \,\rangle|^2\Big| \\ &< 4m(m-1)\varepsilon(\varepsilon+1).\end{aligned}$$

□

Lemma 4.3. *Let $m \geq d$. The function Φ is concave, monotonously increasing and satisfies*

(i) $\Phi(0) = 1$, $\lim_{\alpha\to\infty} \Phi(\alpha) = \infty$,

(ii) $\Phi(\alpha) \leq \dfrac{d}{m^2}\,\mathrm{TFP}(\tilde{\Psi}) + \alpha$ *for all* $\alpha \geq 0$ *with* $\tilde{\Psi}$ *from* (4.4),

(iii) $\Phi(\alpha) = \dfrac{d}{m^2}\,\mathrm{TFP}(\tilde{\Psi}) + \alpha = 1 + \alpha$ *for* $\alpha > 0$ *if and only if a subfamily of Y is a FUNTF.*

Proof. For the concavity let $0 \leq \alpha < \beta$ and $t \in [0, 1]$. Then

$$\begin{aligned}&\Phi(t\alpha + (1-t)\beta) \\ =\ &\min_{\Theta\in(\mathcal{S}^{d-1})^m} \frac{d}{m^2}\,\mathrm{TFP}(\Theta) + (t\alpha + (1-t)\beta)\,\mathrm{P}(\Theta, Y) \\ =\ &\min_{\Theta\in(\mathcal{S}^{d-1})^m} t\left(\frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \alpha\mathrm{P}(\Theta, Y)\right) + (1-t)\left(\frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \beta\mathrm{P}(\Theta, Y)\right) \\ \geq\ &t \min_{\Theta\in(\mathcal{S}^{d-1})^m}\left(\frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \alpha\mathrm{P}(\Theta, Y)\right) + (1-t)\min_{\Psi\in(\mathcal{S}^{d-1})^m}\left(\frac{d}{m^2}\,\mathrm{TFP}(\Psi) + \beta\mathrm{P}(\Psi, Y)\right) \\ =\ &t\Phi(\alpha) + (1-t)\Phi(\beta)\,.\end{aligned}$$

Now suppose there exist $0 \leq \alpha_1 < \alpha_2$ satisfying $\Phi(\alpha_1) \geq \Phi(\alpha_2)$. Due to the concavity of Φ, it follows that $\Phi(\alpha_2) \geq \Phi(\alpha)$ for all $\alpha \geq \alpha_2$ which, by (4.3), leads to the contradiction

$$1 + (\alpha_2 + d) \leq \Phi(\alpha_2 + d) \leq \Phi(\alpha_2) \leq \alpha_2 + d\,.$$

(i) is obvious since $\Phi(\alpha) \geq 1 + \alpha$ for all $\alpha \geq 0$ and

$$\Phi(0) \quad = \quad \min_{\Theta \in (\mathcal{S}^{d-1})^m} \frac{d}{m^2} \operatorname{TFP}(\Theta) = 1.$$

(ii) follows from

$$\Phi(\alpha) = \min_{\Theta \in (\mathcal{S}^{d-1})^m} F_\alpha(\Theta, Y) \leq \min_{\Psi \in Y^m} F_\alpha(\Psi, Y) = \frac{d}{m^2} \operatorname{TFP}(\tilde{\Psi}) + \alpha\,.$$

(iii) has already been shown in the beginning of this section. □

If Y does not contain a FUNTF (which will be most likely the case for real data), the inequality in Lemma 4.3 *(ii)* becomes strict, i.e. $\Phi(\alpha) < \frac{d}{m^2} \operatorname{TFP}(\tilde{\Psi}) + \alpha$ for all $\alpha \in [0, \infty)$.

Proof of Theorem 4.1. Firstly, we show by contradiction that the penalty term P decreases monotonously in α. Suppose that $0 \leq \alpha_1 < \alpha_2$ exist with $\mathrm{P}(\Theta_{\alpha_2}, Y) > \mathrm{P}(\Theta_{\alpha_1}, Y)$ where Θ_{α_1} and Θ_{α_2} are the minimizers in α_1 and α_2, respectively, according to the definition in (4.2). For the sake of simplicity, denote $\mathrm{P}(\Theta_j) = \mathrm{P}(\Theta_{\alpha_j}, Y)$ for $j = 1, 2$. Define the linear functions $h_1,\ h_2 : \mathbb{R} \to \mathbb{R}$,

$$h_j(\alpha) = \frac{d}{m^2} \operatorname{TFP}(\Theta_j) + \alpha \mathrm{P}(\Theta_j), \quad j = 1, 2.$$

Then the optimality of Θ_2 in α_2 leads to

$$\begin{aligned} h_2(\alpha_2) \quad &= \quad \frac{d}{m^2} \operatorname{TFP}(\Theta_2) + \alpha_2 \mathrm{P}(\Theta_2) \\ &\leq \quad \frac{d}{m^2} \operatorname{TFP}(\Theta_1) + \alpha_2 \mathrm{P}(\Theta_1) = h_1(\alpha_2)\,. \end{aligned}$$

It follows from $\mathrm{P}(\Theta_1) < \mathrm{P}(\Theta_2)$ that

$$h_2(0) \quad = \quad \frac{d}{m^2} \operatorname{TFP}(\Theta_2) < \frac{d}{m^2} \operatorname{TFP}(\Theta_1) = h_1(0)$$

which gives

$$h_2(\alpha) \quad < \quad h_1(\alpha) \quad \forall \alpha \in [0, \alpha_2)\,.$$

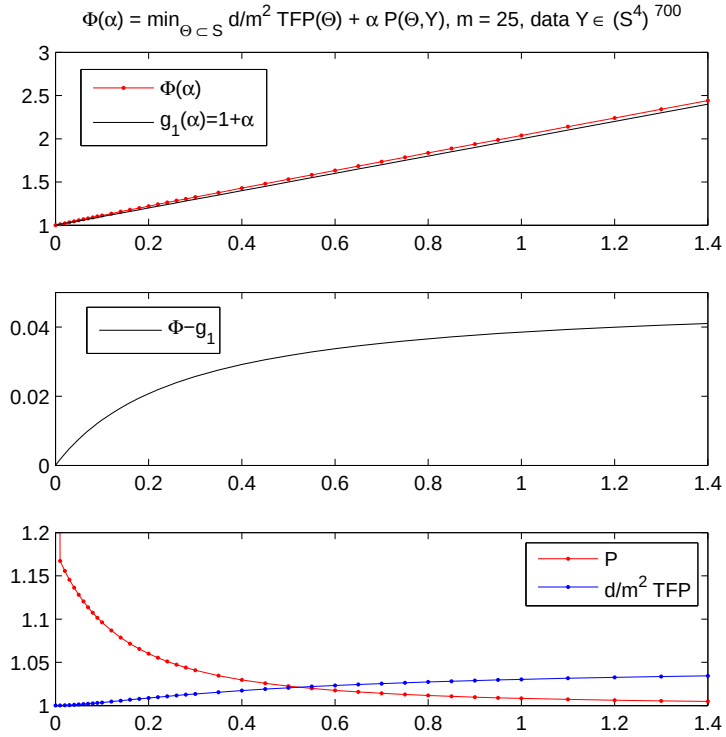


Figure 4.1: Generic plots of Φ, penalty term and frame potential for sample data `AAS.txt` with $N = 700$, $d = 5$ and $m = 25$ (for details, see Section 6.1)

Then $h_2(\alpha_1) < h_1(\alpha_1)$ leads to the contradiction. The monotonicity of TFP can be seen from

$$h_1(\alpha_1) = \frac{d}{m^2}\,\mathrm{TFP}(\Theta_1) + \alpha_1 \mathrm{P}(\Theta_1) \leq \frac{d}{m^2}\,\mathrm{TFP}(\Theta_2) + \alpha_1 \mathrm{P}(\Theta_2) = h_2(\alpha_1)$$

with $\mathrm{P}(\Theta_1) \geq \mathrm{P}(\Theta_2)$. Hence, $\mathrm{P}(\Theta_\alpha)$ decreases and $\mathrm{TFP}(\Theta_\alpha)$ increases monotonously in α.

By Lemma 4.3 and (4.3), Φ converges for $\alpha \to \infty$ to a linear function $g : \mathbb{R} \to \mathbb{R}$, $g(\alpha) = c + \alpha$, $c \in [1, d]$, which implies $\mathrm{P}(\Theta_\alpha) \to 1$ $(\alpha \to \infty)$.

For $\alpha \geq 0$ let $y_{s(\ell)}^{(\alpha)} \in Y$, $\ell = 1, \ldots, m$, satisfy

$$y_{s(\ell)}^{(\alpha)} \quad = \quad \operatorname*{arg\,max}_{1 \leq j \leq N} \langle y_j, \theta_\ell^{(\alpha)} \rangle \tag{4.6}$$

and define the families

$$Y_s^{(\alpha)} \quad = \quad \{ y_{s(\ell)}^{(\alpha)} \mid \ell = 1, \ldots, m \}\,. \tag{4.7}$$

Then (4.4) implies

$$\mathrm{TFP}(\tilde{\Psi}) \quad \leq \quad \mathrm{TFP}(Y_s^{(\alpha)})\,.$$

Moreover, the convergence of $\mathrm{P}(\Theta_\alpha)$ shows that for every $\varepsilon > 0$, there exists $\alpha_0 \geq 0$ satisfying the inequality $\langle y_{s(\ell)}^{(\alpha)}, \theta_\ell^{(\alpha)} \rangle > 1 - \varepsilon^2/2$ for $\alpha > \alpha_0$ and $\ell = 1, \ldots, m$. Then, by polarization,

$$\|y_{s(\ell)}^{(\alpha)} - \theta_\ell^{(\alpha)}\|^2 \;\; < \;\; \varepsilon^2 \qquad \forall \alpha > \alpha_0 \,,$$

which, by Lemma 4.2, completes the proof of Theorem 4.1. □

Remark 4.4. (1) Figure 4.1 shows the behavior of Φ for the sample data from [Spri 11]. The upper plot contains the concave function Φ and the linear function g_1, which bounds Φ from below. The difference $\Phi - g_1$ in the second plot is a strictly increasing concave function. Using the notion of the Non-Tightness from Section 2.3, the difference between the lower bound and g^* from (4.5) is

$$\frac{d}{m^2}\,\mathrm{TFP}(\tilde{\Psi}) + \alpha - g_1(\alpha) = \frac{d}{m^2}\,\mathrm{TFP}(\tilde{\Psi}) - 1 = \frac{d}{m^2}\,\mathrm{NT}(\tilde{\Psi})\,,$$

which acts as an asymptotic upper bound. Finally, the third plot gives an example for the convergence results in Theorem 4.1, showing that $P(\Theta_\alpha)$ decreases monotonously towards 1 whereas $\mathrm{TFP}(\Theta_\alpha)$ is increasing in α.

(2) Suppose $\tilde{\Psi}$ from (4.4) consists of m distinct vectors denoted by $\tilde{\psi}_1, \ldots, \tilde{\psi}_m \in Y^m$. Then $G_{\tilde{\Psi}} = [\tilde{\psi}_1, \ldots, \tilde{\psi}_m]^* \cdot [\tilde{\psi}_1, \ldots, \tilde{\psi}_m] \in \mathbb{R}^{m \times m}$ is the sub-Gramian of size $m \times m$ of the Gramian $G_Y = [y_1, \ldots, y_N]^* \cdot [y_1, \ldots, y_N] \in \mathbb{R}^{N \times N}$ with minimal Frobenius norm. Furthermore, by Proposition 2.18, the corresponding frame matrix $S_{\tilde{\Psi}}$ has minimal spectral variance. △

4.2 Minimal Property and Spherical Dirichlet Cells

One of the major challenges in analyzing the continuous function $\mathrm{F}_\alpha(\cdot, Y)$ from (3.12),

$$\mathrm{F}_\alpha(\Theta, Y) \;\; = \;\; \frac{d}{m^2}\,\mathrm{TFP}(\Theta) + \alpha \left(m + 1 - \sum_{\ell=1}^{m} \max_{1 \leq j \leq N} \langle\, y_j, \theta_\ell \,\rangle \right) \tag{4.8}$$

under the constraints that $\theta_\ell \in \mathcal{S}^{d-1}$ is, that the penalty term prevents global differentiability of $\mathrm{F}_\alpha(\cdot, Y)$ on $\mathcal{S}^{d-1}$. However, as we will show in this paragraph, the function is differentiable in an open neighborhood around the minimum. This allows to omit the max-term under suitable conditions and specify certain subfamilies as in (4.7) of the data Y in order to express the penalty term P in a simpler form. An important means is introduced by the following definition.

Definition 4.5. *Let $\{y_j\}_{j=1,\dots,N} \subset \mathcal{S}^{d-1}$ be a family of (pairwise distinct) vectors in $\mathbb{R}^d$. Then*

$$D_k \;=\; \{v \in \mathcal{S}^{d-1} \,:\, \|y_k - v\|_2 = \min_{j=1,\dots,N} \|y_j - v\|_2\}$$

denotes the (spherical) Dirichlet (or Voronoi) cell of the data vector y_k.

It is easy to see that the Dirichlet cells can be alternatively characterized by

$$D_k \;=\; \{v \in \mathcal{S}^{d-1} \,:\, \langle v, y_k\rangle = \max_{j=1,\dots,N} \langle v, y_j\rangle\}\,. \tag{4.9}$$

Furthermore, according to [Saff 97],

$$\bigcup_{k=1}^{N} D_k = \mathcal{S}^{d-1} \qquad \text{and} \qquad \operatorname{int}(D_j \cap D_k) = \emptyset \;\text{ for }\; j \neq k$$

which is quite obvious. It follows that $\mathcal{S}^{d-1}$ can be partitioned (with respect to the overlapping boundaries) into the N Dirichlet cells and $\partial D_k = D_k \cap \bigcup_{j\neq k} D_j$ in the standard topology of $\mathcal{S}^{d-1}$.

Proposition 4.6. *Let $\{y_j\}_{j=1,\dots,N} \subset \mathcal{S}^{d-1}$ be a family of (pairwise distinct) vectors in $\mathbb{R}^d$. Then for $j \neq k$ and $U_{j,k} = \operatorname{span}\{y_j - y_k\}^\perp$ the following holds:*

(i) The boundaries of the Dirichlet cells satisfy $\partial D_j \cap \partial D_k \subset U_{j,k}$.

(ii) Let $\mathcal{P}_{j,k} : \mathbb{R}^d \to U_{j,k}$ denote the orthogonal projection onto $U_{j,k}$. If $v_{j,k} = \dfrac{y_j + y_k}{\|y_j + y_k\|}$ exists with $v_{j,k} \in \partial D_j \cap \partial D_k$, then

$$v_{j,k} \;=\; \frac{\mathcal{P}_{j,k}(y_j)}{\|\mathcal{P}_{j,k}(y_j)\|} \;=\; \frac{\mathcal{P}_{j,k}(y_k)}{\|\mathcal{P}_{j,k}(y_k)\|}$$

and it holds that

$$\|y_j - v_{j,k}\| = \|y_k - v_{j,k}\| = \min_{w\in\partial D_j\cap\partial D_k} \|y_k - w\| = \min_{w\in\partial D_j\cap\partial D_k} \|y_j - w\|\,.$$

Proof. If $\partial D_j \cap \partial D_k = \emptyset$, *(i)* holds by definition. Now let $w \in \partial D_j \cap \partial D_k \neq \emptyset$. Then from (4.9) we get $\langle w, y_j\rangle = \langle w, y_k\rangle$ and it follows that $w \in \operatorname{span}\{y_j - y_k\}^\perp = U_{j,k}$.

Now, for the proof of *(ii)*, let $u_1, \dots, u_{d-1}$ be an orthonormal basis of $U := U_{j,k}$. Then $\langle u, y_j\rangle = \langle u, y_k\rangle$ for all $u \in U$ and therefore $\mathcal{P}_{j,k}(y_j) = \sum_\ell \langle y_j, u_\ell\rangle\, u_\ell = \mathcal{P}_{j,k}(y_k)$. Moreover,

$\xi := \frac{1}{2}(y_j + y_k)$ satisfies $\xi \in U$, $\xi \neq 0$ and

$$\begin{aligned} \mathcal{P}_{j,k}(y_j) &= \sum_\ell \langle y_j, u_\ell \rangle u_\ell \\ &= \sum_\ell \left(\frac{1}{2} \langle y_j, u_\ell \rangle + \frac{1}{2} \langle y_k, u_\ell \rangle \right) u_\ell \\ &= \sum_\ell \frac{1}{2} \langle y_j + y_k, u_\ell \rangle u_\ell = \xi \,. \end{aligned}$$

It follows that

$$v_{j,k} = \frac{\xi}{\|\xi\|} = \frac{\mathcal{P}_{j,k}(y_j)}{\|\mathcal{P}_{j,k}(y_j)\|} = \frac{\mathcal{P}_{j,k}(y_k)}{\|\mathcal{P}_{j,k}(y_k)\|} \,. \tag{4.10}$$

For $w \in \partial D_j \cap \partial D_k$ it is easy to see by $\|w\| = 1$ and (4.10) that

$$\|w - \xi\|^2 = 1 - 2\langle w, \xi \rangle + \|\xi\|^2 \geq 1 - 2\|\xi\| + \|\xi\|^2 = \|v_{j,k} - \xi\|^2 \,. \tag{4.11}$$

Since *(i)* implies $w - \xi \in U$, the fact that $\xi - y_j \in U^\perp$ and (4.11) lead to

$$\|w - y_j\|^2 = \|w - \xi + \xi - y_j\|^2 = \|w - \xi\|^2 + \|\xi - y_j\|^2 \geq \|v_{j,k} - \xi\|^2 + \|\xi - y_j\|^2 = \|v_{j,k} - y_j\|^2 \,,$$

where we used $v_{j,k} - \xi \in U$ in the last equality. □

In the special case $d = 3$, the proposition shows that the boundaries of the Dirichlet cells are subsets of great circles on the sphere. Note that it is possible that $v_{j,k} \notin \partial D_j \cap \partial D_k$ despite $\partial D_j \cap \partial D_k \neq \emptyset$.

Now, for the following, suppose that $y_1, \ldots, y_N \in \mathcal{S}^{d-1}$ with $y_j \neq y_k$ for $1 \leq j < k \leq N$. Otherwise remove all replicates from the data until the vectors are pairwise distinct and call the new set of data vectors Y. As we will see, the Dirichlet cells $D_1, \ldots, D_N$ of the data Y are of major importance for the differentiability of the function $\mathrm{F}_\alpha(\cdot, Y)$. As in Chapter 2, let $T = [\theta_1, \ldots, \theta_m] \in \mathbb{R}^{d \times m}$.

Lemma 4.7. *If $\theta_\ell \in \mathcal{S}^{d-1} \setminus \bigcup_{j=1}^N \partial D_j$ for all $\ell = 1, \ldots, m$, then the functional $\mathrm{F}_\alpha(\cdot, Y) : \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1} \to \mathbb{R}$ with*

$$F_\alpha(\Theta, Y) = \frac{d}{m^2} \mathrm{TFP}(\Theta) + \alpha \left(m + 1 - \sum_{\ell=1}^m \max_{j=1,\ldots,N} \langle y_j, \theta_\ell \rangle \right)$$

is differentiable in Θ for $\alpha \geq 0$ and the derivative of $\mathrm{F}_\alpha(\cdot, Y)$ in Θ satisfies

$$\nabla F_\alpha(\Theta, Y)^T = \frac{4d}{m^2} TT^*T - \alpha Y_s \tag{4.12}$$

with $Y_s = [y_{s(1)}, \ldots, y_{s(m)}] \in \mathbb{R}^{d \times m}$ and $y_{s(1)}, \ldots, y_{s(m)}$ as in (4.6).

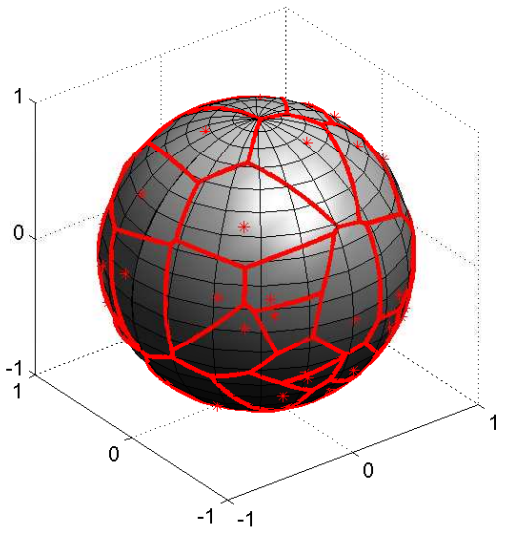


Figure 4.2: Sample data points and their Dirichlet cells on $\mathcal{S}^2$

Proof. If $\theta_\ell \notin \partial D_j$, there is exactly one data point $y_{s(\ell)}$ for each θ_ℓ with

$$\langle y_k, \theta_\ell \rangle < \langle y_{s(\ell)}, \theta_\ell \rangle = \max_{j=1,\dots,N} \langle y_j, \theta_\ell \rangle \ .$$

Hence, with Θ denoting the family $\theta_1, \dots, \theta_m$, the objective function reads as

$$\begin{aligned} F_\alpha(\Theta) &= \frac{d}{m^2}\mathrm{TFP}(\Theta) + \alpha\,(m+1-\sum_{\ell=1}^{m} \max_{j=1,\dots,N} \langle y_j, \theta_\ell \rangle) \\ &= \frac{d}{m^2}\mathrm{TFP}(\Theta) + \alpha\,(m+1-\sum_{\ell=1}^{m} \langle y_{s(\ell)}, \theta_\ell \rangle) \\ &= \frac{d}{m^2}\mathrm{TFP}(\Theta) + \alpha(m+1-\mathrm{trace}\,(Y_s^* T))\,. \end{aligned}$$

This expression obviously shows the differentiability of F_α in T or, respectively, Θ and (4.12) follows from (2.11). □

Thus, if $\theta_\ell \in \mathrm{int}\, D_{s(\ell)}$ for $\ell = 1, \dots, m$, F_α is differentiable in $\Theta \in (\mathcal{S}^{d-1})^m$ with

$$\nabla F_\alpha(\Theta, Y)^T = \frac{4d}{m^2} T T^* T - \alpha Y_s\,.$$

The following theorem connects the minimizers of the PFP functional to the Dirichlet cells $D_1, \dots, D_N$ of the data Y. In fact, it shows that the minimizers of the penalized frame potential cannot be located on the boundaries of the Dirichlet cells generated by the data in the case of strictly positive $\alpha > 0$.

Theorem 4.8. *Let $\alpha > 0$ and $m > d$. If the family of vectors $\Theta = \{\theta_1, \dots, \theta_m\} \subset \mathcal{S}^{d-1}$ represents a local minimum of $F_\alpha(\cdot, Y)$, then $\theta_\ell \notin \partial D_j$ for $\ell = 1, \dots, m$ and $j = 1, \dots, N$.*

Proof. Suppose there is a $\theta_\ell \in \partial D_j$, i.e. there exists $k \in \{1, \dots, N\} \setminus \{j\}$ such that

$$\langle y_j, \theta_\ell \rangle = \langle y_k, \theta_\ell \rangle = \max_{1 \le \mu \le N} \langle y_\mu, \theta_\ell \rangle$$

or, equivalently, $\theta_\ell \in \partial D_j \cap \partial D_k$.

Using the auxiliary function $g : \mathcal{S}^{d-1} \to \mathbb{R}$, $g(\theta) = \dfrac{d}{m^2}\mathrm{TFP}(\theta_1, \dots, \theta_{\ell-1}, \theta, \theta_{\ell+1}, \dots, \theta_m)$, define the two functions $f_{1,\alpha}, f_{2,\alpha} : \mathcal{S}^{d-1} \to \mathbb{R}$ with

$$\begin{aligned} f_{1,\alpha}(\theta) &= g(\theta) + \alpha \left(m + 1 - \langle y_j, \theta \rangle - \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{m} \max_{1 \le \mu \le N} \langle y_\mu, \theta_\nu \rangle\right) \\ &= g(\theta) + \alpha \left(m + 1 - \langle y_j, \theta \rangle - \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{m} \langle y_{s(\nu)}, \theta_\nu \rangle\right) \end{aligned}$$

and

$$f_{2,\alpha}(\theta) = g(\theta) + \alpha \left(m + 1 - \langle y_k, \theta \rangle - \sum_{\substack{\nu=1 \\ \nu \neq \ell}}^{m} \langle y_{s(\nu)}, \theta_\nu \rangle\right),$$

where $y_{s(\nu)}$ are from (4.6). Then obviously $f_{1,\alpha}(\theta) = f_{2,\alpha}(\theta)$ for $\theta \in \partial D_j \cap \partial D_k$ and

$$F_\alpha\left((\theta_1, \dots, \theta_{\ell-1}, \theta, \theta_{\ell+1}, \dots, \theta_m), Y\right) = \begin{cases} f_{1,\alpha}(\theta), & \theta \in D_j\,, \\ f_{2,\alpha}(\theta), & \theta \in D_k\,. \end{cases}$$

Moreover, by Lemma 4.7, both functions are differentiable with

$$\nabla f_{1,\alpha}(\theta)^T = \frac{4d}{m^2} TT^*\theta - \alpha y_j \qquad \text{and} \qquad \nabla f_{2,\alpha}(\theta)^T = \frac{4d}{m^2} TT^*\theta - \alpha y_k\,.$$

Since $\theta_\ell \in \partial D_j \cap \partial D_k$, there exists $y_s \in \{y_j, y_k\}$ such that $\eta := \frac{4d}{m^2} TT^*\theta_\ell - \alpha y_s \notin \operatorname{span}\{\theta_\ell\}$. Let $T = \operatorname{span}\{\theta_\ell\}^\perp$ denote the $(d-1)$-dimensional linear hyperplane which is parallel to the tangent plane to $\mathcal{S}^{d-1}$ at θ_ℓ. Then the orthogonal projection $\mathcal{P}_T$ of η onto T satisfies $v := \mathcal{P}_T(\eta) \neq 0$ and $\langle v, \eta \rangle = \langle v, \eta - v \rangle + \langle v, v \rangle = \|v\|^2$. For $\varepsilon > 0$ let $w := \|\theta_\ell - \varepsilon v\|^{-1}(\theta_\ell - \varepsilon v) \in \mathcal{S}^{d-1}$ which satisfies $w = \theta_\ell - \varepsilon v + \mathcal{O}(\varepsilon^2)$. Letting

$$f_{s,\alpha} = \begin{cases} f_{1,\alpha}, & y_s = y_j\,, \\ f_{2,\alpha}, & y_s = y_k\,, \end{cases}$$

the second-order Taylor expansion of $f_{s,\alpha}$ can be written as

$$\begin{aligned} f_{s,\alpha}(w) &= f_{s,\alpha}(\theta_\ell) + \langle w - \theta_\ell, \eta \rangle + \mathcal{O}(\varepsilon^2) \\ &= f_{s,\alpha}(\theta_\ell) - \varepsilon \langle v, \eta \rangle + \mathcal{O}(\varepsilon^2) \\ &= f_{s,\alpha}(\theta_\ell) - \varepsilon \|v\|^2 + \mathcal{O}(\varepsilon^2)\,. \end{aligned} \tag{4.13}$$

Hence, there exists a $w \in \mathcal{S}^{d-1}$ such that the family $\Theta_w = \{\theta_1, \ldots, \theta_{\ell-1}, w, \theta_{\ell+1}, \ldots, \theta_m\}$ satisfies $F_\alpha(\Theta) > F_\alpha(\Theta_w)$ and therefore Θ cannot be a local minimum of F_α. □

The crucial point in the proof is that for $\theta_\ell \in \partial D_j \cap \partial D_k$, there exists a $y_s \in \{y_j, y_k\}$ such that $\eta \notin \operatorname{span}\{\theta_\ell\}$. Otherwise the inner product $\langle v, \eta \rangle$ in (4.13) would be zero since $v = \mathcal{P}_T(\eta) = 0$ and the existence of a descent direction could not be guaranteed.

Note that the family Θ depends on the choice of the given regularization parameter $\alpha > 0$. Therefore $\theta_1, \ldots, \theta_m$ should also be subscripted with an additional α. However, for ease of notation, we skip the parameter for the single elements. The following corollaries conclude the differentiability of the PFP functional from (3.12) in its minimum.

Corollary 4.9. *If the family of normalized vectors $\Theta_\alpha = \{\theta_1, \ldots, \theta_m\}$ represents a local minimum of $F_\alpha(\cdot, Y)$ for $\alpha > 0$, then for every θ_ℓ there is exactly one*

$$y_{s(\ell)} = \underset{j=1,\ldots,N}{\arg\max} \langle \theta_\ell, y_j \rangle \ .$$

Corollary 4.10. *If the family of normalized vectors $\Theta_\alpha = \{\theta_1, \ldots, \theta_m\}$ represents a local minimum of $F_\alpha(\cdot, Y)$ for $\alpha > 0$, then $F_\alpha(\cdot, Y)$ is differentiable in an open neighborhood of Θ_α.*

4.3 Global Maxima of the Penalized Frame Potential

To complete the chapter on the analysis of the PFP functional $\mathrm{F}_\alpha(\cdot, Y)$ from (3.12), we give a short note on the maxima. As shown in Section 4.2, the minima of $\mathrm{F}_\alpha(\cdot, Y)$ on $\mathcal{S}^{d-1}$ lie, for $\alpha > 0$, in the interior of the data-generated Dirichlet cells $D_1, \ldots, D_N$. Furthermore, we know from Chapter 2 that the Total Frame Potential TFP gets maximized by a family $\Theta = \{\theta_1, \ldots, \theta_m\}$, whenever it holds that $\theta_1, \ldots, \theta_m \in \operatorname{span}\{\theta\}$ for some $\theta \in \mathcal{S}^{d-1}$.

On the other hand, the penalty term P takes its maximal value when $\theta_1, \ldots, \theta_m$ are located as far away from the data Y as possible. So the maximizer has to be on the boundary of a Dirichlet cell. Otherwise the objective value could always be increased. Thus, the maximizers of $\mathrm{F}(\cdot, Y)$ are generally multiplicities of a single vector on a boundary of some Dirichlet cell.

Proposition 4.11. *If* $\theta \in \mathcal{S}^{d-1}$ *satisfies* $\theta = \arg\min_{v \in \mathcal{S}^{d-1}} \max_{1 \le j \le N} \langle y_j, v \rangle$, *then* $\theta \in \operatorname{span}\{y_j - y_k\}^{\perp}$ *for some data vectors* $y_j, y_k \in \mathcal{S}^{d-1}$ *with* $j \neq k$ *and the family* $\Theta = \{\theta_\ell\}_{\ell=1,\ldots,m}$ *with* $\theta_1 = \ldots = \theta_m = \theta$ *maximizes* $F_\alpha(\cdot, Y)$ *for all* $\alpha \geq 0$.

The number of Dirichlet cells that intersect at θ from the Proposition does not depend on the dimension d. Figure 4.3 shows a global maximum which is located in the intersection of three Dirichlet cells. Note that $\theta_1, \ldots, \theta_m$ are perpendicular to a face of the convex hull of the data. Figure 4.4 contains an example where the maximizers lie in an intersection of only two Dirichlet cells on $\mathcal{S}^2$ and $\theta_1, \ldots, \theta_m$ are perpendicular to an edge of the convex hull.

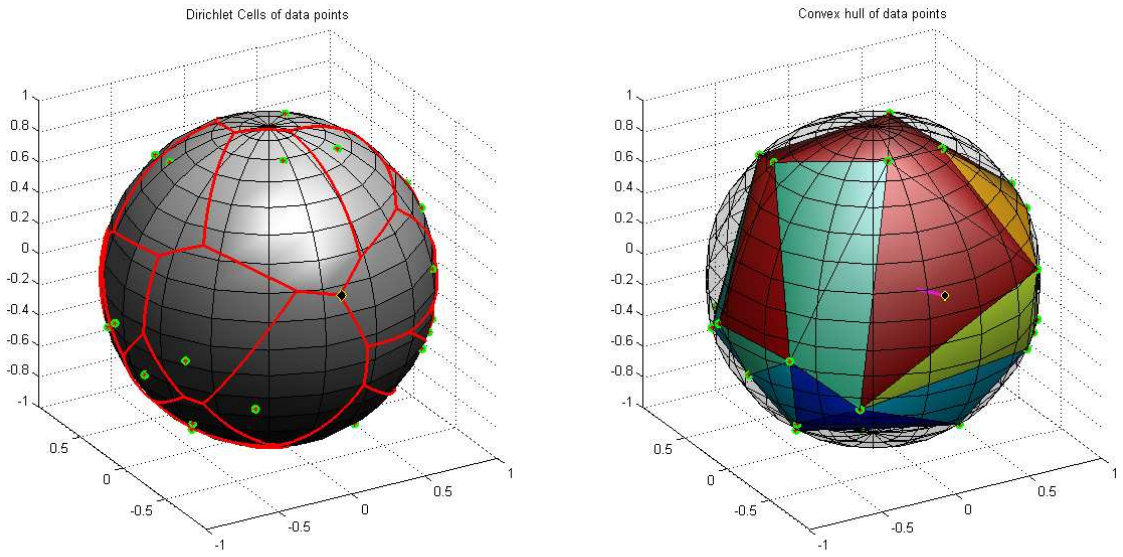


Figure 4.3: Dirichlet cells and convex hull generated by 25 data points (green), $\theta_1 = \ldots = \theta_m = \theta$ (black diamond) in the intersection of three Dirichlet cells

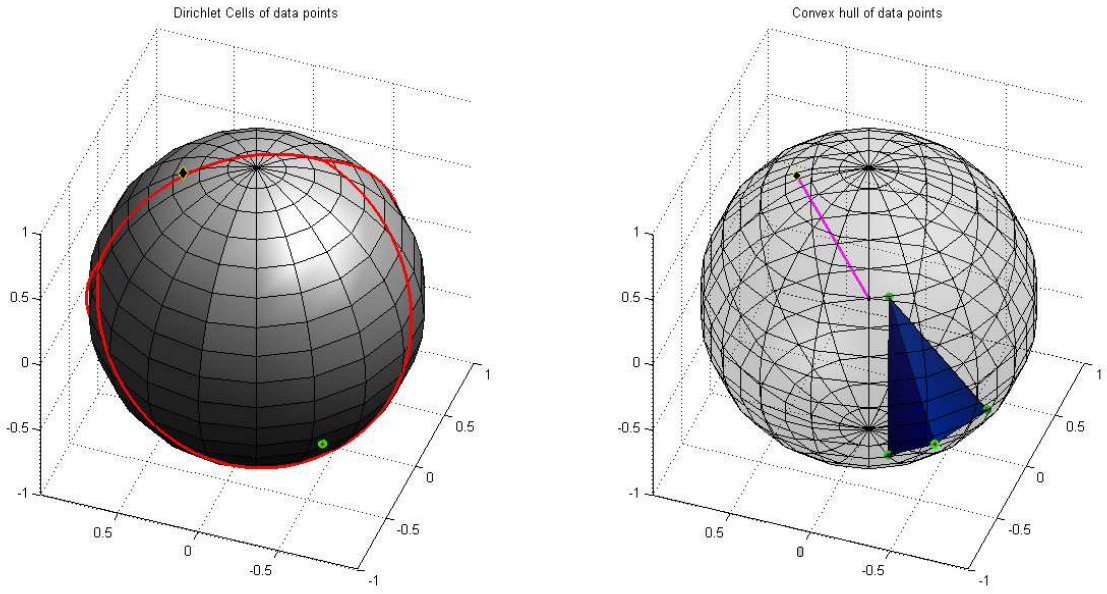


Figure 4.4: Dirichlet cells and convex hull generated by four data points (green), $\theta_1 = \ldots = \theta_m = \theta$ (black diamond) in the intersection of two Dirichlet cells

Chapter 5

The PFP from a Perspective in Nonlinear Programming

As mentioned, global minimization of the penalized frame potential is in general hard to achieve due to the presence of many saddle points. For the characterization of minima of the total frame potential as in Chapter 2, simple techniques from linear algebra are sufficient. On the other hand, in Chapter 4 we showed that the minimizers of the penalized frame potential in (3.12) for positive α are located in the interior of the spherical Dirichlet cells

$$D_j \quad = \quad \left\{ v \in \mathcal{S}^{d-1} \, : \, y_j = \underset{1 \leq k \leq N}{\arg\max} \, \langle \, y_k, v \, \rangle \right\}$$

generated by the given data Y. For $\alpha \to \infty$, the penalty term dominates the total frame potential and the minimum converges towards the most frame-like subfamily $\tilde{\Psi}$ of Y. Both proofs are rather based on arguments from calculus. The purpose of this chapter is to formulate the minimization in terms of optimization theory in order to qualify for the application of methods in this framework.

The outline of the chapter is as follows: we first deal with a formulation of the problem in the terminology of nonlinear programming. In section 5.1 we introduce alternative representations as well as relaxations in order to derive further properties of the penalized frame potential. The relaxations enable us to propose a heuristic method to approximate local minimizers with reasonable computational cost. Section 5.2 provides appropriate dual problems for the

nonlinear problems defined in Section 5.1. Furthermore, we show that under suitable mild relaxations, the minimizer can be computed analytically via the singular values and analyze the duality gaps between the given primal problems and the corresponding dualizations. Section 5.3 deals with the choice of data subsets. We provide an example implying that the optimal subset of the data which generates the Dirichlet cells containing the minimizers depends in general on the regularization parameter α. Section 5.4 reformulates the main problem into a polynomial problem based on a proposal by Lasserre ([Lass 12]). However, it can be seen easily that the number of variables increases heavily in the context of polynomial optimization and it becomes unfeasible to handle the computational effort properly. Finally, Section 5.5 relates the relaxed problems to other well-known problems in the literature.

Throughout this chapter, for ease of notation we concentrate on the case $m \geq d$ (especially in the context of singular values), remarking that most results also cover or are easily transferable to the case $m < d$. First of all, using the notation from Chapter 2, the main problem of minimizing the data-driven penalized frame potential reads as

$$\text{(P)} \quad \begin{cases} \min\limits_{T \in \mathbb{R}^{d\times m}} & \dfrac{d}{m^2} \left\|T^*T\right\|_F^2 + \alpha \left(m + 1 - \displaystyle\sum_{\ell=1}^{m} \max_{1\leq j\leq N} \langle\, y_j, \theta_\ell \,\rangle \right) \\ \text{s.t.} & \operatorname{diag}(T^*T) = (1, \ldots, 1)\,. \end{cases}$$

One of the major challenges for finding the solution stems from the fact that the unit sphere $\mathcal{S}^{d-1}$ is not convex in $\mathbb{R}^d$. This makes the feasible set $(\mathcal{S}^{d-1})^m = \mathcal{S}^{d-1} \times \ldots \times \mathcal{S}^{d-1}$ also a non-convex set. Hence, many classical techniques in nonlinear programming do not apply.

However, by the compactness of $\mathcal{S}^{d-1}$ and continuity of the objective function

$$\mathrm{F}(\Theta, Y) \;=\; \frac{d}{m^2} \left\|T^*T\right\|_F^2 + \alpha \left(m + 1 - \sum_{\ell=1}^{m} \max_{1\leq j\leq N} \langle\, y_j, \theta_\ell \,\rangle \right),$$

existence of extrema is guaranteed. In contrast to the main theorem on the total frame potential in [Bene 03], most extrema are local in the case $\alpha > 0$ as can be seen by simple examples. Figure 5.1 provides such an example with generic data for the case $\alpha = 5$, $d = m = 2$ and $N = 10$. Obviously, the influence of the data permits the existence of minima that are non-global.

A necessary condition for local minima of the penalized frame potential can be formulated again by a Lagrangian approach.

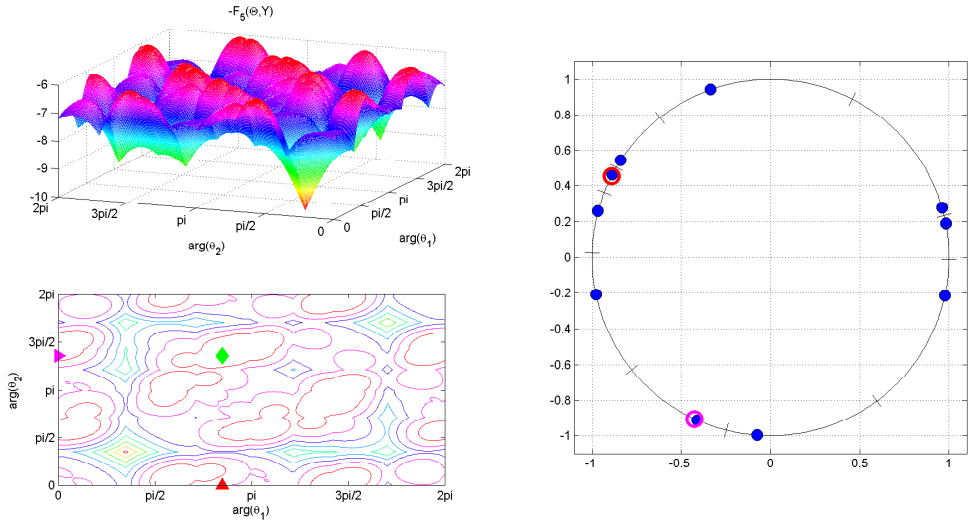


Figure 5.1: Upper left: negative function values of F for $\alpha = 5$ and data $y_1, \dots, y_{10} \in \mathcal{S}^1$; Lower left: contour plot of F and indicators for the arguments of the global minimum; Right: data vectors (blue), the (unique) global minimizers $\theta_1, \theta_2 \in \mathcal{S}^1$ (circles) and boundaries of the Dirichlet cells (dashes)

Theorem 5.1. *Let $Y = \{y_1, \dots, y_N\} \subset \mathcal{S}^{d-1}$ be a family of normalized data vectors. If $T_0 = [\theta_{1,0}, \dots, \theta_{m,0}]$ is a minimizer of* (P)*, there exist $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m) \in \mathbb{R}^{m \times m}$ and $Y_s = [y_{s(1)}, \dots, y_{s(m)}] \in \mathbb{R}^{d \times m}$ with $y_{s(\ell)} = \underset{j=1,\dots,N}{\arg\min} \|\theta_{\ell,0} - y_j\|^2$, $\ell = 1, \dots, m$, such that*

$$\frac{4d}{m^2} T_0 T_0^* T_0 + 2T_0\Lambda \;=\; \alpha Y_s \,. \tag{5.1}$$

Proof. For $\alpha = 0$, it holds that $S_0 = T_0 T_0^* = \frac{m}{d} I_d$ and therefore $\Lambda = \frac{-2}{m} I_m$ satisfies (5.1). If $\alpha > 0$, the matrix Y_s as defined above exists by Corollary 4.9. Analogously to Section 2.2, we define the Lagrange function

$$\mathcal{L}(T, \Lambda) \;=\; \frac{d}{m^2} \|T^*T\|_F^2 - \alpha \sum_{\ell=1}^{m} \max_{1 \le j \le N} \langle y_j, \theta_\ell \rangle + \sum_{k=1}^{m} \lambda_k (\|\theta_k\|^2 - 1) \,,$$

which can be expressed in T_0 according to Corollary 4.10 by the identity

$$\mathcal{L}(T_0, \Lambda) \;=\; \frac{d}{m^2} \|T_0^* T_0\|_F^2 - \alpha \sum_{\ell=1}^{m} \langle y_{s(\ell)}, \theta_{\ell,0} \rangle + \sum_{k=1}^{m} \lambda_k (\|\theta_{k,0}\|^2 - 1) \,. \tag{5.2}$$

Using the trace operator, (5.2) is equivalent to

$$\mathcal{L}(T_0, \Lambda) \;=\; \operatorname{trace}\left(\frac{d}{m^2} (T_0^* T_0)^2 - \alpha T_0^* Y_s + (T_0^* T_0 - I_m)\Lambda \right) . \tag{5.3}$$

Again, according to (2.12) and Lemma 4.7, differentiation leads to the extremal condition in (5.1). □

The $(m \times m)$-matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ consists of the Lagrange multipliers. Theorem 5.1 shows again that for $\alpha = 0$ it holds that $\lambda_1 = \ldots = \lambda_m = -\frac{2}{m}$. We call the matrix $T \in \mathbb{R}^{d \times m}$ α-critical if it satisfies (5.1).

Example 5.2. *Let $d = 2$, $m = N = 3$ and*

$$Y \;=\; \begin{bmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & -1/\sqrt{2} \end{bmatrix} \in \mathbb{R}^{d \times N}.$$

Considering the Lagrange function

$$\begin{aligned} \mathcal{L}(T, \Lambda) &= \|T^*T\|_F^2 + \alpha \|T - Y\|_F^2 + \sum_{k=1}^{m} \lambda_k (\|\theta_k\|^2 - 1) \\ &= \operatorname{trace}\left((T^*T)^2\right) + \alpha \|T - Y\|_F^2 + \operatorname{trace}\left((T^*T - I)\,\Lambda\right), \end{aligned} \tag{5.4}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$ contains the Lagrange multipliers, leads to the extremal condition

$$2TT^*T + T\Lambda \;=\; \alpha\,(Y - T)\,. \tag{5.5}$$

Note that the penalty term is replaced by $\|\cdot - Y\|_F^2$ which is due to $m = N$ and the equivalence of minimization in the norm and maximization of inner products. Moreover, this term can be expressed in terms of the trace operator using the identity $\langle A, B\rangle_F = \operatorname{trace}(B^*A)$ for $A, B \in \mathbb{R}^{d \times m}$ and $\langle A, A\rangle_F = \|A\|_F^2$.

For $\alpha = 0$, it is easy to see that $\eta_1, \eta_2, \eta_3 \in \mathbb{R}^2$ with

$$\eta_1 = \begin{bmatrix} \cos(-\pi/12) \\ \sin(-\pi/12) \end{bmatrix}, \quad \eta_2 = \begin{bmatrix} \cos(7\pi/12) \\ \sin(7\pi/12) \end{bmatrix} \quad \text{and} \quad \eta_3 = \begin{bmatrix} \cos(5\pi/4) \\ \sin(5\pi/4) \end{bmatrix}$$

constitute a FUNTF. Therefore, the minimizer of

$$\mathrm{F}(T) \;=\; \|T^*T\|_F^2 + \alpha \|T - Y\|_F^2$$

is given by $T_0 = [\eta_1, \eta_2, \eta_3]$ or, equivalently, with $c = \cos(\gamma)$ and $s = \sin(\gamma)$ for $\gamma = -\pi/12$ by

$$T_0 \;=\; \begin{bmatrix} c & s & -1/\sqrt{2} \\ s & c & -1/\sqrt{2} \end{bmatrix}. \tag{5.6}$$

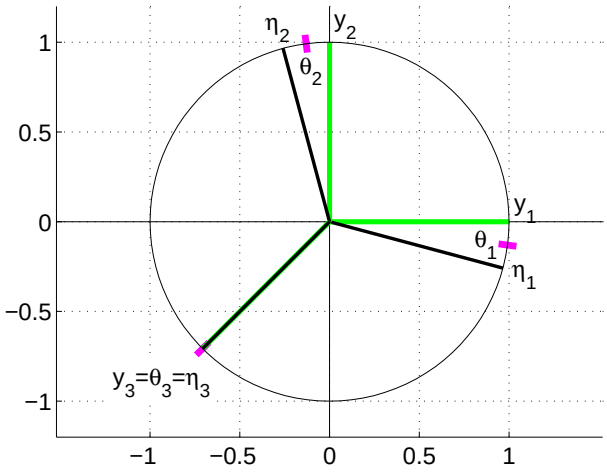


Figure 5.2: Data Y from Example 5.2 (green) and FUNTF η_1, η_2, η_3 (black)

On the other hand, for $\alpha \to \infty$, the matrix $T_0 = Y$ minimizes F*, i.e. T_0 takes again the form in* (5.6) *where $\gamma = 0$.*

Let T_0 from (5.6) *with $\gamma \in I = [-\pi/12, 0]$. Then the first two columns in* (5.5) *read as*

$$\begin{bmatrix} 3c + (4cs+1)s \\ 3s + (4cs+1)c \end{bmatrix} + \lambda_1 \begin{bmatrix} c \\ s \end{bmatrix} = \alpha \begin{bmatrix} 1-c \\ -s \end{bmatrix} \quad \textit{and}$$
$$\begin{bmatrix} 3s + (4cs+1)c \\ 3c + (4cs+1)s \end{bmatrix} + \lambda_2 \begin{bmatrix} s \\ c \end{bmatrix} = \alpha \begin{bmatrix} -s \\ 1-c \end{bmatrix},$$

implying that $\lambda_1 = \lambda_2$ is feasible. The corresponding system of linear equations

$$\begin{bmatrix} 1-c & -c \\ -s & -s \end{bmatrix} \begin{bmatrix} \alpha \\ 3+\lambda_1 \end{bmatrix} = (4cs+1) \begin{bmatrix} s \\ c \end{bmatrix}$$

is solved by $\lambda_1 = (4c+s^{-1})(2c^2-c-1)-3$ and $\alpha = (4c+s^{-1})(1-2c^2)$. Note that $\lambda_1(-\pi/12) = -3$, $\lim_{\gamma\to 0}\lambda_1(\gamma) = -3$ and $\lambda_1(\gamma) \in [-3, -2.9)$ for $\gamma \in I$. Furthermore, the third column in (5.5) *leads to $\lambda_3 = -4(1+cs) = -s(4c+s^{-1}) - 3$, which is monotonously decreasing on I with $\lambda_3(-\pi/12) = -3$ and $\lambda_3(0) = -4$ (see Figure 5.3).*

Hence, for fixed $\gamma \in I$ we find a matrix Λ and $\alpha \geq 0$ such that T_0 from (5.6) *satisfies the extremal condition in* (5.5). *Since the constraints $\|\theta_\ell\|^2 = 1$ are satisfied for $\ell = 1, 2, 3$, the*

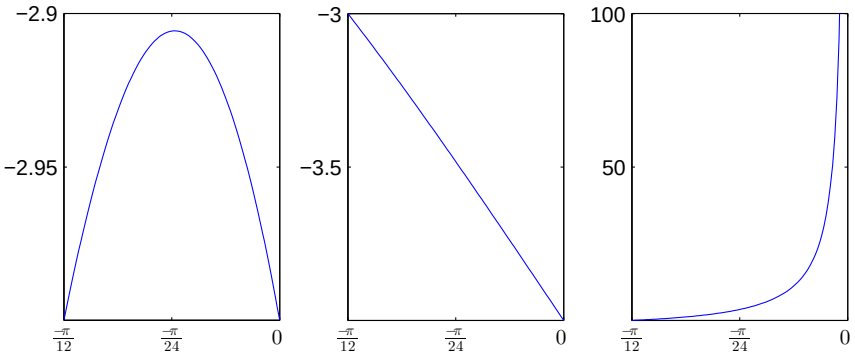


Figure 5.3: $\lambda_1 = \lambda_2$ (left), λ_3 (middle) and α (right) from Example 5.2 for $\gamma \in [-\pi/12, 0]$

function value $\mathcal{L}(T_0, \Lambda)$ *of the Lagrange function equals*

$$\begin{aligned} \mathrm{F}(T_0) &= \frac{9}{2} + 2\left(2cs + \frac{1}{2}\right)^2 + 2\alpha\left((c-1)^2 + s^2\right) \\ &= 5 + 4cs + 8(cs)^2 + 4(4c + s^{-1})(1 - 2c^2)(1 - c)\,. \end{aligned}$$

Note that for $x \geq 0$ *and fixed* $\gamma \in I$ *it holds that* $\mathcal{L}(T_{0,(\cdot)}, \Lambda) : \mathbb{R}_+ \to \mathbb{R}_+$ *with*

$$\mathcal{L}(T_{0,x}, \Lambda) = \|T_{0,x}^* T_{0,x}\|_F^2 + \alpha\|T_{0,x} - Y\|_F^2 + 2\lambda_1(x^2 - 1)$$

and

$$T_{0,x} = \begin{bmatrix} cx & sx & -1/\sqrt{2} \\ sx & cx & -1/\sqrt{2} \end{bmatrix}$$

can be identified by a quartic polynomial $\varphi_\gamma = \mathcal{L}(T_{0,(\cdot)}, \Lambda)$ *with the unique minimum in* $x = 1$ *(see Figure 5.4).*

In the example, the Lagrange multipliers λ_ℓ are in $[-4, -2.9)$ for $\ell = 1, \ldots, m$. However, using the objective function from the main problem (P) leads to the extremal condition in (5.1), which is different from the condition in (5.5). Furthermore, the Lagrange multipliers are generally not bounded as the following Corollary shows.

Corollary 5.3. *The Lagrange multipliers from Theorem 5.1 satisfy* $\lambda_\ell \to \infty$ *for* $\alpha \to \infty$ *and* $\ell = 1, \ldots, m$.

Proof. Let T satisfy (5.1). Using m vector-valued equations instead of one single matrix-valued equation, the necessary condition reads as

$$\frac{4d}{m^2} TT^* \theta_\ell + 2\lambda_\ell \theta_\ell = \alpha\, y_{s(\ell)}\,, \quad \ell = 1, \ldots, m.$$

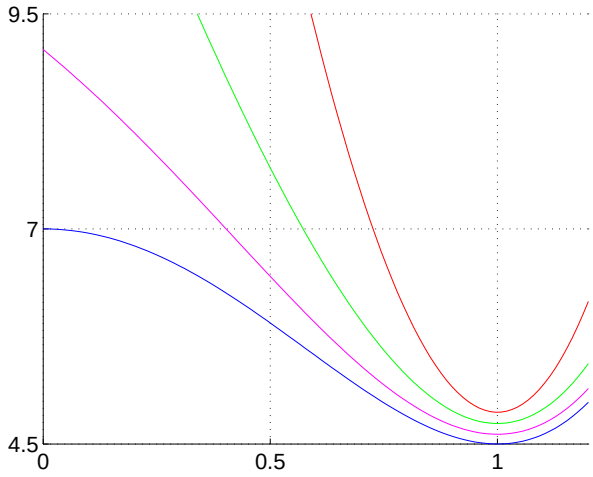


Figure 5.4: Polynomials φ_γ from Example 5.2 for $\gamma = \frac{-\pi}{12}$ (blue), $\frac{-\pi}{16}$ (magenta), $\frac{-\pi}{24}$ (green) and $\frac{-\pi}{48}$ (red)

Multiplication on both sides by θ_ℓ^* shows that

$$\frac{4d}{m^2}\,\|T^*\theta_\ell\|^2 + 2\lambda_\ell = \alpha\,\langle y_{s(\ell)}, \theta_\ell\rangle = \alpha \max_{j=1,\dots,N} \langle y_j, \theta_\ell\rangle\ .$$

Now by $\max_{j=1,\dots,N} \langle y_j, \theta_\ell\rangle \to 1$ for $\alpha \to \infty$ and $\|T^*\theta_\ell\|^2 \le m$ is bounded, the proof is complete. □

5.1 Relaxations of the Main Problem

Since constants do not affect the minimization process, and by using the polarization identity or (3.13), we simply formulate the main primal problem as

$$\text{(P1)} \quad \begin{cases} \min\limits_{T\in\mathbb{R}^{d\times m}} & \|T^*T\|_F^2 + \alpha \sum\limits_{\ell=1}^{m} \min\limits_{1\le j\le N} \|\theta_\ell - y_j\|^2 \\ \text{s.t.} & \operatorname{diag}(T^*T) = (1,\dots,1)\,. \end{cases}$$

Note that the equality constraints

$$\begin{aligned} g_k : \quad \mathbb{R}^{d\times m} &\to \mathbb{R} \\ T = [\theta_1,\dots,\theta_m] &\mapsto g_k(T) = \|\theta_k\|^2 - 1\,, \quad k = 1,\dots,m, \end{aligned}$$

are continuously differentiable in T. By writing $g_k(T) = \theta_k^*\theta_k - 1$, it is easy to see that the gradients $\nabla g_1(T),\dots,\nabla g_m(T)$ are linearly independent for all feasible T. Furthermore, note

that (P1) does not contain any inequality constraints, which reduces the Karush-Kuhn-Tucker conditions to the classical Lagrange approach.

One of the major challenges in solving (P1) arises from the min-term in the sum in the objective function. According to Theorem 2.7, the solutions for $\alpha = 0$ are exactly the FUNTFs. On the other hand, Theorem 4.1 shows that for $\alpha \to \infty$ the solution becomes

$$\tilde{\Psi} \quad = \quad \operatorname*{arg\,min}_{\Psi \in Y^m} \operatorname{TFP}(\Psi)\,. \tag{5.7}$$

In order to simplify the problem (P1), we first consider a fixed data matrix $Y_s \in \mathbb{R}^{d\times m}$ with normalized columns, i.e. $y_j \in \mathcal{S}^{d-1}$ for all j, and propose the alternative problem

$$\text{(P2)} \quad \begin{cases} \min\limits_{T\in\mathbb{R}^{d\times m}} & \|T^*T\|_F^2 + \alpha\,\|T - Y_s\|_F^2 \\ \text{s.t.} & \operatorname{diag}(T^*T) = (1,\ldots,1)\,, \end{cases}$$

where we also used the fact that

$$\|T - Y_s\|_F^2 \quad = \quad \sum_{k=1}^{m} \|\theta_k - y_{s(k)}\|^2\,.$$

Note that there exists $\alpha_0 > 0$ such that the columns of the optimal Y_s contain the vectors of the family $\tilde{\Psi}$ for all $\alpha > \alpha_0$, which makes (P2) equivalent to (P1) for α sufficiently large.

Furthermore, in order to admit even more freedom on the constraints, we replace the requirement that $\theta_\ell \in \mathcal{S}^{d-1}$ for $\ell = 1,\ldots,m$ in (P1) and (P2) by the weaker version $\operatorname{trace}(T^*T) = m$ and introduce the two relaxed problems

$$\text{(P1}*\text{)} \quad \begin{cases} \min\limits_{T\in\mathbb{R}^{d\times m}} & \|T^*T\|_F^2 + \alpha \sum\limits_{\ell=1}^{m} \min\limits_{1\le j\le N} \|\theta_\ell - y_j\|^2 \\ \text{s.t.} & \operatorname{trace}(T^*T) = m\,, \end{cases}$$

and

$$\text{(P2}*\text{)} \quad \begin{cases} \min\limits_{T\in\mathbb{R}^{d\times m}} & \|T^*T\|_F^2 + \alpha\,\|T - Y_s\|_F^2 \\ \text{s.t.} & \operatorname{trace}(T^*T) = m\,. \end{cases}$$

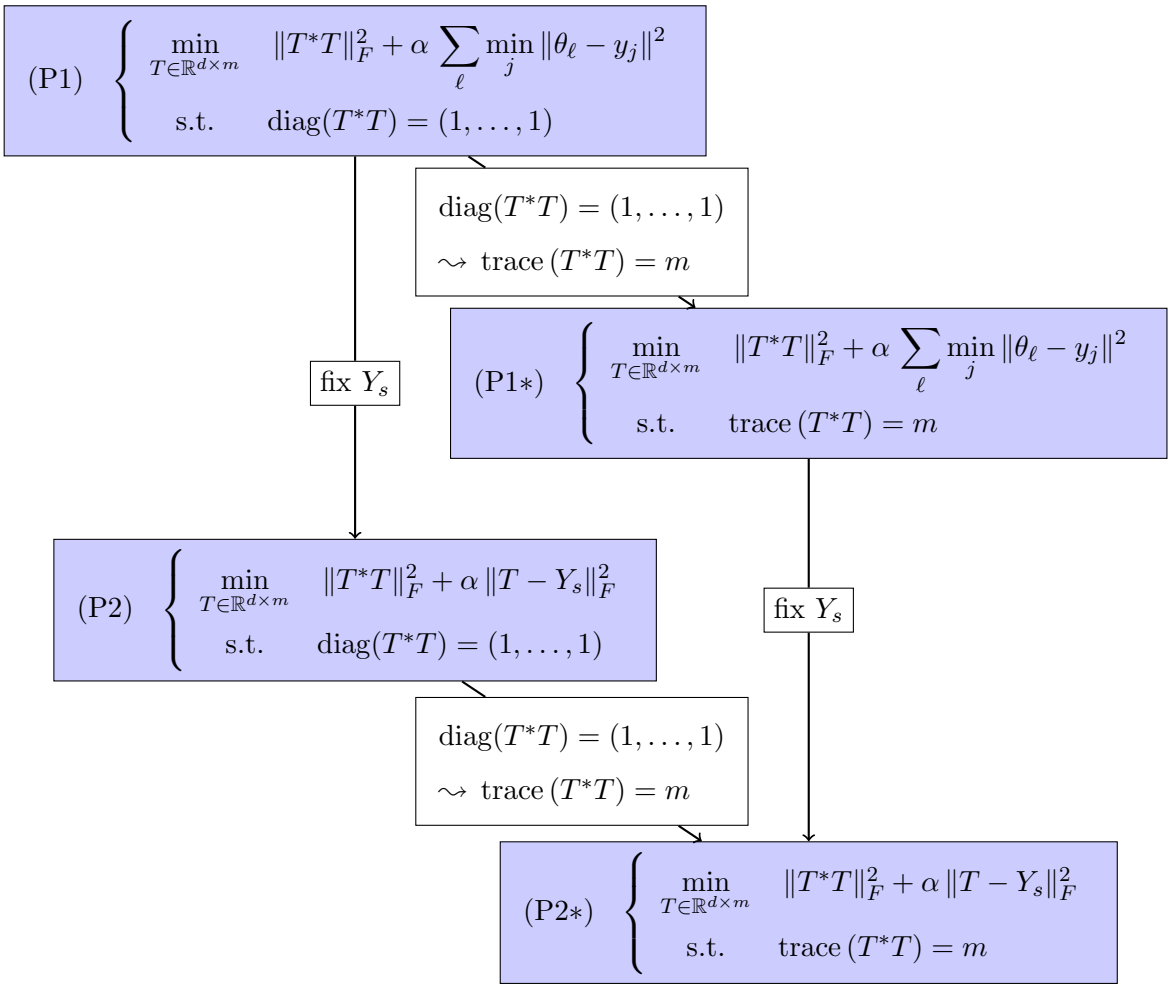


Figure 5.5: Primal problems (P1) and (P2) with relaxations including their respective relations

The diagram in Figure 5.5 visualizes the relations between the four stated optimization problems.

As we will see, (P2∗) is much easier to solve than the first problem (P1) is. Furthermore, in Section 5.2 we characterize the solutions of (P2∗) in terms of the singular values. There is also a strong connection between the dualizations of (P2∗) and (P2). We will use the primal problems in order to formulate an algorithm in order to compute an approximate solution of (P1).

The following theorem is known as the Wielandt-Hoffman-Theorem for singular values and works as a key ingredient for the characterization of minima of (P2∗). The other central aspect will be that the problem (P2∗) can be separated into two consecutive minimization

steps. Further versions of the Wielandt-Hoffman-Theorem exist in the literature, e.g. for the eigenvalues of normal or Hermitian matrices ([Horn 96] Theorem 6.3.5 and Corollary 6.3.8).

Theorem 5.4 (Wielandt-Hoffman, [Horn 96] Corollary 7.3.8)**.** *Let $d, m \in \mathbb{N}$, $d \leq m$, and $A, B \in \mathbb{C}^{d\times m}$ with singular values $\sigma_1^{(A)} \geq \ldots \geq \sigma_d^{(A)} \geq 0$ and $\sigma_1^{(B)} \geq \ldots \geq \sigma_d^{(B)} \geq 0$, respectively. Then*

$$\sum_{j=1}^{d} \left|\sigma_j^{(A)} - \sigma_j^{(B)}\right|^2 \quad \leq \quad \|A - B\|_F^2 \,. \tag{5.8}$$

Since the Frobenius- (or Hilbert-Schmidt-) Norm is invariant under orthogonal transformations, equality in (5.8) holds if $U_A = U_B$ and $V_A = V_B$ where $A = U_A \Sigma_A V_A^*$ and $B = U_B \Sigma_B V_B^*$ are the SVDs of A and B, respectively.

For the following lemma, define for $\lambda \in \mathbb{R}$ the cubic polynomial $p_\lambda : \mathbb{R} \to \mathbb{R}$ by

$$p_\lambda(x) \quad = \quad x^3 + \lambda x \,, \qquad x \in \mathbb{R}, \tag{5.9}$$

and denote $\xi(\lambda)$ as the largest real root of p_λ. More precisely,

$$\xi(\lambda) \quad = \quad \begin{cases} 0, & \lambda \geq 0 \,, \\ \sqrt{-\lambda}, & \lambda < 0 \,. \end{cases} \tag{5.10}$$

Lemma 5.5. *Given $m > 0$ and $c_1 \geq c_2 \geq \ldots \geq c_d \geq 0$ there exist $\lambda \in \mathbb{R}$ and $x_1 \geq x_2 \geq \ldots \geq x_d \geq \xi(\lambda)$ solving the interpolation problem $p_\lambda(x_j) = c_j$ and $\sum_{j=1}^d x_j^2 = m$ with p_λ and $\xi(\lambda)$ from* (5.9) *and* (5.10)*, respectively.*

Moreover, if $c_d > 0$, the sequence $x_1, \ldots, x_d$ and the parameter λ are unique.

Proof. For $\lambda \in \mathbb{R}$, the cubic polynomial p_λ from (5.9) satisfies $p_\lambda(0) = 0$ and $p'_\lambda(0) = \lambda$. Furthermore, p_λ is strictly increasing on $[\xi(\lambda), \infty)$. By the intermediate value theorem, there exists a non-increasing sequence of real numbers $x_1(\lambda) \geq \ldots \geq x_d(\lambda) \geq \xi(\lambda)$ with $p_\lambda(x_j(\lambda)) = c_j$. Note that $x_j(\lambda) = \xi(\lambda)$ if and only if $c_j = 0$.

For $\lambda > 0$, it holds that $\xi(\lambda) = 0$ and $p'_\lambda(0) \to \infty$ as $\lambda \to \infty$. It follows that $\sum_{j=1}^d x_j(\lambda)^2 \to 0$ since $x_j(\lambda) \to 0$, $j = 1, \ldots, d$. On the other hand, we have $\xi(\lambda) \to \infty$ for $\lambda \to -\infty$, which implies $x_j(\lambda) \to \infty$ and therefore $\sum_{j=1}^d x_j(\lambda)^2 \to \infty$. Again, by the intermediate value theorem, $\lambda \in \mathbb{R}$ with $p_\lambda(x_j) = c_j$ and $\sum_{j=1}^d x_j^2 = m$ exists where $x_j = x_j(\lambda)$.

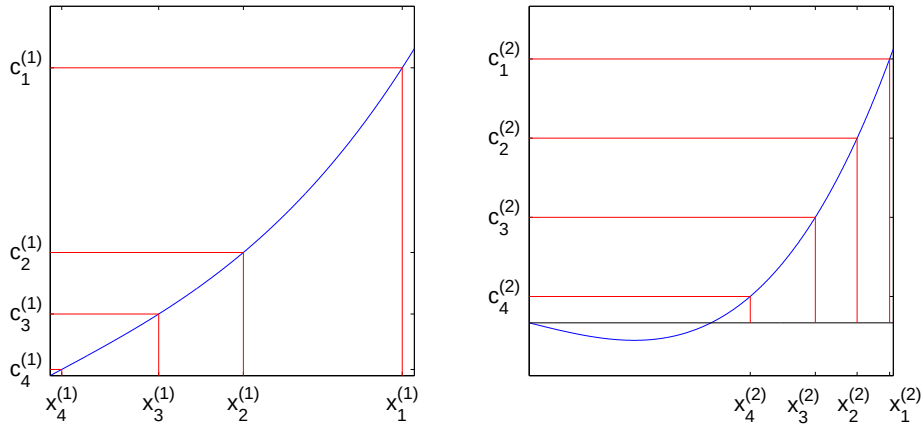


Figure 5.6: Interpolating polynomials $p_\lambda : \mathbb{R} \to \mathbb{R}$, $p_\lambda(x) = x^3 + \lambda x$, with $p_\lambda(x_j^{(r)}) = c_j^{(r)}$ from Lemma 5.5 for two pairwise distinct sequences $c_1^{(r)} \geq \ldots \geq c_d^{(r)} > 0$, $\lambda = \lambda^{(r)}$ for $r = 1, 2$; note that $\xi(\lambda^{(1)}) = 0$ in (5.10) and $\lambda^{(1)} > 0$ (left) whereas $\xi(\lambda^{(2)}) > 0$ and $\lambda^{(2)} < 0$ (right)

In order to show that the solution is unique for $c_1 \geq \ldots \geq c_d > 0$, suppose there exist two distinct solutions $(x, \lambda)^T \in \mathbb{R}^{d+1}$ with $x_1 \geq \ldots \geq x_d > 0$ and $(z, \gamma)^T \in \mathbb{R}^{d+1}$ with $z_1 \geq \ldots \geq z_d > 0$ where $c_j = p_\lambda(x_j) = p_\gamma(z_j)$ for $j = 1, \ldots, d$. Since $x_\nu > z_\nu$ for some $\nu \in \{1, \ldots, d\}$ and $m = \|x\|^2 = \|z\|^2$, there exists $\mu \neq \nu$ with $x_\mu < z_\mu$. By $x_\nu^3 > z_\nu^3$ and

$$0 < c_\nu = x_\nu^3 + \lambda x_\nu = z_\nu^3 + \gamma z_\nu \,,$$

it follows that $\lambda < \gamma$. However, in that case,

$$c_\mu = x_\mu^3 + \lambda x_\mu < z_\mu^3 + \gamma z_\mu = c_\mu \,,$$

which is a contradiction. □

It follows from the proof that the solution for the interpolation problem in Lemma 5.5 can be computed by solving the system of $d + 1$ nonlinear equations

$$\begin{aligned} x_j^3 + \lambda x_j &= c_j, \quad j = 1, \ldots, d \\ \sum_{j=1}^{d} x_j^2 &= m, \end{aligned}$$

containing the $d + 1$ variables $\lambda, x_1, \ldots, x_d$. For numerical computations, Newton's method can be applied, which has also been used for the solutions of the two different interpolation problems in Figure 5.6.

Lemma 5.6 ([Herm 00], Theorem 7.3)**.** *Let $x, y, z \in \mathbb{R}^d$ with $z_j = y_{\tau(j)}$, $j = 1, \ldots, d$, where τ is a permutation of $\{1, \ldots, d\}$. Then $\langle x, z\rangle$ is maximized (over all permutations of $\{1, \ldots, d\}$), if x and z have the same ordering, i.e.*

$$(x_j - x_k)(z_j - z_k) \;\geq\; 0\,, \qquad j, k = 1, \ldots, d\,.$$

Lemma 5.7. *Let $x, y \in \mathbb{R}^d$ with $x_1 \geq \ldots \geq x_d$ and $y_1 \geq \ldots \geq y_d$. Furthermore, let $z_j = y_{\tau(j)}$, $j = 1, \ldots, d$, where τ is a permutation of $\{1, \ldots, d\}$. Then $\|x - z\|_2^2$ is minimized in the sense of Lemma 5.6, if $\tau = \mathrm{id}$ or, equivalently, $z_j = y_j$, $j = 1, \ldots, d$.*

Proof. The statement follows from $\|x - z\|_2^2 = \|x\|_2^2 + \|z\|_2^2 - 2\langle x, z\rangle$ and Lemma 5.6. □

Accordingly, by the Wielandt-Hoffman-Theorem (Theorem 5.4) and Lemma 5.5, the solution of (P2∗) can be characterized in the following way.

Theorem 5.8. *Let $\alpha > 0$ and let $Y_s = \hat{U}\hat{\Sigma}\hat{V}^*$ denote the SVD of $Y_s \in \mathbb{R}^{d\times m}$. Then there exists $\lambda \in \mathbb{R}$ and a matrix of singular values $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d\times m}$ with $\sigma_1 \geq \ldots \sigma_d \geq 0$ satisfying*

$$\begin{aligned} \sigma_j^3 + \frac{\alpha + \lambda}{2}\sigma_j &= \frac{\alpha}{2}\hat{\sigma}_j\,, \quad j = 1, \ldots, d, \\ \sum_{j=1}^{d} \sigma_j^2 &= m\,, \end{aligned} \tag{5.11}$$

such that $T_0 = \hat{U}\Sigma\hat{V}^*$ minimizes (P2∗)*. Furthermore, if* $\mathrm{rank}(Y_s) = d$*, i.e. $\hat{\sigma}_d > 0$, the matrix Σ is uniquely determined and it holds that* $\mathrm{rank}(T_0) = d$*.*

Proof. Firstly, the constraint in (P2∗) can be formulated in terms of the singular values as $m = \mathrm{trace}\,(T^*T) = \|T\|_F^2 = \sum_j \sigma_j^2$. Furthermore, the objective function is equal to

$$\|\Sigma^*\Sigma\|_F^2 + \alpha\,\|U\Sigma V^* - \hat{U}\hat{\Sigma}\hat{V}^*\|_F^2\,,$$

where the regularization term $\|\Sigma^*\Sigma\|_F^2$ is also independent of U and V. Now the choice of $U = \hat{U}$ and $V = \hat{V}$ for the minimizer follows from the Wielandt-Hoffman-Theorem 5.4.

Thus, the problem can be reduced to a problem on the singular values

$$\begin{cases} \min\limits_{\sigma_1 \geq \ldots \geq \sigma_d \geq 0} & \sum\limits_{j=1}^{d} \sigma_j^4 + \alpha\,(\sigma_j - \hat{\sigma}_j)^2 \\ \text{s.t.} & \sum\limits_{\nu=1}^{d} \sigma_\nu^2 = m\,, \end{cases} \tag{5.12}$$

where the objective function equals $\|\Sigma^*\Sigma\|_F^2+\alpha\|\Sigma-\hat{\Sigma}\|_F^2$. Using the single-constraint Lagrange function

$$\hat{\mathcal{L}}(\Sigma,\lambda) \;=\; \sum_{j=1}^{d}\sigma_j^4+\alpha\,(\sigma_j-\hat{\sigma}_j)^2\;+\;\lambda\left(\sum_{\nu=1}^{d}\sigma_\nu^2-m\right), \tag{5.13}$$

which satisfies

$$\nabla\hat{\mathcal{L}}(\Sigma,\lambda) \;=\; \left((4\sigma_j^3+2\alpha(\sigma_j-\hat{\sigma}_j)+2\lambda\sigma_j)_{j=1,\dots,d}\,,\,\sum_{\nu=1}^{d}\sigma_\nu^2-m\right),$$

it holds that (5.11) is necessary for a critical point. Since $\sum_j\sigma_j^2=m$ describes the compact sphere

$$B_{\sqrt{m}}(0) \;=\; \{v\in\mathbb{R}^d\;|\;\|v\|=\sqrt{m}\}\subset\mathbb{R}^d,$$

the continuous objective function in (5.12) takes its minimum and its maximum under the constraint. Suppose, the global minimizer $(\xi_1,\dots,\xi_d)^T\in B_{\sqrt{m}}(0)$ lies in an orthant other than the one described by the cone

$$\mathbb{R}^d_+ \;=\; \{v\in\mathbb{R}^d\;|\;v_j\geq 0,\,j=1,\dots,d\}. \tag{5.14}$$

Then the objective value in (5.12) can be reduced by taking $(|\xi_1|,\dots,|\xi_d|)^T$ without violating the constraint. Therefore the global minimizer of the objective function has to be in $\mathbb{R}^d_+$. Furthermore, it holds that the coordinates of the minimizer have to be in non-increasing order, otherwise rearranging would also decrease the function value in (5.12) by Lemma 5.7 since $\hat{\sigma}_1\geq\ldots\geq\hat{\sigma}_d\geq 0$.

Now, if $\hat{\sigma}_d>0$, according to Lemma 5.5 there exists unique $\lambda\in\mathbb{R}$ such that the system of nonlinear equations in (5.11) has unique solutions $\sigma_1\geq\ldots\geq\sigma_d>0$ satisfying $\sum_j\sigma_j^2=m$. □

Remark 5.9. By application of the Wielandt-Hoffman-Theorem 5.4, the relaxation (P2∗) can be regarded as a separation of the minimization over Σ and V. In general, this separation is not permissible for the minimization of (P) and (P1), respectively.

Note that by letting $\alpha\to 0$ in (P2∗), the singular values will satisfy $\sigma_j\to\sqrt{m/d}$ for $j=1,\dots,d$. Since the columns of V^* will in general not provide that ΣV^* has columns with norm equal to one, the columns of $T_0=[\tilde{\theta}_1^{(0)},\dots,\tilde{\theta}_m^{(0)}]$ do not constitute a FUNTF. However, it is easy to see by

$$\frac{m}{d}\|y\|^2=\langle T_0T_0^*y,y\rangle=\sum_{k=1}^{m}\left|\left\langle y,\tilde{\theta}_k^{(0)}\right\rangle\right|^2\quad\forall\,y\in\mathbb{R}^d,$$

that T_0 constitutes at least a tight frame in the case of $\alpha = 0$. On the other hand, (5.12) again underlines that T_0 becomes Y_s as $\alpha \to \infty$. $\triangle$

Remark 5.10. (1) In personal communication, Shen proposed to solve problem (P1) by an alternating method ([Shen 13]). For fixed $\alpha > 0$, start with a matrix $Y_s^{(0)}$ containing m vectors from the given data Y. Now, minimization of (P2) with $Y_s = Y_s^{(0)}$ leads to a matrix $T_0^{(0)}$ whose column vectors are located in the Dirichlet cells of some subfamily Ψ_1 of the data. Compute this family, let $Y_s^{(1)}$ be generated by arranging the elements of Ψ_1 as its columns and solve (P2) computationally for a new minimizer $T_0^{(1)}$ using $Y_s = Y_s^{(1)}$. Iterative application of these steps leads to an approximate minimizer of (P1). However, convergence of this method towards a minimum of (P1) is not easy to verify. Stop the algorithm, when $Y_s^{(k)}$ equals $Y_s^{(k-1)}$ for the first time during the iteration. Note that the subfamilies Ψ_k need not necessarily contain m distinct vectors and therefore the data matrices are allowed to contain multiple columns.

(2) For a direct implementation of the problem (P1), it is possible to use path following strategies ([Meye 12]). Begin with large $\alpha^{(0)} \gg 0$ and let the starting configuration Y_s for the optimizer of (P1) be the matrix containing the family $\tilde{\Psi}$ from (5.7). Then compute the minimizer $\tilde{T}_0^{(0)}$, replace $\alpha^{(0)}$ by $\alpha^{(1)} < \alpha^{(0)}$ and start the optimizer again with starting configuration $\tilde{T}_0^{(0)}$ in order to get a new minimizer $\tilde{T}_0^{(1)}$. Repeat this step until $\alpha^{(k)} = \alpha$ and set $T_0 = \tilde{T}_0^{(k)}$.

(3) In practice, the requirement of the existence of a full-rank matrix Y_s can always be guaranteed. If the data was located in a subspace of $\mathbb{R}^d$ with dimension less than d, dimension reduction methods could be applied in advance. $\triangle$

As mentioned previously, by using the primal problems (P2$*$) and (P2), we are able to formulate a heuristic method in order to approximate a solution of the main problem (P1).

Algorithm 5.11. Let $d, m \in \mathbb{N}$ and a family $Y = \{y_1, \ldots, y_N\} \subset \mathbb{R}^{d \times N}$ with $y_j \in \mathcal{S}^{d-1}$, $j = 1, \ldots, N$, where $N \gg m > d$.

1. Start with $Y_s^{(0)} = [y_{s(1)}^{(0)}, \ldots, y_{s(m)}^{(0)}] \in \mathbb{R}^{d \times m}$ where the vectors $y_{s(\ell)}^{(0)}$ are also in Y for $\ell = 1, \ldots, m$ and where the frame potential $\|Y_s^{(0)*} Y_s^{(0)}\|_F^2$ is "small".
   Let $q = 0$.

2. Compute the SVD of $Y_s^{(q)}$, i.e. $Y_s^{(q)} = \hat{U}^{(q)}\hat{\Sigma}^{(q)}\hat{V}^{(q)*}$.

3. Solve the system of nonlinear equations

$$\begin{aligned} \sigma_j^3 + \frac{\lambda+\alpha}{2}\sigma_j &= \frac{\alpha}{2}\hat{\sigma}_j^{(q)}, \quad j = 1, \dots, d, \\ \sum_{j=1}^{d} \sigma_j^2 &= m \end{aligned}$$

   by Newton's method in order to get the solution $\sigma_{1,q} \geq \ldots \geq \sigma_{d,q} \geq 0$ and $\lambda_q \in \mathbb{R}$. Denote $T_0^{(q)} = \hat{U}^{(q)}\Sigma_q\hat{V}^{(q)*} = [\tilde{\theta}_1^{(q)}, \dots, \tilde{\theta}_m^{(q)}]$ as the solution of (P2$*$) with $Y_s = Y_s^{(q)}$.

4. Compute $T_1^{(q)} = [\theta_1^{(q)}, \dots, \theta_m^{(q)}]$ where $\theta_k^{(q)} = \dfrac{1}{\|\tilde{\theta}_k^{(q)}\|}\,\tilde{\theta}_k^{(q)}$, $k = 1, \dots, m$.

5. Compute $Y_s^{(q+1)} = [y_{s(1)}^{(q+1)}, \dots, y_{s(m)}^{(q+1)}] \in \mathbb{R}^{d\times m}$ with $y_{s(k)}^{(q+1)} = \underset{1\leq j\leq N}{\arg\min} \|\theta_k^{(q)} - y_j\|^2$.

6. If $Y_s^{(q)} = Y_s^{(q+1)}$, then stop and return $T_1^{(q)}$. Otherwise increment q and go back to 2.

The output $T_1^{(q)}$ is an approximate solution of (P1). The algorithm computes the solution of (P2$*$), normalizes it to $\mathcal{S}^{d-1}$ and updates the optimal data. The loop structure corresponds with the alternating method mentioned in Remark 5.10 in order to solve the main problem (P1).

Remark 5.12. In several fields of application it has become a common technique to minimize combinations of a data-dependent loss or penalty term and a regularization term. Several examples including a classification of the problems stated in the current section will be given in Section 5.5. Note that the stated optimization problems (P1), (P2), (P1$*$) and (P2$*$) are formulated in this manner, all consisting of the total frame potential as regularization term and differing in their penalty terms. In the standard form, the regularization term gets controlled by a parameter $\alpha \geq 0$ which is followed by the summation with the unscaled penalty term. Obviously, controlling the frame potential in our formulations by the adjusted parameter $1/\alpha$ for $\alpha > 0$ leads to the same results as before. However, we want to preserve the relation to the FUNTFs. This goal is obtained by weighting the penalty terms instead of the regularization term which is the reason for the formulations we chose. △

5.2 Dualizations

In this section, we formulate appropriate dualizations for the problems given in Section 5.1 in order to establish further properties. As seen before, for positive α, the problem (P2$*$) has a solution of the form $T_0 = \hat{U}\Sigma\hat{V}^*$ where $Y_s = \hat{U}\hat{\Sigma}\hat{V}^*$ and $\operatorname{rank}(T_0) \geq \operatorname{rank}(Y_s)$.

The standard form of a nonlinear problem over a set X is

$$\text{(P0)} \quad \begin{cases} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & g_k(x) = 0, \qquad k = 1,\dots,m\,, \\ & h_\ell(x) \leq 0, \qquad \ell = 1,\dots,n\,, \end{cases}$$

where the functions $g_1,\dots,g_m : X \to \mathbb{R}$ describe the equality constraints and, accordingly, the functions $h_1,\dots,h_n : X \to \mathbb{R}$ stand for the inequality constraints. For ease of notation, we simply write $g : X \to \mathbb{R}^m$, $g(x) = (g_1(x),\dots,g_m(x))^T$ and $h : X \to \mathbb{R}^n$, $h(x) = (h_1(x),\dots,h_n(x))^T$. Following [Baza 06], the most common dualization for the primal problem (P0) is the corresponding Lagrangian dual problem, which is given by

$$\text{(D0)} \quad \begin{cases} \sup\limits_{\lambda\in\mathbb{R}^m,\,\mu\in\mathbb{R}^n} & \mathcal{L}(\lambda,\mu) \\ \text{s.t.} & \mu_1,\dots,\mu_n \geq 0 \end{cases}$$

with the Lagrangian dual function

$$\mathcal{L}(\lambda,\mu) \;=\; \inf_{x\in X}\; f(x) + \sum_{k=1}^{m} \lambda_k g_k(x) + \sum_{\ell=1}^{n} \mu_\ell h_\ell(x)\,. \tag{5.15}$$

Note that there is no restriction on the signs of λ whereas μ has to be nonnegative, i.e. $\mu \in \mathbb{R}^n_+$ with $\mathbb{R}^n_+$ as in (5.14). The following property of the Lagrangian dual is quite easy to verify. Nevertheless, it implies that the dual can at most have one maximum.

Proposition 5.13 ([Baza 06] Theorem 6.3.1)**.** *Let $X \neq \emptyset$ be compact in $\mathbb{R}^d$ and let $g : X \to \mathbb{R}$ and $h : X \to \mathbb{R}$ be continuous. The Lagrangian dual function $\mathcal{L}$ of a primal problem* (P0) *is concave.*

For further simplification, by letting $\lambda = (\lambda_1,\dots,\lambda_m)^T \in \mathbb{R}^m$, $\mu = (\mu_1,\dots,\mu_n)^T \in \mathbb{R}^n$, the

dual function can be written shortly as

$$\mathcal{L}(\lambda, \mu) \quad = \quad \inf_{x \in X} \; f(x) + \langle \lambda, g(x) \rangle + \langle \mu, h(x) \rangle \tag{5.16}$$

with the standard inner products for the respective spaces.

Theorem 5.14 (Weak Duality Theorem, [Baza 06] Theorem 6.2.1)**.** *Let $x \in X$ be feasible for* (P0) *and let $(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^n_+$ be feasible for* (D0)*. Then $f(x) \geq \mathcal{L}(\lambda, \mu)$.*

Corollary 5.15 ([Baza 06] Corollaries 6.2.1(1, 2))**.** *It holds that*

$$\inf\{f(x) \,:\, x \in X,\, g(x) = 0,\, h(x) \leq 0\} \quad \geq \quad \sup\{\mathcal{L}(\lambda, \mu) \,:\, (\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^n_+\}\,. \tag{5.17}$$

Furthermore, if $f(x_0) = \mathcal{L}(\lambda_0, \mu_0)$ for feasible $x_0 \in X$ and $(\lambda_0, \mu_0) \in \mathbb{R}^m \times \mathbb{R}^n_+$, then x_0 and (λ_0, μ_0) solve (P0) *and* (D0)*, respectively.*

Thus, the optimal objective value of the primal problem is bounded from below by the optimal objective value of the dual problem and vice versa. If both values exist in $\mathbb{R}$ with strict inequality in (5.17), then the difference is denoted as the duality gap.

Considering the primal problems (P2) and (P2$*$) from Section 5.1, let

$$\begin{aligned} F_{2,\alpha}: \; \mathbb{R}^{d \times m} \quad &\to \quad \mathbb{R} \\ T \quad &\mapsto \quad F_{2,\alpha}(T) = \|T^*T\|_F^2 + \alpha \|T - Y_s\|_F^2 \end{aligned}$$

denote the objective function. Then the corresponding dualizations are given by

$$\text{(D2)} \quad \sup_{\substack{\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \\ \lambda_1, \ldots, \lambda_m \in \mathbb{R}}} \mathcal{L}_2(\Lambda) \qquad \text{and} \qquad \text{(D2}*\text{)} \; \sup_{\lambda \in \mathbb{R}} \mathcal{L}_{2*}(\lambda) \tag{5.18}$$

with dual objective functions

$$\mathcal{L}_2(\Lambda) \quad = \quad \inf_{T \in \mathbb{R}^{d \times m}} F_{2,\alpha}(T) + \operatorname{trace}\left((T^*T - I_m)\Lambda\right)\,, \tag{5.19}$$

$$\begin{aligned} \mathcal{L}_{2*}(\lambda) \quad &= \quad \inf_{T \in \mathbb{R}^{d \times m}} F_{2,\alpha}(T) + \lambda(\operatorname{trace}\left(T^*T\right) - m) \\ &= \quad \inf_{T \in \mathbb{R}^{d \times m}} F_{2,\alpha}(T) + \operatorname{trace}\left((T^*T - I_m)\lambda I_m\right)\,. \end{aligned} \tag{5.20}$$

From the fact that $\lambda I_m \in \{\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^{m \times m} \,:\, \lambda_1, \ldots, \lambda_m \in \mathbb{R}\}$, the representations in (5.19) and (5.20) imply that

$$\sup_{\lambda \in \mathbb{R}} \mathcal{L}_{2*}(\lambda) \quad \leq \quad \sup_{\substack{\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \\ \lambda_1, \ldots, \lambda_m \in \mathbb{R}}} \mathcal{L}_2(\Lambda)\,, \tag{5.21}$$

showing that the optimal objective value of (D2) is also bounded from below by the optimal objective value of (D2$*$).

The well-known Strong Duality Theorem (e.g. Theorem 6.2.4 in [Baza 06]) states that under certain constraint qualifications and complexity assumptions there is no duality gap. Since the requirements are not met for the problems (P2) and (P2$*$), the Strong Duality Theorem in its classical form does not apply. However, as Theorem 5.16 shows, the duality gap of (P2$*$) is zero nonetheless.

Theorem 5.16. *Let $\alpha > 0$ and $Y_s = \hat{U}\hat{\Sigma}\hat{V}^*$ with columns $y_j \in \mathcal{S}^{d-1}$, $j = 1, \ldots, m$. For the problem* (P2$*$) *exists no duality gap, i.e. the objective values of* (P2$*$) *and its dual* (D2$*$) *are identical.*

Proof. By Theorem 5.8, a solution of (P2$*$) is given by $T_0 = \hat{U}\Sigma\hat{V}^*$. The matrix Σ and the optimal Lagrange multiplier λ_0 can be computed from (5.11) satisfying the constraint $\|\Sigma\|_F^2 = m$. For the dual problem (D2$*$), it is easy to see by the Wielandt-Hoffman-Theorem (Theorem 5.4) that

$$\begin{aligned}
\sup_{\lambda\in\mathbb{R}} \mathcal{L}_{2*}(\lambda) \quad \geq \quad \mathcal{L}_{2*}(\lambda_0) \quad &= \quad \inf_{T\in\mathbb{R}^{d\times m}} \quad F_{2,\alpha}(T) + \lambda_0(\operatorname{trace}(T^*T) - m) \\
&= \quad \inf_{T\in\mathbb{R}^{d\times m}} \quad \|T^*T\|_F^2 + \alpha\|T - Y_s\|_F^2 + \lambda_0\|T\|_F^2 - \lambda_0 m \\
&= \quad \inf_{\delta_1,\ldots,\delta_d\in\mathbb{R}} \quad \left(\sum_{j=1}^{d} \delta_j^4 + \alpha(\delta_j - \hat{\sigma}_j)^2 + \lambda_0\delta_j^2\right) - \lambda_0 m\,,
\end{aligned}$$

where $\delta_1, \ldots, \delta_d$ denote the singular values of T. Note that an ordering of the form $\delta_1 \geq \ldots \geq \delta_d$ follows from Lemma 5.7 and by the ordering $\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_d$. Minimization of the quadratic distances $(\delta_j - \hat{\sigma}_j)^2$, $j = 1, \ldots, d$, also implies the non-negativity $\delta_d \geq 0$. Furthermore, separability of the problem in $\delta_1, \ldots, \delta_d$ gives

$$\mathcal{L}_{2*}(\lambda_0) \quad = \quad \left(\sum_{j=1}^{d} \inf_{\delta_j\geq 0} \delta_j^4 + \alpha(\delta_j - \hat{\sigma}_j)^2 + \lambda_0\delta_j^2\right) - \lambda_0 m\,. \tag{5.22}$$

Minimization of the inner expressions in (5.22) leads to the system of equations

$$\delta_j^3 + \frac{\alpha + \lambda_0}{2}\delta_j \quad = \quad \frac{\alpha}{2}\hat{\sigma}_j\,, \quad j = 1, \ldots, d. \tag{5.23}$$

Now if $\hat{\sigma}_d > 0$, the solution coincides with the unique solution $\sigma_1 \geq \ldots \geq \sigma_d > 0$ from the

system in (5.11), also satisfying $\sum_j \sigma_j^2 = \text{trace}\,(T^*T) = m$. This leads to the identity

$$\begin{aligned}\mathcal{L}_{2*}(\lambda_0) &= \left(\sum_{j=1}^{d} \sigma_j^4 + \alpha(\sigma_j - \hat{\sigma}_j)^2 + \lambda_0 \sigma_j^2\right) - \lambda_0 m \\ &= \sum_{j=1}^{d} \sigma_j^4 + \alpha(\sigma_j - \hat{\sigma}_j)^2 \quad = F_{2,\alpha}(T_0)\,.\end{aligned}$$

Therefore, applying Corollary 5.15, it holds that

$$F_{2,\alpha}(T_0) \;=\; \inf_{\substack{T\in\mathbb{R}^{d\times m}\\ \text{trace}(T^*T)=m}} F_{2,\alpha}(T) \;\geq\; \sup_{\lambda\in\mathbb{R}} \mathcal{L}_{2*}(\lambda) \;\geq\; \mathcal{L}_{2*}(\lambda_0) \;=\; F_{2,\alpha}(T_0)\,,$$

which completes the proof for the case $\text{rank}(Y_s) = d$.

For rank-deficient Y_s, suppose $\hat{\sigma}_{\ell+1} = \ldots = \hat{\sigma}_d = 0$ for some $\ell \in \{1, \ldots, d-1\}$ and let

$$\mu_0 \;=\; \frac{\alpha + \lambda_0}{2}\,.$$

Then, if $\mu_0 \geq 0$, the solution in (5.23) has to satisfy $\sigma_{\ell+1} = \ldots = \sigma_d = 0$, since $\xi(\mu_0) = 0$ is the only real root of p_{μ_0} from (5.9) and it holds again that $\sum_j \sigma_j^2 = m$.

On the other hand, if $\mu_0 < 0$, then the largest real root $\xi(\mu_0)$ from (5.10) is positive and σ_k can be chosen either as zero or $\xi(\mu_0)$ for $k = \ell+1, \ldots, d$ in order to satisfy (5.23). However, if we let $\sigma_{\ell+1} = \ldots = \sigma_r = \xi(\mu_0) > 0 = \sigma_{r+1} = \ldots = \sigma_d$ for some $r \in \{\ell, \ldots, d-1\}$, then

$$\begin{aligned}\sum_{j=1}^{d} \sigma_j^2 &= \sum_{j=1}^{\ell} \sigma_j^2 + (r-\ell)\,\xi(\mu_0)^2 \\ &< \sum_{j=1}^{\ell} \sigma_j^2 + (d-\ell)\,\xi(\mu_0)^2\,,\end{aligned}$$

where the last expression equals m by Lemma 5.5. Furthermore, $\mu_0 < 0$ implies $\lambda_0 < 0$ and

$$\lambda_0 \left(\sum_{j=1}^{d} \sigma_j^2 \;-\; m\right) \;>\; 0\,.$$

Hence, for the minimization in (5.22), the solution $\sigma_{\ell+1} = \ldots = \sigma_d = \xi(\mu_0)$ has to be chosen. So the minimizer in the case $\hat{\sigma}_d = 0$ also satisfies $\sum_j \sigma_j^2 = m$ leading again to the equality $\mathcal{L}_{2,*}(\lambda_0) = F_{2,\alpha}(T_0)$. □

5.3 Influence of α on the Choice of Optimal Data

As stated previously in Section 4.1, there exists $\alpha_0 \geq 0$ such that for all $\alpha > \alpha_0$, the minimizers of (P1) are located in the Dirichlet cells of the the subfamily $\tilde{\Psi}$ from 5.7 satisfying

$$\tilde{\Psi} \quad = \quad \operatorname*{arg\,min}_{\Psi \in Y^m} \mathrm{TFP}(\Psi)\,.$$

In that case, solving (P1) is equivalent to solving (P2), where the matrix $Y_s \in \mathbb{R}^{d \times m}$ contains the elements of $\tilde{\Psi}$ as column vectors. In general, Y_s depends on α and the transition from (P1) to (P2) becomes more difficult. A natural question is, if α_0 might be zero for all possible data sets Y, which would allow to formulate (P1) directly as (P2) independent of the regularization parameter.

In Example 5.17 we propose two problems in the form of (P2$*$) which differ in the singular value matrices of the respective data. We consider the difference function of the two functions that represent the optimal objective values in α and show that there exists a change of sign for this function. From there it can be concluded that in general $\alpha_0 > 0$.

Remember firstly, that by application of the Wielandt-Hoffman-Theorem 5.4, the problem (P2$*$) can be reduced to a constrained minimization on the singular values (5.12):

$$\begin{cases} \min\limits_{\sigma_1 \geq \ldots \geq \sigma_d \geq 0} & \sum\limits_{j=1}^{d} \sigma_j^4 + \alpha\,(\sigma_j - \hat{\sigma}_j)^2 \\ \text{s.t.} & \sum\limits_{\nu=1}^{d} \sigma_\nu^2 = m \end{cases}$$

with $\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_d \geq 0$ denoting the singular values of the given data matrix Y_s in (P2$*$).

Example 5.17. *For $d = 3$, $m = 9$ define the singular value matrices*

$$\hat{\Sigma}_\gamma \quad = \quad \left[\begin{array}{ccc|c} 2 & 0 & 0 & \\ 0 & 2-\gamma & 0 & \mathbf{0}_{3\times 6} \\ 0 & 0 & \sqrt{5-(2-\gamma)^2} & \end{array}\right]$$

with parameter $\gamma \in [0, \delta]$, $\delta = 2 - \sqrt{5/2}$, and

$$\hat{\Sigma} \quad = \quad \left[\begin{array}{ccc|c} \sqrt{5} & 0 & 0 & \\ 0 & \sqrt{2} & 0 & \mathbf{0}_{3\times 6} \\ 0 & 0 & \sqrt{2} & \end{array}\right]$$

where $\mathbf{0}_{3\times 6}$ *is a block of size* 3×6 *containing only zeros.*

For ease of notation, identify $\hat{\Sigma}_\gamma, \hat{\Sigma}$ *by their "diagonal" entries, i.e.*

$$\begin{aligned} \hat{\Sigma}_\gamma &= \left(2,\, 2-\gamma,\, \sqrt{5-(2-\gamma)^2}\right)^T, \\ \hat{\Sigma} &= \left(\sqrt{5},\, \sqrt{2},\, \sqrt{2}\right)^T \end{aligned}$$

and let

$$\mathcal{D} = \{v \in \mathbb{R}^d : v_1 \geq \ldots \geq v_d \geq 0\},$$

which is isomorphic to the cone $\mathcal{C}$ *of singular value matrices in* $\mathbb{R}^d$

$$\mathcal{C} = \{A \in \mathbb{R}^{d\times m} : a_{1,1} \geq \ldots \geq a_{d,d} \geq 0 \text{ and } a_{j,k} = 0\, \forall\, j \neq k\}.$$

Using the frame potential function $f : \mathcal{D} \to \mathbb{R}$, $f(\Sigma) = \|\Sigma^T\Sigma\|^2$ *and the penalty functions* $g_\gamma, g : \mathcal{D} \to \mathbb{R}$ *with*

$$\begin{aligned} g_\gamma(\Sigma) &= \|\Sigma - \hat{\Sigma}_\gamma\|^2, \\ g(\Sigma) &= \|\Sigma - \hat{\Sigma}\|^2, \end{aligned}$$

define the objective functions $h_{\gamma,\alpha}, h_\alpha : \mathcal{D} \to \mathbb{R}$ *by*

$$\begin{aligned} h_{\gamma,\alpha}(\Sigma) &= f(\Sigma) + \alpha\, g_\gamma(\Sigma) \quad \text{and} \\ h_\alpha(\Sigma) &= f(\Sigma) + \alpha\, g(\Sigma). \end{aligned}$$

Note that $\|\hat{\Sigma}_\gamma\|^2 = \|\hat{\Sigma}\|^2 = m$ *for all* $\gamma \in [0, \delta]$. *The frame potential of* $\hat{\Sigma}_\gamma$ *satisfies*

$$f(\hat{\Sigma}_\gamma) = 33 + 2\gamma\left(\gamma^3 - 8\gamma^2 + 19\gamma - 12\right), \tag{5.24}$$

which is monotonously decreasing on $[0, \delta]$ *with maximal value* $f(\hat{\Sigma}_0) = f(\hat{\Sigma}) = 33$ *and the minimal value* $f(\hat{\Sigma}_\delta) = 57/2$.

For the minimization of the objective functions, consider the functions $\phi_\gamma, \phi : \mathbb{R} \to \mathbb{R}$,

$$\phi_\gamma(\alpha) = \min_{\substack{\Sigma\in\mathcal{D} \\ \|\Sigma\|^2=m}} h_{\gamma,\alpha}(\Sigma), \qquad \phi(\alpha) = \min_{\substack{\Sigma\in\mathcal{D} \\ \|\Sigma\|^2=m}} h_\alpha(\Sigma) \tag{5.25}$$

and let

$$\Sigma^\star_{\gamma,\alpha} = \operatorname*{arg\,min}_{\substack{\Sigma\in\mathcal{D} \\ \|\Sigma\|^2=m}} h_{\gamma,\alpha}(\Sigma), \quad \Sigma^\star_\alpha = \operatorname*{arg\,min}_{\substack{\Sigma\in\mathcal{D} \\ \|\Sigma\|^2=m}} h_\alpha(\Sigma)$$

be the corresponding minimizers depending on $\alpha \in [0, \infty)$*. Now, for* $\alpha \to \infty$*, the minimizers* $\Sigma^\star_{\gamma,\alpha}, \Sigma^\star_\alpha$ *converge to the corresponding data matrices* $\hat{\Sigma}_\gamma$ *and* $\hat{\Sigma}$*, respectively. On the other hand, for* $\alpha \to 0$*, the minimization of* f *becomes dominant, leading to the (FUNTF-) case*

$$\Sigma^\star_{\gamma,0} = \Sigma^\star_0 \;\; = \;\; \Big(\sqrt{3},\, \sqrt{3},\, \sqrt{3}\Big)^T \,. \tag{5.26}$$

Consideration of the gradient

$$\nabla f(\Sigma) \;\; = \;\; 4(\sigma_1^3, \sigma_2^3, \sigma_3^3)$$

gives $\nabla f(\hat{\Sigma}_\gamma) = 4(8, (2-\gamma)^3, (5-(2-\gamma)^2)^{3/2})$ *and* $\nabla f(\hat{\Sigma}) = 4(5\sqrt{5}, 2\sqrt{2}, 2\sqrt{2})$*.*

Furthermore, the Taylor expansion of f *at* $\hat{\Sigma}$ *can be written as*

$$f(\hat{\Sigma} + \varepsilon) \;\; = \;\; f(\hat{\Sigma}) + \nabla f(\hat{\Sigma}) \cdot \varepsilon + \mathcal{O}(\|\varepsilon\|^2)\,, \quad \|\varepsilon\| \to 0\,.$$

The goal is to minimize the product term $\nabla f(\hat{\Sigma}) \cdot \varepsilon$ *under the constraint that* $\|\varepsilon\|^2 = 1$ *and* ε *lies in the tangent plane* $T_{\hat{\Sigma}}\mathcal{S}$ *of the sphere* $\mathcal{S} = \partial B_3(0) = \{v \in \mathbb{R}^d \mid \|v\| = 3\}$ *at* $\hat{\Sigma}$*.*

Applying the Lagrange approach for constrained optimization using the function

$$\mathcal{L}(\varepsilon, \mu) \;\; = \;\; \nabla f(\hat{\Sigma}) \cdot \varepsilon + \mu_1(\|\varepsilon\|^2 - 1) + \mu_2 \langle\, \varepsilon, \hat{\Sigma}\, \rangle$$

yields the minimizers

$$\varepsilon^\star = \frac{1}{3} \begin{pmatrix} -2 \\ \sqrt{5/2} \\ \sqrt{5/2} \end{pmatrix} \qquad \textit{with} \qquad \mu^\star = -4 \begin{pmatrix} \sqrt{5} \\ 11/3 \end{pmatrix}.$$

The same calculation for $\hat{\Sigma}_\gamma$ *with* $\gamma = 0$ *leads to* $\nabla f(\hat{\Sigma}_0) = 4(8, 8, 1)$ *and the minimizers of the corresponding Lagrange function are given by*

$$\varepsilon^\star_\gamma = \frac{1}{\sqrt{18}} \begin{pmatrix} -1 \\ -1 \\ 4 \end{pmatrix} \qquad \textit{with} \qquad \mu^\star_\gamma = -4 \begin{pmatrix} \sqrt{2} \\ 11/3 \end{pmatrix}. \tag{5.27}$$

Hence,

$$\nabla f(\hat{\Sigma}_\gamma) \cdot \varepsilon^\star_\gamma = -8\sqrt{2} \;\; > -8\sqrt{5} = \nabla f(\hat{\Sigma}) \cdot \varepsilon^\star\,, \tag{5.28}$$

implying that there exists a direction at $\hat{\Sigma}$ *where* f *descends faster than at* $\hat{\Sigma}_0$.

Now, choosing $0 < \gamma \ll \delta$ *in order to cause only a small decrease of* f *in* (5.24) *preserves the inequality in* (5.28). *For example, letting* $\gamma = 1/100$, *the frame potential of* $\hat{\Sigma}_\gamma$ *is* $f(\hat{\Sigma}_\gamma) \approx 32.7638 < 33 = f(\hat{\Sigma})$ *and the minimizers in* (5.27) *become*

$$\varepsilon_\gamma^\star \approx \begin{pmatrix} -0.2550 \\ -0.2256 \\ 0.9403 \end{pmatrix} \quad \text{with} \quad \mu_\gamma^\star \approx \begin{pmatrix} -5.6407 \\ -14.5617 \end{pmatrix}$$

which still satisfies the inequality in (5.28) *with* $\nabla f(\hat{\Sigma}_\gamma) \cdot \varepsilon_\gamma^\star \approx -11.2814 > -8\sqrt{5}$.

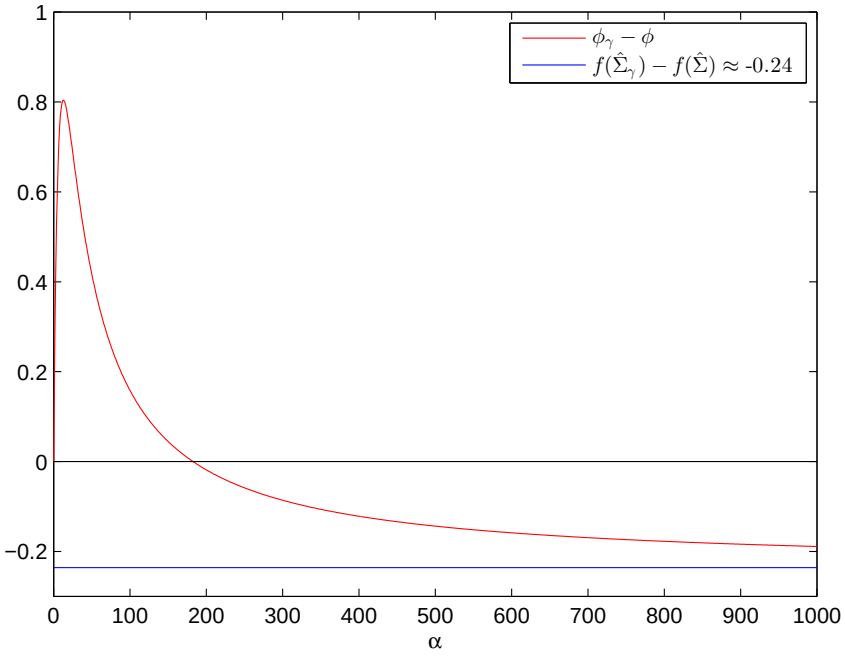


Figure 5.7: Difference function $\phi_\gamma - \phi$ with ϕ_γ, ϕ from (5.25) for $\gamma = 1/100$ (red) and lower bound $f(\hat{\Sigma}_\gamma) - f(\hat{\Sigma})$ (blue)

The result of the minimization for $\alpha \in [0, 1000]$ *can be seen in Figure 5.7, showing that for large* α

$$\hat{\Sigma}_\gamma \quad = \quad \underset{S \in \{\hat{\Sigma}_\gamma, \hat{\Sigma}\}}{\arg\min} \; \min_{\substack{\Sigma \in \mathcal{D} \\ \|\Sigma\|^2 = m}} \; f(\Sigma) + \alpha \|\Sigma - S\|_F^2$$

and, on the other hand, for small $\alpha > 0$,

$$\hat{\Sigma} \quad = \quad \underset{S \in \{\hat{\Sigma}_\gamma, \hat{\Sigma}\}}{\arg\min} \; \min_{\substack{\Sigma \in \mathcal{D} \\ \|\Sigma\|^2 = m}} \; f(\Sigma) + \alpha \|\Sigma - S\|_F^2 \, .$$

Note that the value of the difference function $\phi_\gamma - \phi$ is zero for $\alpha = 0$ which agrees with (5.26). *For $\alpha \to \infty$, we see that $(\phi_\gamma(\alpha) - \phi(\alpha)) \to (f(\hat{\Sigma}_\gamma) - f(\hat{\Sigma}))$.*

Remark 5.18. Let $\gamma = 1/100$ and $\hat{\Sigma}, \hat{\Sigma}_\gamma$ from Example 5.17. Via a system of nonlinear equations, an orthogonal matrix $V \in O(9)$ can be constructed numerically, such that the two data matrices $Y_\gamma, Y \in \mathbb{R}^{3\times 9}$, $Y_\gamma = \hat{\Sigma}_\gamma V^*$ and $Y = \hat{\Sigma} V^*$ with $Y_\gamma \approx$

$$\begin{bmatrix} 0.7253 & -0.7723 & -0.5350 & 0.8054 & 0.7476 & -0.5261 & -0.2078 & 0.7464 & 0.7119 \\ -0.6335 & 0.6065 & 0.7177 & 0.5857 & 0.6211 & 0.7208 & 0.7940 & 0.6217 & 0.6407 \\ 0.2696 & -0.1889 & 0.4457 & -0.0909 & -0.2355 & -0.4512 & 0.5712 & -0.2375 & 0.2878 \end{bmatrix}$$

and $Y \approx$

$$\begin{bmatrix} 0.8109 & -0.8635 & -0.5982 & 0.9005 & 0.8358 & -0.5882 & -0.2324 & 0.8344 & 0.7959 \\ -0.4502 & 0.4310 & 0.5100 & 0.4162 & 0.4414 & 0.5123 & 0.5643 & 0.4418 & 0.4553 \\ 0.3739 & -0.2620 & 0.6181 & -0.1261 & -0.3266 & -0.6258 & 0.7922 & -0.3294 & 0.3991 \end{bmatrix}$$

have unit norm columns. △

5.4 Formulation as a Polynomial Optimization Problem

In the following section, we show briefly how the problem (P1) can be transferred into the field of polynomial optimization. The procedure is the result of personal communication with Jean-Bernard Lasserre ([Lass 12]). An exhaustive discussion on the theory of polynomial optimization and applications can be found in [Lass 10].

Writing the penalty term in (P1) using inner products, we have

$$\text{(P1)} \quad \begin{cases} \min\limits_{\theta_1,\ldots,\theta_m \in \mathbb{R}^d} & \sum\limits_{k,\ell=1}^{m} |\langle \theta_k, \theta_\ell \rangle|^2 - \alpha \sum\limits_{\ell=1}^{m} \max\limits_{1\le j\le N} \langle y_j, \theta_\ell \rangle \\ \text{s.t.} & \|\theta_\ell\|^2 = 1, \qquad \ell = 1,\ldots,m. \end{cases}$$

Let $F : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}$ with

$$F(\Theta) \;=\; \text{TFP}(\Theta) - \alpha \sum_{\ell=1}^{m} \max_{1\le j\le N} \langle y_j, \theta_\ell \rangle$$

denote the objective function. The major idea is to repeatedly use the identity

$$\max\{a,b\} \quad = \quad \frac{a+b}{2} + \frac{|a-b|}{2}$$

to eliminate the max-term from the objective function by using additional constraints. As we will see, the formulation as a polynomial problem leads to a large number of additional variables.

Firstly, introduce for $\ell = 1, \ldots, m$ the variables $z_{1,\ell} \geq 0$ with

$$z_{1,\ell}^2 \quad = \quad (\langle \theta_\ell, y_2 \rangle - \langle \theta_\ell, y_1 \rangle)^2$$

so that $z_{1,\ell} = |\langle \theta_\ell, y_2 \rangle - \langle \theta_\ell, y_1 \rangle|$. Then define the polynomials $q_{1,\ell}$ by the identity

$$q_{1,\ell}(\theta_\ell, z_{1,\ell}) \quad = \quad \frac{1}{2} \left(\langle \theta_\ell, y_2 \rangle + \langle \theta_\ell, y_1 \rangle + z_{1,\ell} \right),$$

which stand for $\max\{\langle \theta_\ell, y_1 \rangle, \langle \theta_\ell, y_2 \rangle\}$.

Afterwards, introduce successively for $\nu = 2, \ldots, N-1$ the variables $z_{\nu,\ell}$ with the polynomial constraints

$$\begin{aligned} z_{\nu,\ell}^2 \quad &= \quad \left(\langle \theta_\ell, y_{\nu+1} \rangle - q_{\nu-1,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu-1,\ell}) \right)^2, \\ z_{\nu,\ell} \quad &\geq \quad 0\,. \end{aligned} \tag{5.29}$$

The variables $z_{\nu,\ell}$ describe the term $|\langle \theta_\ell, y_{\nu+1} \rangle - q_{\nu-1,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu-1,\ell})|$ and the polynomials

$$q_{\nu,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu,\ell}) \quad = \quad \frac{1}{2} \left(\langle \theta_\ell, y_{\nu+1} \rangle + z_{\nu,\ell} + q_{\nu-1,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu,\ell}) \right)$$

represent $\max\{\langle \theta_\ell, y_1 \rangle, \ldots, \langle \theta_\ell, y_{\nu+1} \rangle\}$.

Finally, after $N-1$ steps, the additional variables $z_{1,\ell}, \ldots, z_{N-1,\ell}$ with the corresponding polynomial constraints from (5.29) give a polynomial representation of the functions

$$h_\ell(\Theta) \quad = \quad \max_{1 \leq j \leq N} \langle \theta_\ell, y_j \rangle$$

by the m polynomials $q_{N-1,\ell} \in \mathbb{R}[\theta_\ell, z_{1,\ell}, \ldots, z_{N-1,\ell}]$.

Hence, by [Lass 12], the problem (P1) becomes in the context of polynomial optimization problem the following problem:

$$
\text{(P1)} \quad \left\{ \begin{array}{lll}
\min & F(\Theta) = \displaystyle\sum_{k,\ell=1}^{m} \langle \theta_k, \theta_\ell \rangle^2 - \alpha \sum_{\ell=1}^{m} q_{N-1,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{N-1,\ell}) & \\
\text{s.t.} & \|\theta_\ell\|^2 = 1, \qquad \ell = 1, \ldots, m, & \\
& q_{0,\ell}(\theta_\ell) = \langle \theta_\ell, y_1 \rangle, \qquad \ell = 1, \ldots, m, & \\
& z_{\nu,\ell} \geq 0, \qquad \ell = 1, \ldots, m; \quad \nu = 1, \ldots, N-1, & \\
& z_{\nu,\ell}^2 = \left(\langle \theta_\ell, y_{\nu+1} \rangle - q_{\nu-1,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu-1,\ell}) \right)^2, & \\
& \qquad \ell = 1, \ldots, m; \quad \nu = 1, \ldots, N-1, & \\
& q_{\nu,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu,\ell}) = \frac{1}{2} \left(\langle \theta_\ell, y_{\nu+1} \rangle + z_{\nu,\ell} + q_{\nu-1,\ell}(\theta_\ell, z_{1,\ell}, \ldots, z_{\nu-1,\ell}) \right), & \\
& \qquad \ell = 1, \ldots, m; \quad \nu = 1, \ldots, N-1. &
\end{array} \right.
$$

Note that it is possible to eliminate the polynomials $q_{\nu,\ell}$ from the problem by replacing them directly by their expressions in the variables $\theta_\ell, z_{1,\ell}, \ldots, z_{\nu,\ell}$. However, the problem still consists of $m \cdot d$ variables characterizing the vectors $\theta_\ell \in \mathcal{S}^{d-1}$ and $(N-1) \cdot m$ variables $z_{\nu,\ell} \in \mathbb{R}_{\geq 0}$.

In chapter 5 of [Lass 10], Lasserre proposes methods to solve classes of problems in polynomial optimization by moment-SOS (sums of squares) relaxations. A condensed description of the technique how to handle such semi-definite programming relaxations can be found in [Lass 11]. In our case even the first relaxation is already a large sized SDP. Since the objective function F has a representation as a quartic polynomial in $m((N-1)+d)$ variables, this SDP becomes unfeasible to solve for large N or m ([Lass 12]).

5.5 Related Problems

As mentioned previously in Remark 5.12, the technique of combining regularization terms and loss terms in order to create a tradeoff between approximation and smoothing is often applied in data analysis. One of the earlier examples led to the smoothing splines which were introduced by Schoenberg and Reinsch for the cubic case ([Scho 64, Rein 67]) and are often used in data regression. An excellent generalization for all degrees can be found in de Boor's contribution ([Boor 01]). Technical aspects on this topic can today be found in numerous other publications and academical textbooks, e.g. Section 5.4 in [Hast 01].

For the sparse eigenvalue problem in principal component analysis, d'Aspremont et al. use the common Rayleigh quotient for positive semidefinite matrices as regularization term. The

penalty term is chosen as the negative ℓ_0-norm and the combined functional gets maximized over the unit sphere ([dAsp 08]):

$$\max_{\|x\|_2=1} \; x^T \Sigma x \quad - \quad \alpha \|x\|_0 \, .$$

Other similar functionals in data representation are applied in statistical shrinkage methods and coefficient selection such as ridge regression (Section 3.4 in [Hast 01]) or the lasso by Tibshirani ([Tibs 94]) which differ in the choice of the penalty term. Note that several modifications of the lasso have been proposed, an overview can be found in [Tibs 11]. One of the modifications refers to the field of matrix completion theory in compressive sensing, where Candès et al. [Cand 10], Cai et al. [Cai 10] and Mazumder et al. [Mazu 10] propose to solve

$$\min_{\hat{X}} \; \alpha \|\hat{X}\|_* \quad + \quad \|X - \hat{X}\|_F^2 \, , \quad \alpha \geq 0, \tag{5.30}$$

for a given matrix X of size $d \times m$ with $\|X\|_* = \sum_{j=1}^{d} \sigma_j$ denoting the nuclear norm of $X = U \Sigma V^*$.

In this context, $\| \cdot \|_*$ works as the regularization functional whereas $\|X - \cdot\|_F^2$ is the loss term depending on the given data in X. The objective functional in (5.30) consists of the same quadratic loss term $\|X - \hat{X}\|_F^2$ as in the problems (P2) and (P2$*$). Only the regularization term $\|\hat{X}\|_* = \|(\sigma_1, \ldots, \sigma_d)^T\|_1$ differs from the one in (P2), where $\sum_j \sigma_j^4$ is considered.

The solution of the problem in (5.30) is based on the following lemma:

Lemma 5.19 ([Cai 10])**.** *Suppose $W \in \mathbb{C}^{d \times m}$ is of rank r. Then the solution to*

$$\min_{Z \in \mathbb{C}^{d \times m}} \; \alpha \|Z\|_* \quad + \quad \frac{1}{2} \|W - Z\|_F^2$$

is given by $\hat{Z} = S_\alpha(W) = U \Sigma_\alpha V^*$ where $\Sigma_\alpha = \operatorname{diag}((\sigma_1 - \alpha)_+, \ldots, (\sigma_d - \alpha)_+)$, $t_+ = \max\{t, 0\}$ and $U \Sigma V^*$ is the singular value decomposition of W.

S_α stands for the shrinkage operator performing a soft thresholding on the singular values of W ([Dono 94]). In [Mazu 10], Mazumder et al. present an alternative proof to the one stated in [Cai 10]. However, under application of the Wielandt-Hoffman theorem even this proof can be further condensed:

Proof. Since the objective function is obviously strictly convex, there exists a unique minimizer ([Cai 10]). Moreover, the regularization term depends only on the singular values of Z. Hence, Theorem 5.4 justifies the choice of U and V from the SVD of W as stated in the lemma. So the problem becomes

$$\min_{D} \quad \frac{1}{2}\|\Sigma - D\|_F^2 + \alpha \sum_j \delta_j$$

where D is diagonal with diagonal elements $\delta_j \geq 0$. Since there are no restrictions on the δ_j (except for the non-negativity), the problem can be separated into d problems like in [Mazu 10]:

$$\min_{\delta_j \geq 0} \quad \frac{1}{2}(\sigma_j - \delta_j)^2 + \alpha\delta_j \,.$$

It is easy to see that for $\sigma_j \geq \alpha$ the minimum is $\delta_j = \sigma_j - \alpha$ and $\delta_j = 0$ otherwise. □

Chapter 6

Numerical Results

Finally, we evaluate the performance of the Penalized Frame Potential in terms of the clustering of real data. The algorithm is implemented in Matlab®, the minimization of the function $F_\alpha(\cdot, Y)$ uses the built-in routine `fmincon` from the Optimization Toolbox.

We evaluate the clustering results for both simulated and real data by using tools such as innercluster variance or t- and F-statistics ([Cali 74]). The outline of the last chapter is as follows. Section 6.1 deals with the performance of our method compared with the STEM algorithm from [Erns 05]. In Section 6.2 we apply our method to an example from [Kim 07], which was generated in order to stress characteristic features of the DIB-C algorithm. We introduce this algorithm in a few words. As we will see, our proposed method shows good performance when applied on the mentioned example. Furthermore, we evaluate our method in terms of the Adjusted Rand Index (ARI) which we introduce briefly. The well-known k-means clustering approach showed good performance in terms of the ARI in [Kim 07]. A comparison to the k-means results will be included. Afterwards, we apply modifications of the PFP algorithm in Section 6.3 which have proven useful in the application on real data, especially in gene expression data. In Section 6.4, we extend the proposed method in order to extract features from multispectral data and also include a short example.

We start with a short description of our algorithm that is based on minimization of the Penalized Frame Potential from the previous chapters. The method is realized in the following way. Firstly, we project the given data onto the sphere $\mathcal{S}^{d-1}$ as described in Chapter 3

and call these projections Y. Minimization of $F_\alpha(\cdot, Y)$ on $\mathcal{S}^{d-1}$ leads to (possibly local) minimizers $\theta_1, \ldots, \theta_m$ where m and α have to be chosen a priori. Let $\hat{D}_1, \ldots, \hat{D}_m \subset \mathcal{S}^{d-1}$ denote the Dirichlet cells of those minimizers. We assign a given time series to cluster C_ℓ if its corresponding projection satisfies $y_j \in \hat{D}_\ell$. Furthermore, since the projection $Q : H \cap \mathcal{S}^d \to \mathcal{S}^{d-1}$ from Chapter 3 is orthogonal, we call $Q^T\theta_1, \ldots, Q^T\theta_m \in \mathbb{R}^{d+1}$ cluster prototypes from the PFP algorithm.

In addition, we count the number of data in each cluster and apply a significance test including Bonferroni's correction for multiple testing as in [Erns 05]. Remember that for $\alpha > 0$ by Theorem 4.8, $\theta_\ell \notin \partial D_j$ for $\ell = 1, \ldots, m$ and $j = 1, \ldots, N$ where $D_1, \ldots, D_N \subset \mathcal{S}^{d-1}$ are the Dirichlet cells of the data projections. However, it is obviously still possible that a projection y_j belongs to $n_j > 1$ Dirichlet cells generated by the minimizers. In that case we proceed by assigning the corresponding time series to all n_j nearest clusters and count the assignments by n_j^{-1} as in STEM.

6.1 Performance of the Penalized Frame Potential

One of our major motivations was to create a data-driven method since other algorithms are often data-independent. However, note that the model profiles in STEM ([Erns 05]) or the templates in DIB-C ([Kim 07]) also take some data-specific features into account. In STEM, the model profiles are not allowed to change by more than $c \in \mathbb{N}$ (in general $c = 2$) units between adjacent time points. Since the first value of all models is zero, the extremal values are $-dc$ and dc for time series of length $d + 1$. This restricted behavior simulates real data recorded from biological processes. So there exists a model-based relation of the prototypes from STEM to biological data. Therefore the data-independent solutions from the PFP algorithm with $\alpha = 0$ should not be interpreted as equivalent to the STEM solutions. However, note that in real data the time points at which samples are taken are often not distributed equidistantly which is not considered in the STEM setting.

For our experiments in this section we use the yeast amino acid starvation data from [Gasc 00], available at `http://www.benoslab.pitt.edu/astro/Amino_Acid_Starvation.txt` (accessed July 16, 2013) and filtered the data as described in [Spri 11]:

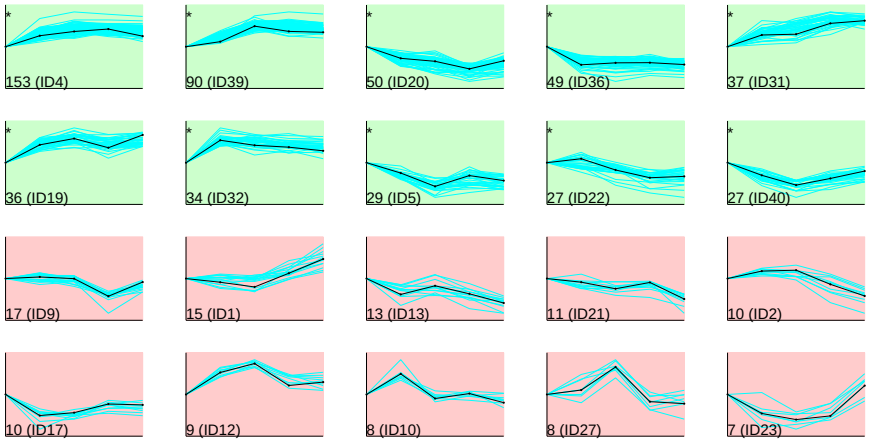


Figure 6.1: 20 largest clusters from PFP clustering with $m = 40$, $\alpha = 2.0$ for AAS example (ten significant clusters, indicated by "$*$"); prototypes (black), data (cyan)

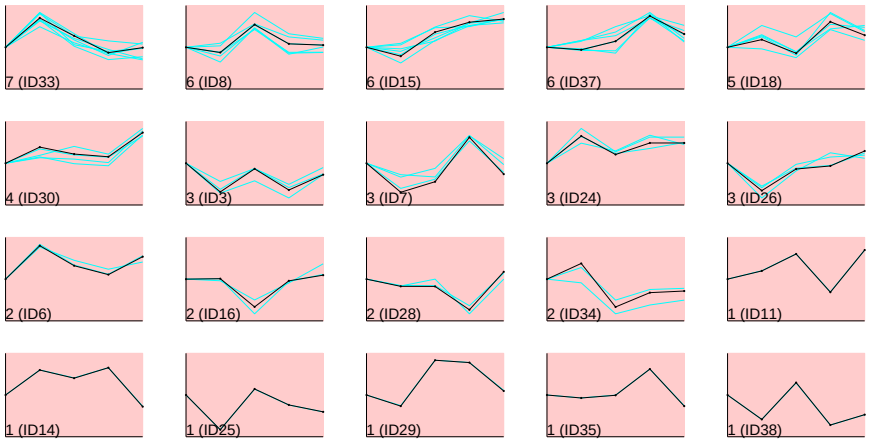


Figure 6.2: 20 smallest clusters from PFP clustering with $m = 40$, $\alpha = 2.0$ for AAS example; prototypes (black), data (cyan), significant clusters indicated by "$*$"

The table contains logarithmic values of responses of 6152 genes to stress by amino acid starvation (AAS) at time points $0.5h$, $1h$, $2h$, $4h$ and $6h$, where we leave out the time point $t = 0$. So the time series have length $d + 1 = 5$. The data are filtered by removing all genes with missing values and genes whose expression levels vary by less than $\varepsilon = 2$ over the whole time interval, giving a total of $N = 700$ short time series for our analysis.

Figures 6.1 and 6.2 show the results of the PFP clustering using $m = 40$ and $\alpha = 2.0$, the

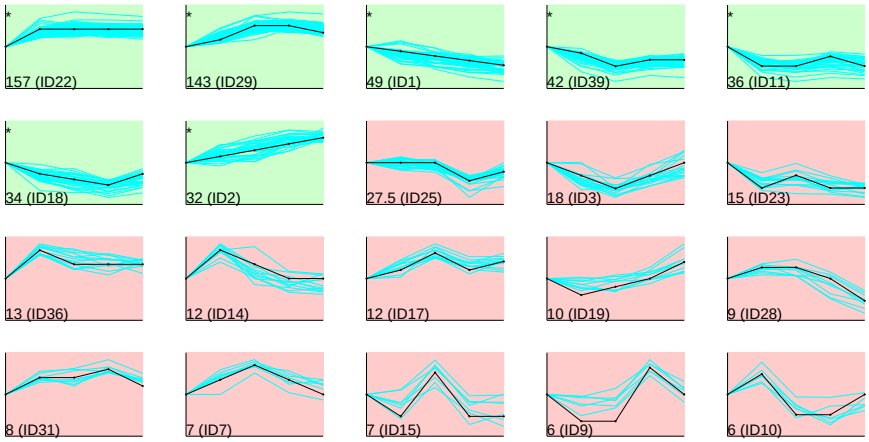


Figure 6.3: 20 largest clusters from STEM clustering with $m = 40$, $c = 2$ for AAS example (seven significant clusters, indicated by "$*$"); prototypes (black), data (cyan)

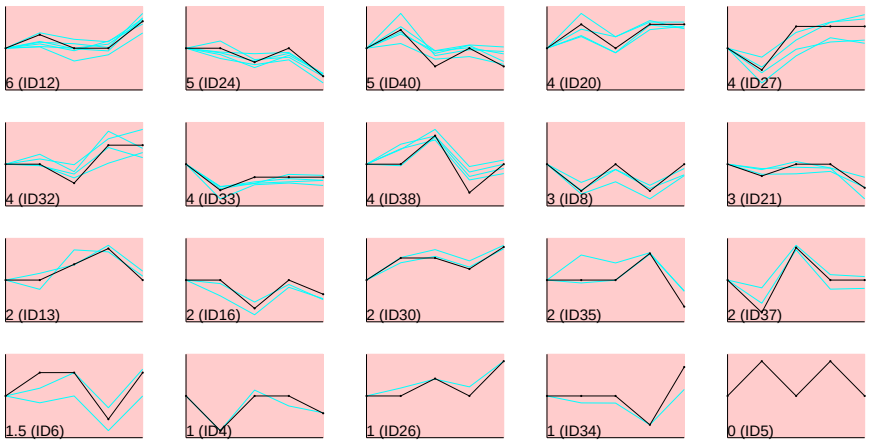


Figure 6.4: 20 smallest clusters from STEM clustering with $m = 40$, $c = 2$ for AAS example; prototypes (black), data (cyan), significant clusters indicated by "$*$"

corresponding statistics are presented in Table 6.1. In the example, the dominant clusters from PFP and STEM (see Figure 6.3) are almost of the same size (153 in PFP and 157 in STEM), whereas the second-largest clusters differ in size (90 in PFP and 143 in STEM). However, Table 6.1 shows that the two largest clusters calculated with PFP are also their respective nearest neighbors which indicates concentration of the data. The corresponding statistics for the STEM results can be found in Table 6.2.

For the evaluation of the clusters, Table 6.1 also gives information on the innercluster variances as well as on t- and F-statistics. For identifying significant clusters we apply a permutation-based test which is also used in [Erns 05]. After calculating all possible $(d+1)!$ permutations of the data, we assign these permutations to the prototypes and count the number of total assignments. If $s_\ell^{(k)}$ denotes the number of data which are assigned to cluster ℓ in permutation k, then, according to [Erns 05], the expected number of data in cluster C_ℓ becomes

$$E_\ell = \frac{1}{(d+1)!} \sum_{k=1}^{(d+1)!} s_\ell^{(k)}, \quad \ell = 1, \ldots, m\,.$$

The number of data in C_ℓ is treated as a binomial random variable X with parameters N and E_ℓ/N and the Bonferroni-corrected p-value becomes $P(X \geq s_\ell^{(1)})$, where $s_\ell^{(1)}$ is the number of original assignments. Finally, C_ℓ is called significant, if $P(X \geq s_\ell^{(1)}) < \lambda/m$ with λ denoting the level of significance. In our examples, we only consider $\lambda = 0.05$.

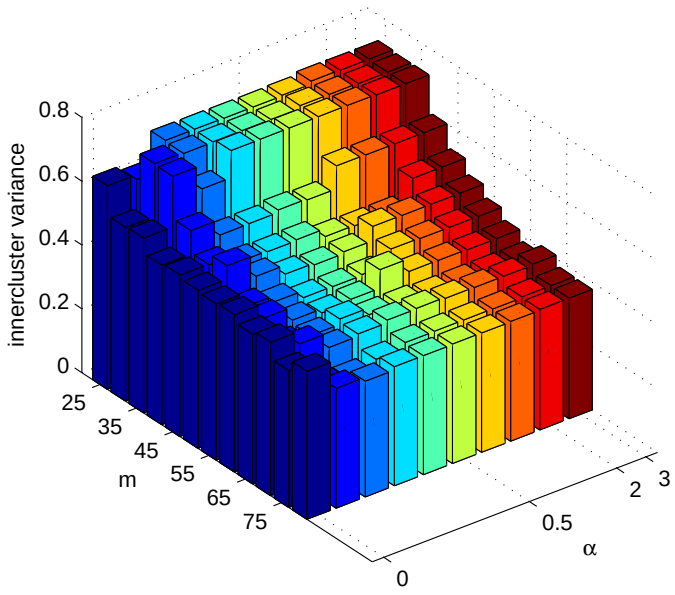


Figure 6.5: Mean innercluster variances of the results from clustering the AAS data with PFP where $m = 25, 30, 35, \ldots, 75, 80$ and $\alpha = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.5, 2.0, 3.0$

A natural question that arises is how to choose the input parameters m and α. For this purpose, we propose to apply a method that is also common in order to calculate an appropriate number of clusters for the k-means approach. After clustering and significance testing for different m and α, compute the innercluster variances for the significant clusters. Figure 6.5 presents the

mean values of the innercluster variances of the significant clusters from the AAS data. As can be seen, the values are essentially identical for $\alpha \geq 1$. Furthermore, from $m = 35$ to $m = 45$ there is a strong decrease for all positive α visible, leading to the conclusion that $\alpha = 1.5$ and $m = 45$ represents a good choice of parameters for the given data. Note that the value increases for $m = 45$ from $\alpha = 1.5$ to $\alpha = 2.0$. The reason is that for $\alpha = 2.0$ there is one more significant cluster than for $\alpha = 1.0$. Besides the mean value one could also take, e.g., the maximal innercluster variance of the significant clusters into consideration.

Ernst et al. propose a greedy algorithm for grouping clusters when the distance of the respective prototypes is below a threshold $\tau > 0$ ([Erns 05]). In our method we adapt this procedure to reduce the effect of overestimating m without giving further results here.

6.2 On an Example for DIB-C

In this section, we apply our method on an example provided by Kim and Kim in [Kim 07] for stressing the performance of the DIB-C algorithm. The DIB-C method is primarily formulated for the clustering of gene expression data. The algorithm classifies the different time series by a certain pattern which relies on the first- and second-order differences between adjacent time points. Note that it is actually capable of considering replicates in gene experiments by using t-statistics for the coding of the data in terms of the classification pattern.

Suppose, we consider N not necessarily normalized data series $y_j \in \mathbb{R}^d$, $j = 1, \ldots, N$, which may be in the log-ratio form described in Chapter 3. In a nutshell, every element of the first-order difference $y_j^{(1)} \in \mathbb{R}^{d-1}$ of y_j is classified by the symbols D, I and N. The symbol D indicates that there exists a decrease in the data values between two adjacent time-points. Correspondingly, I stands for an increase and N is used if there is no significant change present. In a similar way, the symbols A, V and N encode the second-order differences where A and V denote concavity and convexity, respectively. The symbolic pattern for y_j consists of the $d-1$ symbols for the first order differences and the $d-2$ symbols for the second-order differences. Finally, all time series sharing the same symbolic pattern are grouped together building a cluster. For an exhaustive description of the method we refer to [Kim 07].

ID	Clustersize	IC var.	Expectations	p-values	adj. p-values	Significance	F-Statistics	t-Statistics	Neighbor	Distance
4	153	0.167905	20.5083	0	0	1	0.0995259	-1.47928	39	0.164816
39	90	0.12475	8.75833	0	0	1	0.0739455	-2.09011	4	0.164816
20	50	1.02186	11.7333	0	0	1	0.605707	0.910288	36	0.182627
36	49	0.52113	18.675	7.11433e-10	2.84573e-08	1	0.3089	0.729846	17	0.178435
31	37	0.299331	9.55833	1.7959e-12	7.18359e-11	1	0.177428	-0.347816	15	0.207356
19	36	0.196594	19.8333	0.000292496	0.0116998	1	0.116531	-1.5336	30	0.206133
32	34	1.24458	19.1333	0.000595445	0.0238178	1	0.737724	-0.0289545	12	0.206709
5	29	0.648511	13.2917	4.57617e-05	0.00183047	1	0.384406	1.29767	40	0.164396
22	27	1.53642	5.825	2.29516e-11	9.18066e-10	1	0.910712	0.423894	9	0.258853
40	27	0.240406	8.71667	1.25515e-07	5.0206e-06	1	0.142501	2.23127	5	0.164396
9	17	0.507962	40.8333	0.999986	1	0	0.301094	-0.24976	28	0.208299
1	15	0.828108	16.4583	0.579662	1	0	0.490861	0.48373	23	0.241191
13	13	0.275345	15.1667	0.654758	1	0	0.163211	0.536246	36	0.197463
21	11	0.388685	32.95	0.999994	1	0	0.230393	0.411072	13	0.294669
2	10	1.44996	23.6167	0.998813	1	0	0.859466	0.175613	27	0.218162
17	10	0.338329	12.3083	0.68628	1	0	0.200545	1.1341	36	0.178435
12	9	0.382507	9.05833	0.420325	1	0	0.226731	-1.522	32	0.206709
10	8	0.994728	18.6583	0.995634	1	0	0.589625	0.940696	34	0.182673
27	8	0.473005	16.15	0.980782	1	0	0.280374	0.440147	38	0.217912
23	7	0.469252	8.61667	0.630614	1	0	0.27815	1.12473	40	0.206652

Table 6.1: Statistics for the 20 largest clusters from PFP clustering (see Figure 6.1) including innercluster variance (IC var.) and distance to nearest neighbor

ID	Clustersize	IC var.	Expectations	p-values	adj. p-values	Significance	F-Statistics	t-Statistics	Neighbor	Distance
22	157	0.255465	37.7833	0	0	1	0.151427	-1.14109	36	0.209431
29	143	0.144061	9.225	0	0	1	0.0853921	-1.78225	31	0.226426
1	49	1.37316	11.0833	0	0	1	0.813944	0.641295	23	0.292893
39	42	0.684739	9.04167	1.11022e-16	4.44089e-15	1	0.40588	1.39991	18	0.201728
11	36	0.317633	13.7583	1.08568e-07	4.3427e-06	1	0.188277	1.12388	33	0.209431
18	34	0.98064	17.35	0.000100971	0.00403885	1	0.581275	0.705739	39	0.201728
2	32	0.36644	9.84167	3.45835e-09	1.38334e-07	1	0.217208	-0.238872	29	0.272393
25	27.5	0.401879	26.6333	0.420458	1	0	0.238214	-0.358362	38	0.209431
3	18	0.400071	24.3333	0.889079	1	0	0.237142	1.53781	39	0.318615
23	15	0.328014	11.9833	0.152498	1	0	0.19443	1.05588	33	0.209431
36	13	1.68249	14.9917	0.638053	1	0	0.997297	0.517602	14	0.209431
14	12	0.866793	16.225	0.824429	1	0	0.513792	0.528414	36	0.209431
17	12	0.3329	11.6333	0.381826	1	0	0.197327	-0.941757	29	0.26006
19	10	1.03221	19.0417	0.983198	1	0	0.611845	0.388653	26	0.321599
28	9	1.10888	30.825	0.999997	1	0	0.65729	0.396121	7	0.268075
31	8	0.123087	13.4417	0.920803	1	0	0.0729596	-0.419817	13	0.215535
7	7	0.731112	14.9917	0.982795	1	0	0.433367	-0.894309	28	0.268075
15	7	0.811082	13.575	0.961518	1	0	0.480769	0.202284	37	0.209431
9	6	0.157931	20.5917	0.999858	1	0	0.0936134	0.973305	32	0.214286
10	6	0.342331	20.9417	0.999891	1	0	0.202917	0.643427	40	0.357143

Table 6.2: Statistics for the 20 largest clusters from STEM clustering (see Figure 6.3) including innercluster variance (IC var.) and distance to nearest neighbor

The first example in [Kim 07] for the performance of the DIB-C algorithm consists of $N = 180$ time series with $d = 4$ time points. We create $r = 18$ template time series (see Figure 6.6) with integer data values in $[-6, 6]$ by combining only the first-order differences (DDD) and (III) with all possible second-order differences $((AA)$, (AN), (AV), (NA), (NN), (NV), (VA), (VN) and $(VV))$. For example, $(0, 1, 3, 6)$ builds a template for $((III), (VV))$, $(0, 1, 3, 5)$ for $((III), (VN))$ and so forth. Now, reproducing the templates ten times and applying uniform noise in $[-0.015, 0.015]$ as described in [Kim 07] generates the test data. Note that the original approach also considers the cluster generated by the null template $((NNN), (NN))$ which leads to a total of 190 time series. However, since this cluster does not serve any practical purpose, we omit this class.

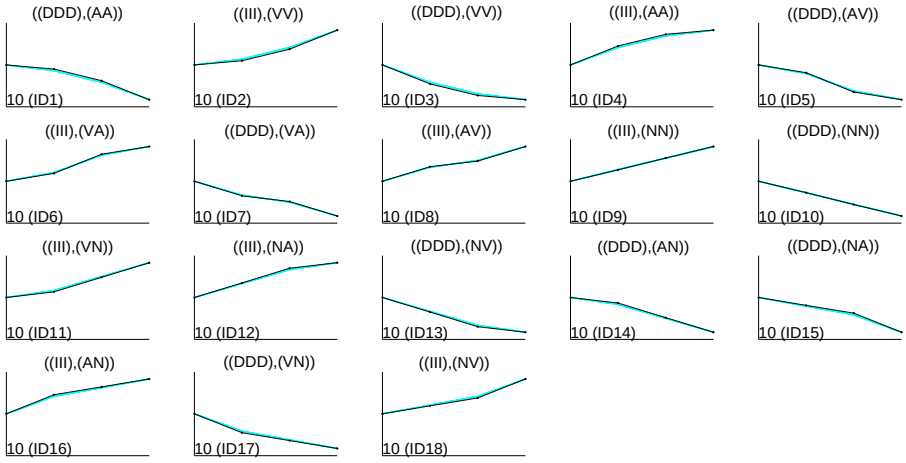


Figure 6.6: Results for the clustering of the experimental data from the DIB-C example using PFP with $m = 18$ and $\alpha = 3.0$; prototypes (black), data (cyan)

Figure 6.6 presents the results of the clustering with the PFP algorithm for the number of clusters $m = 18$ and the regularization parameter $\alpha = 3.0$. As can be seen, the number of time series in each cluster is ten, which agrees with the number of noisy reproductions of the templates. Furthermore, no misclassification occurred since all time series are almost identical with the calculated cluster prototypes.

Due to the relatively low noise, the example becomes very restrictive. However, taking a closer look at the projections of the data onto the sphere in the left chart in Figure 6.7 shows that the data are concentrated in terms of the dissimilarity measure from Chapter 3. It can be seen that by minimization of the PFP it is actually possible to capture the different data clusters.

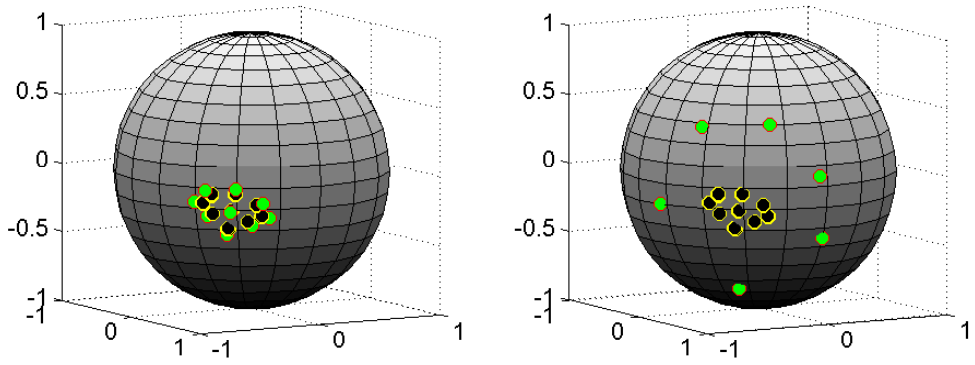


Figure 6.7: Projections onto $\mathcal{S}^2$ of the low-noised data (black) and the prototypes (green) from Figure 6.6 calculated with PFP using $\alpha = 3.0$ (left) and with $\alpha = 0.3$ (right)

The right chart in Figure 6.7 shows the projections of the prototypes calculated with $\alpha = 0.3$. Obviously, the lower weight on the penalty term causes the total frame potential to dominate in the minimization process leading to less data-oriented prototypes. Some prototypes and the other half of the data are essentially located antipodal to the visible points.

As mentioned, the DIB-C algorithm is capable to consider replicates in gene experiments. In order to process replicates in the PFP algorithm, we simply calculate the median over the replicates in each time point which gives the new data. For our second experiment, we generated eight replicates for each of the 180 time series from above and added normal noise using $\mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.06, 0.12, 0.3, 0.6\}$. Since the correct classification is known in advance, the Adjusted Rand Index (ARI) by Hubert and Arabie ([Hube 85]) can be used for the evaluation of the performance of an algorithm:

$$\text{ARI} \;=\; \frac{\sum_{j,k}\binom{n_{jk}}{2} \;-\; \left(\sum_j \binom{n_{j\cdot}}{2}\;\sum_k \binom{n_{\cdot k}}{2}\right) / \binom{n}{2}}{\frac{1}{2}\left(\sum_j \binom{n_{j\cdot}}{2} + \sum_k \binom{n_{\cdot k}}{2}\right) \;-\; \left(\sum_j \binom{n_{j\cdot}}{2}\;\sum_k \binom{n_{\cdot k}}{2}\right) / \binom{n}{2}}\,, \tag{6.1}$$

where
$$\begin{aligned} n &\;\hat{=}\; \text{number of total objects}\,, \\ n_{jk} &\;\hat{=}\; \text{number of objects from class } K_j \text{ in cluster } C_k\,, \\ n_{j\cdot} &\;\hat{=}\; \text{number of objects in class } K_j\,, \\ n_{\cdot k} &\;\hat{=}\; \text{number of objects in cluster } C_k\,. \end{aligned}$$

In practice, these numbers are often listed in a contingency table:

Class / Cluster	C_1	C_2	$\cdots$	C_m	$\sum$
K_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
K_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
K_r	n_{r1}	n_{r2}	$\cdots$	n_{rm}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	n

The maximum ARI is one, its expected value is zero. Therefore, in contrast to the classical Rand Index, negative values are possible. However, a negative ARI indicates "less agreement than expected by chance" ([Frit 10]).

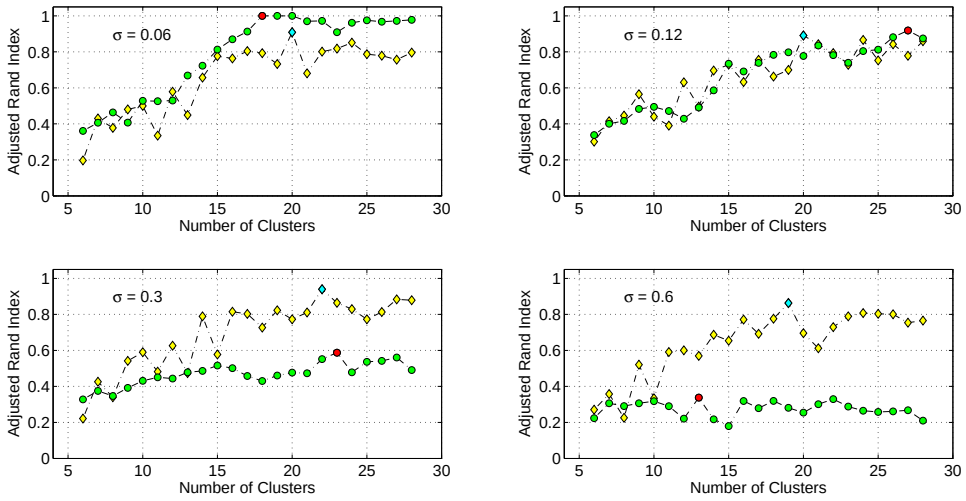


Figure 6.8: ARI for the noise levels $\sigma \in \{0.06, 0.12, 0.3, 0.6\}$ of the replicates for PFP (green circles) and k-means (yellow diamonds) clustering results including indicators for the respective maxima (red/cyan)

In [Kim 07], clustering results of DIB-C, k-means, STEM and Self-Organizing Maps (SOM) for the given example are compared. For low noise levels, DIB-C identifies the correct number of classes and has an ARI of almost one. However, with increasing noise, the DIB-C results deteriorate. The only algorithm that shows constantly good performance over all noise levels is k-means, which is why we compare our method to this algorithm here. The AR indices for the results using k-means and PFP for $m = 6, \ldots, 28$ and the four different noise levels are presented in Figure 6.8. For the lowest noise level $\sigma = 0.06$, the ARI of the PFP clustering peaks with a maximal value of exactly one for $m = 18$, which agrees with the correct number of simulated clusters and is higher than the DIB-C result in [Kim 07]. Note that the AR indices

computed on basis of the PFP results for $\sigma \in \{0.06, 0.12\}$ dominate the ones from k-means. However, in the presence of higher noise levels, identification of the correct classification via the PFP method is obviously not possible.

6.3 Modifications

The key ingredient of our method for computing data-dependent cluster prototypes is the control parameter α. In this section, we present the influence of changes in α on the results. Furthermore, for the application in the clustering of real data we also include additional features which will also be presented in the following.

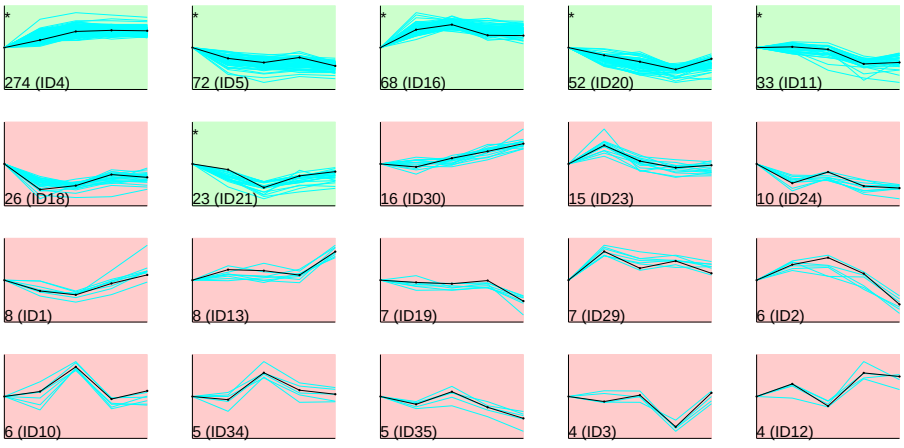


Figure 6.9: 20 largest clusters from PFP clustering with $m = 40$, $\alpha = 0$ for AAS example (six significant clusters, indicated by "$*$"); prototypes (black), data (cyan)

Figure 6.9 presents the twenty largest clusters from the AAS data computed with the data-independent version of the PFP functional. Compared to the results in Figure 6.1 it becomes obvious that the size of the largest cluster increases whereas most of the other clusters shrink. Moreover, the number of significant clusters decreases. Further experiments show that small $\alpha > 0$ already reduces the size of the dominating clusters. However, the matrix

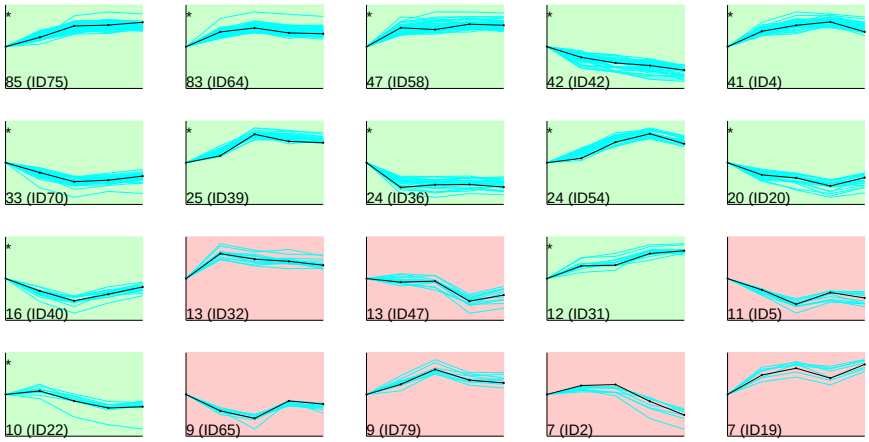


Figure 6.10: 20 largest clusters from PFP clustering with $m = 80$, $\alpha = 2.0$ for AAS example (13 significant clusters, indicated by "$*$"); prototypes (black), data (cyan)

$T = [\theta_1, \ldots, \theta_{40}] \in \mathbb{R}^{4\times 40}$ with

$$T = \begin{bmatrix} -0.4057 & 0.4142 & 0.3513 & -0.8591 & 0.7255 & \cdots & -0.6346 & 0.6854 \\ -0.2902 & 0.4165 & -0.5665 & -0.2171 & 0.4227 & \cdots & -0.6897 & 0.1629 \\ 0.8546 & -0.7868 & 0.0146 & -0.3783 & 0.1113 & \cdots & -0.2554 & 0.6939 \\ 0.1443 & -0.1895 & -0.7453 & 0.2677 & 0.5316 & \cdots & -0.2375 & 0.1490 \end{bmatrix}$$

satisfies $\|\theta_\ell\|_2 = 1$, $\ell = 1, \ldots, 40$. Moreover, the numerical values of the corresponding frame matrix are

$$S^{(0)} = \begin{bmatrix} 10.0000 & -0.0000 & -0.0000 & 0.0000 \\ -0.0000 & 10.0000 & 0.0000 & -0.0000 \\ -0.0000 & 0.0000 & 10.0000 & 0.0000 \\ 0.0000 & -0.0000 & 0.0000 & 10.0000 \end{bmatrix},$$

which is approximately $m/d\, I_d$. Thus, $\theta_1, \ldots, \theta_{40}$ constitute a FUNTF in $\mathbb{R}^4$ by Lemma 2.5. Compared to that, the solution with $\alpha = 2.0$ from Figures 6.1 and 6.2 is, of course, not a FUNTF. In that case, the frame matrix is

$$S^{(2)} = \begin{bmatrix} 12.3593 & 0.8440 & 0.2668 & -0.3999 \\ 0.8440 & 10.0584 & -0.3908 & -0.7798 \\ 0.2668 & -0.3908 & 8.7939 & 0.7866 \\ -0.3999 & -0.7798 & 0.7866 & 8.7884 \end{bmatrix}$$

ID	Clustersize	IC var.	Expectations	p-values	adj. p-values	Significance	F-Statistics	t-Statistics	Neighbor	Distance
75	85	0.16701	6.50833	0	0	1	0.0989951	-1.73236	39	0.0818412
64	83	0.164314	9.78333	0	0	1	0.0973973	-1.70356	79	0.0976394
58	47	0.280375	16.3	7.36392e-11	5.89114e-09	1	0.166192	-0.814742	31	0.104828
42	42	1.11972	4.65833	0	0	1	0.663714	0.880755	20	0.152837
4	41	0.110776	7.18333	0	0	1	0.0656624	-1.59512	64	0.10373
70	33	0.633523	3.775	0	0	1	0.375521	1.40148	20	0.0861074
39	25	0.0596933	3.38333	3.88578e-15	3.10862e-13	1	0.0353832	-2.96206	79	0.0792985
36	24	0.361971	10.1417	4.99563e-05	0.0039965	1	0.214559	0.860502	74	0.0821995
54	24	0.129351	1.825	0	0	1	0.0766728	-1.64891	75	0.0961764
20	20	1.17091	4.2	3.64941e-09	2.91953e-07	1	0.694058	0.739648	70	0.0861074
40	16	0.202884	5.29167	3.62011e-05	0.00289609	1	0.12026	2.13012	70	0.0874667
32	13	0.917674	8.58333	0.0536551	1	0	0.543952	0.623607	64	0.0991393
47	13	0.29738	7.00833	0.0125032	1	0	0.176272	-0.342489	9	0.124688
31	12	0.113057	2.94167	1.23362e-05	0.0009869	1	0.0670148	-0.767419	75	0.090788
5	11	0.891773	4.11667	0.0011169	0.0893521	0	0.528599	1.44643	70	0.0960876
22	10	0.391968	2.7	0.000114764	0.00918115	1	0.232339	0.0377957	52	0.116403
65	9	0.207345	10.3917	0.591263	1	0	0.122904	1.5827	17	0.0963994
79	9	0.087322	4.80833	0.0249529	1	0	0.0517601	-1.37161	39	0.0792985
2	7	1.7258	7.38333	0.458509	1	0	1.02297	0.360513	78	0.0774808
19	7	0.247997	5.85833	0.236111	1	0	0.147	-1.56306	56	0.10654

Table 6.3: Statistics for the 20 largest clusters from PFP clustering with $m = 80$ (see Figure 6.10) including innercluster variance (IC var.) and distance to nearest neighbor

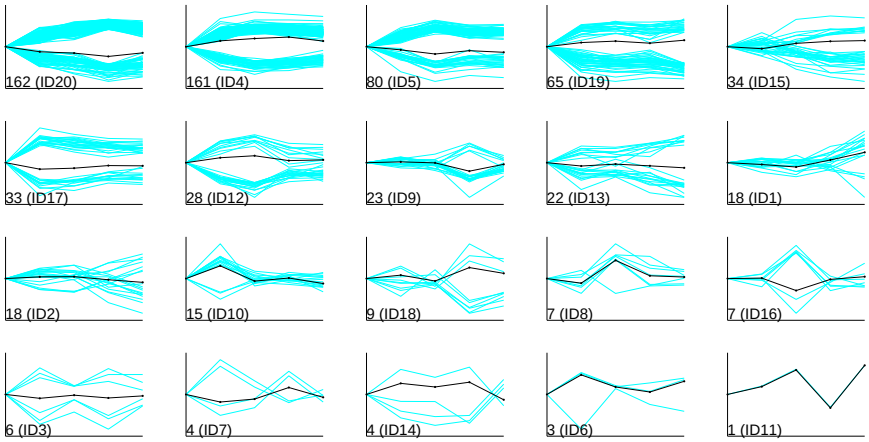


Figure 6.11: 20 largest clusters from PFP clustering with $m = 20$, $\alpha = 2.0$ for AAS example using the assignment rule in (6.2); prototypes (black), data (cyan)

and the frame potential becomes $\|S^{(2)}\|_F^2 \approx 413.14 > 400 \approx \|S^{(0)}\|_F^2$.

The second parameter that has strong influence on the solutions is the number of prototypes m. In Figure 6.10 we see that for $m = 80$ the sizes of the clusters decrease. However, many of the prototypes are similar which can be seen by the smaller distances between neighboring prototypes in Table 6.3 compared to the last column in Table 6.1.

In biological applications it is sometimes of interest to identify data that react contrarily to each other over time. In order to address this problem, we alter the step where the clusters are built. Instead of assigning a given time series to cluster C_ℓ if the corresponding projection satisfies

$$y_j \in \hat{D}_\ell \quad = \quad \left\{ v \in \mathcal{S}^{d-1} \, : \, \theta_\ell = \underset{1 \le k \le m}{\arg\max} \, \langle v, \theta_k \rangle \right\} ,$$

we assign it to cluster C_ℓ if

$$\theta_\ell \quad = \quad \underset{1 \le k \le m}{\arg\max} \, |\langle y_j, \theta_k \rangle| \, . \tag{6.2}$$

The clustering for $m = 20$ using the assignment rule from (6.2) can be seen in Figure 6.11. Note that in this application one is often interested in the rather small clusters.

6.4 Feature Recognition in Multispectral Data

In the final section, we briefly show how the PFP algorithm can be applied to extract certain features from multispectral data. The idea of extending our method to this field was originally suggested by Martin Ehler in personal conversation at the *International Conference on Multivariate Approximation* in Hagen, Germany, in 2011.

According to Jain, "in multispectral imaging there is a sequence of I images $U_i(m,n)$, $i = 1,2,\ldots,I$, where the number I is typically between 2 and 12. It is desired to combine these images to generate a single or a few display images that are representative of their features" ([Jain 89]). For the sake of consistency with the previous chapters, we use the variables d and k instead of I and m, respectively.

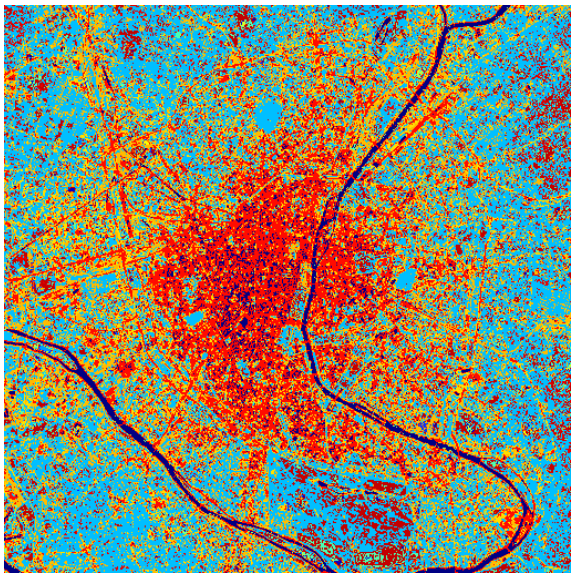
Multispectral data occurs for example in aerial scanning of landscapes using a small number of different frequency bands. The recorded data $U_1,\ldots,U_d$ are stored in a three-dimensional array, which is commonly denoted as a multispectral cube. The objective is to gain information on soil composition or agricultural land use, among others. Each slice of the cube $U_i(k,n)$, where $k = 1,\ldots,N_1$ and $n = 1,\ldots,N_2$, contains the spatial information of the measuring with the same frequency band. Thus, cutting the cube vertically along the frequency domain leads to a large number of $N_1 \cdot N_2$ vectors

$$U(1,1), U(2,1), \ldots, U(N_1,1), U(1,2), \ldots, U(N_1,2), \ldots, U(1,N_2), \ldots, U(N_1,N_2)$$

with $U(k,n) = (U_1(k,n),\ldots,U_d(k,n))^T$ in the low-dimensional space $\mathbb{R}^d$. Clustering with respect to the spatial position allows to extract certain features from the image.

The left picture in Figure 6.12 presents an aerial image of central Paris, France, produced by RGB visualization of three slices of a seven-layered multispectral cube of size $512 \times 512 \times 7$. Cutting the cube vertically leads to 2^{18} vectors in $\mathbb{R}^7$. The right picture is a colored image of the PFP-clustering using $m = 25$ and $\alpha = 8.0$. Each color represents one cluster. Several features in the landscape become prominent. For example, one can clearly identify the Seine (dark blue) or divide the downtown area (red) from the less-populated suburbs (cyan). Also bridges and larger outbound streets are recognized as similar objects and clearly visible.

It should be mentioned that standard algorithms in multispectral data processing provide



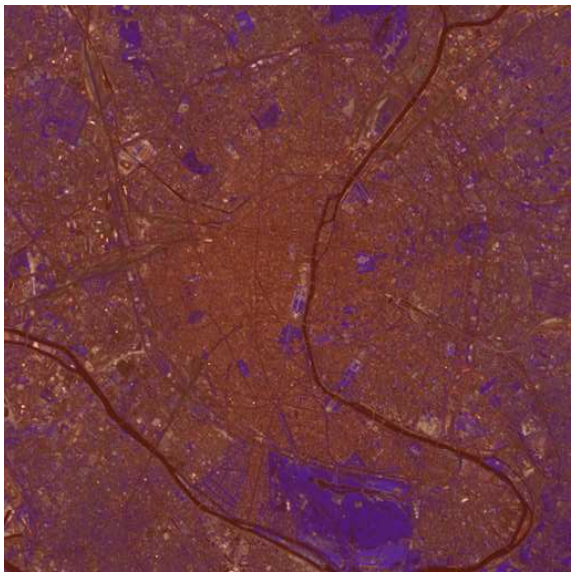


Figure 6.12: RGB image of Paris using three of seven layers from multispectral data of size $512 \times 512 \times 7$ (left), Color representation of the clustering using the PFP method with $m = 25$ and $\alpha = 8.0$ (right)

further features and options for this specific problem such as onboard-processing for satellites. The Multispec package (`https://engineering.purdue.edu/~biehl/MultiSpec/`, accessed July 21, 2013) offers a broad variety of applications. However, note that for comparative purposes the current implementation of the PFP method is capable of giving a solution to the feature extraction problem.

Chapter 7

Brief Discussion and Outlook

As seen in this thesis, the applied methods provide insight into different features of the Penalized Frame Potential. Compared to the first proposal in [Spri 11], we characterized the functional in terms of extremal cases of the regularization parameter α and showed the connection to the spherical Dirichlet cells of the normalizations of the given data. Furthermore, the thorough discussion in the framework of optimization theory proved that under mild relaxations the minimization can be transformed into a problem on the singular values with unique solution under suitable conditions.

However, a number of questions is still open and demands answers in future work:

(Q1) We formulated the PFP for the penalty term that looks for the maximal inner product between the given data and the minimizer Θ. Are there other penalty terms with similar features and reasonable calculational effort for the clustering problem? How can the penalty term be altered in order to adapt the basic concept of penalizing the frame potential to other problems?

(Q2) Chapter 5 showed that for the relaxed problem (P2$*$) there exists no duality gap. Moreover, Example 5.2 provides data on $\mathcal{S}^1$ for which we can characterize the minimizer and the Lagrange multipliers in (P2) and (P1) in terms of a single scalar. We also introduced a class of matrices that are similar to the given minimizer with larger objective values. Therefore, the question remains whether the corresponding optimization prob-

lems contain a duality gap. This question is closely connected to the problem of globally minimizing the PFP.

(Q3) The advantages of the problem (P2$*$) are that one only has to consider a single constraint and that the Wielandt-Hoffman-Theorem 5.4 provides a reduction to an optimization on the singular values. In all other stated problems, the left and right singular matrices of the minimizer T have to be taken into account as well, which in general heavily increases the difficulty of solving it. Is there a possibility to simplify or to find (approximate) solutions for the extremal condition

$$4TT^*T + 2T\Lambda \quad = \quad \alpha Y$$

for general Y ?

Note that there exist iterative methods to approximate solutions for the algebraic Riccati Equation

$$XDX + XA + BX + C \quad = \quad 0$$

with real or complex coefficient matrices A, B, C, D, which appears in filter design and control theory. However, Riccati-type equations are only of quadratic polynomial degree and formulated for quadratic X ([Lanc 95]).

In addition to these mathematical problems, the clustering results from Chapter 6 can be further extended. Even if we provided some practical aspects of the minimization of the PFP for real data, more evaluations with different data sets should be conducted. A future step of a more extensive performance evaluation should include an analysis of the clusters via software tools such as Prima, which is part of the so-called Expander package. It can support more elaborate research by identifying relevant connections within the clustering results of biological data.

Bibliography

[Absi 08] P.-A. Absil, R. Mahony and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[Akhi 66] N. I. Akhiezer and I. M. Glazman. *Theory of Linear Operators in Hilbert spaces*. Vol. 1, Frederick Ungar Publishing Co., New York, 1966.

[Batc 69] B. G. Batchelor and B. R. Wilkins. "Method for Location of Clusters of Patterns to initialise a Learning Machine". *Electronics Letters*, Vol. 5, No. 20, pp. 481–483, 1969.

[Baza 06] M. S. Bazaraa, H. D. Sherali and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, Hoboken, NJ, 3rd Ed., 2006.

[Belk 03] M. Belkin and P. Niyogi. "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation". *Neural Computation*, Vol. 15, No. 6, pp. 1373–1396, 2003.

[Bene 03] J. J. Benedetto and M. Fickus. "Finite Normalized Tight Frames". *Advances in Computational Mathematics*, Vol. 18, pp. 357–385, 2003.

[Bene 10] J. J. Benedetto, W. Czaja and M. Ehler. "Frame Potential Classification Algorithm for Retinal Data". In: K. E. Herold, W. E. Bentley and J. Vossoughi, Eds., *IFMBE Proceedings Series*, pp. 496–499, International Federation for Medical & Biological Engineering, Springer, College Park, Maryland, USA, 2010. 26th Southern Biomedical Engineering Conference.

[Boor 01] C. de Boor. "Calculation of the smoothing spline with weighted roughness measure". *Mathematical Models & Methods in Applied Sciences*, Vol. 11, No. 1, pp. 33–41, 2001.

[Cai 10] J. F. Cai, E. J. Candès and Z. Shen. "A Singular Value Thresholding Algorithm for Matrix Completion". *SIAM Journal on Optimization*, Vol. 20, No. 4, pp. 1956–1982, 2010.

[Cali 74] T. Caliński and J. Harabasz. "A dendrite method for cluster analysis". *Communications in Statistics*, Vol. 3, No. 1, pp. 1–27, 1974.

[Cand 05] E. J. Candès and T. Tao. "Decoding by linear programming". *IEEE Transactions on Information Theory*, Vol. 51, No. 12, pp. 4203–4215, 2005.

[Cand 10] E. J. Candès and T. Tao. "The power of convex relaxation: near-optimal matrix completion". *IEEE Transactions on Information Theory*, Vol. 56, No. 5, pp. 2053–2080, 2010.

[Casa 02a] P. G. Casazza and M. T. Leon. "Existence And Construction Of Finite Tight Frames". Tech. Rep., 2002.

[Casa 02b] P. G. Casazza and M. T. Leon. "Frames With A Given Frame Operator". Tech. Rep., 2002.

[Casa 03] P. G. Casazza and J. Kovačević. "Equal-Norm Tight Frames with Erasures". *Advances in Computational Mathematics*, Vol. 18, No. 2–4, pp. 387–430, 2003.

[Casa 04] P. G. Casazza. "Custom building finite frames". *Contemporary Mathematics*, Vol. 345, pp. 61–86, 2004.

[Casa 09] P. G. Casazza and M. Fickus. "Minimizing Fusion Frame Potential". *Acta Applied Mathematics*, Vol. 107, pp. 7–24, 2009.

[Casa 13] P. G. Casazza and G. Kutyniok. *Finite Frames: Theory and Applications. Applied and Numerical Harmonic Analysis*, Birkhäuser, Boston, 2013.

[Chri 08] O. Christensen. *Frames and Bases: An Introductory Course. Applied and Numerical Harmonic Analysis*, Birkhäuser, Boston, 2008.

[Chui 92] C. K. Chui. *An Introduction to Wavelets*. Vol. 1 of *Wavelet Analysis and Its Applications*, Academic Press, Boston, 1992.

[Conw 93] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Springer, New York, 1993.

[Conw 96] J. H. Conway, R. H. Hardin and N. J. A. Sloane. "Packing Lines, Planes, etc., Packings in Grassmannian Spaces". *Experimental Mathematics*, Vol. 5, pp. 139–159, 1996.

[Czaj 13] W. Czaja and M. Ehler. "Schroedinger Eigenmaps for the Analysis of Bio-Medical Data". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 5, pp. 1274–1280, 2013.

[dAsp 08] A. d'Aspremont, F. Bach and L. El Ghaoui. "Optimal Solutions for Sparse Principal Component Analysis". *Journal of Machine Learning Research*, Vol. 9, pp. 1269–1294, 2008.

[Daub 92] I. Daubechies. *Ten Lectures on Wavelets. Regional Conference Series in Applied Mathematics*, SIAM, 1st Ed., 1992.

[Dono 06] D. L. Donoho. "Compressed sensing". *IEEE Transactions on Information Theory*, Vol. 52, No. 4, pp. 1289–1306, 2006.

[Dono 94] D. L. Donoho and I. M. Johnstone. "Ideal Spatial Adaption by Wavelet Shrinkage". *Biometrika*, Vol. 81, No. 3, pp. 425–455, 1994.

[Duff 52] R. J. Duffin and A. C. Schaeffer. "A Class of Nonharmonic Fourier Series". *Transactions of the American Mathematical Society*, Vol. 72, No. 2, pp. 341–366, 1952.

[Ehle 11a] M. Ehler and J. Galanis. "Frame theory in directional statistics". *Statistics & Probability Letters*, Vol. 81, No. 8, pp. 1046–1051, 2011.

[Ehle 11b] M. Ehler, V. N. Rajapakse, B. R. Zeeberg, B. P. Brooks, J. Brown, W. Czaja and R. F. Bonner. "Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development". *BMC Proceedings*, Vol. 5 Suppl 2, 2011.

[Ehle 12a] M. Ehler. "Random tight frames". *Journal of Fourier Analysis and Applications*, Vol. 18, No. 1, pp. 1–20, 2012.

[Ehle 12b] M. Ehler and K. A. Okoudjou. "Minimization of the p-th probabilistic frame potential". *Journal of Statistical Planning and Inference*, Vol. 142, No. 3, pp. 645–659, 2012.

[Erns 05] J. Ernst, G. J. Nau and Z. Bar-Joseph. "Clustering short time series gene expression data". *Bioinformatics*, Vol. 21, pp. i159–i168, 2005.

[Frit 10] A. Fritsch. *Bayesian Mixtures for Cluster Analysis and Flexible Modeling of Distributions*. PhD thesis, Department of Statistics, Technische Universität Dortmund, 2010.

[Gasc 00] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown. "Genomic expression programs in the response of yeast cells to environmental changes". *Molecular Biology of the Cell*, Vol. 11, pp. 4241–4257, 2000.

[Goya 01] V. Goyal, J. Kovačević and J. A. Kelner. "Quantized Frame Expansions with Erasures". *Applied and Computational Harmonic Analysis*, Vol. 10, pp. 203–233, 2001.

[Goya 98] V. Goyal, M. Vetterli and N. T. Thao. "Quantized Overcomplete Expansions in $\mathbb{R}^n$: Analysis, Synthesis and Algorithms". *IEEE Transactions on Information Theory*, Vol. 44, No. 1, pp. 16–31, 1998.

[Han 00a] D. Han and D. R. Larson. *Frames, Bases and Group Representations*. Vol. 147, Memoirs of the AMS, Providence, RI, 2000.

[Han 00b] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[Hast 01] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning. Springer Series in Statistics*, Springer, New York, 1st Ed., 2001.

[Herm 00] J. Herman, R. Kučera and J. Šimša. *Equations and Inequalities: Elementary Problems and Theorems in Algebra and Number Theory. CMS Books in Mathematics*, Springer, New York, 2000.

[Hern 96] E. Hernández and G. Weiss. *A first course on wavelets*. CRC Press, 1996.

[Hoch 00] B. M. Hochwald, T. L. Marzetta, T. J. Richardson, W. Sweldens and R. L. Urbanke. "Systematic design of unitary space-time constellations". *IEEE Transactions on Information Theory*, Vol. 46, No. 6, pp. 1962–1973, 2000.

[Horn 96] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1st Ed., 1996.

[Hube 85] L. Hubert and P. Arabie. "Comparing Partitions". *Journal of Classification*, Vol. 2, pp. 193–218, 1985.

[Jain 89] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[Kim 07] J. Kim and J. H. Kim. "Difference-based clustering of short time-course microarray data with replicates". *BMC Bioinformatics*, Vol. 8, pp. 253–264, 2007.

[Kova 07a] J. Kovačević and A. Chebira. "Life beyond bases: The advent of frames (Part 1)". *IEEE Signal Processing Magazine*, Vol. 24, No. 4, pp. 86–104, 2007.

[Kova 07b] J. Kovačević and A. Chebira. "Life beyond bases: The advent of frames (Part 2)". *IEEE Signal Processing Magazine*, Vol. 24, No. 5, pp. 115–125, 2007.

[Lanc 95] P. Lancaster and L. Rodman. *Algebraic Riccati Equations. Oxford Science Publications*, Clarendon Press, 1995.

[Lass 10] J. B. Lasserre. *Moments, Positive Polynomials and Their Applications*. Vol. 1 of *Imperial College Press Optimization Series*, Imperial College Press, 2010.

[Lass 11] J. B. Lasserre. "The Moment-SOS approach in Global Optimization". URL: `http://www.lnmb.nl/conferences/2011/programlnmbconference` (Accessed December 18th, 2012), January 2011. Conference Talk.

[Lass 12] J. B. Lasserre. Personal communication, 2012.

[Liao 05] T. W. Liao. "Clustering of time series data - a survey". *Pattern Recognition*, Vol. 38, No. 11, pp. 1857–1874, 2005.

[Mazu 10] R. Mazumder, T. Hastie and R. Tibshirani. "Spectral Regularization Algorithms for Learning Large Incomplete Matrices". *Journal of Machine Learning Research*, Vol. 11, pp. 2287–2322, 2010.

[Meye 12] C. Meyer. Personal communication, 2012.

[Qu 03] Y. Qu, A. Bao-Ling *et al.* "Data Reduction Using a Discrete Wavelet Transform in Discriminant Analysis of Very High Dimensionality Data". *Biometrics*, Vol. 59, pp. 143–151, 2003.

[Rein 67] C. H. Reinsch. "Smoothing by spline functions". *Numerische Mathematik*, Vol. 10, pp. 177–183, 1967.

[Rock 93] R. T. Rockafellar. "Lagrange Multipliers and Optimality". *SIAM*, Vol. 35, No. 2, pp. 183–238, 1993.

[Saff 97] E. Saff and A. B. J. Kuijlaars. "Distributing Many Points on a Sphere". *The Mathematical Intelligencer*, Vol. 19, pp. 5–11, 1997.

[Scho 64] I. J. Schoenberg. "Spline functions and the problem of graduation". *Proceedings of the American Mathematical Society*, Vol. 52, pp. 947–950, 1964.

[Shen 13] Z. Shen. Personal communication, 2013.

[Spri 11] T. Springer, K. Ickstadt and J. Stöckler. "Frame potential minimization for clustering short time series". *Advances in Data Analysis and Classification*, Vol. 5, No. 4, pp. 341–355, 2011.

[Stro 03] T. Strohmer and R. W. Heath Jr. "Grassmannian frames with applications to coding and communications". *Applied Computational Harmonic Analysis*, Vol. 14, No. 3, pp. 257–275, 2003.

[Tamm 30] P. M. L. Tammes. *On the origin of number and arrangement of the places of exit on pollen grains*. PhD thesis, Groningen, 1930.

[Thom 04] J. J. Thomson. "On the Structure of the Atom: an Investigation of the Stability and Periods of Oscillation of a number of Corpuscles arranged at equal intervals around the Circumference of a Circle; with Application of the Results to the Theory of Atomic Structure". *Philosophical Magazine Series 6*, pp. 237–265, 1904.

[Tibs 11] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso: a Retrospective". *Journal of the Royal Statistical Society, Series B*, Vol. 73, pp. 273–282, 2011.

[Tibs 94] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistical Society, Series B*, Vol. 58, pp. 267–288, 1994.

[Vlac 03] M. Vlachos, J. Lin, E. Keogh and D. Gunopulos. "A Wavelet-Based Anytime Algorithm for k-Means Clustering of Time Series". In: *Proceedings of the Workshop on Clustering High Dimensionality Data and Its Applications*, pp. 23–30, 2003.

[Welc 74] L. R. Welch. "Lower Bounds on the Maximum Cross Correlation of Signals". *IEEE Transactions on Information Theory*, Vol. 20, pp. 397–399, 1974.

[Zimm 01] G. Zimmermann. "Normalized Tight Frames in Finite Dimensions". *Recent Progress in Multivariate Approximation*, Vol. 17, pp. 249–252, 2001.