Linear and Nonlinear Inverse Problems in Aerosol Spectroscopy

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Dipl.-Math. Tobias Kyrion

aus Köln

Berichter: Univ.-Prof. Dr. rer. nat. Martin Frank
Univ.-Prof. Dr. rer. nat. Michael Herty

Tag der mündlichen Prüfung: 14.11.2017

Acknowledgements

First and foremost I owe my deepest gratitude to Graham Alldredge, Ph.D. This work gained a lot of quality from your constant interest in and constructive criticism of my ideas, even when you did not have any obligation to work with me any more. If you ever happen to be a professor, it is fair to say that I was your first inofficial Ph.D. student.

Ich danke Herrn Prof. Dr. Hans-Josef Allelein, Herrn Dipl.-Phys. Alexander Ehrlich sowie Herrn Dipl.-Ing. Björn Alexander Krupa und allen anderen beteiligten Projektpartnern für die gelungene Zusammenarbeit im In-Ex Verbundprojekt.

Ich danke meinen Eltern Brigitte und Rolf Friedel Kyrion, meiner Schwester Larissa und meinem Bruder Nils, sowie meiner Großmutter Helga Arnolds für ihren moralischen Beistand während meiner Promotionszeit.

Schließlich danke ich recht herzlich Herrn Prof. Dr. Martin Frank und Herrn Prof. Dr. Michael Herty für die Korrektur dieser Arbeit.

Abstract

In this work we study the evaluation of optical aerosol measurements. Our aim is to reconstruct the size distributions of aerosol particles from optical light extinction measurements in order to obtain a safe measurement technology for potentially harmful aerosols inside a nuclear reactor containment.

The first half of this work is devoted to linear inverse problems. In particular we study the linear integral equation relating aerosol particle size distributions to optical extinction measurements via Mie theory. We derive reconstruction algorithms which work independently from a human operator and thus do not require any monitoring or further adjustments. Based on statistical observations, we derive residual-based methods for finding the appropriate number of discretization points and the regularization parameter for Tikhonov regularization. Since particle size distributions are nonnegative, we apply nonnegativity constraints throughout the whole reconstruction process and all results are derived for constrained regression problems. A special emphasis lies on computational efficiency, since we demand that a single inversion must be completed in less than thirty seconds on a regular notebook.

We compare our method based on the discrepancy principle with a Monte Carlo inversion method, where we also apply nonnegativity constraints. Here the regularization parameter is considered as a model variable and retrieved together with the sought-after size distributions.

Then the discrepany principle strategy is generalized to the case of two-component aerosols, where the aerosol particle material is a mixture of two pure component materials. In addition to the particle size distribution, we retrieve the unknown mixing ratio of the two components.

In the second half of this work we study the nonlinear inverse problem of reconstructing the refractive indices of an aerosol material from measurements of monodisperse aerosols. First we investigate this problem for a fixed light wavelength. We take into account all local minima found here and regard them all as candidate solutions. Then we apply a selection method based on smoothness estimates for refractive index curve sections covering consecutive light wavelengths. The resulting coupled refractive index reconstructions are regularized further using Phillips-Twomey regularization.

Contents

Survey

In Chapter 1 we give a brief introduction into Mie theory on which this work is based on.

Chapter 2 introduces the FASP measurement device. Then we proceed with the mathematical modeling of FASP measurements. We give monotonicity results for the residual of Tikhonov-regularized solutions depending on the regularization parameter. These results are the basis for the discrepancy principle. We also give convergence results for Tikhonov regularization under linear constraints, where the regularized solutions are shown to converge to the true sought-after solution as the noise level approaches zero. Finally we present the Bayesian model selection mechanism for the candidate solutions obtained from the discrepancy principle using a whole set of Morozov safety factors.

In Chapter 3 we compare our retrieval method with established inversion methods in a numerical study with artificial measurement data. Here we solve the forward problem, i.e. our integral equation from Mie theory, with high precision and add zero-mean Gaussian noise to it. We used H_2O as aerosol particle material and air as surrounding medium. The original particle size distributions are from the log-normal, Rosin-Rammler-Sperling-Bennett (RRSB) and Hedrih distribution families.

Chapter 4 is devoted to Monte Carlo inversion methods. We regard the regularization parameter as additional model parameter here using so-called hyperpriors. We develop a Bayesian model selection algorithm for this new problem. For the selected model we perform a Monte Carlo inversion based on a Gibbs sampler. At the end of this chapter we give the results of the Monte Carlo method for the same numerical study from Chapter 3.

In Chapter 5 the results from Chapter 2 are generalized to the case of two-component aerosols. Here we have to identify the correct model depending on the mixing ratio of the two components. We show that we obtain a convergent method, if we select the model with the smallest residual of the unregularized solution, i.e. the best-fitting model.

Chapter 6 contains numerical results for the two-component retrieval algorithm, where various mixing ratios of H_2O and CsI are used. We applied the same original particle size distributions from Chapter 3 to compute our artificial measurement data.

Chapter 7 contains the mathematical treatment of the refractive index reconstruction from FASP measurements of monodisperse aerosols. We begin with convergence and stability results for nonlinear inverse problems, where the model is given as truncated series expansion approximating an infinite series. We first investigate the reconstruction problem for a fixed wavelength. Based on the local minima of the fit function for a single wavelength, we filter out coupled solutions for neighboring wavelengths by minimizing the sum of its squared second finite differences. These coupled solutions are used as start vectors to solve the nonlinear coupled regression problem, where we apply Phillips-Twomey regularization.

In Chapter 8 we present the result of numerical studies for the reconstruction algorithm developed in Chapter 7, where we used the refractive indices of Ag, CsI and H_2O as sought-after original refractive indices of aerosol particle materials.

Statement regarding the good scientific practice

This work was written by Tobias Kyrion - in the following, the author of this work - during his employment at the Chair for Mathematics (CCES) at RWTH Aachen University, between October 2012 and October 2015 and the following one and a half years under the supervision of Prof. Martin Frank. Whenever we make use of existing works by different authors, this is explicitly indicated and we carefully cite the corresponding sources. All computer codes that were used to produce numerical results in this work were written by the author of this work, except the code to compute truncated multivariate normal probabilities written by Alan Genz.

Chapters 2, 3 and 5 were published in [1]. The numerical results in Chapters 3 and 6 were rerun, and thus differ slightly from the numerical results in [1]. The author was introduced into Bayesian statistics by Graham Alldredge, Ph.D., who contributed Section 2.6.2.

Chapters 7 and 8 were published in [2]. The results in Sections 8.14.1 - 8.15.2 were rerun and thus differ slightly from the corresponding results in [2].

Chapter 1

Mie Theory

1.1 Scattering of Light by a Spherical Particle

Gustav Mie (1868 - 1957) found in his famous article [3] from 1908 the exact solution for Maxwell's equation

$$\begin{aligned}\operatorname{div}(\varepsilon \boldsymbol{e}) &= 0,\\ \operatorname{div}(\boldsymbol{b}) &= 0,\\ \operatorname{curl}(\boldsymbol{e}) &= i\omega\mu_0 \boldsymbol{h}\\ \operatorname{curl}(\boldsymbol{h}) &= (-i\varepsilon\varepsilon_0\omega + \sigma)\boldsymbol{e}\end{aligned}$$

for a spherical particle illuminated by light, where ε_0 is the vacuum permittivity, μ_0 the vacuum permeability and

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{e}(\boldsymbol{r}(t))\exp(-i\omega t)\\ \boldsymbol{h} &= \boldsymbol{h}(\boldsymbol{r}(t))\exp(-i\omega t)\end{aligned}$$

are the field components of the harmonic electromagnetic waves under consideration, cf. [4, p. 58]. Here $\boldsymbol{r}(t)$ parameterizes the current location. Furthermore we have $\boldsymbol{b} = \mu_0 \boldsymbol{h}$.

Now with above ansatz, we model the fields $\boldsymbol{e}_{inc}$, $\boldsymbol{h}_{inc}$ of the *incident wave*, the fields $\boldsymbol{e}_{int}$, $\boldsymbol{h}_{int}$ of the waves inside the particle interior and the fields $\boldsymbol{e}_{sca}$, $\boldsymbol{h}_{sca}$ of the scattered waves. We then expand the three fields in *vector spherical harmonics* which are obtained from scalar spherical harmonics. The coefficients of the vector spherical harmonics expansions for the interior and scattered fields are referred to as *Mie coefficients*. They are obtained from Maxwell's boundary conditions on the surface of the sphere

$$\begin{aligned}(\boldsymbol{e}_{inc} + \boldsymbol{e}_{sca}) \times \boldsymbol{n}_r &= \boldsymbol{e}_{int} \times \boldsymbol{n}_r\\ (\boldsymbol{h}_{inc} + \boldsymbol{h}_{sca}) \times \boldsymbol{n}_r &= \boldsymbol{h}_{int} \times \boldsymbol{n}_r,\end{aligned}$$

where $\boldsymbol{n}_r$ is the surface normal vector of a sphere with radius r.

We recapitulate Mie theory in an absorbing medium as presented in [5]. Our first step is to introduce the complex-valued *Riccati-Bessel-functions* $\xi_n : \mathbb{C} \to \mathbb{C}$ and $\psi_n : \mathbb{C} \to \mathbb{C}$ given by

$$\xi_n(z) = \sqrt{\tfrac{\pi}{2}}\sqrt{z}J_{n+\frac{1}{2}}(z) \tag{1.1.1}$$

and

$$\psi_n(z) = \sqrt{\tfrac{\pi}{2}}\sqrt{z}J_{n+\frac{1}{2}}(z) + i\sqrt{\tfrac{\pi}{2}}\sqrt{z}Y_{n+\frac{1}{2}}(z), \tag{1.1.2}$$

with the Bessel functions $J_{n+\frac{1}{2}} : \mathbb{C} \to \mathbb{C}$ and $Y_{n+\frac{1}{2}} : \mathbb{C} \to \mathbb{C}$ of order $n + \frac{1}{2}$ of first and second kind. We define the *size parameter* $\rho = 2\pi\frac{r}{l}$. Then we set $z_{med} := \rho \cdot m_{med}$ and $z_{part} := \rho \cdot m_{part}$. Here and in the following we omit the wavelength dependence of m_{med} and m_{part} for better readability. We introduce the notation $m_{med} = n_{med} + ik_{med}$ and $m_{part} = n_{part} + ik_{part}$.

We introduce the so-called *Mie coefficients*:

$$\begin{aligned}
a_n &:= \frac{m_{part}\dot{\xi}_n(z_{med})\xi_n(z_{part}) - m_{med}\xi_n(z_{med})\dot{\xi}_n(z_{part})}{m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{part}) - m_{med}\psi_n(z_{med})\dot{\xi}_n(z_{part})} \\[2em]
b_n &:= \frac{m_{part}\xi_n(z_{med})\dot{\xi}_n(z_{part}) - m_{med}\dot{\xi}_n(z_{med})\xi_n(z_{part})}{m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{part}) - m_{med}\dot{\psi}_n(z_{med})\xi_n(z_{part})} \\[2em]
c_n &:= \frac{m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{med}) - m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{med})}{m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{part}) - m_{med}\dot{\psi}_n(z_{med})\xi_n(z_{part})} \\[2em]
d_n &:= \frac{m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{med}) - m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{med})}{m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{part}) - m_{med}\psi_n(z_{med})\dot{\xi}_n(z_{part})}
\end{aligned} \tag{1.1.3}$$

With the Mie coefficients we can express the coefficient functions

$$A_n(\rho, m_{med}, m_{part}) := \frac{l}{2\pi m_{part}}\left(|c_n|^2\,\xi_n(z_{part})\overline{\dot{\xi}_n(z_{part})} - |d_n|^2\,\dot{\xi}_n(z_{part})\overline{\xi_n(z_{part})}\right)$$

and

$$B_n(\rho, m_{med}, m_{part}) := \frac{l}{2\pi m_{med}}\left(|a_n|^2\,\dot{\psi}_n(z_{med})\overline{\psi_n(z_{med})} - |b_n|^2\,\psi_n(z_{med})\overline{\dot{\psi}_n(z_{med})}\right),$$

which finally occur in the series expansion of the *Mie extinction efficiency*

$$Q_{ext}(r, l, m_{med}, m_{part}) = \frac{l}{2cI(r,l)}\sum_{n=1}^{\infty}(2n+1)\mathrm{Im}\big(A_n(\rho, m_{med}, m_{part}) + B_n(\rho, m_{med}, m_{part})\big). \tag{1.1.4}$$

Here the quantity $I(r, l)$ is the *average incident intensity* of light with wavelength l for a spherical particle with radius r and c denotes the speed of light in vacuum. The function $I(r, l)$ is given by

$$\begin{aligned}
I(r,l) &= \frac{l^2}{8\pi(k_{med})^2}\frac{n_{med}}{2c}\left(1 + \left(4\pi k_{med}\frac{r}{l} - 1\right)e^{4\pi k_{med}\frac{r}{l}}\right), && \text{for } k_{med} \neq 0 \\
I(r,l) &= \pi r^2\frac{n_{med}}{2c}, && \text{for } k_{med} = 0.
\end{aligned} \tag{1.1.5}$$

Obviously we cannot evaluate (1.1.4) exactly, because we cannot compute an infinite sum due to limited processing resources. Therefore we have to truncate this series expansion. In [6] a commonly used truncation index N_{trunc} is presented, which is given by

$$\begin{aligned} N_{trunc} &= \lceil\, |M\ + 4.05 \cdot M^{\frac{1}{3}} + 2|\, \rceil, \\ \text{with}\ \ M &= \max\lceil |\rho|\,, |\rho \cdot m_{med}|, |\rho \cdot m_{part}| \rceil. \end{aligned} \tag{1.1.6}$$

Chapter 2

Retrieval of Aerosol Particle Size Distributions

2.1 The FASP Measurement Device

The FASP is an optical measurement device for aerosol particle size distributions in rigid environments where the temperature may surpass 200°C and the pressure 8 bar over atmospheric pressure, cf. [7, 8]. The aerosol particles themselves may be acidic as well. The FASP is split into a detector head and into a unit containing an evaluation computer and a light source with different adjustable light wavelengths. The sensitive evaluation and light source unit is connected with the robust detector head via two optical fibers.

The detector head is the only part of the FASP which extends into the containment with the aerosol to be measured and it consists of a pneumatically propelled tube. By moving the tube one can adjust a short or a long measurement path, where the two path lengths are 400 and 800 mm respectively. The sought-after aerosol particle size distributions are reconstructed from the light intensity loss on the gap distance between long and short path, so the FASP works in a similar way to a White cell. The detector head is equipped with two light detectors. The first one can receive light with wavelengths in the infrared domain from 0.8 - 3.4 μm, and the other one in the visible domain from 0.5 - 0.8 μm.

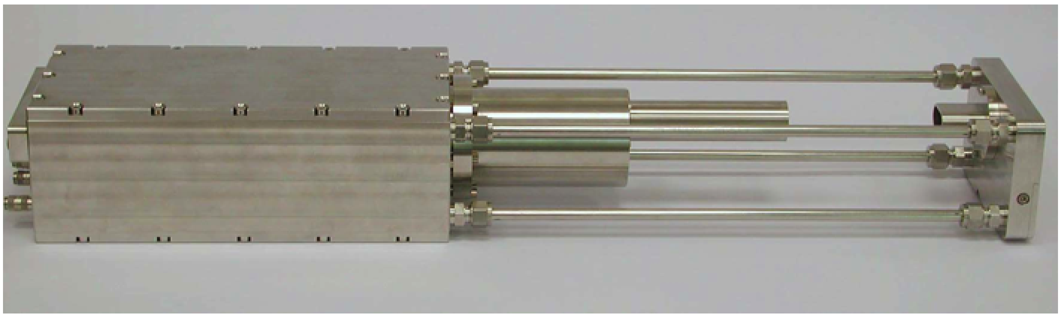


Figure 2.1: The detector head with the movable tube (source: [9])

The ends of the optical fibers have to be floated with protective gas to shield them from harmful aerosol particles. These particle-free sections have to be subtracted from the actual geometric path lengths. This is not problematic since this does not change the gap distance.

Let l denote a current light wavelength used in a measurement, G_{long} the geometric or unfloated long path and G_{short} the geometric short path. The section floated with protective gas is labeled with x. Then the true path lengths are given by $L_{long} := G_{long} - x$ and $L_{short} := G_{short} - x$.

Let $M_{long}(l)$ and $M_{short}(l)$ be the measured intensities for long and short path, both perturbed by detector offsets $O_{long}(l)$ and $O_{short}(l)$ caused by ambient radiation.

Then the intensities cleaned from the detector offsets are given by $I_{long}(l) := M_{long}(l) - O_{long}(l)$ and $I_{short}(l) := M_{short}(l) - O_{short}(l)$.

According to the law of Beer-Lambert we have the relation

$$I_{long}(l) = I_{short}(l) \exp\left(-(L_{long} - L_{short}) \int_0^\infty k(r,l)n(r)dr \right), \tag{2.1.1}$$

where $n(r)$ is the sought-after unknown particle size distribution. The kernel function $k(r,l) := \pi r^2 Q_{ext}(m_{med}(l), m_{part}(l), r, l)$ depends on both complex refractive indices $m_{med}(l)$ and $m_{part}(l)$ of the surrounding medium and the scattering aerosol particles which depend on the wavelength l of the incident light. The Mie extinction efficiency $Q_{ext}(m_{med}(l), m_{part}(l), r, l)$ of a spherical particle with radius r illuminated by light with wavelength l is derived from the general solution to the corresponding boundary value problem for Maxwell's equations and was first introduced in the pioneering article [3]. We adopt the numerical approximation of the Mie extinction efficiency in an absorbing medium from [5]. From all of this follows

$$\int_0^\infty k(r,l)n(r)dr = e(l) \quad \text{with} \quad e(l) = -\frac{\log(I_{long}(l)) - \log(I_{short}(l))}{L_{long} - L_{short}}. \tag{2.1.2}$$

2.2 Modeling of FASP Measurement Data Inversions

Let the measurement data $e(l)$ be an error-contaminated right-hand side for (2.1.2) and $(Kn)(l) := \int_0^\infty k(r,l)n(r)dr$ the compact linear operator with unbounded inverse which maps possible size distributions $n(r)$ to the left-hand side of (2.1.2). We wish to reconstruct $n(r)$ from $e(l)$ by inverting the equation

$$Kn = e. \tag{2.2.1}$$

Here and in the following we omit the dependence on r and l for better readability. We assume that e is given as a vector of finitely many independent Gaussian random variables e_i with standard deviations σ_i and means μ_i, i.e. $e_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. In the framework of Bayesian inference these are our *observed random variables*. Now let $\boldsymbol{n} \in \mathbb{R}^N$ be a discrete approximation to n and $\boldsymbol{K}_N \in \mathbb{R}^{N_l \times N}$ the *kernel matrix* which correspondingly approximates the integral operator K. The details of these discretizations will be given in Section 3.1. We set up the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\sigma}} = \operatorname{diag}(\sigma_1^2, ..., \sigma_{N_l}^2)$. Then the *observed model uncertainty* under the assumption $(\boldsymbol{K}_N \boldsymbol{n})_i = \mu_i$ obeys the probability distribution

$$p_{observed}(\boldsymbol{e}|\boldsymbol{n}) \propto \exp(-\tfrac{1}{2} \|\boldsymbol{\Sigma}_{\boldsymbol{\sigma}}{}^{-\frac{1}{2}}(\boldsymbol{K}_N \boldsymbol{n} - \boldsymbol{e})\|_2^2). \tag{2.2.2}$$

After selecting a subjective *prior distribution* $p_{prior}(\boldsymbol{n})$ which incorporates known *a priori* information about $\boldsymbol{n}$ independent from the observed variable $\boldsymbol{e}$ we use Bayes' rule to obtain the *posterior distribution* $p_{posterior}(\boldsymbol{n}|\boldsymbol{e})$ with

$$p_{posterior}(\boldsymbol{n}|\boldsymbol{e}) \propto p_{observed}(\boldsymbol{e}|\boldsymbol{n}) \times p_{prior}(\boldsymbol{n}). \tag{2.2.3}$$

A more elaborate presentation of this Bayesian framework will be given in Section 2.6.2. By applying a Tikhonov prior distribution

$$p_{prior}(\boldsymbol{n}) \propto \exp(-\tfrac{1}{2}\gamma\|\boldsymbol{n}\|_2^2)I_S(\boldsymbol{n}),$$

where $\gamma \geq 0$ is a regularization parameter and $I_S(\boldsymbol{n})$ is the indicator function of the convex set

$$S := \{\boldsymbol{n} \in \mathbb{R}^N |\ \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b}\} \quad \text{with} \quad \boldsymbol{C} \in \mathbb{R}^{k\times N},\ \boldsymbol{b} \in \mathbb{R}^k,$$

we obtain the posterior distribution

$$p_{posterior}(\boldsymbol{n}) \propto \exp(-\tfrac{1}{2}\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n} - \boldsymbol{e})\|_2^2 - \tfrac{1}{2}\gamma\|\boldsymbol{n}\|_2^2))I_S(\boldsymbol{n}). \tag{2.2.4}$$

The quantity of interest $\boldsymbol{n}$ is estimated by computing the maximizer of the posterior distribution which is called the *maximum a posteriori estimator* (MAP). It is obtained by solving the quadratic programming problem

$$\boldsymbol{n}_{MAP}^{\gamma} := \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n} - \boldsymbol{e})\|_2^2 + \tfrac{1}{2}\gamma\|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b}. \tag{2.2.5}$$

Note that $\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n}-\boldsymbol{e})\|_2^2 \sim \chi^2(N_l)$, which gives $\mathbb{E}(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n}-\boldsymbol{e})\|_2^2) = N_l$.

For a good introduction to Bayesian modeling, see [10].

A classical residual-based inference method is the so-called *discrepancy principle*. After selecting a *Morozov safety factor* τ the regularization parameter γ is determined by demanding $\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n} - \boldsymbol{e})\|_2^2 = \tau N_l$. A common choice for the safety factor is $\tau = 1.1$. We will give a more thorough introduction to the discrepancy principle and some results on it in Section 2.3.

Monte Carlo methods offer another way to evaluate the posterior distribution, cf. [11]. The advantage of Monte Carlo methods is that they take more of the statistical behavior of the observed measurement noise into account because all possible solutions with nonnegligible posterior probabibility are sampled and contribute to the inference result. However these methods require a lot of computational resources, which we cannot afford because our application requires that one FASP measurement data inversion must be completed in under thirty seconds using a regular notebook.

In our hybrid approach we combine the advantages of both methods. We review Tikhonov regularization under linear constraints and derive conditions for the existence of a bijection between the regularization parameter and the residual. If these conditions are fulfilled, we can propose a set of regularization parameters obtained with the discrepancy principle using a set of Morozov safety factors corresponding to high-probability values of the weighted norm of the residual, $\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n} - \boldsymbol{e})\|_2^2 \sim \chi^2(N_l)$. After this a Bayesian model-comparison procedure is applied to these reconstructions, and we rank them according to their posterior probabilities.

We show with numerical simulations that our method satisfies the demands on runtime and accuracy and that it is superior to existing inversion methods based on classical model-selection approaches. In the last section we extend our method to investigate two-component aerosols.

2.3 Tikhonov Regularization under Linear Constraints

Computing the maximum a posteriori estimator leads to a quadratic programming problem of the form

$$\boldsymbol{n}_\gamma := \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{n}-\boldsymbol{r}\|_2^2 + \tfrac{1}{2}\gamma\|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b}, \tag{2.3.1}$$

with $\boldsymbol{K} := \boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}\boldsymbol{K}_N$ and $\boldsymbol{r} := \boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}\boldsymbol{e}$. The function to be minimized is known as the Tikhonov functional.

It is proved in [12] that the residual of the Tikhonov-regularized solution under linear constraints decreases monotonically with the regularization parameter γ. To the best of our knowledge conditions for *strict* monotonicity have not been found yet, so we derive some in the following. The advantage of having a strictly monotonic relation between regularization parameter and residual is that it gives a bijection. Thus we can then identify any regularization parameter γ from the range $[0,\infty)$ with a unique residual value $\|\boldsymbol{K}\boldsymbol{n}_\gamma-\boldsymbol{r}\|_2^2$ from the range $[\|\boldsymbol{K}\boldsymbol{n}_0-\boldsymbol{r}\|_2^2, \|\boldsymbol{K}\boldsymbol{n}_\infty-\boldsymbol{r}\|_2^2)$. Here

$$\boldsymbol{n}_\infty := \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b} \tag{2.3.2}$$

is the *minimum norm element*. As shown in [13] there holds $\lim_{\gamma\to\infty}\boldsymbol{n}_\gamma = \boldsymbol{n}_\infty$. When our monotonicity conditions are satisfied, we obtain a set of distinct regularization parameters by proposing a set of distinct residual values from the range $[\|\boldsymbol{K}\boldsymbol{n}_0-\boldsymbol{r}\|_2^2, \|\boldsymbol{K}\boldsymbol{n}_\infty-\boldsymbol{r}\|_2^2)$. The disadvantageous case of multiple prior distributions corresponding to the same residual value can therefore not occur. Note that in practice the cases $\gamma = 0$ and $\gamma = \infty$ are inadmissible, since then the Tikhonov prior distribution is improper or degenerates to a point mass, so we always restrict ourselves to a finite range $(0, \gamma_{max}]$ with $\gamma_{max} < \infty$.

2.3.1 Necessary Conditions for Strict Monotonicity

The following theorem shows that $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$ for all $\alpha > \beta$ is the only necessary condition needed for strict monotonicity.

Lemma 2.3.1. *Let $\alpha > \beta \geq 0$ be arbitrary and $\boldsymbol{n}_\alpha$ and $\boldsymbol{n}_\beta$ the solutions of* (2.3.1) *for $\gamma = \alpha$ and $\gamma = \beta$ respectively. If there holds $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$ for all $\alpha > \beta$, then the residual $\|\boldsymbol{K}\boldsymbol{n}_\gamma-\boldsymbol{r}\|_2$ is strictly increasing for growing γ.*

Proof. From the first-order necessary Karush-Kuhn-Tucker conditions for the problem (2.3.1) we have that for each γ there exists a vector $\boldsymbol{q}_\gamma \in \mathbb{R}^k$ with

$$\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{n}_\gamma - \boldsymbol{K}^T\boldsymbol{r} + \gamma\boldsymbol{n}_\gamma + \boldsymbol{C}^T\boldsymbol{q}_\gamma = 0 \tag{2.3.3}$$

$$\boldsymbol{C}\boldsymbol{n}_\gamma \leq \boldsymbol{b} \tag{2.3.4}$$

$$\operatorname{diag}\left(\boldsymbol{q}_\gamma\right)(\boldsymbol{C}\boldsymbol{n}_\gamma - \boldsymbol{b}) = 0 \tag{2.3.5}$$

$$\boldsymbol{q}_\gamma \geq 0. \tag{2.3.6}$$

We define the difference vector

$$\boldsymbol{x} := \boldsymbol{n}_\beta - \boldsymbol{n}_\alpha$$

and subtract (2.3.3) for $\gamma = \alpha$ with the same equation for $\gamma = \beta$ to get

$$\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{x} + \beta\boldsymbol{n}_\beta - \alpha\boldsymbol{n}_\alpha + \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right) = 0. \tag{2.3.7}$$

Taking the scalar product of (2.3.7) with $\boldsymbol{n}_\alpha$ and then with $\boldsymbol{n}_\beta$ gives

$$\begin{aligned} &\left\langle \boldsymbol{n}_\alpha, \boldsymbol{K}^T\boldsymbol{K}\boldsymbol{x} + \beta\boldsymbol{n}_\beta - \alpha\boldsymbol{n}_\alpha + \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle = 0 \\ \text{and} \quad &\left\langle \boldsymbol{n}_\beta, \boldsymbol{K}^T\boldsymbol{K}\boldsymbol{x} + \beta\boldsymbol{n}_\beta - \alpha\boldsymbol{n}_\alpha + \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle = 0. \end{aligned}$$

Our next step is to add $(\alpha - \beta)\left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle$ on both sides of the first relation and analogously
$(\alpha - \beta)\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle$ on both sides of the latter relation, which results in

$$\begin{aligned} &\left\langle \boldsymbol{n}_\alpha, \left(\boldsymbol{K}^T\boldsymbol{K} + \beta\boldsymbol{I}\right)\boldsymbol{x} + \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle = (\alpha - \beta)\left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle \\ \text{and} \qquad &\left\langle \boldsymbol{n}_\beta, \left(\boldsymbol{K}^T\boldsymbol{K} + \alpha\boldsymbol{I}\right)\boldsymbol{x} + \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle = (\alpha - \beta)\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle. \end{aligned}$$

Taking the difference of these two equations gives

$$\begin{aligned} &(\alpha - \beta)\big(\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle\big) \\ = \ &\left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \alpha\boldsymbol{I}\right)\boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \beta\boldsymbol{I}\right)\boldsymbol{n}_\alpha\right\rangle + \left\langle \boldsymbol{x}, \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle. \end{aligned}$$

On the one hand this implies

$$\begin{aligned} &(\alpha - \beta)\big(\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle\big) \\ = \ &\left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \beta\boldsymbol{I}\right)\boldsymbol{n}_\beta\right\rangle + (\alpha - \beta)\left\langle \boldsymbol{x}, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \beta\boldsymbol{I}\right)\boldsymbol{n}_\alpha\right\rangle + \left\langle \boldsymbol{x}, \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle \\ = \ &\left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \beta\boldsymbol{I}\right)\boldsymbol{x}\right\rangle + (\alpha - \beta)\left\langle \boldsymbol{x}, \boldsymbol{n}_\beta\right\rangle + \left\langle \boldsymbol{x}, \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle, \end{aligned}$$

while on the other hand

$$\begin{aligned} &(\alpha - \beta)\big(\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle\big) \\ = \ &\left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \alpha\boldsymbol{I}\right)\boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \alpha\boldsymbol{I}\right)\boldsymbol{n}_\alpha\right\rangle - (\beta - \alpha)\left\langle \boldsymbol{x}, \boldsymbol{n}_\alpha\right\rangle + \left\langle \boldsymbol{x}, \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle \\ = \ &\left\langle \boldsymbol{x}, \left(\boldsymbol{K}^T\boldsymbol{K} + \alpha\boldsymbol{I}\right)\boldsymbol{x}\right\rangle + (\alpha - \beta)\left\langle \boldsymbol{x}, \boldsymbol{n}_\alpha\right\rangle + \left\langle \boldsymbol{x}, \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle \end{aligned}$$

holds. Adding these gives

$$\begin{aligned} &2(\alpha - \beta)\big(\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle\big) \\ = \ &\left\langle \boldsymbol{x}, \left(2\boldsymbol{K}^T\boldsymbol{K} + (\alpha + \beta)\boldsymbol{I}\right)\boldsymbol{x}\right\rangle + (\alpha - \beta)\big(\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle\big) + 2\left\langle \boldsymbol{x}, \boldsymbol{C}^T\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)\right\rangle, \end{aligned}$$

and finally we arrive at

$$\begin{aligned} &(\alpha - \beta)\big(\left\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\right\rangle - \left\langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\right\rangle\big) \\ = \ &\left\langle \boldsymbol{n}_\beta - \boldsymbol{n}_\alpha, \left(2\boldsymbol{K}^T\boldsymbol{K} + (\alpha + \beta)\boldsymbol{I}\right)\left(\boldsymbol{n}_\beta - \boldsymbol{n}_\alpha\right)\right\rangle + 2\left\langle \boldsymbol{C}\left(\boldsymbol{n}_\beta - \boldsymbol{n}_\alpha\right), \boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right\rangle. \end{aligned} \tag{2.3.8}$$

Now we consider the term $\left\langle \boldsymbol{C}\left(\boldsymbol{n}_\beta - \boldsymbol{n}_\alpha\right), \boldsymbol{q}_\beta - \boldsymbol{q}_\alpha \right\rangle$. The following four cases can occur:

i-th constraint for the Tikhonov functional for		$(\boldsymbol{C}\boldsymbol{n}_\beta)_i$	$(\boldsymbol{C}\boldsymbol{n}_\alpha)_i$	$(\boldsymbol{C}\boldsymbol{n}_\beta)_i - (\boldsymbol{C}\boldsymbol{n}_\alpha)_i$	$(\boldsymbol{q}_\beta)_i$	$(\boldsymbol{q}_\alpha)_i$	$(\boldsymbol{q}_\beta)_i - (\boldsymbol{q}_\alpha)_i$
$\gamma = \beta$	$\gamma = \alpha$						
active	active	$= (\boldsymbol{b})_i$	$= (\boldsymbol{b})_i$	$= 0$	≥ 0	≥ 0	void
inactive	active	$< (\boldsymbol{b})_i$	$= (\boldsymbol{b})_i$	< 0	$= 0$	≥ 0	≤ 0
active	inactive	$= (\boldsymbol{b})_i$	$< (\boldsymbol{b})_i$	> 0	≥ 0	$= 0$	≥ 0
inactive	inactive	$< (\boldsymbol{b})_i$	$< (\boldsymbol{b})_i$	void	$= 0$	$= 0$	$= 0$

From this we see that all components of the vector

$$\operatorname{diag}\bigl(\boldsymbol{C}\left(\boldsymbol{n}_\beta - \boldsymbol{n}_\alpha\right)\bigr)\left(\boldsymbol{q}_\beta - \boldsymbol{q}_\alpha\right)$$

are nonnegative, and so $\left\langle \boldsymbol{C}\left(\boldsymbol{n}_\beta - \boldsymbol{n}_\alpha\right), \boldsymbol{q}_\beta - \boldsymbol{q}_\alpha \right\rangle \geq 0$. Under the assumption $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$ we have $\boldsymbol{x} \neq 0$, and since the matrix $2\boldsymbol{K}^T\boldsymbol{K} + (\alpha+\beta)\boldsymbol{I}$ is positive definite we finally conclude with (2.3.8) that

$$(\alpha - \beta)\bigl(\langle \boldsymbol{n}_\beta, \boldsymbol{n}_\beta\rangle - \langle \boldsymbol{n}_\alpha, \boldsymbol{n}_\alpha\rangle\bigr) > 0,$$

which is equivalent to $\|\boldsymbol{n}_\beta\|_2^2 > \|\boldsymbol{n}_\alpha\|_2^2$.

We proceed then with

$$\|\boldsymbol{K}\boldsymbol{n}_\alpha - \boldsymbol{r}\|_2^2 - \|\boldsymbol{K}\boldsymbol{n}_\beta - \boldsymbol{r}\|_2^2 = \left\langle \boldsymbol{x}, \boldsymbol{K}^T\boldsymbol{K}\boldsymbol{x}\right\rangle + 2\left\langle -\boldsymbol{x}, \boldsymbol{K}^T\boldsymbol{K}\boldsymbol{n}_\beta - \boldsymbol{K}^T\boldsymbol{r} + \beta\boldsymbol{n}_\beta\right\rangle + 2\beta\left\langle \boldsymbol{x}, \boldsymbol{n}_\beta\right\rangle \tag{2.3.9}$$

by using $\boldsymbol{n}_\alpha = \boldsymbol{n}_\beta - \boldsymbol{x}$. The variational inequality for the Tikhonov functional for $\gamma = \beta$ yields $\left\langle -\boldsymbol{x}, \boldsymbol{K}^T\boldsymbol{K}\boldsymbol{n}_\beta - \boldsymbol{K}^T\boldsymbol{r} + \beta\boldsymbol{n}_\beta\right\rangle \geq 0$. Moreover we have

$$\langle \boldsymbol{x}, \boldsymbol{n}_\beta\rangle \;=\; \langle \boldsymbol{n}_\beta - \boldsymbol{n}_\alpha, \boldsymbol{n}_\beta\rangle \;\geq\; \|\boldsymbol{n}_\beta\|_2^2 - \|\boldsymbol{n}_\alpha\|_2\|\boldsymbol{n}_\beta\|_2 \;>\; 0.$$

In summary we have shown $\|\boldsymbol{K}\boldsymbol{n}_\alpha - \boldsymbol{r}\|_2^2 > \|\boldsymbol{K}\boldsymbol{n}_\beta - \boldsymbol{r}\|_2^2$. □

Remark 2.3.2. From (2.3.9) follows that all $\boldsymbol{n}_\gamma$ with $\|\boldsymbol{K}\boldsymbol{n}_\gamma - \boldsymbol{r}\|_2 = \tau$ for an arbitrary but fixed τ must coincide. This means in other words that if the residual of the regularized solutions “gets stuck” at some value τ, the solutions $\boldsymbol{n}_\gamma$ are constant for these values of γ. In the next section we derive conditions which prevent this case.

2.3.2 Sufficient Conditions for Strict Monotonicity

In this section we derive sufficient conditions for $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$ for $\alpha > \beta$, hence by Lemma 2.3.1 for strict monotonicity. In particular we focus on constraints of the form $\boldsymbol{C}\boldsymbol{n} \geq 0$ with $\boldsymbol{C} \in \mathbb{R}^{k \times N}$ and $k \leq N$, i.e. on generalized nonnegativity constraints. For this specific type of constraints we have that for the minimum norm solution $\boldsymbol{n}_\infty$ defined in (2.3.2) that $\boldsymbol{n}_\infty \equiv 0$ holds, which gives according to [13] the relation $\|\boldsymbol{K}\boldsymbol{n}_\alpha - \boldsymbol{r}\|_2 \leq \|\boldsymbol{r}\|_2$ for all $\alpha \geq 0$.

Theorem 2.3.3. *Let $\boldsymbol{n}_\alpha$ be given by*

$$\boldsymbol{n}_\alpha := \underset{\boldsymbol{n} \in \mathbb{R}^N}{\operatorname{argmin}} \ \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{n} - \boldsymbol{r}\|_2^2 + \tfrac{1}{2}\alpha\|\boldsymbol{n}\|_2^2 \quad s.t. \quad -\boldsymbol{C}\boldsymbol{n} \leq 0, \tag{2.3.10}$$

with $\boldsymbol{C} \in \mathbb{R}^{k\times N}$ having full row rank $k \leq N$. If $\|\boldsymbol{K}\boldsymbol{n}_\alpha - \boldsymbol{r}\|_2 < \|\boldsymbol{r}\|_2$, or equivalently $\boldsymbol{n}_\alpha \neq 0$ for all $\alpha \in [0,\infty)$ according to Lemma 2.3.1 and Remark 2.3.2, then we have $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$ for all $\alpha > \beta$.

Proof. Let $\alpha > \beta$. Let $\boldsymbol{C}^\alpha_{act}$ denote the submatrix of $\boldsymbol{C}$ with active constraints in (2.3.10) for the regularization parameter α.

We first consider the case $\boldsymbol{C}^\alpha_{act} \neq \boldsymbol{C}^\beta_{act}$. We obtain $\boldsymbol{C}(\boldsymbol{n}_\alpha - \boldsymbol{n}_\beta) \neq 0$, i.e. $\boldsymbol{n}_\alpha - \boldsymbol{n}_\beta \notin \ker(\boldsymbol{C})$. This gives directly $\boldsymbol{n}_\alpha - \boldsymbol{n}_\beta \neq 0$.

Now we turn to the case $\boldsymbol{C}^\alpha_{act} = \boldsymbol{C}^\beta_{act}$. The first-order necessary conditions for a minimizer in (2.3.10) are given by

$$\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{n} - \boldsymbol{K}^T\boldsymbol{r} + \alpha\boldsymbol{n} - \boldsymbol{C}^T\boldsymbol{q}_\alpha = 0, \tag{2.3.11}$$

where $\boldsymbol{q}_\alpha \geq 0$. Let us assume $\boldsymbol{n}_\alpha = \boldsymbol{n}_\beta$. Then taking the difference of (2.3.11) for the parameters α and β yields

$$(\alpha - \beta)\boldsymbol{n}_\alpha - \boldsymbol{C}^T(\boldsymbol{q}_\alpha - \boldsymbol{q}_\beta) = 0. \tag{2.3.12}$$

Let us first consider the subcase that none of the constraints is active. Then we have $\boldsymbol{q}_\alpha = \boldsymbol{q}_\beta = 0$, which implies $(\alpha - \beta)\boldsymbol{n}_\alpha = 0$. This contradicts $\boldsymbol{n}_\alpha \neq 0$, so we must have $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$. Now we turn to the subcase that at least one constraint is active. Let $\boldsymbol{q}^\alpha_{act}$ and $\boldsymbol{q}^\beta_{act}$ be the subvectors of $\boldsymbol{q}_\alpha$ and $\boldsymbol{q}_\beta$ corresponding to active constraints. Then we can rewrite the last equation as

$$(\alpha - \beta)\boldsymbol{n}_\alpha - {\boldsymbol{C}^\alpha_{act}}^T(\boldsymbol{q}^\alpha_{act} - \boldsymbol{q}^\beta_{act}) = 0,$$

where we remember that $\boldsymbol{C}^\alpha_{act}$ is obtained from $\boldsymbol{C}$ by canceling its i-th row when the constraint $-(\boldsymbol{C}\boldsymbol{n})_i \leq 0$ is inactive and thus $(\boldsymbol{q}_\alpha)_i = (\boldsymbol{q}_\beta)_i = 0$ holds. Our next step is to multiply this equation from the left with $\boldsymbol{C}^\alpha_{act}$. By construction of $\boldsymbol{C}^\alpha_{act}$ we have $\boldsymbol{C}^\alpha_{act}\boldsymbol{n}_\alpha = 0$ and therefore

$$-\boldsymbol{C}^\alpha_{act}{\boldsymbol{C}^\alpha_{act}}^T(\boldsymbol{q}^\alpha_{act} - \boldsymbol{q}^\beta_{act}) = 0.$$

Since $\boldsymbol{C}^\alpha_{act}{\boldsymbol{C}^\alpha_{act}}^T$ has full rank, this implies $\boldsymbol{q}^\alpha_{act} = \boldsymbol{q}^\beta_{act}$ and hence $\boldsymbol{q}_\alpha = \boldsymbol{q}_\beta$. Inserting this finding back into (2.3.12) gives $\boldsymbol{n}_\alpha = 0$, which contradicts our assumption $\boldsymbol{n}_\alpha \neq 0$ for all $\alpha \in [0,\infty)$. Therefore we must also have $\boldsymbol{n}_\alpha \neq \boldsymbol{n}_\beta$ in this subcase. □

2.4 The Discrepancy Principle

With the next Theorem we summarize our previous results.

Theorem 2.4.1. *Let the conditions of Theorem 2.3.3 be fulfilled and let $\boldsymbol{n}_\infty$ be the minimum norm solution defined in* (2.3.2). *Define $r_0 := \|\boldsymbol{K}\boldsymbol{n}_0 - \boldsymbol{r}\|_2$ and $r_\infty := \|\boldsymbol{K}\boldsymbol{n}_\infty - \boldsymbol{r}\|_2$. Then there exist for any τ from $[r_0, r_\infty)$ a unique γ from $[0,\infty)$ such that $\|\boldsymbol{K}\boldsymbol{n}_\gamma - \boldsymbol{r}\|_2 = \tau$. The residual grows strictly monotonically with γ.*

□

Remark 2.4.2. The discrepancy principle carries directly over to generalized Tikhonov regularization, where the prior distribution is given by

$$p_{prior}(\boldsymbol{n}) \propto \exp(-\tfrac{1}{2}\gamma \boldsymbol{n}^T \boldsymbol{R}\boldsymbol{n}) I_S(\boldsymbol{n}),$$

where $\boldsymbol{R}$ is a positive definite regularization matrix and $I_S(\boldsymbol{n})$ is the indicator function of $S = \{\boldsymbol{n} \in \mathbb{R}^N | -\boldsymbol{C}\boldsymbol{n} \leq 0\}$. Here we have to solve the quadratic programming problem

$$\min_{\boldsymbol{n}\in\mathbb{R}^N} \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{n} - \boldsymbol{r}\|_2^2 + \tfrac{1}{2}\gamma \boldsymbol{n}^T \boldsymbol{R}\boldsymbol{n} \quad \text{s.t.} \quad -\boldsymbol{C}\boldsymbol{n} \leq 0.$$

Let $\boldsymbol{R} = \boldsymbol{U}^T\boldsymbol{U}$ the Cholesky decomposition. Then the substitution $\boldsymbol{n} = \boldsymbol{U}^{-1}\boldsymbol{v}$ transforms the above quadratic programming problem into the standard form (2.3.1).

2.5 Convergence Analysis

At this point we review some classical convergence criteria for parameter-choice strategies for Tikhonov regularization under linear constraints. With convergence we mean that the regularized reconstructions approach the true solution of the noise-free linear inverse problem as the noise level goes to 0. We decompose the noisy data vector $\boldsymbol{r}$ into

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{e}_{true} + \boldsymbol{\delta}) \\ \text{with} \quad \boldsymbol{\delta} &= (\delta_1, ..., \delta_{N_l})^T, \quad \delta_i \sim \mathcal{N}(0, \sigma_i^2). \end{aligned}$$

We carry out our convergence analysis under following assumption.

Assumption 2.5.1. *The covariance matrix* $\boldsymbol{\Sigma_\sigma}$ *has the simple form*

$$\boldsymbol{\Sigma_\sigma} = \delta^2 \cdot \operatorname{diag}(\sigma_1^2, ..., \sigma_{N_l}^2) =: \delta^2 \cdot \boldsymbol{\Sigma},$$

where $\delta \geq 0$ *is an arbitrary but fixed noise level and* σ_1, ..., σ_{N_l} *are fixed.*

Now instead of maximizing the posterior probability (2.2.4) directly, we use the fact that

$$\exp(-\tfrac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_N\boldsymbol{n} - (\boldsymbol{e}_{true} + \boldsymbol{\delta}))\|_2^2 - \tfrac{1}{2}\gamma\delta^2\|\boldsymbol{n}\|_2^2) I_S(\boldsymbol{n})$$

has the same maximizer. To obtain the function above we scaled the argument of the exponential in (2.2.4) with the noise level δ^2. For simpler notation we redefine for all the following

$$\boldsymbol{K} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{K}_N, \quad \boldsymbol{r} := \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{e}_{true} + \boldsymbol{\delta}) \quad \text{and} \quad \alpha = \gamma\delta^2.$$

This means that we work with versions of $\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{e}_{true} + \boldsymbol{\delta})$ and $\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}\boldsymbol{K}_N$ where the noise magnitude δ^2 is scaled out. So instead of solving (2.3.1), we now solve

$$\min_{\boldsymbol{n}\in\mathbb{R}^N} \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{n} - \boldsymbol{r}\|_2^2 + \tfrac{1}{2}\alpha\|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b}. \tag{2.5.1}$$

We have to point out that the back-scaled parameter $\gamma = \alpha/\delta^2$ must be used for the statistical computations for the posterior probabilities in Section 2.6. So we always compute the parameter α first from (2.5.1) and then obtain γ from it. We can already see here that Bayesian model-selection computations are not feasible for very small noise levels δ, since the parameter γ diverges as δ tends to 0. Another reason for skipping the model selection step for δ approaching 0 is that the entries of the covariance matrix $\boldsymbol{\Sigma_\sigma}$ get closer to 0 as well here, which causes problems in the statistical computations which will follow in Section 2.6. We recommend to switch to the classical discrepany principle in this case.

Now we present the standard convergence rate for Tikhonov regularization.

Proposition 2.5.2. *If the noise-free true solution $\boldsymbol{n}_0$ is an element of the feasible set of* (2.3.1)*, then the regularized solutions $\boldsymbol{n}_\alpha$ of the noise-free problem satisfies*

$$\|\boldsymbol{K}(\boldsymbol{n}_0 - \boldsymbol{n}_\alpha)\|_2 = \mathcal{O}(\alpha^{\frac{1}{2}}) \tag{2.5.2}$$

as α goes to 0*. Thus* $\lim_{\alpha\to 0} \boldsymbol{n}_\alpha = \boldsymbol{n}_0$.

Proof. Rearranging (2.3.8) for $\beta = 0$ gives the error representation

$$\|\boldsymbol{K}(\boldsymbol{n}_0 - \boldsymbol{n}_\alpha)\|_2^2 = \langle \boldsymbol{n}_0 - \boldsymbol{n}_\alpha, \alpha\boldsymbol{n}_\alpha - \boldsymbol{C}^T(\boldsymbol{q}_0 - \boldsymbol{q}_\alpha)\rangle.$$

Now since $\langle \boldsymbol{n}_0 - \boldsymbol{n}_\alpha, \boldsymbol{C}^T(\boldsymbol{q}_0 - \boldsymbol{q}_\alpha)\rangle \geq 0$ and $\|\boldsymbol{n}_\alpha\|_2 \leq \|\boldsymbol{n}_0\|_2$ hold, we can therefore estimate

$$\|\boldsymbol{K}(\boldsymbol{n}_0 - \boldsymbol{n}_\alpha)\|_2^2 \leq \alpha\|\boldsymbol{n}_0\|_2^2$$

which gives the first result. The second assertion was proved in [13]. □

Proposition 2.5.3. *Let $\boldsymbol{r}$ and $\tilde{\boldsymbol{r}}$ be two different data vectors for* (2.5.1) *and let $\boldsymbol{n}_\alpha$ and $\tilde{\boldsymbol{n}}_\alpha$ be the corresponding regularized solutions of* (2.5.1) *for the parameter α. Then*

$$\|\boldsymbol{K}(\boldsymbol{n}_\alpha - \tilde{\boldsymbol{n}}_\alpha)\|_2 \leq \|\boldsymbol{r} - \tilde{\boldsymbol{r}}\|_2 \quad \text{and} \quad \|\boldsymbol{n}_\alpha - \tilde{\boldsymbol{n}}_\alpha\|_2 \leq \frac{\|\boldsymbol{r} - \tilde{\boldsymbol{r}}\|_2}{\alpha^{\frac{1}{2}}}. \tag{2.5.3}$$

Proof. We give the proof from [13]. The solutions $\boldsymbol{n}_\alpha$ and $\tilde{\boldsymbol{n}}_\alpha$ fulfill the variational inequalities

$$\begin{aligned} &\langle \boldsymbol{K}^T\boldsymbol{K}\boldsymbol{n}_\alpha - \boldsymbol{K}^T\boldsymbol{r} + \alpha\boldsymbol{n}_\alpha, \tilde{\boldsymbol{n}}_\alpha - \boldsymbol{n}_\alpha\rangle \geq 0 \\ \text{and} \quad &\langle \boldsymbol{K}^T\boldsymbol{K}\tilde{\boldsymbol{n}}_\alpha - \boldsymbol{K}^T\tilde{\boldsymbol{r}} + \alpha\tilde{\boldsymbol{n}}_\alpha, \boldsymbol{n}_\alpha - \tilde{\boldsymbol{n}}_\alpha\rangle \geq 0. \end{aligned}$$

Adding them gives

$$\begin{aligned} \|\boldsymbol{K}(\tilde{\boldsymbol{n}}_\alpha - \boldsymbol{n}_\alpha)\|_2^2 + \alpha\|\tilde{\boldsymbol{n}}_\alpha - \boldsymbol{n}_\alpha\|_2^2 &\leq \langle \tilde{\boldsymbol{r}} - \boldsymbol{r}, \boldsymbol{K}(\tilde{\boldsymbol{n}}_\alpha - \boldsymbol{n}_\alpha)\rangle \\ &\leq \|\tilde{\boldsymbol{r}} - \boldsymbol{r}\|_2\|\boldsymbol{K}(\tilde{\boldsymbol{n}}_\alpha - \boldsymbol{n}_\alpha)\|_2, \end{aligned}$$

and the desired results follow from the last inequality. □

Finally we show under which conditions the regularized solutions $\boldsymbol{n}_\alpha^\delta$ of the noisy problem (2.5.1) converge to the true solution $\boldsymbol{n}_0$ of the noise-free problem for $\delta \to 0$. In preparation we note that for the weighted residual with noise level $\boldsymbol{\delta}$

$$\begin{aligned} \|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2 \sim \chi^2(N_l) \quad \text{thus} \quad &\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2\big) = \delta^2 \cdot \mathbb{E}\big(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2\big) = N_l\delta^2 = \mathcal{O}(\delta^2), \\ \text{i.e.} \quad &\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2\big) \leq \Big(\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2\big)\Big)^{\frac{1}{2}} = \mathcal{O}(\delta). \end{aligned}$$

We set $\boldsymbol{r}_{true} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_{true}$. Then for the expected value we have

$$\mathbb{E}(\|\boldsymbol{r} - \boldsymbol{r}_{true}\|_2) = \mathbb{E}(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2) = \mathcal{O}(\delta).$$

Theorem 2.5.4. *If we have* $\mathbb{E}(\|\boldsymbol{r}-\boldsymbol{r}_{true}\|_2)=\mathcal{O}(\delta)$ *and* $\alpha(\delta)$ *has the properties* $\lim_{\delta\to 0}\ \alpha(\delta)=0$ *and* $\lim_{\delta\to 0}\ \frac{\delta^2}{\alpha(\delta)}=0$, *then* $\lim_{\delta\to 0}\ \mathbb{E}\big(\|\boldsymbol{n}^{\delta}_{\alpha(\delta)}-\boldsymbol{n}_0\|_2\big)=0$ *holds.*

Proof. We have

$$\mathbb{E}\big(\|\boldsymbol{n}^{\delta}_{\alpha(\delta)}-\boldsymbol{n}_0\|_2\big)\leq\mathbb{E}\big(\|\boldsymbol{n}^{\delta}_{\alpha(\delta)}-\boldsymbol{n}_{\alpha(\delta)}\|_2\big)+\mathbb{E}(\|\boldsymbol{n}_{\alpha(\delta)}-\boldsymbol{n}_0\|_2),$$

where $\boldsymbol{n}_{\alpha(\delta)}$ is the regularized solution for the noise-free data $\boldsymbol{r}_{true}$. Having $\mathbb{E}(\|\boldsymbol{r}-\boldsymbol{r}_{true}\|_2)=\mathcal{O}(\delta)$, we can further estimate using Proposition 2.5.3

$$\mathbb{E}\big(\|\boldsymbol{n}^{\delta}_{\alpha(\delta)}-\boldsymbol{n}_0\|_2\big)\leq\mathbb{E}(\|\boldsymbol{n}_{\alpha(\delta)}-\boldsymbol{n}_0\|_2)+\frac{\mathcal{O}(\delta)}{\alpha^{\frac{1}{2}}}.$$

Then the result follows with Proposition 2.5.2. □

In the unconstrained case, as pointed out in [13], the standard convergence rate from Proposition 2.5.2 can be improved to the rate $o(\alpha^{\mu})$ under the assumption of a so-called *source condition*

$$\boldsymbol{n}_0=(\boldsymbol{K}^*\boldsymbol{K})^{\mu}\boldsymbol{v},$$

for some vector $\boldsymbol{v}$ and $\mu<\frac{1}{2}$. This result was generalized to the constrained case in [14]. More general source conditions of the form

$$\boldsymbol{n}_0=g(\boldsymbol{K}^*\boldsymbol{K})\boldsymbol{v},$$

where $g:\mathbb{R}\to\mathbb{R}$ is a monotonic function, are studied in [15].

The parameter μ can be regarded as a measure of smoothness of the vector $\boldsymbol{n}_0$. In practical applications however, it is not known. In [16] a parameter choice strategy is derived, which does not need the exact knowledge of a source condition, but it is assumed here that the residual obeys certain decay rates as the noise level δ approaches zero.

Another important issue is the discretization of the operator equation (2.2.1). The quality of the inversion results relies strongly on a proper choice of the model space dimension N, for instance a too coarse discretization is inappropriate for the reconstruction of an oscillatory and thus rather non-smooth function. This problem treated in [17], where an adaptive parameter choice strategy based on estimates of the approximation error $\|K^*K-\boldsymbol{K}^*_N\boldsymbol{K}_N\|$ is introduced. Based on an idea from [18], the regularization parameter α is selected from a geometric sequence $\alpha_0 q^i$, $i=1,...,M$ here with $\alpha_0=\delta^2$, $q>1$ and $q^{M-1}\alpha_0\leq 1<q^M\alpha_0$. Similar adaptive methods were derived in [19] and [20]. These methods have still the drawback, that the noise level δ must be known quite well. Therefore we derive in the next section an adaptive method both for the model space and regularization parameter selection, which is based on statistical considerations.

2.6 The Retrieval Method

Suppose we have discretized our linear operator with a Galerkin collocation method on a set of m different grids. Each grid has N_k collocation points with $N_1<...<N_m$ and we have computed a discrete approximation $\boldsymbol{K}_k$ to K for each grid. The approximation $\boldsymbol{n}_k$ of the sought-after function n lies in $\mathbb{R}^{N_k}$. For each grid we apply a Tikhonov prior with nonnegativity constraints on the observed model uncertainty

such that we have according to the previously derived results a bijection between attainable residuals and regularization parameters.

Because $\delta_1, ..., \delta_{N_l}$ are normally distributed, it follows with $\mathbf{\Sigma_\sigma}^{-\frac{1}{2}} \boldsymbol{K}_k \boldsymbol{n}_k = \boldsymbol{e}_{true}$ that

$$\|\mathbf{\Sigma_\sigma}^{-\frac{1}{2}} (\boldsymbol{K}_k \boldsymbol{n}_k - (\boldsymbol{e}_{true} + \boldsymbol{\delta})) \|_2^2 \sim \chi^2(N_l),$$

$$\text{and thus} \quad \mathbb{E}\big(\|\mathbf{\Sigma_\sigma}^{-\frac{1}{2}} (\boldsymbol{K}_k \boldsymbol{n}_k - (\boldsymbol{e}_{true} + \boldsymbol{\delta})) \|_2^2\big) = N_l, \quad \forall k \in \{1, ..., m\}.$$

2.6.1 Model Generation

In the literature on the discrepancy principle, e.g. in [21], the error estimate N_l is multiplied with a factor τ near 1 which is known as *Morozov's safety parameter*. Now we interpret it here statistically as high-probability values of the observed distribution of the weighted residual. Of course we do not select just one single value for τ, instead we select a grid of Morozov safety parameters $\tau_1, ..., \tau_s$. The following example of the $\chi^2(48)$ probability density functions illustrates this strategy:

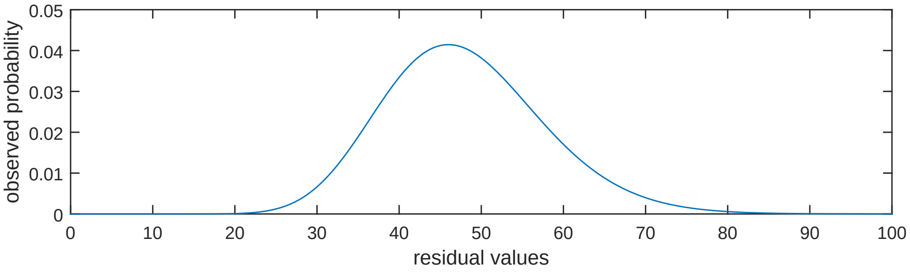


Figure 2.2: $\chi^2(48)$ probability density function

It is indeed a unimodal distribution with residual values having a nonnegliglible probability ranging from 30 to 70. For $N_l = 48$ this corresponds to values of τ ranging from ca. 0.6 to ca. 1.5. Therefore proposing just a single residual value for the discrepancy principle ($1.1N_l$ would be a common choice) excludes many probable reconstructions corresponding to other residual values, such that the posterior probability exploration is limited. Moreover the danger of under- or overregularization would be high.

As in the previous section we use the normalized version $\mathbf{\Sigma}$ of the covariance matrix $\mathbf{\Sigma_\sigma}$. This means that we try to fit the normalized residuals

$$\|\mathbf{\Sigma}^{-\frac{1}{2}} (\boldsymbol{K}_k \boldsymbol{n}_k - (\boldsymbol{e}_{true} + \boldsymbol{\delta})) \|_2^2$$

to the values $\tau N_l \delta^2$ where the values for τ run through the grid of preselected Morozov safety factors. In practice the noise magnitude δ^2 is taken as the biggest measurement sample mean and $\mathbf{\Sigma}$ is estimated from $\mathbf{\Sigma_\sigma}$ by normalizing it with the estimate for δ^2.

For the following we set $\boldsymbol{e}_{real} := (\tilde{e}_1, ..., \tilde{e}_{N_l})^T$, hence this is the vector of the realizations of the random variables $e_1, ..., e_{N_l}$. With these preparations the model generation step proceeds as follows:

Algorithm 1 Model Generation

```
$MaxDisc = 3$
$SolutionSets = \{\}$
$ApproxSets = \{\}$
$PriorSets = \{\}$
$TauSets = \{\}$
$DiscCntr = 0$
estimate $\sigma_1^2$, ..., $\sigma_{N_l}^2$ from the sample means approximating the standard deviations of $e_1$, ..., $e_{N_l}$.
$\delta^2 := \max\{\sigma_1^2, ..., \sigma_{N_l}^2\}$
$\boldsymbol{\Sigma} := \delta^{-2} \cdot \operatorname{diag}(\sigma_1^2, ..., \sigma_{N_l}^2)$
for $i = 1$ to $m$ do
    $S_i = \{\}$
    $A_i = \{\}$
    $P_i = \{\}$
    $T_i = \{\}$
    $\boldsymbol{n}_{lsqnng} = \underset{\boldsymbol{n} \in \mathbb{R}^{N_i}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n} - \boldsymbol{e}_{real})\|_2^2$ s.t. $\boldsymbol{n} \geq 0$
    $R_{lsqnng} = \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n}_{lsqnng} - \boldsymbol{e}_{real})\|_2^2$
    for $j = 1$ to $s$ do
        if $R_{lsqnng} < \tau_j N_l \delta^2 \ \wedge \ \tau_j N_l \delta^2 < \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_{real}\|_2^2$ then
            compute $\gamma_{ij}$ such that
            $\boldsymbol{n}_{trial} = \underset{\boldsymbol{n} \in \mathbb{R}^{N_i}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n} - \boldsymbol{e}_{real})\|_2^2 + \frac{1}{2}\gamma_{ij}\boldsymbol{n}^T\boldsymbol{R}_i\boldsymbol{n}$ s.t. $\boldsymbol{n} \geq 0$
            with $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n}_{trial} - \boldsymbol{e}_{real})\|_2^2 = \tau_j N_l \delta^2$
        end if
        if $\boldsymbol{n}_{trial}$ exists then
            $S_i = S_i \cup \{\boldsymbol{n}_{trial}\}$
            $A_i = A_i \cup \{\boldsymbol{K}_i\}$
            $P_i = P_i \cup \{\gamma_{ij}\boldsymbol{R}_i\}$
            $T_i = T_i \cup \{\tau_j\}$
        end if
    end for
    if $S_i$, $A_i$, $P_i$ and $T_i$ not empty then
        $SolutionSets = SolutionSets \cup \{S_i\}$
        $ApproxSets = ApproxSets \cup \{A_i\}$
        $PriorSets = PriorSets \cup \{P_i\}$
        $TauSets = TauSets \cup \{T_i\}$
        $DiscCntr = DiscCntr + 1$
    end if
    if $DiscCntr == MaxDisc$ then
        break
    end if
end for
```

The outer loop runs through the discretization levels beginning with the coarsest one. This approach is in accordance with the principle of *Occam's razor*, where among all possible explanations of a problem simpler ones are preferred over more complicated ones. For a detailed explanation why Bayesian model selection implements Occam's razor, cf. [22]. Another motivation is *regularization by discretization*, which means that the approximate problems for the operator inversion are for coarser discretizations less ill-conditioned than for finer discretizations. But by using the discrepancy principle we ensure that the models selected are not too coarse by demanding that the model has to fit the data, which means that the residuals may not be too big. Convergence of the finite dimensional regularized approximations to the solutions of a linear operator equation to its solution under quite general assumptions was shown in [23].

For each i-th discretization level in the outer loop, the inner loop runs through the preselected grid of Morozov safety factors, where for each factor τ_j the computation of a regularized solution $\boldsymbol{n}_{trial}$ with residual $\tau_j N_l$ is attempted. In line 18 it is checked if the discrepancy principle is applicable. If it is possible to compute $\boldsymbol{n}_{trial}$, this reconstruction is stored in the container S_i and the approximation $\boldsymbol{K}_i$ to K in A_i. The prior information given by the regularization parameter γ_{ij} and the regularization matrix $\frac{1}{2}\boldsymbol{R}_i$ are stored in P_i and the residual parameter τ_j in T_i. These matrices will be used to compute the Bayesian posterior probabilities for the model selection in the next section.

If in the current discretization level the containers with reconstructions, operator approximation matrices, prior informations and residual parameters are not empty, they are be added to the containers *SolutionSets*, *ApproxSets*, *PriorSets* and *TauSets* respectively. Note that we have limited the maximal number of admissible discretization levels to three. On the one hand this is done to save computational effort, but on the other hand it turns out that the posterior probabilities get too similar and thus not clearly or reliably distinguishable when using too many finely discretized models.

2.6.2 Model Selection

In this section we apply the Bayesian model selection framework as introduced in [24]. Since we assume that the data is given by independent Gaussian random variables, the observed model uncertainty is a multivariate Gaussian distribution. For any of the approximations $\boldsymbol{K}_k$ to the operator K with $k \in \{1, ..., m\}$ it is given by

$$p(\boldsymbol{e}|\boldsymbol{n}, N_k, \boldsymbol{K}_k) = (2\pi)^{-\frac{N_l}{2}} \left|\det(\boldsymbol{\Sigma_\sigma})\right|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\boldsymbol{K}_k\boldsymbol{n}-\boldsymbol{e})^T\boldsymbol{\Sigma}_{\boldsymbol{\sigma}}^{-1}(\boldsymbol{K}_k\boldsymbol{n}-\boldsymbol{e})). \quad (2.6.1)$$

Here the vector $\boldsymbol{n} \in \mathbb{R}^{N_k}$ represents all possible reconstructions for the current discretization.

We know beforehand that our reconstruction must be nonnegative and that it is smooth. We put this prior knowledge into our reconstruction method by setting up the Bayesian conditional prior probability which is determined by

$$p(\boldsymbol{n}|N_k, \boldsymbol{K}_k, \boldsymbol{R}_k, \gamma_{kj}) = C_{kj}^{-1} \exp(-\tfrac{1}{2}\gamma_{kj}\boldsymbol{n}^T\boldsymbol{R}_k\boldsymbol{n})I_{\geq 0}(\boldsymbol{n}), \quad (2.6.2)$$

where $I_{\geq 0}(\boldsymbol{n})$ is the indicator function of the first quadrant of $\mathbb{R}^{N_k}$, $\boldsymbol{R}_k$ is the regularization matrix and γ_{kj} is the regularization parameter. All these quantities were computed and stored in the model generation procedure in the previous section.

If $\boldsymbol{R}_k$ is regular and positive definite, the normalizing constant

$$C_{kj} = \int_{[0,\infty)^{N_k}} \exp(-\tfrac{1}{2}\gamma_{kj}\boldsymbol{n}^T\boldsymbol{R}_k\boldsymbol{n})d\boldsymbol{n} \tag{2.6.3}$$

is well-defined. For *Tikhonov regularization*, where $\boldsymbol{R}_k = \boldsymbol{I}_{N_k}$ holds, we have a closed form expression for it, namely

$$C_{kj} = \left(\frac{\pi}{2\gamma_{kj}}\right)^{\frac{N_k}{2}}.$$

In *minimal first differences regularization* with zero boundary conditions the regularization matrix is given by

$$\boldsymbol{R}_k = \boldsymbol{H}_k^T\boldsymbol{H}_k \quad \text{with} \quad \boldsymbol{H}_k = \begin{pmatrix} -1 & & & & \\ 1 & -1 & & & \\ & \ddots & \ddots & & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix}.$$

For *Twomey regularization* with eliminated zero boundary conditions we have

$$\boldsymbol{R}_k = \boldsymbol{H}_k^T\boldsymbol{H}_k \quad \text{with} \quad \boldsymbol{H}_k = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

Here $\boldsymbol{R}_k$ is a positive definite tridiagonal matrix. For the latter two regularization methods C_{kj} must be computed numerically.

Remember that the container *SolutionSets* stores reconstructions from at most 3 discretization levels. We let the index i run through all discretization levels in *SolutionSets* and the index j through all residual parameters captured in the i-th level. Then with Bayes' rule the posterior model probabilities are

$$p(N_k, \gamma_{kt}|\boldsymbol{e}) = \frac{p(\boldsymbol{e}|N_k, \gamma_{kt})p(N_k, \gamma_{kt})}{\sum_i \sum_j p(\boldsymbol{e}|N_i, \gamma_{ij})p(N_i, \gamma_{ij})}$$

where with (2.6.1)-(2.6.3) we have

$$\begin{aligned} &p(\boldsymbol{e}|N_i, \gamma_{ij}) \\ =&\int_{\mathbb{R}^{N_i}} p(\boldsymbol{e}, \boldsymbol{n}|N_i, \gamma_{ij})d\boldsymbol{n} \\ =&\int_{\mathbb{R}^{N_i}} p(\boldsymbol{e}|\boldsymbol{n}, N_i)p(\boldsymbol{n}|N_i, \gamma_{ij})d\boldsymbol{n} \\ =&\int_{[0,\infty)^{N_i}} B^{-1}C_{ij}^{-1}\exp(-\tfrac{1}{2}\left\|\boldsymbol{\Sigma}_{\boldsymbol{\sigma}}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n} - \boldsymbol{e})\right\|_2^2 - \tfrac{1}{2}\gamma_{ij}\boldsymbol{n}^T\boldsymbol{R}_i\boldsymbol{n})d\boldsymbol{n}, \end{aligned} \tag{2.6.4}$$

where

$$B = (2\pi)^{\frac{N_l}{2}}\left|\det(\boldsymbol{\Sigma}_{\boldsymbol{\sigma}})\right|^{\frac{1}{2}}$$

$$\text{and} \quad C_{ij} = \int_{[0,\infty)^{N_i}} \exp(-\tfrac{1}{2}\gamma_{ij}\boldsymbol{n}^T\boldsymbol{R}_i\boldsymbol{n})d\boldsymbol{n}.$$

We assumed that the model matrix $\boldsymbol{K}_i$ and the regularization matrix $\boldsymbol{R}_i$ were implicitly given by each discretization level N_i, i.e. we actually have $p(\boldsymbol{e}|N_i,\gamma_{ij}) = p(\boldsymbol{e}|N_i,\boldsymbol{K}_i,\boldsymbol{R}_i,\gamma_{ij})$. For simplicity of notation these were omitted. Note that the prior model probabilities $p(N_i,\gamma_{ij})$ are still free.

For the computation of the above integrals of multivariate Gaussian densities over the first quadrant of each model space $\mathbb{R}^{N_i}$ we applied an effective pseudo-random integration method described in [25] which implements the routines presented in [26] and [27].

We now turn to the prior model probabilities $p(N_i,\gamma_{ij})$. As mentioned in the beginning of Section 2.3 we assume that a $\gamma_{min} > 0$ and a $\gamma_{max} < \infty$ exist which give a lower and an upper bound for the regularization parameters γ in order to exclude improper or point-mass priors for the cases $\gamma = 0$ or $\gamma = \infty$. This assumption is independent of the discretization level. We further assume the discretization level to be independent and uniformly distributed. Thus we are taking a *noninformative* prior, and so the prior model probabilities cancel out and do not affect the posterior probabilities.

Now everything is prepared to perform the model selection. To compute integrals of the form

$$\int_{[0,\infty)^N} \exp(-\tfrac{1}{2}(\boldsymbol{n}^T\boldsymbol{H}\boldsymbol{n} - 2\boldsymbol{n}^T\boldsymbol{v} + q))d\boldsymbol{n},$$

where N is the dimension of the square matrix $\boldsymbol{H}$, we apply the method from [25]. It actually can only evaluate intgrals of the form

$$\frac{1}{\sqrt{\det(\boldsymbol{W})(2\pi)^N}} \int_{a_1}^{b_1} \ldots \int_{a_N}^{b_N} \exp(-\tfrac{1}{2}\boldsymbol{n}^T\boldsymbol{W}^{-1}\boldsymbol{n})d\boldsymbol{n},$$

where the cases $a_i = -\infty$ and $b_i = \infty$ are allowed. So we have to perform a simple affine transformation using the Cholesky factorization $\boldsymbol{H} = \boldsymbol{U}^T\boldsymbol{U}$:

$$\begin{aligned} &\int_{[0,\infty)^N} \exp(-\tfrac{1}{2}(\boldsymbol{n}^T\boldsymbol{H}\boldsymbol{n} - 2\boldsymbol{n}^T\boldsymbol{v} + q))d\boldsymbol{n} \\ = &\left(\exp(-\tfrac{1}{2}(q - \boldsymbol{v}^T\boldsymbol{H}^{-1}\boldsymbol{v}))\sqrt{\det(\boldsymbol{H}^{-1})(2\pi)^N}\right) \\ &\cdot \frac{1}{\sqrt{\det(\boldsymbol{H}^{-1})(2\pi)^N}} \int_{\left\{\boldsymbol{z}\in\mathbb{R}^N \mid\, \boldsymbol{z} \geq -\boldsymbol{H}^{-1}\boldsymbol{v}\right\}} \exp(-\tfrac{1}{2}\boldsymbol{z}^T\boldsymbol{H}\boldsymbol{z})d\boldsymbol{z}. \end{aligned}$$

The model selection algorithm is as follows.

Algorithm 2 Model Selection

```
1: get $S_1, ..., S_{MaxDisc}$ from $SolutionSets$
2: get $A_1, ..., A_{MaxDisc}$ from $ApproxSets$
3: get $P_1, ..., P_{MaxDisc}$ from $PriorSets$
4: get $T_1, ..., T_{MaxDisc}$ from $TauSets$
5: $m_1 = |S_1|, ..., m_{MaxDisc} = |S_{MaxDisc}|$
6: $m_{total} = \sum_{k=1}^{MaxDisc} m_k$
7: $B = (2\pi)^{\frac{N_l}{2}} \left|\det(\boldsymbol{\Sigma_\sigma})\right|^{\frac{1}{2}}$
8: $P_{post} = \{\}$
9: for $i = 1$ to $MaxDisc$ do
10:     for $j = 1$ to $m_i$ do
11:         $\boldsymbol{K}_{ij} = A_i(j)$
12:         $\boldsymbol{R}_{ij} = \frac{1}{\delta^2} P_i(j)$
13:         $C_{ij} = \int_{[0,\infty)^{N_i}} \exp(-\frac{1}{2}\boldsymbol{n}^T \boldsymbol{R}_{ij}\boldsymbol{n}) d\boldsymbol{n}$
14:         $M_{ij} = \int_{[0,\infty)^{N_i}} \exp(-\frac{1}{2}\left\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_{ij}\boldsymbol{n} - \boldsymbol{e}_{real})\right\|_2^2 - \frac{1}{2}\boldsymbol{n}^2\boldsymbol{R}_{ij}\boldsymbol{n}) d\boldsymbol{n}$
15:         $P_{post} = P_{post} \cup \left\{M_{ij}/(B \cdot C_{ij})\right\}$
16:     end for
17: end for
18: $SumP_{post} = \sum_{k=1}^{m_{total}} P_{post}(k)$
19: for $i = 1$ to $MaxDisc$ do
20:     $P_{post}(i) = P_{post}(i)/SumP_{post}$
21: end for
22: $S_{total} = S_1 \cup ... \cup S_{MaxDisc}$
23: sort $S_{total}(1), ..., S_{total}(m_{total})$ according to $P_{post}(1), ..., P_{post}(m_{total})$
```

In the first lines of the model-selection algorithm the containers for computed reconstructions, operator approximation matrices, prior matrices and residual parameters are loaded for each examined discretization level. They store the results of the model-generation algorithm from Section 2.6. In the case of too noisy or improper data it might happen that in the model generation step none of the models can fit the data. Then all containers are empty and the model selection algorithm has to be aborted. For simplicity we assume that the model-generation step was successful.

The double loop in lines 7-17 performs the multidimensional integrations needed in (2.6.3) and (2.6.4). In line 15 these integrals are used for the unnormalized posterior probabilities $p(N_i, \gamma_{ij}|\boldsymbol{e})$ from (2.6.4). Note that the prior model probabilities $p(N_i, \gamma_{ij})$ do not appear in the algorithm, since they are selected to be uniform and thus cancel out in the normalizing step performed in lines 18 - 14. At last all reconstructions are sorted according to their posterior probabilities.

We have to be careful not to forget to normalize the regularization matrices $P_i(j)$ with the estimated noise level δ^2 as in line 12 because all statistical computations have to be carried out using the unnormalized covariance matrix $\boldsymbol{\Sigma_\sigma}$. For very small noise levels we recommend to skip the model selection step completely due to instabilities in the statistical computations mentioned above. It is sufficient to use only the coarsest model generated with the commonly used value $\tau = 1.1$ in this case.

Chapter 3

Numerical Results

3.1 Simulation of Aerosol Spectroscopy Measurements

We applied our algorithm to a simplified version of problem (2.1.2), where we assumed that we know the minimal and maximal particle radii r_{min} and r_{max}. This led to the integral equation

$$\int_{r_{min}}^{r_{max}} k(r,l)n(r)dr = e(l). \tag{3.1.1}$$

For the kernel function $k(r,l)$ from Mie theory we selected H_2O as the material for the scattering particles and air for the medium.

In our simulations we assumed $r_{min} = 0.01$ μm and $r_{max} = 7.0$ μm. In practice the extinction function can only be measured for a finite number of light wavelengths $l_1, ..., l_{N_l}$. In our simulations we used the grid of 48 wavelengths composed of 8 linearly spaced wavelengths from $0.6 - 0.8$ μm, 8 from $1.1 - 1.3$ μm, 8 from $1.6 - 1.8$ μm, 16 from $2.1 - 2.5$ μm and 8 from $3.1 - 3.3$ μm. These five intervals were chosen to exclude wavelengths where light absorption by ambient water can occur which distorts the measured extinctions $e(l)$ heavily. That is, the selected wavelengths cover the so-called *optical window* which is free from this unwanted physical effect.

We generated artificial extinction values $e(l_i)$ for the selected $l_1, ..., l_{N_l}$ by solving the forward problem, which means inserting an original 'true' size distributions $n(r)$ into the integral equation (3.1.1). To avoid the *inverse crime* we used a very fine grid with 10001 points and the composite Simpson rule to compute the resulting integrals.

We performed three simulation runs consiting of 1000 single inversions for each of the three size distribution families for different noise levels. For each single inversion we generated a set of 300 noisy extinctions from the artificial true extinction values by adding zero-mean Gaussian noise where the standard deviations were taken to be 5%, 15% and 30% respectively of the true extinction values $e(l_i)$. This means that a vector $\boldsymbol{e}$ of noisy extinctions for each single simulated measurement in the first simulation run was modeled as

$$(\boldsymbol{e})_i = e(l_i) + \delta_i \quad \text{with } \delta_i \sim \mathcal{N}(0, (0.05 \cdot e(l_i))^2), \quad i = 1, ..., N_l.$$

For the other two simulation runs we computed

$$(\boldsymbol{e})_i = e(l_i) + \delta_i \quad \text{with } \delta_i \sim \mathcal{N}(0, (0.15 \cdot e(l_i))^2), \quad i = 1, ..., N_l$$

and

$$(\boldsymbol{e})_i = e(l_i) + \delta_i \quad \text{with } \delta_i \sim \mathcal{N}(0, (0.3 \cdot e(l_i))^2), \quad i = 1, ..., N_l.$$

respectively. We used the sample means and variances of these 300 artificial noisy extinctions to do inferences about the simulated Gaussian noise.

For the discretization of (3.1.1) we used a Galerkin collocation method with linear basis functions on an integration grid with $N_r = 300$ equidistant points. Alternative discretization methods include Legendre basis polynomials (see [28]) and Bernstein polynomials (see [29]), but are rather seldom used. We generated our model spaces by selecting collocation grids as near equidistant subgrids of the integration grid where the number of grid points N_{col} ranged from 3 (coarsest discretization level) to 50 (finest discretization level). For the collocation grids we set up linearly spaced 'pre-collocation grids' with N_{col} points first and then performed a nearest-neighbor-fitting of their points to the integration grid, such that they became subgrids. Since we are considering size distributions which attain small values at the minimal and maximal radii, we assumed zero boundary conditions. This effectively reduced the number of unknowns N in each model space from $N = 3, ..., 50$ to $N = 1, ..., 48$ and—more importantly—prevented the reconstructed size distributions from sheering out at the smallest radius value, which would have been a not reasonable behavior, physically speaking. It was important that the dimension N of each model space never succeeded the number of measurements $N_l = 48$, such that the resulting regression problems were fully or overdetermined.

Let $r_1, ..., r_{N_r}$ denote the integration grid points. Let $\{r_1 = c_1 < ... < c_{N_{col}} = r_{N_r}\} \subset \{r_1, ..., r_{N_r}\}$ be a collocation grid. The *triangular basis funktions* $b_k(r)$, $k = 1, ..., N_{col}$ are the piecewise linear functions on the intervals $[c_1, c_2]$, ..., $[c_{N_{col}-1}, c_{N_{col}}]$ which fulfill

$$b_k(c_j) = \delta_{kj}, \text{ for } j = 1, ..., N_{col}.$$

We approximated the sought-after function $n(r)$ with the linear combination

$$n(r) = \sum_{k=1}^{N_{col}} n_k b_k(r), \tag{3.1.2}$$

where the weights $n_2, ..., n_{N_{col}-1} \in \mathbb{R}$ are free variables and $n_1 = n_{N_{col}} = 0$ holds because of the zero boundary conditions.

Inserting (3.1.2) into (3.1.1) yields the linear system of equations for the unknown weights

$$\sum_{k=1}^{N_{col}} n_k \int_{r_{min}}^{r_{max}} k(r, l_i) b_k(r) dr = e(l_i), \quad i = 1, ..., N_l. \tag{3.1.3}$$

We applied the composite trapezoidal rule with the integration grid $r_1, ..., r_{Nr}$ on the integrals defining the coefficients in above linear system. The resulting coefficient matrix is the matrix $\boldsymbol{K}_N$ from Section 2.6 which approximates the integral operator from the left-hand side of (3.1.1).

3.2 Numerical Study

We performed a numerical study for our reconstruction algorithm with model size distributions from the log-normal, Rosin-Rammler-Sperling-Bennett (RRSB) and

Hedrih families, where each of these size distribution families has certain free parameters. We varied the parameters in domains giving physically reasonable size distributions and generated noise in the same order of magnitude as observed in real experimental FASP measurements. Therefore the numerical results should give good estimates of the quality of the reconstructions compared to real size distributions. In the same simulation runs we compared our algorithm with existing reconstruction methods.

3.2.1 Applied Methods

For all inversion methods applied in our numerical study we selected for the priors Tikhonov, minimal first differences and Phillips-Twomey regularization from Section 2.6.2.

In our inversion method we set the Morozov safety factor grid to

$$\tau_1 = 0.6, \tau_2 = 0.7, ..., \tau_{12} = 1.7.$$

We refer to this as the *constrained method* in the following.

To see that the constraints in the constrained method are worth the computational effort, we compared it with its counterpart without constraints, which we call the *unconstrained method.* It performs the same model generation step based on the discrepancy principle with the same Morozov safety factors grid, but the constraints in (2.3.1) were dropped. The computations for the model selection are much easier here, since the integrals of the multivariate Gaussian distributions over the parameter spaces can be evaluated analytically.

By reducing the grid of Morozov safety factors in the constrained method simply to the classical value $\tau = 1.1$ we obtained another method participating in our numerical study. We call it the *Morozov method.* The comparison with it shows whether the grid of Morozov safety factors is justified or not.

We also implemented a classical model-selection method for the unconstrained problem which is independent of the prior. Here we compared the three coarsest models where the discrepancy principle was applicable with the *Bayesian Information Criterion* (BIC), which was first introduced in [30]. The model with the lowest BIC-value

$$-2\Big(-\tfrac{1}{2} N_l \log(2\pi) - \tfrac{1}{2}\log(\det(\boldsymbol{\Sigma}_{\boldsymbol{\sigma}})) - \tfrac{1}{2}\|\boldsymbol{\Sigma}_{\boldsymbol{\sigma}}^{-\frac{1}{2}}(\boldsymbol{K}_N \boldsymbol{n}_{ml} - \boldsymbol{e})\|_2^2\Big) + N\log(N_l),$$

where $\boldsymbol{n}_{ml}$ is the unconstrained maximum-likelihood solution, is selected here. We call this method the *BIC method.*

3.2.2 Model Size Distributions

We generated the simulated measurement data vectors $\boldsymbol{e}_{true}$ by inserting one of the following three model size distributions adopted from [31] into our integral equation (3.1.1):

1. *log-normal distribution*

$$n(r) = \frac{A}{\sqrt{2\pi}\sigma r} \exp\left(-\frac{1}{2\sigma^2}\big(\log(r) - \log(\mu)\big)^2\right) \tag{3.2.1}$$

with amplitude A, standard deviation σ, and mean μ.

2. *Rosin-Rammler-Sperling-Bennet* (RRSB) *distribution*

$$n(r) = \frac{AN}{\nu}\Big(\frac{r}{\nu}\Big)^{N-1} \exp\bigg(- \Big(\frac{r}{\nu}\Big)^{N}\bigg) \tag{3.2.2}$$

   with amplitude A, exponent N, and mean ν.

3. *Hedrih distribution*

$$n(r) = \frac{128Ar^3}{3\eta^4} \exp\bigg(- \frac{4r}{\eta}\bigg) \tag{3.2.3}$$

   with amplitude A and mean μ.

For each simulated size distribution we set the amplitude to $A = 10^4$. We choose the remaining parameters so that the relation

$$n(r_{max}) \leq Tol \tag{3.2.4}$$

with $r_{max} = 7.0$ μm and $Tol = 10$ was satisfied. This is to be consistent with the assumption, that we can neglect the tails of the distributions and truncate them at the maximal radius r_{max}. Furthermore we assumed the modal value of the log-normal and RRSB distributions to be greater or equal to 1.0 μm in order to exclude too peaked distributions. For each of the above three model size distributions we looped in our simulations through a set of 100 possible parameters satisfying (3.2.4).

For the log-normal distributions we first selected for the mean σ a linearly spaced grid with ten points ranging from 0.2 to 0.5, i.e. $\sigma_k = 0.2 + 0.3\frac{k-1}{9}$, $k = 1, ..., 10$. Then we saw after a lengthy calculation that (3.2.4) is equivalent to

$$r_{max} \exp\Bigg(- \Bigg(- 2\sigma^2 \log\bigg(\frac{\sqrt{2\pi} r_{max}\sigma Tol}{A}\bigg)\Bigg)^{\frac{1}{2}}\Bigg) \geq \mu.$$

The modal value of the log-normal distribution is $r_{mod} = \exp\big(\log(\mu) - \sigma^2\big)$, so $r_{mod} \geq 1.0$ is equivalent to $\mu \geq 1.0 \exp\big(\sigma^2\big)$. Using the last two inequalities we selected

$$\begin{aligned} \mu_{kj} &= 1.0 \exp\big(\sigma_k^2\big) + \frac{j-1}{9}\Big(v_k - 1.0\exp\big(\sigma_k^2\big)\Big) \\ \text{with} \quad v_k &= r_{max}\exp\Bigg(- \Bigg(- 2\sigma_k^2 \log\bigg(\frac{\sqrt{2\pi} r_{max}\sigma_k Tol}{A}\bigg)\Bigg)^{\frac{1}{2}}\Bigg), \\ k &= 1, ..., 10, \ j = 1, ..., 10. \end{aligned}$$

These are the 100 parameters used for the log-normal distributions.

For the RRSB distributions we took for the exponents N the integer values $N_k = k + 2$, $k = 1, ..., 10$. We computed the auxiliary variables p_k as the real-valued solutions of the equations

$$p_k \exp(-p_k) = \frac{r_{max}Tol}{AN_k}$$

being greater than one. With some algebra one can see that

$$\nu \leq r_{max} \cdot p_k^{-\frac{1}{N_k}}$$

is then equivalent to (3.2.4) for the RRSB distribution. The modal value of the RRSB distribution is $r_{mod} = \nu \left(\frac{N-1}{N}\right)^{\frac{1}{N}}$, therefore $r_{mod} \geq 1.0$ is equivalent to $\nu \geq 1.0 \cdot \left(\frac{N-1}{N}\right)^{-\frac{1}{N}}$. Using the last two inequalities we selected

$$\nu_{kj} = 1.0 \cdot \left(\frac{N_k - 1}{N_k}\right)^{-\frac{1}{N_k}} + \frac{j-1}{9}\left(r_{max} \cdot p_k^{-\frac{1}{N_k}} - 1.0 \cdot \left(\frac{N_k - 1}{N_k}\right)^{-\frac{1}{N_k}}\right),$$
$$k = 1, ..., 10,\ j = 1, ..., 10.$$

Thus we have 100 parameters for the RRSB distributions.

For the Hedrih distribution we found that (3.2.4) is equivalent to $\eta \leq \eta_{max}$ with $\eta_{max} \approx 2.0566$. Thus we took for η the values

$$\eta_k = 0.8 + \frac{k-1}{99}\left(\eta_{max} - 0.8\right), \quad k = 1, ..., 100.$$

For each of the three size distribution classes we simulated ten artificial noisy measurement-data vetors $\boldsymbol{e}$ as described in Section 3.1 for each of the corresponding 100 parameters. This resulted in total in 1000 single simulated FASP experiments for one model size distribution class.

For every inversion we computed the L^2-error of the obtained reconstruction relative to the original size distribution and measured the total run time needed for the inversion. The computations were performed on a notebook with a 2.27 GHz CPU and 3.87 GB accessible primary memory.

3.2.3 Average L^2-Errors

Results for 5% Noise

Log-Normal Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	11.8763	11.9621	12.3108
Morozov	12.6919	13.2348	13.9335
unconstrained	16.1952	20.0564	21.9649
BIC	32.3190	34.3861	35.6112

RRSB Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	9.9690	8.9646	8.5195
Morozov	11.9921	11.7512	11.4955
unconstrained	18.1053	20.1539	22.0244
BIC	38.1494	39.1321	40.5063

Hedrih Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	12.0214	12.0852	12.1321
Morozov	13.2794	12.4251	12.6591
unconstrained	11.2122	11.0906	10.9498
BIC	16.0117	15.2573	15.2073

Results for 15% **Noise**

Log-Normal Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	17.3371	17.1004	17.6913
Morozov	25.7848	26.0661	27.3864
unconstrained	23.0265	26.7038	29.7148
BIC	33.0129	35.1238	37.1033

RRSB Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	14.3285	12.9690	12.7290
Morozov	18.1593	17.6314	17.9012
unconstrained	23.7878	26.7608	30.5864
BIC	41.7431	43.9701	47.3254

Hedrih Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	11.9472	11.7113	11.5986
Morozov	16.5099	15.0745	14.4839
unconstrained	20.8872	20.7861	20.8880
BIC	28.0989	27.4491	27.3066

Results for 30% Noise

Log-Normal Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	21.1133	21.4006	22.1933
Morozov	29.1050	29.2021	30.7157
unconstrained	31.2302	34.0586	37.5831
BIC	56.3298	58.9926	62.5693

RRSB Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	17.2089	16.4395	16.5887
Morozov	23.8178	23.1906	23.5631
unconstrained	29.1748	33.2577	37.6546
BIC	79.9881	82.1931	85.5154

Hedrih Distribution			
method	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	14.7340	14.6443	14.4238
Morozov	25.4313	23.7145	22.8068
unconstrained	36.9309	36.8277	36.9877
BIC	41.5024	40.7930	40.4949

3.2.4 Average Run Times

Results for 5% Noise

Log-Normal Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	3.2192	3.2690	3.2746
Morozov	0.6947	0.7047	0.7076
unconstrained	0.2884	0.2812	0.2814
BIC	0.0614	0.0596	0.0596

RRSB Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	5.5637	5.6545	5.6046
Morozov	1.0963	1.1128	1.1019
unconstrained	0.3445	0.3405	0.3402
BIC	0.0747	0.0754	0.0741

Hedrih Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	2.4919	2.5088	2.5231
Morozov	0.7259	0.7287	0.7317
unconstrained	0.2740	0.2680	0.2688
BIC	0.0644	0.0634	0.0614

Results for 15% Noise

Log-Normal Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	2.5520	2.5825	2.5803
Morozov	0.6174	0.6267	0.6283
unconstrained	0.2854	0.2793	0.2791
BIC	0.0579	0.0594	0.0570

RRSB Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	4.2281	4.2824	4.2545
Morozov	0.9406	0.9510	0.9428
unconstrained	0.3297	0.3229	0.3233
BIC	0.0716	0.0719	0.0696

Hedrih Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	2.7098	2.7091	2.7207
Morozov	0.6011	0.6043	0.6066
unconstrained	0.3353	0.3290	0.3291
BIC	0.0631	0.0634	0.0624

Results for 30% Noise

Log-Normal Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	1.4135	1.4187	1.4061
Morozov	0.3552	0.3607	0.3594
unconstrained	0.1831	0.1771	0.1772
BIC	0.0399	0.0399	0.0389

RRSB Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	2.3742	2.4065	2.3756
Morozov	0.5312	0.5385	0.5337
unconstrained	0.2061	0.1990	0.2003
BIC	0.0453	0.0456	0.0442

Hedrih Distribution			
method	average run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	1.8456	1.8404	1.8187
Morozov	0.3689	0.3734	0.3712
unconstrained	0.2215	0.2182	0.2183
BIC	0.0366	0.0366	0.0355

3.2.5 Average Model Space Dimensions

Results for 5% Noise

Log-Normal Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	10.0640	10.1820	10.0870
Morozov	11.7670	11.8700	11.8290
unconstrained	8.4030	8.7550	8.9030
BIC	10.3480	10.3480	10.3480

RRSB Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	13.8260	13.8380	13.8850
Morozov	15.2820	15.3260	15.3040
unconstrained	12.6050	13.2180	13.2140
BIC	11.8800	11.8800	11.8800

Hedrih Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	8.4470	8.4640	8.5260
Morozov	10.2560	10.2500	10.2800
unconstrained	6.7190	6.8320	7.0430
BIC	9.2290	9.2290	9.2290

Results for 15% Noise

Log-Normal Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	7.7360	7.7330	7.6800
Morozov	9.7380	9.6910	9.6450
unconstrained	6.3660	6.5220	6.5390
BIC	8.0330	8.0330	8.0330

RRSB Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	11.4150	11.3050	11.2430
Morozov	12.9790	12.9130	12.8390
unconstrained	10.2130	10.8250	10.7910
BIC	9.7990	9.7990	9.7990

Hedrih Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	6.4520	6.5130	6.5450
Morozov	8.0920	8.0720	8.0910
unconstrained	5.3050	5.3250	5.3400
BIC	7.7890	7.7890	7.7890

Results for 30% Noise

Log-Normal Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	6.5100	6.4480	6.4090
Morozov	8.4620	8.3670	8.3520
unconstrained	5.2170	5.3040	5.3630
BIC	7.2400	7.2400	7.2400

RRSB Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	10.1140	9.9110	9.7500
Morozov	11.3290	11.2280	11.1740
unconstrained	8.7600	9.5200	9.3600
BIC	8.6240	8.6240	8.6240

Hedrih Distribution			
method	average model space dimensions		
	Tikhonov	min. first fin. diff.	Twomey
constrained	5.9610	5.9730	5.9980
Morozov	7.4950	7.4850	7.4510
unconstrained	4.7750	4.7880	4.8140
BIC	7.2950	7.2950	7.2950

3.2.6 Extreme Cases

If the relative error of the reconstruction (compared with the original size distribution) is equal or even greater than 100 percent, we regard the inversion as failed. Note that the inversion methods returned $\boldsymbol{n} \equiv 0$ by default if none of the kernel matrices in any of the model spaces would yield a reconstruction. Now we list how many times the inversion methods failed in our test runs. To see how trustworthy the results are we present the worst case L^2 errors as well. Finally we display the worst case run times.

Results for 5% Noise

Log-Normal Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	5	5	5
unconstrained	0	0	0
BIC	13	13	13

RRSB Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	28	28	28
unconstrained	0	0	0
BIC	20	21	22

Hedrih Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	3	3	3
unconstrained	0	0	0
BIC	5	5	5

Results for 15% Noise

Log-Normal Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	11	11	11
unconstrained	0	0	0
BIC	12	12	12

RRSB Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	41	41	42
unconstrained	0	0	0
BIC	22	22	23

Hedrih Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	14	14	14
unconstrained	0	0	0
BIC	9	9	9

Results for 30% Noise

3.2.7 Reconstruction Failures

Log-Normal Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	31	31	33
unconstrained	0	0	0
BIC	13	15	15

RRSB Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	3	2	1
Morozov	59	60	61
unconstrained	0	0	0
BIC	37	39	39

Hedrih Distribution			
method	number of L^2-errors ≥ 100 % (out of 1000)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	0	0	0
Morozov	25	25	27
unconstrained	0	0	0
BIC	14	14	16

3.2.8 Worst Case Reconstruction Errors

Results for 5% Noise

Log-Normal Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	26.9719	26.4728	26.4760
Morozov	100	100	100
unconstrained	42.2197	47.5723	46.9064
BIC	$7.9063 \cdot 10^3$	$7.9198 \cdot 10^3$	$7.9401 \cdot 10^3$

RRSB Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	62.2097	42.7859	43.4484
Morozov	106.5653	117.9899	131.8554
unconstrained	67.2122	76.6402	77.1507
BIC	$7.1217 \cdot 10^3$	$7.4831 \cdot 10^3$	$7.9567 \cdot 10^3$

Hedrih Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	19.6190	19.6144	19.6157
Morozov	174.6250	174.6358	174.6417
unconstrained	29.3923	21.0755	17.9936
BIC	$2.0139 \cdot 10^3$	$2.0153 \cdot 10^3$	$2.0172 \cdot 10^3$

Results for 15% Noise

Log-Normal Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	49.4987	49.4428	49.4388
Morozov	$6.0234 \cdot 10^3$	$6.0240 \cdot 10^3$	$6.0249 \cdot 10^3$
unconstrained	65.9515	61.5108	60.4107
BIC	$4.6867 \cdot 10^3$	$4.6868 \cdot 10^3$	$4.6872 \cdot 10^3$

RRSB Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	70.1768	63.5123	64.0486
Morozov	184.4034	201.3391	255.3595
unconstrained	80.7321	80.6531	82.7335
BIC	$6.3531 \cdot 10^3$	$6.3536 \cdot 10^3$	$6.3542 \cdot 10^3$

Hedrih Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	30.8181	29.8020	29.7911
Morozov	100.0000	100.0000	100.0000
unconstrained	34.3466	34.2846	34.2059
BIC	$4.8311 \cdot 10^3$	$4.9191 \cdot 10^3$	$5.0158 \cdot 10^3$

Results for 30% Noise

Log-Normal Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	80.2893	48.4626	48.4429
Morozov	$2.3420 \cdot 10^3$	$2.3491 \cdot 10^3$	$2.3542 \cdot 10^3$
unconstrained	70.5992	64.1337	66.8937
BIC	$7.1044 \cdot 10^3$	$7.2810 \cdot 10^3$	$7.5079 \cdot 10^3$

RRSB Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	131.5244	131.5263	104.9113
Morozov	182.0357	197.1594	252.0434
unconstrained	84.6883	81.1661	82.4180
BIC	$2.9711 \cdot 10^4$	$2.9806 \cdot 10^4$	$2.9932 \cdot 10^4$

Hedrih Distribution			
method	worst case L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	39.7993	37.6784	37.6685
Morozov	100.0000	100.0000	115.0592
unconstrained	57.7572	57.7192	57.6711
BIC	$5.0893 \cdot 10^3$	$5.0995 \cdot 10^3$	$5.1149 \cdot 10^3$

3.2.9 Worst Case Run Times

Results for 5% Noise

Log-Normal Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	12.0411	12.4570	12.5885
Morozov	4.5225	4.5381	4.5857
unconstrained	0.5769	0.5995	0.5856
BIC	0.1389	0.1142	0.1320

RRSB Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	25.6129	26.2979	27.2205
Morozov	3.7168	3.7758	3.8567
unconstrained	0.6483	0.6258	0.5978
BIC	0.1252	0.1320	0.1388

Hedrih Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	8.1922	8.6590	9.3658
Morozov	4.3966	4.4999	4.6788
unconstrained	1.2404	0.9995	1.0106
BIC	0.1964	0.1286	0.1162

Results for 15% Noise

Log-Normal Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	7.4443	7.6258	7.8005
Morozov	4.6163	4.1752	4.3086
unconstrained	0.5563	0.5765	0.5585
BIC	0.1068	0.1120	0.1706

RRSB Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	25.2151	23.6935	23.0974
Morozov	3.9076	4.0044	3.9774
unconstrained	0.9480	0.8766	0.7568
BIC	0.1198	0.1299	0.1376

Hedrih Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	13.3920	10.5595	8.7664
Morozov	3.9038	4.0082	4.0133
unconstrained	0.8075	0.6807	0.7742
BIC	0.1068	0.1137	0.1263

Results for 30% Noise

Log-Normal Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	5.4749	5.6907	5.6535
Morozov	3.6386	3.7187	3.7566
unconstrained	0.6710	0.6553	0.6529
BIC	0.2008	0.3046	0.1433

RRSB Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	13.4702	14.9904	13.6324
Morozov	3.3278	3.4159	3.3878
unconstrained	0.8178	0.5565	0.7891
BIC	0.1130	0.1187	0.1710

Hedrih Distribution			
method	worst case run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
constrained	6.3211	5.9184	6.6355
Morozov	6.4399	6.6852	6.1017
unconstrained	0.7770	0.7160	0.8296
BIC	0.1284	0.0960	0.1195

3.3 Conclusion

The constrained method had the smallest average L^2-errors and close to zero failure rates. Only for the RRSB distributions at 30% noise were three, two, and one failures out of 1000 inversions recorded for the different priors, respectively. The overall worst case reconstruction error of 131.5263% was only moderately above 100%. It needed the longest run times from all methods, but even the overall worst case run time of 27.2205 seconds was below our thirty-second requirement. The difference of the average L^2-errors depending on the three priors we applied was not very prominent. For the constrained method the differences were at most ca. 2%, where none of the priors could be determined as the best performing one. For the other inversion methods the L^2-errors behave similarly depending on the priors, but here the differences were more prominent, but always below 6%. A common trend was that these differences rose for higher noise levels. From our observations, none of the three priors we used could be identified clearly as that one giving the best regularization.

For the Morozov method the average L^2-errors were for each noise level higher than those of the constrained method. The differences were growing for higher noise levels. For 5% noise they were in the 1 to 2% range, for 15% in the 4 to 10% range and for 30% in the 6 to 11% range. For each noise level, the average run times represented only about one fifth of those of the constrained method. However, for each noise level the numbers of failures was significantly higher. Especially for RRSB distributions this was up to 3, 4 and 6% respectively. The overall worst case L^2-error of $6.0249 \cdot 10^3$% was clearly higher than 100%.

The run times of the unconstrained method were always one third to one half of the Morozov method run times. The unconstrained method did not show any reconstruction failures at all. The overall worst case L^2-error was a relatively moderate 84.6883%, but the average L^2-errors were always bigger than the Morozov method L^2-errors reaching 2 to 14% for 30% noise and already 1.5 to 3 times as big as the constrained method L^2-errors.

The BIC method was by far the fastest one with run times of only a few hundredths of a second, but the average L^2 errors growing from ca. 15 to 40% for 5% noise to ca. 40 to 80% for 30% noise were rather poor. The overall worst case L^2-error was even $2.9932 \cdot 10^4$%.

For every inversion method the average model space dimension was declining for growing noise levels, which is reflected by the higher average run times for smaller noise levels.

For practical FASP experiments we conclude that the constrained method performed best, because its average L^2-errors were smallest, had virtually no failures, and clearly satisfied our thirty-seconds run-time limit even in the worst cases. It showed the best convergence behavior for descending noise levels as well.

Chapter 4

Markov Chain Monte Carlo Methods

4.1 Hyperprior Distributions

An alternative to the discrepancy principle for determining the regularization parameter γ are *hyperprior distributions.* As outlined in [21] the problem of selecting the regularization parameter is made here a part of the inference problem. We recall that using the discrepancy principle γ is obtained from fitting the residual of the regularized solution to an estimate of the noise magnitude. This approach fully determines γ such that it is regarded as given for the subsequent retrieval of the sought-after entity $\boldsymbol{n}$ then. This means that γ is implicitly given as a function of the noisy data $\boldsymbol{e}$. In the hyperprior approach instead the regularization parameter γ and $\boldsymbol{n}$ are retrieved simultaneously. This is done by considering the posterior densitiy to be of the form $p_{posterior}(\gamma, \boldsymbol{n}|\boldsymbol{e})$. Now the actual hyperprior is a density $p_{hyperprior}(\gamma)$ for the regularization parameter, which is believed to be a reasonable model for it. It is important to note that in contrast to the discrepancy principle the hyperprior distrubution is independent from the noisy data $\boldsymbol{e}$. Now taking into account the hyperprior, the joint prior density $p_{prior}(\gamma, \boldsymbol{n})$ is given by

$$p_{prior}(\gamma, \boldsymbol{n}) = p_{prior}(\boldsymbol{n}|\gamma) \times p_{hyperprior}(\gamma). \tag{4.1.1}$$

Recalling Bayes' theorem, we have

$$\begin{aligned} p_{posterior}(\gamma, \boldsymbol{n}|\boldsymbol{e}) &\propto p_{observed}(\boldsymbol{e}|\boldsymbol{n}) \times p_{prior}(\gamma, \boldsymbol{n}) \\ &\propto p_{observed}(\boldsymbol{e}|\boldsymbol{n}) \times p_{prior}(\boldsymbol{n}|\gamma) \times p_{hyperprior}(\gamma). \end{aligned}$$

Our observed probability density was of the form

$$p_{observed}(\boldsymbol{e}|\boldsymbol{n}) \propto \exp\left(-\tfrac{1}{2}\left(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_k\boldsymbol{n} - \boldsymbol{e})\|_2^2\right)\right), \tag{4.1.2}$$

with the noisy data vector $\boldsymbol{e}$, the kernel matrix $\boldsymbol{K}_k$ for the model space dimension N_k, i.e. $\boldsymbol{n} \in \mathbb{R}^{N_k}$, and the covariance matrix $\boldsymbol{\Sigma_\sigma}$. As prior distribution we selected

$$p_{prior}(\boldsymbol{n}|\gamma) = \left(\frac{\pi}{2\gamma}\right)^{-\frac{N_k}{2}} \exp(-\tfrac{1}{2}\gamma\|\boldsymbol{n}\|_2^2) I_{\geq 0}(\boldsymbol{n}).$$

Following [21] we took for the hyperprior a *Rayleigh distribution*

$$p_{hyperprior}(\gamma) = \frac{\gamma}{\gamma_0^2} \exp\Big(-\tfrac{1}{2}\Big(\frac{\gamma}{\gamma_0}\Big)^2\Big).$$

Therefore we have

$$p_{prior}(\gamma, \boldsymbol{n}) \propto \gamma^{\frac{N_k+2}{2}} \exp\Big(-\tfrac{1}{2}\Big(\gamma\|\boldsymbol{n}\|_2^2 + \Big(\frac{\gamma}{\gamma_0}\Big)^2\Big)\Big) I_{\geq 0}(\boldsymbol{n}),$$

which yields

$$p_{posterior}(\gamma, \boldsymbol{n}|\boldsymbol{e}) \propto \exp\Big(-\tfrac{1}{2}\Big(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_k\boldsymbol{n}-\boldsymbol{e})\|_2^2+\gamma\|\boldsymbol{n}\|_2^2+\Big(\frac{\gamma}{\gamma_0}\Big)^2-(N_k+2)\log(\gamma)\Big)\Big) I_{\geq 0}(\boldsymbol{n}). \tag{4.1.3}$$

4.2 Model Selection

In Section 2.6.2 we used the discrepancy principle for the model selection as well. Starting a the most coarse discretization, our approach was to refine the discretization successively and to apply the discrepancy principle for a whole set of Morozov safety factors τ, until a set of candidate solutions was obtained. Thus these candidate solutions lived on the coarsest discretizations possible on which the discrepancy principle was applicable.

For the hypperprior approach this adaptive strategy is not available. Here we have to take into account all discretizations with the number of discretization points ranging from 3 to 50. For the fully Bayesian model selection strategy presented here we use model averaging, i.e. we compute integrals for the model posterior distributions. An alternative Bayesian model selection strategy based on MAP-estimators was given in [32]. Another approach for the model selection is to incorporate it in the Monte Carlo posterior evaluation using a so-called *reversible jump method*, cf. [33]. Here the drawn samples may lie in different model spaces. Our strategy is different, i.e. we first perform the model selection and then keep the model order fixed.

We also use eliminated zero boundary conditions, so the model space dimensions actually range from 1 to 48. We begin with deriving the model probabilities $p(N_k|\boldsymbol{e})$ using a hyperprior. In contrast to the discrepancy principle approach, we do not consider γ as a function of $\boldsymbol{e}$. Thus we do not consider it as given alongside $\boldsymbol{e}$ as we did when computing the model probabilities $p(N_k|\gamma, \boldsymbol{e})$ for the discrepancy principle. In the hyperprior approach we consider γ as a free sought-after parameter as well, therefore the model probabilities must be based on the full joint probability density $p(\boldsymbol{e}, \gamma, \boldsymbol{n}|N_k)$. Then we have to marginalize over γ, too, in order to obtain $p(N_k|\boldsymbol{e})$, which can be seen in detail in from

$$\begin{aligned} p(N_k|\boldsymbol{e}) &= \frac{p(\boldsymbol{e}|N_k)p(N_k)}{\sum_i p(\boldsymbol{e}|N_i)p(N_i)} \\ \text{with} \quad p(\boldsymbol{e}|N_i) &= \int_0^\infty \int_{\mathbb{R}^{N_i}} p(\boldsymbol{e}, \gamma, \boldsymbol{n}|N_i) d\boldsymbol{n} d\gamma \\ &= \int_0^\infty \int_{\mathbb{R}^{N_i}} p(\boldsymbol{e}|N_i, \boldsymbol{n}) p(\gamma, \boldsymbol{n}|N_i) d\boldsymbol{n} d\gamma \end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \int_{\mathbb{R}^{N_i}} p(\boldsymbol{e}|N_i, \boldsymbol{n}) p(\boldsymbol{n}|\gamma, N_i) p(\gamma|N_i) d\boldsymbol{n} d\gamma \\
&= \int_0^\infty \int_{[0,\infty)^{N_i}} B^{-1} C_i^{-1} \gamma_0^{-2} \exp\Big(-\tfrac{1}{2}\Big(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n} - \boldsymbol{e})\|_2^2 \\
&\qquad + \gamma\|\boldsymbol{n}\|_2^2 + \Big(\frac{\gamma}{\gamma_0}\Big)^2 - (N_i + 2)\log(\gamma)\Big)\Big) d\boldsymbol{n} d\gamma,
\end{aligned}
$$

$$
\text{with} \quad B = (2\pi)^{\frac{N_l}{2}} \big|\det(\boldsymbol{\Sigma_\sigma})\big|^{\frac{1}{2}} \quad \text{and} \quad C_i = \Big(\frac{\pi}{2}\Big)^{\frac{N_i}{2}}.
$$

Here B is the normalizing constant of $p(\boldsymbol{e}|N_i, \boldsymbol{n}) = p_{observed}(\boldsymbol{e}|\boldsymbol{n})$, the constant C_i the factor of $\gamma^{-\frac{N_i}{2}}$ in the normalizing constant of $p(\boldsymbol{n}|\gamma, N_i) = p_{prior}(\boldsymbol{n}|\gamma)$ and γ_0^{-2} the normalizing constant of $p(\gamma|N_i) = p_{hyperprior}(\gamma)$.

The integrals $p(\boldsymbol{e}|N_i)$ cannot be evaluated exactly. We introduce the auxiliary function

$$
f(\gamma, \boldsymbol{n}) := \tfrac{1}{2}\Big(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n} - \boldsymbol{e})\|_2^2 + \gamma\|\boldsymbol{n}\|_2^2 + \Big(\frac{\gamma}{\gamma_0}\Big)^2 - (N_i + 2)\log(\gamma)\Big)
$$

and approximate the integrands with

$$
p(\boldsymbol{e}, \gamma, \boldsymbol{n}|N_i) \approx \exp\Big(- f(\gamma', \boldsymbol{n}') - \tfrac{1}{2}\big((\gamma, \boldsymbol{n}^T) - (\gamma', \boldsymbol{n}'^T)\big) Hess_f(\gamma', \boldsymbol{n}') \big((\gamma, \boldsymbol{n}^T) - (\gamma', \boldsymbol{n}'^T)\big)^T\Big),
$$

where

$$
(\gamma', \boldsymbol{n}'^T)^T := \operatorname*{argmin}_{(\gamma, \boldsymbol{n}^T)^T \in [0,\infty)\times\mathbb{R}^{N_i}} f(\gamma, \boldsymbol{n}),
$$

and the Hessian is given by

$$
Hess_f(\gamma, \boldsymbol{n}) = \begin{pmatrix} \dfrac{1}{\gamma_0^2} + \dfrac{N_i + 2}{2}\dfrac{1}{\gamma^2} & \boldsymbol{n}^T \\ \boldsymbol{n} & \boldsymbol{K}_i^T \boldsymbol{\Sigma_\sigma}^{-1} \boldsymbol{K}_i + \gamma \boldsymbol{I} \end{pmatrix}.
$$

This approximation is commonly known as “Laplace’s method”, cf. [34].

Now the resulting integrals of a multivariate Gaussian distribution over the first quadrant of $\mathbb{R}^{N_i+1}$ are feasible for Genz’ method from Chapter 2. For this the approximations

$$
\begin{aligned}
p(\boldsymbol{e}|N_i) &\approx B^{-1} C_i^{-1} \gamma_0^{-2} \exp\big(- f(\gamma', \boldsymbol{n}')\big) \\
&\quad \times \int_0^\infty \int_{[0,\infty)^{N_i}} \exp\Big(-\tfrac{1}{2}\big((\gamma, \boldsymbol{n}^T) - (\gamma', \boldsymbol{n}'^T)\big) Hess_f(\gamma', \boldsymbol{n}')\big((\gamma, \boldsymbol{n}^T) - (\gamma', \boldsymbol{n}'^T)\big)^T\Big) d\boldsymbol{n} d\gamma \\
&= B^{-1} C_i^{-1} \gamma_0^{-2} \exp\big(- f(\gamma', \boldsymbol{n}')\big) \\
&\quad \times \int_{\left\{\boldsymbol{z} \in \mathbb{R}^{N_i+1} \;|\; \boldsymbol{z} \geq -(\gamma', \boldsymbol{n}'^T)^T\right\}} \exp\Big(-\tfrac{1}{2}\boldsymbol{z}^T Hess_f(\gamma', \boldsymbol{n}')\boldsymbol{z}\Big) d\boldsymbol{z}
\end{aligned}
$$

have to be transformed into the form of the very last integral.

The computation of γ' and $\boldsymbol{n}'$ is basically a one-dimensional problem. The first order optimality system for a minimum of $f(\gamma, \boldsymbol{n})$ is given by

$$
\tfrac{1}{2}\|\boldsymbol{n}\|_2^2 + \frac{\gamma}{\gamma_0^2} - \frac{N_i + 2}{2}\frac{1}{\gamma} = 0 \tag{4.2.1}
$$

$$\big(\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i+\gamma\boldsymbol{I}\big)\boldsymbol{n}-\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{e}=0. \tag{4.2.2}$$

Solving (4.2.2) for $\boldsymbol{n}$ and inserting the solution into (4.2.1) gives that γ' is a root of the scalar function

$$g(\gamma):=\tfrac{1}{2}\big\|\big(\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i+\gamma\boldsymbol{I}\big)^{-1}\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{e}\big\|_2^2+\frac{\gamma}{\gamma_0^2}-\frac{N_i+2}{2}\frac{1}{\gamma}. \tag{4.2.3}$$

We have

$$\frac{d}{d\gamma}g(\gamma)=-\boldsymbol{e}^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i\big(\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i+\gamma\boldsymbol{I}\big)^{-3}\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{e}+\frac{1}{\gamma_0^2}+\frac{N_i+2}{2}\frac{1}{\gamma^2}, \tag{4.2.4}$$

so the sought-after root can be found efficiently with Newton's method. It is left to the reader to verify that the root of $g(\gamma)$ is unique.

As done in Chapter 2 we use uniform model prior probabilities $p(N_i)$. Then the model selection algorithm for hyperpriors is as follows.

Algorithm 3 Model Selection for Hyperpriors

```
$N_l = 48$
$P_{post} = \{\}$
for $i = 1$ to $N_l$ do
    Get $\gamma'$ as solution of
    $\tfrac{1}{2}\big\|\big(\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i+\gamma\boldsymbol{I}\big)^{-1}\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{e}\big\|_2^2+\frac{\gamma}{\gamma_0^2}-\frac{N_i+2}{2}\frac{1}{\gamma}=0.$
    $\boldsymbol{n}'=\big(\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i+\gamma'\boldsymbol{I}\big)^{-1}\boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{e}$
    $f_{max}=\tfrac{1}{2}\Big(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_i\boldsymbol{n}'-\boldsymbol{e})\|_2^2+\gamma'\|\boldsymbol{n}'\|_2^2+\Big(\frac{\gamma'}{\gamma_0}\Big)^2-(N_i+2)\log(\gamma')\Big)$
    $\boldsymbol{H}=\begin{pmatrix}\frac{1}{\gamma_0^2}+\frac{N_i+2}{2}\frac{1}{\gamma'^2} & \boldsymbol{n}'^T\\ \boldsymbol{n}' & \boldsymbol{K}_i^T\boldsymbol{\Sigma_\sigma}^{-1}\boldsymbol{K}_i+\gamma'\boldsymbol{I}\end{pmatrix}$
    $M_i=(2\pi)^{\frac{N_l}{2}}\big|\det(\boldsymbol{\Sigma_\sigma})\big|^{\frac{1}{2}}\left(\frac{\pi}{2}\right)^{\frac{N_i}{2}}\gamma_0^{-2}\exp(-f_{max})$
        $\times\int_{\left\{\boldsymbol{z}\in\mathbb{R}^{N_i+1}\ |\ \boldsymbol{z}\geq-(\gamma',\boldsymbol{n}'^T)^T\right\}}\exp\Big(-\tfrac{1}{2}\boldsymbol{z}^T\boldsymbol{H}\boldsymbol{z}\Big)d\boldsymbol{z}$
    $P_{post}=P_{post}\cup\{M_i\}$
end for
$SumP_{post}=\sum_{i=1}^{N_l}P_{post}(i)$
for $i = 1$ to $N_l$ do
    $P_{post}(i)=P_{post}(i)/SumP_{post}$
end for
```

For the discretization with the highest posterior probability $P_{post}(i)$ we proceed with computing the *conditional mean estimator*

$$\boldsymbol{n}_{CM}:=\int_{\mathbb{R}^{N_i}\times\mathbb{R}}\boldsymbol{n}p_{posterior}(\gamma,\boldsymbol{n}|\boldsymbol{e})d\boldsymbol{n}d\gamma. \tag{4.2.5}$$

The high-dimensional integration needed to compute $\boldsymbol{n}_{CM}$ requires draws from the posterior distribution obtained with a Monte Carlo method, which we will derive in the next section.

4.3 Markov Chain Monte Carlo Methods

We give here a very short introduction into Markov Chain Monte Carlo methods and follow [21]. More detailed background information can be found in e.g. [35].

Let μ be a probability measure and let f be a measurable function on $\mathbb{R}^N$. The basic idea of Monte Carlo methods is to construct a sequence $\boldsymbol{x}_i$, $i = 1, ..., N_{draws}$ which is distributed according to μ. Then we can use this sequence to approximate

$$\int_{\mathbb{R}^N} f(\boldsymbol{n})\mu(d\boldsymbol{n}) \approx \frac{1}{N_{draws}} \sum_{i=1}^{N_{draws}} f(\boldsymbol{x}_i).$$

The sequence is obtained with a so-called *probability transition kernel* $P : \mathbb{R}^N \times \mathfrak{B}$, where $\mathfrak{B} = \mathfrak{B}(\mathbb{R}^N)$ denotes the Borel sets over $\mathbb{R}^N$. It has the properties that for each $B \in \mathfrak{B}$, the mapping $\mathbb{R}^N \to [0,1]$, $\boldsymbol{x} \mapsto P(\boldsymbol{x}, B)$ is a measurable function, and that for each $\boldsymbol{x} \in \mathbb{R}^N$, the mapping $\mathfrak{B} \to [0,1]$, $B \mapsto P(\boldsymbol{x}, B)$ is a probability distribution. Now a *time homogenous Markov chain* is a stochastic process $\{X_j\}_{j=1}^{\infty}$, where its measures μ_{X_j} fulfill

$$\mu_{X_{j+1}}(B_{j+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_j) = \mu_{X_{j+1}}(B_{j+1}|\boldsymbol{x}_j) = P(\boldsymbol{x}_j, B_{j+1}).$$

Thus given $X_1 = \boldsymbol{x}_1$, ..., $X_j = \boldsymbol{x}_j$, the distribution for the new iterate $\boldsymbol{x}_{j+1}$ only depends on its direct predecessor $\boldsymbol{x}_j$. This dependency does not vary with time.

The measure μ is an *invariant measure* of P if for all $B \in \mathfrak{B}$

$$\mu P(B) = \int_{\mathbb{R}^N} P(\boldsymbol{x}, B)\mu(d\boldsymbol{x}) = \mu(B)$$

holds. The transition kernel P is *irreducible* with respect to μ if for each $\boldsymbol{x} \in \mathbb{R}^N$ and $B \in \mathfrak{B}$ with $\mu(B) > 0$, there exists a $k \in \mathbb{N}$ with $P^k(\boldsymbol{x}, B) > 0$, where $P^k(\boldsymbol{x}, B)$ is for $k \geq 2$ and a sequence $\boldsymbol{x}_j$ and B_j iteratively defined by

$$P^k(\boldsymbol{x}_j, B) = \int_{\mathbb{R}^N} P(\boldsymbol{x}_{j+k-1}, B_{j+k})P^{k-1}(\boldsymbol{x}_j, d\boldsymbol{x}_{j+k-1}).$$

So this means for any $\boldsymbol{x}$, that there is for any start point a nonzero probability, that $\boldsymbol{x}$ is reached by the iterated transition kernel in a finite number of steps. A transition kernel P is *periodic*, if there exists a sequence of sets B_1, ..., $B_m \in \mathfrak{B}$ with $P(\boldsymbol{x}, B_{j+1(\text{mod}m)}) = 1$ for all $\boldsymbol{x} \in B_j$ for $j = 1, ..., m$. Thus the Markov chain remains in a loop forever, if its start point lies within one of the sets B_1, ..., B_m. The kernel P is *aperiodic*, if it is not periodic. Now the following proposition forms the basis of Markov chain Monte Carlo algorithms.

Proposition 4.3.1. *Let P be a transition kernel with invariant measure μ. We assume P to be irreducible with respect to μ and aperiodic. Then there holds for any time homogenous Markov chain $\{X_j\}_{j=1}^{\infty}$ generated with P that the iterated measures P^k fulfill for any start point $\boldsymbol{x} \in \mathbb{R}^N$*

$$\lim_{k\to\infty} P^k(\boldsymbol{x}, B) = \mu(B) \quad \textit{for all} \quad B \in \mathfrak{B}.$$

Furthermore holds for any function which is measurable with respect to μ that

$$\lim_{N_{draws}\to\infty} \frac{1}{N_{draws}} \sum_{j=1}^{N_{draws}} f(X_j) = \int_{\mathbb{R}^N} f(\boldsymbol{x})\mu(d\boldsymbol{x})$$

almost certainly.

□

We omit the proof of this proposition here.

There is an infinite number of possibilities for the construction of the transition kernel. For probability distributions, for which all its conditional distributions are known and which have a simply connected support - such distributions often occur in problems of Bayesian inference -, the so called *Gibbs sampler* is a commonly used method for constructing a transition kernel. We will outline it in the next section.

4.4 Gibbs sampler

Using the Gibbs sampler method, the iterates of the Markov chain are drawn from conditional probabilities of the porbability density $p_{posterior}(\gamma, \boldsymbol{n}|\boldsymbol{e})$ under investigation. Here the model vector $(\gamma, \boldsymbol{n}^T)^T$ is partitioned into subvectors $(\boldsymbol{v}_1^T, ..., \boldsymbol{v}_s^T)^T = (\gamma, \boldsymbol{n}^T)^T$. Then for each $j \in \{1, ..., s\}$ the subvector $\boldsymbol{v}_j$ is drawn from the conditional probability $p_{posterior}(\gamma, \boldsymbol{v}_j|\boldsymbol{e}, (\boldsymbol{v}_1^T, ..., \boldsymbol{v}_{j-1}^T, \boldsymbol{v}_{j+1}^T, ..., \boldsymbol{v}_s^T)^T)$, where the complement vector $(\boldsymbol{v}_1^T, ..., \boldsymbol{v}_{j-1}^T, \boldsymbol{v}_{j+1}^T, ..., \boldsymbol{v}_s^T)^T)$ of $\boldsymbol{v}_j$ is given. In the following we outline the full Gibbs sampler method for (4.1.3), which means that the subvectors $\boldsymbol{v}_j$ are given scalars given by the single components of $(\gamma, \boldsymbol{n}^T)^T$. A more detailed introduction into the Gibbs sampler method can be found in [21, pp. 98-105]. It is also proved there, that the strategy of iteratively drawing from conditional distributions gives a transition kernel.

Updating γ

Conditioned on $\boldsymbol{n} \in [0, \infty)^{N_k}$, the posterior probability density is of the form

$$p_{posterior}(\gamma|\boldsymbol{n}, \boldsymbol{e}) \propto \exp\left(-\tfrac{1}{2}\left(\gamma\|\boldsymbol{n}\|_2^2 + \left(\frac{\gamma}{\gamma_0}\right)^2 - (N_k + 2)\log(\gamma)\right)\right).$$

Its cumulative distribution $F_{posterior}(\gamma|\boldsymbol{n}, \boldsymbol{e}) = \int_0^\gamma p_{posterior}(\gamma|\boldsymbol{n}, \boldsymbol{e})d\gamma$ can not be evaluated analytically. Therefore we use a simple approximate sampling method, where we first truncate the unnormalized density function

$$f_{poseterior}(\gamma|\boldsymbol{n}, \boldsymbol{e}) := \exp\left(-\tfrac{1}{2}\left(\gamma\|\boldsymbol{n}\|_2^2 + \left(\frac{\gamma}{\gamma_0}\right)^2 - (N_k + 2)\log(\gamma)\right)\right)$$

at some $\gamma_{max} = 2^{t_{max}}$, with $t_{max} \geq 1$ being the smallest integer with $f_{posterior}\left(2^{t_{max}}|\boldsymbol{n}, \boldsymbol{e}\right) < 10^{-100}$. Then we approximate the truncated unnormalized density with a piecewise linear function using 10000 intervals of the same length covering $[0, \gamma_{max}]$. We apply the inverse cdf method using this piecewise linear function in order to obtain approximate draws from $p_{posterior}(\gamma|\boldsymbol{n}, \boldsymbol{e})$.

Updating $\boldsymbol{n}$

For $i \in \{1, ..., N_k\}$ we introduce the subvector $(\boldsymbol{n})_{-i} := (n_1, ..., n_{i-1}, n_{i+1}, ..., n_{N_k})^T$ of $\boldsymbol{n}$. Then the conditional probability densities are of the form

$$p_{posterior}(n_i|\gamma, (\boldsymbol{n})_{-i}, \boldsymbol{e}) \propto \exp\left(-\tfrac{1}{2}(a_i n_i^2 + b_i n_i + c_i)\right) I_{\geq 0}(n_i), \tag{4.4.1}$$

where a_i, b_i and c_i depend on γ and $(\boldsymbol{n})_{-i}$. For computational efficiency we derive in the following explicit equations for a_i, b_i and c_i in terms of $\boldsymbol{K}_k$, $\boldsymbol{e}$, $\boldsymbol{\Sigma_\sigma}$, $(\boldsymbol{n})_{-i}$ and γ. We begin with defining

$$\boldsymbol{H} := \boldsymbol{K}_k^T \boldsymbol{\Sigma_\sigma}^{-1} \boldsymbol{K}_k, \qquad \boldsymbol{b} := 2\boldsymbol{K}_k^T \boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}} \boldsymbol{e} \qquad \text{and} \qquad c := \|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}} \boldsymbol{e}\|_2^2.$$

Then we have

$$\boldsymbol{n}^T \left(\boldsymbol{H} + \gamma \boldsymbol{I}\right) \boldsymbol{n} - \boldsymbol{b}^T \boldsymbol{n} + c = \|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}} (\boldsymbol{K}_k \boldsymbol{n} - \boldsymbol{e})\|_2^2 + \gamma \|\boldsymbol{n}\|_2^2.$$

Now we denote with $(\boldsymbol{H})_{i,-i}$ the i-th row of $\boldsymbol{H}$ with the i-th column canceled and with $(\boldsymbol{H})_{-i,-i}$ the matrix obtained from $\boldsymbol{H}$ by canceling both its i-th row and i-th column. Our next step is to factor out the monomials n_i^2 and n_i in the left hand side of above equation. Then by equating their coefficients the explicit relations

$$\begin{aligned}
a_i &= (\boldsymbol{H})_{ii} + \gamma \\
b_i &= 2(\boldsymbol{H})_{i,-i} (\boldsymbol{n})_{-i} - (\boldsymbol{b})_i \\
c_i &= (\boldsymbol{n})_{-i}^T (\boldsymbol{H})_{-i,-i} (\boldsymbol{n})_{-i} + \gamma \|(\boldsymbol{n})_{-i}\|_2^2 - (\boldsymbol{b})_{-i}^T (\boldsymbol{n})_{-i} + \|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}} \boldsymbol{e}\|_2^2
\end{aligned}$$

are readily obtained.

Let us now turn to the cumulative distribution function $F(n_i|\gamma, \boldsymbol{e}, (\boldsymbol{n})_{-i})$. To simplify notation, we define for a function $f : \mathbb{R} \to \mathbb{R}$ the operator

$$\bar{\mathrm{I}}[f(s)](x) := \Big(\int_0^\infty f(s) ds \Big)^{-1} \int_0^x f(s) ds,$$

where we assume that both integrals involving $f(s)$ exist. Thus we obtain

$$\begin{aligned}
F_{posterior}(n_i|\gamma, \boldsymbol{e}, (\boldsymbol{n})_i) &= \bar{\mathrm{I}}\big[\exp\left(-\tfrac{1}{2}(a_i s^2 + b_i s + c_i)\right)\big](n_i) \\
&= \bar{\mathrm{I}}\Big[\exp\Big(-\tfrac{1}{2}\Big(\sqrt{a_i} s + \frac{b_i}{2\sqrt{a_i}}\Big)^2\Big)\Big](n_i) \\
&= \Big(\int_{L_i}^\infty \exp(-t^2) dt\Big)^{-1} \int_{L_i}^{U_i} \exp(-t^2) dt
\end{aligned}$$

with

$$L_i := \frac{b_i}{2\sqrt{2a_i}} \qquad \text{and} \qquad U_i := \frac{\sqrt{a_i}}{\sqrt{2}} n_i + \frac{b_i}{2\sqrt{2a_i}}.$$

The second equation was obtained by completing the square in the quadratic $a_i s^2 + b_i s + c_i$. Then the constant rest term not depending on s canceled out in the integrals in the numerator and denominator, such that only the squared term is left in their integrands. The third equation was obtained from a linear change of the variable s.

We now introduce the *error function* $\operatorname{erf}(n_i) := \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^{n_i} \exp\left(-t^2\right) dt$.

Using $\lim_{n_i \to \infty} \operatorname{erf}(n_i) = 1$, the last representation of $F(n_i|\gamma, \boldsymbol{e}, \boldsymbol{n}_i)$ can be expressed as

$$F_{posterior}(n_i|\gamma, \boldsymbol{e}, \boldsymbol{n}_i) = \Big(1 - \operatorname{erf}\Big(\frac{b_i}{2\sqrt{2a_i}}\Big)\Big)^{-1} \Big(\operatorname{erf}\Big(\frac{\sqrt{a_i}}{\sqrt{2}} n_i + \frac{b_i}{2\sqrt{2a_i}}\Big) - \operatorname{erf}\Big(\frac{b_i}{2\sqrt{2a_i}}\Big)\Big).$$

Let u be a sample from a uniform distribution on $[0,1]$. Then a sample from $p(n_i|\gamma,(\boldsymbol{n})_{-i},\boldsymbol{e})$ can be obtained by solving $F(n_i|\gamma,\boldsymbol{e},(\boldsymbol{n})_{-i}) = u$ for n_i. With previous computations this solution is given by

$$n_i = \frac{\sqrt{2}}{\sqrt{a_i}}\operatorname{erf}^{-1}\Big((1-u)\operatorname{erf}\Big(\frac{b_i}{2\sqrt{2a_i}}\Big) + u\Big) - \frac{b_i}{2a_i}.$$

The results of this section are summarized in

Algorithm 4 Gibbs Sampler

```
 1: $\boldsymbol{H} = \boldsymbol{K}_k^T \boldsymbol{\Sigma_\sigma}^{-1} \boldsymbol{K}_k$
 2: $\boldsymbol{b} := 2\boldsymbol{K}_k^T \boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}} \boldsymbol{e}$
 3: Compute the MAP estimator
    $(\gamma_{MAP}, \boldsymbol{n}_{MAP}^T)^T = \underset{\gamma\in\mathbb{R},\, \boldsymbol{n}\in\mathbb{R}^k}{\operatorname{argmin}} \ \frac{1}{2}\Big(\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{K}_k\boldsymbol{n} - \boldsymbol{e})\|_2^2 + \gamma\|\boldsymbol{n}\|_2^2 + \Big(\frac{\gamma}{\gamma_0}\Big)^2$
        $- (N_k + 2)\log(\gamma)\Big)$
        s.t. $\gamma \geq 0, \quad \boldsymbol{n} \geq 0$
 4: $\gamma^0 = \gamma_{MAP}$
 5: $\boldsymbol{n}^0 = \boldsymbol{n}_{MAP}$
 6: $N_{draws} = 5000$
 7: for $i = 1$ to $N_{draws}$ do
 8:     Truncate $f(\gamma) := \exp\Big(-\frac{1}{2}\Big(\gamma\|\boldsymbol{n}^{i-1}\|_2^2 + \Big(\frac{\gamma}{\gamma_0}\Big)^2 - (N_k + 2)\log(\gamma)\Big)\Big)$ at
        $\gamma_{max} = 2^{t_{max}}$ with $t_{max} = \min\{t \in \mathbb{N},\ t \geq 1 \mid f(2^t) < 10^{-100}\}$.
 9:     Approximate the truncated version of $f(\gamma)$ with a piecewise linear
        function $f_{approx}(\gamma)$ using 10000 linearly spaced subintervals of $[0, \gamma_{max}]$.
10:     Draw $u \sim U([0,1])$.
11:     $\gamma^i = G^{-1}(u)$ with $G(\gamma) = \big(\int_0^{\gamma_{max}} f_{approx}(s)ds\big)^{-1} \int_0^{\gamma} f_{approx}(s)ds$
12:     for $j = 1$ to $N_k$ do
13:         $\boldsymbol{n}_j^i = \big(n_1^i, ..., n_{j-1}^i, n_{j+1}^{i-1}, ..., n_{N_k}^{i-1}\big)^T$
14:         $a_j^i = (\boldsymbol{H})_{jj} + \gamma^i$
15:         $b_j^i = 2(\boldsymbol{H})_{j,-j}\boldsymbol{n}_j^i - (\boldsymbol{b})_j$
16:         Draw $u \sim U([0,1])$.
17:         $n_j^i = \frac{\sqrt{2}}{\sqrt{a_j^i}}\operatorname{erf}^{-1}\Big((1-u)\operatorname{erf}\Big(\frac{b_j^i}{2\sqrt{2a_j^i}}\Big) + u\Big) - \frac{b_j^i}{2a_j^i}$
18:     end for
19: end for
```

In lines 4 and 5 we initialize γ and $\boldsymbol{n}$ with their corresponding MAP estimators which have been found in the preceding model selection step. The MAP estimators lie within the domain of convergence of the Gibbs sampler method, so initial burn-in iterations to reach convergence are not needed. We computed $N_{draws} = 5000$ samples for any model space dimension N_k determined in the model selection step.

4.5 Numerical Study

We conducted a numerical study with the same test cases as in Section 3.2.2, i.e. with the same simulated original particle size distributions from the log-normal, RRSB and Hedrih families specified by the same sets of parameters, and with the same amount of noise contaminating the simulated original spectral extinctions. For each of the three noise levels, we simulated for each of the three families 1000 single inversions. For each single inversion, we calculated 300 noisy spectral extinctions with additive Gaussian noise with a standard deviation of 5%, 15% and 30% respectively of the true extinctions, and we used the sample means and variances of the them as estimates for the true extinctions and noise magnitudes.

In the following we present the results for a Rayleigh hyperprior with $\gamma_0 = 0.1$. We first show average and maximal run times for the model selection step. Then we list the average and maximal run times for the Gibbs sampler. We proceed with the average and maximal L^2-errors of the MAP estimators followed by the average and maximal L^2-errors of the conditional mean estimators. At last we show the average and maximal model space dimensions of the results.

4.5.1 L^2-errors of MAP Estimators

Results for 5% Noise

	L^2-errors (%)		
	Log-Normal	RRSB	Hedrih
average	28.7463	120.2392	11.6742
worst case	175.8651	$1.6128 \cdot 10^3$	40.1555

Results for 15% Noise

	L^2-errors (%)		
	Log-Normal	RRSB	Hedrih
average	61.3485	297.4809	14.7444
worst case	329.1157	$1.4520 \cdot 10^3$	45.1685

Results for 30% Noise

	L^2-errors (%)		
	Log-Normal	RRSB	Hedrih
average	90.0913	323.6232	20.7106
worst case	617.1629	$1.1379 \cdot 10^3$	51.9149

4.5.2 L^2-errors of the Conditional Mean Estimators

Results for 5% Noise

	L^2-errors (%)		
	Log-Normal	RRSB	Hedrih
average	12.6675	16.7886	10.8433
worst case	30.5679	113.1329	20.7521

Results for 15% Noise

	L^2-errors (%)		
	Log-Normal	RRSB	Hedrih
average	18.0044	24.2764	13.1940
worst case	39.9891	216.8520	32.3436

Results for 30% Noise

	L^2-errors (%)		
	Log-Normal	RRSB	Hedrih
average	22.4450	29.2653	19.2887
worst case	46.3712	149.3014	42.3618

4.5.3 Run Times for the Model Selection Step

Results for 5% Noise

	run times (s)		
	Log-Normal	RRSB	Hedrih
average	11.6462	12.0560	11.9135
worst case	28.1454	29.2136	29.7096

Results for 15% Noise

	run times (s)		
	Log-Normal	RRSB	Hedrih
average	11.5484	11.3946	12.0545
worst case	21.0221	22.9508	30.2464

Results for 30% Noise

	run times (s)		
	Log-Normal	RRSB	Hedrih
average	10.5608	11.0053	11.1530
worst case	14.1052	15.8992	14.5782

4.5.4 Run Times for the Gibbs Sampler

Results for 5% Noise

	run times (s)		
	Log-Normal	RRSB	Hedrih
average	14.0753	16.0902	13.8658
worst case	19.2091	22.7198	27.4881

Results for 15% Noise

	run times (s)		
	Log-Normal	RRSB	Hedrih
average	12.9457	14.1758	13.4292
worst case	30.7609	28.1368	25.7760

Results for 30% Noise

	run times (s)		
	Log-Normal	RRSB	Hedrih
average	11.3974	12.8834	12.0432
worst case	16.3708	20.3767	17.2901

4.5.5 Model Space Dimensions

Results for 5% Noise

	average model space dimensions		
	Log-Normal	RRSB	Hedrih
average	9.7920	12.2500	8.2740

Results for 15% Noise

	average model space dimensions		
	Log-Normal	RRSB	Hedrih
average	7.9990	10.6220	7.4360

Results for 30% Noise

	average model space dimensions		
	Log-Normal	RRSB	Hedrih
average	7.1170	9.6930	7.0190

4.5.6 Reconstruction Failures of the Conditional Mean Estimators

Results for 5% Noise

	number of L^2-errors ≥ 100 % (out of 1000)		
	Log-Normal	RRSB	Hedrih
number (out of 1000)	0	1	0

Results for 15% Noise

	number of L^2-errors ≥ 100 % (out of 1000)		
	Log-Normal	RRSB	Hedrih
number (out of 1000)	0	13	0

Results for 30% Noise

	number of L^2-errors ≥ 100 % (out of 1000)		
	Log-Normal	RRSB	Hedrih
number (out of 1000)	0	18	0

4.6 Conclusion

Clearly the MAP-estimators are only of practical use in order to save burn-in iterations for the Gibbs sampler, since their L^2-errors are too big. Regarding the average and worst case L^2-errors, the results of the Gibbs sampler are matched best by the results of the Morozov method from Chapter 3, i.e. by the results of the discrepancy principle under nonnegativity constraints using only the expected value of the residual multiplied with the Morozov factor $\tau = 1.1$ as noise estimate. Only for RRSB distributions the results are better matched by those of the unconstrained method

The total average run time, i.e. the sum of the run times of the model step and the gibbs sampler, was highest for RRSB distributions. It declined from ca. 28 seconds for 5% noise to ca. 24 seconds for 30% noise. The total run times were much higher than those of any of the methods from Chapter 2, but still fulfill our 30 seconds time limit.

In the next chapter we will investigate a more complex model selection problem for the retrieval of particle size distributions of two-component aerosols. Here we will only consider retrieval methods based on the discrepancy principle for their superior efficiency.

Chapter 5

Two-Component Aerosols

In the preceding sections it was assumed that the aerosol particles consist of a known material, and therefore the refractive indices $m_{part}(l)$ needed to compute the extinction efficiency $Q_{ext}(m_{med}(l), m_{part}(l), r, l)$ were given exactly as well. But this is not generally the case in real experimental measurements where typically both size distributions and optical properties of scattering particles are unknown. In the ideal case we could set up an additional device for measuring the aerosol refractive indices and perform a two-stage measurement process, where the first step is to retrieve the refractive indices as preparation for the second step of reconstructing the size distribution, but this is not practical. Indeed all measurement techniques for optical properties of aerosol particles demand a pretreatment of the aerosol itself such as vaporizing it into its gas phase or transforming it into a monodisperse aerosol. This would make the FASP too inefficient to be of practical use.

In real applications we simply want to examine some aerosol components of particular interest. Thus we assume that the aerosol to be investigated is a mixture of a small number of known materials, such that only the problem remains to retrieve the volume fractions of these materials in the whole composite aerosol. As an initial explorative step into this general problem we further assume that the aerosol is made up of only two materials.

To compute the refractive indices of composite aerosols from those of their pure components so-called *mixing rules* are used. Some of these are compared in [36]. Let $m_1 = k_1 + in_1$ and $m_2 = k_2 + in_2$ be the refractive indices of two aerosol components for a wavelength l of the incident light. We adopt the most commonly used rule, the *Lorentz-Lorenz rule.* Here the total refractive $m_{tot} = k_{tot} + in_{tot}$ is obtained from the relation

$$\frac{m_{tot}^2 - 1}{m_{tot}^2 + 2} = f_1 \frac{m_1^2 - 1}{m_1^2 + 2} + f_2 \frac{m_2^2 - 1}{m_2^2 + 2}, \tag{5.0.1}$$

where f_1 and f_2 are the volume fractions of the components.

Now our new problem is to invert the parameter-dependent integral equation

$$\int_{r_{min}}^{r_{max}} k_p(r, l) n(r) dr = e(l), \tag{5.0.2}$$

where the sought-after parameter $p \in [0, 1]$ characterizes the unknown volume fractions. Let $m_p(l)$ denote the solution m_{tot} of (5.0.1) with $f_1 = p$ and $f_2 = 1 - p$. Then the p-dependent kernel function is given by

$$k_p(r, l) = \pi r^2 Q_{ext}(m_{med}(l), m_p(l), r, l).$$

Mathematically this means that in addition to inverting it we have to identify the "right" integral operator K_p from the set

$$\left\{ (K_p n)(l) := \int_{r_{min}}^{r_{max}} k_p(r,l) n(r) dr \,\middle|\, p \in [0,1] \right\}.$$

We can easily check that $k_p(r,l)$ depends continuously on p and therefore so do the discrete approximations $\boldsymbol{K}_{k,p}$ to K_p as well. Furthermore the continuous dependence of regularized solutions on the data for problems of this type was shown in [37]. We again make Assumption 7.2.1. So by setting

$$\boldsymbol{K}_p := \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{K}_{k,p} \quad \text{and} \quad \boldsymbol{r} = \boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{e}_{true} + \boldsymbol{\delta})$$

as in Section 2.5 we obtain the p-parametrized quadratic programming problem

$$\min_{\boldsymbol{n} \in \mathbb{R}^N} \tfrac{1}{2} \|\boldsymbol{K}_p \boldsymbol{n} - \boldsymbol{r}\|_2^2 + \tfrac{1}{2}\gamma \|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b}. \tag{5.0.3}$$

as in Section 2.2 for the computation of the maximum a posteriori solution.

5.1 Fraction Retrieval for two Aerosol Components

For the determination of the parameter p we modify the adaptive model-generation algorithm from Section 2.6. As a preparation we prove a continuity result.

Proposition 5.1.1. *The minimizer $\boldsymbol{n}_p$ of* (5.0.3) *for $\gamma = 0$ depends continuously on the kernel matrix $\boldsymbol{K}_p$.*

Proof. Let $p_1, p_2 \in [0,1]$ be arbitrary. We write

$$\boldsymbol{K}_{p_1} =: \boldsymbol{K} \quad \text{and} \quad \boldsymbol{K}_{p_2} =: \boldsymbol{K} + \boldsymbol{S},$$

hence $\boldsymbol{S} = \boldsymbol{K}_{p_2} - \boldsymbol{K}_{p_1}$. From the continuous dependence of $\boldsymbol{K}_p$ on p we have

$$\lim_{p_2 \to p_1} \boldsymbol{S} = 0. \tag{5.1.1}$$

The first-order necessary conditions for the minimizers $\boldsymbol{n}_{p_1}$ and $\boldsymbol{n}_{p_2}$ of (5.0.3) for $p = p_1$ and $p = p_2$ are given by the relations

$$\boldsymbol{K}^T \boldsymbol{K} \boldsymbol{n}_{p_1} - \boldsymbol{K}^T \boldsymbol{r} + \boldsymbol{C}^T \boldsymbol{q}_{p_1} = 0 \tag{5.1.2}$$

$$\text{and} \quad (\boldsymbol{K}^T\boldsymbol{K} + \boldsymbol{K}^T\boldsymbol{S} + \boldsymbol{S}^T\boldsymbol{K} + \boldsymbol{S}^T\boldsymbol{S})\boldsymbol{n}_{p_2} - (\boldsymbol{K}+\boldsymbol{S})^T\boldsymbol{r} + \boldsymbol{C}^T\boldsymbol{q}_{p_2} = 0, \tag{5.1.3}$$

with vectors $\boldsymbol{q}_{p_1} \geq 0$, $\boldsymbol{q}_{p_2} \geq 0$. Subtracting (5.1.2) from (5.1.3) yields

$$\boldsymbol{K}^T\boldsymbol{K}(\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}) + (\boldsymbol{K}^T\boldsymbol{S} + \boldsymbol{S}^T\boldsymbol{K} + \boldsymbol{S}^T\boldsymbol{S})\boldsymbol{n}_{p_2} - \boldsymbol{S}^T\boldsymbol{r} + \boldsymbol{C}^T(\boldsymbol{q}_{p_2} - \boldsymbol{q}_{p_1}) = 0.$$

Forming the scalar product with $\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}$ yields

$$\begin{aligned} &\langle \boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}, \boldsymbol{K}^T\boldsymbol{K}(\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1})\rangle + \langle \boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}, (\boldsymbol{K}^T\boldsymbol{S} + \boldsymbol{S}^T\boldsymbol{K} + \boldsymbol{S}^T\boldsymbol{S})\boldsymbol{n}_{p_2}\rangle \\ &- \langle \boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}, \boldsymbol{S}^T\boldsymbol{r}\rangle + \langle \boldsymbol{C}(\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}), \boldsymbol{q}_{p_2} - \boldsymbol{q}_{p_1}\rangle = 0. \end{aligned}$$

With (5.1.1) we obtain in the limit $p_2 \to p_1$

$$\lim_{p_2 \to p_1} \Big(\langle \boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}, \boldsymbol{K}^T \boldsymbol{K}(\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}) \rangle + \langle \boldsymbol{C}(\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}), \boldsymbol{q}_{p_2} - \boldsymbol{q}_{p_1} \rangle \Big) = 0.$$

A calculation as in the proof of Lemma 2.3.1 shows

$$\langle \boldsymbol{C}(\boldsymbol{n}_{p_2} - \boldsymbol{n}_{p_1}), \boldsymbol{q}_{p_2} - \boldsymbol{q}_{p_1} \rangle \geq 0,$$

which finally implies

$$\lim_{p_2 \to p_1} \boldsymbol{n}_{p_2} = \boldsymbol{n}_{p_1}.$$

□

From the last proposition we directly obtain an existence result for an optimal p.

Corollary 5.1.2. *For $\gamma = 0$ the residual $\|\boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{r}\|_2$ of the minimizer of* (5.0.3) *depends continuously on p, so there exists a $p \in [0,1]$ for which it attains its minimal value.* □

Our next step is to find a condition for uniqueness of this minimizer for $\gamma = 0$.

Proposition 5.1.3. *Let $\gamma = 0$ and $p \in [0,1]$ be such that $\|\boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{r}\|_2^2$ minimizes all Tikhonov functionals in* (5.0.3) *over the parameter range* $[0,1]$*. Lets $s \in [0,1]$, $s \neq p$, be arbitrary. Then if*

$$\langle \boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{K}_s \boldsymbol{n}_s, \boldsymbol{r} \rangle \neq 0 \tag{5.1.4}$$

holds, the minimizing parameter p is unique.

Proof. The necessary conditions for $\boldsymbol{n}_p$ and $\boldsymbol{n}_s$ to be a minimizer of (5.0.3) are given by

$$\boldsymbol{K}_p^T \boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{K}_p^T \boldsymbol{r} + \boldsymbol{C}^T \boldsymbol{q}_p = 0 \tag{5.1.5}$$

$$\text{and} \quad \boldsymbol{K}_s^T \boldsymbol{K}_s \boldsymbol{n}_s - \boldsymbol{K}_s^T \boldsymbol{r} + \boldsymbol{C}^T \boldsymbol{q}_s = 0 \tag{5.1.6}$$

with vectors $\boldsymbol{q}_p \geq 0$, $\boldsymbol{q}_s \geq 0$. Assume

$$\|\boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{r}\|_2^2 = \|\boldsymbol{K}_s \boldsymbol{n}_s - \boldsymbol{r}\|_2^2,$$

which is equivalent to

$$\langle \boldsymbol{n}_p, \boldsymbol{K}_p^T \boldsymbol{K}_p \boldsymbol{n}_p \rangle - \langle \boldsymbol{n}_s, \boldsymbol{K}_s^T \boldsymbol{K}_s \boldsymbol{n}_s \rangle = 2 \langle \boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{K}_s \boldsymbol{n}_s, \boldsymbol{r} \rangle. \tag{5.1.7}$$

We form the scalar products of (5.1.5) with $\boldsymbol{n}_p$ and of (5.1.6) with $\boldsymbol{n}_s$. Then forming the difference of the resulting equations gives

$$\langle \boldsymbol{n}_p, \boldsymbol{K}_p^T \boldsymbol{K}_p \boldsymbol{n}_p \rangle - \langle \boldsymbol{n}_s, \boldsymbol{K}_s^T \boldsymbol{K}_s \boldsymbol{n}_s \rangle - \langle \boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{K}_s \boldsymbol{n}_s, \boldsymbol{r} \rangle + \langle \boldsymbol{C} \boldsymbol{n}_p, \boldsymbol{q}_p \rangle - \langle \boldsymbol{C} \boldsymbol{n}_s, \boldsymbol{q}_s \rangle = 0.$$

Inserting (5.1.7) yields

$$\langle \boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{K}_s \boldsymbol{n}_s, \boldsymbol{r} \rangle + \langle \boldsymbol{C} \boldsymbol{n}_p, \boldsymbol{q}_p \rangle - \langle \boldsymbol{C} \boldsymbol{n}_s, \boldsymbol{q}_s \rangle = 0.$$

From (2.3.5) with $\boldsymbol{b} = 0$ we conclude $\langle \boldsymbol{C} \boldsymbol{n}_p, \boldsymbol{q}_p \rangle = 0$ and analogously $\langle \boldsymbol{C} \boldsymbol{n}_s, \boldsymbol{q}_s \rangle = 0$. But then

$$\langle \boldsymbol{K}_p \boldsymbol{n}_p - \boldsymbol{K}_s \boldsymbol{n}_s, \boldsymbol{r} \rangle = 0,$$

which contradicts (5.1.4). Thus if (5.1.4) holds, the minimizing parameter p must be unique.

□

Condition (5.1.4) demands that the kernel matrices $\boldsymbol{K}_s$ are sufficiently different to $\boldsymbol{K}_p$ so that we get distinguishable residuals of the unregularized solutions. If an $s \in [0,1]$ happens to exist with $\boldsymbol{K}_s = \boldsymbol{K}_p$, condition (5.1.4) cannot be fulfilled. Unfortunately we are not currently able to check this condition a priori.

We conclude this section with investigating how the unregularized residuals behave for moderate noise levels. In the following the superscript δ marks solutions of (5.0.3) for a data vector $\boldsymbol{r}$ contaminated with noise.

Proposition 5.1.4. *We assume that condition* (5.1.4) *holds for any noise vector satisfying* $0 \leq \|\boldsymbol{\delta}\|_2 \leq \delta$. *Let* $\boldsymbol{n}_t$ *be the minimizer for the true noise-free model, i.e. the parameter* $t \in [0,1]$ *yields the minimal residual* $\|\boldsymbol{K}_t\boldsymbol{n}_t - \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_{true}\|_2^2$ *over the whole parameter interval* $[0,1]$. *Let* $p = p(\boldsymbol{\delta}) \in [0,1]$ *be the parameter yielding the minimal unregularized residual* $\|\boldsymbol{K}_p\boldsymbol{n}_p^\delta - \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{e}_{true} + \boldsymbol{\delta})\|_2^2$ *for the noisy data vector* $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{e}_{true} + \boldsymbol{\delta})$. *Then holds* $\lim_{\|\boldsymbol{\delta}\|_2 \to 0} p(\boldsymbol{\delta}) = t$.

Proof. To shorten notation we write $\boldsymbol{r}_{true} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_{true}$ and $\boldsymbol{\rho} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}$. Let $\boldsymbol{n}_t^\delta$ be the minimizer for the parameter t and the noisy data vector $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{e}_{true} + \boldsymbol{\delta})$, i.e.

$$\boldsymbol{n}_t^\delta = \underset{\boldsymbol{n}\in\mathbb{R}^N}{\operatorname{argmin}} \ \tfrac{1}{2}\|\boldsymbol{K}_t\boldsymbol{n} - (\boldsymbol{r}_{true} + \boldsymbol{\rho})\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \leq \boldsymbol{b}.$$

Then we have the first order necessary conditions

$$\boldsymbol{K}_t^T\boldsymbol{K}_t\boldsymbol{n}_t - \boldsymbol{K}_t^T\boldsymbol{r}_{true} + \boldsymbol{C}\boldsymbol{q}_t = 0 \tag{5.1.8}$$

$$\boldsymbol{K}_t^T\boldsymbol{K}_t\boldsymbol{n}_t^\delta - \boldsymbol{K}_t^T(\boldsymbol{r}_{true} + \boldsymbol{\rho}) + \boldsymbol{C}\boldsymbol{q}_t^\delta = 0, \tag{5.1.9}$$

with vectors $\boldsymbol{q}_t^\delta \geq 0$, $\boldsymbol{q}_t \geq 0$. Subtracting (5.1.9) from (5.1.8) and scalar multiplying the result with $\boldsymbol{n}_t - \boldsymbol{n}_t^\delta$ gives

$$\langle \boldsymbol{n}_t - \boldsymbol{n}_t^\delta, \boldsymbol{K}_t^T\boldsymbol{K}_t(\boldsymbol{n}_t - \boldsymbol{n}_t^\delta)\rangle + \langle \boldsymbol{n}_t - \boldsymbol{n}_t^\delta, \boldsymbol{K}_t^T\boldsymbol{\rho}\rangle + \langle \boldsymbol{C}(\boldsymbol{n}_t - \boldsymbol{n}_t^\delta), \boldsymbol{q}_t - \boldsymbol{q}_t^\delta\rangle = 0.$$

As in the proof of Lemma 2.3.1 this establishes

$$\begin{aligned}\|\boldsymbol{K}_t(\boldsymbol{n}_t - \boldsymbol{n}_t^\delta)\|_2^2 &\leq \langle \boldsymbol{n}_t^\delta - \boldsymbol{n}_t, \boldsymbol{K}_t^T\boldsymbol{\rho}\rangle \\ &\leq \|\boldsymbol{K}_t(\boldsymbol{n}_t - \boldsymbol{n}_t^\delta)\|_2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2,\end{aligned}$$

which gives

$$\|\boldsymbol{K}_t(\boldsymbol{n}_t - \boldsymbol{n}_t^\delta)\|_2 = \mathcal{O}(\|\boldsymbol{\delta}\|_2).$$

Now since the parameter p minimizes the residuals for the noisy vector $\boldsymbol{r}_{true} + \boldsymbol{\rho}$ we can estimate

$$\begin{aligned}\|\boldsymbol{K}_p\boldsymbol{n}_p^\delta - (\boldsymbol{r}_{true} + \boldsymbol{\rho})\|_2 &\leq \|\boldsymbol{K}_t\boldsymbol{n}_t^\delta - (\boldsymbol{r}_{true} + \boldsymbol{\rho})\|_2 \\ &\leq \|\boldsymbol{K}_t\boldsymbol{n}_t - \boldsymbol{r}_{true}\|_2 + \|\boldsymbol{K}_t(\boldsymbol{n}_t^\delta - \boldsymbol{n}_t)\|_2 + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2.\end{aligned}$$

With the previous finding we see that the upper bound in the last inequality converges to the residual $\|\boldsymbol{K}_t\boldsymbol{n}_t - \boldsymbol{r}_{true}\|_2$ for $\|\boldsymbol{\delta}\|_2 \to 0$. Thus we obtain in the limit

$$\lim_{\|\boldsymbol{\delta}\|_2 \to 0} \|\boldsymbol{K}_p\boldsymbol{n}_p^\delta - (\boldsymbol{r}_{true} + \boldsymbol{\rho})\|_2 \leq \|\boldsymbol{K}_t\boldsymbol{n}_t - \boldsymbol{r}_{true}\|_2.$$

By definition of t we have

$$\|\boldsymbol{K}_t\boldsymbol{n}_t - \boldsymbol{r}_{true}\|_2 \leq \lim_{\|\boldsymbol{\delta}\|_2 \to 0} \|\boldsymbol{K}_p\boldsymbol{n}_p^\delta - (\boldsymbol{r}_{true} + \boldsymbol{\rho})\|_2,$$

so condition (5.1.4) finally implies $\lim_{\|\boldsymbol{\delta}\|_2 \to 0} p(\boldsymbol{\delta}) = t$. □

5.2 Convergence Analysis

In this section we show that the regularized solutions from the retrieved aerosol fraction converge to the true solution from the true fraction as the noise level approaches zero. This means that we generalize Theorem 2.5.4 to the case where the underlying true linear operator must be identified from a known set of possible operators.

Theorem 5.2.1. *Under Assumption 7.2.1, if condition 5.1.4 is satisfied for all noise vectors $\boldsymbol{\delta}$ of random variables, then we have for any $\alpha(\delta)$ with the properties $\lim_{\delta\to 0}\alpha(\delta) = 0$ and $\lim_{\delta\to 0}\frac{\delta^2}{\alpha(\delta)} = 0$ that $\lim_{\delta\to 0}\mathbb{E}\big(\|\boldsymbol{n}_p^{\delta,\alpha(\delta)} - \boldsymbol{n}_t\|_2\big) = 0$. Here $\boldsymbol{n}_p^{\delta,\alpha(\delta)}$ is the regularized solution for the retrieved fraction parameter $p = p(\boldsymbol{\delta})$.*

Proof. We again use the notations $\boldsymbol{r}_{true} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_{true}$ and $\boldsymbol{\rho} := \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}$. Let $p = p(\boldsymbol{\delta})$ be the fraction parameter retrieved by minimizing the unregularized residuals. We write

$$\begin{aligned}
\boldsymbol{n}_p^{\delta,\alpha} &:= \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{K}_p\boldsymbol{n} - (\boldsymbol{r}_{true} + \boldsymbol{\rho})\|_2^2 + \tfrac{1}{2}\alpha\|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \le \boldsymbol{b},\\
\boldsymbol{n}_p^{\alpha} &:= \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{K}_p\boldsymbol{n} - \boldsymbol{r}_{true}\|_2^2 + \tfrac{1}{2}\alpha\|\boldsymbol{n}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \le \boldsymbol{b},\\
\boldsymbol{n}_p &:= \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{K}_p\boldsymbol{n} - \boldsymbol{r}_{true}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \le \boldsymbol{b},\\
\boldsymbol{n}_t &:= \underset{\boldsymbol{n}\,\in\,\mathbb{R}^N}{\operatorname{argmin}}\ \tfrac{1}{2}\|\boldsymbol{K}_t\boldsymbol{n} - \boldsymbol{r}_{true}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{C}\boldsymbol{n} \le \boldsymbol{b}.
\end{aligned}$$

Then we have the estimate

$$\mathbb{E}\big(\|\boldsymbol{n}_p^{\delta,\alpha} - \boldsymbol{n}_t\|_2\big) \le \mathbb{E}\big(\|\boldsymbol{n}_p^{\delta,\alpha} - \boldsymbol{n}_p^{\alpha}\|_2\big) + \mathbb{E}\big(\|\boldsymbol{n}_p^{\alpha} - \boldsymbol{n}_p\|_2\big) + \mathbb{E}(\|\boldsymbol{n}_p - \boldsymbol{n}_t\|_2).$$

For the first term in the upper bound, the estimate

$$\mathbb{E}\big(\|\boldsymbol{n}_p^{\delta,\alpha} - \boldsymbol{n}_p^{\alpha}\|_2\big) = \frac{\mathcal{O}(\delta)}{\alpha^{\frac{1}{2}}}$$

follows from Proposition 2.5.3. For the second term, Proposition 2.5.2 gives

$$\lim_{\alpha\to 0}\ \mathbb{E}\big(\|\boldsymbol{n}_p^{\alpha} - \boldsymbol{n}_p\|_2\big) = 0.$$

Finally, for the third term from Proposition 5.1.4 follows

$$\lim_{\delta\to 0}\ \mathbb{E}(\|\boldsymbol{n}_p - \boldsymbol{n}_t\|_2) = 0.$$

This altogether proves our claim. □

5.3 The Retrieval method

5.3.1 Model Generation

Proposition 5.1.3 motivates us to use the unregularized residuals as model generation criterion, which means that we determine those parameters s, where they are small. In presence of moderate measurement noise in $\boldsymbol{e}$ these parameters lie in the vicinity of the unique true parameter p as was shown in the proof of Proposition 5.1.4. In the following we discuss the model generation algorithm extended for two-component aerosols. As in Section 2.6 we compute collocation grids with $N_1 < ... < N_m$ points and select a grid of Morozov safety factors $\tau_1 < ... < \tau_s$. Furthermore the refractive indices $k_1(l_1) + in_1(l_1), ..., k_1(l_{N_l}) + in_1(l_{N_l})$ and $k_2(l_1) + in_2(l_1), ..., k_2(l_{N_l}) + in_2(l_{N_l})$ of two pure aerosol components depending on wavelengths $l_1, ..., l_{N_l}$ are given.

Algorithm 5 Model Generation for Two-Component Aerosols

```
$MaxDisc = 1$
$SolutionSets = \{\}$
$ApproxSets = \{\}$
$PriorSets = \{\}$
$MixRatioSets = \{\}$
$TauSets = \{\}$
$DiscCntr = 0$
estimate $\sigma_1^2$, ..., $\sigma_{N_l}^2$ from the sample means approximating the standard deviations of $e_1$, ..., $e_{N_l}$.
$\delta^2 := \max\{\sigma_1^2, ..., \sigma_{N_l}^2\}$
$\mathbf{\Sigma} := \delta^{-2} \cdot \mathrm{diag}(\sigma_1^2, ..., \sigma_{N_l}^2)$
$p_i = \frac{i-1}{N_{frac}-1}$, $i = 1, ..., N_{frac}$
$N_{frac} = 201$
$N_{mean} = 5$
$I_{min} = \{\}$
for $i = 1$ to $N_{frac}$ do
    for $j = 1$ to $N_l$ do
        compute $k_{tot}(l_j) + in_{tot}(l_j)$ from $k_1(l_j) + in_1(l_j)$ and $k_2(l_j) + in_2(l_j)$ using (5.0.1) with $f_1 = p_i$ and $f_2 = 1 - p_i$
    end for
    for $k = 1$ to $m$ do
        compute kernel matrix $\boldsymbol{K}_{ik}$ for $p_i$ and the collocation grid with $N_k$ points using $k_{tot}(l_1) + in_{tot}(l_1), ..., k_{tot}(l_{N_l}) + in_{tot}(l_{N_l})$
    end for
end for
for $k = 1$ to $m$ do
    $S_k = \{\}$
    $A_k = \{\}$
    $P_k = \{\}$
    $M_k = \{\}$
    $T_k = \{\}$
    $R = \{\}$
    $RM_{min} = \infty$
    for $i = 1$ to $N_{frac}$ do
        $\boldsymbol{n}_{lsqnng} = \underset{\boldsymbol{n} \in \mathbb{R}^{N_k}}{\mathrm{argmin}} \frac{1}{2} \|\mathbf{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_{ik}\boldsymbol{n} - \boldsymbol{e}_{real})\|_2^2$ s.t. $\boldsymbol{n} \geq 0$
        $R = R \cup \left\{\|\mathbf{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_{ik}\boldsymbol{n}_{lsqnng} - \boldsymbol{e}_{real})\|_2^2\right\}$
    end for
```

```
35:     for $i = 1$ to $N_{frac} - N_{mean} + 1$ do
36:         $RM = \text{mean}(R(i), R(i+1), ..., R(i + N_{mean} - 1))$
37:         if $RM < RM_{min}$ then
38:             $RM_{min} = RM$
39:             $t_{min} = \{i, i+1, ..., i + N_{mean} - 1\}$
40:         end if
41:     end for
42:     $t_{cur} = \{t_{min}(1), t_{min}(3), t_{min}(5)\}$
43:     $N_{cur} = |t_{cur}|$
44:     for $i = 1$ to $N_{cur}$ do
45:         for $j = 0$ to $s$ do
46:             if $R(t_{cur}(i)) < \tau_j N_l \delta^2 \ \wedge \ \tau_j N_l \delta^2 < \|\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{e}_{real}\|_2^2$ then
47:                 compute $\gamma_{kij}$ such that
48:                 $\boldsymbol{n}_{trial} = \underset{\boldsymbol{n} \in \mathbb{R}^{N_i}}{\text{argmin}} \ \frac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_{t_{cur}(i),k}\boldsymbol{n} - \boldsymbol{e}_{real})\|_2^2$
                            $+ \frac{1}{2}\gamma_{kij}\boldsymbol{n}^T \boldsymbol{R}_k \boldsymbol{n} \ \text{ s.t. } \boldsymbol{n} \geq 0$
49:                 with $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{K}_{t_{cur}(i),k}\boldsymbol{n}_{trial} - \boldsymbol{e}_{real})\|_2^2 = \tau_j N_l \delta^2$
50:             end if
51:             if $\boldsymbol{n}_{trial}$ exists then
52:                 $S_k = S_k \cup \{\boldsymbol{n}_{trial}\}$
53:                 $A_k = A_k \cup \{\boldsymbol{K}_{t_{cur}(i),k}\}$
54:                 $P_k = P_k \cup \{\gamma_{kij}\boldsymbol{R}_k\}$
55:                 $M_k = M_k \cup \{p_{t_{cur}(i)}\}$
56:                 $T_k = T_k \cup \{\tau_j\}$
57:             end if
58:         end for
59:     end for
60:     if $S_k$, $A_k$, $P_k$, $M_k$ and $T_k$ not empty then
61:         $SolutionSets = SolutionSets \cup \{S_k\}$
62:         $ApproxSets = ApproxSets \cup \{A_k\}$
63:         $PriorSets = PriorSets \cup \{P_k\}$
64:         $MixRatioSets = MixRatioSets \cup \{M_k\}$
65:         $TauSets = TauSets \cup \{T_k\}$
66:         $DiscCntr = DiscCntr + 1$
67:     end if
68:     if $DiscCntr == MaxDisc$ then
69:         break
70:     end if
71: end for
```

In line 11 the aerosol fraction parameter interval $[0, 1]$ is approximated with a linearly spaced grid. For each discrete aerosol fraction p_i the approximation $\boldsymbol{K}_{ik}$ to the linear operator K_{p_i} is computed in lines 15 to 22 for all model space orders N_k.

In line 23 the main loop for the model generation begins. Note that we first run through all model orders from 1 to m beginning with the coarsest models before we iterate through all aerosol fractions p_i. This means that we perform the residual-based search strategy motivated in Proposition 5.1.3 for each model space separately, where we start with the coarsest model and refine it if necessary.

In lines 31-41 the residuals of the unregularized reconstructions are calculated, and a scan to find the minimal mean of N_{mean} solutions corresponding to successive parameters $p_i, p_{i+1}, ..., p_{i+N_{mean}-1}$ is performed. A subset of the indices $i, i+1, ..., i+N_{mean}-1$ corresponding to the residuals with minimal mean is selected in line 42. By filtering out some of the models corresponding to the parameters $p_i, p_{i+1}, ..., p_{i+N_{mean}-1}$ with small residuals we ensure that the models to be compared are not too similar. The selected indices are used for the actual model generation in lines 44-59. Here we loop through all preselected Morozov safety parameters $\tau_1, ..., \tau_s$ and we propose with them the possible residual values $\tau_j N_l$ for the discrepancy principle. In line 46 it is checked if the discrepancy principle is applicable.

If the model generation step is successful, the obtained reconstructions accompanied by their kernel and regularization matrices and their aerosol fraction and residual parameters are stored in the containers S_k, A_k, P_k, M_k and T_k in lines 51-57.

Finally if the model generation is successful for $MaxDisc$ model spaces, the model generation loop is terminated in line 69.

5.3.2 Model Selection

Not only the model generation procedure has to be generalized to the case of a two-component aerosol, but also the model-selection framework presented in Section 2.6.2 needs to be generalized as well. Here we are not just comparing models with different model spaces but also with different underlying operators K_p. Thus prior probabilities are also needed for the parameters p which determine the linear operators—or more precisely their approximations—to be compared. Let k label the model dimensions N_k, Let i run through the indices for the aerosol-fraction parameters p_i, where i depends on k, and let j run through all Morozov safety parameters τ_j used for the model generation, where j depends on k and i. Then we can compute the model posterior probabilities by

$$p(N_k, \boldsymbol{K}_{ik}, \gamma_{kij}|\boldsymbol{e}) = \frac{p(\boldsymbol{e}|N_k, \boldsymbol{K}_{ik}, \gamma_{kij})p(N_k, \boldsymbol{K}_{ik}, \gamma_{kij})}{\sum_u \sum_{v(u)} \sum_{w(u,v)} p(\boldsymbol{e}|N_u, \boldsymbol{K}_{vu}, \gamma_{uvw})p(N_u, \boldsymbol{K}_{vu}, \gamma_{uvw}))} \tag{5.3.1}$$

We assume that $p(\boldsymbol{K}_{vu})$ and $p(N_u, \gamma_{uvw})$ are independent and thus

$$p(N_u, \boldsymbol{K}_{vu}, \gamma_{uvw}) = p(N_u, \gamma_{uvw})p(\boldsymbol{K}_{vu}).$$

We select $p(\boldsymbol{K}_{vu})$ to be uniform and adopt $p(N_u, \gamma_{uvw})$ from Section 2.6.2. This leads to

$$p(N_u, \boldsymbol{K}_{vu}, \gamma_{uvw}) = \frac{1}{N_{total}}, \tag{5.3.2}$$

where N_{total} is the total number of triplets $\big(u,\ v(u),\ w(u,v)\big)$.

Then the model-selection algorithm proceeds in the same way as Algorithm 2, so we do not restate here. The differences to Section 2.6.2 are that we have already set $MaxDisc = 1$ in the model generation step and that the single container A_1 stores kernel matrices approximating different operators K_{p_i}. While in principle the algorithm could continue to compare different discretizations, this only lead to worse results in our simulations. Therefore, once the algorithm finds a discretization level for which reconstructions are at all possible for any of the safety factors, we stop the refinement and simply focus on the problem of identifying the volume fraction.

Chapter 6

Numerical Results

6.1 Numerical Study

We conducted a numerical study of our inversion algorithm with almost the same settings as the last section but extended for the retrieval of volume fractions of a two-component aerosol. We used the same wavelength grid as in Sections 3.1 and simulated the same model size distributions as in Section 3.2.2. We selected air as ambient medium as well. We extended the grid of Morozov safety parameters to

$$\tau_1 = 0.5, \tau_2 = 0.6, ..., \tau_{16} = 2.0.$$

If when running through all model spaces none of these safety factors yielded a solution, we performed in this extreme case an another run of the model generation step using a second grid of safety factors given by

$$\tau_1 = 2.5, \tau_2 = 3.0, ..., \tau_6 = 5.0.$$

This time we did not just simulate an original aerosol consisting purely of H_2O but instead generated with (5.0.1) refractive indices of H_2O and CsI mixtures for the scattering particles. Here the volume fractions of H_2O ranged through a set of preselected percentages, namely

$$0,\ 11,\ 22,\ 33,\ 44,\ 56,\ 67,\ 78,\ 89\ \text{ and }\ 100.$$

For each of the 100 parameters for the log-normal, RRSB or Hedrih distributions we also now have the above 10 fractions. This results in a total of 1000 cases to simulate.

As preparation to run Algorithm 5 we computed the kernel matrices depending on the water volume fraction parameter $p \in [0, 1]$ for $p_i = \frac{i-1}{100}$, $i = 1, ..., 101$ and interpolated each kernel matrix entry with a cubic spline on a linearly spaced grid with 201 points covering $[0, 1]$ to increase further the resolution in p. Thus we have $N_{frac} = 201$ in Algorithm 5.

Again we simulated the same noise levels as in Section 3.1. Thus the noisy measurement data vector $\boldsymbol{e}$ was modeled with

$$(\boldsymbol{e})_i = e(l_i) + \delta_i \quad \text{with } \delta_i \sim \mathcal{N}(0, (0.05 \cdot e(l_i))^2), \quad i = 1, ..., N_l,$$

$$(\boldsymbol{e})_i = e(l_i) + \delta_i \quad \text{with } \delta_i \sim \mathcal{N}(0, (0.15 \cdot e(l_i))^2), \quad i = 1, ..., N_l$$

and

$$(\boldsymbol{e})_i = e(l_i) + \delta_i \quad \text{with } \delta_i \sim \mathcal{N}(0, (0.3 \cdot e(l_i))^2), \quad i = 1, ..., N_l.$$

respectively.

To investigate the quality of the reconstructions in each simulation run, we computed their L^2-errors relative to the original size distribution. We list them separately for each of the ten original water fractions. We proceed this way for all of our simulation results.

Furthermore we determined the deviations of the reconstructed water volume fractions from the original ones, e.g. when the original fraction was 22% and $p_{recon} \in [0, 1]$ the retrieved fraction parameter, we calculated the deviation by $|22 - 100 \cdot p_{recon}|\%$. This showed us how well one can investigate the unknown two-component aerosol only from FASP measurements using our extended inversion algorithm.

We also report how often the inversions failed. There were two main reasons for inversion failures: the first when the relative L^2 was greater than or equal to 100%, the second when the fraction deviation was greater than or equal to 50%. In both cases the reconstruction cannot give any reasonable information about the true size distribution and the true scattering material anymore. Note that in our simulations we returned by default $\boldsymbol{n} \equiv 0$ and $p_{recon} = 0.5$ when no reconstruction could be found in any of the model spaces. For brevity we only list those original fractions where inversion failures occurred

Finally we list the average and worst case inversion run times over all 1000 simulations for all three simulation runs.

6.1.1 Average L^2-Errors

Results for 5% Noise

Log-Normal Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	33.4495	33.5512	33.8810
11%	29.9288	30.0234	31.8153
22%	28.8686	29.0432	31.0142
33%	24.8269	26.1235	27.7701
44%	22.5902	24.2364	26.1419
56%	21.0371	21.8765	24.5631
67%	19.1780	19.9413	22.7730
78%	19.0107	19.1835	21.0428
89%	18.6772	19.2620	20.0989
100%	18.0467	18.9292	18.9216

RRSB Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	37.6093	36.2611	35.7504
11%	31.7431	30.5070	30.0359
22%	30.3894	29.1295	29.0800
33%	28.7004	27.2851	26.9338
44%	24.2003	23.6463	24.9326
56%	21.4835	21.1434	21.2237
67%	19.7283	19.5032	18.7670
78%	17.0828	16.6510	16.5460
89%	14.2999	14.1295	14.1319
100%	11.6901	11.4887	11.5191

Hedrih Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	16.3524	16.7309	17.1452
11%	16.9073	16.8683	17.1026
22%	15.9820	15.5398	16.1297
33%	13.7607	13.6466	13.8722
44%	15.7127	14.6788	14.6119
56%	14.9451	14.7220	14.6970
67%	14.7707	14.5446	14.2365
78%	16.8178	15.9551	15.6820
89%	13.6688	13.4196	13.4240
100%	11.4425	11.3837	11.0599

Results for 15% Noise

Log-Normal Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	41.0230	42.0280	42.8570
11%	38.7119	40.3354	40.7482
22%	35.4868	37.0136	37.6257
33%	38.2937	39.1035	39.6033
44%	35.1116	35.7339	36.3366
56%	35.3125	36.2310	36.8520
67%	31.4433	32.8912	34.6204
78%	34.1495	35.4525	36.7260
89%	28.2679	29.9144	31.6657
100%	24.9730	26.8085	27.5983

RRSB Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	46.9113	47.2809	47.6162
11%	44.5891	45.7270	46.0245
22%	42.8317	43.0837	41.9885
33%	40.0047	40.7705	40.6499
44%	36.8078	37.8087	39.0816
56%	35.1608	34.9177	34.9506
67%	34.2437	34.6667	34.6924
78%	30.5743	29.9018	30.2997
89%	26.6096	25.5681	25.6317
100%	25.1971	25.4078	26.8239

Hedrih Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	38.2540	41.0444	41.7418
11%	37.2035	39.2980	39.8581
22%	37.5875	39.2181	39.3586
33%	37.9361	39.3276	40.1512
44%	37.1656	38.9567	40.1287
56%	35.6547	37.0807	38.1747
67%	36.7073	38.0401	39.0624
78%	37.3575	38.5249	39.4832
89%	32.9935	34.3648	35.3986
100%	21.0929	22.2989	23.4942

Results for 30% **Noise**

Log-Normal Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	54.4440	56.1149	56.5318
11%	53.4176	55.1921	55.4526
22%	50.2123	51.9879	52.9042
33%	48.5834	50.5190	51.9092
44%	48.2961	49.2341	51.0833
56%	49.6869	50.4403	52.3943
67%	48.7960	49.8042	51.4531
78%	45.3692	46.7035	48.4497
89%	40.7916	42.5978	44.2974
100%	37.8053	40.2252	40.6318

RRSB Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	60.9539	60.1571	59.4861
11%	55.8505	55.8186	55.8801
22%	56.6660	56.7041	57.2950
33%	51.7956	51.7272	52.4263
44%	55.0476	55.4631	54.7889
56%	53.3087	53.4469	52.7929
67%	56.3399	56.0415	55.7153
78%	43.7841	44.3589	44.7646
89%	35.7766	35.2697	35.2719
100%	33.4715	34.4714	35.2654

Hedrih Distribution			
original water volume percent	average L^2-errors (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	61.2199	63.0944	64.9315
11%	62.1051	63.5453	65.3789
22%	60.2048	62.0754	63.7821
33%	60.6939	62.3668	64.2677
44%	58.9614	60.5740	62.9731
56%	59.2670	60.5463	63.1971
67%	60.0236	60.9762	62.9238
78%	55.6534	56.7551	58.9474
89%	48.4864	49.9086	52.0646
100%	35.5582	37.4365	39.5205

6.1.2 Average Water Fraction Deviation

Results for 5% Noise

Log-Normal Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	11.0750	10.9550	11.0950
11%	7.6500	7.3200	7.5200
22%	6.3450	6.2750	6.4250
33%	4.4000	4.4000	4.4500
44%	3.5750	3.8350	3.8150
56%	3.2700	3.1100	3.1500
67%	2.5050	2.4550	2.4750
78%	2.3850	2.3050	2.1850
89%	2.0100	2.0400	1.7700
100%	1.2550	1.0150	0.7150

RRSB Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	6.7000	6.7000	6.6600
11%	5.2200	5.1700	5.0900
22%	4.6400	4.5400	4.5400
33%	3.7100	3.7000	3.5300
44%	3.6200	3.5400	3.5400
56%	3.2050	3.1750	3.1550
67%	2.4650	2.4150	2.3350
78%	1.7750	1.8350	1.8750
89%	1.3250	1.3250	1.3050
100%	0.4650	0.3850	0.3450

Hedrih Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	4.8800	4.9200	5.0600
11%	6.3850	6.3550	6.2450
22%	4.1300	4.1900	4.3300
33%	4.3650	4.4750	4.3750
44%	3.3100	3.4500	3.4400
56%	3.0450	3.0150	2.9950
67%	1.8100	1.9300	1.8000
78%	2.7250	2.4950	2.4350
89%	1.9350	1.8850	1.7350
100%	0.7250	0.5450	0.4050

Results for 15% **Noise**

Log-Normal Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	22.7300	22.8500	22.8500
11%	14.9650	15.4850	15.6050
22%	11.4200	11.7200	11.7800
33%	11.0250	10.9550	11.0150
44%	9.4150	9.1050	9.3650
56%	8.4350	8.5450	8.3250
67%	6.6650	6.6950	6.7750
78%	7.8100	7.7600	7.7300
89%	4.6300	4.9900	4.9900
100%	1.8050	1.6450	1.2650

RRSB Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	13.0650	12.6850	12.7250
11%	14.9600	15.0300	14.9300
22%	10.7500	10.8400	10.7700
33%	10.8900	10.9000	10.9900
44%	10.3200	10.2500	10.2600
56%	6.9550	6.8350	6.7950
67%	5.8300	5.6800	5.8000
78%	4.4500	4.3800	4.4000
89%	3.1850	3.1450	3.1450
100%	1.9350	1.9150	2.0750

Hedrih Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	19.4000	20.0200	20.2000
11%	15.1000	15.5200	15.5300
22%	15.9050	16.2850	16.3650
33%	13.4650	13.9250	14.0850
44%	11.4050	11.7850	12.0450
56%	8.8800	9.2600	9.4500
67%	9.3000	9.4000	9.6000
78%	8.3150	8.5550	8.6850
89%	4.6500	5.1300	5.3700
100%	1.1300	0.9300	0.3900

Results for 30% Noise

Log-Normal Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	33.3450	33.5650	33.4850
11%	26.8900	27.2100	27.1000
22%	22.2350	22.7650	23.0250
33%	18.3700	18.9100	19.1100
44%	17.7700	17.6600	17.9100
56%	17.0850	16.8650	17.2450
67%	14.6700	14.6800	14.7200
78%	11.5950	11.6450	11.8250
89%	6.9650	7.0850	7.3650
100%	2.2650	2.2250	1.7250

RRSB Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	22.6650	22.5850	22.4050
11%	22.3800	22.3900	22.5000
22%	19.5000	19.4100	19.4600
33%	17.3000	17.1100	17.1300
44%	18.0600	18.0100	17.9900
56%	13.3100	13.3100	13.3400
67%	11.6500	11.4500	11.3700
78%	8.1000	7.9800	7.8400
89%	5.3800	5.2600	5.2500
100%	3.0300	2.8900	2.7700

Hedrih Distribution			
original water volume percent	average water fraction deviation (%)		
	Tikhonov	min. first fin. diff.	Twomey
0%	41.2100	41.8100	41.6500
11%	39.4300	39.8100	39.7100
22%	32.4950	32.9750	32.8750
33%	29.2950	29.6150	29.7750
44%	24.3600	24.1600	24.8800
56%	21.8000	22.1300	22.5100
67%	18.3750	18.5250	19.0650
78%	13.7000	13.7400	14.0800
89%	6.8900	6.9700	7.3700
100%	1.0700	0.9900	0.4100

6.1.3 Average Model Space Dimensions

Results for 5% **Noise**

Log-Normal Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
7.5550	7.5550	7.5550

RRSB Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
11.3940	11.3940	11.3940

Hedrih Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
6.3330	6.3330	6.3330

Results for 15% Noise

Log-Normal Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
4.9840	4.9840	4.9840

RRSB Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
7.9840	7.9840	7.9840

Hedrih Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
4.5890	4.5890	4.5890

Results for 30% Noise

Log-Normal Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
3.8240	3.8240	3.8240

RRSB Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
6.5260	6.5260	6.5260

Hedrih Distribution		
average model space dimensions		
Tikhonov	min. first fin. diff.	Twomey
3.7000	3.7000	3.7000

6.1.4 Extreme Cases

When the deviation of the retrieved aerosol fraction from the true one exceeded 50% or the L^2-error between reconstruction and true solution was bigger than 100% we had to regard the reconstruction as failed. We now list when these failures occurred. There were no failed simulations with the Hedrih distribution.

6.1.5 Reconstruction Failures

Results for 5% Noise

Log-Normal Distribution			
original water volume percent	number of L^2-errors ≥ 100 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	1	1	1
11%	1	0	0
44%	1	0	0

RRSB Distribution			
original water volume percent	number of L^2-errors ≥ 100 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	5	6	6
11%	1	1	0
33%	1	1	1
44%	1	0	0
67%	1	1	1

Results for 15% Noise

Log-Normal Distribution			
original water volume percent	number of L^2-errors ≥ 100 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	1	1	1

RRSB Distribution			
original water volume percent	number of L^2-errors ≥ 100 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	9	8	7
11%	4	4	3
22%	3	3	1
33%	3	3	2
44%	2	2	2
56%	3	2	2
67%	1	1	1
78%	1	1	0
89%	1	0	0

Results for 30% Noise

Log-Normal Distribution			
original water volume percent	number of L^2-errors ≥ 100 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
89%	1	1	1
100%	2	2	2

RRSB Distribution			
original water volume percent	number of L^2-errors ≥ 100 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	12	12	12
11%	6	5	5
22%	7	7	6
33%	4	4	5
44%	10	11	9
56%	10	9	8
67%	12	13	12
78%	2	1	1
100%	2	2	2

6.1.6 Water-Fraction Retrieval Failures

Results for 5% Noise

Log-Normal Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	3	3	3
11%	3	3	3
22%	1	1	1

RRSB Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	1	1	1

Results for 15% Noise

Log-Normal Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	7	7	7
11%	1	1	1
22%	3	3	2
33%	2	2	2
44%	1	0	0

RRSB Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	7	7	7
11%	9	9	7
22%	5	5	5
33%	6	6	6
44%	1	1	1

Hedrih Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	7	7	7
11%	2	2	0

Results for 30% Noise

Log-Normal Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	18	18	18
11%	12	11	9
22%	5	5	5
33%	1	1	1
44%	2	2	2
56%	1	1	1

RRSB Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	17	17	17
11%	17	17	17
22%	14	14	14
33%	7	7	7
44%	5	4	4
56%	3	3	3
67%	2	2	2
78%	1	1	1

Hedrih Distribution			
original water volume percent	number of deviations ≥ 50 % (out of 100)		
	Tikhonov	min. first fin. diff.	Twomey
0%	34	34	34
11%	31	31	28
22%	3	4	4
33%	6	6	6
44%	6	4	4

6.1.7 Average and Worst-Case Run Times

Results for 5% Noise

	Log-Normal Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	1.9760	1.9942	2.0068
worst case	7.7510	7.2250	7.1985

	RRSB Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	2.9324	2.9625	2.9790
worst case	33.3764	33.3267	34.9001

	Hedrih Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	1.4476	1.4508	1.4606
worst case	4.7932	4.8286	4.8890

Results for 15% Noise

	Log-Normal Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	1.5145	1.5206	1.5166
worst case	6.2287	6.4040	6.4904

	RRSB Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	3.3324	3.3626	3.3634
worst case	27.5733	28.3437	27.9566

	Hedrih Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	2.9680	2.9574	2.9615
worst case	8.1061	8.0749	8.2823

Results for 30% Noise

	Log-Normal Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	1.4768	1.4764	1.4714
worst case	5.2418	5.1896	5.1899

	RRSB Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	1.9579	1.9837	1.9885
worst case	19.8418	20.5086	20.6216

	Hedrih Distribution		
	run times (s)		
	Tikhonov	min. first fin. diff.	Twomey
average	1.1961	1.1893	1.1864
worst case	3.7253	3.6821	3.8738

6.2 Conclusion

A common trend in the results is that the average L^2-error decreases with increasing original water volume fraction. This was observed for all noise levels. For all noise levels the average L^2-errors were for 0% water 2 to 3 times as big as for 100% water.. This behavior can also be seen in the numbers of reconstruction failures, which predominantly occurred for small water percentages.

The water volume fractions deviations behaved in a similar way. They decreased for increasing original fractions, which means that the quality of the water fraction retrieval was improving towards higher original fractions. Water fraction retrieval failures predominantly occurred for samll water percentages and their number rose for higher noise levels. For Hedrih distributions more than 30% of the inversions were affected by a water fraction retrieval failure for 30% noise and for original water percentages of 0 and 11%.

Again the differences in the deviations depending on the priors were only marginal.

The worst case run times succeeded our thirty-seconds limit only for RRSB distributions. Even in the extreme cases they stayed below 35 seconds, which is still acceptable.

We can conclude that with the settings made in previous section the analysis of two-component aerosols is possible satisfying our demands on run time and accuracy. The results for 5% noise are of comparable quality than the results for 30% noise for single-component aerosols in Chapter 3.

Chapter 7

Spectroscopic Measurements of Refractive Indices

7.1 Modeling of Refrative Index Reconstructions

The following chapters provide an algorithm for the reconstruction of refractive indices from spectral measurements of monodisperse aerosols. The experiments we are conducting are similar to the experiments presented in [38] with the difference that we are using air as surrounding medium for the aerosol particles and that temperature and pressure may approach 200°C and 8 bar respectively. For these rigid conditions reliable databases for refractive indices do not exist up to now. These refractive index databases are needed for the measurement of particle size distributions of polydisperse aerosols using the FASP.

As outlined in Chapter 2 the FASP measures light intensities $I_{long}(l)$ and $I_{short}(l)$ having passed a long and a short measurement path length L_{long} and L_{short} respectively. The evaluations of the FASP measurements are based on the relation

$$\int_0^\infty k(r,l)n(r)dr = e(l) \quad \text{with} \quad e(l) = -\frac{\log(I_{long}(l)) - \log(I_{short}(l))}{L_{long} - L_{short}}, \qquad (7.1.1)$$

where $k(r,l) := \pi r^2 Q_{ext}(m_{med}(l), m_{part}(l), r, l)$ is the so-called kernel function, l is the wavelength of the incident light, r is the radius of the spherical scattering particle and $m_{med}(l)$ and $m_{part}(l)$ are the refractive indices of the surrounding medium and the particle material depending on the wavelength l. The function $Q_{ext}(m_{med}(l), m_{part}(l), r, l)$ is the *Mie extinction efficiency* from [5]. The function $n(r)$ is the size distribution of the scattering particels. The right-hand side $e(l)$ in (7.1.1) is denoted as the *spectral extinction.*

Now if $n(r)$ is the size distribution of a *monodisperse aerosol*, where all particles possess the same radius r_m, it is given by $n(r) = n\delta(r - r_m)$, where n is the total number of particles and $\delta(r - r_m)$ is a Dirac delta distribution truncated on the positive half-axis. Inserting this into (7.1.1) gives

$$n\pi r_m^2 Q_{ext}(m_{med}(l), m_{part}(l), r_m, l) = e(l), \qquad (7.1.2)$$

hence the Mie extinction efficiency is measured directly at the radius r_m.

The Mie extinction efficiency is given as an infinite series, i.e.

$$Q_{ext}(m_{med}(l), m_{part}(l), r, l) = \sum_{n=1}^{\infty} q_n(m_{med}(l), m_{part}(l), r, l).$$

The computation of the coefficient functions $q_n(m_{med}(l), m_{part}(l), r, l)$ will be discussed in Section 1.

It is clear that in practical computations $Q_{ext}(m_{med}(l), m_{part}(l), r, l)$ can only be approximated by a truncated series, because only the computation of a finite number of the $q_n(m_{med}(l), m_{part}(l), r, l)$'s is practically feasible.

We now fix a wavelength l. The complex refractive index $m_{part}(l)$ for the wavelength l is reconstructed from FASP measurements of several monodisperse aerosols with particle radii r_1, ..., r_N. Let $q(r_1, l)$, ..., $q(r_N, l)$ denote the measured spectral extinctions $e(l)$ corresponding to the particle radii r_1, ..., r_N. We assume that they are contaminated by additive Gaussian noise, i.e. $q(r_i, l) = q_{true}(r_i, l) + \delta_i$ with $\delta_i \sim \mathcal{N}(0, s_i^2)$ for $i = 1, ..., N$. Furthermore we assume that the standard deviations s_i can be estimated from measurements sufficiently accurately, such that we can regard them as known. We have

$$q_{true}(r_i, l) = n_i \pi r_i^2 \sum_{n=1}^{\infty} q_n(m_{med}(l), m_{part}(l), r_i, l),$$

where n_i is the number of particles having the same radius r_i. Then a reconstrution of $m_{part}(l)$ is obtained from the set of solutions $M(l)$ of the nonlinear regression problem

$$M(l) := \underset{m \in \mathbb{C}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2\left(\frac{s_i}{n_i}\right)^2} \left(\pi r_i^2 \sum_{n=1}^{N_{tr}} q_n(m_{med}(l), m, r_i, l) - \frac{q(r_i, l)}{n_i} \right)^2. \qquad (7.1.3)$$

Note that $M(l)$ contains in general more than one solution, especially when $q(r_i, l)$ is perturbed by measurement noise. We discuss nonlinear regression problems with truncated series expansions such as (7.1.3) in Section 7.2.

For solving (7.1.3) we use a global optimization strategy presented in Section 7.3 to generate reasonable candidates for start values for a local solver for a regularized version of (7.1.3). Section 7.6 provides a selection method to find a unique start value out of the candidates. In order to apply a gradient-based local solver we must know the derivatives of the Mie extinction efficiency series, which are discussed in Section A.

7.2 Nonlinear Regression using Truncated Series Expansions

We wish to reconstruct the refractive indices of a particle material from spectral measurements by solving a nonlinear regression problem of the form

$$X_{t,\delta} := \underset{\boldsymbol{x} \in \mathbb{R}^D}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} \left(\sum_{n=1}^{t} a_n^i(\boldsymbol{x}) - \sum_{n=1}^{\infty} a_n^i(\boldsymbol{x}_{true}) - \delta_i \right)^2, \qquad (7.2.1)$$

where $t \in \mathbb{N}$ is a finite truncation index and $\delta_i \sim \mathcal{N}(0, s_i^2)$. We have to confine ourselves to a finite truncation index t, because it is practically not feasible to compute all coefficient functions $a_n^i(\boldsymbol{x})$ for $i = 1, ..., N$. We also wish to keep the truncation index as small as possible in order to save computational effort. Therefore this section is devoted to study the influence of the truncation index on the accuracy of the reconstructions.

Remember that N represents the number of particle radii r_i of the different monodisperse aerosols we are investigating. We still assume that for each radius r_i the standard deviations s_i are determined well enough from a set of experiments, such that they can be regarded as known. Throughout this paper we assume that the feasible set Ω is compact.

We define the functions $\boldsymbol{f}_t : \mathbb{R}^D \to \mathbb{R}^N$ and $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^N$ by

$$\boldsymbol{f}_t(\boldsymbol{x}) := \left(\sum_{n=1}^{t} a_n^1(\boldsymbol{x}),\ ...,\ \sum_{n=1}^{t} a_n^N(\boldsymbol{x}) \right)^T$$

$$\text{and} \quad \boldsymbol{f}(\boldsymbol{x}) := \left(\sum_{n=1}^{\infty} a_n^1(\boldsymbol{x}),\ ...,\ \sum_{n=1}^{\infty} a_n^N(\boldsymbol{x}) \right)^T .$$

We set $\boldsymbol{e} := \boldsymbol{f}(\boldsymbol{x}_{true}) + \boldsymbol{\delta}$ with $\boldsymbol{\delta} := (\delta_1, ..., \delta_N)^T$. Then the observed probability density is given by

$$p_{observed}(\boldsymbol{e}|\boldsymbol{x}) := (2\pi)^{-\frac{N_l}{2}} \left|\det(\boldsymbol{\Sigma_\sigma})\right|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{e})\|_2^2)$$

with the covariance matrix $\boldsymbol{\Sigma_\sigma} := \operatorname{diag}\left(\sigma_1^2, ..., \sigma_N^2\right)$. We know a priori that the vector $\boldsymbol{x}$ specifying our model $\boldsymbol{f}_t(\boldsymbol{x})$ lies within the set Ω. This knowledge can be expressed with the prior probability density

$$p_{prior}(\boldsymbol{x}) := (\operatorname{vol}(\Omega))^{-1} I_\Omega(\boldsymbol{x}),$$

where I_Ω is the indicator function of Ω. Now $X_{t,\delta}$ is the set of MAP-estimators of the posterior probability density, i.e.

$$X_{t,\delta} := \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p_{posterior}(\boldsymbol{x}|\boldsymbol{e})$$

$$\text{with} \quad p_{posterior}(\boldsymbol{x}|\boldsymbol{e}) \propto p_{observed}(\boldsymbol{e}|\boldsymbol{x}) p_{prior}(\boldsymbol{x}) \propto \exp(-\tfrac{1}{2}\|\boldsymbol{\Sigma_\sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{e})\|_2^2) I_\Omega(\boldsymbol{x}). \tag{7.2.2}$$

We carry out all the following investigations under the next assumption on the covariance matrix:

Assumption 7.2.1. *The covariance matrix $\boldsymbol{\Sigma_\sigma}$ has the simple form*

$$\boldsymbol{\Sigma_\sigma} = \delta^2 \cdot \operatorname{diag}(\sigma_1^2, ..., \sigma_N^2) =: \delta^2 \cdot \boldsymbol{\Sigma},$$

where $\delta \geq 0$ is an arbitrary but fixed noise level and σ_1, ..., σ_N are fixed.

To simplify notations we introduce the two functions $\boldsymbol{f}_t : \mathbb{R}^D \to \mathbb{R}^N$ and $\boldsymbol{g}_t : \mathbb{R}^D \to \mathbb{R}^N$ depending on the truncation index t and defined by

$$(\boldsymbol{f}_t(\boldsymbol{x}))_i := \sum_{n=1}^{\lfloor t \rfloor} a_n^i(\boldsymbol{x}) + \big(t - \lfloor t \rfloor\big) a_{\lfloor t \rfloor + 1}^i(\boldsymbol{x})$$

$$\text{and} \quad (\boldsymbol{g}_t(\boldsymbol{x}))_i := (\boldsymbol{f}(\boldsymbol{x}))_i - (\boldsymbol{f}_t(\boldsymbol{x}))_i\,, \quad \text{for} \quad i = 1, ..., N.$$

In the following we will investigate how an element $\boldsymbol{x}_{t,\delta}$ of the set $X_{t,\delta}$ depends on the truncation index t. We change to a continuous truncation index here, i.e. we change from now on from (7.2.1) to the new regression problem

$$\begin{aligned} X_{t,\delta} &:= \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^D} F_{t,\delta}(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{x} \in \Omega, \\ \text{with} \quad F_{t,\delta}(\boldsymbol{x}) &:= \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2^2 \end{aligned} \tag{7.2.3}$$

where the truncation index $t \geq 0$ is allowed to be non-integer.

As a preparation we prove the following technical lemma, which will form the basis of our continuity and convergence results.

Lemma 7.2.2. *Let the twice continuously differentiable function $F : \mathbb{R}^N \to \mathbb{R}$ have a strict local minimum $\boldsymbol{x}_0$ inside a compact set $S \subset \mathbb{R}^N$. Let the function $h : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ have the property $\lim_{\varepsilon\to 0} h(\boldsymbol{x}, \varepsilon) = 0$ for all $\boldsymbol{x} \in S$ and let $h(\boldsymbol{x}, \varepsilon)$ be twice continuously differentiable with respect to $\boldsymbol{x}$ and continuous in ε. Furthermore we assume that the local minima $\boldsymbol{x}_\varepsilon$ of $F_\varepsilon(\boldsymbol{x}) := F(\boldsymbol{x}) + h(\boldsymbol{x}, \varepsilon)$ are strict for any $\varepsilon > 0$. Then there exists a sequence of local minima $\boldsymbol{x}_\varepsilon$ of $F_\varepsilon(\boldsymbol{x})$ with $\lim_{\varepsilon\to 0} \boldsymbol{x}_\varepsilon = \boldsymbol{x}_0$.*

Proof. The strategy of the proof is to construct for given ε a neighborhood of $\boldsymbol{x}_0$ which must contain a local minimizer $\boldsymbol{x}_\varepsilon$ of the perturbed function $F_\varepsilon(\boldsymbol{x})$. By sending ε to 0, this neighborhood shrinks down to the local minimum $\boldsymbol{x}_0$ itself, thus yielding the convergence of $\boldsymbol{x}_\varepsilon$ to $\boldsymbol{x}_0$. To have this neighborhood shrink down to $\boldsymbol{x}_0$, it is crucially important that $\boldsymbol{x}_0$ must be a strict local minimum.

We define $d(\varepsilon) := \sup_{\boldsymbol{x}\in S} |h(\boldsymbol{x}, \varepsilon)|$. From $\lim_{\varepsilon\to 0} h(\boldsymbol{x}, \varepsilon) = 0$ for all $\boldsymbol{x} \in S$ follows $\lim_{\varepsilon\to 0} d(\varepsilon) = 0$. Let us now introduce the function $F^-(\boldsymbol{x}) := F(\boldsymbol{x}) - d(\varepsilon)$. Obviously $\boldsymbol{x}_0$ is also a local minimum of $F^-(\boldsymbol{x})$, so for ε sufficiently small there exists a neighborhood $U_{2d(\varepsilon)}(\boldsymbol{x}_0) \subset S$ of $\boldsymbol{x}_0$ with

$$F^-(\boldsymbol{x}) \geq F^-(\boldsymbol{x}_0) \quad \text{and} \quad F^-(\boldsymbol{x}) - F^-(\boldsymbol{x}_0) \leq 2d(\varepsilon) \quad \text{for all} \quad \boldsymbol{x} \in U_{2d(\varepsilon)}(\boldsymbol{x}_0).$$

In particular we have

$$\forall \boldsymbol{x} \in \partial U_{2d(\varepsilon)}(\boldsymbol{x}_0) : \; F^-(\boldsymbol{x}) = F^-(\boldsymbol{x}_0) + 2d(\varepsilon) = F(\boldsymbol{x}_0) + d(\varepsilon).$$

Let us assume that there exists an $\boldsymbol{x} \in \partial U_{2d(\varepsilon)}(\boldsymbol{x}_0)$ with

$$F_\varepsilon(\boldsymbol{x}) < F(\boldsymbol{x}_0) + d(\varepsilon) = F^-(\boldsymbol{x}) = F(\boldsymbol{x}) - d(\varepsilon).$$

Then $F_\varepsilon(\boldsymbol{x}) = F(\boldsymbol{x}) + h(\boldsymbol{x}, \varepsilon)$ implies $-d(\varepsilon) > h(\boldsymbol{x}, \varepsilon)$, hence $-h(\boldsymbol{x}, \varepsilon) > d(\varepsilon) \geq -h(\boldsymbol{x}, \varepsilon)$ by definition of $d(\varepsilon)$, contradiction. Therefore we conclude

$$\forall \boldsymbol{x} \in \partial U_{2d(\varepsilon)}(\boldsymbol{x}_0) : \; F_\varepsilon(\boldsymbol{x}) \geq F(\boldsymbol{x}_0) + d(\varepsilon). \tag{7.2.4}$$

Since $F_\varepsilon(\boldsymbol{x})$ is continuous and $\overline{U}_{2d(\varepsilon)}(\boldsymbol{x}_0)$ is compact for ε small enough, there exists an $\boldsymbol{x}_\varepsilon \in \overline{U}_{2d(\varepsilon)}(\boldsymbol{x}_0)$ with

$$F_\varepsilon(\boldsymbol{x}_\varepsilon) = \min_{\boldsymbol{x}\in\overline{U}_{2d(\varepsilon)}(\boldsymbol{x}_0)} F_\varepsilon(\boldsymbol{x}).$$

Let us assume $F_\varepsilon(\boldsymbol{x}_\varepsilon) > F(\boldsymbol{x}_0) + d(\varepsilon)$. Then by definition of $\boldsymbol{x}_\varepsilon$ we get in particular

$$F(\boldsymbol{x}_0) + h(\boldsymbol{x}_0, \varepsilon) = F_\varepsilon(\boldsymbol{x}_0) \geq F_\varepsilon(\boldsymbol{x}_\varepsilon) > F(\boldsymbol{x}_0) + d(\varepsilon),$$

i.e. $h(\boldsymbol{x}_0, \varepsilon) > d(\varepsilon) \geq h(\boldsymbol{x}_0, \varepsilon)$, contradiction. It follows

$$F_\varepsilon(\boldsymbol{x}_\varepsilon) \leq F(\boldsymbol{x}_0) + d(\varepsilon) \quad \text{and} \quad F_\varepsilon(\boldsymbol{x}_0) \leq F(\boldsymbol{x}_0) + d(\varepsilon), \tag{7.2.5}$$

where the latter follows with a proof by contradiction as well.

If it happens to hold that $F_\varepsilon(\boldsymbol{x}_\varepsilon) = F(\boldsymbol{x}_0) + d(\varepsilon)$, then we also have $F_\varepsilon(\boldsymbol{x}_0) = F(\boldsymbol{x}_0) + d(\varepsilon)$. Otherwise we have $F_\varepsilon(\boldsymbol{x}_\varepsilon) < F(\boldsymbol{x}_0) + d(\varepsilon)$ and then (7.2.4) implies that $\boldsymbol{x}_\varepsilon$ cannot lie on $\partial U_{2d(\varepsilon)}(\boldsymbol{x}_0)$, thus it must lie within the interior of $U_{2d(\varepsilon)}(\boldsymbol{x}_0)$. So in any case (7.2.5) gives that $U_{2d(\varepsilon)}(\boldsymbol{x}_0)$ must contain a local minimizer $\boldsymbol{x}_\varepsilon$ of $F_\varepsilon(\boldsymbol{x})$.

Now $\lim_{\varepsilon\to 0} d(\varepsilon) = 0$ gives $\lim_{\varepsilon \to 0} \boldsymbol{x}_\varepsilon = \boldsymbol{x}_0$. The existence of the last limit is guaranteed by the fact that $\boldsymbol{x}_0$ is strict and the claim is proved.

□

Proposition 7.2.3. *Let all coefficient functions $a_n^i(\boldsymbol{x})$ be twice continuously differentiable and bounded on Ω. We assume that each local minimum $\boldsymbol{x}_{t,\delta}$ of the right hand side function $F_{t,\delta}(\boldsymbol{x})$ in* (7.2.3) *is strict and lies in the interior of Ω. Then each local minimum depends continuously on the truncation index t.*

Proof. To prove the claim, one could be tempted to apply the implicit function theorem on the equation $\boldsymbol{d}(t, \boldsymbol{x}_{t,\delta}) = 0$ with $\boldsymbol{d}(s, \boldsymbol{x}) := \nabla \boldsymbol{f}_s(\boldsymbol{x})$. This would give that the local minima are parameterized by a function $\boldsymbol{m}(s)$ with the property $\boldsymbol{m}(t) = \boldsymbol{x}_{t,\delta}$, where s is from an environment $U(t)$ of t. The problem with this approach is that it requires continuous differentiability of $\boldsymbol{d}(s, \boldsymbol{x})$ in the truncation parameter s. Thus the continuous truncation we are using would need more complicated methods such as spline interpolation of the partial sums, which would increase the overall computational effort.

Therefore we use in the following a more direct approach to prove the claim. Let $\varepsilon > 0$ be arbitrary. First we consider an integer truncation index $t \in \mathbb{N}$, i.e. we have $t = \lfloor t \rfloor$. Now for ε small enough, we get $\lfloor t + \varepsilon \rfloor = t$ and $\lfloor t - \varepsilon \rfloor = t - 1$. This gives

$$\begin{aligned} \left(\boldsymbol{f}_{t+\varepsilon}(\boldsymbol{x})\right)_i &= \left(\boldsymbol{f}_t(\boldsymbol{x})\right)_i + \varepsilon a_{t+1}^i(\boldsymbol{x}) \\ \text{and} \quad \left(\boldsymbol{f}_{t-\varepsilon}(\boldsymbol{x})\right)_i &= \left(\boldsymbol{f}_{t-1}(\boldsymbol{x})\right)_i + (1-\varepsilon) a_t^i(\boldsymbol{x}) \\ &= \left(\boldsymbol{f}_t(\boldsymbol{x})\right)_i - \varepsilon a_t^i(\boldsymbol{x}). \end{aligned}$$

As next step we turn to an noninteger truncation index t. In this case, we can always select ε small enough such that $\lfloor t + \varepsilon \rfloor = \lfloor t \rfloor$ and $\lfloor t - \varepsilon \rfloor = \lfloor t \rfloor$ respectively hold. This yields

$$\begin{aligned} \left(\boldsymbol{f}_{t+\varepsilon}(\boldsymbol{x})\right)_i &= \left(\boldsymbol{f}_t(\boldsymbol{x})\right)_i + \varepsilon a_{\lfloor t \rfloor+1}^i(\boldsymbol{x}) \\ \text{and} \quad \left(\boldsymbol{f}_{t-\varepsilon}(\boldsymbol{x})\right)_i &= \left(\boldsymbol{f}_t(\boldsymbol{x})\right)_i - \varepsilon a_{\lfloor t \rfloor+1}^i(\boldsymbol{x}). \end{aligned}$$

Now we introduce the function

$$\boldsymbol{a}(\boldsymbol{x}) := \begin{cases} \left(a_t^1(\boldsymbol{x}), ..., a_t^N(\boldsymbol{x})\right)^T, & \text{for} \quad t - \varepsilon,\ t \in \mathbb{N} \\ \left(a_{\lfloor t \rfloor+1}^1(\boldsymbol{x}), ..., a_{\lfloor t \rfloor+1}^N(\boldsymbol{x})\right)^T, & \text{else.} \end{cases}$$

For $F_{t,\delta}(\boldsymbol{x}) = \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2^2$ this yields

$$F_{t+\varepsilon,\delta}(\boldsymbol{x}) = F_{t,\delta}(\boldsymbol{x}) + 2\varepsilon\big\langle\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{a}(\boldsymbol{x}),\ \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\big\rangle + \varepsilon^2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{a}(\boldsymbol{x})\|_2^2$$

$$\text{and}\quad F_{t-\varepsilon,\delta}(\boldsymbol{x}) = F_{t,\delta}(\boldsymbol{x}) - 2\varepsilon\big\langle\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{a}(\boldsymbol{x}),\ \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\big\rangle + \varepsilon^2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{a}(\boldsymbol{x})\|_2^2.$$

Therefore we obtain both for $F_{t+\varepsilon,\delta}(\boldsymbol{x})$ and $F_{t-\varepsilon,\delta}(\boldsymbol{x})$ a decomposition of the form $F_{t+\varepsilon,\delta}(\boldsymbol{x}) = F_{t,\delta}(\boldsymbol{x}) + h^{\varepsilon}_{t,\delta}(\boldsymbol{x})$ and $F_{t-\varepsilon,\delta}(\boldsymbol{x}) = F_{t,\delta}(\boldsymbol{x}) + h^{\varepsilon}_{t,\delta}(\boldsymbol{x})$ respectively, where the function $h^{\varepsilon}_{t,\delta}(\boldsymbol{x})$ is appropriately selected according to above findings. We can readily check $\lim_{\varepsilon\to 0}|h^{\varepsilon}_{t,\delta}(\boldsymbol{x})| = 0$ for all $\boldsymbol{x} \in \Omega$ from the boundedness of the $a^i_n(\boldsymbol{x})$'s. Then the result follows from Lemma 7.2.2.

□

Corollary 7.2.4. *Let t_1 and t_2 be truncation indices with $t_1 < t_2$. Let $\boldsymbol{x}_{t_1,\delta}$ be a local minimizer of (7.2.1). Let $\gamma \in [0,1]$ and define $t_\gamma := t_1 + \gamma\,(t_2 - t_1)$. Then beginning at $\gamma = 0$ one can successively find local minimizers $\boldsymbol{x}_{t_\gamma,\delta}$ for the truncation index t_γ using numerical continuation, see [39]. Here for $\gamma_1 < \gamma_2$ the minimizer $\boldsymbol{x}_{t_{\gamma_1},\delta}$ is used as a start vector to compute the next minimizer $\boldsymbol{x}_{t_{\gamma_2},\delta}$. The next parameter γ_2 has to be sufficiently close to γ_1, such that the start vector $\boldsymbol{x}_{t_{\gamma_1},\delta}$ still lies within the domain of convergence for Newton's method.* □

We use Corollary 7.2.4 to compute $\boldsymbol{x}_{t,\delta}$ for increasing truncation index t in a stable way. If we would keep t as integer and increase it in integer steps, we might leave the domain of convergence in the continuation method. Therefore we increase them using a smaller step width.

In the following we investigate how well the minimizers $\boldsymbol{x}_{t,\delta}$ of the noise-contaminated regression problem (7.2.3) with truncated series expansions approximate the minimizers $\boldsymbol{x}_{\infty,0}$ of the noise-free and untruncated problem

$$\boldsymbol{x}_{\infty,0} := \underset{\boldsymbol{x}\in\mathbb{R}^D}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2\sigma_i^2}\left(\sum_{n=1}^{\infty} a^i_n(\boldsymbol{x}) - \sum_{n=1}^{\infty} a^i_n(\boldsymbol{x}_{true})\right)^2 \quad \text{s.t.}\quad \boldsymbol{x}\in\Omega. \tag{7.2.6}$$

Proposition 7.2.5. *Let the noise vector $\boldsymbol{\delta}$ fulfill $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2 = 0$ and let the functions $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{f}_{t_\delta}(\boldsymbol{x})$ be bounded on Ω. Assume $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x})\|_2^2 = 0$ for all $\boldsymbol{x}\in\Omega$. Then for any strict minimizer $\boldsymbol{x}_{\infty,0}$ of the right hand side function of (7.2.6) in the interior of Ω exist minimizers $\boldsymbol{x}_{t_\delta,\delta}$ of (7.2.3) with $\lim_{\delta\to 0}\boldsymbol{x}_{t_\delta,\delta} = \boldsymbol{x}_{\infty,0}$. Here we also assume the $\boldsymbol{x}_{t_\delta,\delta}$'s to be strict for all $\delta > 0$.*

Proof. With the notation introduced before we can write

$$\boldsymbol{x}_{\infty,0} \in X_{\infty,0} := \underset{\boldsymbol{x}\in\mathbb{R}^D}{\operatorname{argmin}}\, F_{\infty,0}(\boldsymbol{x}) \quad \text{s.t.}\quad \boldsymbol{x}\in\Omega$$

$$\text{with}\quad F_{\infty,0}(\boldsymbol{x}) := \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2^2.$$

From the decomposition $\boldsymbol{f}_{t_\delta}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}_{t_\delta}(\boldsymbol{x})$ we obtain

$$F_{t_\delta,\delta}(\boldsymbol{x}) = \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\,(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\,\|_2^2 - 2\big\langle\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}),\ \boldsymbol{\Sigma}^{-\frac{1}{2}}\,(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\,\big\rangle + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x})\|_2^2.$$

Then a further decomposition of the first term on the right hand side yields

$$\begin{aligned} F_{t_\delta,\delta}(\boldsymbol{x}) = F_{\infty,0}(\boldsymbol{x}) &- 2\big\langle \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta},\ \boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true})\right)\big\rangle + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2 \\ &-2\big\langle \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}),\ \boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta}\right)\big\rangle + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x})\|_2^2 \\ &=: F_{\infty,0}(\boldsymbol{x}) + H_{t_\delta,\delta}(\boldsymbol{x}). \end{aligned}$$

From the limit $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x})\|_2^2 = 0$, the limit $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2 = 0$ and the boundedness of $\boldsymbol{f}(\boldsymbol{x})$ follows $\lim_{\delta\to 0}|H_{t_\delta,\delta}(\boldsymbol{x})| = 0$ for arbitrary but fixed $\boldsymbol{x}\in\Omega$.

Then the existence of the $\boldsymbol{x}_{t_\delta,\delta}$'s follows from Lemma 7.2.2.

□

At last we study how the minimizers $\boldsymbol{x}_{t_\delta,\delta}$ of (7.2.3) behave for $\delta\to 0$. We begin with a preparing corollary.

Corollary 7.2.6. *Let the assumptions of Proposition 7.2.5 hold. Then we have for any local minimizer $\boldsymbol{x}_{t_\delta,\delta}$ of (7.2.3) approximating a local minimizer $\boldsymbol{x}_{\infty,0}$ of (7.2.6) for $\delta\to 0$ with $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{\infty,0}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 = 0$ that*

$$\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 = 0.$$

Proof. The assumptions of Proposition 7.2.5 give

$$\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2 = 0 \quad\text{and}\quad \lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta})\|_2^2 = 0.$$

We have by continuity of $\boldsymbol{f}(\boldsymbol{x})$ that $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 = 0$. Then

$$\begin{aligned} \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta})\|_2 + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2 \\ \geq \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 \end{aligned}$$

gives the desired result. □

Proposition 7.2.7. *Let the assumptions of Proposition 7.2.5 hold. Assume that the local minimizers $\boldsymbol{x}_{\infty,0}$ of (7.2.6) with $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{\infty,0}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 = 0$ form a discrete set $S_{\infty,0}$. Then the set $L_{\infty,0}$ consisting of the limits $\lim_{\delta\to 0}\boldsymbol{x}_{t_\delta,\delta}$ of local minimizers $\boldsymbol{x}_{t_\delta,\delta}$ of (7.2.3) with $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 = 0$ coincides with $S_{\infty,0}$ and there exists a noise level δ_{max} such that all minimizers $\boldsymbol{x}_{t_\delta,\delta}$ approximating $S_{\infty,0}$ are isolated for all $\delta\leq\delta_{max}$.*

Proof. On the one hand from Proposition 7.2.5 we know that there exists a sequence $\boldsymbol{x}_{t_\delta,\delta}$ of minimizers of (7.2.3) with $\lim_{\delta\to 0}\boldsymbol{x}_{t_\delta,\delta} = \boldsymbol{x}_{\infty,0}$. Then $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{\infty,0}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 = 0$ and Corollary 7.2.6 give $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 = 0$, which implies $S_{\infty,0}\subseteq L_{\infty,0}$.

On the other hand holds for $\boldsymbol{x}_{t_\delta,\delta}$ with $\lim_{\delta\to 0}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 = 0$ that

$$\begin{aligned} \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta})\|_2 + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2 \\ \geq \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 \end{aligned}$$

which implies $\lim_{\delta\to 0} \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 = 0$. In particular this means by continuity of $\boldsymbol{f}(\boldsymbol{x})$ that the vector $\lim_{\delta\to 0} \boldsymbol{x}_{t_\delta,\delta}$ must be a local minimizer of $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2$. Thus we have also shown $L_{\infty,0} \subseteq S_{\infty,0}$.

In the following we number all elements of $S_{\infty,0}$ with the index k, i.e. we write $\boldsymbol{x}^k_{\infty,0}$ for $k = 1, ..., |S_{\infty,0}|$. Similarly we number all minimizers $\boldsymbol{x}_{t_\delta,\delta}$ with $\lim_{\delta\to 0} \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 = 0$ approximating the $\boldsymbol{x}^k_{\infty,0}$'s with $\boldsymbol{x}^k_{t_\delta,\delta}$, i.e. $\lim_{\delta\to 0} \boldsymbol{x}^k_{t_\delta,\delta} = \boldsymbol{x}^k_{\infty,0}$ for $k = 1, ..., |S_{\infty,0}|$. Define

$$D_{min} := \min_{i\neq j} \|\boldsymbol{x}^i_{\infty,0} - \boldsymbol{x}^j_{\infty,0}\|_2.$$

Since $\lim_{\delta\to 0} \boldsymbol{x}^k_{t_\delta,\delta} = \boldsymbol{x}^k_{\infty,0}$, we can find an error levels δ^k_{max} such that

$$\|\boldsymbol{x}^k_{t_\delta,\delta} - \boldsymbol{x}^k_{\infty,0}\|_2 < \tfrac{1}{2} D_{min} \quad \text{for} \quad k = 1, ..., |S_{\infty,0}|,$$

which holds for all $0 \leq \delta \leq \delta^k_{max}$ for each k. Then for all $0 \leq \delta \leq \delta_{max} := \min_k\{\delta^k_{max}\}$ the $\boldsymbol{x}^k_{t_\delta,\delta}$'s must have pairwise mutual distances greater than zero.

□

Now Proposition 7.2.7 gives that the number of local minima $\boldsymbol{x}^k_{t_\delta,\delta}$ remains constant if the noise level δ is small enough. It also yields that these local minima then form a set of separated continuous curves parametrized in δ.

At last we wish to have an estimate of the convergence of the local minima $\boldsymbol{x}^k_{t_\delta,\delta}$ of the truncated and noise contaminated problem to the local minima $\boldsymbol{x}^k_{\infty,0}$ of the noise-free and untruncated problem, which is useful for practical computations.

Proposition 7.2.8. *Let the derivatives of $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}_t(\boldsymbol{x})$ be bounded for any $t \geq 0$. Then for the noise level δ small enough, we can bound for any local minimum $\boldsymbol{x}^k_{t_\delta,\delta}$ the approximation error $\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2$ with a positively weighted linear combination of the residual $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2$, the truncation error $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2$ and the noise estimate $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2$.*

Proof. The first order necessary conditions for a local minimum of $F_{t_\delta,\delta}(\boldsymbol{x}) = F_{\infty,0}(\boldsymbol{x}) + H_{t_\delta,\delta}(\boldsymbol{x})$ at $\boldsymbol{x}^k_{t_\delta,\delta}$ and a local minimum of $F_{\infty,0}(\boldsymbol{x})$ at $\boldsymbol{x}^k_{\infty,0}$ yield in particular

$$\begin{aligned} \left\langle \nabla F_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}) + \nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta}),\ \boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta} \right\rangle &\geq 0 \\ \text{and} \quad \left\langle \nabla F_{\infty,0}(\boldsymbol{x}^k_{\infty,0}),\ \boldsymbol{x}^k_{t_\delta,\delta} - \boldsymbol{x}^k_{\infty,0} \right\rangle &\geq 0, \end{aligned}$$

Adding the last two inequalities yields

$$\left\langle \nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta}),\ \boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta} \right\rangle \geq \left\langle \nabla F_{\infty,0}(\boldsymbol{x}^k_{\infty,0}) - \nabla F_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}),\ \boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta} \right\rangle. \quad (7.2.7)$$

Since $\nabla F_{\infty,0}(\boldsymbol{x})$ is totally differentiable at $\boldsymbol{x}^k_{\infty,0}$ we obtain

$$\nabla F_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}) = \nabla F_{\infty,0}(\boldsymbol{x}^k_{\infty,0}) + \mathrm{Hess}_{F_{\infty,0}}(\boldsymbol{x}^k_{\infty,0}) \left(\boldsymbol{x}^k_{t_\delta,\delta} - \boldsymbol{x}^k_{\infty,0} \right) + \boldsymbol{w}_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}, \boldsymbol{x}^k_{\infty,0}),$$

where $\boldsymbol{w}_{\infty,0}(\boldsymbol{x}, \boldsymbol{x}^k_{\infty,0})$ fulfills

$$\|\boldsymbol{w}_{\infty,0}(\boldsymbol{x}, \boldsymbol{x}^k_{\infty,0})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}^k_{\infty,0}\|_2 \epsilon_{\infty,0}(\boldsymbol{x}, \boldsymbol{x}^k_{\infty,0}) \quad \text{with} \quad \lim_{\boldsymbol{x}\to\boldsymbol{x}^k_{\infty,0}} \epsilon_{\infty,0}(\boldsymbol{x}, \boldsymbol{x}^k_{\infty,0}) = 0.$$

Since $\mathrm{Hess}_{F_{\infty,0}}(\boldsymbol{x}^k_{\infty,0})$ is positive definite, the expression $\big(\langle \boldsymbol{x}, \mathrm{Hess}_{F_{\infty,0}}(\boldsymbol{x}^k_{\infty,0})\, \boldsymbol{x}\rangle\big)^{\frac{1}{2}}$ gives a norm on $\mathbb{R}^D$. Because of the equivalence of all norms in $\mathbb{R}^D$, there exists a constant $C^k_{\infty,0} > 0$ with

$$\Big(\big\langle \boldsymbol{x}, \mathrm{Hess}_{F_{\infty,0}}(\boldsymbol{x}^k_{\infty,0})\, \boldsymbol{x}\big\rangle\Big)^{\frac{1}{2}} \geq C^k_{\infty,0}\|\boldsymbol{x}\|_2 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^D.$$

Since $\lim_{\delta\to 0} \boldsymbol{x}^k_{t_\delta,\delta} = \boldsymbol{x}^k_{\infty,0}$ we can find a noise level ρ^k_{max} such that $|\epsilon_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}, \boldsymbol{x}^k_{\infty,0})| \leq d^k_{\infty,0}$ for all $\delta \leq \rho^k_{max}$, where $d^k_{\infty,0}$ is a constant with $0 \leq d^k_{\infty,0} < \big(C^k_{\infty,0}\big)^2$. Then using

$$\big\langle \boldsymbol{w}_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}, \boldsymbol{x}^k_{\infty,0}), \boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\big\rangle \leq \epsilon_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}, \boldsymbol{x}^k_{\infty,0})\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2^2$$

and (7.2.7) we can estimate

$$\begin{aligned}
&\|\nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2 \\
&\geq \big\langle \nabla F_{\infty,0}(\boldsymbol{x}^k_{\infty,0}) - \nabla F_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}), \boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\big\rangle \\
&= \big\langle \mathrm{Hess}_{F_{\infty,0}}(\boldsymbol{x}^k_{\infty,0})\left(\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\right) - \boldsymbol{w}_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}, \boldsymbol{x}^k_{\infty,0}),\ \boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\big\rangle \\
&\geq \left(\big(C^k_{\infty,0}\big)^2 - \epsilon_{\infty,0}(\boldsymbol{x}^k_{t_\delta,\delta}, \boldsymbol{x}^k_{\infty,0})\right)\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2^2,
\end{aligned}$$

i.e. this gives

$$\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2 \leq \left(\big(C^k_{\infty,0}\big)^2 - d^k_{\infty,0}\right)^{-1}\|\nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2 \tag{7.2.8}$$

for all $\delta \leq \rho^k_{max}$.

We have

$$\begin{aligned}
\nabla H_{t_\delta,\delta}(\boldsymbol{x}) = \ & 2\Big(\mathrm{Jac}^T_{\boldsymbol{g}_{t_\delta}}(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}) - \mathrm{Jac}^T_{\boldsymbol{g}_{t_\delta}}(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta}) \\
& -\mathrm{Jac}^T_{\boldsymbol{f}}(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}) - \mathrm{Jac}^T_{\boldsymbol{f}}(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\Big),
\end{aligned}$$

i.e. we find that

$$\begin{aligned}
\|\nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2 \leq 2\Big(&\|\mathrm{Jac}^T_{\boldsymbol{g}_{t_\delta}}(\boldsymbol{x}^k_{t_\delta,\delta})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2 \\
&+ \|\mathrm{Jac}^T_{\boldsymbol{g}_{t_\delta}}(\boldsymbol{x}^k_{t_\delta,\delta})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 \\
&+ \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\big) \\
&+ \|\mathrm{Jac}^T_{\boldsymbol{f}}(\boldsymbol{x}^k_{t_\delta,\delta})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2 \\
&+ \|\mathrm{Jac}^T_{\boldsymbol{f}}(\boldsymbol{x}^k_{t_\delta,\delta})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2\Big),
\end{aligned}$$

which gives the result.

□

Corollary 7.2.9. *Let the derivatives of $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}_t(\boldsymbol{x})$ be bounded for any $t \geq 0$. Let the truncation indices t_δ depend on the vector of independent Gaussian random variables $\boldsymbol{\delta}$ with $\lim_{\delta\to 0}\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2^2\big) = 0$ such that $\lim_{\delta\to 0}\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x})\|_2^2\big) = 0$ holds for all arbitrary but fixed $\boldsymbol{x} \in \Omega$. Then we have for all minimizers $\boldsymbol{x}^k_{\infty,0}$ of*

(7.2.6) *with* $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x}^k_{\infty,0}) - \boldsymbol{f}(\boldsymbol{x}_{true}))\|_2 = 0$ *that for* δ *sufficiently small there exist minimizers* $\boldsymbol{x}^k_{t_\delta,\delta}$ *of* (7.2.3) *with*

$$\lim_{\delta\to 0} \mathbb{E}\big(\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2\big) = 0.$$

Proof. Proposition 7.2.5 establishes the existence of the $\boldsymbol{x}^k_{t_\delta,\delta}$'s. Corollary 7.2.6 gives

$$\lim_{\delta\to 0} \mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2\big) = 0.$$

Proposition 7.2.7 gives that for δ sufficiently small there exists a constant $K^k_{\infty,0}$ with

$$\mathbb{E}\big(\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2\big) \leq K^k_{\infty,0}\mathbb{E}\big(\|\nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\big).$$

Set $S^k_1 := \sup_{\boldsymbol{x}\in\Omega} \|\mathrm{Jac}^T_{\boldsymbol{g}_{t_\delta}}(\boldsymbol{x})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2 < \infty$ and $S^k_2 := \sup_{\boldsymbol{x}\in\Omega} \|\mathrm{Jac}^T_{\boldsymbol{f}}(\boldsymbol{x})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_2 < \infty$. Then the estimate for $\|\nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2$ in the proof of Proposition 7.2.7 gives

$$\begin{aligned}\mathbb{E}\big(\|\nabla H_{t_\delta,\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\big) \leq 2\Big(& S^k_1\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\big)\\ &+ S^k_1\Big(\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2\big)\\ &+ \mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\big)\Big)\\ &+ S^k_2\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2\big)\\ &+ S^k_2\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2\big)\Big),\end{aligned}$$

which proves the claim since Assumption 7.2.1 gives $\mathbb{E}\big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\delta}\|_2\big) \leq \sqrt{N}\delta$.

□

The strategy for our retrieval algorithm is to start with an initial guess for the truncation index t_{start} and try to find all local minima $\boldsymbol{x}^k_{t_{start},\delta}$. Then the truncation index is gradually increased and starting from $\boldsymbol{x}^k_{t_{start},\delta}$ the continuation method is applied to find finally the local minima $\boldsymbol{x}^k_{t_\delta,\delta}$. Motivated by Propositions 7.2.8 and 7.2.9 only those local minima are considered to be possible approximations to our sought-after refractive index, where the residual $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2$ and an estimate of the truncation error $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{g}_{t_\delta}(\boldsymbol{x}^k_{t_\delta,\delta})\|_2$ are both reasonably small. The latter serves also as a stopping criterion for the continuation method

The initial guess t_{start} has to be selected with care. On the one hand if it is to small, the model is to inaccurate and the retrieval of the sought-after local minima can not be guaranteed. On the other hand if it is too big, computational effort is wasted, since too many Mie coefficient functions with almost vanishing magnitudes and thus essentially not changing the local minima are computed.

7.3 Generation of Candidate Solutions

We now return to our regression problem (7.1.3). For $i = 1, ..., N$ we see that the measured extinctions normalized by the number of particles n_i with radius r_i, i.e. the quantity $\frac{e_i}{n_i}$, is Gaussian-distributed with mean $\frac{1}{n_i} q_{true}(r_i, l)$ and standard deviation $\sigma_i := \frac{s_i}{n_i}$. In the following we fix a wavelength l, i.e. we reconstruct the sought-after particle refractive index $m_{part}(l)$ wavelength by wavelength. In the following the unit both for particle radii and light wavelengths is μm.

We make use of the function $\boldsymbol{q}_{N_{tr}} : \mathbb{R}^2 \to \mathbb{R}^N$ defined by

$$\boldsymbol{q}_{N_{tr}}(\boldsymbol{x}) := \left(\pi r_1^2 \sum_{n=1}^{N_{tr}} q_n(\boldsymbol{x}, r_1, l), \; ..., \; \pi r_N^2 \sum_{n=1}^{N_{tr}} q_n(\boldsymbol{x}, r_N, l) \right)^T , \tag{7.3.1}$$

where $R = \{r_1, ..., r_N\}$ is the particle radius grid and N_{tr} the truncation index to be used. We allow non-integer truncation indices N_{tr} as well, where the non-integer truncation is done like in Proposition 7.2.3. Here the expression $q_n(\boldsymbol{x}, r_k, l)$ is a short notation of $q_n(m_{med}(l), (\boldsymbol{x})_1 + (\boldsymbol{x})_2 i, r_k, l)$ from Section 7.1, where the sought-after refrative index $m_{part}(l)$ is identified with the vector $\boldsymbol{x}$ here, i.e. $m_{part}(l) = (\boldsymbol{x})_1 + (\boldsymbol{x})_2 i$. So its computation follows Chapter 1.

In the following the refractive index search area is given by the rectangle $\Omega := [0, 20] \times [0, 40]$, which means that we only consider refractive indices of particle materials whose real parts lie in the interval $[0, 20]$ and its imaginary parts in the interval $[0, 40]$. This rather large search area makes the algorithm suitable for a wide range of aerosol materials.

Algorithm 6 Reconstruction of Refractive Indices

```
$b_{real} = 20$
$b_{imag} = 40$
$N_{real} = 81$
$N_{imag} = 161$
$c_i = (i-1)\frac{b_{real}}{N_{real}-1}$ for $i = 1, ..., N_{real}$
$d_i = (i-1)\frac{b_{imag}}{N_{imag}-1}$ for $i = 1, ..., N_{imag}$
$R = \{0.1, 0.2, 0.3\}$
$N = 3$
$N_{tr} = 3$
$S_{start} = \{\}$
estimate $\sigma_1^2$, ..., $\sigma_N^2$ from the sample means approximating the standard deviations of $\frac{e_1}{n_1}$, ..., $\frac{e_N}{n_N}$.
$\delta^2 := \max\{\sigma_1^2, ..., \sigma_N^2\}$
$\boldsymbol{\Sigma} := \delta^{-2} \cdot \mathrm{diag}(\sigma_1^2, ..., \sigma_N^2)$
for $i = 1$ to $N_{real}$ do
    for $j = 1$ to $N_{imag}$ do
        compute the Hessian $\boldsymbol{H}(c_i, d_j)$ of $F(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{q}_{N_{tr}}(\boldsymbol{x}) - \boldsymbol{e}_{real}\right)\|_2^2$ at $\boldsymbol{x} = (c_i, d_j)^T$
```

```
17:         if $\boldsymbol{H}(c_i, d_j)$ is positive definite then
18:             use $(c_i, d_j)^T$ as start vector to compute
19:             $\boldsymbol{x}_{new} = \underset{\boldsymbol{x}\in\mathbb{R}^2}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{q}_{N_{tr}}(\boldsymbol{x}) - \boldsymbol{e}\right)\|_2^2 \quad \text{s.t.} \quad \boldsymbol{x} \in [0, b_{real}] \times [0, b_{imag}]$
20:             if $S_{start}$ is empty $\vee\ \frac{\|\boldsymbol{x}-\boldsymbol{x}_{new}\|_2}{\|\boldsymbol{x}\|_2} \geq 10^{-2} \quad \forall \boldsymbol{x} \in S_{start}$ then
21:                 $S_{start} = S_{start} \cup \{\boldsymbol{x}_{new}\}$
22:             end if
23:         end if
24:     end for
25: end for
26: $N_{start} = |S_{start}|$
27: $S_{out} = \{\}$
28: $\tau = 3$
29: for $i = 1$ to $N_{start}$ do
30:     $c = N_{tr}$
31:     $Tol_{rel} = 10^{-3}$
32:     $D_{rel} = \infty$
33:     while $D_{rel} > Tol_{rel}$ do
34:         for $p = 1$ to $10$ do
35:             use the vector $S_{start}(i)$ as start vector to compute
36:             $\boldsymbol{x}_{new} = \underset{\boldsymbol{x}\in\mathbb{R}^2}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{q}_{c+\frac{p}{10}}(\boldsymbol{x}) - \boldsymbol{e}\right)\|_2^2$
                        s.t. $\boldsymbol{x} \in [0, b_{real}] \times [0, b_{imag}]$
37:             $Res_{cur} = \|\boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{q}_{c+\frac{p-1}{10}}(S_{start}(i)) - \boldsymbol{e}\right)\|_2^2$
38:             $Res_{new} = \|\boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{q}_{c+\frac{p}{10}}(\boldsymbol{x}_{new}) - \boldsymbol{e}\right)\|_2^2$
39:             $D_{rel} = \frac{|Res_{cur} - Res_{new}|}{Res_{cur}}$
40:             $S_{start}(i) = \boldsymbol{x}_{new}$
41:         end for
42:         $c = c + 1$
43:     end while
44:     if $Res_{new} < \tau N \delta^2$ then
45:         if $S_{out}$ is empty $\vee\ \frac{\|\boldsymbol{x}-\boldsymbol{x}_{new}\|_2}{\|\boldsymbol{x}\|_2} \geq 10^{-2} \quad \forall \boldsymbol{x} \in S_{out}$ then
46:             $S_{out} = S_{out} \cup \{\boldsymbol{x}_{new}\}$
47:         end if
48:     end if
49: end for
```

In the first loop from lines 14 - 25 a search for local minima of the fit function $F(\boldsymbol{x})$ defined in line 16 for the truncation index $N_{tr} = 3$ is performed. The loop runs through all grid points $(c_i, d_j)^T$ of the search grid defined in lines 5 - 6. If the Hessian of $F(\boldsymbol{x})$ at some grid point $(c_i, d_j)^T$ is positive definite, this point might lie in the vicinity of a local minimum. The Hessian is computed exactly, where the second partial derivatives of the Mie extinction efficiency with respect to the real and imaginary part of the scattering material needed here are computed using the product rule approach from Section A. So we use $(c_i, d_j)^T$ as start point for a local solver in this case. In line 20 we only accept a new local minimum if it is sufficiently different from the local minima already found. Then it is stored in the container S_{start}. This simple global search strategy can find all local minima if the search grid is fine enough.

The second loop from lines 29 - 49 uses the local minima found in the first loop as start points for the continuation method following Proposition 7.2.3 and Corollary 7.2.4. We found that a step width of 0.1 is for our problem a well-balanced choice between too big step widths rendering the continuation method unstable and too small step widths making it computationally inefiicient. With the stopping criterion $D_{rel} \leq Tol_{rel}$ of the while-loop it is approximately checked if the magnitude of the remainder term is small enough. Finally in line 44 it is checked if the residual is small enough. In our implementation we did another run of lines 44 - 48 with $\tau = 5$ and $\tau = 7$ respectively, if none of the reconstructions had a squared residual smaller than $\tau N_r \delta^2$ for the previous τ. This had to be done, because the parameter τ has to be selected carefully in order to estimate the bound on $\mathbb{E}\big(\|\boldsymbol{x}^k_{\infty,0} - \boldsymbol{x}^k_{t_\delta,\delta}\|_2\big)$ derived in the proof of Corollary 7.2.9 correctly.

7.4 Comparison with Established Truncation Index Heuristics

As solution of the forward problem we generated for a discrete set of wavelengths l_1, ..., l_{N_l} unperturbed spectral extinctions normalized with the number of particles of the monodisperse aerosol by computing

$$(\boldsymbol{e}_{true})_{i,j} := \pi r_i^2 \sum_{n=1}^{N_{tr}} q_n(m_{med}(l_j), m_{part}(l_j), r_i, l_j), \quad \text{for} \quad i = 1, ..., N, \quad j = 1, ..., N_l$$

with $m_{part}(l_i)$ taken as the refractive indices of Ag, H_2O and CsI. Here we used the truncation index

$$\begin{aligned} \rho &= 2\pi \frac{r}{l}, \\ M &= \max\{|\rho|, |\rho \cdot m_{part}(l)|, |\rho \cdot m_{med}(l)|\}, \\ N_{tr} &:= \lceil |M + 4.05 \cdot M^{\frac{1}{3}} + 2| \rceil \end{aligned} \tag{7.4.1}$$

introduced in [6].

For particle size distribution reconstructions as outlined in Chapters 2 and 5 we need particle refractive indices for five optical windows, see [40], so the wavelength grid of interest consists of five ranges. These ranges are given by 8 linearly spaced wavelengths from $0.6 - 0.8$ μm, 8 from $1.1 - 1.3$ μm, 8 from $1.6 - 1.8$ μm, 16 from $2.1 - 2.5$ μm and 8 from $3.1 - 3.3$ μm, so we have in total $N_l = 48$ wavelengths.

For each of the 48 wavelengths we generated noisy spectral extinctions $\boldsymbol{e}$ by adding zero-mean Gaussian noise to $\boldsymbol{e}_{true}$, i.e.

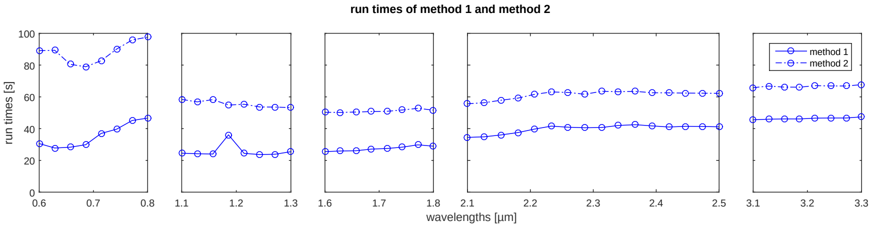
$$(\boldsymbol{e})_{i,j} = (\boldsymbol{e}_{true})_{i,j} + \delta_{i,j} \quad \text{with} \quad \delta_{i,j} \sim \mathcal{N}(0, (0.05 \cdot (\boldsymbol{e}_{true})_{i,j})^2), \quad i = 1, ..., N \quad j = 1, ..., N_l.$$

Here the standard deviations were taken to be 5% of the original extinction values. We computed each mean $(\boldsymbol{e}_{real})_{i,j}$ of the noisy spectral extinctions with a sample size of $N_s = 300$.

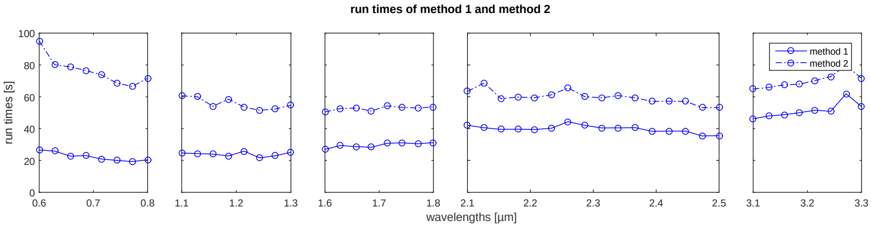
In the following Algortihm 6 is referred to as method 1. On the same simulated spectral extinctions we let Algorithm 6 run up to line 25, but with the difference that at each evaluation of $\boldsymbol{q}_{N_{tr}}(\boldsymbol{x})$ we directly took the trunction index from (7.4.1). We denote this approach with method 2. We now display the average run times of method 1 and method 2 for 10 sweeps through all 48 wavelenghs.

7.4.1 Run Times

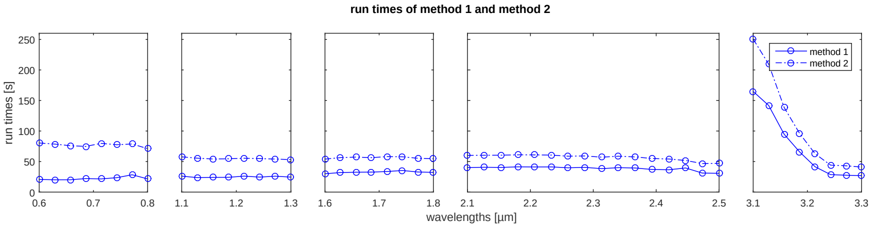
Results for Ag



results for CsI



Results for H_2O



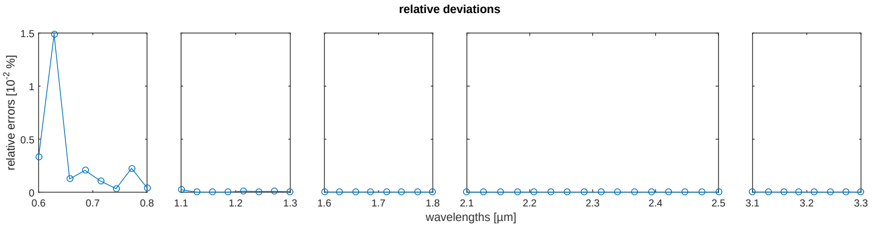
7.4.2 Maximal Relative Deviations

For the 10 simulation runs we list the maximal relative deviations

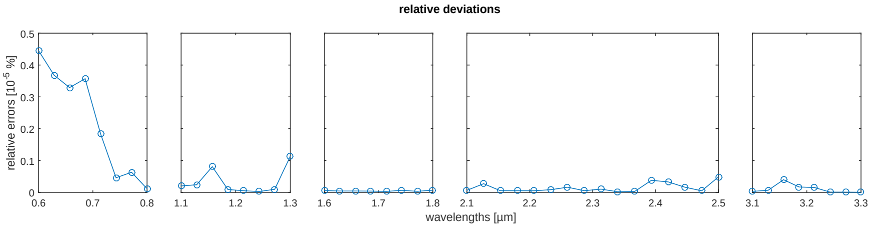
$$100 \cdot \frac{\| \left(n^1_{part}(l_j), k^1_{part}(l_j)\right)^T - \left(n^2_{part}(l_j), k^2_{part}(l_j)\right)^T \|_2}{\| \left(n^1_{part}(l_i), k^1_{part}(l_j)\right)^T \|_2}$$

of the refractive index reconstructions $\left(n^2_{part}(l_j), k^2_{part}(l_j)\right)^T$ from method 2 from $\left(n^1_{part}(l_j), k^1_{part}(l_j)\right)^T$ of method 1 for $j = 1, ..., 48$. At each wavelength, multiple local minima can be detected by both methods. For the relative deviations we always selected the local minima forming the smoothest reconstructions on each optical window in the sense of Section 7.6.

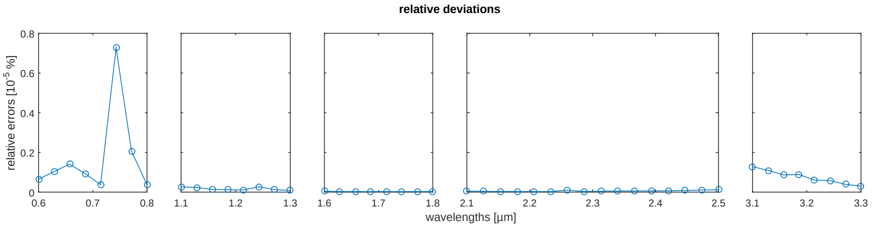
Results for Ag



results for CsI



Results for H_2O



7.4.3 Conclusion

For Ag, the average total run time over all 48 wavelengths for method 1 was 44.0167% less than for method 2, for H_2O 43.9808% and for CsI 44.7322%, i.e. method 1 is almost two times faster than method 2. The results are of the same quality, since their relative deviations are just small fractions of percentages.

The continuation method approach saves run time significantly with the same quality of the results compared to using the truncation index (7.4.1) all the time.

7.5 Candidate Selection and Regularization

So far we have solved the regression problem (7.2.3) without any regularization, thus the obtained refractive index reconstructions might still be too error-contaminated to be of practical use. A widely used regularization strategy for nonlinear regression problems is Tikhonov regularization, which yields the regularized regression problem

$$\boldsymbol{x}_\gamma := \underset{\boldsymbol{x}\in\mathbb{R}^D}{\operatorname{argmin}} \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2^2 + \gamma\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \quad \text{s.t.} \quad \boldsymbol{x} \in \Omega \qquad (7.5.1)$$

when we apply it on (7.2.3), cf. [41]. Here γ is a regularization parameter and $\boldsymbol{x}^*$ is an estimate of the sought-after true solution. In many cases the unregularized problem has a whole set of minimizers, thus the vector $\boldsymbol{x}^*$ works also as a selection criterion. Now if a reasonable $\boldsymbol{x}^*$ is found, the regularization parameter γ can be determined with the discrepancy principle, i.e. γ is computed such that

$$\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}_\gamma) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\|_2 = R(\delta),$$

where $R(\delta)$ is an estimate of the residual of the “true” solution which depends on the noise level δ. For this task monotonicity in the residual of $\boldsymbol{x}_\gamma$ is established in [41].

The problem of finding a good estimate $\boldsymbol{x}^*$ still remains. In [42] an alternative implementable parameter choice strategy without the need of an $\boldsymbol{x}^*$ is derived. Applied on our problem it gives

$$\begin{aligned} &\gamma\big\langle\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}_\gamma) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta}),\ \mathrm{J}_\gamma^{-1}(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{f}_t(\boldsymbol{x}_\gamma) - \boldsymbol{f}(\boldsymbol{x}_{true}) - \boldsymbol{\delta})\big\rangle = R(\delta), \\ \text{with} \quad &\mathrm{J}_\gamma := \gamma\mathrm{I} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{Jac}_{\boldsymbol{f}_t}(\boldsymbol{x}_\gamma)\mathrm{Jac}_{\boldsymbol{f}_t}(\boldsymbol{x}_\gamma)^T\boldsymbol{\Sigma}^{-\frac{1}{2}}. \end{aligned}$$

This method has the drawback that the matrix J_γ needs to be inverted, which may lead to instabilities.

Nevertheless the quality of the regularized solutions still depends strongly on the start values for solving (7.5.1). We know about our sought-after refractive indices that they form smooth curves on each of the five optical windows. The complex refractive index curves of most materials can be described using the so-called Lorentz-oscillator-model, cf. [4]. Here points with bigger curvatures only occur at so-called resonance frequencies corresponding to some isolated resonance wavelengths. Motivated by these facts we derive in the following a method to find reasonable start values for Phillips-Twomey-regularization out of the results of Algorithm 6, which will be outlined in Section 7.7.

7.6 Finding the Smoothest Coupled Solutions

We have the problem of identifying the best approximation to the sought-after true particle material refractive index $\boldsymbol{x}_{true}$ out of a set of multiple solutions obtained with Algorithm 6. This problem is referred to as *input selection problem.* In principle a Bayesian selection mechanism as described in [43] can be applied here, but for our regression problem it is infeasible due to the severe nonlinearity of the Mie extinction series. Instead we use in the context of the framework from [44] a *filter mechanism.* This means we define a measure of the quality for the candidate solutions and select the best ones. We achieve this by coupling the solutions, which means that we combine solutions from neighboring wavelengths l in each of the five optical windows in order to obtain a unique solution for every optical window. We know about the complex refractive index curves to be retrieved that they are smooth, hence we expect their sum of the squared second finite differences both in the real and imaginary parts to be small.

Let l_1, ..., l_s denote the wavelengths of any of our five wavelength ranges. Let N_1, ..., N_s be the number of solutions found for all the wavelengths. We denote with $\boldsymbol{x}_j^i$ the j-th solution found for wavelength l_i for $i = 1, ..., s$ and $j = 1, ..., N_i$. Now we wish to find the smoothest combined solution from all possible combinations $\boldsymbol{x}_{j_1}^1$, ..., $\boldsymbol{x}_{j_s}^s$ for $j_i = 1, ..., N_i$, hence we have a total number of $\prod_{i=1}^s N_i$ combinations. Here we measure smoothness of a combination $\boldsymbol{x}_{j_1}^1$, ..., $\boldsymbol{x}_{j_s}^s$ with the sum

$$S := \sum_{i=2}^{s-1} \left(\left(\left(\boldsymbol{x}_{j_{i-1}}^{i-1}\right)_1 - 2\left(\boldsymbol{x}_{j_i}^{i}\right)_1 + \left(\boldsymbol{x}_{j_{i+1}}^{i+1}\right)_1 \right)^2 + \left(\left(\boldsymbol{x}_{j_{i-1}}^{i-1}\right)_2 - 2\left(\boldsymbol{x}_{j_i}^{i}\right)_2 + \left(\boldsymbol{x}_{j_{i+1}}^{i+1}\right)_2 \right)^2 \right)$$

of its second finite differences both in the real parts $\left(\boldsymbol{x}_{j_i}^i\right)_1$ and its imaginary parts $\left(\boldsymbol{x}_{j_i}^i\right)_2$, which means that we regard a combination the smoother the smaller its sum S is.

We encounter the problem that the total number of possible combinations $\prod_{i=1}^s N_i$ might get too big to iterate through all combinations in the search for the smoothest one in acceptable time. Therefore we propose a greedy algorithm, which uses each second finite difference as start point to find a smooth combination.

Algorithm 7 Detection of the Smoothest Combination

```
$S_{min} = \infty$
$S_{cur} = 0$
$Comb = \{\}$
$SmoothestCombination = \{\}$
for $z = 2$ to $s - 1$ do
    for $c1 = 1$ to $N_{z-1}$ do
        for $c2 = 1$ to $N_z$ do
            for $c3 = 1$ to $N_{z+1}$ do
                $S_{cur} = \left(\left(\boldsymbol{x}_{c1}^{z-1}\right)_1 - 2\left(\boldsymbol{x}_{c2}^{z}\right)_1 + \left(\boldsymbol{x}_{c3}^{z+1}\right)_1\right)^2 + \left(\left(\boldsymbol{x}_{c1}^{z-1}\right)_2 - 2\left(\boldsymbol{x}_{c2}^{z}\right)_2 + \left(\boldsymbol{x}_{c3}^{z+1}\right)_2\right)^2$
                $Comb(z-1) = \boldsymbol{x}_{c1}^{z-1}$
                $Comb(z) = \boldsymbol{x}_{c2}^{z}$
                $Comb(z+1) = \boldsymbol{x}_{c3}^{z+1}$
                for $k = z-2$ to $1$ do
                    $D_{min} = \infty$
                    $D_{cur} = \infty$
                    $\boldsymbol{x}_{min} = (0,0)^T$
                    $\boldsymbol{x}_{mid} = Comb(k+1)$
                    $\boldsymbol{x}_{right} = Comb(k+2)$
                    for $j = 1$ to $N_k$ do
                        $D_{cur} = \Big(\big(\boldsymbol{x}_j^k\big)_1 - 2\big(\boldsymbol{x}_{mid}\big)_1 + \big(\boldsymbol{x}_{right}\big)_1\Big)^2$
                            $+ \Big(\big(\boldsymbol{x}_j^k\big)_2 - 2\big(\boldsymbol{x}_{mid}\big)_2 + \big(\boldsymbol{x}_{right}\big)_2\Big)^2$
                        if $D_{cur} < D_{min}$ then
                            $D_{min} = D_{cur}$
                            $\boldsymbol{x}_{min} = \boldsymbol{x}_j^k$
                        end if
                    end for
                    $S_{cur} = S_{cur} + D_{min}$
                    $Comb(k) = \boldsymbol{x}_{min}$
                end for
                for $k = z+2$ to $s$ do
                    $D_{min} = \infty$
                    $D_{cur} = \infty$
                    $\boldsymbol{x}_{min} = (0,0)^T$
                    $\boldsymbol{x}_{mid} = Comb(k-1)$
                    $\boldsymbol{x}_{left} = Comb(k-2)$
                    for $j = 1$ to $N_k$ do
                        $D_{cur} = \Big(\big(\boldsymbol{x}_{left}\big)_1 - 2\big(\boldsymbol{x}_{mid}\big)_1 + \big(\boldsymbol{x}_j^k\big)_1\Big)^2$
                            $+ \Big(\big(\boldsymbol{x}_{left}\big)_2 - 2\big(\boldsymbol{x}_{mid}\big)_2 + \big(\boldsymbol{x}_j^k\big)_2\Big)^2$
                        if $D_{cur} < D_{min}$ then
                            $D_{min} = D_{cur}$
                            $\boldsymbol{x}_{min} = \boldsymbol{x}_j^k$
                        end if
                    end for
                    $S_{cur} = S_{cur} + D_{min}$
                    $Comb(k) = \boldsymbol{x}_{min}$
                end for
                if $S_{cur} < S_{min}$ then
                    $S_{min} = S_{cur}$
                    $SmoothestCombination = Comb$
                end if
            end for
        end for
    end for
end for
```

The main loop spanning over the lines 5 - 52 iterates through all positions $z = 2, ..., s-1$ of a middle point for a second finite difference both in the real and imaginary part. At each position z the inner loops beginning in lines 6 - 8 iterate through all possible second finite differences which can be formed out of the vectors $\boldsymbol{x}_{c1}^{z-1}$, $\boldsymbol{x}_{c2}^{z}$ and $\boldsymbol{x}_{c3}^{z+1}$ for $c1 = 1, ..., N_{z-1}$, $c2 = 1, ..., N_z$ and $c3 = 1, ..., N_{z+1}$, i.e. they loop through all of their middle points and left and right neighbors at position z. In lines 9 - 12 the variable S_{cur} is initialized with the sum of the squared second finite differences in the real and imaginary parts of the current vectors $\boldsymbol{x}_{c1}^{z-1}$, $\boldsymbol{x}_{c2}^{z}$ and $\boldsymbol{x}_{c3}^{z+1}$ and the positions $z-1$, z and $z+1$ of the array $Comb$ are filled with the current vectors. For $z \geq 3$ the loop in lines 13 - 28 successively fills the positions $k = z-2, z-3, ..., 1$ of the array $Comb$. At each new position k the minimal sum of the two squared second finite differences in the real and imaginary part D_{min} is determined in lines 19 - 25, where the middle and right point are fixed and taken as the leftmost two vectors from the array $Comb$ and the right point runs through all $\boldsymbol{x}_j^k$ for $j = 1, ..., N_k$. After the vector $\boldsymbol{x}_{min}$ giving out of all of the $\boldsymbol{x}_j^k$'s giving the minimal sum D_{min} is found, the D_{min} is added to S_{cur} and $\boldsymbol{x}_{min}$ is stored in k-th entry $Comb(k)$. In a similar way the loop in lines 29 - 44 succesively fills the positions $k = z+2, z+3, ..., s$ for $z \leq s-2$. This time the left and middle point are fixed and taken as the rightmost two points of the array $Comb$, whereas the left point iterates through all $\boldsymbol{x}_j^k$ for $j = 1, ..., N_k$. Again the vector $\boldsymbol{x}_{min}$ is that one of the $\boldsymbol{x}_j^k$'s giving the minimal sum D_{min} and it is stored in $Comb(k)$. As well the sum D_{min} is added to S_{cur}.

In the above procedure every triple of neighboring vectors from the results of Algorithm 6 is considered to possibly lie on the sought-after smoothest combination with the smallest sum of all squared second differences. The three vectors are used as start points to find a smooth combination with a greedy strategy, where only a vector is added to the current combination, if it gives the smallest sum D_{min} at the left or right end of the growing set with vectors already added, until the first and last position are reached.

Finally from all of the combinations constructed this way the smoothest one with the smallest sum S_{min} out of all S_{cur}'s is selected in lines 45 - 48 to be the final output $SmoothestCombination$.

Define $N_{total} := \sum_{j=1}^{s} N_j$. Then the total number of operations needed for Algorithm 7 can be estimated by $\mathcal{O}\big(N_{total} \sum_{j=2}^{s-1} N_{j-1} N_j N_{j+1}\big)$ which is considerably less than the $\mathcal{O}\big(\prod_{j=1}^{s} N_j\big)$ operations needed by the naive method of iterating through all possible combinations.

7.7 Further Regularization of Coupled Solutions

Not only for determining the smoothest refractive index curve reconstructions formed from the results of Algorithm 6 the coupled view on the solutions is beneficial - it also leads to further improvement of the results by Twomey-regularization. Let us investigate the coupled approach in a probability theoretical setting. Here we reuse the notations introduced in Section 7.2, i.e. we let $\boldsymbol{x}^1$, ..., $\boldsymbol{x}^s$ denote a set of solution for any of the five optical windows. Then the joint posterior probability density of

$\boldsymbol{x}^1$, ..., $\boldsymbol{x}^s$ is given by

$$p(\boldsymbol{x}^1,...,\boldsymbol{x}^s|\boldsymbol{e}^1,...,\boldsymbol{e}^s) = \prod_{j=1}^{s} p(\boldsymbol{x}^j|\boldsymbol{e}^j) \propto \exp\left(-\tfrac{1}{2}\sum_{j=1}^{s}\|\boldsymbol{\Sigma}_j^{-\frac{1}{2}}(\boldsymbol{f}_t^j(\boldsymbol{x}^j)-\boldsymbol{e}^j)\|_2^2\right)\prod_{j=1}^{s} I_\Omega(\boldsymbol{x}^j), \tag{7.7.1}$$

where $\boldsymbol{e}^j$ is the data vector for the j-th wavelength l_j having N entries with N being the size of the radius grid. Moreover $\boldsymbol{\Sigma}_j$ is the scaled covariance matrix for l_j and $\boldsymbol{f}_t^j(\boldsymbol{x})$ is the applied model depending on l_j. Note that we initially have differing truncation indices t_1, ..., t_s. Since the coefficient functions of the truncated model function $\boldsymbol{f}_{t_j}(\boldsymbol{x})$ are decaying fast for each t_j, it is convenient to change to the same truncation index $t := \max\{t_1,...,t_s\}$ for all wavelengths l_1, ..., l_s. The errors introduced by doing so are negliglible. It is easy to show that maximizing the joint density (7.7.1) is equivalent to maximize all single densities $p(\boldsymbol{x}^j|\boldsymbol{e}^j)$ independently, i.e. a joint MAP-estimator

$$\boldsymbol{x}_{opt}^1,...,\boldsymbol{x}_{opt}^s = \operatorname*{argmax}_{\boldsymbol{x}^1,...,\boldsymbol{x}^s} p(\boldsymbol{x}^1,...,\boldsymbol{x}^s|\boldsymbol{e}^1,...,\boldsymbol{e}^s)$$

consists of the single MAP-estimators

$$\boldsymbol{x}_{opt}^j = \operatorname*{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{e}^j)$$

for $j = 1,...,s$. This means the the results of Algorithm 6 can be used to construct MAP-estimators for the joint posterior probabilty density.

This behavior changes when we replace the joint prior probabilty density

$$p_{prior}(\boldsymbol{x}^1,...,\boldsymbol{x}^s) = (\mathrm{vol}(\Omega))^{-s}\prod_{j=1}^{s} I_\Omega(\boldsymbol{x}^j)$$

with

$$p_{prior}(\boldsymbol{x}^1,...,\boldsymbol{x}^s) \propto \exp\left(-\tfrac{1}{2}\gamma S(\boldsymbol{x}^1,...,\boldsymbol{x}^s)\right)\prod_{j=1}^{s} I_\Omega(\boldsymbol{x}^j),$$

where

$$\begin{aligned} S(\boldsymbol{x}^1,...,\boldsymbol{x}^s) :=& \sum_{i=2}^{s-1}\left(\left(\left(\boldsymbol{x}^{i-1}\right)_1 - 2\left(\boldsymbol{x}^{i}\right)_1 + \left(\boldsymbol{x}^{i+1}\right)_1\right)^2 + \left(\left(\boldsymbol{x}^{i-1}\right)_2 - 2\left(\boldsymbol{x}^{i}\right)_2 + \left(\boldsymbol{x}^{i+1}\right)_2\right)^2\right) \\ &+\rho\sum_{i=1}^{s}\left(\left(\boldsymbol{x}^i\right)_1^2 + \left(\boldsymbol{x}^i\right)_2^2\right), \end{aligned}$$

where γ is a regularization parameter and ρ is a parameter specifying the amount Tikhonov regularization.

In the new prior distribution we use a combination of Tikhonov and Phillips-Twomey-regularization both in the real and imaginary parts. Here we apply a small amount of Tikhonov-regularization by setting $\rho = 10^{-8}$, such that the resulting regularization operator gets regular. This means that the regularized regression problem (7.7.2) can be transformed into standard Tikhonov form and that the monotonicity results from [41] are still valid. These results were generalized to a statistical setting in [45]. Each second finite difference is clearly a function of three neighboring points, therefore a decoupled computation of the joint MAP-estimator

$$\boldsymbol{x}_{opt}^1, ..., \boldsymbol{x}_{opt}^s := \underset{\boldsymbol{x}^1,...,\boldsymbol{x}^s}{\operatorname{argmax}}\, p(\boldsymbol{x}^1, ..., \boldsymbol{x}^s | \boldsymbol{e}^1, ..., \boldsymbol{e}^s) \tag{7.7.2}$$

with

$$p(\boldsymbol{x}^1, ..., \boldsymbol{x}^s | \boldsymbol{e}^1, ..., \boldsymbol{e}^s) \propto \exp\left(-\tfrac{1}{2} \textstyle\sum_{j=1}^s \| \boldsymbol{\Sigma}_j^{-\frac{1}{2}} (\boldsymbol{f}_t^j(\boldsymbol{x}^j) - \boldsymbol{e}^j) \|_2^2 - \tfrac{1}{2}\gamma S(\boldsymbol{x}^1, ..., \boldsymbol{x}^s) \right) \prod_{j=1}^s I_\Omega(\boldsymbol{x}^j)$$

for each wavelength seperately is not possible anymore after changing to the new prior density. However the result vectors $\boldsymbol{x}^1$, ..., $\boldsymbol{x}^s$ from Algorithm 7 form a good start vector to solve the nonlinear regression problem (7.7.2).

We selected the regularization parameter γ using the discrepancy principle, i.e. we compute γ such that the regularized solution

$$\boldsymbol{x}_\gamma^1, ..., \boldsymbol{x}_\gamma^s := \underset{\boldsymbol{x}^1,...,\boldsymbol{x}^s}{\operatorname{argmin}} \sum_{j=1}^s \| \boldsymbol{\Sigma}_j^{-\frac{1}{2}} (\boldsymbol{f}_t^j(\boldsymbol{x}^j) - \boldsymbol{e}^j) \|_2^2 + \gamma S(\boldsymbol{x}^1, ..., \boldsymbol{x}^s) \quad \text{s.t.} \quad \boldsymbol{x}^j \in \Omega, \quad j = 1, ..., s$$

fulfills a relation of the form

$$\sum_{j=1}^s \| \boldsymbol{\Sigma}_j^{-\frac{1}{2}} (\boldsymbol{f}_t^j(\boldsymbol{x}_\gamma^j) - \boldsymbol{e}^j) \|_2^2 = R(\delta),$$

where $R(\delta)$ is a proposed residual value depending on the noise level δ. In Chapter 2 several different residual values are proposed for a fixed model discretization and a set of regularization parameters is obtained from those using the discrepancy principle. The pairings of model discretizations and regularization parameters obtained this way are compared by their Bayesian posterior probabilities. In the case that the posterior probabilities can be approximated by Gaussians quite well, these probabilities can be computed approximately with Monte Carlo integration methods, see Section 4.2 and [46]. Due to the highly nonlinear behavior of our model $\boldsymbol{f}_t(\boldsymbol{x})$ such integration methods are not feasible here. Therefore we simplified the posterior exploration in such a way that only one residual value is proposed.

Since each observed probability density $p(\boldsymbol{e}^j | \boldsymbol{x}^j)$ for $j = 1, ..., s$ is Gaussian, the joint observed density $p(\boldsymbol{e}^1, ..., \boldsymbol{e}^1 | \boldsymbol{x}^1, ..., \boldsymbol{x}^s) = \prod_{j=1}^s p(\boldsymbol{e}^j | \boldsymbol{x}^j)$ is Gaussian as well. We have $\boldsymbol{x}^j \in \mathbb{R}^2$, thus the sum of residuals $\sum_{j=1}^s \| \boldsymbol{\Sigma}_j^{-\frac{1}{2}} (\boldsymbol{f}_t^j(\boldsymbol{x}^j) - \boldsymbol{e}^j) \|_2^2$ running through all wavelengths in the optical window is $\chi^2(2s)$-distributed. This yields

$$\mathbb{E}\Big(\sum_{j=1}^s \| \boldsymbol{\Sigma}_j^{-\frac{1}{2}} (\boldsymbol{f}_t^j(\boldsymbol{x}^j) - \boldsymbol{e}^j) \|_2^2 \Big) = 2s.$$

Now a widely proposed residual value for the discrepancy principle is $\tau \cdot 2s$, where $\tau = 1.1$ is the so-called Morozov safety factor. This choice is prone to under- or overregularization since the residual value corresponding to the "true" solution might be much smaller or bigger than $2\tau s$. Therefore we proposed a residual value which depends more dynamically on the observed behovior of the residual. For more general Morozov discrepancy principles, where τ fulfills $1 < \tau_1 \leq \tau \leq \tau_2$ with

$\tau_1 < \tau_2$, convergence of the regularized solutions to a minimizer of the nonlinear noise-free fit function was established in [47] under quite general conditions.

Let $\boldsymbol{x}_0^1$, ..., $\boldsymbol{x}_0^s$ denote the unregularized solutions, i.e. the results of Algorithm 7. Then their squared residual is given by $R_0 := \sum_{j=1}^{s} \|\boldsymbol{\Sigma}_j^{-\frac{1}{2}}(\boldsymbol{f}_t^j(\boldsymbol{x}_0^j) - \boldsymbol{e}^j)\|_2^2$. We first proposed

$$R(\delta) = \max\{2\tau_1 s,\ \tau_1 R_0\},$$

where we selected $\tau_1 = 1.1$. This means that the residual of the regularized solution is beginning at R_0 at least increased by the factor τ_1, which avoids underregularization. If it then happens that $\frac{R(\delta)}{R_0} > \theta$ with $\theta = 1.5$ the proposed residual is most likely too big and overregularization occurs. In this case we corrected $R(\delta)$ by setting

$$R(\delta) = \max\{2\tau_2 s,\ \theta R_0\}$$

with $\tau_2 = 0.9$.

Chapter 8

Numerical Results

8.1 Simulations of Refractive Index Reconstructions

To see how reliable our proposed reconstruction algortihm is, we performed for each of the scatterer materials Ag, H_2O a numerical study with 100 sweeps through all 48 wavelenghts of the five optical windows with the same settings as in Section 7.4. We found out that the radii $r_1 := 0.1$ μm, $r_2 := 0.2$ μm and $r_3 := 0.3$ μm contain the most information about the refractive indices. This was found by keeping our 48 wavelengths fixed and comparing the quality of inversion results under varying aerosol particle radii. Bigger radii did not improve the results in our simulations and refractive index reconstructions only using bigger radii even turned out to be too unstable. A more thorough treatment of this problem can be found in [48], where a covariance eigenvalue analysis is used. Although not directly comparable with our study of uncoated particles, the coated radii 0.0975 μm, 0.2305 μm and 0.11 μm carrying the most information content found in this study are roughly comparable to our radii.

We computed original spectral extinctions

$$(\boldsymbol{e}_{true})_{i,j} := \pi r_i^2 \sum_{n=1}^{N_{tr}} q_n(m_{med}(l_i), m_{part}(l_i), r_j, l_i), \quad i = 1, ..., 48, \quad j = 1, ..., 3$$

for Ag, H_2O and CsI and added zero-mean Gaussian noise to it in order to obtain the simulated noisy spectral extinctions

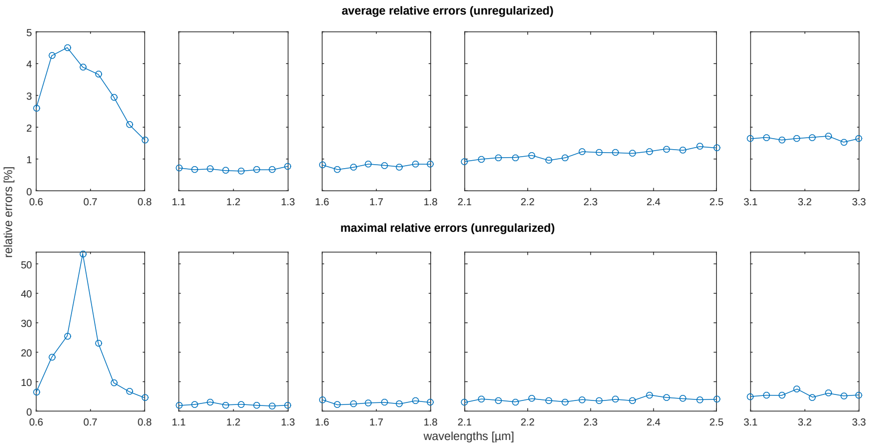
$$(\boldsymbol{e})_{i,j} = (\boldsymbol{e}_{true})_{i,j} + \delta_{i,j} \quad \text{with } \delta_{i,j} \sim \mathcal{N}(0, (0.05 \cdot (\boldsymbol{e}_{true})_{i,j})^2), \quad i = 1, ..., 48, \quad j = 1, ..., 3.$$

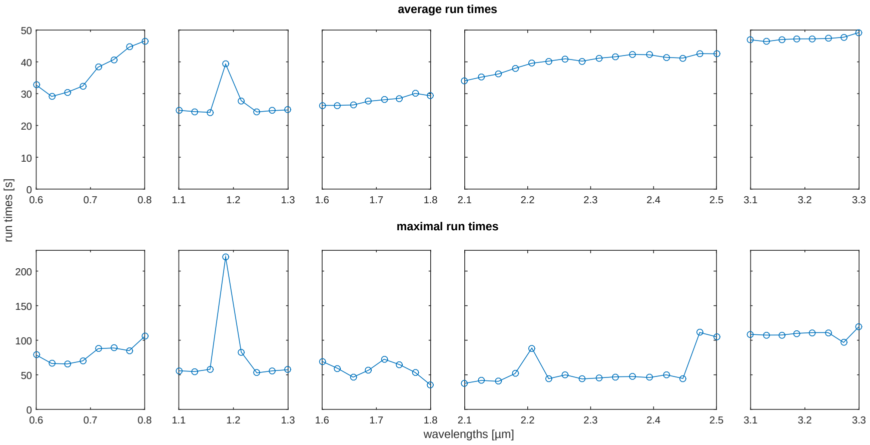
The standard deviations were taken to be 5% of the true spectral extinctions. Real experiments using 500 wavelengths were contaminated by Gaussian noise with 30% of the true spectral extinctions as standard deviations. We expect that switching to 48 wavelengths and thus increasing the time resolution of the measurements will lower the standard deviations to a small percentage. We used a sample size of $N_s = 300$ to compute each mean $(\boldsymbol{e}_{real})_{i,j}$ of noisy spectral extinctions.

In the following the results are presented separately for each of the three materials. The uppermost plot presents the relative errors of the unregularized solutions obtained with Algorithm 7 from the original scatterer refractive indices. The next plot displays the run times of Algorithm 6, which returned all local minima of (7.2.3). These candidate solutions served as input for Algorithm 7. Then the relative errors of the regularized solutions are presented. Finally the relative errors of the average of the regularized solutions are shown.

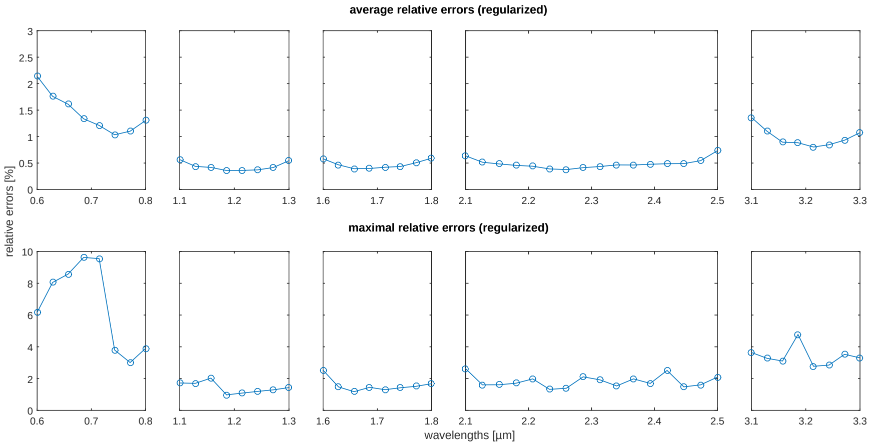
8.2 Results for Ag

8.2.1 Results of Algorithms 6 and 7

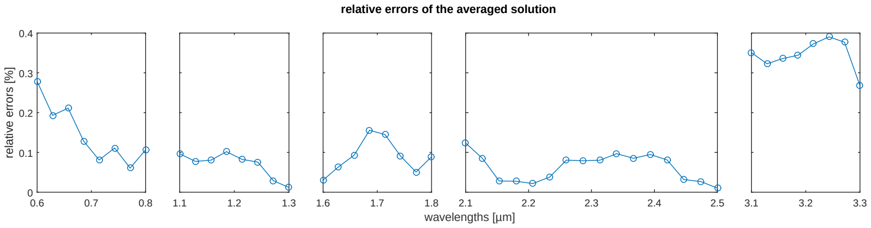




8.2.2 Relative Errors of the Regularized Solutions

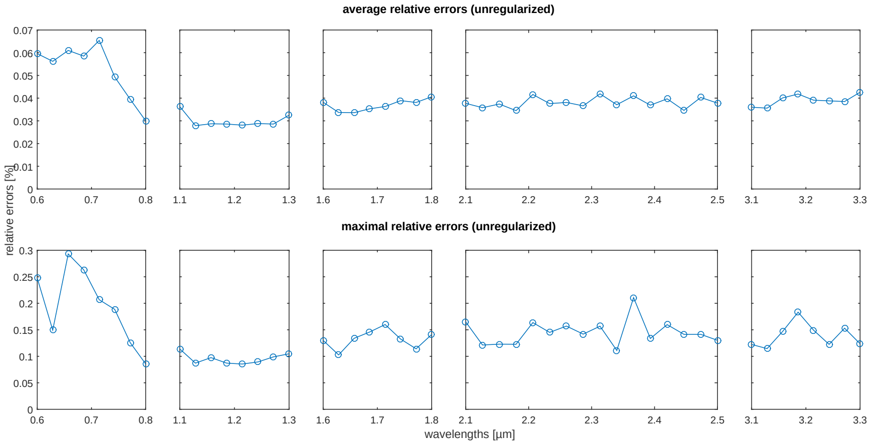


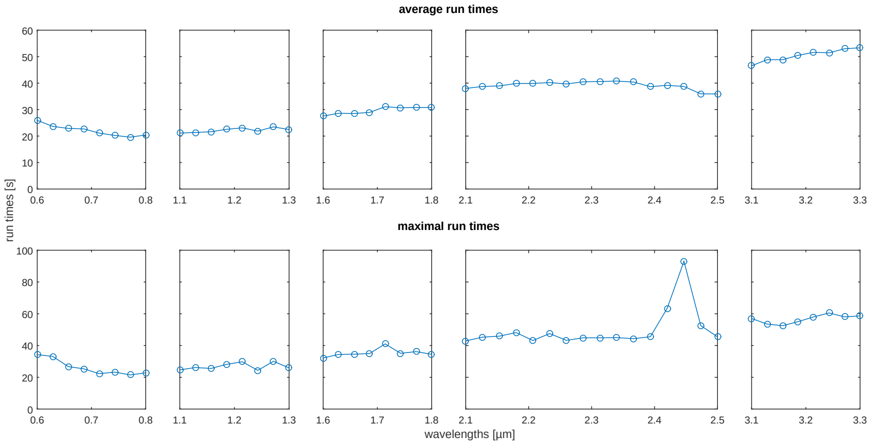
8.2.3 Relative Errors of the Average of the Regularized Solutions



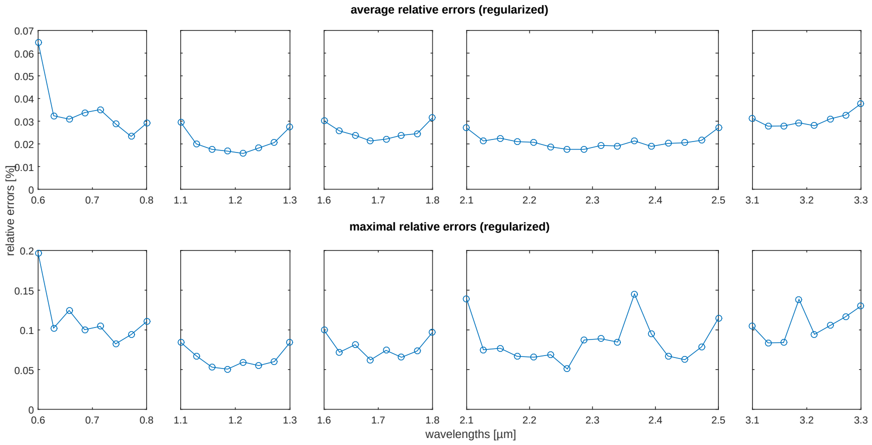
8.3 Results for CsI

8.3.1 Results of Algorithms 6 and 7

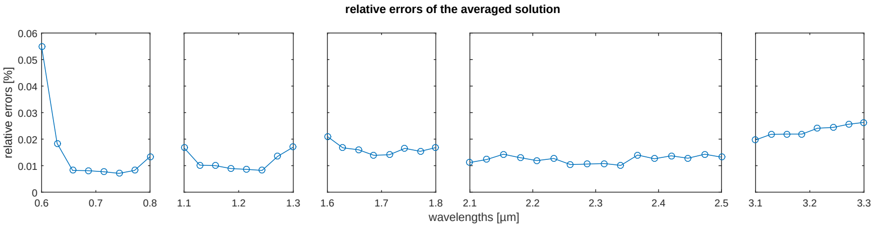




8.3.2 Relative Errors of the Regularized Solutions

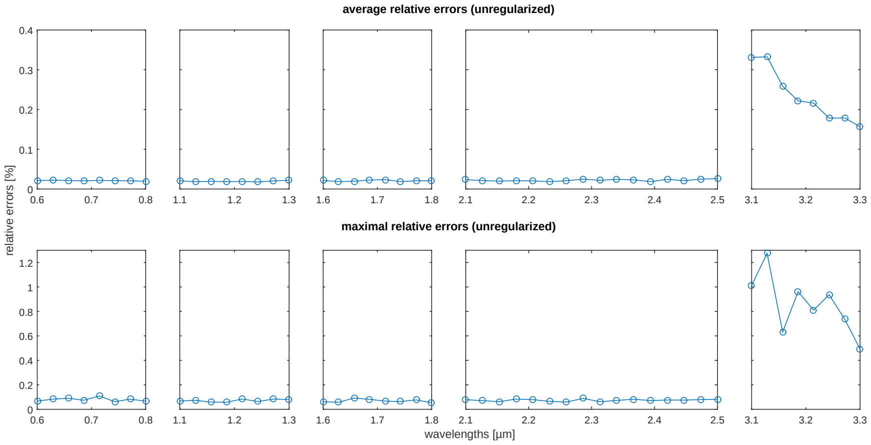


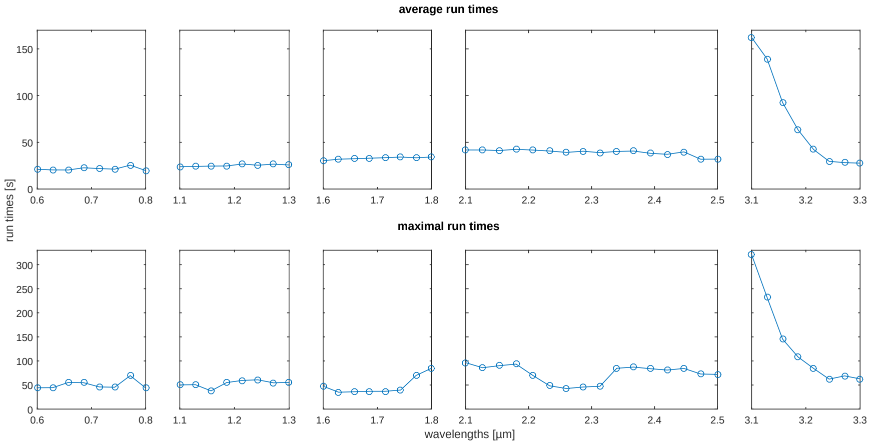
8.3.3 Relative Errors of the Average of the Regularized Solutions



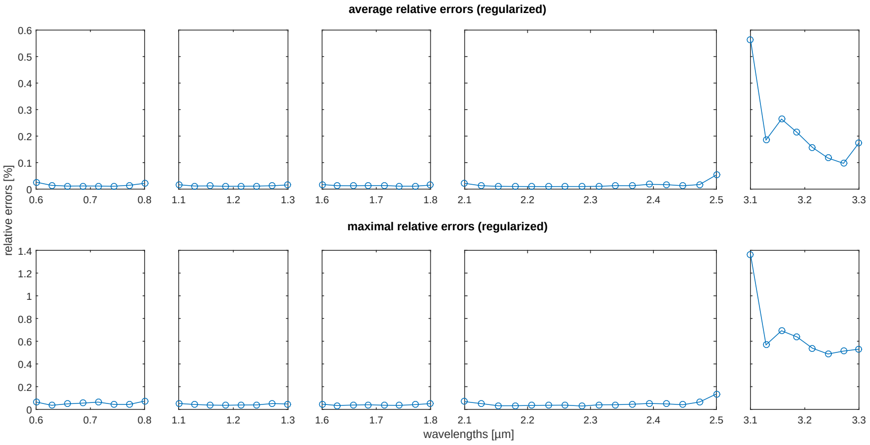
8.4 Results for H_2O

8.4.1 Results of Algorithms 6 and 7

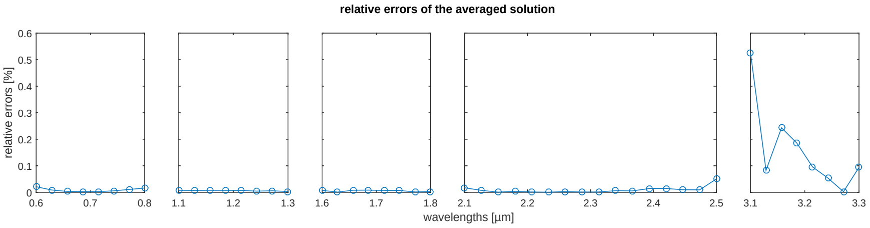




8.4.2 Relative Errors of the Regularized Solutions



8.4.3 Relative Errors of the Average of the Regularized Solutions



8.5 Conclusion

The severest relative errors can be observed for Ag. For the initial unregularized solutions they lie between 1 and 5% on average and can go up to ca. 53% in the extreme cases as one can see in the leftmost subplot for the first optical window. The run times of Algorithm 6 lie between 30 and 50 seconds in the average case and can rise up to 200 seconds in the extreme cases. A typical sweep through all 48 wavelengths needed ca. 30 minutes in total and this value was very much the same for all three materials. For Ag the regularization procedure effectively reduced the relative errors such that they are in the range between 0.5 and 2.2% on average and are below 10% in the extreme cases. Finally one can see in the last plot that the relative errors of the average of all 100 regularized solutions are all below 0.4%.

For CsI the relative errors of the unregularized solutions are already quite small and lie between 0.03 and 0.065% on average and rise only up to 0.3% in the extreme cases. The run times of Algorithm 6 are typically in the range from 25 to 55 seconds and are always below 95 seconds. The regularization of the solutions brought only a small improvement of the results here such that the relative errors did not change much. They are still in the same range from 0.03 and 0.065% on average but only reach up to ca. 0.2% now. The relative errors of the average of the 100 regularized solutions are between 0.01 and 0.055%.

Also for H_2O the relative errors of the unregularized solutions are comparably small and are below 0.35% on average and still below 1.3% in the extreme cases. Especially the rightmost subplot for the last optical window shows the biggest relative errors, whereas for all the other optical windows the relative errors are below 0.03% on average and below 0.15% in the extreme cases. A similar behavior can be observed for the run times of Algorithm 6. For the first four optical windows they are between 20 and 45 seconds on average and below 100 seconds in the extreme cases, whereas for the last optical window they are between 30 and 170 seconds on average and can even rise up to 350 seconds. For H_2O the regularization procedure improves the relative errors only slightly for the first four optical windows and even increases them for the last optical window such that they can rise up to ca. 0.06% on average and 1.4% in the extreme cases. The relative errors of the average of the 100 regularized solutions are virtually zero for the first four optical windows and below 0.55% for the last optical window.

8.6 Higher Noise Levels

To see how our proposed reconstruction algortihm behaves for higher noise levels, we performed for each of the scatterer materials Ag, H_2O and CsI two numerical studies with 10 sweeps through all 48 wavelenghts of the five optical windows with the same settings as in Section 7.4. We computed original spectral extinctions

$$(\boldsymbol{e}_{true})_{i,j} := \pi r_i^2 \sum_{n=1}^{N_{tr}} q_n(m_{med}(l_i), m_{part}(l_i), r_j, l_i), \quad i = 1, ..., 48, \quad j = 1, ..., 3$$

for all wavelengths and added zero-mean Gaussian noise to it in order to obtain the simulated noisy spectral extinctions

$$(\boldsymbol{e})_{i,j} = (\boldsymbol{e}_{true})_{i,j} + \delta_{i,j} \quad \text{with} \quad \delta_{i,j} \sim \mathcal{N}(0, (0.15 \cdot (\boldsymbol{e}_{true})_{i,j})^2), \quad i = 1, ..., 48, \quad j = 1, ..., 3.$$

for the first study and

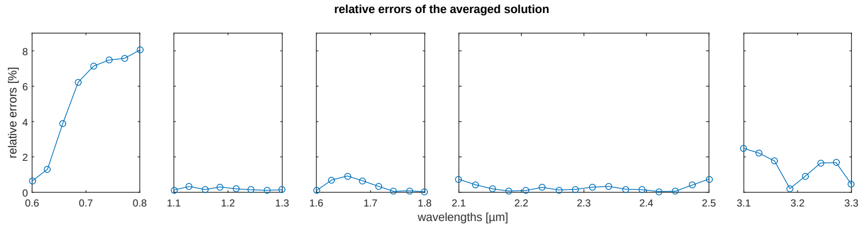
$$(\boldsymbol{e})_{i,j} = (\boldsymbol{e}_{true})_{i,j} + \delta_{i,j} \quad \text{with} \quad \delta_{i,j} \sim \mathcal{N}(0, (0.3 \cdot (\boldsymbol{e}_{true})_{i,j})^2), \quad i = 1, ..., 48, \quad j = 1, ..., 3.$$

for the second. The standard deviations were taken to be 15% and 30% respectively of the true spectral extinctions. We used a sample size of $N_s = 300$ to compute each means $(\boldsymbol{e}_{real})_{i,j}$ of noisy spectral extinctions.

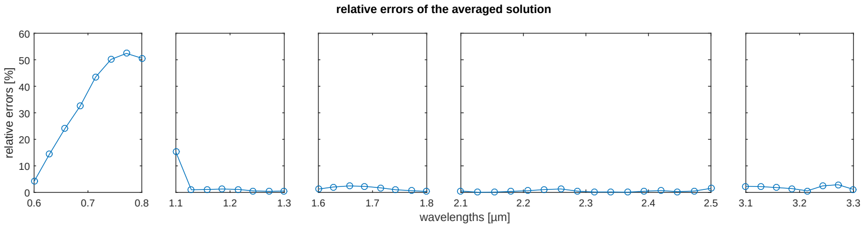
For brevity we only present the relative errors of the average of the 10 regularized solutions.

8.6.1 Results for Ag

Results for 15% Noise

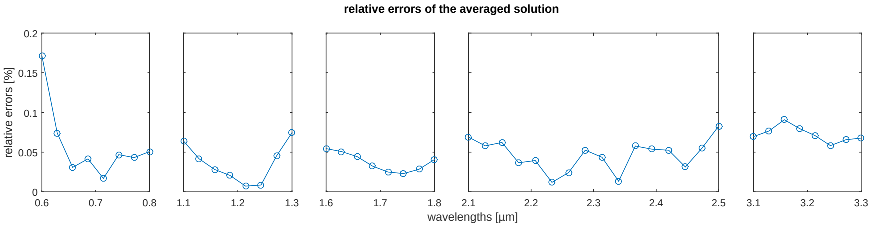


Results for 30% Noise

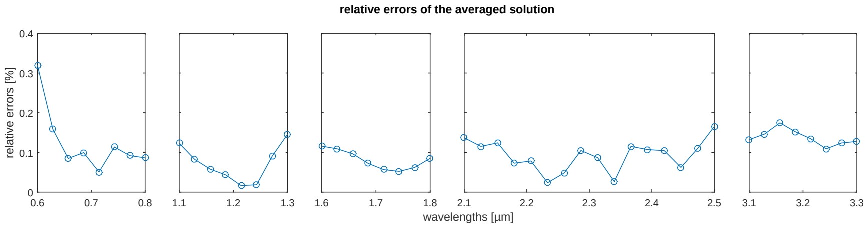


8.6.2 Results for CsI

Results for 15% Noise

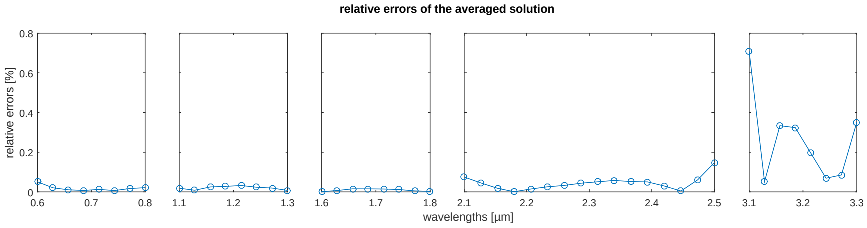


Results for 30% Noise

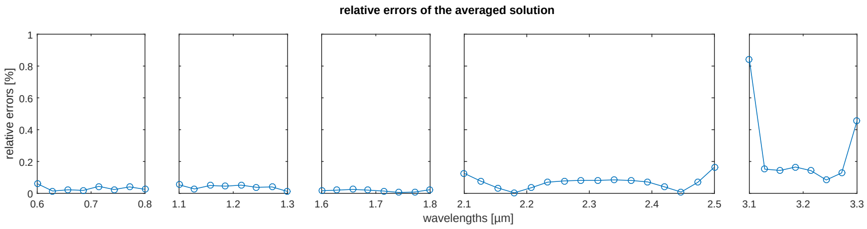


8.6.3 Results for H_2O

Results for 15% Noise



Results for 30% Noise



8.7 Conclusion

Whereas the relative errors for CsI and H_2O are still below 1%, they can rise up to ca. 53% for Ag. Therefore the reconstructed refractive indices for Ag under this noise level are most likely not of practical use. This shows that the FASP measurements of monodisperse aerosols must be sufficiently accurate in order to retrieve the scatterer refractive indices from them.

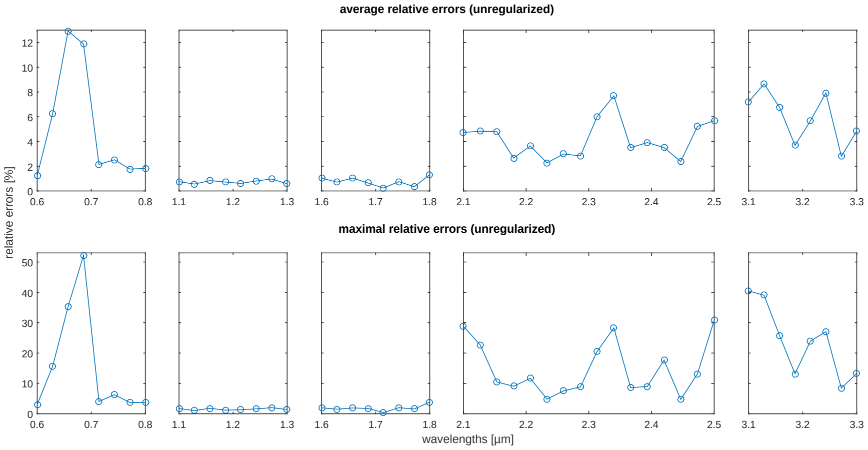
8.8 Comparison with Genetic Algorithms

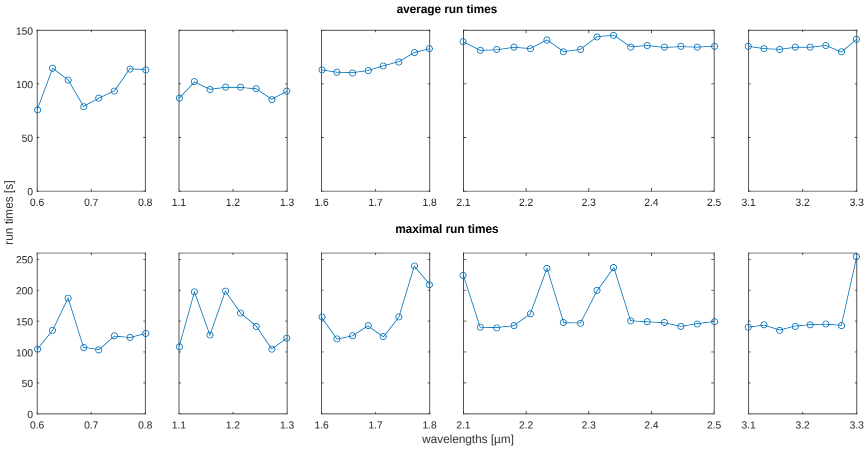
Of course the sequential search strategy is not the only way to find the candiate solutions. In this section we present the results obtained with a genetic algorithm. We used the same settings as in previous sections, i.e. we generated for each refractive index retrieval artificial measurement data consisting of 300 single measurements perturbed by zero-mean Gaussian noise with a standard deviation of 5% of the true extinction values.

We applied the MATLAB function “ga” with its standard settings. For each refractive index retrieval we performed five sweeps of the “ga” function in order to capture all local minima of the fit function of interest.

8.9 Results for Ag

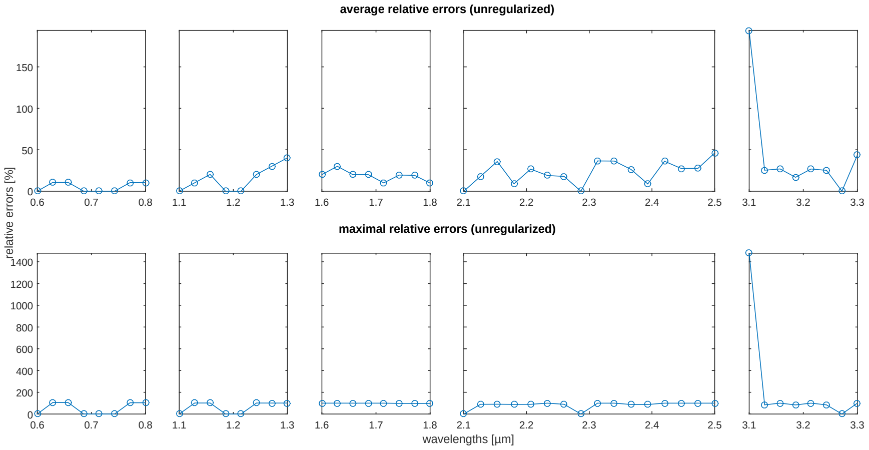
8.9.1 Results of the Genetic Algorithm and Algorithm 7

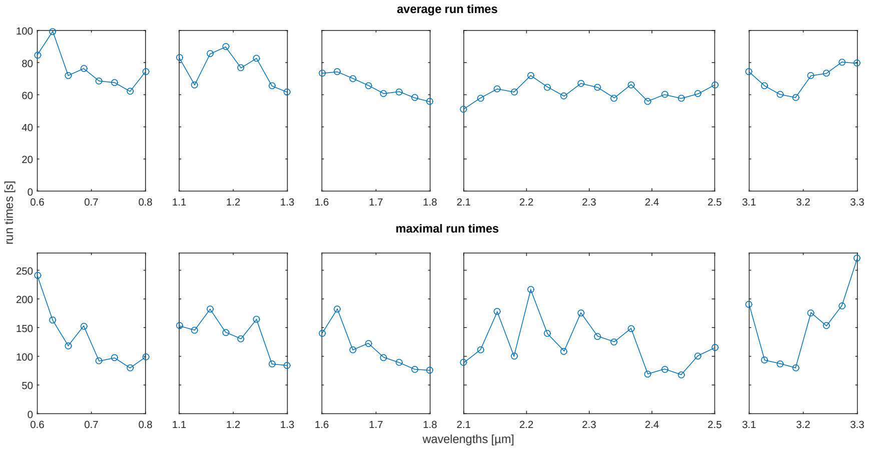




8.10 Results for CsI

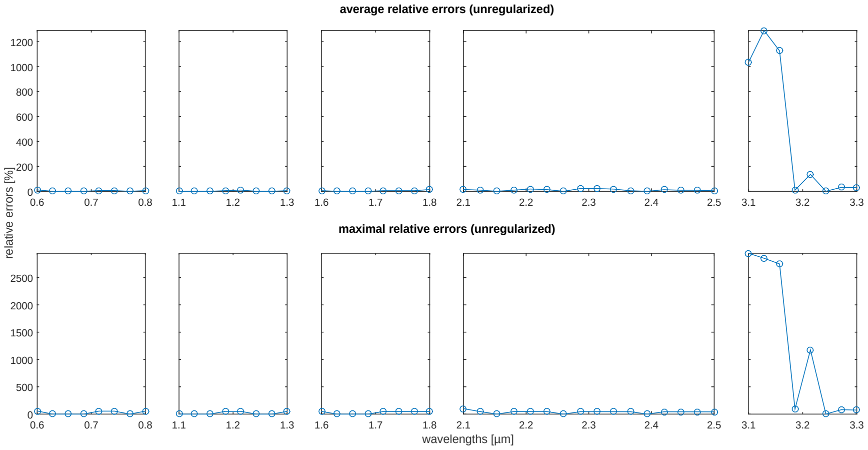
8.10.1 Results of the Genetic Algorithm and Algorithm 7

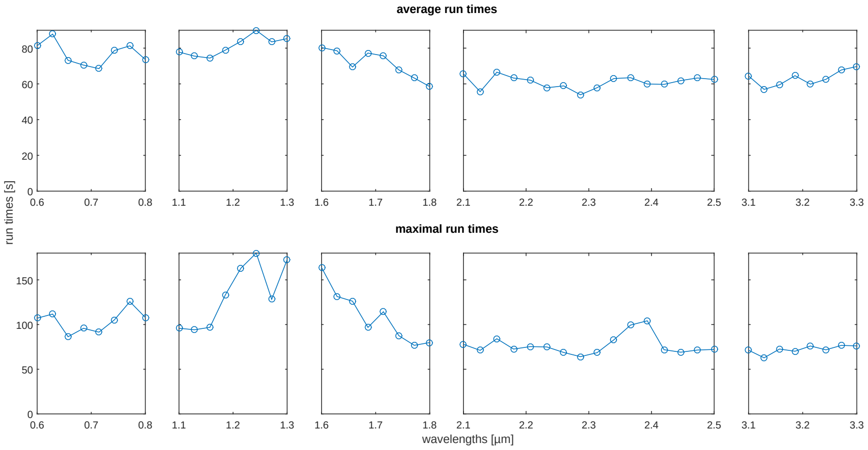




8.11 Results for H_2O

8.11.1 Results of the Genetic Algorithm and Algorithm 7





8.12 Conclusion

The results show that genetic algorithms are not superior to the sequential search strategy, both in run time and quality of the results.

Remark 8.12.1. For the application, i.e. the reconstruction of refractive indices from spectral measurements of homogeneously internally mixed particles, a sequential search strategy is feasible and sufficient. However if more complicated models for the particles under consideration are applied, e.g. a core-plus-shell models, the dimensionality of the regression problems for the refractive index retrieval grows. For the core-plus-shell model the complex refractive indices both of the particle cores and shells must be determined, therefore we deal here with a four-dimensional problem. In this case a sequential search strategy is infeasible. A remedy to this problem is given in [48], where the technique of simulated annealing is applied on the core-plus-shell model.

8.13 Numerical Study for Reconstructed Refractive Indices

We performed four numerical studies for two-component aerosols with log-normal, RRSB and Hedrih model size distributions as outlined in Chapter 5. The aerosol particles were assumed to be homogeneously internally mixed, such that only one effective refractive index was retrieved. One component of the simulated aerosols was H_2O with volume fractions of 0, 11, 22, 33, 44, 56, 67, 78, 89 and 100%. In the first two studies we simulated mixtures of H_2O and CsI, where we used the original aerosol component refractive indices for the first study. For the second study we used the average of the 100 regularized solutions from Section 8.1. We did the same for the third and fourth study, but here we simulated mixtures of H_2O and Ag. In the third study we utilized the original aerosol component refractive indices and for the fourth the average of the 100 regularized solutions from Section 8.1.

We applied the same reconstruction methods described in Chapter 5 under the same settings, i.e. for each reconstruction we generated 300 artificial noisy measurements for all 48 wavelengths, where the measurement error was simulated as additive zero-mean Gaussian noise. For each wavelength, the standard deviations were taken as 5% of the solutions of the forward problem. In Chapter 5 three different regularization methods, namely Tikhonov, minimal first differences and Twomey regularization, were compared and their results turned out to be very similar. Therefore we only used Tikhonov regularization in the following. The results for the first study were directly adopted from Chapter 5.

For every inversion we computed the L^2-error of the obtained reconstruction relative to the original size distribution and measured the total run time needed for the inversion. The computations were performed on a notebook with a 2.27 GHz CPU and 3.87 GB accessible primary memory.

8.14 Results for Mixtures of H_2O and CsI

8.14.1 Noise-free Refractive Indices

original H_2O volume percent	average L^2-errors (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	33.4495	37.6093	16.3524
11%	29.9288	31.7431	16.9073
22%	28.8686	30.3894	15.9820
33%	24.8269	28.7004	13.7607
44%	22.5902	24.2003	15.7127
56%	21.0371	21.4835	14.9451
67%	19.1780	19.7283	14.7707
78%	19.0107	17.0828	16.8178
89%	18.6772	14.2999	13.6688
100%	18.0467	11.6901	11.4425

original H_2O volume percent	average fraction deviation (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	11.0750	6.7000	4.8800
11%	7.6500	5.2200	6.3850
22%	6.3450	4.6400	4.1300
33%	4.4000	3.7100	4.3650
44%	3.5750	3.6200	3.3100
56%	3.2700	3.2050	3.0450
67%	2.5050	2.4650	1.8100
78%	2.3850	1.7750	2.7250
89%	2.0100	1.3250	1.9350
100%	1.2550	0.4650	0.7250

8.14.2 Noisy Refractive Indices

original H_2O volume percent	average L^2-errors (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	33.9736	37.2845	17.1607
11%	29.6079	30.1367	16.6469
22%	27.9444	30.3778	16.2839
33%	24.0048	31.1303	13.8363
44%	22.1817	24.0214	15.7084
56%	20.0710	20.9873	15.4322
67%	18.1141	20.9204	14.5403
78%	19.0356	15.5429	17.9794
89%	18.6182	13.4797	12.7544
100%	18.4766	11.0450	12.1062

original H_2O volume percent	average fraction deviation (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	12.5550	7.3150	3.7050
11%	6.9950	5.3950	6.7600
22%	6.2250	4.5100	4.7450
33%	4.0750	3.9700	4.1700
44%	4.0450	3.5550	3.4850
56%	2.9700	2.9500	3.3050
67%	2.2700	2.6250	2.2200
78%	2.4900	1.7400	2.9450
89%	2.0850	1.0850	1.8150
100%	1.3650	0.4200	0.6150

8.15 Results for Mixtures of H_2O and Ag

8.15.1 Noise-free Refractive Indices

original H_2O volume percent	average L^2-errors (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	63.1229	72.4297	57.4379
11%	49.8838	60.2852	53.6202
22%	40.7656	63.2290	39.5531
33%	56.6018	67.4771	48.7732
44%	54.8652	70.7186	53.1320
56%	45.8326	66.9322	37.4831
67%	37.6511	55.0038	24.7958
78%	30.5058	44.5771	15.5999
89%	24.7593	27.6324	17.1080
100%	18.6930	9.2744	11.2146

original H_2O volume percent	average fraction deviation (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	0	0.8100	0
11%	0.2650	0.3550	0.3150
22%	0.5200	2.6050	0.6250
33%	16.1300	15.8000	11.2300
44%	12.7550	13.5150	14.1100
56%	8.5500	9.6650	9.1500
67%	6.3700	7.0100	5.5600
78%	3.1150	3.2150	1.8550
89%	1.3750	1.3800	1.1300
100%	0.3600	0	0.0400

8.15.2 Noisy Refractive Indices

original H_2O volume percent	average L^2-errors (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	64.9875	69.1656	59.7450
11%	50.2221	60.3690	54.9776
22%	40.4326	74.3954	40.8412
33%	55.8575	65.4610	49.8945
44%	54.6888	70.3043	51.8924
56%	47.8998	63.6031	38.3468
67%	39.5417	58.6511	24.9801
78%	33.7818	43.1114	16.6046
89%	22.5077	26.3631	15.2523
100%	17.8098	10.3706	11.5581

original H_2O volume percent	average fraction deviation (%)		
	Log-Normal Distribution	RRSB Distribution	Hedrih Distribution
0%	0	0	0
11%	0.2700	0.4000	0.3400
22%	0.4950	3.1800	1.6650
33%	15.0350	15.0700	11.1850
44%	12.1250	13.9850	14.0800
56%	9.0900	8.7800	8.7850
67%	6.4150	7.3600	5.7950
78%	3.0850	3.3050	2.0650
89%	1.2050	1.4350	1.0300
100%	0.2550	0.0250	0.0800

8.16 Conclusion

The resuts of the first and second study only differ by ca. 3% at most and behave very similarly. The same is for the third and fourth study. These numerical results indicate that 100 FASP measurement sweeps consisting of 300 single measurements with an accuracy as in Section 8.1 are sufficient to determine aerosol refractive indices in such a quality, that they are suitable for particle size distribution reconstructions for two-component homogeneously internally mixed aerosols using the FASP. The particle radii of the three monodisperse aerosols generated for the refractive indices retrieval need to be 0.1 μm, 0.2 μm and 0.3 μm respectively.

Chapter 9

Summary and Outlook

In this work we derived inversion methods for single- and two-component aerosols, which satisfy our demands on run time and accuracy, i.e. a single inversion can be completed in under 30 seconds on a regular notebook and the reconstruction errors are below 100% for realistc noise levels. They are adaptive methods based on the statistical investigation of the residual. We established the convergence of our inversion results to the true sought-after particle size distribution for declining noise level theoretically and showed the convergence in numerical studies dealing with appropriate original particle size distributions. In the same numerical studies we compared our methods with existing inversion methods - among those a Monte Carlo method based on a Gibbs sampler - and found that it peformed better regarding the quality of the inversion results.

We also worked on the problem of retrieving aerosol refractive indices from measurements of mondisperse arosols. We obtained an effective reconstruction method by investigating the behavior of the reconstructions depending on the truncation index of the Mie extinction efficiency.

In this work we confined ourselves to internally mixed particles. For experimental applications this simplification might not always be suitable, because aerosol particles may have a more complicated structure, e.g. they may have a core-plus-shell structure, where core and shell consist of different non-mixing materials. It is still open if the methods derived in this work can be extended to those more complicated cases.

Appendix A

Derivatives of the Truncated Mie Efficiency Series

The problem of computing the partial derivatives of the truncated Mie efficiency series with respect to the real and imaginary parts n_{part} and k_{part} of the refractive index of the particle material can be reduced to the problem of computing the derivatives of its coefficient functions A_n and B_n. These coefficient functions in turn depend on the Mie coefficients a_n, b_n, c_n and d_n. Thus we first compute the derivatives of the Mie coefficients with respect to n_{part} and k_{part}, and then apply the product rule to obtain the derivatives of A_n and B_n. Similar to [49] we use recurrence relations of Bessel functions to compute the derivatives. Here we also give the second derivatives.

The Bessel functions $J_\alpha(z)$ and $Y_\alpha(z)$ for an arbitrary weight α fulfill the recurrence relations

$$\frac{d}{dz}\big(z^\alpha J_\alpha(z)\big) = z^\alpha J_{\alpha-1}(z) \quad \text{and} \quad \frac{d}{dz}\big(z^\alpha Y_\alpha(z)\big) = z^\alpha Y_{\alpha-1}(z), \tag{A.0.1}$$

cf. [50]. For the Bessel functions occurring in the Riccati-Bessel-functions $\xi_n(z)$ and $\psi_n(z)$ follows from this with the weight $\alpha = n + \frac{1}{2}$ that

$$\dot{\xi}_n(z) = \sqrt{\tfrac{\pi}{2}}\sqrt{z}\left(J_{n-\frac{1}{2}}(z) - \frac{n}{z}J_{n+\frac{1}{2}}(z)\right) \tag{A.0.2}$$

$$\text{and} \quad \dot{\psi}_n(z) = \sqrt{\tfrac{\pi}{2}}\sqrt{z}\left(J_{n-\frac{1}{2}}(z) - \frac{n}{z}J_{n+\frac{1}{2}}(z)\right) + \sqrt{\tfrac{\pi}{2}}\sqrt{z}\left(Y_{n-\frac{1}{2}}(z) - \frac{n}{z}Y_{n+\frac{1}{2}}(z)\right) i \tag{A.0.3}$$

for $z \neq 0$.

We apply (A.0.1) a second time to get $\ddot{\xi}_n(z)$, which yields

$$\ddot{\xi}_n(z) = \sqrt{\tfrac{\pi}{2}}\frac{\sqrt{z}}{z^2}\left(n(n+1)J_{n+\frac{1}{2}}(z) + (1-2n)J_{n-\frac{1}{2}}(z) + z^2 J_{n-\frac{3}{2}}(z)\right).$$

For an arbitrary weight α we have the recurrence relation

$$J_{\alpha-1}(z) = \frac{2\alpha}{z}J_\alpha(z) - J_{\alpha+1}(z),$$

see [50], and we use it to eliminate the term $J_{n-\frac{3}{2}}(z)$ in the expression for $\ddot{\xi}_n(z)$. Then also $J_{n-\frac{1}{2}}(z)$ cancels out, such that we obtain the representation

$$\ddot{\xi}_n(z) = \sqrt{\tfrac{\pi}{2}}\frac{\sqrt{z}}{z^2}\left(n(n+1) - z^2\right) J_{n+\frac{1}{2}}(z) \tag{A.0.4}$$

only involving $J_{n+\frac{1}{2}}(z)$. Applying (A.0.1) on (A.0.4) gives

$$\dddot{\xi}_n(z) = \sqrt{\tfrac{\pi}{2}}\frac{\sqrt{z}}{z^3}\left(\left(n(n+1) - z^2\right) zJ_{n-\frac{1}{2}}(z) + \left(z^2 - n^2 - 3n - 2\right) nJ_{n+\frac{1}{2}}(z)\right). \tag{A.0.5}$$

The Bessel function values

$$\begin{aligned} & J_{0+\frac{1}{2}}(z_{med}), ..., J_{N_{trunc}+\frac{1}{2}}(z_{med}), \quad Y_{0+\frac{1}{2}}(z_{med}), ..., Y_{N_{trunc}+\frac{1}{2}}(z_{med}) \\ \text{and} \quad & J_{0+\frac{1}{2}}(z_{part}), ..., J_{N_{trunc}+\frac{1}{2}}(z_{part}). \end{aligned}$$

already computed for a function evaluation of the truncated Mie extinction efficiency can be reused for their derivatives.

At last we recapitulate the Cauchy-Riemann equations in its complex form. For a holomorphic function $f : \mathbb{C} \to \mathbb{C}$ with $f(z) = f(x + iy) = u(x, y) + iv(x, y)$ holds

$$\dot{f}(z) = \frac{d}{dz}f(z) = \frac{\partial}{\partial x}f(x + iy) = -i\frac{\partial}{\partial y}f(x + iy). \tag{A.0.6}$$

From this follows

$$u_x = \mathrm{Re}\big(\dot{f}(z)\big), \quad u_y = -\mathrm{Im}\big(\dot{f}(z)\big), \quad v_x = \mathrm{Im}\big(\dot{f}(z)\big) \quad \text{and} \quad v_y = \mathrm{Re}\big(\dot{f}(z)\big). \tag{A.0.7}$$

Now everything is prepared to differentiate the squared magnitudes of the Mie coefficients a_n, b_n, c_n and d_n with respect to n_{part} and k_{part}. These derivatives will be used to compute the derivatives of the truncated Mie extinction efficiency with the chain rule.

A.1 Derivatives of $|a_n|^2$

First Derivatives

First we write the squared norm of the Mie coefficient a_n as $|a_n|^2 = a_n\overline{a_n}$, which gives

$$\begin{aligned} \frac{\partial}{\partial n_{part}} |a_n|^2 &= \left(\frac{\partial}{\partial n_{part}}a_n\right)\overline{a_n} + a_n\left(\frac{\partial}{\partial n_{part}}\overline{a_n}\right) \\ \text{and} \quad \frac{\partial}{\partial k_{part}} |a_n|^2 &= \left(\frac{\partial}{\partial k_{part}}a_n\right)\overline{a_n} + a_n\left(\frac{\partial}{\partial k_{part}}\overline{a_n}\right). \end{aligned}$$

We write

$$\begin{aligned} a_n = \frac{E_1}{D_1} \quad \text{with} \quad & E_1 := m_{part}\dot{\xi}_n(z_{med})\xi_n(z_{part}) - m_{med}\xi_n(z_{med})\dot{\xi}_n(z_{part}) \\ \text{and} \quad & D_1 := m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{part}) - m_{med}\psi_n(z_{med})\dot{\xi}_n(z_{part}), \end{aligned}$$

which yields

$$\begin{aligned} & \frac{d}{dm_{part}}a_n = \frac{1}{D_1^2}\left(\left(\frac{d}{dm_{part}}E_1\right)D_1 - E_1\left(\frac{d}{dm_{part}}D_1\right)\right) \\ \text{with} \quad & \frac{d}{dm_{part}}E_1 = \dot{\xi}_n(z_{med})\xi_n(z_{part}) + m_{part}\dot{\xi}_n(z_{med})\rho\dot{\xi}_n(z_{part}) - m_{med}\xi_n(z_{med})\rho\ddot{\xi}_n(z_{part}) \end{aligned}$$

$$\text{and}\quad \frac{d}{dm_{part}}D_1 = \dot{\psi}_n(z_{med})\xi_n(z_{part}) + m_{part}\dot{\psi}_n(z_{med})\rho\dot{\xi}_n(z_{part}) - m_{med}\psi_n(z_{med})\rho\ddot{\xi}_n(z_{part}).$$

Furthermore follow from (A.0.6) the relations

$$\begin{aligned} \frac{\partial}{\partial n_{part}}a_n &= \frac{d}{dm_{part}}a_n \\ \text{and}\quad \frac{\partial}{\partial k_{part}}a_n &= \left(\frac{d}{dm_{part}}a_n\right)i. \end{aligned}$$

Although $\overline{a_n}$ is not holomorphic with respect to m_{part}, we can still compute the partial derivatives $\frac{\partial}{\partial n_{part}}\overline{a_n}$ and $\frac{\partial}{\partial k_{part}}\overline{a_n}$. We obtain using (A.0.7) the relations

$$\begin{aligned} \frac{\partial}{\partial n_{part}}\overline{a_n} &= \overline{\frac{\partial}{\partial n_{part}}a_n} \\ \text{and}\quad \frac{\partial}{\partial k_{part}}\overline{a_n} &= \overline{\frac{\partial}{\partial k_{part}}a_n}. \end{aligned}$$

This completes the computations of $\frac{\partial}{\partial n_{part}}\left|a_n\right|^2$ and $\frac{\partial}{\partial k_{part}}\left|a_n\right|^2$.

Second Derivatives

We have that

$$\begin{aligned} \frac{\partial^2}{\partial n_{part}^2}\left|a_n\right|^2 &= \left(\frac{\partial^2}{\partial n_{part}^2}a_n\right)\overline{a_n} + 2\left(\frac{\partial}{\partial n_{part}}a_n\right)\left(\frac{\partial}{\partial n_{part}}\overline{a_n}\right) + a_n\left(\frac{\partial^2}{\partial n_{part}^2}\overline{a_n}\right), \\ \frac{\partial^2}{\partial n_{part}\partial k_{part}}\left|a_n\right|^2 &= \left(\frac{\partial^2}{\partial n_{part}\partial k_{part}}a_n\right)\overline{a_n} + \left(\frac{\partial}{\partial n_{part}}a_n\right)\left(\frac{\partial}{\partial k_{part}}\overline{a_n}\right) \\ &\quad + \left(\frac{\partial}{\partial k_{part}}a_n\right)\left(\frac{\partial}{\partial n_{part}}\overline{a_n}\right) + a_n\left(\frac{\partial^2}{\partial n_{part}\partial k_{part}}\overline{a_n}\right) \\ \text{and}\quad \frac{\partial}{\partial k_{part}}\left|a_n\right|^2 &= \left(\frac{\partial^2}{\partial k_{part}^2}a_n\right)\overline{a_n} + 2\left(\frac{\partial}{\partial k_{part}}a_n\right)\left(\frac{\partial}{\partial k_{part}}\overline{a_n}\right) + a_n\left(\frac{\partial^2}{\partial k_{part}^2}\overline{a_n}\right). \end{aligned}$$

In order to obtain the partial derivatives of a_n and $\overline{a_n}$ we first compute

$$\begin{aligned} \frac{d^2}{dm_{part}^2}a_n = \frac{1}{D_1^3}\Bigg(&\left(\left(\frac{d^2}{dm_{part}^2}E_1\right)D_1 - E_1\left(\frac{d^2}{dm_{part}^2}D_1\right)\right)D_1 \\ &- 2\left(\left(\frac{d}{dm_{part}}E_1\right)D_1 - E_1\left(\frac{d}{dm_{part}}D_1\right)\right)\left(\frac{d}{dm_{part}}D_1\right)\Bigg) \end{aligned}$$

$$\begin{aligned} \text{with}\quad \frac{d^2}{dm_{part}^2}E_1 &= 2\dot{\xi}_n(z_{med})\rho\dot{\xi}_n(z_{part}) + m_{part}\dot{\xi}_n(z_{med})\rho^2\ddot{\xi}_n(z_{part}) \\ &\quad - m_{med}\xi_n(z_{med})\rho^2\dddot{\xi}_n(z_{part}) \\ \text{and}\quad \frac{d^2}{dm_{part}^2}D_1 &= 2\dot{\psi}_n(z_{med})\rho\dot{\xi}_n(z_{part}) + m_{part}\dot{\psi}_n(z_{med})\rho^2\ddot{\xi}_n(z_{part}) \\ &\quad - m_{med}\psi_n(z_{med})\rho^2\dddot{\xi}_n(z_{part}). \end{aligned}$$

Then (A.0.6) and (A.0.7) give

$$\begin{aligned}
\frac{\partial^2}{\partial n_{part}^2} a_n &= \frac{d^2}{dm_{part}^2} a_n, \\
\frac{\partial^2}{\partial n_{part} \partial k_{part}} a_n &= \left(\frac{d^2}{dm_{part}^2} a_n \right) i, \\
\frac{\partial^2}{\partial k_{part}^2} a_n &= -\frac{d^2}{dm_{part}^2} a_n, \\
\frac{\partial^2}{\partial n_{part}^2} \overline{a_n} &= \overline{\frac{\partial^2}{\partial n_{part}^2} a_n}, \\
\frac{\partial^2}{\partial n_{part} \partial k_{part}} \overline{a_n} &= \overline{\frac{\partial^2}{\partial n_{part} \partial k_{part}} a_n}, \\
\text{and} \quad \frac{\partial^2}{\partial k_{part}^2} \overline{a_n} &= \overline{\frac{\partial^2}{\partial k_{part}^2} a_n}.
\end{aligned}$$

This completes the computations of the second partial derivatives of $|a_n|^2$ with respect to n_{part} and k_{part}.

A.2 Derivatives of $|b_n|^2$

First Derivatives

The compuations are completely analogous to those for $|a_n|^2$, but here we use

$$\begin{aligned}
b_n = \frac{E_2}{D_2} \quad \text{with} \quad & E_2 := m_{part} \xi_n(z_{med}) \dot{\xi}_n(z_{part}) - m_{med} \dot{\xi}_n(z_{med}) \xi_n(z_{part}) \\
\text{and} \quad & D_2 := m_{part} \psi_n(z_{med}) \dot{\xi}_n(z_{part}) - m_{med} \dot{\psi}_n(z_{med}) \xi_n(z_{part}),
\end{aligned}$$

and

$$\begin{aligned}
\frac{d}{dm_{part}} E_2 &= \xi_n(z_{med}) \dot{\xi}_n(z_{part}) + m_{part} \xi_n(z_{med}) \rho \ddot{\xi}_n(z_{part}) - m_{med} \dot{\xi}_n(z_{med}) \rho \dot{\xi}_n(z_{part}) \\
\text{and} \quad \frac{d}{dm_{part}} D_2 &= \psi_n(z_{med}) \dot{\xi}_n(z_{part}) + m_{part} \psi_n(z_{med}) \rho \ddot{\xi}_n(z_{part}) - m_{med} \dot{\psi}_n(z_{med}) \rho \dot{\xi}_n(z_{part}).
\end{aligned}$$

Second Derivatives

Again the computations are analogous to those for $|a_n|^2$. Here we need

$$\begin{aligned}
\frac{d^2}{dm_{part}^2} E_2 &= 2 \xi_n(z_{med}) \rho \ddot{\xi}_n(z_{part}) + m_{part} \xi_n(z_{med}) \rho^2 \dddot{\xi}_n(z_{part}) \\
&\quad - m_{med} \dot{\xi}_n(z_{med}) \rho^2 \ddot{\xi}_n(z_{part}) \\
\text{and} \quad \frac{d^2}{dm_{part}^2} D_2 &= 2 \psi_n(z_{med}) \rho \ddot{\xi}_n(z_{part}) + m_{part} \psi_n(z_{med}) \rho^2 \dddot{\xi}_n(z_{part}) \\
&\quad - m_{med} \dot{\psi}_n(z_{med}) \rho^2 \ddot{\xi}_n(z_{part}).
\end{aligned}$$

A.3 Derivatives of $|c_n|^2$

First Derivatives

Here we use

$$\begin{aligned} c_n = \frac{E_3}{D_3} \quad \text{with} \quad & E_3 := m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{med}) - m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{med}) \\ \text{and} \quad & D_3 := m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{part}) - m_{med}\dot{\psi}_n(z_{med})\xi_n(z_{part}), \end{aligned}$$

and the derivatives

$$\begin{aligned} & \frac{d}{dm_{part}} E_3 = \psi_n(z_{med})\dot{\xi}_n(z_{med}) - \dot{\psi}_n(z_{med})\xi_n(z_{med}) \\ \text{and} \quad & \frac{d}{dm_{part}} D_3 = \psi_n(z_{med})\dot{\xi}_n(z_{part}) + m_{part}\psi_n(z_{med})\rho\ddot{\xi}_n(z_{part}) \\ & \qquad - m_{med}\dot{\psi}_n(z_{med})\rho\dot{\xi}_n(z_{part}). \end{aligned}$$

Second Derivatives

Here we need

$$\begin{aligned} & \frac{d^2}{dm_{part}^2} E_3 = 0 \\ \text{and} \quad & \frac{d^2}{dm_{part}^2} D_3 = 2\psi_n(z_{med})\rho\ddot{\xi}_n(z_{part}) + m_{part}\psi_n(z_{med})\rho^2\dddot{\xi}_n(z_{part}) \\ & \qquad - m_{med}\dot{\psi}_n(z_{med})\rho^2\ddot{\xi}_n(z_{part}). \end{aligned}$$

A.4 Derivatives of $|d_n|^2$

First Derivatives

We make use of

$$\begin{aligned} d_n = \frac{E_4}{D_4} \quad \text{with} \quad & E_4 := m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{med}) - m_{part}\psi_n(z_{med})\dot{\xi}_n(z_{med}) \\ \text{and} \quad & D_4 := m_{part}\dot{\psi}_n(z_{med})\xi_n(z_{part}) - m_{med}\psi_n(z_{med})\dot{\xi}_n(z_{part}), \end{aligned}$$

and the derivatives

$$\begin{aligned} & \frac{d}{dm_{part}} E_4 = \dot{\psi}_n(z_{med})\xi_n(z_{med}) - \psi_n(z_{med})\dot{\xi}_n(z_{med}) \\ \text{and} \quad & \frac{d}{dm_{part}} D_4 = \dot{\psi}_n(z_{med})\xi_n(z_{part}) + m_{part}\dot{\psi}_n(z_{med})\rho\dot{\xi}_n(z_{part}) \\ & \qquad - m_{med}\psi_n(z_{med})\rho\ddot{\xi}_n(z_{part}). \end{aligned}$$

Second Derivatives

At last we need

$$
\begin{aligned}
&\frac{d^2}{dm_{part}^2}E_4 = 0 \\
\text{and} \quad &\frac{d^2}{dm_{part}^2}D_4 = 2\dot{\psi}_n(z_{med})\rho\dot{\xi}_n(z_{part}) + m_{part}\dot{\psi}_n(z_{med})\rho^2\ddot{\xi}_n(z_{part}) \\
&\qquad - m_{med}\psi_n(z_{med})\rho^2\dddot{\xi}_n(z_{part}).
\end{aligned}
$$

A.5 Derivatives of $\mathrm{Im}\,(A_n)$

First Derivatives

For a holomorphic function $f(x+iy)$ we can easily deduce from (A.0.7) that

$$
\frac{\partial}{\partial x}\mathrm{Im}\,(f(x+iy)) = \mathrm{Im}\left(\frac{\partial}{\partial x}f(x+iy)\right) \quad \text{and} \quad \frac{\partial}{\partial y}\mathrm{Im}\,(f(x+iy)) = \mathrm{Im}\left(\frac{\partial}{\partial y}f(x+iy)\right).
$$

This gives with respect to (A.0.6)

$$
\begin{aligned}
&\frac{\partial}{\partial n_{part}}\mathrm{Im}\,(A_n) = \mathrm{Im}\left(\frac{\partial}{\partial n_{part}}A_n\right) \\
\text{and} \quad &\frac{\partial}{\partial k_{part}}\mathrm{Im}\,(A_n) = \mathrm{Im}\left(\frac{\partial}{\partial k_{part}}A_n\right)
\end{aligned}
$$

Therefore we only need to compute the partial derivatives of A_n. The first derivative with respect to n_{part} is given by

$$
\begin{aligned}
\frac{\partial}{\partial n_{part}}A_n = \frac{l}{2\pi}\Bigg(&\frac{\partial}{\partial n_{part}}\left(|c_n|^2\right)U_1 + |c_n|^2\frac{\partial}{\partial n_{part}}U_1 \\
&- \frac{\partial}{\partial n_{part}}\left(|d_n|^2\right)U_2 - |d_n|^2\frac{\partial}{\partial n_{part}}U_2\Bigg),
\end{aligned}
$$

$$
\text{where} \quad U_1 = \frac{\xi_n(z_{part})\overline{\dot{\xi}_n(z_{part})}}{m_{part}},
$$

$$
\frac{\partial}{\partial n_{part}}U_1 = \left(\frac{1}{m_{part}}\rho\dot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\xi_n(z_{part})\right)\overline{\dot{\xi}_n(z_{part})} + \frac{1}{m_{part}}\xi_n(z_{part})\rho\overline{\ddot{\xi}_n(z_{part})}
$$

$$
\text{and} \quad U_2 = \frac{\dot{\xi}_n(z_{part})\overline{\xi_n(z_{part})}}{m_{part}},
$$

$$\frac{\partial}{\partial n_{part}} U_2 = \left(\frac{1}{m_{part}} \rho \ddot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2} \dot{\xi}_n(z_{part}) \right) \overline{\xi_n(z_{part})} + \frac{1}{m_{part}} \dot{\xi}_n(z_{part}) \rho \overline{\dot{\xi}_n(z_{part})}$$

Analogously we obtain

$$\begin{aligned} \frac{\partial}{\partial k_{part}} A_n = \frac{l}{2\pi} \Bigg(& \frac{\partial}{\partial k_{part}} \left(|c_n|^2 \right) U_1 + |c_n|^2 \frac{\partial}{\partial k_{part}} U_1 \\ & - \frac{\partial}{\partial n_{part}} \left(|d_n|^2 \right) U_2 - |d_n|^2 \frac{\partial}{\partial k_{part}} U_2 \Bigg), \end{aligned}$$

where
$$\begin{aligned} \frac{\partial}{\partial k_{part}} U_1 = & \left(\frac{1}{m_{part}} \rho \dot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2} \xi_n(z_{part}) \right) \overline{\dot{\xi}_n(z_{part})} \ i \\ & + \frac{1}{m_{part}} \xi_n(z_{part}) \rho \overline{\left(\ddot{\xi}_n(z_{part}) \right) i} \end{aligned}$$

and
$$\begin{aligned} \frac{\partial}{\partial k_{part}} U_2 = & \left(\frac{1}{m_{part}} \rho \ddot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2} \dot{\xi}_n(z_{part}) \right) \overline{\xi_n(z_{part})} \ i \\ & + \frac{1}{m_{part}} \dot{\xi}_n(z_{part}) \rho \overline{\left(\dot{\xi}_n(z_{part}) \right) i}. \end{aligned}$$

Second Derivatives

We start with

$$\begin{aligned} \frac{\partial^2}{\partial n_{part}^2} \mathrm{Im}\left(A_n\right) &= \mathrm{Im}\left(\frac{\partial^2}{\partial n_{part}^2} A_n \right), \\ \frac{\partial^2}{\partial n_{part} \partial k_{part}} \mathrm{Im}\left(A_n\right) &= \mathrm{Im}\left(\frac{\partial^2}{\partial n_{part} \partial k_{part}} A_n \right), \\ \text{and} \quad \frac{\partial^2}{\partial k_{part}^2} \mathrm{Im}\left(A_n\right) &= \mathrm{Im}\left(\frac{\partial^2}{\partial k_{part}^2} A_n \right). \end{aligned}$$

Again we only have to compute the second partial derivatives of A_n. We have

$$\begin{aligned} \frac{\partial^2}{\partial n_{part}^2} A_n = \frac{l}{2\pi} \Big(& \frac{\partial^2}{\partial n_{part}^2} \left(|c_n|^2\right) U_1 + 2 \frac{\partial}{\partial n_{part}} \left(|c_n|^2\right) \frac{\partial}{\partial n_{part}} U_1 + |c_n|^2 \frac{\partial^2}{\partial n_{part}^2} U_1 \\ & \frac{\partial^2}{\partial n_{part}^2} \left(|d_n|^2\right) U_2 + 2 \frac{\partial}{\partial n_{part}} \left(|d_n|^2\right) \frac{\partial}{\partial n_{part}} U_2 + |d_n|^2 \frac{\partial^2}{\partial n_{part}^2} U_2 \Big). \end{aligned}$$

The two new terms needed here are

$$\begin{aligned} \frac{\partial^2}{\partial n_{part}^2} U_1 = & \left(\frac{1}{m_{part}} \rho^2 \ddot{\xi}_n(z_{part}) - \frac{2}{m_{part}^2} \rho \dot{\xi}_n(z_{part}) + \frac{2}{m_{part}^3} \xi_n(z_{part}) \right) \overline{\dot{\xi}_n(z_{part})} \\ & + 2 \left(\frac{1}{m_{part}} \rho \dot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2} \xi_n(z_{part}) \right) \rho \overline{\ddot{\xi}_n(z_{part})} \end{aligned}$$

$$+\frac{1}{m_{part}}\rho^2\xi_n(z_{part})\overline{\dddot{\xi}_n(z_{part})}$$

and
$$\begin{aligned}\frac{\partial^2}{\partial n_{part}^2}U_2 = &\left(\frac{1}{m_{part}}\rho^2\dddot{\xi}_n(z_{part}) - \frac{2}{m_{part}^2}\rho\ddot{\xi}_n(z_{part}) + \frac{2}{m_{part}^3}\dot{\xi}_n(z_{part})\right)\overline{\xi_n(z_{part})} \\ &+ 2\left(\frac{1}{m_{part}}\rho\ddot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\dot{\xi}_n(z_{part})\right)\overline{\rho\dot{\xi}_n(z_{part})} \\ &+ \frac{1}{m_{part}}\rho^2\dot{\xi}_n(z_{part})\overline{\ddot{\xi}_n(z_{part})}.\end{aligned}$$

The mixed derivative is given by

$$\begin{aligned}\frac{\partial^2}{\partial n_{part}\partial k_{part}}A_n = \frac{l}{2\pi}\Big(&\frac{\partial^2}{\partial n_{part}\partial k_{part}}\left(|c_n|^2\right)U_1 + \frac{\partial}{\partial n_{part}}\left(|c_n|^2\right)\frac{\partial}{\partial k_{part}}U_1 \\ &+\frac{\partial}{\partial k_{part}}\left(|c_n|^2\right)\frac{\partial}{\partial n_{part}}U_1 + |c_n|^2\frac{\partial^2}{\partial n_{part}\partial k_{part}}U_1 \\ &\frac{\partial^2}{\partial n_{part}\partial k_{part}}\left(|d_n|^2\right)U_2 + \frac{\partial}{\partial n_{part}}\left(|d_n|^2\right)\frac{\partial}{\partial k_{part}}U_2 \\ &+\frac{\partial}{\partial k_{part}}\left(|d_n|^2\right)\frac{\partial}{\partial n_{part}}U_2 + |d_n|^2\frac{\partial^2}{\partial n_{part}\partial k_{part}}U_2\Big).\end{aligned}$$

Here we have to complete the computations with

$$\begin{aligned}\frac{\partial^2}{\partial n_{part}\partial k_{part}}U_1 = &\left(\frac{1}{m_{part}}\rho^2\ddot{\xi}_n(z_{part}) - \frac{2}{m_{part}^2}\rho\dot{\xi}_n(z_{part})\ i + \frac{2}{m_{part}^3}\xi_n(z_{part}\ i\right)\overline{\dot{\xi}_n(z_{part})} \\ &+\left(\frac{1}{m_{part}}\rho\dot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\xi_n(z_{part})\right)\overline{\rho\ddot{\xi}_n(z_{part})\ i} \\ &+\left(\frac{1}{m_{part}}\rho\dot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\xi_n(z_{part})\right)\overline{\rho\ddot{\xi}_n(z_{part})}\ i \\ &+\frac{1}{m_{part}}\rho^2\xi_n(z_{part})\overline{\dddot{\xi}_n(z_{part})\ i}\end{aligned}$$

and
$$\begin{aligned}\frac{\partial^2}{\partial n_{part}\partial k_{part}}U_2 = &\left(\frac{1}{m_{part}}\rho^2\dddot{\xi}_n(z_{part}) - \frac{2}{m_{part}^2}\rho\ddot{\xi}_n(z_{part})\ i + \frac{2}{m_{part}^3}\dot{\xi}_n(z_{part}\ i\right)\overline{\xi_n(z_{part})} \\ &+\left(\frac{1}{m_{part}}\rho\ddot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\dot{\xi}_n(z_{part})\right)\overline{\rho\dot{\xi}_n(z_{part})\ i} \\ &+\left(\frac{1}{m_{part}}\rho\ddot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\dot{\xi}_n(z_{part})\right)\overline{\rho\dot{\xi}_n(z_{part})}\ i \\ &+\frac{1}{m_{part}}\rho^2\dot{\xi}_n(z_{part})\overline{\ddot{\xi}_n(z_{part})\ i}.\end{aligned}$$

Finally the second derivative with respect to the imaginary part is given by

$$\frac{\partial^2}{\partial k_{part}^2}A_n = \frac{l}{2\pi}\Big(\frac{\partial^2}{\partial k_{part}^2}\left(|c_n|^2\right)U_1 + 2\frac{\partial}{\partial k_{part}}\left(|c_n|^2\right)\frac{\partial}{\partial k_{part}}U_1 + |c_n|^2\frac{\partial^2}{\partial k_{part}^2}U_1$$

$$\frac{\partial^2}{\partial k_{part}^2}\left(|d_n|^2\right)U_2 + 2\frac{\partial}{\partial k_{part}}\left(|d_n|^2\right)\frac{\partial}{\partial k_{part}}U_2 + |d_n|^2\frac{\partial^2}{\partial k_{part}^2}U_2\Big).$$

Here we need

$$\begin{aligned}\frac{\partial^2}{\partial k_{part}^2}U_1 = &\left(-\frac{1}{m_{part}}\rho^2\ddot{\xi}_n(z_{part}) + \frac{2}{m_{part}^2}\rho\dot{\xi}_n(z_{part}) - \frac{2}{m_{part}^3}\xi_n(z_{part})\right)\overline{\dot{\xi}_n(z_{part})}\\ &+ 2\left(\frac{1}{m_{part}}\rho\dot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\xi_n(z_{part})\right)\rho\overline{\ddot{\xi}_n(z_{part})\ i}\ i\\ &- \frac{1}{m_{part}}\rho^2\xi_n(z_{part})\overline{\dddot{\xi}_n(z_{part})}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial k_{part}^2}U_2 = &\left(-\frac{1}{m_{part}}\rho^2\dddot{\xi}_n(z_{part}) + \frac{2}{m_{part}^2}\rho\ddot{\xi}_n(z_{part}) - \frac{2}{m_{part}^3}\dot{\xi}_n(z_{part})\right)\overline{\xi_n(z_{part})}\\ &+ 2\left(\frac{1}{m_{part}}\rho\ddot{\xi}_n(z_{part}) - \frac{1}{m_{part}^2}\dot{\xi}_n(z_{part})\right)\rho\overline{\dot{\xi}_n(z_{part})\ i}\ i\\ &- \frac{1}{m_{part}}\rho^2\dot{\xi}_n(z_{part})\overline{\ddot{\xi}_n(z_{part})}.\end{aligned}$$

A.6 Derivatives of $\operatorname{Im}(B_n)$

First Derivatives

The derivatives of $\operatorname{Im}(B_n)$ are much easier to compute, since the dependence on n_{part} and k_{part} lies only in $|a_n|^2$ and $|b_n|^2$ here. Here we also begin with

$$\frac{\partial}{\partial n_{part}}\operatorname{Im}(B_n) = \operatorname{Im}\left(\frac{\partial}{\partial n_{part}}B_n\right)$$

and

$$\frac{\partial}{\partial k_{part}}\operatorname{Im}(B_n) = \operatorname{Im}\left(\frac{\partial}{\partial k_{part}}B_n\right).$$

Therefore we need

$$\frac{\partial}{\partial n_{part}}B_n = \frac{l}{2\pi}\left(\frac{\partial}{\partial n_{part}}\left(|a_n|^2\right)\frac{\dot{\psi}_n(z_{med})\overline{\psi_n(z_{med})}}{m_{med}} - \frac{\partial}{\partial n_{part}}\left(|b_n|^2\right)\frac{\psi_n(z_{med})\overline{\dot{\psi}_n(z_{med})}}{m_{med}}\right)$$

and

$$\frac{\partial}{\partial k_{part}}B_n = \frac{l}{2\pi}\left(\frac{\partial}{\partial k_{part}}\left(|a_n|^2\right)\frac{\dot{\psi}_n(z_{med})\overline{\psi_n(z_{med})}}{m_{med}} - \frac{\partial}{\partial k_{part}}\left(|b_n|^2\right)\frac{\psi_n(z_{med})\overline{\dot{\psi}_n(z_{med})}}{m_{med}}\right).$$

Second Derivatives

The last derivatives needed are

$$\begin{aligned}
\frac{\partial^2}{\partial n_{part}^2}\mathrm{Im}\left(B_n\right) &= \mathrm{Im}\left(\frac{\partial^2}{\partial n_{part}^2}B_n\right), \\
\frac{\partial^2}{\partial n_{part}\partial k_{part}}\mathrm{Im}\left(B_n\right) &= \mathrm{Im}\left(\frac{\partial^2}{\partial n_{part}\partial k_{part}}B_n\right), \\
\text{and} \quad \frac{\partial^2}{\partial k_{part}^2}\mathrm{Im}\left(B_n\right) &= \mathrm{Im}\left(\frac{\partial^2}{\partial k_{part}^2}B_n\right).
\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial^2}{\partial n_{part}^2}B_n = \frac{l}{2\pi}\Bigg(&\frac{\partial^2}{\partial n_{part}^2}\left(|a_n|^2\right)\frac{\dot{\psi}_n(z_{med})\overline{\psi_n(z_{med})}}{m_{med}} \\
&-\frac{\partial^2}{\partial n_{part}^2}\left(|b_n|^2\right)\frac{\psi_n(z_{med})\overline{\dot{\psi}_n(z_{med})}}{m_{med}}\Bigg), \\
\frac{\partial^2}{\partial n_{part}\partial k_{part}}B_n = \frac{l}{2\pi}\Bigg(&\frac{\partial^2}{\partial n_{part}\partial k_{part}}\left(|a_n|^2\right)\frac{\dot{\psi}_n(z_{med})\overline{\psi_n(z_{med})}}{m_{med}} \\
&-\frac{\partial^2}{\partial n_{part}\partial k_{part}}\left(|b_n|^2\right)\frac{\psi_n(z_{med})\overline{\dot{\psi}_n(z_{med})}}{m_{med}}\Bigg) \\
\text{and} \quad \frac{\partial^2}{\partial k_{part}^2}B_n = \frac{l}{2\pi}\Bigg(&\frac{\partial^2}{\partial k_{part}^2}\left(|a_n|^2\right)\frac{\dot{\psi}_n(z_{med})\overline{\psi_n(z_{med})}}{m_{med}} \\
&-\frac{\partial^2}{\partial k_{part}^2}\left(|b_n|^2\right)\frac{\psi_n(z_{med})\overline{\dot{\psi}_n(z_{med})}}{m_{med}}\Bigg).
\end{aligned}$$

Appendix B

Basic Mathematical Tools

B.1 Optimization

Theorem B.1.1. *(Karush-Kuhn-Tucker Theorem) For a minimization problem of the form*

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\, f(\boldsymbol{x}) \quad s.t. \quad \boldsymbol{h}(\boldsymbol{x}) = 0, \quad \boldsymbol{g}(\boldsymbol{x}) \leq 0$$

with $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^p$ exist vectors $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \in \mathbb{R}^p$ with

$$\begin{aligned}
&\boldsymbol{\mu} \geq 0 \\
&\nabla f(\boldsymbol{x}^*) + \boldsymbol{\lambda}^T \mathrm{Jac}_{\boldsymbol{g}}(\boldsymbol{x}^*)^T + \boldsymbol{\mu}^T \mathrm{Jac}_{\boldsymbol{h}}(\boldsymbol{x}^*)^T = 0 \\
&\boldsymbol{\mu}^T \boldsymbol{g} f(\boldsymbol{x}^*) = 0,
\end{aligned}$$

if the matrix $\left(\mathrm{Jac}_{\boldsymbol{h}}(\boldsymbol{x}^*)^T,\ \nabla \boldsymbol{g}_{j\in J(\boldsymbol{x}^*)}(\boldsymbol{x}^*)\right)$ has full rank, where

$$J(\boldsymbol{x}^*) = \{j \mid 1 \leq j \leq p,\ \boldsymbol{g}_j(\boldsymbol{x}^*) = 0\}$$

is the so-called set of active inequality constraints.

B.2 Probability Theory

Definition B.2.1. *Let Ω be a set. A σ-**algebra** over Ω is a system $\mathcal{F}(\Omega)$ of subsets of Ω with $\Omega \in \mathcal{F}(\Omega)$, $\Omega \setminus A \in \mathcal{F}(\Omega)$ for all $A \in \mathcal{F}(\Omega)$ and $\cup_{i=1}^{\infty} A_i \in \mathcal{F}(\Omega)$ for all countable sequences $A_1, A_2, ...$ in $\mathcal{F}(\Omega)$.*

Definition B.2.2. *A **measure** μ on Ω is a function $\mu : \mathcal{F}(\Omega) \to [0, \infty)$ with the properties $\mu(\emptyset) = 0$, $\mu(A) \geq 0$ for all $A \in \mathcal{F}(\Omega)$ and $\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ for all pairwise disjoint sequences $A_1, A_2, ...$ in $\mathcal{F}(\Omega)$.*

Definition B.2.3. *A measure p on Ω that additionally fulfills $p(\Omega) = 1$ is called a **probability measure**.*

Definition B.2.4. *A triple Ω, $\mathcal{F}(\Omega)$, p is called a **probability space**, if Ω is a set, $\mathcal{F}(\Omega)$ a σ-algebra over Ω and p a probability measure on $\mathcal{F}(\Omega)$.*

Definition B.2.5. *A* ***random variable*** *on a probability space* Ω, $\mathcal{F}(\Omega)$, p *is a function* $X : \Omega \to \mathbb{R}$ *that satisfies* $\{\omega \subseteq \Omega \mid X(\omega) \leq x\} \in \mathcal{F}(\Omega)$ *for all* $x \in \mathbb{R}$*.*

Definition B.2.6. *The* ***cumulative distribution function*** $F : \mathbb{R} \to [0,1]$ *of a random variable* X *is given by* $F(x) = \mathbb{P}\,(X \leq x)$*.*

Definition B.2.7. *A random variable* X *is called* ***continuous****, if its distribution function can be written as*

$$F(x) = \int_{-\infty}^{x} f(s)ds,$$

where the function $f : \mathbb{R} \to [0, \infty)$ *is called its probability density function.*

Definition B.2.8. *Two random variables* X *and* Y *are called* ***independent****, if their* ***joint distribution*** $F : \mathbb{R}^2 \to [0,1]$ *given by*

$$F(x,y) = \mathbb{P}\,(x \leq X \wedge y \leq Y)$$

can be factorized into

$$F(x,y) = F_X(x)F_Y(y),$$

where $F_X : \mathbb{R} \to [0,1]$ *is the cumulative distribution function of* X *and* $F_Y : \mathbb{R} \to [0,1]$ *the cumulative distribution function of* Y*. This is equivalent to*

$$f(x,y) = f_X(x)f_Y(y),$$

where $f(x,y) : \mathbb{R}^2 \to [0,\infty)$ *is the joint probability density function,* $f_X : \mathbb{R} \to [0,\infty)$ *the probability density function of* X *and* $f_Y : \mathbb{R} \to [0,\infty)$ *the probability density function of* Y*.*

Definition B.2.9. *A* ***random vector*** X *is a random variable over* $\mathbb{R}^n$*. Its* ***joint distribution function*** $F : \mathbb{R}^n \to [0,1]$ *is given by* $F(\boldsymbol{x}) = \mathbb{P}(X \leq \boldsymbol{x})$*, where the inequality is understood componentwise here.*

Definition B.2.10. *The* ***expected value*** $\mathbb{E}(X)$ *of a continuous random variable* X *with probability density function* $f(x)$ *is defined as*

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Definition B.2.11. *A random variable* X *is* ***integrable****, if* $\mathbb{E}(X) < \infty$*. For an integrable random variable* X *its* ***variance*** *is given by*

$$\mathrm{var}(X) = \mathbb{E}\big(\,(X - \mathbb{E}(X))^2\,\big).$$

Furthermore, its ***standard deviation*** $\sigma(X)$ *is given by*

$$\sigma(X) = \sqrt{\mathrm{var}(X)}.$$

Theorem B.2.12. *Let for* $n \in \mathbb{N}$ *the random variables* X_1*, ...,* X_n *be independently distributed. We assume that they all have the same standard deviation* σ*. Then the standard deviation of the mean* $\frac{1}{n}\sum_{i=1}^{n} X_i$ *is given by* $\frac{\sigma}{\sqrt{n}}$*.*

Theorem B.2.13. *(Central Limit Theorem) Let the random variables X_1, ..., X_n be independent and equally distributed. We assume that the common expected values and standard deviations exist and are finite, i.e.*

$$\mu := \mathbb{E}(X_i) < \infty \quad \textit{and} \quad \sigma := \sigma(X_i) < \infty$$

for $i = 1, ..., n$. Then we have for every real number x

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\tfrac{1}{2}t^2)dt.$$

Theorem B.2.14. *(Inverse transform sampling) Let $u \sim U([0,1])$ be a uniform random variable and X be a random variable with cumulative distribution function $F : \mathbb{R} \to [0,1]$. Then the random variable $F^{-1}(u)$ is distributed like X.*

Definition B.2.15. *The **conditional probability** of a random variable A given a random variable B is specified by*

$$p(A|B) = \frac{p(A \cap B)}{p(B)},$$

where $p(A \cap B)$ is the **joint probability** of A and B. Here $A \cap B$ is the intersection of sets, where A and B live on.

Theorem B.2.16. *(Bayes' Theorem) The conditional probability of a random variable A given a random variable B can be expressed as*

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

Bibliography

[1] G. Alldredge and T. Kyrion, “Robust inversion methods for aerosol spectroscopy,” *Inverse Problems in Science and Engineering*, vol. 25, pp. 710–748, 2017.

[2] T. Kyrion, “Reconstruction of Refractive Indices from Spectral Measurements of Monodisperse Aerosols,” *Inverse Problems in Science and Engineering*, 2017.

[3] G. Mie, “Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen,” *Annalen der Physik, Vol. 25*, 1908.

[4] M. Quinten, *Optical Properties of Nanoparticle Systems.* WILEY-VCH Verlag, 2010.

[5] Q. Fu and W. Sun, “Mie theory for light scattering by a spherical particle in an absorbing medium,” *Applied Optics, Vol. 40, Issue 9*, 2001.

[6] W. Wiscombe, “Improved mie scattering algorithms,” *Applied Optics, Vol. 19, Issue 9*, 1980.

[7] B. A. Krupa, P.-M. Steffen, J. Kobalz, and H.-J. Allelein, “Development and qualification of an aerosol generator for investigations under thermal-hydraulic severe accident boundary conditions,” *Proceedings of the 16th International Topical Meeting on Nuclear Reactor Thermal Hydraulics NURETH-16, Chicago, Illinois, USA, 30 Aug 2015 - 4 Sep 2015*, 2015.

[8] B. A. Krupa, H.-J. Allelein, A. Dreizler, V. Ebert, M. Frank, and D. Steiger, “Development and Qualification of Innovative Measurement Devices for Multi-Component Aerosols and Relative Humidity by Using Extinction Photometry,” *Proceedings of Annual Meeting on Nuclear Technology (AMNT14), Frankfurt a. Main Germany, May 6 -8, Vol. 45*, 2014.

[9] W. Salzmann, “Bedienungsanleitung FASP,” *Fraunhofer IPM*, 2011.

[10] M. Dashti and A. M. Stuart, “The Bayesian Approach to Inverse Problems,” *lecture notes*, pp. 1–105, 2014.

[11] E. Greenberg and S. Chib, “Markov Chain Monte Carlo Simulation Methods in Econometrics,” *Econometric Theory*, vol. 12, 1996.

[12] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems.* Springer Science and Business Media, 1996.

[13] A. Neubauer, "Tikhonov regularization of ill-posed linear operator equations on closed convex sets," *J. Approx. Theory*, vol. 53, pp. 304–320, 1988.

[14] J. Flemming and B. Hofmann, "Convergence rates in constrained Tikhonov regularization: equivalence of projected source conditions and variational inequalities," *Inverse Problems*, pp. 085001–11, 2011.

[15] M. T. Nair, E. Schock, and U. Tautenhahn, "Morozovs Discrepancy Principle under General Source Conditions," *Journal for Analysis and its Applications*, vol. 22, no. 1, p. 199214, 2003.

[16] J. Flemming and B. Hofmann, "Regularization without preliminary knowledge of smoothness and error behaviour," *Euro. Jnl of Applied Mathematics*, pp. 303–317, 2005.

[17] M. T. Naira and S. V. Pereverzev, "Regularized collocation method for Fredholm integral equations of the first kind," *Journal of Complexity, Vol. 23, Issues 46*, pp. 454–467, 2007.

[18] S. Pereverzev and E. Schock, "On the Adaptive Selection of the Parameter in Regularization of Ill-Posed Problems," *SIAM J. Numer. Anal., Vol. 43, Issue 5*, pp. 2060–2076, 2005.

[19] P. Mathé and S. V. Pereverzev, "Discretization strategy for linear ill-posed problems in variable Hilbert scales," *Inverse Problems*, vol. 19, p. 12631277, 2003.

[20] A. G. an Barbara Kaltenbacher and B. Vexler, "Efficient computation of the Tikhonov regularization parameter by goal oriented adaptive discretization," *Inverse Problems*, vol. 24, 2008.

[21] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems.* Springer, 2005.

[22] D. J. C. MacKay, "Bayesian Interpolation," *Neural Computation*, vol. 25, no. 3, pp. 415–447, 1992.

[23] C. W. Groetsch and A. Neubauer, "Convergence of a general projection method for an operator equation of the first kind," *Houston journal of mathematics*, vol. 14, no. 2, pp. 201–208, 1988.

[24] E. I. George, H. Chipman, and R. E. McCulloch, "The Practical Implementation of Bayesian Model Selection," *IMS Lecture Notes*, vol. 38, 2001.

[25] A. Genz, "Numerical Computation of Multivariate Normal Probabilities," *J. of Computational and Graphical Stat.*, vol. 1, 1992.

[26] H. Niederreiter, "On a Number-Theoretical Integration Method," *Aequationes Mathematicae, Vol. 8*, 1972.

[27] R. Cranley and T. Patterson, "Randomization of Number Theoretic Methods for Multiple Integration," *SIAM J Numer Anal, Vol. 13*, 1976.

[28] S. Aminsadrabad, "Numerical Solution of Integral Equations with Legendre Basis," *Int. J. Contemp. Math. Sciences*, vol. 6, pp. 1131 – 1135, 2011.

[29] A. Shirin and M. S. Islam, "Numerical Solutions of Fredholm Integral Equations Using Bernstein Polynomials," *J. Sci. Res.*, vol. 2, pp. 264–272, 2010.

[30] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics, Vol. 6, Issue 2*, 1978.

[31] N. Riefler and T. Wriedt, "Intercomparison of Inversion Algorithms for Particle-Sizing Using Mie Scattering," *Particle and Particle Systems Characterization*, 2008.

[32] A. Mohammad-Djafari, "Model selection for inverse problems: Best choice of basis functions and model order selection," *arXiv preprint*, pp. 1–18, 2001.

[33] M. Laine and J. Tamminen, "Aerosol model selection and uncertainty modelling by adaptive MCMC technique," *Atmos. Chem. Phys.*, pp. 7697–7707, 2008.

[34] L. Wasserman, "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, 2000.

[35] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis.* CRC Press, 2014.

[36] S. Sharma, P. B. Patel, R. S. Patel, and J. J. Vora, "Density and Comparative Refractive Index Study on Mixing Properties of Binary Liquid Mixtures of Eucalyptol with Hydrocarbons at 303.15, 308.15 and 313.15K," *E-Journal of Chemistry, Vol. 4, Issue 3*, 2006.

[37] F. Colonius and K. Kunisch, "Output Least Squares Stability in Elliptic Systems," *Appl. Math. Optim.*, vol. 19, pp. 33–63, 1989.

[38] A. A. Riziq, C. Erlick, E. Dinar, and Y. Rudich, "Optical properties of absorbing and non-absorbing aerosols retrieved by cavity ring down (crd) spectroscopy," *Atmos. Chem. Phys., Vol. 7*, 2007.

[39] P. Deuflhard and A. Hohmann, *Numerical Mathematics I: an algorithmically oriented introduction.* de Gruyter, 2002.

[40] R. M. Goody and Y. L. Yung, *Atmospheric Radiation. Theoretical Basis, second edition.* Oxford University Press, New York, 1989.

[41] H. W. Engl, K. Kunisch, and A. Neubauer, "Convergence rates for Tikhonov regularisation of non-linear ill-posed problems," *Inverse Problems*, vol. 5, pp. 523–540, 1989.

[42] O. Scherzer, H. W. Engl, and K.Kunisch, "Optimal a Posteriori Parameter Choice for Tikhonov Regularization for Solving Nonlinear Ill-Posed Problems," *SIAM J. Numer. Anal.*, vol. 6, pp. 1796–1838, 1993.

[43] T. V. Gestel, M. Espinoza, J. Suykens, C. Brasseur, and B. D. Moor, "Bayesian input selection for nonlinear regression with LS-SVMS," *IFAC Proceedings Volumes*, vol. 36, no. 16, pp. 555–560, 2003.

[44] R. May, G. Dandy, and H. Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks," *chapter in the book: Artificial Neural Networks - Methodological Advances and Biomedical Applications*, pp. 19–44, 2011.

[45] N. Bissantz, T. Hohage, and A. Munk, "Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise," *Inverse Problems*, vol. 20, no. 6, p. 1773 1789, 2004.

[46] D. J. C. MacKay, "Bayesian Non-Linear Modeling for the Prediction Competition," *chapter in the book: Maximum Entropy and Bayesian Methods*, pp. 221–234, 1994.

[47] S. W. Anzengruber and R. Ramlau, "Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators," *Inverse Problems*, vol. 26, no. 2, 2009.

[48] C. Erlick, M. Haspel, and Y. Rudich, "Simultaneous retrieval of the complex refractive indices of the core and shell of coated aerosol particles from extinction measurements using simulated annealing," *Appl. Opt.*, pp. 4393–4402, 2011.

[49] R. G. Grainger, J. Lucas, G. E. Thomas, and G. B. L. Ewen, "Calculation of Mie derivatives," *Applied Optics*, vol. 43, no. 28, pp. 5386–93, 2004.

[50] M. Abromovitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications, 1972.