Reformulation of the model using curvelets

F. Parzer

October 29, 2020

Abstract

Overview

Section 1 contains a short summary of the relevant informations of the problem as we understood it.

In Section 2, a formulation of the model is presented where the distribution function of the galaxy is approximated using curvelets.

Finally, in Section 3, the optimization problem is formulated. We also present the computational aspects and discuss how approximate uncertainty quantification in form of covariance estimates and confidence intervals could be performed.

Contents

1. Formulation of the model

1.1. Problem summary. The problem at hand is to estimate the line-of-sight velocity distribution of a galaxy.

A galaxy is modelled as a collection G of *stellar populations*. Each stellar population consists of a number of stars and is characterized by two parameters. The first parameter is the *metallicity* $m_p \in \mathbb{R}$. As the logarithm of a ratio, it is a real number. The second parmeter is the age $a_p > 0$ of the stellar population in Gyr (Gigayears). In the *MILES simple stellar population model*, these two numbers determine the wavelength spectrum of the stellar population. A wavelength spectrum is a function $s : I \times \Theta \to \mathbb{R}^+$, where $I \subset \mathbb{R}^+$ is the range of relevant wavelengths and $\Theta \subset \mathbb{R} \times \mathbb{R}^+$ is the space of possible values of the parameter $\boldsymbol{\theta}_p = (m_p, a_p)$.

By the *distribution function of the galaxy* we mean the histogram of the parameter $\boldsymbol{\theta}$ with respect to stellar mass, i.e. a function $f : \Theta \to \mathbb{R}^+$ that maps to each parameter value $\boldsymbol{\theta}$ the total stellar mass of all stellar populations with the corresponding value. The function f necessarily satisfies

$$f(\boldsymbol{\theta}) \geq 0 \text{ and } \int_\Theta f(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = M_{\text{total}}, \tag{1.1}$$

where M_{total} is the total stellar mass of the galaxy. From practical experience, f is typically a multimodal function which is characterized by a few sickle-shaped "hills" and otherwise close to zero. This corresponds to the physical fact that the distribution of the stellar mass is concentrated at a few clusters with similar age and metallicity. Thus, we expect that in a localized basis or frame of functions (see Section 2.1), f can be well-approximated by only a few basis elements. That is, we expect *sparsity* in the discretization.

The *composite spectrum* of the stellar population of a galaxy is then given through a superposition of the spectra of all stellar populations:

$$S(\lambda; f) = \int_\Theta f(\theta) s(\lambda; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{1.2}$$

The Doppler effect shifts the observed spectrum. The size of this shift is determined by the component of the velocity of the stellar population along the line-of-sight from that population to earth. Therefore, in order to know the unshifted spectrum, we have to estimate the *line-of-sight velocity distribution* (LOSVD) of the galaxy. This is a collection $(v_p)_{p \in G} \subset \mathbb{R}$ of real numbers, where each v_p represents the line-of-sight velocity component of the stellar population p. One can normalize the LOSVD by estimating first the line-of-sight velocity of the galaxy (is this the average velocity of all stellar populations, or something else?), and then shift the LOSVD by this amount. Thus, the LOSVD is centered around zero.

Observed LOSVDs are typically modelled by *Gauss-Hermite expansions*. We define

$$L(v, \boldsymbol{\theta}_v) := \mathcal{N}(\hat{v}; 0, 1) \left[\sum_{m=0}^{M} h_m H_m(\hat{v}) \right],$$

where $\mathcal{N}(\cdot; 0, 1)$ denotes the density function of the standard normal distribution, and where furthermore

$$\begin{aligned} \hat{v} &:= \frac{v - V}{\sigma}, \\ H_m(x) &:= \frac{H_m^{\text{phys}}(x)}{\sqrt{m! 2^m}}, \\ H_m^{\text{phys}}(x) &:= (-1)^n e^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} [\mathrm{e}^{-x^2}], \end{aligned}$$

and $\boldsymbol{\theta}_v$ is the collection of parameters given by

$$\boldsymbol{\theta}_v = (V, \sigma, h_0, \ldots, h_M)$$

Thus, taking into account the Doppler effect and the LOSVD of the galaxy, the shifted spectrum is given by

$$\bar{y}(\lambda) = \int_{-\infty}^{\infty} \frac{1}{1 + v/c} S\left(\frac{\lambda}{1 + v/c}; f \right) L(v; \boldsymbol{\theta}) \, \mathrm{d}v. \tag{1.3}$$

1.2. The observation operator. We condense the above model in an observation operator

$$\mathcal{G} : \mathbb{F} \times \mathbb{R}^{M+2} \longrightarrow \mathbb{R}^{N_y}, \qquad \boldsymbol{\mathcal{G}}(f, \boldsymbol{\theta}_v) = \bar{\boldsymbol{y}}(f, \boldsymbol{\theta}_v),$$

which maps the parameters f and $\boldsymbol{\theta}_v$ to the discrete, shifted composite spectrum $\bar{\boldsymbol{y}} \in \mathbb{R}^{N_y}$. Here, N_y is the number of wavelet bins, and the vector $\bar{\boldsymbol{y}}$ is obtained from the function $\bar{y}$ through a discretization operation $\boldsymbol{P}_{N_y}$, i.e.

$$\bar{\boldsymbol{y}} = \boldsymbol{P}_{N_y}(\bar{y}).$$

We assume that the distribution function f lies in function space $\mathbb{F}$, given by

$$\mathbb{F} = \left\{ f \in L^2(\Theta) : \; f \geq 0, \int_\Theta f(\theta_s) \, \mathrm{d}\theta_s = M_{\text{total}} \right\},$$

where $\Theta \subset \mathbb{R} \times \mathbb{R}^+$ is the range of possible values of the metallicity and age.

The continuous spectrum $\bar{y}$ is computed from

$$\bar{y}(\lambda; f, \boldsymbol{\theta}_v) = \int_{-\infty}^{\infty} \frac{1}{1 + v/c} S\left(\frac{\lambda}{1 + v/c}; f \right) L(v; \boldsymbol{\theta}) \, \mathrm{d}v,$$

where

$$S(\lambda; f) = \int_\Theta f(\theta) s(\lambda; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},$$

and

$$L(v, [V, \sigma, h_0, \ldots, h_M]) := \mathcal{N}(\frac{v - V}{\sigma}; 0, 1) \left[\sum_{m=0}^{M} h_m H_m(\frac{v - V}{\sigma}) \right].$$

If we insert these expressions in the formula for $\bar{y}$, we obtain an explicit expression for $\boldsymbol{\mathcal{G}}(f, \boldsymbol{\theta}_v)$:

$$\boldsymbol{\mathcal{G}}(f, \boldsymbol{\theta}_v) = \int_{-\infty}^{\infty} \frac{1}{1 + v/c} \left(\int_\Theta f(\boldsymbol{\theta}) s(\frac{\lambda}{1 + v/c}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \right) \mathcal{N}(\frac{v - V}{\sigma}; 0, 1) \left(\sum_{m=0}^{M} h_m H_m(\frac{v - V}{\sigma}) \right) \mathrm{d}v. \tag{1.4}$$

1.3. The measurement error. We will assume that there is only additive measurement error, i.e. that the actual measurement is given by

$$\boldsymbol{y} = \boldsymbol{\mathcal{G}}(f, \boldsymbol{\theta}_v) + \boldsymbol{\epsilon}.$$

We will furthermore assume that

- ϵ is multivariate normally distributed (on $\mathbb{R}^{N_y}$),
- uncorrelated per wavelength bin,
- and that the expected noise level is equal for each wavelength bin.

That is, we assign the noise distribution

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \delta^2 \mathbb{I}_{N_y}), \tag{1.5}$$

where $\mathbb{I}_{N_y} \in \mathbb{R}^{N_y \times N_y}$ is the N_y-dimensional identity matrix and $\delta > 0$ is the noise level. Typically, the signal-to-noise ratio lies in the range $[30, 300]$.

1.4. The prior distribution. The expected sparsity of the distribution function f is represented in our model by assigning a sparsity-promoting prior. We choose a Laplace-like prior that also accounts for the additional constraints (1.1) that f has to satisfy. Formally, we would like to write this in terms of a probability density

$$f \sim \frac{1}{C} \exp\left(-\sqrt{2} \left\| \Sigma_1^{-1/2} f \right\|_1 \right),$$

and then restrict this function to only those values that satisfy the necessary conditions (1.1) (recall that the factor $\sqrt{2}$ is just a scaling factor that ensures that the distribution has covariance Σ_1 and not $\frac{1}{2}\Sigma_1$.

However, since f is an infinite-dimensional object and probability density functions on infinite-dimensional spaces are not well-defined, the only way to make rigorous sense of a prior probability on the distribution function $f : \Theta \to \mathbb{R}^+$ is as a probability measure on the infinite-dimensional function space $L^2(\Theta)$. In our particular case, this probability measure is given by

$$\mu_0(\,\mathrm{d}f \mid \Sigma_1) = \frac{1}{C_0}\mathrm{Lap}(\,\mathrm{d}f \mid 0, \Sigma_1)\mathbb{1}_{\mathbb{F}}(f), \tag{1.6}$$

where $\mathrm{Lap}(\,\mathrm{d}f|0, \Sigma_1)$ is the infinite-dimensional Laplace-measure on $L^2(\Theta)$ with mean 0 (in order to promote sparsity) and covariance operator Σ_1 (see `IEKF11.pdf` for the exact construction of the infinite-dimensional Laplace measure). The indicator function $\mathbb{1}_{\mathbb{F}}(f)$ is equal to 1 if $f \in \mathbb{F}$, and 0 else. Consequently, it projects the prior probability on the constraints (1.1).

This measure-theoretic formulation can be treated as a formal detail and does not affect our actual implementation, since we will only deal with finite-dimensional discretizations of f. However, it might be relevant for rigorous interpretation of some results.

For the other parameter $\boldsymbol{\theta}_v$, which is given by the vector $(V, \sigma, h_0, \dots h_M)$, we will assume a simple uncorrelated Gaussian prior, i.e.

$$\boldsymbol{\theta}_v \sim \mathcal{N}(\bar{\boldsymbol{\theta}}_v, \boldsymbol{\Sigma}_2), \tag{1.7}$$

where $\bar{\boldsymbol{\theta}}_v = (\bar{V}, \bar{\sigma}, \bar{h}_0, \dots \bar{h}_M)$ is the prior mean and $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\delta_v, \delta_{h_0}, \dots, \delta_{h_M}) \in \mathbb{R}^{(M+2)\times(M+2)}$ is a diagonal covariance matrix. The parameters $\bar{\boldsymbol{\theta}}_v$ and $\boldsymbol{\Sigma}_2$ have to be chosen by us, and we should evaluate in numerical experiments how this choice affects our inference.

If we assign the priors (1.6) and (1.7) and use the assumption (1.5), then the statistical formulation of our inverse problem is

$$\boldsymbol{y} = \boldsymbol{\mathcal{G}}(f, \boldsymbol{\theta}_v) + \boldsymbol{\epsilon}, \tag{1.8}$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \delta^2 \mathbb{I}_{N_y}), \tag{1.9}$$
$$f \sim \mu_0(\,\mathrm{d}f \mid \Sigma_1), \tag{1.10}$$
$$\boldsymbol{\theta}_v \sim \mathcal{N}(\bar{\boldsymbol{\theta}}_v, \boldsymbol{\Sigma}_2), \tag{1.11}$$
$$f \perp\!\!\!\perp \boldsymbol{\theta}_v, \tag{1.12}$$

where $f \perp\!\!\!\perp \boldsymbol{\theta}_v$ means that we assume that the unknown parameters f and $\boldsymbol{\theta}_v$ are independent.

2. Discretization of the distribution function

As mentioned above, the distribution function f is expected to have a particular shape: It is negligibly small on most of its domain, except for a few curved hills in which most of its mass is concentrated. Because of this fact, in our first meeting we came up with the idea of finding a sparse approximation of f using *curvelets*.

2.1. Curvelet preliminaries. Curvelets are a useful modification of wavelets for sparse geometrical image representation. Their advantage over wavelets is that they allow for a better representation of directional components of images, i.e. curved edges. For some exemplary applications of curvelets to astronomy, see [15] and [9].

Mathematical definition

Mathematically, curvelets yield a discretization of the function space $L^2(\mathbb{R}^2)$. This discretization is particularly suited for the approximation of images that are smooth except for discontinuities along C^2 curves. For a very good review on curvelets in signal processing, see [11].

We define the curvelet dictionary as follows: Consider the so-called *scaled Meyer windows*

$$V(t) := \begin{cases} 1, & |t| \leq \frac{1}{3}, \\ \cos\left[\frac{\pi}{2}\nu(3\,|t|-1)\right], & \frac{1}{3} \leq |t| \leq \frac{2}{3}, \\ 0, & \text{else}, \end{cases} \qquad t \in \mathbb{R}, \tag{2.1}$$

and

$$W(r) := \begin{cases} \cos\left[\frac{\pi}{2}\nu(5-6r)\right], & \frac{2}{3} \leq r \leq \frac{5}{6}, \\ 1, & \frac{5}{6} \leq r \leq \frac{4}{3}, \\ \cos\left[\frac{\pi}{2}\nu(3r-4)\right], & \frac{4}{3} \leq r \leq \frac{5}{3}, \\ 0, & \text{else}, \end{cases} \tag{2.2}$$

where $\nu \in C^\infty(\mathbb{R})$ an be any smooth function that satsfies

- $\nu \equiv 0$ on $(-\infty, 0]$ and $\nu \equiv 1$ on $[1, \infty)$,
- $\nu(x) + \nu(1-x) = 1$, for all $x \in \mathbb{R}$.

Next, we define the *a-scaled window*

$$U_a(r\omega) := a^{3/4} W(ar) V(\frac{\omega}{\sqrt{a}}), \qquad a \in [0,1], \quad r \geq 0, \quad \omega \in [0, 2\pi).$$

This window is used as a function in polar coordinates on the frequency domain, i.e. for a frequency $\xi = [\xi^1, \xi^2]^\top \in \mathbb{R}^2$ we define

$$r := \sqrt{|\xi^1|^2 + |\xi^2|^2}, \quad \text{and} \quad \omega := \arctan\left(\frac{\xi^1}{\xi^2}\right).$$

Then, we define the base curvelet in frequency domain as

$$\hat{\varphi}_{a,0,0}(\xi) := U_a(\xi), \qquad \xi \in \mathbb{R}^2. \tag{2.3}$$

Via rotation and translation, this generates the curvelet family

$$\varphi_{a,b,\theta}(x) := \varphi_{a,0,0}(R_\theta(x-b)),, \tag{2.4}$$

where

$$R_\theta := \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \theta \in [0, 2\pi), \tag{2.5}$$

is the 2-dimensional rotation around the angle θ.

The discrete curvelet frame

The curvelet family $(\varphi_{a,b,\theta})$ defines a frame of $L^2(\mathbb{R}^2)$. Recall that frames generalize the concept of a basis to not necessarily linearly independent sequences. Thus, they allow for *redundancy* in the representation of an object. Frames are defined as follows (see, for example, [12]):

Definition 2.1 Let $\mathbb{H}$ be a Hilbert space. A sequence $(\phi_i)_{i \in I}$ is a *frame* of $\mathbb{H}$ if there are two constants $A, B > 0$, with $A \leq B$, such that

$$A\,\|f\|^2 \leq \sum_{i \in I} |\langle f, \phi_i \rangle|^2 \leq B\,\|f\|^2, \qquad \forall f \in \mathbb{H}. \tag{2.6}$$

If $A = B$, then the frame is called *tight*.

Since the frame constant A must be strictly positive, condition (2.6) implies that the cardinality of the index set I cannot be smaller than the dimension of the space $\mathbb{H}$. Frames generalize the concept of a basis in the further sense that they allow for coefficient expansions of elements of $\mathbb{H}$. The next result can be found as theorem 5.5 in [12].

Theorem 2.2 *Let $(\phi_n)_{i\in I}$ be a frame of a Hilbert space $\mathbb{H}$. Then, for every $f \in \mathbb{H}$, we have*

$$f = \sum_{i\in I} \langle f, \phi_i\rangle \, \tilde{\phi}_i = \sum_{i\in I} \left\langle f, \tilde{\phi}_i \right\rangle \phi_i,$$

where $\tilde{\phi}_i := (\Phi^*\Phi)^{-1}\phi_i$ are the elements of the so-called dual frame*, and $\Phi : \mathbb{H} \longrightarrow \mathbb{R}^I$ is the* frame analysis operator*, given by*

$$(\Phi f)_i := \langle f, \phi_i\rangle \, .$$

If the frame is tight with constant A, then $\tilde{\phi}_i = \frac{1}{A}\phi_i$, for all $i \in I$.

In particular, this shows that any function f can be reconstructed from its frame coefficients.
Finally, the next theorem demonstrates the usual construction of a curvelet frame of $L^2(\mathbb{R}^2)$.

Theorem 2.3 *Let*

$$\begin{aligned}
&a_j := 2^{-j},\\
&\theta_{j,l} := \frac{\pi l 2^{-\lceil j/2\rceil}}{2},\\
&b_k^{j,l} := R_{\theta_{j,l}}^{-1} \begin{bmatrix} k^1/2^j \\ k^2/2^{j/2} \end{bmatrix}, \qquad k \in \mathbb{Z}^2, \quad l \in \{0, 1, \ldots, 4\cdot 2^{-\lceil j/2\rceil} - 1\}, \quad j \in \mathbb{N}.
\end{aligned}$$

We define the scaled window function U_j by

$$U_j(r,\omega) := 2^{-3j/4} W(2^{-j}r) V\left(\frac{2 - 2^{-\lceil j/2\rceil}\omega}{\pi}\right).$$

and the discrete family of curvelets by

$$\begin{aligned}
&\hat{\phi}_{j,0,0}(\xi) := U_j(\xi),\\
&\phi_{j,k,l}(x) := \phi_{j,0,0}(R_{\theta_{j,l}}(x - b_k^{j,l})),\\
&\phi_{-1,k,0} := \phi_{-1}(x - k).
\end{aligned}$$

Let $I_c = \left\{ \boldsymbol{i} = (j,k,l) : \, k \in \mathbb{Z}^2, l \in \{0, 1, \ldots, 4\cdot 2^{-\lceil j/2\rceil} - 1\}, j \in \mathbb{N} \right\}$. Then $(\phi_{\boldsymbol{i}})_{\boldsymbol{i}\in I_c}$ is a tight frame of $L^2(\mathbb{R}^2)$ with frame bound $A = 1$.

For a proof of this statement and more properties of curvelets, see [8].
Since the frame bound of the curvelet frame is 1, a particularly simple coefficient expansion holds for all $f \in L^2(\mathbb{R}^2)$:

$$f = \sum_{\boldsymbol{i}\in I_c} \langle f, \phi_{\boldsymbol{i}}\rangle_2 \, \phi_{\boldsymbol{i}}, \qquad \forall f \in L^2(\mathbb{R}^2), \tag{2.7}$$

where $\langle \cdot, \cdot \rangle_2$ denotes the L^2-inner product. Furthermore, the curvelet expansion satisfies Parseval's identity,

$$\sum_{\boldsymbol{i}\in I_c} |\langle f, \phi_{\boldsymbol{i}}\rangle_2|^2 = \|f\|_2^2 \, .$$

Similar to the fast Fourier transform, there exists a very efficient algorithm for computing the curvelet coefficients $(f, \phi_{\boldsymbol{i}})$ of a given function f, called the *fast curvelet transform*. If f is given as a discrete image $(f(n_1, n_2))$ of $N \cdot N$ pixels, then the computational complexity of the discrete curvelet transform is $O(N^2 \log N)$ [7].

2.2. The curvelet model. We can now reformulate our model from Section 1.3 using curvelets. Suppose the distribution f has the curvelet expansion

$$f = \sum_{\boldsymbol{i} \in I_c} f_{\boldsymbol{i}} \phi_{\boldsymbol{i}}, \tag{2.8}$$

where $f_{\boldsymbol{i}} = \langle f, \phi_{\boldsymbol{i}} \rangle$. Let $\tilde{f} := \Phi f = (f_{\boldsymbol{i}})_{\boldsymbol{i} \in I_c}$. If we insert (2.8) in (1.2), we have

$$\begin{aligned} S(\lambda, f) &= \sum_{\boldsymbol{i} \in I_c} f_{\boldsymbol{i}} \underbrace{\int \phi(\theta_s) s(\lambda; \theta_s) \, \mathrm{d}\theta_s}_{=: s_i(\lambda)} \\ &=: S(\lambda, \tilde{f}). \end{aligned}$$

Then, we can rewrite the observation operator as

$$\mathcal{G}(\tilde{f}, \theta_v) = \sum_{\boldsymbol{i} \in I_c} f_{\boldsymbol{i}} \int_{-\infty}^{\infty} \frac{1}{1 + v/c} s_{\boldsymbol{i}} \left(\frac{1}{1 + v/c} \right) L(v; \theta_v) \, \mathrm{d}v. \tag{2.9}$$

2.3. The finite-dimensional approximate model. Suppose next we treat f as a discrete image $\boldsymbol{f}[n_1, n_2]$. Correspondingly, we let $\boldsymbol{\Sigma}_1$ be the corresponding projection of the (infinite-dimensional) covariance matrix on the discretization. We can compute its curvelet coefficients $\tilde{\boldsymbol{f}} = \boldsymbol{\Phi} \boldsymbol{f}$ using the fast curvelet transform. Since curvelets are not necessarily nonnegative, the nonnegativity condition in (1.1) does not directly translate into a nonnegativity condition on the curvelet coefficients. Instead, we have the condition

$$\boldsymbol{\Phi}^{-1} \tilde{\boldsymbol{f}} \geq 0,$$

where $\boldsymbol{\Phi}^{-1}$ is the finite-dimensional curvelet reconstruction operator.

The integral condition in (1.1) translates to

$$\sum_{\boldsymbol{i}} w_{\boldsymbol{i}} \tilde{\boldsymbol{f}}_{\boldsymbol{i}} = M_{\text{total}}, \tag{2.10}$$

where $w_{\boldsymbol{i}} = \int \phi_{\boldsymbol{i}}(\theta_s) \, \mathrm{d}\theta_s$.

Note that we have only a finite number of wavelength bins so that our measurement is actually m-dimensional, where m is the number of bins. Thus, we need no additional discretization of the measurements.

Our discretized model then looks like this:

$$\boldsymbol{y} = \boldsymbol{\mathcal{G}}(\tilde{\boldsymbol{f}}, \boldsymbol{\theta}_v) + \boldsymbol{\epsilon}, \tag{2.11}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \delta^2 \mathbb{I}_m), \tag{2.12}$$

$$\tilde{\boldsymbol{f}} \sim \mathrm{Lap}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_1), \tag{2.13}$$

$$\boldsymbol{\theta}_v \sim \mathcal{N}(\bar{\boldsymbol{\theta}}_v, \boldsymbol{\Sigma}_2), \tag{2.14}$$

$$\tilde{\boldsymbol{f}} \perp\!\!\!\perp \boldsymbol{\theta}_v, \tag{2.15}$$

where $\tilde{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Phi} \boldsymbol{\Sigma}_1 \boldsymbol{\Phi}^\top$ is the covariance matrix for the curvelet coefficients and $\mathrm{Lap}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_1)$ is the Laplace measure (see (1.6)). Its probability density function is given by

$$p_0(\tilde{\boldsymbol{f}}) = \frac{1}{C_0} \exp\left(-\sqrt{2} \left\| \tilde{\boldsymbol{\Sigma}}_1^{-1/2} \tilde{\boldsymbol{f}} \right\|_1 \right) \mathbb{1}_{\mathbb{F}}(\tilde{\boldsymbol{f}}).$$

where the set $\mathbb{F} = \left\{ \tilde{\boldsymbol{g}} = [g_1, \ldots, g_M] : \boldsymbol{\Phi}^{-1} \tilde{\boldsymbol{g}} \geq \boldsymbol{0}, \sum_i w_i \tilde{g}_i = M_{\text{total}} \right\}$ is the set of all vectors $\tilde{\boldsymbol{g}}$ of curvelet coefficients such that $\boldsymbol{g} = \boldsymbol{\Phi}^{-1} \tilde{\boldsymbol{g}}$ satisfies the constraints (1.1), and C_0 is a normalization constant.

3. THE COMPUTATIONAL APPROACH

Since any sampling-based approach to uncertainty quantification is likely to be computationally infeasible, we have to make due with mode-based inference. That means we compute point estimates for $\tilde{\boldsymbol{f}}$ and $\boldsymbol{\theta}_v$ through maximum-a-posteriori estimation, and perform rudimentary uncertainty quantification using mode-based estimates of the posterior covariance and confidence regions.

By Bayes' rule, the posterior probability density p_{post} for the finite-dimensional model (2.11)-(2.15) satisfies

$$p_{\text{post}}(\tilde{\boldsymbol{f}}, \boldsymbol{\theta}_v | y) \propto \exp\left(-\frac{1}{2\delta^2} \left\| \boldsymbol{y} - \mathcal{G}(\tilde{\boldsymbol{f}}, \boldsymbol{\theta}_v) \right\|_2^2 - \sqrt{2} \left\| \Sigma_1^{-1/2} \tilde{\boldsymbol{f}} \right\|_1 - \frac{1}{2} \left\| \boldsymbol{\Sigma}_2^{-1/2} (\boldsymbol{\theta}_v - \bar{\boldsymbol{\theta}}_v) \right\|_2^2 \right) \mathbb{1}_{\mathbb{F}}(\tilde{\boldsymbol{f}}),$$

where $\mathbb{F}$ is the set of feasible points defined above.

Any mode of p_{post} is a minimizer of the following constrained minimization problem

$$\begin{aligned} \min \quad & \frac{1}{2\delta^2} \left\| \boldsymbol{y} - \mathcal{G}(\tilde{\boldsymbol{f}}, \boldsymbol{\theta}_v) \right\|_2^2 + \sqrt{2} \left\| \boldsymbol{\Sigma}_1^{-1/2} \tilde{\boldsymbol{f}} \right\|_1 + \frac{1}{2} \left\| \boldsymbol{\Sigma}_2^{-1/2} (\boldsymbol{\theta}_v - \bar{\boldsymbol{\theta}}_v) \right\|_2^2, && (3.1) \\ \text{subject to} \quad & \boldsymbol{\Phi}^{-1} \tilde{\boldsymbol{f}} \geq 0, && (3.2) \\ & \sum_i w_i \tilde{f}_i = M_{\text{total}}. && (3.3) \end{aligned}$$

If we want point estimates, we have to solve this optimization problem.

3.1. Transform into nonlinear least-squares problem. In general, l^2-regularized least-squares problems are easier to solve than l^1-regularized least-squares problems. However, we can transform the latter into the former.

Consider a generic l^1-regularized nonlinear least-squares problem

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x}) \right\|_2^2 + \sqrt{2} \left\| \boldsymbol{\Sigma}^{-1/2} \boldsymbol{x} \right\|_1. \tag{3.4}$$

(In statistics, this is called "LASSO"). The problem (3.4) can be reparametrized into an l^2-regularized problem of the form

$$\min_{\boldsymbol{z}} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{F} \circ \boldsymbol{T}(\boldsymbol{z}) \right\|_2^2 + \frac{1}{2} \left\| \boldsymbol{\Sigma}^{-1/2} \boldsymbol{z} \right\|_2^2, \tag{3.5}$$

(in statistics, this is called "ridge regression") where $\boldsymbol{T}$ is a suitable nonlinear transformation.

Consider the univariate transform $\tau : \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$\tau(x) = \frac{1}{2\sqrt{2}} \operatorname{sign}(x) \left| x \right|^2. \tag{3.6}$$

This function is continuous and bijective (see [16]).

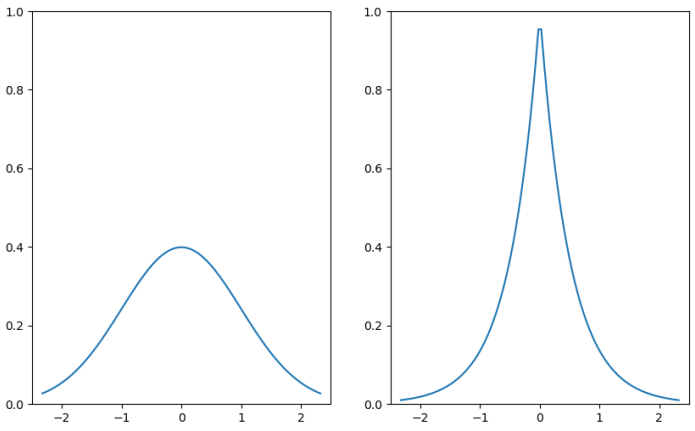

FIGURE 1. *Probability density functions of the standard normal distribution (left) and the Laplace distribution, each with mean 0 and covariance 1 (right).*

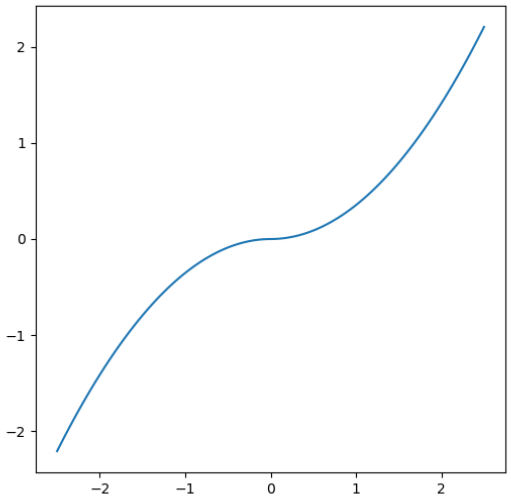

FIGURE 2. *The transform τ defined in (3.6). It transports a standard normal distribution to a Laplace distribution with mean 0 and covariance 1.*

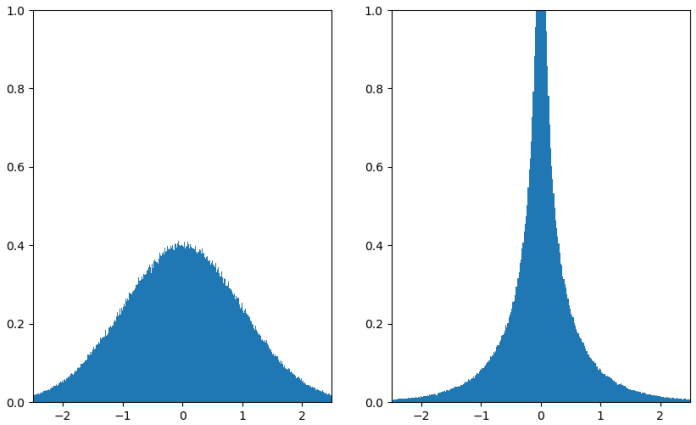

FIGURE 3. *Left: Histogram of 10^6 independent samples of a standard normal distribution. Right: The image of those samples under the function τ.*

The transform τ can easily be extended to $\mathbb{R}^n$. We define the multivariate transform $\boldsymbol{T} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{\Sigma}^{1/2}\tau(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}), \tag{3.7}$$

where $\tau(\boldsymbol{x}) = [\tau(x_1), \ldots, \tau(x_n)]$ denotes componentwise application of τ. Proposition 3.5 in [16] then implies that with this choice for the transform T, problems (3.4) and (3.5) are equivalent. (Note that this is not true for arbitrary transformations. It is a well-known fact (and problem) of mode-based inference that maximum-a-posteriori estimators are not invariant under general reparametrization.)

If we apply this reparametrization trick to our particular problem (3.1)-(3.3), we obtain the constrained nonlinear least-squares problem

$$\min_{\boldsymbol{g}} \quad \frac{1}{2\delta^2}\left\|\boldsymbol{y} - \boldsymbol{\mathcal{G}}(\boldsymbol{T}(\boldsymbol{g}), \boldsymbol{\theta}_v)\right\|_2^2 + \frac{1}{2}\left\|\tilde{\boldsymbol{\Sigma}}_1^{-1/2}\boldsymbol{g}\right\|_2 + \frac{1}{2}\left\|\boldsymbol{\Sigma}_2^{-1/2}(\boldsymbol{\theta}_v - \bar{\boldsymbol{\theta}}_v)\right\|_2^2, \tag{3.8}$$

$$\text{subject to} \quad \boldsymbol{\Phi}^{-1}\boldsymbol{T}(\boldsymbol{g}) \geq \boldsymbol{0}, \tag{3.9}$$

$$\sum_i w_i T(g)_i = M_{\text{total}}. \tag{3.10}$$

where $\boldsymbol{g} = \boldsymbol{T}^{-1}(\tilde{\boldsymbol{f}})$. This can be written more concisely as

$$\begin{aligned} \min \quad & \frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{\mathcal{H}}(\boldsymbol{z})\|_2^2 + \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}\right\|, && (3.11) \\ \text{subject to} \quad & \boldsymbol{T}(\boldsymbol{z}) \in \mathbb{F}, && (3.12) \end{aligned}$$

where $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2] = [\boldsymbol{g}, \boldsymbol{\theta}_v]$, $\boldsymbol{\mathcal{H}}(\boldsymbol{z}) = \boldsymbol{\mathcal{G}}(\boldsymbol{T}(\boldsymbol{z}_1), \boldsymbol{z}_2)$, $\mathbb{F}$ is defined as above, and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{bmatrix}.$$

3.2. The constrained Gauss-Newton method. So, we have reduced our inverse problem to the solution of the finite-dimensional constrained nonlinear least-squares problem (3.8)-(3.10). Probably the most popular method for solving nonlinear least-squares problems is the Gauss-Newton method, and it can be generalized to problems with constraints. One such generalization that has seen a lot of applications in parameter estimation is the *constrained Gauss-Newton method* (CGN) (usually called the generalized Gauss-Newton method ("verallgemeinertes Gauß-Newton-Verfahren") by the people from Heidelberg, although there are a couple of different methods under that name). It is an extension of the classical Gauss-Newton method and was developed by Bock and colleagues for parameter estimation with ODEs (see [3] and [4]).

Consider a generic nonlinear least-squares problem with equality- and inequality constraints.

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|\boldsymbol{F}_1(\boldsymbol{x})\|_2^2, \\ \text{subject to} \quad & \boldsymbol{F}_2(\boldsymbol{x}) = \boldsymbol{0}, \\ & \boldsymbol{F}_3(x) \geq \boldsymbol{0}. \end{aligned}$$

(Note that our transformed problem (3.8)-(3.10) fits into this framework.) The CGN iteration is given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x}_k,$$

where $\boldsymbol{x}_0$ is an initial guess and $\Delta\boldsymbol{x}_k$ is a solution of the linearized subproblem

$$\begin{aligned} \min_{\Delta\boldsymbol{x}} \quad & \|\boldsymbol{F}_1(x_k) + \boldsymbol{F}_1'(\boldsymbol{x}_k)\Delta\boldsymbol{x}\|_2^2, \\ \text{such that} \quad & \boldsymbol{F}_2(\boldsymbol{x}_k) + \boldsymbol{F}_2'(\boldsymbol{x}_k)\Delta\boldsymbol{x} = \boldsymbol{0}, \\ & \boldsymbol{F}_3(\boldsymbol{x}_k) + \boldsymbol{F}_3'(\boldsymbol{x}_k)\Delta\boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

This is a constrained linear least-squares problem, which can be solved using, for example, an active set approach. We omit a detailed description of the CGN method here, but it is available, for example, in the dissertation of Bock [4] (you can ask me if you want the pdf).

3.3. Uncertainty quantification. Suppose we have computed a MAP estimator $\boldsymbol{z}^*$, i.e. a solution of (3.11)-(3.12). Then, the transformed estimator

$$\begin{bmatrix} \tilde{\boldsymbol{f}}^* \\ \boldsymbol{\theta}_v^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}(\boldsymbol{z}_1^*) \\ \boldsymbol{z}_2^* \end{bmatrix}$$

is a MAP estimate for our original problem.

We want to perform uncertainty quantification by computing Bayesian *highest probability-density regions* (HPD regions). The definition of the HPD region is distinct from the frequentist confidence interval, although both regions often overlap in practive. Given an unknown parameter $\boldsymbol{x} \in \mathbb{R}^n$ and a constant $\alpha \in (0, 1)$, an $(1-\alpha)$-HPD region C_α is any smallest subset of $\mathbb{X}$ such that

$$\mathbb{P}(\boldsymbol{x} \in C_\alpha \mid \boldsymbol{y}) \geq 1 - \alpha.$$

That is, C_α is a minimal region that contains the parameter $\boldsymbol{x}$ with probability $1-\alpha$, given the measurement $\boldsymbol{y}$. Normally, such a region could be computed using a Markov chain Monte Carlo approach. However, since we do not want to use Monte Carlo methods (due to the fact that they do not scale well enough), we

have to use the next-best thing, which consists in mode-based inference. That is, we try to approximate C_α using local properties of the posterior distribution at the MAP estimator $\boldsymbol{x}^*$.

The limitation of this strategy comes from the fact that we are approximating global properties of the posterior distribution using only local information at the mode. Intuitively, this performs reasonably well if the greater part of the posterior probability mass is concentrated around a single point.

Mode-based covariance estimation for unconstrained maximum-a-posteriori inference is well-known and widely-used. If $\boldsymbol{x}^*$ is a MAP estimator for the statistical inverse problem

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{\mathcal{G}}(\boldsymbol{x}) + \boldsymbol{\xi},\\ \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{x}_0, \boldsymbol{\Sigma}),\\ \boldsymbol{\xi} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}),\end{aligned}$$

then the posterior covariance $\boldsymbol{\Sigma}_{\text{post}}$ can be approximated by

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\text{post}} &= \boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{G}}'(\boldsymbol{x}^*)^*\boldsymbol{\Gamma}^{-1}\boldsymbol{\mathcal{G}}'(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{1/2} + \mathbb{I}\right)^{-1}\boldsymbol{\Sigma}^{1/2}\\ &= \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\mathcal{G}}'(\boldsymbol{x}^*)^*\boldsymbol{\Gamma}^{-1}\boldsymbol{\mathcal{G}}'(x^*)\right)^{-1}.\end{aligned}$$

One can then compute an approximation $\tilde{C}_\alpha^{(1)}$ of C_α using $\hat{\boldsymbol{\Sigma}}_{\text{post}}$ (see [10] and [1, section 2.3]). The approximate HPD region $\tilde{C}_\alpha^{(1)}$ is defined as

$$\tilde{C}_\alpha^{(1)} = \left\{\, \boldsymbol{x} \in \mathbb{R}^n : \left\|\hat{\boldsymbol{\Sigma}}_{\text{post}}^{-1/2}(\boldsymbol{x} - \boldsymbol{x}*)\right\| \le \chi_n^2(1-\alpha) \,\right\},$$

where $\chi_n^2(1-\alpha)$ is the $(1-\alpha)$-quantile of the chi-squared distribution with n degrees of freedom. These estimates can be generalized to the constrained case as in [5].

Once we have computed $\hat{\boldsymbol{\Sigma}}_{\text{post}}$, we have to transform it back to a covariance for the original problem. This can be done by a Monte Carlo approach. If $\mathbf{Cov}(\boldsymbol{g}) = \hat{\boldsymbol{\Sigma}}_{\text{post}}$ and $\boldsymbol{f} = \boldsymbol{\Phi}^{-1}\boldsymbol{T}(\boldsymbol{g})$, we can simply sample $\boldsymbol{g}_1, \dots, \boldsymbol{g}_K \sim \mathcal{N}(\boldsymbol{g}_*, \hat{\boldsymbol{\Sigma}}_{\text{post}}$, set $\boldsymbol{f}_k = \boldsymbol{\Phi}^{-1}\boldsymbol{T}(\boldsymbol{g}_k)$ for $k = 1, \dots, K$, and estimate

$$\mathbf{Cov}(f) \approx \frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{f}_k - \bar{\boldsymbol{f}})(\boldsymbol{f}_k - \bar{f})^\top,$$

where $\bar{\boldsymbol{f}} = \frac{1}{K}\boldsymbol{f}_k$ is the sample mean.

An alternative estimator that works for linear problems is based on the *concentration of measure* phenomenon. This term describes a collection of mathematical and empirical results which state that the mass of high-dimensional probability distributions tends to be concentrated on lower-dimensional manifolds. Pereyra [13] used this phenomenon to derive an approximation $\tilde{C}_\alpha^{(2)}$ of the $(1-\alpha)$-HPD region. It is given by

$$\tilde{C}_\alpha^{(2)} = \{\, \boldsymbol{x} \in \mathbb{R}^n : \Phi(\boldsymbol{x}) \le \Phi(\boldsymbol{x}^*) + n(\tau_\alpha + 1) \,\},$$

where Φ is the MAP cost functional and $\tau_\alpha = \sqrt{\frac{16\log(3/\alpha)}{n}}$. It is then shown in [13] that $\tilde{C}_\alpha^{(2)}$ is a conservative estimate of C_α, i.e. we have

$$C_\alpha \subset \tilde{C}_\alpha^{(2)}.$$

This is a very nice result, because it guarantees that we do not underestimate the uncertainty in our results. Alas, it only applies to log-concave probability distributions (e.g. Gaussians), which in our case is equivalent to the assumption that $\boldsymbol{\mathcal{G}}$ is linear. (But I mailed Pereyra and he said that if $\boldsymbol{\mathcal{G}}$ is not too nonlinear around $\boldsymbol{x}_*$, $\tilde{C}_\alpha^{(2)}$ should still be a useful approximation of C_α, especially if the regularization term is l^1, as is the case in our situation! He suggested to use a convex approximation of the cost functional around $\boldsymbol{x}_*$, which we obtain for example by linearizing the observation operator.) We note also that in our situation, $C_\alpha^{(2)}$ can be computed for the non-transformed problem, since all we need is the MAP estimator.

To summarize, we have presented two mode-based approximations of the posterior HPD region. Both essentially assume that the model is linear, but we hope that they still give useful results in our mildly nonlinear case. The idea behind the first approximation $C_\alpha^{(1)}$ is classical (it can found in research from 60

years ago [2]) and well-established as a tool in nonlinear regression [1]. The second approximation $C_\alpha^{(2)}$ is very modern (2017) and there are yet no papers that have applied it to nonlinear inverse problems, and only a couple studies in the linear case (for example, [6] and [14]). Since both $C_\alpha^{(1)}$ and $C_\alpha^{(2)}$ can be essentially computed for free once we have obtained the MAP estimator $\boldsymbol{x}_*$, we suggest to use both and see how they differ. It would also be wise to perform a preliminary experiment where MCMC-sampling is still doable. Then, the credible region computed by MCMC can serve as a gold standard with respect to which we can compare the performance of the approximations $C_\alpha^{(1)}$ and $C_\alpha^{(2)}$.

4. What remains to be done

If we agree on the computational method, we can finally start implementing.

- *Prior hyperparameters*: We have some hyperparameters that have to be tuned. Among them are the prior covariances Σ_1 and Σ_2 and the prior mean $\bar{\theta}_v$ for the hyperparameters of the Gauss-Hermite interpolation. We should start with something simple, like choosing Σ_1 and Σ_2 to be diagonal matrices, and then adapt if we have first numerical results.
- *Curvelet transform and optimization method*: There is already a good implementation of the fast curvelet transform in CurveLab (`www.curvelet.org/software`), which is freely available for non-commercial use. For the CGN method, I know some people that I could mail for code, but it will probably be easier in the long run if we implement it by ourselves.
- *Uncertainty quantification*: As discussed above, computation of confidence intervals are something that can easily be added once we have implemented the CGN method. Both of the presented approximations do not add much computational cost to the overall method, and so we can simply use both and compare them, either with each other, or with the results from the already existing Hamiltonian Monte Carlo implementation.

4.1. Open questions.

(i) How does the discretization of the measurement P_{N_y} look like (see Section 1.2)? Is it accurate to write the discretized measurement $\bar{\boldsymbol{y}}$ as

$$\bar{y}_j = \int_{I_j} \frac{1}{1+v/c} S\left(\frac{\lambda}{1+v/c}; f\right) L(v;\boldsymbol{\theta})\,\mathrm{d}v,$$

where $I_j = [\lambda_j, \lambda_{j+1}) \subset \mathbb{R}^+$ is the j-th wavelength bin? How did you implement it?

(ii) As an alternative to curvelets, we could also use a Gaussian mixture model, i.e. try to represent $\boldsymbol{f}$ by a superposition of Gaussians, which can be generically written as

$$\sum_{k=1}^{K} \phi_k \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{\Sigma}_k), \qquad \phi_1, \dots, \phi_K \in \mathbb{R}.$$

This would simplify the later implementation, since for example the nonnegativity condition $f \geq 0$ is much easier to ensure for a Gaussian mixture (just choose $\phi_k \geq 0$ for all k). Furthermore, the parameters of a Gaussian mixture model are much easier to interpret than the parameters of a curvelet model. So, Gaussian mixtures could also be a viable alternative. But this hinges of course on the following question:

Is the distribution function of a galaxy well-modelled by a mixture of Gaussians?

4.2. Wishlist. This is a non-exhaustive list of things that we need, or that would be at least very helpful, before we can start implementing our approach.

(i) *Observation operator*: An implementation of the input-output operator $\mathcal{G}$ given by (1.4). Ideally an implemented function that takes $\boldsymbol{f}$ and $\boldsymbol{\theta}_v$ as input and returns the corresponding measurement $\bar{\boldsymbol{y}}$. It would be helpful if this part is well-documented, so that we can adapt it quickly to the curvelet formulation.

(ii) *Fake data simulator*: It would also be vital for testing to have a program that can simulate distribution functions of a galaxy. Ideally, the program would randomly generate realizations of $\boldsymbol{f}$ that look (at least approximately) like distribution functions of real galaxies. Only then can we check how well the sparsity-promoting regularization is working.

(iii) *Access to the Monte Carlo implementation*: It would also be very helpful as a reference to have access to the already existing Hamiltonian Monte Carlo implementation for the problem, including a description of the made choices (what are the prior distributions that you used, etc.). The program should be in a form where we can give it a noisy measurement $\boldsymbol{y}$ as input (and maybe also tune the hyperparameters of the prior) and it automatically generates samples from the posterior distribution of $\boldsymbol{f}$ and $\boldsymbol{\theta}_v$, given $\boldsymbol{y}$.

References

[1] D. M. Bates and D. G. Watts. "Nonlinear Regression Analysis and Its Applications". John Wiley & Sons, Inc., 1988. DOI: 10.1002/9780470316757 (cited on pages 11, 12).

[2] E. M. L. Beale. "Confidence Regions in Non-Linear Estimation". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 22.1 (1960), pp. 41–76. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1960.tb00353.x (cited on page 12).

[3] H. G. Bock. "Numerical Treatment of Inverse Problems in Chemical Reaction Kinetics". In: *Springer Series in Chemical Physics*. Springer Berlin Heidelberg, 1981, pp. 102–125. DOI: 10.1007/978-3-642-68220-9_8 (cited on page 10).

[4] H. G. Bock. "Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen". PhD thesis. Universität Bonn, 1985 (cited on page 10).

[5] H. G. Bock, E. Kostina, and O. Kostyukova. "Covariance Matrices for Parameter Estimates of Constrained Parameter Estimation Problems". In: *SIAM Journal on Matrix Analysis and Applications* 29.2 (2007), pp. 626–642. ISSN: 0895-4798. DOI: 10.1137/040617893 (cited on page 11).

[6] X. Cai, M. Pereyra, and J. D. McEwen. "Uncertainty quantification for radio interferometric imaging: II. MAP estimation". In: *Monthly Notices of the Royal Astronomical Society* 480.3 (2018), pp. 4170–4182. ISSN: 0035-8711. DOI: 10.1093/mnras/sty2015 (cited on page 12).

[7] E. Candès, L. Demanet, D. Donoho, and L. Ying. "Fast discrete curvelet transforms". In: *Multiscale Modeling & Simulation* 5 (2006), pp. 861–899. ISSN: 1540-3459 (cited on page 6).

[8] E. J. Candès and D. L. Donoho. "Continuous curvelet transform II: Discretization and frames". In: *Applied and Computational Harmonic Analysis* 19.2 (2005), pp. 198–222. ISSN: 1063-5203. DOI: 10.1016/j.acha.2005.02.004 (cited on page 6).

[9] P. Lambert, S. Pires, J. Ballot, R. A. García, J.-L. Starck, and S. Turck-Chièze. "Curvelet analysis of asteroseismic data". In: *Astronomy & Astrophysics* 454.3 (2006), pp. 1021–1027. DOI: 10.1051/0004-6361:20054541 (cited on page 4).

[10] D. Lu, M. Ye, and M. C. Hill. "Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification". In: *Water Resources Research* 48.9 (2012). ISSN: 0043-1397. DOI: 10.1029/2011wr011289 (cited on page 11).

[11] J. Ma and G. Plonka. "The Curvelet Transform". In: *IEEE Signal Processing Magazine* 27.2 (2010), pp. 118–133. DOI: 10.1109/msp.2009.935453 (cited on page 4).

[12] S. G. Mallat. "A wavelet tour of signal processing: the sparse way". 3rd ed. Amsterdam; Boston: Elsevier/Academic Press, 2009 (cited on page 5).

[13] M. Pereyra. "Maximum-a-Posteriori Estimation with Bayesian Confidence Regions". In: *SIAM Journal on Imaging Sciences* 10.1 (2017), pp. 285–302. ISSN: 1936-4954. DOI: 10.1137/16m1071249 (cited on page 11).

[14] A. Repetti, M. Pereyra, and Y. Wiaux. "Scalable Bayesian Uncertainty Quantification in Imaging Inverse Problems via Convex Optimization". In: *SIAM Journal on Imaging Sciences* 12.1 (2019), pp. 87–118. ISSN: 1936-4954. DOI: 10.1137/18m1173629 (cited on page 12).

[15] J. L. Starck, D. L. Donoho, and E. J. Candès. "Astronomical image representation by the curvelet transform". In: *Astronomy & Astrophysics* 398.2 (2003), pp. 785–800. DOI: 10.1051/0004-6361:20021571 (cited on page 4).

[16] C. A. Zarzer. "On Tikhonov regularization with non-convex sparsity constraints". In: *Inverse Problems* 25 (2009), p. 025006. ISSN: 0266-5611 (cited on pages 8, 9).