Astronomy & Astrophysics manuscript no. D3PO
November 11, 2013

Denoising, Deconvolving, and Decomposing Photon Observations

Derivation of the D^{3}PO Algorithm

Marco Selig[1,2] and Torsten A. Enßlin[1,2]

[1] Max Planck Institut für Astrophysik (Karl-Schwarzschild-Straße 1, D-85748 Garching, Germany)
[2] Ludwig-Maximilians-Universität München (Geschwister-Scholl-Platz 1, D-80539 München, Germany)

Received 07 Nov. 2013 / Accepted DD MMM. YYYY

ABSTRACT

The analysis of astronomical images is a non-trivial task. The D^{3}PO algorithm addresses the inference problem of denoising, deconvolving, and decomposing photon observations. The primary goal is the simultaneous reconstruction of the diffuse and point-like photon flux from a given photon count image. In order to discriminate between these morphologically different signal components, a probabilistic algorithm is derived in the language of information field theory based on a hierarchical Bayesian parameter model. The signal inference exploits prior information on the spatial correlation structure of the diffuse component and the brightness distribution of the spatially uncorrelated point-like sources. A maximum *a posteriori* solution and a solution minimizing the Gibbs free energy of the inference problem using variational Bayesian methods are discussed. Since the derivation of the solution does not dependent on the underlying position space, the implementation of the D^{3}PO algorithm uses the NIFTy package to ensure operationality on various spatial grids and at any resolution. The fidelity of the algorithm is validated by the analysis of simulated data, including a realistic high energy photon count image showing a $32 \times 32\,\mathrm{arcmin}^2$ observation with a spatial resolution of 0.1 arcmin. In all tests the D^{3}PO algorithm successfully denoised, deconvolved, and decomposed the data into a diffuse and a point-like signal estimate for the respective photon flux components.

Key words. methods: data analysis – methods: numerical – methods: statistical – techniques: image processing – gamma-rays: general – X-rays: general

1. Introduction

An astronomical image might display multiple superimposed features, such as "point sources", "compact objects", "diffuse emission", or "background radiation". The raw photon count images delivered by high energy telescopes are far from perfect suffering from shot noise and distortions due to instrumental effects. The analysis of such astronomical observations demands elaborate denoising, deconvolution and decomposition strategies.

The data obtained by the detection of individual photons is subject to Poissonian shot noise which is more severe for low count rates. This cumbers the discrimination of faint sources against noise, and makes their detection exceptionally challenging. Furthermore, uneven or incomplete survey coverage and complex instrumental response functions leave imprints in the photon data. As a result, the data set might exhibit gaps and artificial distortions rendering the clear recognition of different features a difficult task. Especially point-like sources are afflicted by the instrument's point spread function (PSF) that smoothes them out in the observed image, and therefore can cause fainter ones to vanish completely in the background noise.

In addition to such noise and convolution effects, it is the superposition of the different objects that makes their separation ambiguous, if possible at all. In astrophysics, photon emitting objects are commonly divided into two morphological classes, diffuse sources and point sources. Diffuse sources span rather smoothly across large fractions of an image, and exhibit apparent internal correlations. Point sources, on the contrary, are local features that, if observed perfectly, would only appear in one pixel of the image. In this work, we will not distinguish between diffuse sources and background, both are diffuse contributions. Intermediate cases, which are sometimes classified as "extended" or "compact" sources, are also not considered here.

The question arises, how to reconstruct the original source contributions, the individual signals, that caused the observed photon data. This task is an ill-posed inverse problem without an unique solution. There are a number of heuristic and probabilistic approaches to this problem.

SExtractor (Bertin & Arnouts 1996) is one of the heuristic kind and the most prominent tool for identifying sources in astronomy. Its popularity is mostly based on its speed and easy operability. CLEAN (Högbom 1974) is commonly used in radio astronomy and attempts a deconvolution assuming there are only contributions from point sources. Decomposition techniques for diffuse backgrounds, based on the analysis of angular power spectra, have recently been proposed by Hensley et al. (2013). Other successful techniques exploit wavelet transformations (based on the work by Haar 1910, 1911) and thresholding for source separation (e.g., González-Nuevo et al. 2006).

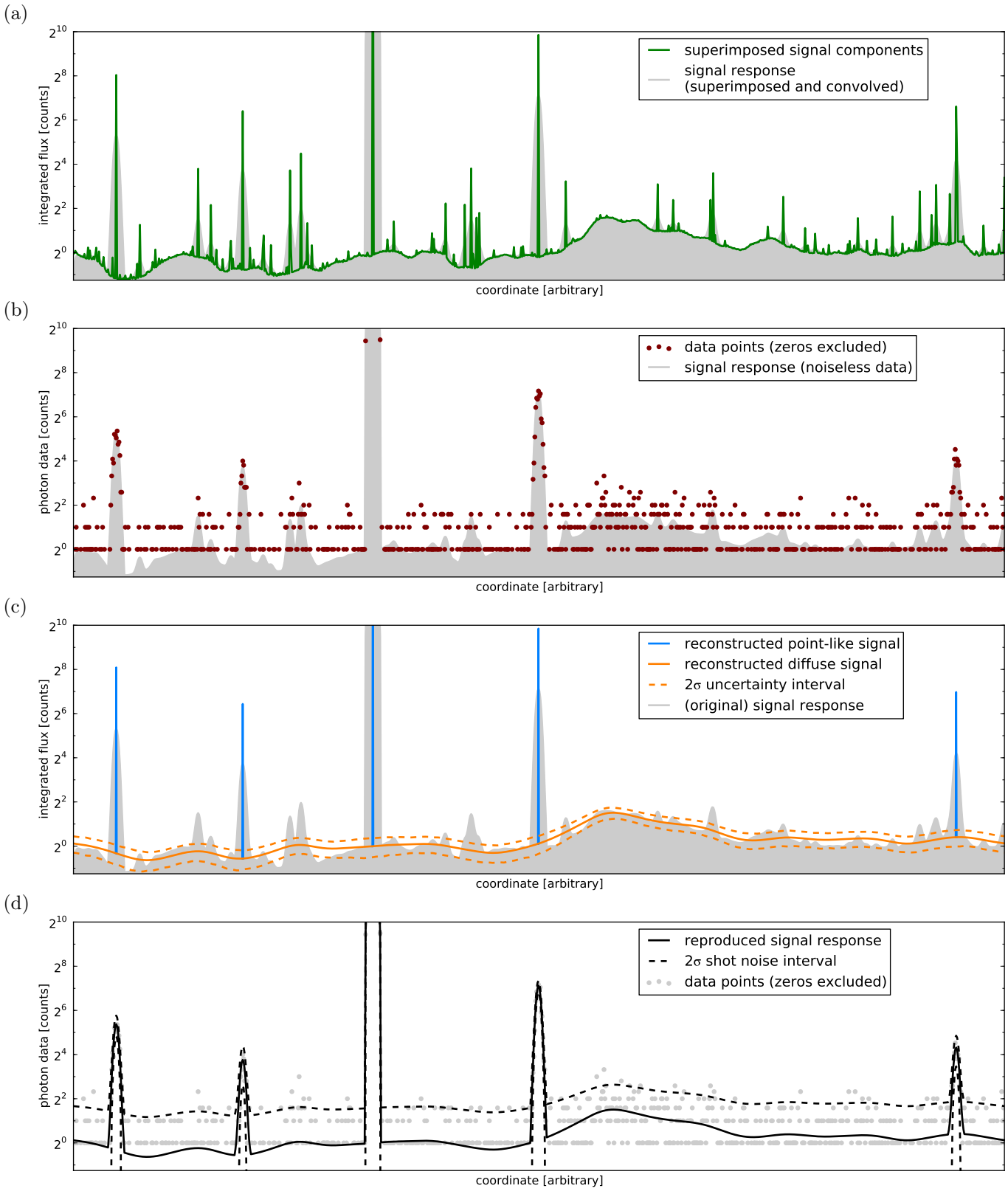

Fig. 1. Illustration of a 1D reconstruction scenario with 1024 pixels. Panel (a) shows the superimposed diffuse and point-like signal components (green solid line) and its observational response (gray contour). Panel (b) shows again the signal response representing noiseless data (gray contour) and the generated Poissonian data (red markers). Panel (c) shows the reconstruction of the point-like signal component (blue solid line), the diffuse one (orange solid line), its 2σ reconstruction uncertainty interval (orange dashed line), and again the original signal response (gray contour). The point-like signal comprises 1024 point-sources of which only 5 are not invisibly faint. Panel (d) shows the reproduced signal response representing noiseless data (black solid line), its 2σ shot noise interval (black dashed line), and again the data (gray markers).

Inference methods, in contrast, investigate the probabilistic relation between the data and the signals. Here, the signals of interest are the different source contributions. Probabilistic approaches allow a transparent incorporation of model and *a priori* assumptions, but often result in computationally heavier algorithms. As an initial attempt, a maximum likelihood analysis was proposed by Valdes (1982). In later work, maximum entropy methods were applied to the INTEGRAL/SPI data (Strong 2003), and a Bayesian model was used to analyze the ROSAT data (Guglielmetti et al. 2009). The background-source separation technique of this latter approach is based on a two-component mixture model that reconstructs (extended) sources and background concurrently. The fast algorithm PowellSnakesI/II by Carvalho et al. (2009, 2012) is capable of analyzing multi-frequency data sets and detecting point-like objects within diffuse emission regions. It relies on matched filters using PSF templates and Bayesian filters exploiting, among others, priors on source position, size, and number. PowellSnakes II has been successfully applied to the Planck data (Planck Collaboration et al. 2011).

The strategy presented in this work aims at the simultaneous reconstruction of two signals, the diffuse and point-like photon flux. Both fluxes contribute equally to the observed photon counts, but their morphological imprints are very different. The proposed algorithm, derived in the framework of information field theory (IFT) (Enßlin et al. 2009; Enßlin 2012), therefore incorporates prior assumptions in form of a hierarchical parameter model. The fundamentally different morphologies of diffuse and point-like contributions reflected in different prior correlations and statistics. The exploitation of these different prior models is key to the signal decomposition.

The diffuse and point-like signal are treated as two separate signal fields. A signal field represents an original signal appearing in nature; e.g., the physical photon flux distribution of one source component as a function of real space or sky position. In theory, a field has infinitely many degrees of freedom being defined on a continuous position space. In computational practice, however, a field needs of course to be defined on a finite grid. It is desirable that the signal field is reconstructed independently from the grid's resolution, except for potentially unresolvable features.[1] Notice that the point-like signal field hosts one point source in every pixel, however, most of them might be invisibly faint. Hence, a complicated determination of the number of point sources, as many algorithms perform, is not required in our case.

Furthermore, the proposed algorithm reconstructs the harmonic power spectrum of the diffuse component from the data itself, and provides uncertainty information on both inferred signals. The derivation of the algorithm makes use of a wide range of Bayesian methods that are discussed below in detail with regard to their implications and applicability.

Fig. 1 illustrates a reconstruction scenario in one dimension, where the coordinate could be an angle or position (or time, or energy) in order to represent a 1D sky (or a time series, or an energy spectrum). The numerical implementation uses the NIFTy[2] package (Selig et al. 2013). NIFTy permits that an algorithm can be set up abstractly, independent of the finally chosen topology, dimension, or resolution of the underlying position space. In this way, a 1D prototype code can be used for development, and then just be applied in 2D, 3D, or even on the sphere.

The remainder of this paper is structured as follows. Sec. 2 discusses the inference on photon observations; i.e., the underlying model and prior assumptions. The D^{3}PO algorithm solving this inference problem by denoising, deconvolution, and decomposition is derived in Sec. 3. In Sec. 4 the algorithm is demonstrated in a numerical application on simulated high energy photon data. We conclude in Sec. 5.

2. Inference on Photon Observations

2.1. Signal Inference

Here, a signal is defined as an unknown quantity of interest that one wants to learn about. The most important information source on a signal is usually the data obtained in an observation to measure the signal. Inferring a signal from an observational data set poses a fundamental problem due to the presence of noise in the data and the ambiguity that several possible signals could have produced the same data, even in the case of negligible noise.

Any realistic experiment involves measurement noise that is the sum of imprints from all occurring random processes. In order not to draw false conclusions, it is crucial to understand the sources and probabilistic properties of the noise, and to consider them in the data analysis. Furthermore, the experimental data may suffer from incomplete information that renders the signal reconstruction ambiguous. For example, given some image data like photon counts, we want to infer the underlying photon flux distribution. This physical flux is a continuous scalar field that varies with respect to time, energy, and observational position. The measured photon count data, however, is restricted by its spatial and energy binning, as well as its limitations in energy range and observation time. Basically, all data sets are finite for practical reasons, and therefore cannot capture all of the infinitely many degrees of freedom of a continuous signal field.

There is no exact solution to such signal inference problems, since there might be (infinitely) many signal field configurations that could lead to the same data. This is why a probabilistic data analysis, which does not pretend to calculate the correct field configuration but provides expectation values and uncertainties of the signal field, is appropriate for signal inference.

Given a data set $\boldsymbol{d}$, the *a posteriori* probability distribution $P(\boldsymbol{s}|\boldsymbol{d})$ judges how likely a potential signal $\boldsymbol{s}$ is. This posterior is given by Bayes' theorem,

$$P(\boldsymbol{s}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{s})\,P(\boldsymbol{s})}{P(\boldsymbol{d})}, \tag{1}$$

[1] If the resolution of the reconstruction would be increased gradually, the diffuse signal field might exhibit more and more small scale features until the information content of the given data is exhausted. From this point on, any further increase in resolution would not change the signal field significantly. In a similar manner, the localization accuracy and number of detections of point sources might increase with the resolution until all relevant information of the data was captured. All higher resolution grids can then be regarded as acceptable representations of the continuous position space.

[2] NIFTy homepage `http://www.mpa-garching.mpg.de/ift/nifty/`

as a combination of the likelihood $P(\boldsymbol{d}|\boldsymbol{s})$, the signal prior $P(\boldsymbol{s})$, and the evidence $P(\boldsymbol{d})$, which serves as a normalization. The likelihood characterizes how likely it is to measure data set $\boldsymbol{d}$ from a given signal field $\boldsymbol{s}$. It covers all processes that are relevant for the measurement of $\boldsymbol{d}$. The prior describes the knowledge about $\boldsymbol{s}$ without considering the data, and should, in general, be less restrictive than the likelihood.

IFT is a Bayesian framework for the inference of signal fields exploiting mathematical methods for theoretical physics. A signal field, $\boldsymbol{s} = s(x)$, is a function of a continuous position x in some position space Ω. In order to avoid a dependence of the reconstruction on the partition of Ω, the according calculus regarding fields is geared to preserve the continuum limit, cf. (Enßlin 2012; Selig et al. 2013). In general, we are interested in the *a posteriori* mean estimate $\boldsymbol{m}$ of the signal field given the data, and its (uncertainty) covariance $\boldsymbol{D}$, defined as

$$\boldsymbol{m} = \langle \boldsymbol{s} \rangle_{(\boldsymbol{s}|\boldsymbol{d})} = \int \mathcal{D}\boldsymbol{s}\ \boldsymbol{s}\ P(\boldsymbol{s}|\boldsymbol{d}), \tag{2}$$

$$\boldsymbol{D} = \left\langle (\boldsymbol{m} - \boldsymbol{s})(\boldsymbol{m} - \boldsymbol{s})^\dagger \right\rangle_{(\boldsymbol{s}|\boldsymbol{d})}, \tag{3}$$

where $\dagger$ denotes adjunction and $\langle \cdot \rangle_{(\boldsymbol{s}|\boldsymbol{d})}$ the expectation value with respect to the posterior probability distribution $P(\boldsymbol{s}|\boldsymbol{d})$.[3]

In the following, the posterior of the physical photon flux distribution of two morphologically different source components given a photon count data set is build up piece by piece according to Eq. (1).

2.2. Poissonian Likelihood

The images provided by astronomical high energy telescopes typically consist of integer photon counts that are binned spatially into pixels. Let d_i be the number of detected photons, also called events, in pixel i, where $i \in \{1, \ldots, N_{\mathrm{pix}}\} \subset \mathbb{N}$.

The kind of signal field we would like to infer from such data is the causative photon flux distribution. The photon flux, $\boldsymbol{\rho} = \rho(x)$, is defined for each position x on the observational space Ω. In astrophysics, this space Ω is typically the $\mathcal{S}^2$ sphere representing an all-sky view, or a region within $\mathbb{R}^2$ representing an approximately plane patch of the sky. The flux $\boldsymbol{\rho}$ might express different morphological features, which can be classified into a diffuse and point-like component. The exact definitions of the diffuse and point-like flux should be specified *a priori*, without knowledge of the data, and are addressed in Sec. 2.3.1 and 2.3.3, respectively. At this point it shall suffices to say that the diffuse flux varies smoothly on large spatial scales, while the flux originating from point sources is fairly local. These two flux components are superimposed,

$$\boldsymbol{\rho} = \boldsymbol{\rho}_{\mathrm{diffuse}} + \boldsymbol{\rho}_{\mathrm{point-like}} = \rho_0 \left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right), \tag{4}$$

where we introduced the dimensionless diffuse and point-like signal fields, $\boldsymbol{s}$ and $\boldsymbol{u}$, and the constant ρ_0 which absorbs the physical dimensions of the photon flux; i.e., events per area per energy and time interval. The exponential function in Eq. (4) is applied componentwise and ensures a strictly positive photon flux.

A measurement apparatus observing the photon flux $\boldsymbol{\rho}$ is expected to detect a certain number of photons $\boldsymbol{\lambda}$. This process can be modeled by a linear response operator $\boldsymbol{R}_0$ as follows,

$$\boldsymbol{\lambda} = \boldsymbol{R}_0 \boldsymbol{\rho} = \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right), \tag{5}$$

where $\boldsymbol{R} = \boldsymbol{R}_0 \rho_0$. This reads for pixel i,

$$\lambda_i = \int_\Omega \mathrm{d}x\ R_i(x) \left(\mathrm{e}^{s(x)} + \mathrm{e}^{u(x)}\right). \tag{6}$$

The response operator $\boldsymbol{R}_0$ comprises all aspects of the measurement process; i.e., all instrument response functions. This includes the survey coverage, which describes the instrument's overall exposure to the observational area, and the instrument's PSF, which describes how a point source is imaged by the instrument.

The superposition of different components and the transition from continuous coordinates to some discrete pixelization, cf. Eq. (6), cause a severe loss of information about the original signal fields. In addition to that, measurement noise distorts the signal's imprint in the data. The individual photon counts per pixel can be assumed to follow a Poisson distribution $\mathcal{P}$ each. Therefore, the likelihood of the data $\boldsymbol{d}$ given an expected number of events $\boldsymbol{\lambda}$ is modeled as a product of statistically independent Poisson processes,

$$P(\boldsymbol{d}|\boldsymbol{\lambda}) = \prod_i \mathcal{P}(d_i, \lambda_i) = \prod_i \frac{1}{d_i!}\ \lambda_i^{d_i} \mathrm{e}^{-\lambda_i}. \tag{7}$$

The Poisson distribution has a signal-to-noise ratio of $\sqrt{\boldsymbol{\lambda}}$ which scales with the expected number of photon counts. Therefore, Poissonian shot noise is most severe in regions with low photon fluxes. This makes the detection of faint sources in high energy astronomy a particularly challenging task, as X- and γ-ray photons are sparse.

The likelihood of photon count data given a two component photon flux is hence described by the Eqs. (7) and (5). Rewriting this likelihood $P(\boldsymbol{d}|\boldsymbol{s}, \boldsymbol{u})$ in form of its negative logarithm yields the information Hamiltonian $H(\boldsymbol{d}|\boldsymbol{s}, \boldsymbol{u})$,[4]

$$\begin{aligned} H(\boldsymbol{d}|\boldsymbol{s}, \boldsymbol{u}) &= -\log P(\boldsymbol{d}|\boldsymbol{s}, \boldsymbol{u}) && (8)\\ &= H_0 + \mathbf{1}^\dagger \boldsymbol{\lambda} - \boldsymbol{d}^\dagger \log(\boldsymbol{\lambda}) && (9)\\ &= H_0 + \mathbf{1}^\dagger \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right) - \boldsymbol{d}^\dagger \log\left(\boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right)\right), && (10) \end{aligned}$$

where the ground state energy H_0 comprises all terms constant in $\boldsymbol{s}$ and $\boldsymbol{u}$, and $\mathbf{1}$ is a constant data vector being 1 everywhere.

2.3. Prior Assumptions

The diffuse and point-like signal fields, $\boldsymbol{s}$ and $\boldsymbol{u}$, contribute equally to the likelihood defined by Eq. (10), and thus leaving it completely degenerate. On the mere basis of the likelihood, the full data set could be explained by the diffuse

[3] This expectation value is computed by a path integral, $\int \mathcal{D}\boldsymbol{s}$, over the complete phase space of the signal field $\boldsymbol{s}$; i.e. all possible field configurations.

[4] Throughout this work we define $H(\,\cdot\,) = -\log P(\,\cdot\,)$, and absorb constant terms into a normalization constant H_0 in favor of clarity.

signal alone, or by point-sources only, or any other conceivable combination. For this reason, priors are introduced.

The priors introduced in the following address the morphology of the different photon flux contributions, and define "diffuse" and "point-like" in the first place. These priors aid the reconstruction by providing some remedy for the degeneracy of the likelihood. For a decomposition of the total photon flux, the introduction of priors is imperative. Nevertheless, the reconstruction of the total photon flux $\boldsymbol{\rho}$ is still primarily data driven.

2.3.1. Diffuse Component

The diffuse photon flux, $\boldsymbol{\rho}^{(s)} = \rho_0 \mathrm{e}^{\boldsymbol{s}}$, is strictely positive and might vary in intensity over several orders of magnitude. Its morphology shows cloudy patches with smooth fluctuations across spatial scales; i.e., one expects similar values of the diffuse flux in neighboring locations. In other words, the diffuse component exhibits spatial correlations. A log-normal model for $\boldsymbol{\rho}^{(s)}$ satisfies those requirements according to the maximum entropy principle (Oppermann et al. 2012). If the diffuse photon flux follows a multi-variant log-normal distribution, the diffuse signal field $\boldsymbol{s}$ obeys a multi-variant Gaussian distribution $\mathcal{G}$,

$$P(\boldsymbol{s}|\boldsymbol{S}) = \mathcal{G}(\boldsymbol{s},\boldsymbol{S}) = \frac{1}{\sqrt{\det[2\pi\boldsymbol{S}]}}\ \exp\left(-\frac{1}{2}\boldsymbol{s}^\dagger \boldsymbol{S}^{-1}\boldsymbol{s}\right), \quad (11)$$

with a given covariance $\boldsymbol{S} = \left\langle \boldsymbol{s}\boldsymbol{s}^\dagger \right\rangle_{(\boldsymbol{s}|\boldsymbol{S})}$. This covariance describes the strength of the spatial correlations, and thus the smoothness of the fluctuations.

A convenient parameterization of the covariance $\boldsymbol{S}$ can be found, if the signal field $\boldsymbol{s}$ is *a priori* not known to distinguish any position or orientation axis; i.e., its correlations only depend on relative distances. This is equivalent to assume $\boldsymbol{s}$ to be statistically homogeneous and isotropic. Under this assumption, $\boldsymbol{S}$ is diagonal in the harmonic basis[5] of the position space Ω such that

$$\boldsymbol{S} = \sum_k \mathrm{e}^{\tau_k} \boldsymbol{S}_k, \quad (12)$$

where τ_k are spectral parameters and $\boldsymbol{S}_k$ are projections onto a set of disjoint harmonic subspaces of Ω. These subspaces are commonly denoted as spectral bands or harmonic modes. The set of spectral parameters, $\boldsymbol{\tau} = \{\tau_k\}_k$, is then the logarithmic power spectrum of the diffuse signal field $\boldsymbol{s}$ with respect to the chosen harmonic basis denoted by k.

However, the diffuse signal covariance is in general unknown *a priori*. This requires the introduction of another prior for the covariance, or for the set of parameters $\boldsymbol{\tau}$ describing it adequately. This approach of hyperpriors on prior parameters creates a hierarchical parameter model.

2.3.2. Unknown Power Spectrum

The lack of knowledge of the power spectrum, requires its reconstruction from the same data the signal is inferred from (Wandelt et al. 2004; Jasche et al. 2010; Enßlin & Frommert 2011). Therefore, two *a priori* constraints for the spectral parameters $\boldsymbol{\tau}$, which describe the logarithmic power spectrum, are incorporated in the model.

The power spectrum is unknown and might span over several orders of magnitude. This implies a logarithmically uniform prior for each element of the power spectrum, and a uniform prior for each spectral parameter τ_k, respectively. Let us initially assume independent inverse-Gamma distributions $\mathcal{I}$ for the individual elements,

$$P(\mathrm{e}^{\boldsymbol{\tau}}|\boldsymbol{\alpha},\boldsymbol{q}) = \prod_k \mathcal{I}(\mathrm{e}^{\tau_k},\alpha_k,q_k) \quad (13)$$

$$= \prod_k \frac{q_k^{\alpha_k-1}}{\Gamma(\alpha_k-1)}\ \mathrm{e}^{-\left(\alpha_k\tau_k + q_k \mathrm{e}^{-\tau_k}\right)}, \quad (14)$$

and hence

$$P_{\mathrm{un}}(\boldsymbol{\tau}|\boldsymbol{\alpha},\boldsymbol{q}) = \prod_k \mathcal{I}(\mathrm{e}^{\tau_k},\alpha_k,q_k)\left|\frac{\mathrm{d}\mathrm{e}^{\tau_k}}{\mathrm{d}\tau_k}\right| \quad (15)$$

$$\propto \exp\Big(-(\boldsymbol{\alpha}-\mathbf{1})^\dagger\boldsymbol{\tau} - \boldsymbol{q}^\dagger \mathrm{e}^{-\boldsymbol{\tau}}\Big), \quad (16)$$

where $\boldsymbol{\alpha} = \{\alpha_k\}_k$ and $\boldsymbol{q} = \{q_k\}_k$ are the shape and scale parameters, and Γ denotes the Gamma function. In the limit of $\alpha_k \to 1$ and $q_k \to 0\ \forall k$, the inverse-Gamma distributions become asymptotically flat on a logarithmic scale, and thus P_{un} constant.[6] Small non-zero scale parameters, $0 < q_k$, provide lower limits for the power spectrum that, in practice, lead to more stable inference algorithms.

So far, the variability of the individual elements of the power spectrum is accounted for, but the question about their correlations has not been addressed. Empirically, power spectra of a diffuse signal field do not exhibit wild fluctuation or change drastically over neighboring modes. They rather show some sort of spectral smoothness. Moreover, for diffuse signal fields that were shaped by local and causal processes, we might expect a finite correlation support in position space. This translates into a smooth power spectrum. In order to incorporate spectral smoothness, we employ a prior introduced by Enßlin & Frommert (2011); Oppermann et al. (2012). This prior is based on the second logarithmic derivative of the spectral parameters $\boldsymbol{\tau}$, and favors power spectra that obey a power-law. It reads

$$P_{\mathrm{sm}}(\boldsymbol{\tau}|\boldsymbol{\sigma}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\tau}^\dagger \boldsymbol{T}\boldsymbol{\tau}\right), \quad (17)$$

with

$$\boldsymbol{\tau}^\dagger\boldsymbol{T}\boldsymbol{\tau} = \int \mathrm{d}(\log k)\frac{1}{\sigma_k^2}\left(\frac{\partial^2\tau_k}{\partial(\log k)^2}\right)^2, \quad (18)$$

where $\boldsymbol{\sigma} = \{\sigma_k\}_k$ are Gaussian standard deviations specifying the tolerance against deviation from a power-law behavior of the power spectrum. In the limit of $\sigma_k \to \infty\ \forall k$, no smoothness is enforced upon the power spectrum.

The resulting prior for the spectral parameters is given by the product of the priors discussed above,

$$P(\boldsymbol{\tau}|\boldsymbol{\alpha},\boldsymbol{q},\boldsymbol{\sigma}) = P_{\mathrm{un}}(\boldsymbol{\tau}|\boldsymbol{\alpha},\boldsymbol{q})\ P_{\mathrm{sm}}(\boldsymbol{\tau}|\boldsymbol{\sigma}). \quad (19)$$

The parameters $\boldsymbol{\alpha},\boldsymbol{q}$ and $\boldsymbol{\sigma}$ are considered to be given as part of the hierarchical Bayesian model, and provide a flexible handle to model our knowledge on the scaling and smoothness of the power spectrum.

[5] The basis in which the Laplace operator is diagonal is denoted harmonic basis. If Ω is a n-dimensional Euclidean space $\mathbb{R}^n$ or Torus $\mathcal{T}^n$, the harmonic basis is the Fourier basis; if Ω is the $\mathcal{S}^2$ sphere, the harmonic basis is the spherical harmonics basis.

[6] If $P(\tau_k = \log z) = \text{const.}$, then a substitution yields $P(z) = P(\log z)\ |\mathrm{d}(\log z)/\mathrm{d}z| \propto z^{-1} \sim \mathcal{I}(z, \alpha \to 1, q \to 0)$.

2.3.3. Point-like Component

The point-like photon flux, $\boldsymbol{\rho}^{(u)} = \rho_0 \mathrm{e}^{\boldsymbol{u}}$, is supposed to originate from very distant astrophysical sources. These sources appear morphologically point-like to an observer because their actual extent is negligible due to the extreme distances. This renders point sources to be spatially local phenomena. The photon flux contributions of neighboring point sources can (to zeroth order approximation) be assumed to be statistically independent of each other. Even if the two sources are very close on the observational plane, their physical distance might be huge. Therefore, any *a priori* spatial correlation between point sources are ignored. Statistically independent priors for the photon flux contribution of each point-source are assumed in the following.

Due to the spatial locality of a point source, the corresponding photon flux signal is supposed to be confined to a single spot, too. If the point-like signal field, defined over a continuous position space Ω, is discretized properly[7], this spot is sufficiently identified by an image pixel in the reconstruction. A discretization, $\rho(x \in \Omega) \to (\rho_x)_x$, is an inevitable step since the algorithm is to be implemented in a computer environment anyway. Nevertheless, we have to ensure that the *a priori* assumptions do not depend on the chosen discretization but satisfy the continuous limit.

Therefore, the prior for the point-like signal component factorizes spatially,

$$P(\boldsymbol{\rho}^{(u)}) = \prod_x P(\rho_x^{(u)}), \tag{20}$$

but the functional form of the priors are yet to be determined. This model allows the point-like signal field to host one point source in every pixel. Most of these point sources are expected to be invisibly faint contributing negligibly to the total photon flux. However, the point sources which are just identifiable from the data are pinpointed in the reconstruction. In this approach, there is no necessity for a complicated determination of the number and position of sources.

For the construction of a prior, it further needs to be considered that the photon flux is a strictely positive quantity. Thus, a simple exponential prior,

$$P(\rho_x^{(u)}) \propto \exp\left(-\rho_x^{(u)}/\rho_0\right), \tag{21}$$

has been suggested (e.g., Guglielmetti et al. 2009). It has the advantage of being (easily) analytically treatable, but its physical implications are questionable. This distribution strongly suppresses high photon fluxes in favor of lower ones. The maximum entropy prior, which is also often applied, is even worse because it corresponds to a brightness distribution,[8]

$$P(\rho_x^{(u)}) \propto \left(\rho_x^{(u)}/\rho_0\right)^{\left(-\rho_x^{(u)}/\rho_0\right)}. \tag{22}$$

The following (rather crude) consideration might motivate a more astrophysical prior. Say the universe hosts a homogeneous distribution of point sources. The number of point sources would therefore scale with the observable volume; i.e., with distance cubed. Their apparent brightness, which is reduced due to the spreading of the light rays; i.e., a proportionality to the distance squared. Consequently, a power-law behavior between the number of point sources and their brightness with a slope $\beta = \frac{3}{2}$ is to be expected (Fomalont 1968; Malyshev & Hogg 2011). However, such a plain power-law diverges at 0, and is not necessarily normalizable. Furthermore, galactic and extragalactic sources can not be found in arbitrary distances due to the finite size of the Galaxy and the cosmic (back) light cone. Imposing an exponential cut-off above 0 onto the power-law yields an inverse-Gamma distribution, which has been shown to be an appropriate prior for point-like photon fluxes (Guglielmetti et al. 2009; Carvalho et al. 2009, 2012).

The prior for the point-like signal field is therefore derived from a product of independent inverse-Gamma distributions,[9]

$$P(\boldsymbol{\rho}^{(u)}|\boldsymbol{\beta},\boldsymbol{\eta}) = \prod_x \mathcal{I}(\rho_x^{(u)}, \beta_x, \rho_0\eta_x) \tag{23}$$

$$= \prod_x \frac{(\rho_0\eta_x)^{\beta_x - 1}}{\Gamma(\beta_x - 1)} \left(\rho_x^{(u)}\right)^{-\beta_x} \exp\left(-\frac{\rho_0\eta_x}{\rho_x^{(u)}}\right), \tag{24}$$

yielding

$$P(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\eta}) = \prod_x \mathcal{I}(\rho_0 \mathrm{e}^{u_x}, \beta_k, \rho_0\eta_k) \left|\frac{\mathrm{d}\rho_0 \mathrm{e}^{u_x}}{\mathrm{d}u_x}\right| \tag{25}$$

$$\propto \exp\Big(-(\boldsymbol{\beta} - \mathbf{1})^\dagger \boldsymbol{u} - \boldsymbol{\eta}^\dagger \mathrm{e}^{-\boldsymbol{u}}\Big), \tag{26}$$

where $\boldsymbol{\beta} = \{\beta_x\}_x$ and $\boldsymbol{\eta} = \{\eta_x\}_x$ are the shape and scale parameters. The latter is responsible for the cut-off of vanishing fluxes, and should be chosen adequately small in analogy to the spectral scale parameters $\boldsymbol{q}$. The determination of the shape parameters is more difficile. The geometrical argument above suggests a universal shape parameter, $\beta_x = \frac{3}{2}\ \forall x$. A second argument for this value results from demanding *a priori* independence of the discretization. If we choose a coarser resolution that would add up the flux from two point sources at merged pixels, then our prior should still be applicable. The universal value of $\frac{3}{2}$ indeed fulfills this requirement as shown in App. A. There it is also shown that η has to be chosen resolution dependent, though.

2.4. Parameter Model

Fig. 2 gives an overview of the parameter hierarchy of the suggested Bayesian model. The data $\boldsymbol{d}$ is given, and the diffuse signal field $\boldsymbol{s}$ and the point-like signal field $\boldsymbol{u}$ shall be reconstructed from that data. The logarithmic power spectrum $\boldsymbol{\tau}$ is a set of nuisance parameters that also need to be reconstructed from the data in order to accurately model the diffuse flux contributions. The model parameters form the top layer of this hierarchy and are given to the reconstruction algorithm. This set of model parameters can be boiled down to five scalars, namely α, q, σ, β, and η, if one defines $\boldsymbol{\alpha} = \alpha\mathbf{1}$, etc. The incorporation of the scalars in

[7] The numerical discretization of information fields is described in great detail in Selig et al. (2013).

[8] The so-called maximum entropy regularization $\sum_x (\rho_x^{(u)}/\rho_0) \log(\rho_x^{(u)}/\rho_0)$ of the log-likelihood can be regarded as log-prior, cf. Eqs. (20) and (22).

[9] A possible extension of this prior model that includes spatial correlations would be an inverse-Wishart distribution for $\mathrm{diag}[\boldsymbol{\rho}^{(u)}]$.

the inference is possible in theory, but this would increase the computational complexity dramatically.

We discussed reasonable values for these scalars to be chosen *a priori*. If additional information sources, such as theoretical power spectra or object catalogs, are available the model parameters can be adjusted accordingly. In Sec. 4, different parameter choices for the analysis of simulated data are investigated.

3. Denoising, Deconvolution, and Decomposition

The likelihood model, describing the measurement process, and the prior assumptions for the signal fields and the power spectrum of the diffuse component yield a well-defined inference problem. The corresponding posterior is given by

$$P(\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{u}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{s},\boldsymbol{u})\; P(\boldsymbol{s}|\boldsymbol{\tau})\; P(\boldsymbol{\tau}|\alpha,q,\sigma)\; P(\boldsymbol{u}|\beta,\eta)}{P(\boldsymbol{d})}, \tag{27}$$

which is a complex form of Bayes' theorem (1).

Ideally, we would now calculate the *a posteriori* expectation values and uncertainties according to Eqs. (2) and (3) for the diffuse and point-like signal fields, $\boldsymbol{s}$ and $\boldsymbol{u}$, as well as for the logarithmic spectral parameters $\boldsymbol{\tau}$. However, an analytical computation of these expectation values is not possible due to the complexity of the posterior.

Numerical approaches involving Markov chain Monte Carlo methods (Metropolis & Ulam 1949; Metropolis et al. 1953) are possible, but hardly feasible due to the huge parameter phase space. Nevertheless, similar problems have been addressed by elaborate sampling techniques (Wandelt et al. 2004; Jasche et al. 2010).

Here, two approximative algorithms with lower computational costs are derived. The first one uses the maximum *a posteriori* (MAP) approximation, the second one minimizes the Gibbs free energy of an approximate posterior ansatz in the spirit of variational Bayesian methods. The fidelity and accuracy of these two algorithms are compared in a numerical application in Sec. 4.

3.1. Posterior Maximum

A naive approximation of the posterior expectation value is its maximum; i.e., approximating the mean by the mode of the distribution. This approximation holds for symmetric, single peaked distributions, but can perform poorly in other cases (e.g., Enßlin & Frommert 2011).

Instead of the complex posterior distribution, it is convenient to consider the information Hamiltonian, defined by its negative logarithm,

$$\begin{aligned} H(\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{u}|\boldsymbol{d}) &= -\log P(\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{u}|\boldsymbol{d}) && (28)\\ &= H_0 + \mathbf{1}^\dagger \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right) - \boldsymbol{d}^\dagger \log\left(\boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right)\right)\\ &\quad + \frac{1}{2}\log\left(\det\left[\boldsymbol{S}\right]\right) + \frac{1}{2}\boldsymbol{s}^\dagger \boldsymbol{S}^{-1}\boldsymbol{s} && (29)\\ &\quad + (\boldsymbol{\alpha}-\mathbf{1})^\dagger\boldsymbol{\tau} + \boldsymbol{q}^\dagger \mathrm{e}^{-\boldsymbol{\tau}} + \frac{1}{2}\boldsymbol{\tau}^\dagger \boldsymbol{T}\boldsymbol{\tau}\\ &\quad + (\boldsymbol{\beta}-\mathbf{1})^\dagger\boldsymbol{u} + \boldsymbol{\eta}^\dagger \mathrm{e}^{-\boldsymbol{u}}, \end{aligned}$$

where all terms constant in $\boldsymbol{s}$, $\boldsymbol{\tau}$, and $\boldsymbol{u}$ have been absorbed into a ground state energy H_0, cf. Eqs. (7), (11), (19), and (26), respectively.

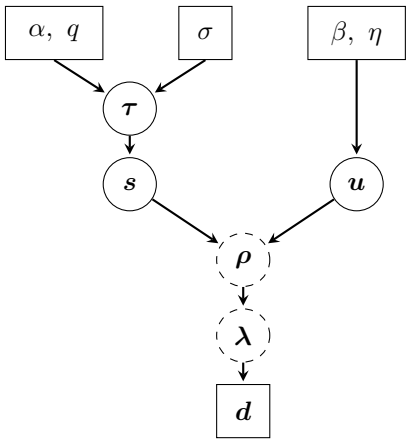

Fig. 2. Graphical model of the model parameters α, q, σ, β, and η, the logarithmic spectral parameters $\boldsymbol{\tau}$, the diffuse signal field $\boldsymbol{s}$, the point-like signal field $\boldsymbol{u}$, the total photon flux $\boldsymbol{\rho}$, the expected number of photons $\boldsymbol{\lambda}$, and the observed photon count data $\boldsymbol{d}$.

The MAP solution, which maximizes the posterior, minimizes the Hamiltonian. This minimum can thus be found by taking the first (functional) derivatives of the Hamiltonian with respect to $\boldsymbol{s}$, $\boldsymbol{\tau}$, and $\boldsymbol{u}$ and equating them with zero. Unfortunately, this yields a set of implicit, self-consistent equations rather than an explicit solution. However, these equations can be solved by an iterative minimization of the Hamiltonian using a steepest descent method for example, see Sec. 3.3 for details.

In order to better understand the structure of the MAP solution, let us consider the minimum $(\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{u}) = (\boldsymbol{m}^{(s)},\boldsymbol{\tau}^\star,\boldsymbol{m}^{(u)})$. The resulting filter formulas for the diffuse and point-like signal field read

$$\left.\frac{\partial H}{\partial \boldsymbol{s}}\right|_{\min} = \mathbf{0} = (\mathbf{1}-\boldsymbol{d}/\boldsymbol{l})^\dagger\, \boldsymbol{R} * \mathrm{e}^{\boldsymbol{m}^{(s)}} + \boldsymbol{S}^{\star\,-1}\boldsymbol{m}^{(s)}, \tag{30}$$

$$\left.\frac{\partial H}{\partial \boldsymbol{u}}\right|_{\min} = \mathbf{0} = (\mathbf{1}-\boldsymbol{d}/\boldsymbol{l})^\dagger\, \boldsymbol{R} * \mathrm{e}^{\boldsymbol{m}^{(u)}} + \boldsymbol{\beta} - \mathbf{1} - \boldsymbol{\eta} * \mathrm{e}^{-\boldsymbol{m}^{(u)}}, \tag{31}$$

with

$$\boldsymbol{l} = \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{m}^{(s)}} + \mathrm{e}^{\boldsymbol{m}^{(u)}}\right), \tag{32}$$

$$\boldsymbol{S}^\star = \sum_k \mathrm{e}^{\tau_k^\star}\boldsymbol{S}_k. \tag{33}$$

Here, $*$ and / denote componentwise multiplication and division, respectively. The first term in Eq. (30) and (31), which comes from the likelihood, vanishes in case $\boldsymbol{l} = \boldsymbol{d}$. Notice that $\boldsymbol{l}$ describes the most likely number of photon counts, not the expected number of photon counts $\boldsymbol{\lambda} = \langle \boldsymbol{d} \rangle_{(\boldsymbol{d}|\boldsymbol{s},\boldsymbol{u})}$, cf. Eqs. (5) and (7). For this reason, the MAP solution tends to overfitting; i.e., noise features are partly assigned to the signal fields in order to achieve a unnecessary closer agreement with the data.

The second derivative of the Hamiltonian describes the curvature around the minimum, and therefore approximates the (inverse) uncertainty covariance,

$$\left.\frac{\partial^2 H}{\partial \boldsymbol{s}\partial \boldsymbol{s}^\dagger}\right|_{\min} \approx \boldsymbol{D}^{(s)\,-1}, \qquad \left.\frac{\partial^2 H}{\partial \boldsymbol{u}\partial \boldsymbol{u}^\dagger}\right|_{\min} \approx \boldsymbol{D}^{(u)\,-1}. \tag{34}$$

The closed form of $\boldsymbol{D}^{(s)}$ and $\boldsymbol{D}^{(u)}$ is given explicitly in App. B.

The filter formula for the power spectrum, which is derived from a first derivate of the Hamiltonian with respect to $\boldsymbol{\tau}$, yields

$$\mathrm{e}^{\boldsymbol{\tau}^\star} = \frac{\boldsymbol{q} + \frac{1}{2}\left(\mathrm{tr}\left[\boldsymbol{m}^{(s)}\boldsymbol{m}^{(s)\dagger}\boldsymbol{S}_k^{-1}\right]\right)_k}{\boldsymbol{\gamma} + \boldsymbol{T}\boldsymbol{\tau}^\star}, \tag{35}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\alpha} - \mathbf{1}) + \frac{1}{2}\left(\mathrm{tr}\left[\boldsymbol{S}_k\boldsymbol{S}_k{}^{-1}\right]\right)_k$. This formula is in accordance with the results by Enßlin & Frommert (2011); Oppermann et al. (2012). It has been shown by the former authors that such a filter exhibits a perception threshold; i.e., on scales where the signal-response-to-noise ratio drops below a certain bound the reconstructed signal power becomes vanishingly low. This threshold can be cured by a better capture of the *a posteriori* uncertainty structure.

3.2. Posterior Approximation

In order to overcome the analytical infeasibility as well as the perception threshold, we seek an approximation to the true posterior. Instead of approximating the expectation values of the posterior, approximate posteriors are investigated in this section. In case the approximation is good, the expectation values of the approximate posterior should then be close to the real ones.

The posterior given by Eq. (27) is inaccessible due to the entanglement of the diffuse signal field $\boldsymbol{s}$, its logarithmic power spectrum $\boldsymbol{\tau}$, and the point-like signal field $\boldsymbol{u}$. The involvement of $\boldsymbol{\tau}$ can been simplified by a mean field approximation,

$$P(\boldsymbol{s}, \boldsymbol{\tau}, \boldsymbol{u}|\boldsymbol{d}) \approx Q = Q_s(\boldsymbol{s}, \boldsymbol{u}|\boldsymbol{\mu}, \boldsymbol{d})\; Q_\tau(\boldsymbol{\tau}|\boldsymbol{\mu}, \boldsymbol{d}), \tag{36}$$

where $\boldsymbol{\mu}$ denotes an abstract "mean field" mediating some information between the signal field tuple $(\boldsymbol{s}, \boldsymbol{u})$ and $\boldsymbol{\tau}$ that are separated by the product ansatz in Eq. (36). This mean field is usually only needed implicitly for the derivation, an explicit formula can be found in App. C.3, though.

Since the *a posteriori* mean estimates for the signal fields and their uncertainty covariances are of primary interest, a Gaussian approximation for Q_s that accounts for correlation between $\boldsymbol{s}$ and $\boldsymbol{u}$ would be sufficient. Hence, our previous approximation is extended by setting

$$Q_s(\boldsymbol{s}, \boldsymbol{u}|\boldsymbol{\mu}, \boldsymbol{d}) = \mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{D}), \tag{37}$$

with

$$\boldsymbol{\varphi} = \begin{pmatrix} \boldsymbol{s} - \boldsymbol{m}^{(s)} \\ \boldsymbol{u} - \boldsymbol{m}^{(u)} \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}^{(s)} & \boldsymbol{D}^{(su)} \\ \boldsymbol{D}^{(su)\dagger} & \boldsymbol{D}^{(u)} \end{pmatrix}. \tag{38}$$

This Gaussian approximation is also a convenient choice in terms of computational complexity due to its simple analytic structure.

The goodness of the approximation $P \approx Q$ can be quantified by an information theoretical measure, see App. C.1. The Gibbs free energy of the inference problem,

$$G = \left\langle H(\boldsymbol{s}, \boldsymbol{\tau}, \boldsymbol{u}|\boldsymbol{d})\right\rangle_Q - \left\langle -\log Q(\boldsymbol{s}, \boldsymbol{\tau}, \boldsymbol{u}|\boldsymbol{d})\right\rangle_Q, \tag{39}$$

which is equivalent to the Kullback-Leibler divergence $D_{\mathrm{KL}}(Q, P)$, is chosen as such a measure (Enßlin & Weig 2010).

In favor of comprehensibility, let us suppose the solution for the logarithmic power spectrum $\boldsymbol{\tau}^\star$ is known for the moment. The Gibbs free energy is then calculated by plugging in the Hamiltonian, and evaluating the expectation values[10],

$$\begin{aligned} G &= G_0 + \left\langle H(\boldsymbol{s}, \boldsymbol{u}|\boldsymbol{d})\right\rangle_{Q_s} - \frac{1}{2}\log\left(\det\left[\boldsymbol{D}\right]\right) && (40)\\ &= G_1 + \mathbf{1}^\dagger\boldsymbol{l} - \boldsymbol{d}^\dagger\left\{\log(\boldsymbol{l}) - \sum_{\nu=2}^{\infty}\frac{(-1)^\nu}{\nu}\left\langle\left(\boldsymbol{\lambda}/\boldsymbol{l} - 1\right)^\nu\right\rangle_{Q_s}\right\} \\ &\quad + \frac{1}{2}\boldsymbol{m}^{(s)\dagger}\boldsymbol{S}^{\star-1}\boldsymbol{m}^{(s)} + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{D}^{(s)}\boldsymbol{S}^{\star-1}\right] && (41)\\ &\quad + (\boldsymbol{\beta} - \mathbf{1})^\dagger\boldsymbol{m}^{(u)} + \boldsymbol{\eta}^\dagger\mathrm{e}^{-\boldsymbol{m}^{(u)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(u)}} \\ &\quad - \frac{1}{2}\log\left(\det\left[\boldsymbol{D}\right]\right), \end{aligned}$$

with

$$\boldsymbol{\lambda} = \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{s}} + \mathrm{e}^{\boldsymbol{u}}\right), \tag{42}$$

$$\boldsymbol{l} = \left\langle\boldsymbol{\lambda}\right\rangle_{Q_s} = \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{m}^{(s)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(s)}} + \mathrm{e}^{\boldsymbol{m}^{(u)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(u)}}\right), \tag{43}$$

$$\boldsymbol{S}^\star = \sum_k \mathrm{e}^{\tau_k^\star}\boldsymbol{S}_k, \text{ and} \tag{44}$$

$$\widehat{\boldsymbol{D}} = \mathrm{diag}\left[\boldsymbol{D}\right]. \tag{45}$$

Here, G_0 and G_1 carry all terms independent of $\boldsymbol{s}$ and $\boldsymbol{u}$. In comparison to the Hamiltonian given in Eq. (29), there are a number of correction terms that now also consider the uncertainty covariances of the signal estimates properly. For example, the expectation values of the photon fluxes differ comparing $\boldsymbol{l}$ in Eq. (32) and (43) where it now describes the expectation value of $\boldsymbol{\lambda}$ over the approximate posterior. In case $\boldsymbol{l} = \boldsymbol{\lambda}$ the explicit sum in Eq. (41) vanishes. Since this sum includes powers of $\left\langle\boldsymbol{\lambda}^{\nu>2}\right\rangle_{Q_s}$ its evaluation would require all entries of $\boldsymbol{D}$ to be known explicitely. In order to keep the algorithm computationally feasible, this sum shall furthermore be neglected. This is equivalent to truncating the corresponding expansion at second order; i.e., $\nu = 2$. It can be shown that, in consequence of this approximation, the cross-correlation $\boldsymbol{D}^{(su)}$ equals zero, and $\boldsymbol{D}$ becomes block diagonal.

Without these second order terms, the Gibbs free energy reads

$$\begin{aligned} G &= G_1 + \mathbf{1}^\dagger\boldsymbol{l} - \boldsymbol{d}^\dagger\log(\boldsymbol{l}) \\ &\quad + \frac{1}{2}\boldsymbol{m}^{(s)\dagger}\boldsymbol{S}^{\star-1}\boldsymbol{m}^{(s)} + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{D}^{(s)}\boldsymbol{S}^{\star-1}\right] && (46)\\ &\quad + (\boldsymbol{\beta} - \mathbf{1})^\dagger\boldsymbol{m}^{(u)} + \boldsymbol{\eta}^\dagger\mathrm{e}^{-\boldsymbol{m}^{(u)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(u)}} \\ &\quad - \frac{1}{2}\log\left(\det\left[\boldsymbol{D}^{(s)}\right]\right) - \frac{1}{2}\log\left(\det\left[\boldsymbol{D}^{(u)}\right]\right). \end{aligned}$$

[10] The second likelihood term in Eq. (41), $\boldsymbol{d}^\dagger\log(\boldsymbol{\lambda})$, is thereby expanded according to

$$\begin{aligned} \log(x) &= \log\langle x\rangle - \sum_{\nu=2}^{\infty}\frac{(-1)^\nu}{\nu}\left\langle\left(\frac{x}{\langle x\rangle} - 1\right)^\nu\right\rangle \\ &\approx \log\langle x\rangle + \mathcal{O}\left(\left\langle x^2\right\rangle\right), \end{aligned}$$

under the assumption $x \approx \langle x\rangle$.

Minimizing the Gibbs free energy with respect to $\boldsymbol{m}^{(s)}$, $\boldsymbol{m}^{(u)}$, $\boldsymbol{D}^{(s)}$, and $\boldsymbol{D}^{(u)}$ would optimize the fitness of the posterior approximation $P \approx Q$, and enable us to compute expectation values of the diffuse and point-like photon flux straight forwardly,

$$\left\langle \boldsymbol{\rho}^{(s)} \right\rangle_P \approx \left\langle \boldsymbol{\rho}^{(s)} \right\rangle_Q = \rho_0 \mathrm{e}^{\boldsymbol{m}^{(s)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(s)}}, \tag{47}$$

$$\left\langle \boldsymbol{\rho}^{(s)} \right\rangle_P \approx \left\langle \boldsymbol{\rho}^{(u)} \right\rangle_Q = \rho_0 \mathrm{e}^{\boldsymbol{m}^{(u)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(u)}}. \tag{48}$$

Filter formulas for the Gibbs solution can be derived by taking the derivative of G with respect to the approximate mean estimates,

$$\frac{\partial G}{\partial \boldsymbol{m}^{(s)}} = \boldsymbol{0} = (\boldsymbol{1} - \boldsymbol{d}/\boldsymbol{l})^\dagger \, \boldsymbol{R} * \mathrm{e}^{\boldsymbol{m}^{(s)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(s)}} + \boldsymbol{S}^{\star -1} \boldsymbol{m}^{(s)}, \tag{49}$$

$$\frac{\partial G}{\partial \boldsymbol{m}^{(u)}} = \boldsymbol{0} = (\boldsymbol{1} - \boldsymbol{d}/\boldsymbol{l})^\dagger \, \boldsymbol{R} * \mathrm{e}^{\boldsymbol{m}^{(u)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(u)}} + \boldsymbol{\beta} - \boldsymbol{1} - \boldsymbol{\eta} * \mathrm{e}^{-\boldsymbol{m}^{(u)} + \frac{1}{2}\widehat{\boldsymbol{D}}^{(u)}}, \tag{50}$$

This filter formulas again account for the uncertainty of the mean estimates in comparison to the MAP filter formulas in Eq. (30) and (31). The uncertainty covariances can be constructed by either taking the second derivatives,

$$\frac{\partial^2 G}{\partial \boldsymbol{m}^{(s)} \partial \boldsymbol{m}^{(s)\dagger}} \approx \boldsymbol{D}^{(s)-1}, \quad \frac{\partial^2 G}{\partial \boldsymbol{m}^{(u)} \partial \boldsymbol{m}^{(u)\dagger}} \approx \boldsymbol{D}^{(u)-1}, \tag{51}$$

or setting the first derivatives of G with respect to the uncertainty covariances equal to zero matrices,

$$\frac{\partial G}{\partial D^{(s)}_{xy}} = 0, \qquad \frac{\partial G}{\partial D^{(u)}_{xy}} = 0. \tag{52}$$

The closed form of $\boldsymbol{D}^{(s)}$ and $\boldsymbol{D}^{(u)}$ is given explicitly in App. B.

So far, the logarithmic power spectrum $\boldsymbol{\tau}^\star$, and with it $\boldsymbol{S}^\star$, have been supposed to be known. The mean field approximation in Eq. (36) does not specify the approximate posterior $Q_\tau(\boldsymbol{\tau}|\boldsymbol{\mu}, \boldsymbol{d})$, but it can be retrieved by variational Bayesian methods (Jordan et al. 1999; Wingate & Weber 2013), according to the procedure detailed in App. C.2. The subsequent App. C.3 discusses the derivation of an solution for $\boldsymbol{\tau}$ by extremizing Q_τ. This result, which was also derived in Oppermann et al. (2012), applies to the inference problem discussed here, yielding

$$\mathrm{e}^{\boldsymbol{\tau}^\star} = \frac{\boldsymbol{q} + \frac{1}{2}\left(\mathrm{tr}\left[\left(\boldsymbol{m}^{(s)}\boldsymbol{m}^{(s)\dagger} + \boldsymbol{D}^{(s)}\right)\boldsymbol{S}_k^{-1}\right]\right)_k}{\boldsymbol{\gamma} + \boldsymbol{T}\boldsymbol{\tau}^\star}. \tag{53}$$

Again, this solution includes a correction term in comparison to the MAP solution in Eq. (35). Since $\boldsymbol{D}^{(s)}$ is positive definite, it contributes positive to the (logarithmic) power spectrum, and therefore reduces the possible perception threshold further.

Notice that this is a minimal Gibbs free energy solution that maximizes Q_τ. A proper calculation of $\langle \boldsymbol{\tau} \rangle_{Q_\tau}$ might include further correction terms, but their derivation is not possible in closed form. Moreover, the above used diffuse signal covariance $\boldsymbol{S}^{\star -1}$ should be replaced by $\left\langle \boldsymbol{S}^{-1} \right\rangle_{Q_\tau}$ adding further correction terms to the filter formulas.

In order to keep the computational complexity on a feasible level, all these higher order corrections are not considered here. The detailed characterization of their implications and implementation difficulties is left for future investigation.

3.3. Imaging Algorithm

The problem of denoising, deconvolving, and decomposing photon observations is a non-trivial task. Therefore, this section discusses the implementation of the D^{3}PO algorithm given the two sets of filter formulas derived in Sec. 3.1 and 3.2, respectively.

The information Hamiltonian, or equivalently the Gibbs free energy, are scalar quantities defined over a huge phase space of possible field and parameter configurations including, among others, the elements of $\boldsymbol{m}^{(s)}$ and $\boldsymbol{m}^{(u)}$. If we only consider those, and no resolution refinement from data to signal space, two numbers need to be inferred from one data value. Including $\boldsymbol{\tau}$ and the uncertainty covariances $\boldsymbol{D}^{(s)}$ and $\boldsymbol{D}^{(u)}$ in the inference, the problem of underdetermined degrees of freedom gets worse. This is reflected in the possibility of a decent number of local minima in the manifold landscape of the codomain of the Hamiltonian, or Gibbs free energy, respectively. The complexity of the inference problem goes back to the, in general, non-linear entanglement between the individual parameters.

The D^{3}PO algorithm is based on an iterative optimization scheme, where an optimization is repeated on certain subsets of the problem instead of the full problem at once. Each subset optimization is then designed individually, see below. So far, a step-by-step guide of the algorithm looks like the following.

1. Initialize the algorithm with starting values; e.g., $m^{(s)}_x = m^{(u)}_x = 0$, $D^{(s)}_{xy} = D^{(u)}_{xy} = \delta_{xy}$, and $\tau^\star_k = k^{-2}$.
   Those values are arbitrary. In principle, the optimization is not sensible to the starting values, but rather inappropriate values can cripple the algorithm for numerical reasons. This behavior goes back to the high non-linearity of the inference problem.
2. Optimize the initial point-like signal field $\boldsymbol{m}^{(u)}$ preliminarily.
   The brightest, most obvious, point-like sources, which are visible in the data image by eye, dominate the disagreement between data and current guess. This causes high values for the considered potential; i.e., the information Hamiltonian or Gibbs free energy, respectively. The gradient of the potential can be computed according to Eq. (31) or (50). Its minima will be at the expected position of the brightest point source which has not been reconstructed, yet. It is therefore very efficient to increase $\boldsymbol{m}^{(u)}$ at this location directly until the sign of the gradient flips, and repeat this until the obvious point-like sources are fit.
3. Optimize the current point-like signal field $\boldsymbol{m}^{(u)}$.
   This task can be done by a steepest descent minimization of the potential combined with a line search following the Wolfe conditions (Nocedal & Wright 2006). The potentials can be computed according to Eq. (29) or (41) neglecting terms independent of $\boldsymbol{m}^{(u)}$, and the gradient according to Eq. (31) or (50). A more sophisticated minimization scheme, such as a non-linear conjugate gradient (Shewchuk 1994), is conceivable but would require

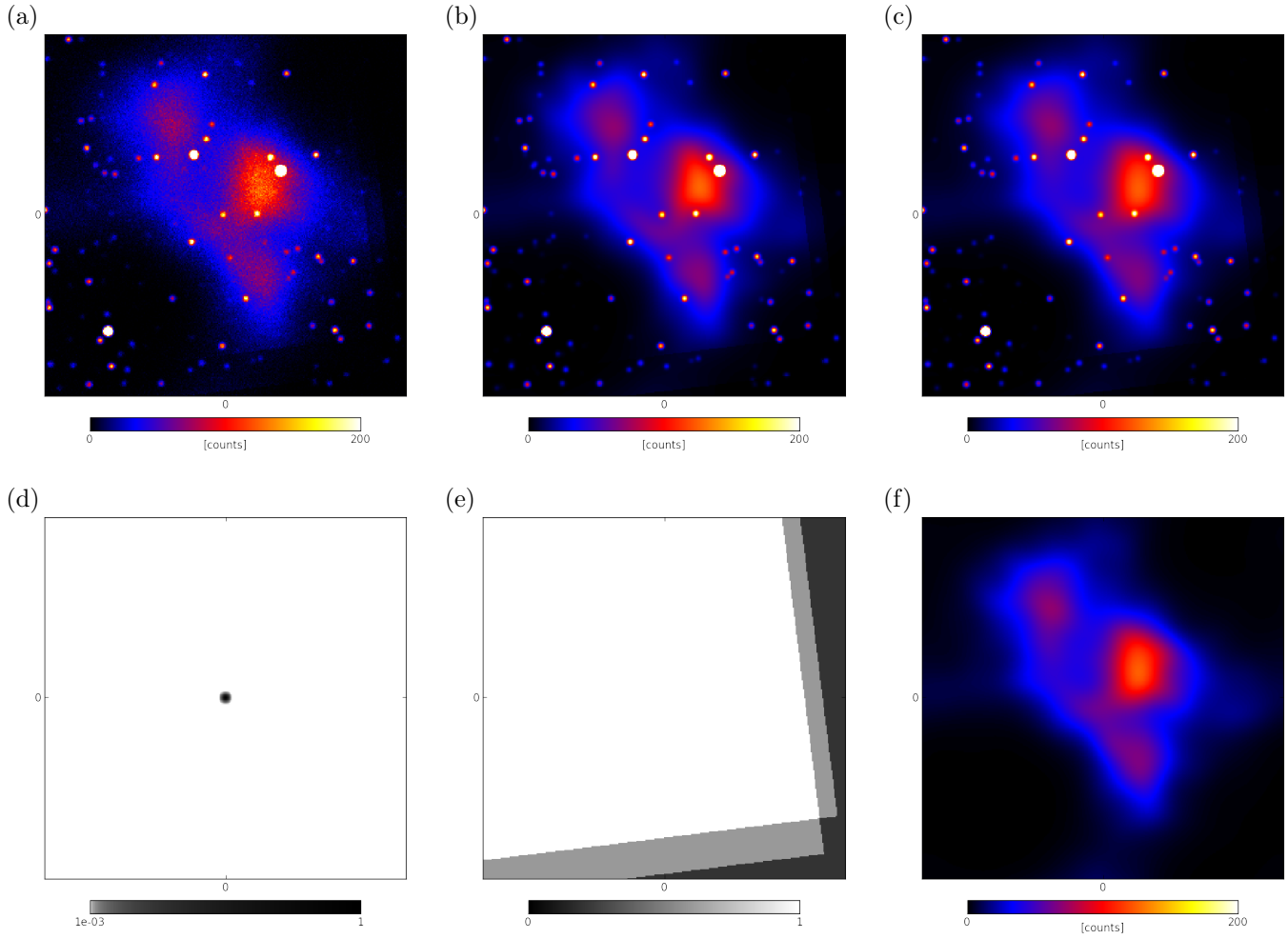

Fig. 3. Illustration of the data and noiseless reproductions from reconstructed signals. Panel (a) shows the data from a mock observation of a $32 \times 32\,\mathrm{arcmin}^2$ patch of the sky with a resolution of 0.1 arcmin corresponding to a total of 102 400 pixels. The Data had been convolved with a Gaussian-like PSF (FWHM ≈ 0.2 arcmin $= 2$ pixels, finite support of 1.1 arcmin $= 11$ pixels) and masked due to an uneven exposure. Panel (b) shows the reproduced signal response of a reconstruction using a MAP approach. Panel (c) shows the reproduction of a reconstruction using a Gibbs approach. Panel (d) shows the centered convolution kernel. Panel (e) shows the exposure mask. Panel (f) shows the unmasked diffuse contribution of the above panel (c).

the application of the full Hessian, cf. step 4. In the first run, it might be sufficient to restrict the optimization to the locations identified in step 2.

4. Update the current point-like uncertainty variance $\widehat{\boldsymbol{D}}^{(u)}$ in case of a Gibbs approach.
   It is not feasible to compute the full uncertainty covariance $\boldsymbol{D}^{(u)}$ explicitly in order to extract its diagonal. A more elegant way is to apply a probing technique relying on the application of $\boldsymbol{D}^{(u)}$ to random fields that project out the diagonal (Hutchinson 1989; Selig et al. 2012). The uncertainty covariance is given as the inverse Hessian by Eq. (34) or (51), and should be symmetric and positive definite. For that reason, it can be applied to a field using a conjugate gradient (Shewchuk 1994). However, if the current phase space position is far away from the minimum, the Hessian is not necessarily positive definite. One way to overcome this temporal instability, would be to introduce a Levenberg damping in the Hessian (inspired by Transtrum et al. 2010; Transtrum & Sethna 2012).
5. Optimize the current diffuse signal field guess $\boldsymbol{m}^{(s)}$.
   An analog scheme as in step 3 using steepest descent and Wolfe conditions is appropriate. The potentials can be computed according to Eq. (29) or (41) neglecting terms independent of $\boldsymbol{m}^{(s)}$, and the gradient according to Eq. (30) or (49), respectively. It has proven useful to first ensure a convergence on large scales; i.e., small harmonic modes k. This can be done repeating steps 5, 6, and 7 for all $k < k_{\mathrm{max}}$ with growing k_{max} using the corresponding projections $\boldsymbol{S}_k$.
6. Update the current diffuse uncertainty $\widehat{\boldsymbol{D}}^{(s)}$ in case of a Gibbs approach in analogy to step 4.
7. Optimize the current logarithmic power spectrum $\boldsymbol{\tau}^{\star}$.
   This is done by solving Eq. (35) or (53). The trace term can be computed analog to the diagonal; e.g., by probing. Given this, the equation can be solved efficiently by a Newton-Raphson method.
8. Repeat the steps 3 till 7 until convergence.
   This scheme will take several iterations until the algorithm reaches the desired convergence level. Therefore, it is not required to achieve a convergence to the final accuracy level in all subsets in all iterations. It is advis-

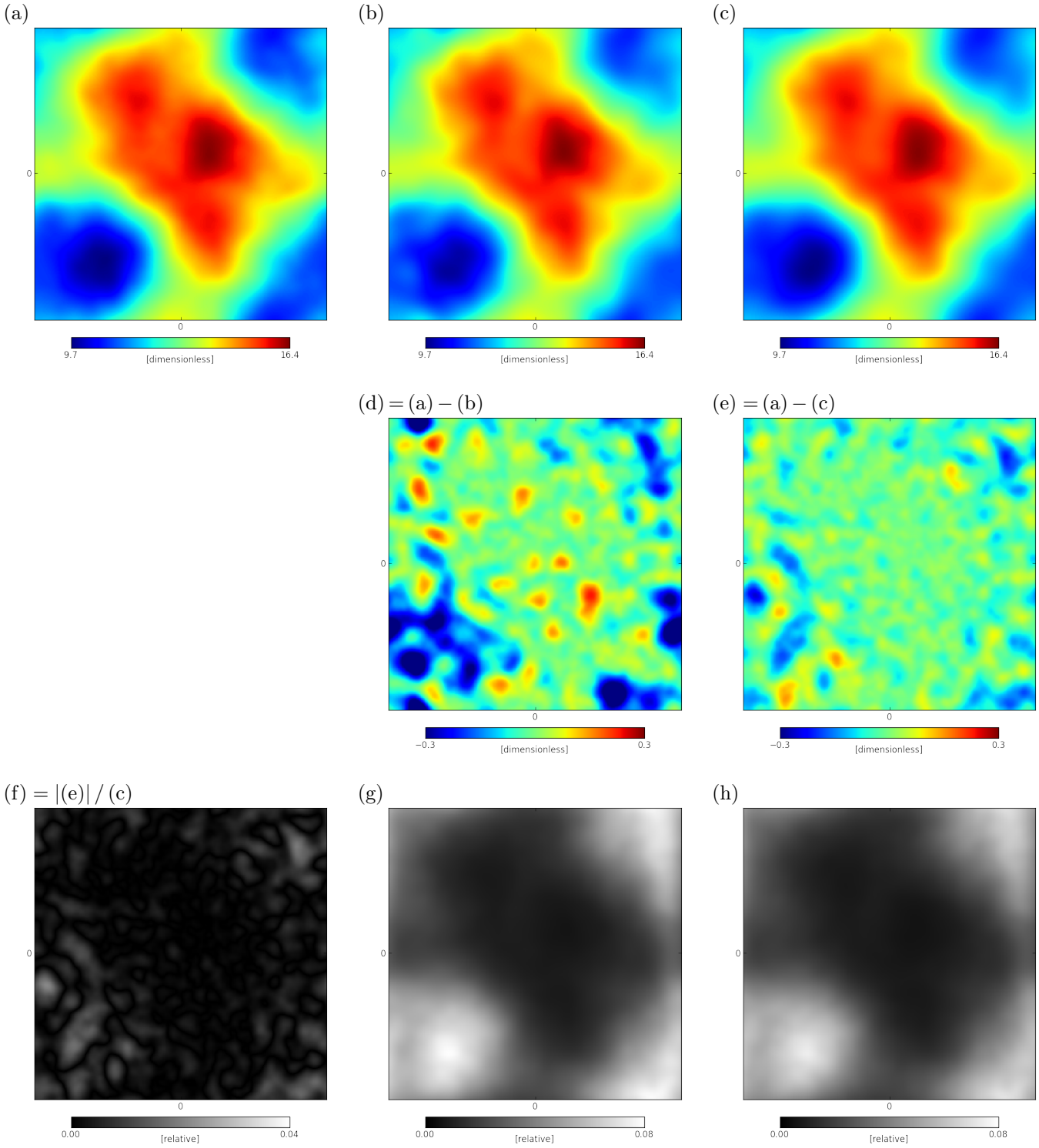

Fig. 4. Illustration of the reconstruction of the diffuse signal field $\boldsymbol{s}$ and its uncertainty. The top panels show diffuse signal fields. Panel (a) shows the original simulation, panel (b) the reconstruction using a MAP approach, and panel (c) the reconstruction using a Gibbs approach. The panels (d) and (e) show the differences between original and reconstruction. Panel (f) shows the relative difference. The panels (g) and (h) show the relative uncertainty of the above reconstructions.

able to start with weak convergence criteria in the first loop and increase them gradually.

A few remarks are in order.

The phase space of possible signal field configurations is tremendously huge. It is therefore impossible to judge if the algorithm has converged to the global or some local minima, but this does not matter if both yield reasonable results that do not differ substantially.

The algorithm starts with the reconstruction of the point-like signal field. This raises the chance of explaining diffuse features by point sources. Starting with the diffuse components instead would in turn give raise to the opposite

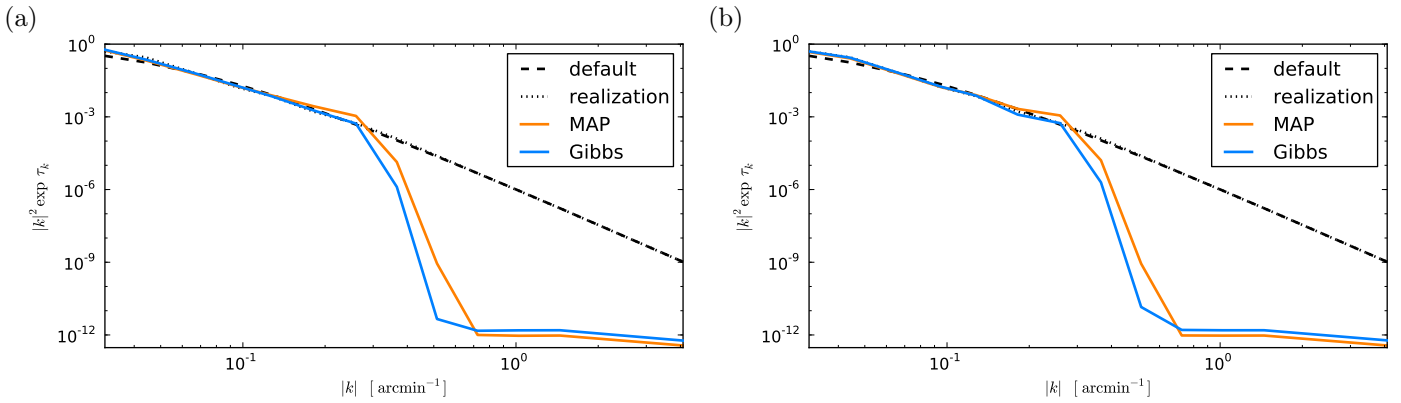

Fig. 5. Illustration of the reconstruction of the logarithmic power spectrum $\boldsymbol{\tau}$. Both panels show the default power spectrum (black dashed line), and the simulated realization (black dotted line). Panel (a) shows the reconstructed power spectra using a MAP (orange solid line) and Gibbs approach (blue solid line) for a chosen σ parameter of 10. Panel (b) shows the reconstructions for a σ of 1000.

bias. To avoid such biases, it is advisable to restart the algorithm partially. To be more precise, we propose to discard the current reconstruction of $\boldsymbol{m}^{(u)}$ after finishing step 7 for the first time, then start the second iteration again with step 2, and to discard the current $\boldsymbol{m}^{(s)}$ before step 5.

The above scheme exploits a few numerical techniques, such as probing or Levenberg damping, that are described in great detail in the given references. The code of our implementation of the D^{3}PO algorithm will be made public in the future.

4. Numerical Application

Exceeding the simple 1D scenario illustrated in Fig. 1, the D^{3}PO algorithm is now applied to a realistic, but simulated, data set. The data set represents a high energy observation with a field of view of $32 \times 32\,\mathrm{arcmin}^2$ and a resolution of 0.1 arcmin; i.e., the photon count image comprises 102 400 pixels. The instrument response includes the convolution with a Gaussian-like PSF with a FWHM of roughly 0.2 arcmin, and an uneven survey mask due to the inhomogeneous exposure of the virtual instrument. The data image and those characteristics are shown in Fig. 3.

In addition, Fig. 3 shows two reproduced signal responses of reconstructed signal fields. The reconstructions used a MAP and a Gibbs approach, respectively. Both show a very good agreement with the actual data. Notice that only the quality of denoising is visible, since the signal response shows the convolved and superimposed signal fields.

The last panel of Fig. 3 shows solely the diffuse contribution to the deconvolved photon flux as reconstructed using the Gibbs approach. There, all point-like contributions as well as noise and instrumental effects have been removed presenting a denoised, deconvolved and decomposed reconstruction result.

Fig. 4 illustrates the diffuse signal field reconstructions. The original and the reconstructions agree well, and the strongest deviations are found in the areas with low amplitudes. With regard to the exponential ansatz in Eq. (4), it is not surprising that the inference on the signal fields is more sensible to higher values than to lower ones. For example, a small change in the diffuse signal field, $\boldsymbol{s} \to (1 \pm \epsilon)\boldsymbol{s}$, translates into a factor in the photon flux, $\boldsymbol{\rho}^{(s)} \to \boldsymbol{\rho}^{(s)} \mathrm{e}^{\pm \epsilon \boldsymbol{s}}$, that scales exponentially with the amplitude of the diffuse signal field.

The Gibbs solution shows less deviations from the original signal than the MAP solution due to the overfitting by the latter. This includes overestimates in noisy regions with low flux intensities, as well as underestimates at locations where point-like contributions dominate the total flux. The latter indicates an overfitting of the point-like component by the MAP approach.

The diffuse photon fluxes can be computed directly from the reconstruction,

$$\left\langle \rho_x^{(s)} \right\rangle_{(\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{u}|\boldsymbol{d})} \overset{\mathrm{MAP}}{\approx} \rho_0 \mathrm{e}^{m_x^{(s)}} \tag{54}$$

$$\overset{\mathrm{Gibbs}}{\approx} \rho_0 \mathrm{e}^{m_x^{(s)} + \frac{1}{2} D_{xx}^{(s)}}, \tag{55}$$

The uncertainty of the reconstruction can be calculated as for any log-normal distribution,

$$\left\langle {\rho_x^{(s)}}^2 \right\rangle_{(\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{u}|\boldsymbol{d})} - \left\langle \rho_x^{(s)} \right\rangle_{\ldots}^2 \approx \left\langle \rho_x^{(s)} \right\rangle_{\ldots}^2 \left(\mathrm{e}^{D_{xx}^{(s)}} - 1 \right), \tag{56}$$

where the square root of the latter term would describe the relative uncertainty, which is illustrated in the lower panels of Fig. 4. In accordance with the above argument, the uncertainty is large in areas with low amplitudes. Further, the uncertainty is also slightly larger in areas with less observational exposure, cf. with the exposure mask shown in Fig. 3.

The reconstruction of the power spectrum, as shown in Fig. 5, gives further indications of the reconstruction quality of the diffuse component. The simulation used a default power spectrum of

$$\mathrm{e}^{\tau_k} = 42\,(k+1)^{-7}. \tag{57}$$

This power spectrum was on purpose chosen to deviate from a strict power-law supposed by the smoothness prior.

From Fig. 5 it is apparent that the reconstructed power spectra track the original well up to a harmonic mode k of roughly $0.4\,\mathrm{arcmin}^{-1}$. Beyond that point, the reconstructed power spectra fall steeply until they hit a lower boundary set by the model parameter q, which was here set to 10^{-12}. This drop-off point at $0.4\,\mathrm{arcmin}^{-1}$ corresponds to a physical wavelength of roughly 2.5 arcmin, and thus (half-phase) fluctuations on a spatial distances below 1.25 arcmin. The Gaussian-like PSF of the virtual observatory has a finite

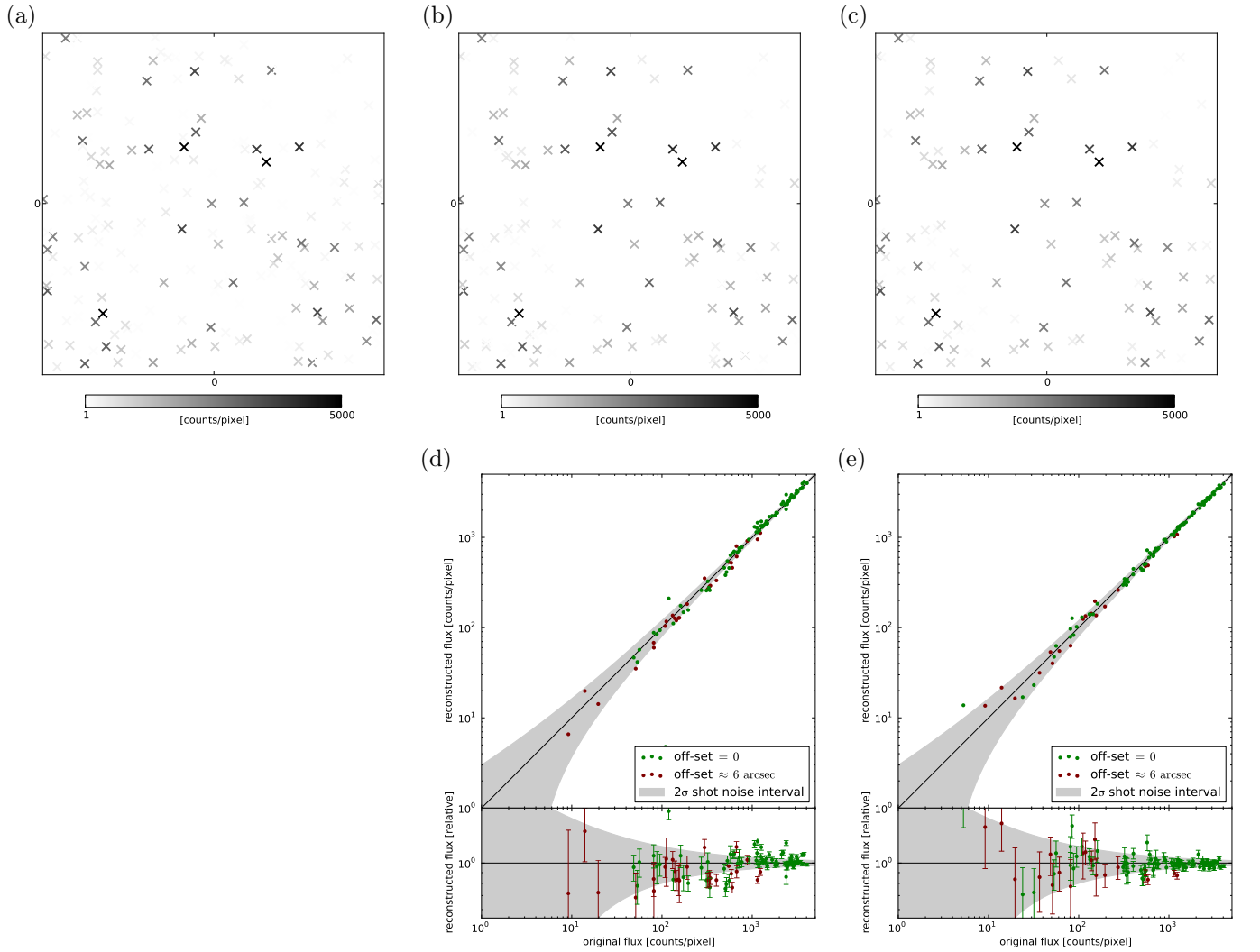

Fig. 6. Illustration of the reconstruction of the point-like signal field $\boldsymbol{u}$ and its uncertainty. The top panels show the location (markers) and intensity (gray scale) of the point-like photon fluxes. Panel (a) shows the original simulation, panel (b) the reconstruction using a MAP approach, and panel (c) the reconstruction using a Gibbs approach. The bottom panels (d) and (e) show the match between original and reconstruction in absolute and relative fluxes, the 2σ shot noise interval (gray contour), as well as some reconstruction uncertainty estimate (error bars).

support of 1.1 arcmin. The lack of reconstructed power indicates that the algorithm assigns features on spatial scales smaller than the PSF support preferably to the point-like component. This is reasonable since the point-like signal can solely cause PSF-like shaped imprints in the data image. However, this does not mean that there is a strict distinction between the components due to their spatial extend, because the consideration of noise effects blurs out such boundaries, of course.

The differences between the reconstruction using a MAP and a Gibbs approach are subtle. The drop-off point is apparently at higher k for the former. The difference in the reconstruction formulas given by Eqs. (35) and (53) is an additive trace term involving $\boldsymbol{D}^{(s)}$, which is positive definite. Therefore, a reconstructed power spectrum using the Gibbs approach is never below the MAP solution given the same $\boldsymbol{m}^{(s)}$. However, the reconstruction of the signal field follows different filter formulas. Since the Gibbs approach considers the uncertainty covariance $\boldsymbol{D}^{(s)}$ properly, it can present a more conservative solution, cf. Eq. (55). In turn, the MAP solution tends to overfit by absorbing some noise power into $\boldsymbol{m}^{(s)}$ as discussed in Sec. 3. Thus, the higher MAP power spectrum in Fig. 5 seems to be caused by a higher level of noise remnants in the signal estimate.

The influence of the choice of the model parameter σ is also shown in Fig. 5. Neither a smoothness prior with $\sigma = 10$, nor a weak one with $\sigma = 1000$ influences the reconstruction of the power spectrum substantially in this case.[11] The latter choice, however, exhibits some more fluctuations in order to better track the concrete realization.

The results for the reconstruction of the point-like component are illustrated in Fig. 6. Overall, the reconstructed point-like signal field and the corresponding photon flux are in good agreement with the original ones. The point-sources have been located with an accuracy of ± 0.1 arcmin, which is less than the FWHM of the PSF. The localization tends to be more precise for higher flux values because of the lower signal-to-noise ratio. The reconstructed intensities

[11] For a discussion of further log-normal reconstruction scenarios please refer to the work by Oppermann et al. (2012).

match the simulated ones well, although the MAP solution shows a spread that exceeds the expected shot noise uncertainty interval. This is again an indication of the overfitting known for the MAP solution. Moreover, neither reconstruction shows a bias towards higher or lower fluxes.

The uncertainty estimates for the point-like photon flux $\boldsymbol{\rho}^{(u)}$ obtained from $\boldsymbol{D}^{(u)}$ in analogy to Eq. (56) are, in general, consistent with the deviations from the original and the shot noise uncertainty, cf. Fig. 6. They show a reasonable scaling being higher for lower fluxes and vice versa. However, some uncertainties seem to be underestimated. There are different reasons for this.

On the one hand, the Hessian approximation for $\boldsymbol{D}^{(u)}$ in Eq. (34) or (51) is in so far poor as that the curvature of the considered potential does not always describe the uncertainty of the point-like component adequately. The data constrains the flux intensity of a point source strongly, especially if the point source is a bright one. Furthermore, a change of the corresponding signal field value will have an asymmetric effect on the flux intensity. Both effects result in a rather narrow, asymmetric dip in the manifold landscape of the considered potential that is not well described by a quadratic approximation. On the other hand, the approximation leading to vanishing cross-correlation $\boldsymbol{D}^{(su)}$, takes away the possibility of communicating uncertainties between diffuse and point-like components. However, omitting the used simplification or incorporating higher order corrections would render the algorithm too computationally expensive. The fact that the Gibbs solution, which takes $\boldsymbol{D}^{(u)}$ into account, shows improvements backs up this argument.

The original point-like photon flux has been drawn according to the prior $P(\boldsymbol{u}|\beta = \frac{3}{2}, \eta = 1)$, cf. Eq. (26). Several reconstructions with $\beta \in \{1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4}, 2\}$ and $\eta \in \{10^{-6}, 10^{-4}, 10^{-2}, 1\}$ have been carried out, without leading to significantely different results. The $\beta = 1$ case that corresponds to an logarithmically flat prior on $\boldsymbol{u}$ showed a tendency to fit more noise features by point-like contributions. The reconstructions shown in Fig. 6 used $\beta = \frac{3}{2}$ and $\eta = 10^{-4}$. Although the choice of the model parameters β and η has some influence on the reconstruction results, this is very modest.

In summary, the D^{3}PO algorithm is capable of denoising, deconvolving and decomposing photon observations by reconstructing the diffuse and point-like signal field, as well as the logarithmic power spectrum of the former. The reconstruction using a MAP and Gibbs approach perform flawlessly, except for the estimate of the uncertainty of the point-like component. The MAP approach shows signs of overfitting, but those are not overwhelming. Considering the simplicity of the MAP approach that goes along with a numerically faster performance, this shortcome seems acceptable.

Due to the iterative scheme of the algorithm, a combination of the MAP approach for the signal fields and a Gibbs approach for the power spectrum is possible.

5. Conclusions & Summary

The D^{3}PO algorithm for the denoising, deconvolving and decomposing photon observations has been derived. It allows for the simultaneous reconstruction of the diffuse and point-like photon fluxes, as well as the diffuse power spectrum, from data images that are exposed to Poissonian shot noise and effects of the instrument response functions.

The theoretical foundation is a hierarchical Bayesian parameter model embedded in the framework of IFT. The model comprises *a priori* assumptions for the signal fields that account for the different statistics and correlations of the morphologically different components. The diffuse photon flux is assumed to obey multi-variant log-normal statistics, where the covariance is described by a power spectrum. The power spectrum is *a priori* unknown and reconstructed from the data along with the signal. Therefore hyperpriors on the (logarithmic) power spectra have been introduced, including a spectral smoothness prior (Enßlin & Frommert 2011; Oppermann et al. 2012). The point-like photon flux, in contrast, is assumed to factorize spatially in independent inverse-Gamma distributions implying a (regularized) power-law behavior of the amplitudes of the flux.

An adequate description of the noise properties in terms of a likelihood, here a Poisson distribution, and the incorporation of all instrumental effects into the response operator renders the denoising and deconvolution task possible. The strength of the proposed model is the exploitation of those priors, especially for the task of component separation. However, the model is not parameter free. The model comes down to five scalar parameters, for which all *a priori* defaults can be motivated, and of which none is driving the inference predominantely.

The performace of the D^{3}PO algorithm has been demonstrated in realistic simulations carried out in 1D and 2D. The implementation relies on the NIFTy package (Selig et al. 2013), which allows for the application regardless of the underlying position space.

In the 2D application example, a high energy observation of a $32 \times 32\,\mathrm{arcmin}^2$ patch of the sky has been analyzed. The D^{3}PO algorithm successfully denoised, deconvolved and decomposed the data image. The analysis yielded a detailed reconstruction of the diffuse photon flux and its logarithmic power spectrum, as well as the precise localization of the point sources and accurate determination of their flux intensities.

The D^{3}PO algorithm should be applicable to a wide range of inference problems appearing in astronomical imaging and related fields. Concrete applications in high energy astrophysics, for example, the analysis of data from the Chandra X-ray observatory or the Fermi γ-ray space telescope, are currently considered by the authors. In this regard, the public release of the D^{3}PO code is planned.

Acknowledgments

We thank Niels Oppermann and Henrik Junklewitz for the insightful discussions and productive comments.

Furthermore, we thank the DFG Forschergruppe 1254 "Magnetisation of Interstellar and Intergalactic Media: The Prospects of Low-Frequency Radio Observations" for travel support in order to present this work at their annual meeting in 2013.

Some of the results in this publication have been derived using the NIFTy package (Selig et al. 2013). This research has made use of NASA's Astrophysics Data System.

References

Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393
Carvalho, P., Rocha, G., & Hobson, M. P. 2009, MNRAS, 393, 681
Carvalho, P., Rocha, G., Hobson, M. P., & Lasenby, A. 2012, MNRAS, 427, 1384
Caticha, A. 2008, ArXiv e-prints physics.data-an/0808.0012
Caticha, A. 2011, in American Institute of Physics Conference Series, Vol. 1305, American Institute of Physics Conference Series, ed. A. Mohammad-Djafari, J.-F. Bercher, & P. Bessiére, 20–29
Enßlin, T. A. 2012, ArXiv e-prints
Enßlin, T. A. & Frommert, M. 2011, Phys. Rev. D, 83, 105014
Enßlin, T. A., Frommert, M., & Kitaura, F. S. 2009, Phys. Rev. D, 80, 105005
Enßlin, T. A. & Weig, C. 2010, Phys. Rev. E, 82, 051112
Enßlin, T. A. & Weig, C. 2012, Phys. Rev. E, 85, 033102
Fomalont, E. B. 1968, Bull. Astron. Inst. Netherlands, 20, 69
Giron, Francisco Javier, C. C. d. 2001, RACSAM, 95, 39
González-Nuevo, J., Argüeso, F.and Lopez-Caniego, M., Toffolatti, L., et al. 2006, Notices of the Royal Astronomical Society, 1603
Guglielmetti, F., Fischer, R., & Dose, V. 2009, MNRAS, 396, 165
Haar, A. 1910, Mathematische Annalen, 69, 331
Haar, A. 1911, Mathematische Annalen, 71, 38
Hensley, B. S., Pavlidou, V., & Siegal-Gaskins, J. M. 2013, MNRAS, 433, 591
Högbom, J. A. 1974, A&AS, 15, 417
Hutchinson, M. F. 1989, Communications in Statistics - Simulation and Computation, 18, 1059
Iatsenko, D., Stefanovska, A., & McClintock, P. V. E. 2012, Phys. Rev. E, 85, 033101
Jasche, J., Kitaura, F. S., Wandelt, B. D., & Enßlin, T. A. 2010, MNRAS, 406, 60
Jaynes, E. T. 1957, Physical Reviews, 106 and 108, 620
Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. 1999, Machine Learning, 37, 183
Kullback, S. & Leibler, R. A. 1951, The Annals of Mathematical Statistics, 22, 79
Malyshev, D. & Hogg, D. W. 2011, ApJ, 738, 181
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087
Metropolis, N. & Ulam, S. 1949, J. Am. Stat. Assoc., 44, 335
Nocedal, J. & Wright, S. J. 2006, Numerical optimization
Oppermann, N., Selig, M., Bell, M. R., & Enßlin, T. A. 2012
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2011, A&A, 536, A7
Selig, M., Bell, M. R., J. H., Oppermann, N., et al. 2013, A&A, 554, A26
Selig, M., Oppermann, N., & Enßlin, T. A. 2012, Phys. Rev. E, 85, 021134
Shewchuk, J. R. 1994, Technical report, Carnegie Mellon University, Pittsburgh, PA
Strong, A. W. 2003, A&A, 411, L127
Transtrum, M. K., Machta, B. B., & Sethna, J. P. 2010, Physical Review Letters, 104, 060201
Transtrum, M. K. & Sethna, J. P. 2012, ArXiv e-prints
Valdes, F. 1982, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 331, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 465–472
Wandelt, B. D., Larson, D. L., & Lakshminarayanan, A. 2004, Phys. Rev. D, 70, 083511
Wingate, D. & Weber, T. 2013, ArXiv e-prints

Appendix A: Point Source Stacking

In Sec. 2.3.3, a prior for the point-like signal field has been derived under the assumption that the photon flux of point sources is independent between different pixels and identically inverse-Gamma distributed,

$$\rho_x^{(u)} \curvearrowleft \mathcal{I}\left(\rho_x^{(u)}, \beta = \frac{3}{2}, \rho_0\eta\right) \quad \forall x, \tag{A.1}$$

with the shape and scale parameters, β and η. It can be shown that, for $\beta = \frac{3}{2}$, the sum of N such variables still obeys an inverse-Gamma distribution,

$$\rho_N^{(u)} = \sum_x^N \rho_x^{(u)} \tag{A.2}$$

$$\rho_N^{(u)} \curvearrowleft \mathcal{I}\left(\rho_N^{(u)}, \beta = \frac{3}{2}, N^2\rho_0\eta\right). \tag{A.3}$$

For a proof see (Giron 2001).

In the case of $\beta = \frac{3}{2}$, the power-law behavior of the prior becomes independent of the discretization of the continuous position space. This means that the slope of the distribution of $\rho_x^{(u)}$ remains unchanged notwithstanding that we refine or coarsen the resolution of the reconstruction. However, the scale parameter η needs to be adapted for each resolution; i.e., $\eta \to N^2\eta$ if N pixels are merged.

Appendix B: Covariance & Curvature.

The covariance $\boldsymbol{D}$ of a Gaussian $\mathcal{G}(\boldsymbol{s}-\boldsymbol{m}, \boldsymbol{D})$ describes the uncertainty associated with the mean $\boldsymbol{m}$ of the distribution. It can be computed by second moments or cumulants according to Eq. (3), or in this Gaussian case as the inverse Hessian of the corresponding information Hamiltonian,

$$\begin{aligned}\left.\frac{\partial^2 H}{\partial\boldsymbol{s}\partial\boldsymbol{s}^\dagger}\right|_{\boldsymbol{s}=\boldsymbol{m}} &= \left.\frac{\partial^2}{\partial\boldsymbol{s}\partial\boldsymbol{s}^\dagger}\left(\frac{1}{2}(\boldsymbol{s}-\boldsymbol{m})^\dagger\boldsymbol{D}^{-1}(\boldsymbol{s}-\boldsymbol{m})\right)\right|_{\boldsymbol{s}=\boldsymbol{m}} \\ &= \boldsymbol{D}^{-1}.\end{aligned} \tag{B.1}$$

In Sec. 3, uncertainty covariances for the diffuse signal field $\boldsymbol{s}$ and the point-like signal field $\boldsymbol{u}$ have been derived that are here given in closed form.

The MAP uncertainty covariances introduced in Sec. 3.1 are approximated by inverse Hessians. According to Eq. (34), they read

$$\begin{aligned}{D^{(s)}_{xy}}^{-1} \approx &\left\{\sum_i \left(1-\frac{d_i}{l_i}\right) R_{ix} \mathrm{e}^{m_x^{(s)}}\right\}\delta_{xy} \\ &+ \sum_i \frac{d_i}{l_i^2}\left(R_{ix}\mathrm{e}^{m_x^{(s)}}\right)\left(R_{iy}\mathrm{e}^{m_y^{(s)}}\right) + {S^\star_{xy}}^{-1},\end{aligned} \tag{B.2}$$

and

$$\begin{aligned}{D^{(u)}_{xy}}^{-1} \approx &\left\{\sum_i \left(1-\frac{d_i}{l_i}\right) R_{ix} \mathrm{e}^{m_x^{(u)}} + \eta\,\mathrm{e}^{m_x^{(u)}}\right\}\delta_{xy} \\ &+ \sum_i \frac{d_i}{l_i^2}\left(R_{ix}\mathrm{e}^{m_x^{(u)}}\right)\left(R_{iy}\mathrm{e}^{-m_y^{(u)}}\right),\end{aligned} \tag{B.3}$$

with

$$l_i = \int \mathrm{d}x\; R_{ix}\left(\mathrm{e}^{m_x^{(s)}} + \mathrm{e}^{m_x^{(u)}}\right). \tag{B.4}$$

The corresponding covariances derived in the Gibbs approach according to Eq. (51), yield

$$\begin{aligned}{D^{(s)}_{xy}}^{-1} \approx &\left\{\sum_i \left(1-\frac{d_i}{l_i}\right) R_{ix} \mathrm{e}^{m_x^{(s)}+\frac{1}{2}D_{xx}^{(s)}}\right\}\delta_{xy} \\ &+ \sum_i \frac{d_i}{l_i^2}\left(R_{ix}\mathrm{e}^{m_x^{(s)}+\frac{1}{2}D_{xx}^{(s)}}\right)\left(R_{iy}\mathrm{e}^{m_y^{(s)}+\frac{1}{2}D_{xx}^{(s)}}\right) \\ &+ {S^\star_{xy}}^{-1},\end{aligned} \tag{B.5}$$

and

$$D^{(u)\,-1}_{xy} \approx \left\{ \sum_i \left(1 - \frac{d_i}{l_i}\right) R_{ix} \mathrm{e}^{m^{(u)}_x + \frac{1}{2} D^{(u)}_{xx}} + \eta\, \mathrm{e}^{-m^{(u)}_x + \frac{1}{2} D^{(u)}_{xx}} \right\} \delta_{xy} \tag{B.6}$$

$$+ \sum_i \frac{d_i}{l_i^2} \left(R_{ix} \mathrm{e}^{m^{(u)}_x + \frac{1}{2} D^{(u)}_{xx}} \right) \left(R_{iy} \mathrm{e}^{m^{(u)}_y + \frac{1}{2} D^{(u)}_{xx}} \right),$$

with

$$l_i = \int \mathrm{d}x\; R_{ix} \left(\mathrm{e}^{m^{(s)}_x + \frac{1}{2} D^{(s)}_{xx}} + \mathrm{e}^{m^{(u)}_x + \frac{1}{2} D^{(u)}_{xx}} \right). \tag{B.7}$$

They are identical up to the $+\frac{1}{2} D_{xx}$ terms in the exponents. On the one hand, this reinforces the approximations done in Sec. 3.2. On the other hand, this shows that higher order correction terms might alter the uncertainty covariances further, cf. Eq. (41). The concrete impact of these correction terms is difficult to judge, since they introduce terms involving D_{xy} that couple all elements of $\boldsymbol{D}$ in an implicit manner.

Notice that the inverse Hessian describes the curvature of the potential, its interpretation as uncertainty is, strictly speaking, only valid for quadratic potentials. However, in most cases it is a sufficient approximation.

The Gibbs approach provides an alternative by equating the first derivative of the Gibbs free energy with respect to the convariace with zero. Following Eq. (52), the covariances read

$$D^{(s)\,-1}_{xy} = \left\{ \sum_i \left(1 - \frac{d_i}{l_i}\right) R_{ix} \mathrm{e}^{m^{(s)}_x + \frac{1}{2} D^{(s)}_{xx}} \right\} \delta_{xy} \tag{B.8}$$

$$+ {S^\star_{xy}}^{-1}, \tag{B.9}$$

and

$$D^{(u)\,-1}_{xy} = \left\{ \sum_i \left(1 - \frac{d_i}{l_i}\right) R_{ix} \mathrm{e}^{m^{(u)}_x + \frac{1}{2} D^{(u)}_{xx}} \right. \tag{B.10}$$

$$\left. + \eta\, \mathrm{e}^{-m^{(u)}_x + \frac{1}{2} D^{(u)}_{xx}} \right\} \delta_{xy}. \tag{B.11}$$

Compared to the above solutions, there is one term missing indicating that they already lack first order corrections. For this reasons, the solutions obtained from the inverse Hessians are used in the D^{3}PO algorithm.

Appendix C: Posterior Approximation

Appendix C.1: Information Theoretical Measure

If the full posterior $P(\boldsymbol{z}|\boldsymbol{d})$ of an inference problem is so complex that an analytic handling is infeasible, an approximate posterior Q might be used instead. The fitness of such an approximation can be quantified by an asymmetric measure for which different terminologies appear in the literature.

First, the Kullback-Leibler divergence,

$$D_{\mathrm{KL}}(Q,P) = \int \mathcal{D}\boldsymbol{z}\; Q(\boldsymbol{z}|\boldsymbol{d}) \log \frac{Q(\boldsymbol{z}|\boldsymbol{d})}{P(\boldsymbol{z}|\boldsymbol{d})} \tag{C.1}$$

$$= \left\langle \log \frac{Q(\boldsymbol{z}|\boldsymbol{d})}{P(\boldsymbol{z}|\boldsymbol{d})} \right\rangle_Q, \tag{C.2}$$

defines mathematically an information theoretical distance, or divergence, which is minimal if a maximal cross information between P and Q exists (Kullback & Leibler 1951).

Second, the information entropy,

$$S_{\mathrm{E}}(Q,P) = -\int \mathcal{D}\boldsymbol{z}\; P(\boldsymbol{z}|\boldsymbol{d}) \log \frac{P(\boldsymbol{z}|\boldsymbol{d})}{Q(\boldsymbol{z}|\boldsymbol{d})} \tag{C.3}$$

$$= \left\langle -\log \frac{P(\boldsymbol{z}|\boldsymbol{d})}{Q(\boldsymbol{z}|\boldsymbol{d})} \right\rangle_P \tag{C.4}$$

$$= -D_{\mathrm{KL}}(P,Q),$$

is derived under the maximum entropy principle (Jaynes 1957) from fundamental axioms demanding locality, coordinate invariance and system independence, cf. Caticha (2008, 2011).

Third, the (approximate) Gibbs free energy (Enßlin & Weig 2010),

$$G = \left\langle H(\boldsymbol{z}|\boldsymbol{d}) \right\rangle_Q - S_{\mathrm{B}}(Q) \tag{C.5}$$

$$= \left\langle -\log P(\boldsymbol{z}|\boldsymbol{d}) \right\rangle_Q - \left\langle -\log Q(\boldsymbol{z}|\boldsymbol{d}) \right\rangle_Q \tag{C.6}$$

$$= D_{\mathrm{KL}}(Q,P),$$

describes the difference between the internal energy $\langle H(\boldsymbol{z}|\boldsymbol{d}) \rangle_Q$ and the Boltzmann-Shannon entropy $S_{\mathrm{B}}(Q) = S_{\mathrm{E}}(1,Q)$. The derivation of the Gibbs free energy is based on the principles of thermodynamics[12].

The Kullback-Leibler divergence, information entropy, and the Gibbs free energy are equivalent measures that allow one to assess the approximation $Q \approx P$. Alternatively, a parametrized proposal for Q can be pinned down by extremizing the measure of choice with respect to the parameters.

Appendix C.2: Calculus of Variations

The information theoretical measure can be interpreted as an action to which the principle of least action applies. This concept is the basis for variational Bayesian methods (Jordan et al. 1999; Wingate & Weber 2013), which enable among others the derivation of approximate posterior distributions.

Let $\boldsymbol{z}$ be a set of multiple signal fields, $\boldsymbol{z} = \{\boldsymbol{z}^{(i)}\}_{i \in \mathbb{N}}$, $\boldsymbol{d}$ a given data set, and $P(\boldsymbol{z}|\boldsymbol{d})$ the posterior of interest. In practice, such a problem is often addressed by a mean field approximation that factorizes the variational posterior Q,

$$P(\boldsymbol{z}|\boldsymbol{d}) \approx Q = \prod_i Q_i(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}, \boldsymbol{d}). \tag{C.7}$$

Here, the mean field $\boldsymbol{\mu}$, which mimics the effect of all $\boldsymbol{z}^{(i \neq j)}$ onto $\boldsymbol{z}^{(j)}$, has been introduced. The approximation in Eq. (C.7) shifts any possible entanglement between the $\boldsymbol{z}^{(i)}$ within P into the dependence of $\boldsymbol{z}^{(i)}$ on $\boldsymbol{\mu}$ within Q_i.

A variation, $\delta_j = \delta / \delta Q_j(\boldsymbol{z}^{(j)}|\boldsymbol{\mu}, \boldsymbol{d})$, of the Gibbs free energy with respect to one approximate posterior

[12] In Eq. (C.5), a unit temperature is implied, cf. discussion by Enßlin & Weig (2010); Iatsenko et al. (2012); Enßlin & Weig (2012)

$Q_j(\boldsymbol{z}^{(j)}|\boldsymbol{\mu},\boldsymbol{d})$, yields

$$\delta_j G = \frac{\delta}{\delta Q_j(\boldsymbol{z}^{(j)}|\boldsymbol{\mu},\boldsymbol{d})} \left\{ \left\langle H(\boldsymbol{z}|\boldsymbol{d}) \right\rangle_Q - \left\langle -\log Q \right\rangle_Q \right\} \quad \text{(C.8)}$$

$$= \frac{\delta}{\delta Q_j(\boldsymbol{z}^{(j)}|\boldsymbol{\mu},\boldsymbol{d})} \left\{ \left\langle H(\boldsymbol{z}|\boldsymbol{d}) \right\rangle_Q + \sum_i \left\langle \log Q_i(\boldsymbol{z}^{(i)}|\boldsymbol{\mu},\boldsymbol{d}) \right\rangle_{Q_i} \right\} \quad \text{(C.9)}$$

$$= \left\langle H(\boldsymbol{z}|\boldsymbol{d})\Big|_{\boldsymbol{z}^{(j)}} \right\rangle_{\prod Q_{i\neq j}} + \log Q_j(\boldsymbol{z}^{(j)}|\boldsymbol{\mu},\boldsymbol{d}) + \text{const.} \quad \text{(C.10)}$$

Following the principle of least action, any variation must vanish; i.e., $\delta_j G = 0$. This defines a solution for the approximate posterior Q_j, where the constant term in Eq. (C.10) ensures the correct normalization[13] of Q_j,

$$Q_j(\boldsymbol{z}^{(j)}|\boldsymbol{\mu},\boldsymbol{d}) \propto \exp\left(- \left\langle H(\boldsymbol{z}|\boldsymbol{d})\Big|_{\boldsymbol{z}^{(j)}} \right\rangle_{\prod Q_{i\neq j}} \right). \quad \text{(C.11)}$$

Although the parts $\boldsymbol{z}^{(i\neq j)}$ are integrated out, Eq. (C.11) is no marginalization since the integration is performed on the level of the (negative) logarithm of a probability distribution. The success of the mean field approach might be that this integration is often more well-behaved in comparison to the corresponding marginalization. However, the resulting equations for the Q_i depend on each other, and thus need to be solved self-consistently.

A maximum *a posteriori* solution for $\boldsymbol{z}^{(j)}$ can then be found by minimizing an effective Hamiltonian,

$$\underset{\boldsymbol{z}^{(j)}}{\operatorname{argmax}}\, P(\boldsymbol{z}|\boldsymbol{d}) = \underset{\boldsymbol{z}^{(j)}}{\operatorname{argmin}}\, H(\boldsymbol{z}|\boldsymbol{d}) \quad \text{(C.12)}$$

$$\approx \underset{\boldsymbol{z}^{(j)}}{\operatorname{argmin}} \left\langle H(\boldsymbol{z}|\boldsymbol{d})\Big|_{\boldsymbol{z}^{(j)}} \right\rangle_{\prod Q_{i\neq j}}. \quad \text{(C.13)}$$

Since the posterior is approximated by a product, the Hamiltonian is approximated by a sum, and each summand depends on solely one variable in the partition of the latent variable $\boldsymbol{z}$.

Appendix C.3: Example

In this section, the variational method is demonstrated with an exemplary posterior of the following form,

$$P(\boldsymbol{s},\boldsymbol{\tau}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{s})}{P(\boldsymbol{d})}\, P(\boldsymbol{s}|\boldsymbol{\tau})\, P(\boldsymbol{\tau}) \quad \text{(C.14)}$$

$$= \frac{P(\boldsymbol{d}|\boldsymbol{s})}{P(\boldsymbol{d})}\, \mathcal{G}(\boldsymbol{s},\boldsymbol{S})\, P_{\mathrm{un}}(\boldsymbol{\tau}|\boldsymbol{\alpha},\boldsymbol{q})\, P_{\mathrm{sm}}(\boldsymbol{\tau}|\boldsymbol{\sigma}), \quad \text{(C.15)}$$

where $P(\boldsymbol{d}|\boldsymbol{s})$ stands for an arbitrary likelihood describing how likely the data $\boldsymbol{d}$ can be measured from a signal $\boldsymbol{s}$, and $\boldsymbol{S} = \sum_k \mathrm{e}^{\tau_k} \boldsymbol{S}_k$ for a parametrization of the signal covariance. This posterior is equivalent to the one derived in Sec. 2 in order to find a solution for the logarithmic power spectrum $\boldsymbol{\tau}$. Here, any explicit dependence on the point-like signal field $\boldsymbol{u}$ is veiled in favor of clarity.

[13] The normalization could be included by usage of Lagrange multipliers; i.e., by adding a term $\sum_i \lambda_i \big(1 - \int \mathcal{D}\boldsymbol{z}^{(i)}\, Q_i(\boldsymbol{z}^{(i)}|\boldsymbol{\mu},\boldsymbol{d})\big)$ to the Gibbs free energy in Eq. (C.8).

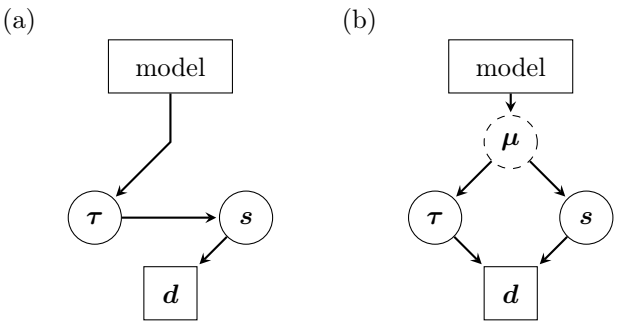

Fig. C.1. Graphical model for the variational method applied to the example posterior in Eq. (C.14). Panel (a) shows the graphical model without, and panel (b) with the mean field $\boldsymbol{\mu}$.

The corresponding Hamiltonian reads

$$H(\boldsymbol{s},\boldsymbol{\tau}|\boldsymbol{d}) = -\log P(\boldsymbol{s},\boldsymbol{\tau}|\boldsymbol{d}) \quad \text{(C.16)}$$

$$= H_0 + \frac{1}{2}\sum_k \left(\varrho_k \tau_k + \mathrm{tr}\left[\boldsymbol{s}\boldsymbol{s}^\dagger \boldsymbol{S}_k^{-1}\right] \mathrm{e}^{-\tau_k} \right) \quad \text{(C.17)}$$

$$+ (\boldsymbol{\alpha}-\boldsymbol{1})^\dagger \boldsymbol{\tau} + \boldsymbol{q}^\dagger \mathrm{e}^{-\boldsymbol{\tau}} + \frac{1}{2}\boldsymbol{\tau}^\dagger \boldsymbol{T} \boldsymbol{\tau},$$

where $\varrho_k = \mathrm{tr}\left[\boldsymbol{S}_k \boldsymbol{S}_k{}^{-1}\right]$ and all terms constant in $\boldsymbol{\tau}$, including the likelihood $P(\boldsymbol{d}|\boldsymbol{s})$, have been absorbed into H_0.

For an arbitrary likelihood it might not be possible to marginalize the posterior over $\boldsymbol{s}$ analytically. However, an integration of the Hamiltonian over $\boldsymbol{s}$ might be feasible since the only relevant term is quadratic in $\boldsymbol{s}$. As, on the one hand, the prior $P(\boldsymbol{s}|\boldsymbol{\tau})$ is Gaussian and, on the other hand, a posterior mean $\boldsymbol{m}$ and covariance $\boldsymbol{D}$ for the signal field $\boldsymbol{s}$ suffice, cf. Eq. (2) and (3), let us assume a Gaussian approximation for Q_s; i.e., $Q_s = \mathcal{G}(\boldsymbol{s}-\boldsymbol{m},\boldsymbol{D})$.

We now introduce a mean field approximation, denoted by $\boldsymbol{\mu}$, by changing the causal structure as depicted in Fig. C.1. With the consequential approximation of the posterior,

$$P(\boldsymbol{s},\boldsymbol{\tau}|\boldsymbol{d}) \approx \mathcal{G}(\boldsymbol{s}-\boldsymbol{m},\boldsymbol{D})\, Q_\tau(\boldsymbol{\tau}|\boldsymbol{\mu},\boldsymbol{d}), \quad \text{(C.18)}$$

we can calculate the effective Hamiltonian for $\boldsymbol{\tau}$ as

$$\left\langle H(\boldsymbol{s},\boldsymbol{\tau}|\boldsymbol{d})\Big|_{\boldsymbol{\tau}} \right\rangle_{Q_s} = H_0 + \boldsymbol{\gamma}^\dagger \boldsymbol{\tau} + \frac{1}{2}\boldsymbol{\tau}^\dagger \boldsymbol{T}\boldsymbol{\tau} + \boldsymbol{q}^\dagger \mathrm{e}^{-\boldsymbol{\tau}} \quad \text{(C.19)}$$

$$+ \frac{1}{2}\sum_k \mathrm{tr}\left[\left\langle \boldsymbol{s}\boldsymbol{s}^\dagger \right\rangle_{Q_s} \boldsymbol{S}_k^{-1} \right] \mathrm{e}^{-\tau_k}$$

$$= H_0 + \boldsymbol{\gamma}^\dagger \boldsymbol{\tau} + \frac{1}{2}\boldsymbol{\tau}^\dagger \boldsymbol{T}\boldsymbol{\tau} + \boldsymbol{q}^\dagger \mathrm{e}^{-\boldsymbol{\tau}} \quad \text{(C.20)}$$

$$+ \frac{1}{2}\sum_k \mathrm{tr}\Big[\left(\boldsymbol{m}\boldsymbol{m}^\dagger + \boldsymbol{D}\right) \boldsymbol{S}_k^{-1} \Big] \mathrm{e}^{-\tau_k},$$

where $\boldsymbol{\gamma} = (\boldsymbol{\alpha}-\boldsymbol{1}) + \frac{1}{2}\boldsymbol{\varrho}$.

The nature of the mean field $\boldsymbol{\mu}$ can be derived from the coupling term in Eq. (C.17) that ensures an information flow between $\boldsymbol{s}$ and $\boldsymbol{\tau}$,

$$\boldsymbol{\mu} = \begin{pmatrix} \left\langle \mathrm{tr}\left[\boldsymbol{s}\boldsymbol{s}^\dagger \boldsymbol{S}_k^{-1}\right] \right\rangle_{Q_s} \\ \left\langle \sum_k \mathrm{e}^{-\tau_k} \boldsymbol{S}_k^{-1} \right\rangle_{Q_\tau} \end{pmatrix} = \begin{pmatrix} \mathrm{tr}\left[\left(\boldsymbol{m}\boldsymbol{m}^\dagger + \boldsymbol{D}\right) \boldsymbol{S}_k^{-1} \right] \\ \left\langle \boldsymbol{S}^{-1} \right\rangle_{Q_\tau} \end{pmatrix} \quad \text{(C.21)}$$

Hence, the mean field effect on τ_k is given by the above trace, and the mean field effect on $\boldsymbol{s}$ is described by $\left\langle \boldsymbol{S}^{-1} \right\rangle_{Q_\tau}$.

Extremizing Eq. (C.20) yields

$$\mathrm{e}^{\boldsymbol{\tau}} = \frac{\boldsymbol{q} + \frac{1}{2}\left(\mathrm{tr}\left[\left(\boldsymbol{m}\boldsymbol{m}^\dagger + \boldsymbol{D}\right)\boldsymbol{S}_k^{-1}\right]\right)_k}{\boldsymbol{\gamma} + \boldsymbol{T}\boldsymbol{\tau}}. \tag{C.22}$$

This formula is in agreement with the critical filter formula (Enßlin & Frommert 2011; Oppermann et al. 2012). In case a Gaussian likelihood and no smoothness prior is assumed, it is the exact maximum of the true posterior with respect to the (logarithmic) power spectrum.