https://doi.org/10.1007/s11222-022-10152-9

Proximal nested sampling for high-dimensional Bayesian model selection

Xiaohao Cai[1,2] · Jason D. McEwen[1,3] · Marcelo Pereyra[4]

Received: 15 April 2022 / Accepted: 9 September 2022 / Published online: 5 October 2022

Abstract

Bayesian model selection provides a powerful framework for objectively comparing models directly from observed data, without reference to ground truth data. However, Bayesian model selection requires the computation of the marginal likelihood (model evidence), which is computationally challenging, prohibiting its use in many high-dimensional Bayesian inverse problems. With Bayesian imaging applications in mind, in this work we present the *proximal nested sampling* methodology to objectively compare alternative Bayesian imaging models for applications that use images to inform decisions under uncertainty. The methodology is based on nested sampling, a Monte Carlo approach specialised for model comparison, and exploits proximal Markov chain Monte Carlo techniques to scale efficiently to large problems and to tackle models that are log-concave and not necessarily smooth (e.g., involving ℓ_1 or total-variation priors). The proposed approach can be applied computationally to problems of dimension $\mathcal{O}(10^6)$ and beyond, making it suitable for high-dimensional inverse imaging problems. It is validated on large Gaussian models, for which the likelihood is available analytically, and subsequently illustrated on a range of imaging problems where it is used to analyse different choices of dictionary and measurement model.

Keywords Nested sampling · MCMC sampling · Marginal likelihood · Bayesian evidence · Inverse problems · Proximal optimisation · Model selection

1 Introduction

High-dimensional inverse problems are ubiquitous in the data and imaging sciences, as well as in the physical and engineering sciences more generally. Due to limitations of the data observation process and measurement noise, or even just due to the nature of the problem at hand, most inverse problems encountered are seriously ill-conditioned or ill-posed (canonical examples include, e.g., medical and radio interferometric imaging; Durmus et al. 2018; Cai et al. 2019; Zhou et al. 2020; Lunz et al. 2021). Developing better methodology for solving challenging inverse problems is a significant focus of the community. The Bayesian statistical framework is currently one of the predominant frameworks to perform inference in inverse problems (Robert and Casella 2004; Pereyra et al. 2016). The choice of the Bayesian model used has a profound impact on the solutions delivered, as alternative models can lead to significantly different point estimations and uncertainty quantification results.

In this article we develop methodology to objectively compare alternative Bayesian models in performing inference in the regime of high-dimensional inverse problems, directly form the observed data and in the absence of ground truth. Motivated by applications in computational imaging, we focus on the comparison of models with posterior distributions that are log-concave and potentially not smooth. In this context, model selection has been traditionally addressed through benchmark experiments involving ground truth data and expert supervision. However, for many applications it is difficult and expensive to produce reliable ground truth

✉ Xiaohao Cai
x.cai@soton.ac.uk

Jason D. McEwen
jason.mcewen@ucl.ac.uk

Marcelo Pereyra
m.pereyra@hw.ac.uk

[1] Mullard Space Science Laboratory (MSSL), University College London (UCL), Dorking RH5 6NT, UK

[2] School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

[3] Alan Turing Institute, London NW1 2DB, UK

[4] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

data. Moreover, for many problems it is simply impossible. Bayesian model selection provides a framework for selecting the most appropriate model directly from the observed data in an objective manner and without reference to ground truth data.

Bayesian model selection requires the computation of the *marginal likelihood* of the data – the average likelihood of a model over its prior probability space – which is also called the *Bayesian evidence*. This quantity is a key ingredient of model selection statistics such as Bayes factors and likelihood ratio tests (Robert 2007). The computation of the marginal likelihood for high-dimensional models is highly non-trivial because it requires the computation of integrals over the (high-dimensional) solution space. For example, in the context of Bayesian imaging problems, the dimension is given by the number of parameters (e.g. pixels) of interest, which frequently reach sizes of $\mathcal{O}(10^5)$–$\mathcal{O}(10^6)$ and beyond. For such settings, the evaluation of the marginal likelihood has been previously considered to be computationally intractable.

Broadly speaking, general purpose Monte Carlo methods can only handle model selection tasks for problems of dimension $\mathcal{O}(10)$ to $\mathcal{O}(10^2)$ (for reviews see Clyde et al. 2007; Friel and Wyse 2012; Llorente et al. 2020). Nested sampling (Skilling 2006), a state-of-the-art Monte Carlo strategy designed specifically for model selection, has enabled model selection for moderate dimensional problems of size $\mathcal{O}(10^2)$ to $\mathcal{O}(10^3)$ (Mukherjee et al. 2006; Feroz and Hobson 2008; Feroz et al. 2009; Brewer et al. 2011; Feroz and Skilling 2013; Handley et al. 2015). To the best of our knowledge, model selection for larger problems is currently possible only for models with very specific structures (e.g., conditionally Gaussian models; Harroue 2020).

In this work, we address the difficult computation of the marginal likelihood by proposing a new methodology that carefully integrates nested sampling (Skilling 2006) with proximal Markov chain Monte Carlo (MCMC) (Pereyra 2016; Durmus et al. 2018). This leads to a *proximal nested sampling* methodology specialised for comparing high-dimensional posterior distributions that are log-concave but potentially not smooth. The proposed approach can be applied computationally to log-concave models of dimension $\mathcal{O}(10^6)$ and beyond, making it suitable for model comparison in Bayesian imaging problems. We demonstrate the approach with a range of scientific imaging applications.

The remainder of the article is organised as follows. In Sect. 2 we recall the Bayesian model selection approach, highlight the associated computational challenges, and discuss proximal MCMC methodology for Bayesian computation for inverse problems with an underlying convex geometry. Section 3 recalls the standard nested sampling method. Our proposed proximal nested sampling framework is presented in general form in Sect. 4. In Sect. 5 explicit forms of proximal nested sampling are presented for common forms of the likelihood and prior that arise in imaging sciences. Experimental results validating the proposed method and showcasing its use in scientific imaging applications are reported in Sect. 6. Finally, we conclude in Sect. 7.

2 Bayesian inference for high-dimensional inverse problems

In this section we briefly recall the Bayesian decision-theoretic approach to model comparison, introduce some elements of convex analysis which are essential for our method, and review proximal MCMC methods, which are an important component of the proximal nested sampling methodology proposed in Sect. 4. We conclude the section by briefly explaining the computational difficulties encountered in high-dimensional Bayesian model selection and why it is necessary to develop new methodology for this task. Readers familiar with Bayesian model selection and with proximal MCMC methodology may prefer to skip this section and continue reading from Sect. 3.

2.1 Bayesian estimation and model selection

Let $\Omega \subseteq \mathbb{R}^d$. We consider the estimation of a quantity of interest $x \in \Omega$ from observed data y. Bayesian methods address such problems by postulating a statistical model $\mathcal{M}$ relating x and y, from which estimators of x and other inferences can be derived. More precisely, $\mathcal{M}$ defines a joint probability distribution $p(x, y|\mathcal{M})$ specified via the decomposition $p(x, y|\mathcal{M}) = p(y|x, \mathcal{M})p(x|\mathcal{M})$, where $p(y|x, \mathcal{M})$ denotes the likelihood of x for the observed data y, and the marginal $p(x|\mathcal{M})$ is the so-called prior of x. Following Bayes' theorem, inferences on $x|y$ are then based on the posterior distribution

$$p(x|y, \mathcal{M}) = \frac{p(y|x, \mathcal{M})p(x|\mathcal{M})}{p(y|\mathcal{M})}, \tag{1}$$

which models our beliefs about x after observing y. With applications in Bayesian imaging sciences in mind, we focus on posterior distributions that are log-concave and assume that the potential function $x \mapsto -\log p(x|y, \mathcal{M})$ is convex lower semicontinuous (l.s.c.) on Ω, but possibly not smooth. This is an important class of models in modern Bayesian imaging sciences because it leads to point estimators that are by construction well-posed and that can be efficiently estimated by using scalable proximal convex optimisation and stochastic sampling methods (Kaipio and Somersalo 2005; Robert and Casella 2004; Pereyra et al. 2016).

We condition on $\mathcal{M}$ explicitly in (1) because our focus is model selection, where one entertains several alternative posterior distributions for $x|y$ stemming from different underlying modelling assumptions. As a result, rather than the posterior $p(x|y, \mathcal{M})$, our main object of interest is the *marginal likelihood* or *model evidence*

$$p(y|\mathcal{M}) = \int_\Omega p(y, x|\mathcal{M})\mathrm{d}x = \int_\Omega p(y|x, \mathcal{M})p(x|\mathcal{M})\mathrm{d}x, \tag{2}$$

which measures the likelihood of the observed data under model $\mathcal{M}$, and which we use to objectively compare different models relating x and y (Robert 2007). Notice that the likelihood of the observed data y under the model $\mathcal{M}$ is essentially the expectation (or average value) of the likelihood function $p(y|x, \mathcal{M})$ with respect to (w.r.t.) the prior $p(x|\mathcal{M})$. Therefore, a model that allocates its prior mass to solutions that agree with the observed data achieves a large marginal likelihood value. Conversely, a low marginal likelihood value indicates that only a small proportion of the solutions favoured by the prior agree with the observed data. In other words, the marginal likelihood (2) measures the degree to which the observed data is in agreement with the assumptions of the model, and in doing so it provides a goodness-of-fit summary. Moreover, because all priors have the same total probability mass (i.e., $\int_\Omega p(x)\mathrm{d}x = 1$), the likelihood (2) naturally incorporates Occams's razor, trading off model simplicity and accuracy and penalising over-fitting (Robert 2007).

Bayesian model selection arises from the common and natural inquiry of which model is the most suitable to analyse $x|y$ from a set of models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ available. For simplicity and without loss of generality, we suppose two alternative models $\mathcal{M}_1$ and $\mathcal{M}_2$ (the generalisation to additional models is straightforward). From Bayesian decision theory, to objectively compare the two models in settings without ground truth available, one should calculate the *Bayes factor* (Robert 2007)

$$\rho_{12} = \frac{p(\mathcal{M}_1|y)}{p(\mathcal{M}_2|y)} \frac{p(\mathcal{M}_2)}{p(\mathcal{M}_1)} \tag{3}$$

where $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ denote the prior probabilities assigned to the two competing models, and where, from Bayes' theorem, we have that for any $i \in \{1, 2\}$

$$p(\mathcal{M}_i|y) = \frac{p(y|\mathcal{M}_i)p(\mathcal{M}_i)}{p(y|\mathcal{M}_1)p(\mathcal{M}_1) + p(y|\mathcal{M}_2)p(\mathcal{M}_2)}. \tag{4}$$

By developing (3) we can easily express the Bayes factor as the likelihood ratio

$$\rho_{12} = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_2)}, \tag{5}$$

highlighting that ρ_{12} is invariant to choice of the prior probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$. If one assumes $p(\mathcal{M}_1) = p(\mathcal{M}_2) = 1/2$ to reflect the absence of prior information, then the factor also coincides with the posterior probability ratio $p(\mathcal{M}_1|y)/p(\mathcal{M}_2|y)$.

Being a likelihood ratio, the factor ρ_{12} is straightforward to read: if $\rho_{12} \gg 1$, we prefer model $\mathcal{M}_1$ over the alternative $\mathcal{M}_2$; conversely, if $\rho_{12} \ll 1$, we prefer model $\mathcal{M}_2$; and if $\rho_{12} \approx 1$, we do not prefer either, inasmuch as the data y are insufficient for us to make an informed judgement. The fact that ρ_{12} is a likelihood ratio is also appealing from a frequentist viewpoint, as it is associated with the most powerful test for these two model hypotheses (Casella and Berger 2002).

Unfortunately, calculating ρ_{12} is generally not possible in large-scale settings because the dimensionality of x renders the marginal likelihoods $p(y|\mathcal{M}_1)$ and $p(y|\mathcal{M}_2)$ computationally intractable. More precisely, the marginal likelihoods are doubly-intractable because they require computing two intractable integrals over the space of solutions Ω: the marginalisation of x denoted explicitly in (2); and the normalising constant of the priors $p(x|\mathcal{M}_i)$ when these are not available analytically, which otherwise implicitly also requires integrating over Ω.

It is worth emphasising at this point that this major difficulty related to model selection is not encountered when performing inferences with the posteriors $p(x|y, \mathcal{M}_1)$ and $p(x|y, \mathcal{M}_2)$ individually, as one can use MCMC methods to sample from $p(x|y, \mathcal{M})$ without ever having to evaluate the marginal likelihood $p(y|\mathcal{M})$. As a result, efficient Bayesian model selection remains an open problem in many areas of science and engineering that have widerly adopted Bayesian inference techniques for point estimation and uncertainty quantification.

In the following we briefly recall MCMC sampling methods derived from the overdamped Langevin diffusion process, particularly proximal MCMC techniques specialised for large models that are log-concave, and explain why it is necesary to modify them to enable efficient model comparison.

2.2 Bayesian computation and proximal MCMC methods

2.2.1 Convex analysis

Let $f : \mathbb{R}^d \to [-\infty, +\infty]$. The function f is said to be proper if there exists $x_0 \in \mathbb{R}^d$ such that $f(x_0) < +\infty$. Denote for all $M \in \mathbb{R}$, $\{f \leq M\} = \{z \in \mathbb{R}^d \mid f(z) \leq M\}$. The function f is l.s.c. if for all $M \in \mathbb{R}$, $\{f \leq M\}$ is a closed subset of $\mathbb{R}^d$. For $k \geq 0$, denoted by $\mathcal{C}^k(\mathbb{R}^d)$ the set of k-times continuously differentiable functions. For $f \in \mathcal{C}^1(\mathbb{R}^d)$, denote by ∇f the gradient of f. We say that $f \in \mathcal{C}^1(\mathbb{R}^d)$ is a Lipschitz continuously differentiable function if there exists

$C \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq C\|x - y\|$.

Given a convex, proper, l.s.c. function $h : \mathbb{R}^d \to (-\infty, +\infty]$ and $\lambda > 0$, the proximal operator (Bauschke and Combettes 2011) associated with function h at $x \in \mathbb{R}^d$ is defined as

$$\operatorname{prox}_h^\lambda(x) = \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \left\{h(u) + \|u - x\|_2^2/2\lambda\right\}. \tag{6}$$

When $\lambda = 1$, we denote $\operatorname{prox}_h^1(x)$ by $\operatorname{prox}_h(x)$ for simplicity.

Let $\mathcal{K}$ be a closed convex set in $\mathbb{R}^d$ and let $\chi_\mathcal{K}$ be the characteristic function for $\mathcal{K}$, defined by $\chi_\mathcal{K}(x) = 0$ if $x \in \mathcal{K}$ and $+\infty$ otherwise. The proximal operator of $\chi_\mathcal{K}$ is the projection onto $\mathcal{K}$, given by

$$\operatorname{proj}_\mathcal{K}(x) = \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \left\{\chi_\mathcal{K}(u) + \|u - x\|_2^2/2\right\}. \tag{7}$$

The convex conjugate of function h, denoted by h^*, is defined as

$$h^*(x) = \sup_{u \in \mathbb{R}^d} \left\{x^\top u - h(u)\right\}. \tag{8}$$

Its proximal operator can be related to the proximal operator of h by

$$\operatorname{prox}_{h^*}(x) = x - \operatorname{prox}_h(x). \tag{9}$$

The λ-Moreau-Yosida envelope of h (Bauschke and Combettes 2011) is given for any $x \in \mathbb{R}^d$ and $\lambda > 0$ by

$$h^\lambda(x) = \min_{u \in \mathbb{R}^d} \left\{h(u) + \|u - x\|_2^2/2\lambda\right\}. \tag{10}$$

The envelope h^λ is continuously differentiable with Lipschitz gradient. In particular, using the proximal operator, the gradient of h^λ can be written

$$\nabla h^\lambda(x) = \left(x - \operatorname{prox}_h^\lambda(x)\right)/\lambda, \tag{11}$$

with λ simultaneously controlling the Lipschitz constant of ∇h^λ as well as the error between h and its smooth approximation h^λ. This approximation error can be made arbitrarily small by reducing λ, at the expense of deteriorating the regularity of ∇h^λ, and consequently the speed of convergence of proximal Bayesian computation algorithms rely on h^λ.

2.2.2 Proximal Langevin MCMC sampling

Consider the problem of calculating probabilities or expectations with respect to (w.r.t.) some distribution $\pi(\mathrm{d}x)$ which admits a density $\pi(x)$ w.r.t. the usual d-dimensional Lebesgue measure. In the context of Bayesian inference, this is typically the posterior $p(x|y, \mathcal{M})$. Evaluating expectations and probabilities w.r.t. π is non-trivial in problems of moderate and high dimension because of the integrals involved, which are usually beyond the scope of analytical and deterministic numerical integration schemes. These calculations are further complicated when the normalising constant of π is not known, as this requires evaluating an additional d-dimensional integral. Monte Carlo sampling methods address these difficulties by simulating a set of samples from π followed by Monte Carlo stochastic integration to compute probabilities and expectations w.r.t. π. While there are different ways of simulating samples from π, we focus on MCMC strategies where one proceeds by constructing a Markov chain that has π as its invariant stationary distribution. Again, there are different methods for constructing such Markov chains (see Robert and Casella 2004 for an excellent introduction to MCMC methodology and Green et al. 2015 for a survey of recent developments in the Bayesian computation literature).

The fastest provably convergent MCMC methods for Bayesian inference models can be derived from the Langevin diffusion process, which we recall below. For simplicity, rather than presenting the approach in full generality, we focus our presentation on proximal overdamped Langevin sampling for non-smooth models, which we later use in the proximal nested sampling method proposed in Sect. 4. For a more exhaustive introduction to the topic please see Vargas et al. (2020, Section 2) and references therein.

Assume that π admits a decomposition $\pi(x) \propto \exp\{-f(x) - g(x)\}$ for all $x \in \mathbb{R}^d$, where $f \in \mathcal{C}^1(\mathbb{R}^d)$ with ∇f Lipschitz continuous with constant L_f, and where g is a proper l.s.c. function that is convex on $\mathbb{R}^d$ but potentially non-smooth (e.g., g could encode constraints on the solution space and involve non-smooth regularisers such as the ℓ_1 norm). To simulate from π, we construct the overdamped Langevin stochastic differential equation (SDE) on $\mathbb{R}^d$ given by Durmus et al. (2018)

$$\mathrm{d}X_t = -[\nabla f(\mathrm{X}_t) + \nabla g^\lambda(\mathrm{X}_t)]\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0 = x_0, \tag{12}$$

where $(W_t)_{t \geq 0}$ is a d-dimensional Brownian motion, g^λ is the Moreau-Yosida envelop of g given by (10), $\lambda > 0$ is a smoothing parameter that we will discuss later, and $x_0 \in \mathbb{R}^d$. When $x \to f(x) + g^\lambda(x)$ is convex, the SDE has a unique strong solution and X_t converges exponentially fast (as $t \to \infty$) to an invariant measure that is in the neighbourhood of π.

To use (12) for Bayesian computation, we use a numerical solver to compute a discrete-time approximation of X_t over some time period $t \in [0, T]$; the resulting discrete sample path constitutes our set of Monte Carlo samples. In particular, in this article we use the conventional Euler-Maruyama approximation

$$X_{n+1} = X_n - \frac{\delta}{2}\nabla f(X_n) - \frac{\delta}{2}\nabla g^{\lambda}(X_n) + \sqrt{\delta} Z_{n+1}, \qquad (13)$$

where $\delta \in [0, 1/(L_f + 1/\lambda)]$ is a given stepsize and $(Z_n)_{n\geq 1}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables. This MCMC method is known as the Moreau-Yosida unadjusted Langevin algorithm (MYULA) (Durmus et al. 2018). The Markov chain (13) is usually implemented by using (11) and reads

$$X_{n+1} = X_n - \frac{\delta}{2}\nabla f(X_n) - \frac{\delta}{2\lambda}\left(X_n - \text{prox}_g^{\lambda}(X_n)\right) + \sqrt{\delta} Z_{n+1}. \qquad (14)$$

The smoothing parameter λ and the stepsize δ jointly control a bias-variance trade-off between the asymptotic estimation errors and non-asymptotic errors associated with using a finite number of iterations. In this article, we use $\lambda = 1/L_f$ and $\delta = 0.8/(L_f + 1/\lambda)$ as recommended in Durmus et al. (2018) (recall that ∇f is Lipschitz continuous with constant L_f, please see Durmus et al. 2018; Vargas et al. 2020 for further details).

The samples generated by (14) can be directly used for biased Monte Carlo estimation (Durmus et al. 2018). Alternatively, at the expense of additional computation, one can supplement each iteration of MYULA with an MH (Metropolis-Hastings) correction step to asymptotically remove the approximation errors related to the discretisation of the SDE and the use of g^{λ} instead of g, leading to a type of Metropolis-adjusted Langevin algorithm (MALA) (see Pereyra 2016 for details).

2.3 Estimation of marginal likelihoods and Bayes factors

Let $\{X_n\}_{n=1}^N$ be a set of samples from π (or an approximation of π), generated by using a proximal MCMC method or otherwise. Following a Monte Carlo integration approach, the expectation of any function $\phi : \mathbb{R}^d \to \mathbb{R}$ w.r.t. π is approximated by

$$\hat{\text{E}}_{\pi}(\phi) = \frac{1}{N}\sum_{n=1}^{N} \phi(X_n)\,, \qquad (15)$$

which, under assumptions, converges to the truth $\text{E}_{\pi}(\phi) = \int_{\Omega} \phi(x)\pi(x)\text{d}x$ as N increases (or to a biased estimate if the samples are not exactly from π). The accuracy of Monte Carlo estimates depends of course on the number of samples N and on the properties of the MCMC method used, but it also depends crucially on the variance $\text{Var}_{\pi}(\phi)$. Unfortunately, $\text{Var}_{\pi}(\phi)$ is often very large for the kinds of functions ϕ required for estimating the marginal likelihood (2) (and in some cases $\text{Var}_{\pi}(\phi)$ is not even defined), leading to Monte Carlo estimators of the marginal that behave poorly (Newton and Raftery 1994). As a result, it is difficult to use the samples $\{X_n\}_{n=1}^N$ to perform model selection. Several strategies have been proposed to address the aforementioned difficulty and derive well-posed estimators for the marginal likelihood (2) and the Bayes factor (3) (for reviews of classical methods see Friel and Wyse 2012; Clyde et al. 2007).

One avenue is to generate samples from a sequence of distributions bridging π to some tractable reference π_0 such as the prior distribution or a Gaussian approximation of π, e.g., thermodynamic integration (O'Ruanaidh and Fitzgerald 1996) and annealed important sampling (Neal 2001). Such strategies struggle with large problems because the number of intermediate distributions grows quickly as d increases.

Another promising approach to derive computationally efficient estimators is to construct Rao-Blackwellized estimators by carefully introducing auxiliary variables, as proposed in the seminal papers Chib (1995) and Chib and Jeliazkov (2001). This strategy has been successfully applied recently to signal and image processing models that are conditionally Gaussian given conjugate model hyper-parameters (Harroue 2020). Some generalisations are possible, but constructing efficient Rao-Blackwellized estimators for more general classes of models, e.g., of the form (1), is highly non-trivial.

An alternative natural strategy for stable Monte Carlo estimators for (2) and (3) is to construct a truncated estimator by first using the samples $\{X_n\}_{n=1}^N$ to identify a suitable truncating set $\mathcal{A}$, followed by a sample average (15) only with the samples verifying $X_n \in \mathcal{A}$ (Brosse et al. 2017). Although by construction well-posed, truncated estimators need to be de-biased by using the volume of $\mathcal{A}$, which is usually very expensive to compute when the dimension d is large. From the results of Brosse et al. (2017), we believe that this strategy is unlikely to produce scalable methods suitable for large problems. One can circumvent or simplify the calculation of the volume of $\mathcal{A}$ (e.g., see Durmus et al. 2018), but in our experience the resulting estimators become unstable and are difficult to use.

Another alternative approach, which is agnostic to the sampling method, is the harmonic mean estimator (Newton and Raftery 1994); although, in its original form the variance of the estimator can be very poorly behaved such that the estimator can be highly inaccurate in practice. Strategies to resolve this issue have been developed in the recently proposed learnt harmonic mean estimator (McEwen et al. 2022), which has been shown to be highly effective and can scale to dimension $\mathcal{O}(10^3)$ and beyond. Nevertheless, it may be challenging to scale this approach to the high-dimensional settings considered in this paper.

One can also consider the widely used Laplace's method (Tierney and Kadane 1986), which relies on the assumption that the posterior distribution can be adequately approximated by a Gaussian distribution. Unfortunately, this is a strong assumption that often leads to inaccurate estimates in inverse problems that are ill-conditioned or ill-posed, particularly if $d \geq \dim(y)$. Many other alternatives are described in the literature, e.g., the Savage-Dickey density ratio (Trotta 2007) and Reversible Jump MCMC (Green 1995), which are mainly useful for nested or small models. It is worth mentioning that there are also some model selection strategies that do not rely on the computation of the marginal likelihood (see, e.g., Kamary et al. 2018; Pereyra and McLaughlin 2016); however these are usually very computationally intensive.

Finally, nested sampling provides a distinctively different approach for efficiently estimating (2) and (3) (Skilling 2006). The key idea underpinning nested sampling is the re-parameterisation of the marginal likelihood (2) as a one-dimensional integral of the likelihood with respected to the enclosed prior volume. This greatly reduces the computation costs involved, provided that one can efficiently sample from the prior distribution subject to a hard constraint on the likelihood value. Nested sampling therefore shifts the computational challenge from the direct evaluation of a high-dimensional integral to sampling of the prior subject to a hard likelihood constraint. The generation of samples is challenging and previous works have considered a range of sampling strategies. For example, conventional MCMC sampling (Skilling 2006), rejection sampling (e.g. Mukherjee et al. 2006; Feroz and Hobson 2008; Feroz et al. 2009), slice sampling (e.g. Handley et al. 2015), and more advanced MCMC samplers such as Galilean Monte Carlo (Feroz and Skilling 2013) and diffusive nested sampling (Brewer et al. 2011). Following over a decade of active research, nested sampling is now a well-established technique for computing the marginal likelihood that has found widespread application, particularly in astronomy (e.g. Feroz and Hobson 2008; Feroz et al. 2009; Trotta 2007). Nevertheless, broadly speaking, current nested sampling techniques remain restricted to moderate dimensional problems of size $\mathcal{O}(10^2)$ to $\mathcal{O}(10^3)$.

With imaging problems in mind, this article presents an efficient nested sampling methodology specifically designed for high-dimensional log-concave models of the form (1). A significant novelty of the proposed approach is that we address the difficult generation of samples by using a proximal MCMC technique that is naturally suited for dealing with high-dimensional log-concave distributions subject to hard convex constraints. Moreover, the proximal nature of the method straightforwardly allows the use of the non-smooth priors that are frequently encountered in imaging (e.g., involving the ℓ_1 and total-variation regularisers), which would not be easily addressed by using alternative gradient-based samplers. Section 3 below reviews the nested sampling approach. The proposed proximal nested sampling methodology is presented in Sect. 4.

3 Nested sampling

For ease of notation, given a model $\mathcal{M}$, let $\mathcal{L}(x) = p(y|x, \mathcal{M})$ denote the likelihood function, $\pi(x) = p(x|\mathcal{M})$ the prior, and

$$\mathcal{Z} = p(y|\mathcal{M}) = \int_{\Omega} \mathcal{L}(x)\pi(x)\mathrm{d}x, \tag{16}$$

the marginal likelihood or evidence associated with a given model $\mathcal{M}$ (to simplify notation, we henceforth omit the dependence of $\mathcal{Z}$ and $\mathcal{L}$ on y).

Nested sampling (Skilling 2006) was proposed specifically to facilitate the efficient evaluation of $\mathcal{Z}$ for Bayesian model selection, while also supporting posterior inferences. As mentioned previously, the calculation of the multidimensional marginal likelihood integral (16) is generally computationally intractable. Nested sampling addresses this difficulty by cleverly converting (16) to a one-dimensional integral by re-parameterising the likelihood in terms of the enclosed prior volume. In addition, nested sampling involves the prior via simulation and hence does not require knowledge of the prior normalising constant. As a result, it also circumvents the second level of intractability of $\mathcal{Z}$ that arises in imaging problems.

Let $\Omega_{L^*} = \{x|\mathcal{L}(x) > L^*\}$, which groups the parameter space Ω into a series of nested subspaces according to the level-set or iso-likelihood contour $\mathcal{L}(x) = L^* \geq 0$. Note that $\Omega_{L^*=0} = \Omega$, since the likelihood values cannot be negative. Define the prior volume ξ by

$$\xi(L^*) = \int_{\Omega_{L^*}} \pi(x)\mathrm{d}x. \tag{17}$$

Note that $\xi(0) = 1$ and $\xi(L_{\max}) = 0$, where $L_{\max}$ is the maximum of the likelihood in Ω. Let $\mathcal{L}^{\dagger}(\xi)$ be the inverse of the prior volume $\xi(L^*)$ such that $\mathcal{L}^{\dagger}(\xi(L^*)) = L^*$[1], and assume it is a monotonically decreasing function of ξ (which, when $\mathcal{L}$ is continuous and π has connected support, is satisfied theoretically and up to practical numerical considerations that can be trivially overcome; Sivia and Skilling 2006). The marginal likelihood integral (16) can then be rewritten as

$$\mathcal{Z} = \int_0^1 \mathcal{L}^{\dagger}(\xi)\mathrm{d}\xi, \tag{18}$$

[1] In other words, $\mathcal{L}^{\dagger}$ is a tail quantile function such that, for any $L^* > 0$, the inverse of $\mathcal{L}^{\dagger}(\xi(L^*))$ represents the probability that a draw x from the prior π will have a likelihood $\mathcal{L}(x) > L^*$.

which is a one-dimensional integral over the prior volume ξ.

To evaluate (18) in practice it is necessary to compute likelihood level-sets (iso-contours) L_i, which correspond to prior volumes $0 < \xi_i \leq 1$ satisfying (17). A strategy to generate the likelihoods L_i and associated prior volumes ξ_i is discussed in Sect. 3.2. Once the likelihoods $L_i = \mathcal{L}^{\dagger}(\xi_i)$ are obtained, (18) can be used to evaluate the marginal likelihood, where $\{\xi_i\}_{i=0}^{N}$ is a sequence of decreasing prior volumes, i.e.,

$$0 < \xi_N < \cdots < \xi_1 < \xi_0 = 1. \tag{19}$$

After discretising the integral (18) and associating each likelihood L_i a quadrature weight w_i, the marginal likelihood can be computed numerically using standard quadrature methods to give

$$\mathcal{Z} \approx \sum_{i=1}^{N} L_i w_i. \tag{20}$$

The simplest assignment of the quadrature weights is $w_i = \xi_{i-1} - \xi_i$. The trapezium rule can also be used, i.e., $w_i = (\xi_{i-1} + \xi_{i+1})/2$. The approximation error related to the discretisation of (18) can be made arbitrarily small by increasing N.

3.1 Posterior inferences

Posterior inferences can be easily computed once $\mathcal{Z}$ is found. Any sample taken randomly in the prior volume interval (ξ_{i-1}, ξ_i) is simply assigned an importance weight

$$p_i = \frac{L_i w_i}{\mathcal{Z}}. \tag{21}$$

Samples with the assigned weights $\{p_i\}$ can then be used to calculate posterior inferences such as the posterior moments, probabilities, and credible regions.

3.2 Marginal likelihood evaluation

We now recall the basic procedure of the standard nested sampling framework for evaluating the marginal likelihood, i.e. to compute the summation (20). In particular, it is necessary to generate samples of the likelihoods L_i and to estimate the corresponding enclosed prior volume ξ_i.

Firstly, set the iteration number $i = 0$, the prior volume $\xi_0 = 1$, and draw N_{live} *live* samples of the unknown image x from the prior distribution $\pi(x)$. Secondly, remove the sample with the smallest likelihood, say L_{i+1}, from the live set and replace it with a new sample. This new sample is again drawn from the prior, but constrained to a higher likelihood than L_{i+1}.

It is necessary to then determine the prior volume ξ_{i+1} enclosed by the likelihood level-set (iso-contour) defined by L_{i+1}. This is estimated in a stochastic manner. The enclosed prior volume for each step i can be estimated by a shrinkage ratio (random variable) t_{i+1}, i.e. by $\xi_{i+1} = t_{i+1}\xi_i$, where t_{i+1} follows the distribution[2]

$$p(t) = N_{\text{live}} t^{N_{\text{live}}-1}. \tag{22}$$

Repeat the above step (removing the sample with the smallest likelihood and estimating the updated prior volume) until the entire prior volume (and the nested shells of likelihood) has been traversed. We finally obtain $\{L_i\}$ and $\{\xi_i\}$ which can then be used to compute the marginal likelihood by (20). Moreover, we also simultaneously obtain a set of samples of the parameter x comprising all the discarded (dead) samples and the N_{live} final live samples, which can be used for posterior parameter inferences (refer to Sect. 3.1 for further detail).

The volume prior at step i of the nested sampling algorithm, is $\xi_i = \prod_{k=1}^{i} t_k$; recall that t_k is the shrinkage ratio and is independently distributed following the probability density function given in (22). Since the mean and standard deviation of $\log t$ are respectively

$$E(\log t) = -1/N_{\text{live}} \quad \text{and} \quad \sigma(\log t) = 1/N_{\text{live}}, \tag{23}$$

we have

$$\log \xi_i \approx -i/N_{\text{live}} \pm \sqrt{i}/N_{\text{live}}. \tag{24}$$

Ignoring uncertainty, one thus takes

$$\xi_i = \exp(-i/N_{\text{live}}). \tag{25}$$

A convergence criteria for the nested sampling algorithm should be adopted. Terminating the algorithm too early or late should be avoided to ensure the marginal likelihood is estimated accurately without unnecessary additional computational cost. One stopping criterion is that the difference in marginal likelihood estimates between two iterations falls below a predefined threshold, while another is to ensure a sufficient number of dead samples is used.

The pseudo code for the nested sampling algorithm is given in Algorithm 1. Observe that the most challenging task in the nested sampling algorithm is drawing samples from the prior with the hard constraint that samples lie within Ω_{L_i}, i.e. within the space defined by the likelihood level-set (see lines 8–10 in Algorithm 1). This constrained sampling step is

[2] The probability distribution (22) is for the largest of N_{live} samples drawn uniformly from the interval [0, 1]. This follows since the parameter x is uniformly sampled from the prior $\pi(x)$ and $\{\xi_i\}$ are uniformly distributed (by the relation $\mathrm{d}\xi = \pi(x)\mathrm{d}x$).

relatively easy in small problems but can become very computationally challenging as problem dimension increases. As a result, nested sampling is usually restricted to problems of moderate size.

Algorithm 1 Nested sampling algorithm

```
Initialization: Data $Y$. Set $\mathcal{Z} = 0$, $\xi_0 = 1$ and $i = 0$. Draw $N_{\text{live}}$ samples $\{x_n\}_{n=1}^{N_{\text{live}}}$ from the prior distribution $\pi(x)$ in the prior space $\Omega$.
Output: Evidence $\mathcal{Z}$ and posterior probabilities $\{p_i\}$.
for $i = 1, \ldots,$ until the stopping criterion reached
    - Find the lowest likelihood, say $L_i$, in the set of live samples.
    - Compute weight $w_i = (\xi_{i-1} - \xi_{i+1})/2$, where $\xi_i = \exp(-i/N_{\text{live}})$.
    - Update evidence by $\mathcal{Z} = \mathcal{Z} + L_i w_i$.
    - Draw a new sample from the prior distribution $\pi(x)$ in the restricted parameter space $\Omega_{L_i}$, and replace the individual sample associated with the lowest likelihood $L_i$ in the set of live samples.
end for
Update the evidence by $\mathcal{Z} = \mathcal{Z} + \frac{w_{i+1}}{N_{\text{live}}} \sum_{n=1}^{N_{\text{live}}} \mathcal{L}(x_n)$.
Compute the posterior probability for each individual sample $p_i = L_i w_i / \mathcal{Z}$.
```

3.3 Error estimation

If the prior volumes $\{\xi_i\}$ considered in the discretised integral (20) used to evaluate the marginal likelihood could be assigned exactly, then the only error in the estimate of the marginal likelihood would be due to the discretisation of the integral, which is trivially $\mathcal{O}(1/N^2)$ and negligible when N is sufficiently large. However, since the shrinkage ratio t_i is generated randomly, each prior volume ξ_i is then assigned approximately, which tends to overwhelm the error brought by the discretisation of the integral and will therefore cause the dominant source of uncertainty in the final computed evidence $\mathcal{Z}$. This uncertainty, fortunately, can be estimated easily. We recall below the error estimation scheme presented in Skilling (2006) using the entropy of the prior volumes. This approach is highly efficient since it does not require any additional sampling.

Let $\mathcal{P}(\xi) = \mathcal{L}(\xi)/\mathcal{Z}$ be the posterior distribution regarding the prior volume ξ. Then the negative relative entropy H can be defined as

$$H = \int \mathcal{P}(\xi) \log[\mathcal{P}(\xi)] \mathrm{d}\xi \approx \sum_{i=1}^{N} \frac{L_i w_i}{\mathcal{Z}} \log\left(\frac{L_i}{\mathcal{Z}}\right), \quad (26)$$

which can be computed directly from the obtained likelihoods $\{L_i\}$, weights $\{w_i\}$ and the evidence $\mathcal{Z}$. Following Skilling (2006), the standard deviation of the uncertainty of $\log \mathcal{Z}$ using the nested sampling algorithm reads $\sqrt{H/N_{\text{live}}}$, i.e.,

$$\log \mathcal{Z} = \log\left(\sum_{i=1}^{N} L_i w_i\right) \pm \sqrt{\frac{H}{N_{\text{live}}}}. \quad (27)$$

In Chopin and Robert (2010), it is established that, under some regularity conditions, the approximation error is asymptotically Gaussian in the limit $N \to \infty$ and vanishes at the usual Monte Carlo rate $\mathcal{O}(N^{-1/2})$. Moreover, the error scales approximately linearly with the model dimension d.

4 Proximal nested sampling framework

The main difficulty in applying nested sampling to large inverse problems is to efficiently simulate from the prior distribution subject to a hard likelihood constraint. More precisely, at iteration i, the samples from the prior are constrained to the region Ω_{L_i} defined by the likelihood level-set corresponding to L_i (i.e. where a new sample must have a likelihood value greater than L_i at iteration i).

In this section we present our proposed *proximal nested sampling* method to address this challenging constrained sampling problem. Moreover, the proximal nature of the sampling method ensures that non-differentiable distributions, such as popular sparsity-promoting priors involving the ℓ_1 norm, are supported. We first present the methodology of proximal nested sampling for arbitrary log-concave distributions of the form (1). Explicit forms of proximal nested sampling for common choices of priors and likelihoods in imaging sciences are presented in Sect. 5.

4.1 General constrained sampling problem

Following (1) and adopting the notation of Sect. 3, assume that the prior and the likelihood are of the form $\pi(x) = \exp(-f(x))$ and $\mathcal{L}(x) = \exp(-g(x))$, where f and g are convex l.s.c. (lower semicontinuous) functions on Ω.

We consider sampling from the prior $\pi(x)$, such that $\mathcal{L}(x) > L^*$ for some generic likelihood value $L^* > 0$. Let $\iota_{L^*}(x)$ and $\chi_{L^*}(x)$ be the indicator function and characteristic function, respectively, defined as

$$\iota_{L^*}(x) = \begin{cases} 1, & \mathcal{L}(x) > L^*, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and}$$

$$\chi_{L^*}(x) = \begin{cases} 0, & \mathcal{L}(x) > L^*, \\ +\infty, & \text{otherwise}. \end{cases} \quad (28)$$

Since log is monotonic, $\mathcal{L}(x) > L^*$ is equivalent to $g(x) < \tau$, where

$$\tau = -\log L^*. \quad (29)$$

Let $\mathcal{B}_\tau := \{x \mid g(x) < \tau\}$. Then it is apparent that $\chi_{L^*}(x)$, as a constraint for x, is equivalent to $\chi_{\mathcal{B}_\tau}(x)$, where

$$\chi_{\mathcal{B}_\tau}(x) = \begin{cases} 0, & x \in \mathcal{B}_\tau, \\ +\infty, & \text{otherwise.} \end{cases} \tag{30}$$

Let $\pi_{L^*}(x) = \pi(x)\iota_{L^*}(x)$ represent the prior distribution with the hard likelihood constraint $\mathcal{L}(x) > L^*$. Since $\iota_{L^*}(x) = \exp(-\chi_{L^*}(x))$, then we have

$$\begin{aligned} \pi_{L^*}(x) &= \pi(x)\iota_{L^*}(x) \\ &= \exp(-f(x))\exp(-\chi_{L^*}(x)) \\ &= \exp(-[f(x) + \chi_{L^*}(x)]) \\ &= \exp(-[f(x) + \chi_{\mathcal{B}_\tau}(x)]). \end{aligned} \tag{31}$$

Note that taking logarithm of $\pi_{L^*}(x)$ reads

$$-\log \pi_{L^*}(x) = f(x) + \chi_{\mathcal{B}_\tau}(x). \tag{32}$$

In the following section we introduce our proximal nested sampling algorithm for parameter x to sample from the constrained prior distribution $\exp(-[f(x) + \chi_{\mathcal{B}_\tau}(x)])$.

4.2 Drawing a sample from the constrained prior

Sampling distributions over Ω is usually challenging because of the dimensionality involved. Sampling from the constrained prior (32) is particularly difficult because of the hard constraint that $x \in \mathcal{B}_\tau$, encoded in the characteristic function $\chi_{\mathcal{B}_\tau}(x)$. Sampling is further complicated if the log-prior $f(x)$ is not Lipschitz differentiable over Ω (e.g. for non-differentiable sparsity-promoting priors), since high-dimensional sampling methods rely heavily on gradient information. To circumvent these issues we adopt a proximal MCMC approach, which is particularly suitable for high-dimensional distributions that are log-concave but not smooth. More precisely, in a manner akin to Durmus et al. (2018), we use the unadjusted Langevin algorithm (ULA) MCMC sampling strategy combined with Moreau-Yosida approximations of non-differential terms, followed by Metropolis Hastings correction step to control the approximations made, as described in Pereyra (2016).

Using the ULA iterative formula, for each given τ (recall that τ corresponds to a likelihood value L^* by $\tau = -\log L^*$; see (29)), we can generate the following Markov chain

$$x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla\big[f(x^{(k)}) + \chi_{\mathcal{B}_\tau}(x^{(k)})\big] + \sqrt{\delta}w^{(k+1)}, \tag{33}$$

where $\delta > 0$ is the step size and $w^{k+1} \sim \mathcal{N}(0, \mathbf{1}_K)$ (a K-sequence of standard Gaussian random variables).

The non-differentiable characteristic function $\chi_{\mathcal{B}_\tau}(x)$ can be approximated by its Moreau-Yosida envelope $\chi^\lambda_{\mathcal{B}_\tau}(x)$, with approximation controlled by $\lambda > 0$. It is straightforward to show that

$$\chi^\lambda_{\mathcal{B}_\tau}(x) = \frac{1}{2\lambda}\|x - x^*\|_2^2, \tag{34}$$

where x^* is the closest point in $\mathcal{B}_\tau$ to x, given by the projection of x onto $\mathcal{B}_\tau$, i.e. $x^* = \text{proj}_{\mathcal{B}_\tau}(x) = \text{prox}_{\chi_{\mathcal{B}_\tau}}(x)$. Critically, the λ-Moreau-Yosida envelope is $\frac{1}{\lambda}$-Lipschitz differentiable. Its gradient can be calculated directly from (34) or by noting (11), yielding

$$\nabla\chi^\lambda_{\mathcal{B}_\tau}(x) = (x - x^*)/\lambda = (x - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x))/\lambda. \tag{35}$$

Replacing the characteristic function by its Moreau-Yosida approximation in (33) , and noting the gradient (35), yields

$$\begin{aligned} x^{(k+1)} = x^{(k)} - \frac{\delta}{2}\nabla f(x^{(k)}) - \frac{\delta}{2\lambda}\big[x^{(k)} - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\big] \\ +\sqrt{\delta}w^{(k+1)}. \end{aligned} \tag{36}$$

When $f(x)$ is differentiable its gradient can be computed directly (we consider the case where $f(x)$ is non-differentiable shortly). For differential log-priors $f(x)$, (36) provides the general strategy for sampling from the prior subject to the hard likelihood constraint (with a subsequent Metropolis-Hasting step as discussed below).

If the sample $x^{(k)}$ is already in $\mathcal{B}_\tau$, i.e. $x \in \mathcal{B}_\tau$, the term $\big[x^{(k)} - \text{prox}^\lambda_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\big]$ disappears and the Markov chain iteration simply involves taking a noisy step to descent the gradient. In contrast, if $x^{(k)}$ is not in $\mathcal{B}_\tau$, i.e. $x \notin \mathcal{B}_\tau$, then a step is taken in the direction $-\big[x^{(k)} - \text{prox}^\lambda_{\chi_{\mathcal{B}_\tau}}(x^{(k)})\big]$, which acts to move the next iteration in the Markov chain in the direction of the projection of $x^{(k)}$ onto the convex set $\mathcal{B}_\tau$. This term therefore acts to push the Markov chain back into the constraint set $\mathcal{B}_\tau$ if it wanders outside of it, although due to the Moreau-Yosida approximation of $\chi_{\mathcal{B}_\tau}$ it does not guarantee the constraint is satisfied (the subsequent Metropolis-Hasting step does guarantee the hard likelihood constraint is satisfied as discussed below).

When $f(x)$ is non-differentiable, it may be approximated by its differentiable Moreau-Yosida envelope $f^\lambda(x)$. By noting (11), the gradient of the term involving the sum of the two Moreau-Yosida approximations then reads

$$\begin{aligned} &\nabla(f^\lambda(x) + \chi^\lambda_{\mathcal{B}_\tau}(x)) = (x - \text{prox}^\lambda_f(x))/\lambda \\ &+(x - \text{prox}_{\chi_{\mathcal{B}_\tau}}(x))/\lambda. \end{aligned} \tag{37}$$

Here we have used the same regularisation parameter $\lambda > 0$ for both approximations for notational brevity, although clearly different parameters can be considered for $f^{\lambda}(x)$ and $\chi^{\lambda}_{\mathcal{B}_\tau}(x)$ if desired.

Replacing in (33) both $f(x)$ and $\chi_{\mathcal{B}_\tau}(x)$ by their Moreau-Yosida approximations, and noting the gradient (37), yields

$$x^{(k+1)} = (1 - \frac{\delta}{\lambda})x^{(k)} + \frac{\delta}{2\lambda}\mathrm{prox}^{\lambda}_{f}(x^{(k)}) + \frac{\delta}{2\lambda}\mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)}) + \sqrt{\delta}w^{(k+1)}. \tag{38}$$

For non-differentiable log-concave priors, (38) provides the general strategy for sampling from the prior subject to the hard likelihood constraint.

To summarise, given a proper initial sample, say $x^{(0)}$, we generate a Markov chain by iteratively applying the Markov kernel (36) if f is Lipschitz differentiable or the regularised surrogate (38) if it is not, which allows drawing samples from the prior that are likely to be within the likelihood iso-contour L^*. This is the main challenge in nested sampling.

The Markov chains generated by ULA-type kernels exhibit some bias resulting from the discretisation of the Langevin stochastic differential equation and from the use of Moreau-Yosida regularisations. This bias can be asymptotically removed by introducing a Metropolis-Hasting correction step to ensure convergence to the required target density. In detail, at each iteration, a new candidate x' generated using formula (36) or (38) is then accepted with probability

$$\min\left\{1, \frac{q(x^{(k)}|x')\pi_{L^*}(x')}{q(x'|x^{(k)})\pi_{L^*}(x^{(k)})}\right\}, \tag{39}$$

where $q(\cdot|\cdot)$ is a transition kernel, which we define by a Gaussian related to the ULA random component (following Pereyra 2016), i.e.,

$$q(x'|x^{(k)}) \sim \exp\Big(-\frac{\big(x' - x^{(k)} - \frac{\delta}{2}\nabla\log\pi_{L^*}(x^{(k)})\big)^2}{2\delta}\Big). \tag{40}$$

If the candidate sample x' is outside of $\mathcal{B}_\tau$, i.e. $x' \notin \mathcal{B}_\tau$, then $\pi_{L^*}(x') = 0$ and according to the Metropolis-Hasting update the candidate will not be accepted, ensuring the hard likelihood constraint is satisfied.

We summarise our proximal technique to draw an individual sample from the prior under the hard likelihood constraint in Algorithm 2.

Algorithm 2 Proximal individual sample draw algorithm

```
ProxSampleDraw($x^{(0)}$, $L^*$)
Initialization: $k = 0$, $K_{\text{gap}}$.
Input: $x^{(0)}$, $L^*$ (starting point of Markov chain and likelihood threshold).
Output: Individual sample $x_{\text{new}}$ fulfilling the constraint $\mathcal{L}(x_{\text{new}}) > L^*$.
Compute $\tau = -\log L^*$.
for $k = 1, \ldots$
  - Compute $x^{(k)}$ using the iterative formula (36) if $f$ is differentiable; otherwise (38).
  - Metropolis-Hasting step following (39) to remove the estimation bias.
  if $\mathcal{L}(x^{(k)}) > L^*$ and $k \geq K_{\text{gap}}$
    break.
  end if
end for
Set $x_{\text{new}} = x^{(k)}$.
```

4.3 Initialisation from the unconstrained prior

The initialisation of the nested sampling method is to draw N_{live} samples $\{x_n\}_{n=1}^{N_{\text{live}}}$ from the prior distribution $\pi(x)$ in the prior space Ω. If the log-prior $f(x)$ is differentiable this may be applied trivially with the ULA iterative formula. Otherwise $f(x)$ may again be approximated by its Moreau-Yosida envelope and samples from the prior can be generated by the iterative formula

$$x^{(k+1)} = (1 - \frac{\delta}{2\lambda})x^{(k)} + \frac{\delta}{2\lambda}\mathrm{prox}^{\lambda}_{f}(x^{(k)}) + \sqrt{\delta}w^{(k+1)}. \tag{41}$$

To draw N_{live} samples from the prior, it is necessary to first discard initial samples generated before converging on the target prior distribution. Initial samples corresponding to a number of burn-in iterations, say K_{burn}, are discarded. Due to correlations between samples and the algorithm's memory footprint, the chain is thinned by discarding a number of intermediate iterations between samples (the chain's thinning factor), say $(K_{\text{gap}} - 1)$. That is, only the K_{gap}-th sample generated by the iterative formula is kept. Only 1-in-K_{gap} samples are stored when $k > K_{\text{burn}}$ and $\mathtt{mod}(k - K_{\text{burn}}, K_{\text{gap}}) = 0$, where $\mathtt{mod}(\cdot, \cdot)$ represents modulus after division. A Metropolis-Hasting step can also be introduced here to remove the estimation bias. We summarise the technique for drawing N_{live} live samples from the prior in Algorithm 3.

4.4 Proximal nested sampling algorithm

After embedding Algorithms 2 and 3 into Algorithm 1 (i.e., the standard nested sampling algorithm), we obtain our proposed proximal nested sampling algorithm, which is summarised in Algorithm 4. Recall that Algorithm 2 generates a new single sample from the prior subject to the hard likelihood constraint, which is used to replace the sample with the lowest likelihood value in the live sample set. We suggest using a sample randomly selected from the live sample set as a starting point for Algorithm 2.

Algorithm 3 Proximal algorithm of drawing live samples (from prior)

```
Initialization: $N_{\mathrm{live}}$, $K_{\mathrm{burn}}$, $K_{\mathrm{gap}}$, and $x^{(0)}$.
Output: $N_{\mathrm{live}}$ live samples $\{x_n\}_{n=1}^{N_{\mathrm{live}}}$ (draw from the prior with no constraint).
for $k = 1, \ldots, K_{\mathrm{burn}}$
  - Compute $x^{(k)}$ using the iterative formula (41).
end for
$n = 1$;
for $k = K_{\mathrm{burn}} + 1, \ldots, K_{\mathrm{burn}} + N_{\mathrm{live}} K_{\mathrm{gap}}$
  - Compute $x^{(k)}$ using the iterative formula (41).
  - Metropolis-Hasting step to remove the estimation bias.
  if $\mathrm{mod}(k - K_{\mathrm{burn}}, K_{\mathrm{gap}}) = 0$
    $x_n = x^{(k)}$; $n = n + 1$.
  end if
end for
```

So far we have presented the proximal nested sampling framework in its most general form for arbitrary log-concave distributions, which is based on the iterative formula (36) or (38) to sample from the constrained prior. These iterative formula involve computing proximal operators related to the log-prior and likelihood constraint, which we have not yet considered in further detail. In principle computing proximal operators involves solving a minimisation problem, although in many scenarios this can be solved analytically or otherwise efficient iterative algorithms can be used. In the following section we consider explicit forms of proximal nested sampling for common forms of the prior and likelihood, outlining explicitly how the required proximal operators can be computed.

Algorithm 4 Proximal nested sampling algorithm

```
Initialization: Data $Y$. Set $\mathcal{Z} = 0, \xi_0 = 1$ and $i = 0$. Using Algorithm 3 to draw $N_{\mathrm{live}}$ samples $\{x_n\}_{n=1}^{N_{\mathrm{live}}}$ from the prior distribution $\pi(x)$ in the prior space $\Omega$.
Output: Evidence $\mathcal{Z}$ and posterior probabilities $\{p_i\}$.
for $i = 1, \ldots,$ until the stopping criterion reached
  - Find the lowest likelihood, say $L_i$, in the set of live samples.
  - Compute weight $w_i = (\xi_{i-1} - \xi_{i+1})/2$, where $\xi_i = \exp(-i/N_{\mathrm{live}})$.
  - Update evidence by $\mathcal{Z} = \mathcal{Z} + L_i w_i$.
  - Randomly select a sample, say $x^{(0)}$, from the set of live samples.
  - Use Algorithm 2 to draw a new sample $x_{\mathrm{new}} = \mathtt{ProxSampleDraw}(x^{(0)}, L_i)$ from the prior distribution $\pi(x)$ in the restricted parameter space $\Omega_{L_i}$, and replace the individual sample $x_{i,\mathrm{low}}$ by the newly drawn sample $x_{\mathrm{new}}$.
end for
Update the evidence by $\mathcal{Z} = \mathcal{Z} + \sum_{n=1}^{N_{\mathrm{live}}} \mathcal{L}(x_n) w_{i+1} / N_{\mathrm{live}}$.
Compute the posterior probability for each individual sample $p_i = L_i w_i / \mathcal{Z}$.
```

Before concluding this section, we note that the proposed proximal nested sampling method summarised in Algorithm 4 seeks to provide a Bayesian model selection strategy that is computationally efficient, simple, robust, and easy to deploy, as opposed to a strategy that seeks to deliver optimal performance by using adaptive methods or by leveraging model-specific properties. For example, for some models with favourable factorisation properties, better results would be obtained by replacing ULA by a Gibbs sampler (see e.g. Lucka 2016). Similarly, for models that are close to isotropic, one could replace ULA with a proximal Markov kernel derived from the underdamped Langevin SDE, which includes a Hamiltonian term (see e.g. Melidonis et al. 2022[3]).

Such methods scale more efficiently to large models than the overdamped Langevin method used in this paper, but they are less robust to anisotropy, which is a common feature in Bayesian inverse problems. Moreover, one could also consider using an adaptive MALA kernel with a matrix-valued step-size taking into account second-order properties of the posterior distribution (Pereyra et al. 2016). Lastly, because the proposed proximal nested sampling method has been specifically designed for large models that are log-concave, it is not equipped with mechanisms to handle multi-modality. For problems involving multi-modality, we would recommend modifying the Markov kernel either by using some form of annealing (Neal 2001), or by using an adaptive importance sampling scheme (Martino et al. 2017). However, as mentioned previously, performing model selection for models that are both large and multi-modal is very difficult and remains an important perspective for future work.

5 Explicit forms of proximal nested sampling

In the general proximal nested sampling framework presented in Sect. 4 we considered arbitrary log-concave terms for the prior and likelihood and did not consider further how to compute the proximal operators related to those terms. We now exemplify our proposed proximal nested sampling framework with explicit forms for common priors and likelihoods used in high-dimensional signal and image processing problems. In particular, we outline explicitly how to compute the required proximal operators.

For illustration, we focus on sparsity-promoting priors corresponding to $f(x) = \mu \|\boldsymbol{\Psi}^{\dagger} x\|_1$, where $\boldsymbol{\Psi}^{\dagger} \in \mathbb{C}^{p \times d}$ represents a sparsifying transform, and Gaussian likelihoods corresponding to $g(x) = \|y - \boldsymbol{\Phi} x\|_2^2 / 2\sigma^2$, where $y \in \mathbb{C}^m$ denotes measured data, $x \in \mathbb{R}^d$ the underlying parameters, and $\boldsymbol{\Phi} \in \mathbb{C}^{m \times d}$ the measurement operator (model), although other common priors are also considered. For simplicity,

[3] Note that we focus on proximal MCMC kernels since purely gradient-based MCMC methods based on the Langevin or Hamiltonian dynamics are not directly applicable to the non-smooth models considered in this paper. They might fail to be geometrically ergodic, in which case the nested sampling scheme would also behave poorly (see Betancourt 2011 for an example of a nested sampling method based on Hamiltonian dynamics).

although not essential, we assume $\mathbf{\Psi}$ is an orthonormal transformation, i.e., $\mathbf{\Psi}^\dagger\mathbf{\Psi} = \mathbf{\Psi}\mathbf{\Psi}^\dagger = I$.

From the iterative forms given in (36), (38) and (41), on which our proximal nested sampling framework is based, it is necessary to compute two proximal operators: $\mathrm{prox}_f^\lambda(x)$ and $\mathrm{prox}_{\chi_{\mathcal{B}_\tau}}(x)$, related to the prior and likelihood, respectively (recall that the definition of $\chi_{\mathcal{B}_\tau}$ is related to likelihood function g; see (30)). In the following we calculate these two proximal operators for explicit expressions of $f(x)$ and $g(x)$ and show the corresponding explicit forms of the iterative formulas of (36), (38) and (41).

5.1 Proximal operator for the prior

When $f(x)$ represents a flat prior or $f(x) = \mu\|\mathbf{\Psi}^\dagger x\|_2^2$ (Gaussian prior) it is differentiable with gradient

$$\nabla f(x) = 0 \quad \text{or} \quad \nabla f(x) = 2\mu\mathbf{\Psi}\mathbf{\Psi}^\dagger x = 2\mu x, \tag{42}$$

respectively (here we use $\mathbf{\Psi}\mathbf{\Psi}^\dagger = I$). Obviously, there is no need to use the Moreau-Yosida envelope $\nabla f^\lambda(x)$ to approximate $\nabla f(x)$ when $f(x)$ is differentiable.

When $f(x)$ represents a sparsity-promoting Laplacian-type prior $f(x) = \mu\|\mathbf{\Psi}^\dagger x\|_1, \forall x' \in \mathbb{R}^d$, we have

$$\begin{aligned}\mathrm{prox}_f^\lambda(x') &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\mu\|\mathbf{\Psi}^\dagger x\|_1 + \|x - x'\|_2^2/2\lambda\right\}\\ &= x' + \mathbf{\Psi}\left(\mathrm{prox}_{\|\cdot\|_1}^{\lambda\mu}(\mathbf{\Psi}^\dagger x') - \mathbf{\Psi}^\dagger x'\right)\\ &= x' + \mathbf{\Psi}\left(\mathrm{soft}_{\lambda\mu}(\mathbf{\Psi}^\dagger x') - \mathbf{\Psi}^\dagger x'\right),\end{aligned} \tag{43}$$

where the second line follows by standard properties of the proximal operator (Combettes and Pesquet 2011) and where $\mathrm{soft}_\lambda(x) = (\mathrm{soft}_\lambda(x_1), \mathrm{soft}_\lambda(x_2), \cdots)$ is the soft-thresholding operator defined by

$$\mathrm{soft}_\lambda(x_i) = \begin{cases}0, & |x_i| < \lambda,\\ x_i(|x_i| - \lambda)/|x_i|, & \text{otherwise.}\end{cases} \tag{44}$$

5.2 Proximal operator for the likelihood

Consider the Gaussian likelihood corresponding to $g(x) = \|y - \mathbf{\Phi}x\|_2^2/2\sigma^2$. Recall that $\chi_{\mathcal{B}_\tau}(x) = 0$ if $x \in \{x \mid g(x) < \tau\}$ and otherwise $\chi_{\mathcal{B}_\tau}(x) = +\infty$. We are to solve

$$\begin{aligned}\mathrm{prox}_{\chi_{\mathcal{B}_\tau}}^\lambda(x') &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\chi_{\mathcal{B}_\tau}(x) + \|x - x'\|_2^2/2\lambda\right\}\\ &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\chi_{\mathcal{B}_\tau}(x) + \|x - x'\|_2^2\right\}\\ &= \mathrm{proj}_{\chi_{\mathcal{B}_\tau}}(x'),\end{aligned} \tag{45}$$

which is a projection onto set $\mathcal{B}_\tau$.

For the case where the measurement operator is the identity, $\mathbf{\Phi} = I$, (e.g. denoising problems) then problem (45) is the projection onto the ℓ_2 ball with radius $\sqrt{2\tau\sigma^2}$. In this case the proximal (projection) operator has closed-form solution

$$\mathrm{proj}_{\chi_{\mathcal{B}_\tau}}(x) = \begin{cases}x, & \text{if } x \in \mathcal{B}_\tau,\\ \frac{x-y}{\|x-y\|_2}\sqrt{2\tau\sigma^2} + y, & \text{otherwise.}\end{cases} \tag{46}$$

For the case where the measurement operator is not the identity, $\Phi \neq I$, problem (45) is equivalent to finding an $x \in \mathbb{R}^d$ satisfying

$$\min_{x\in\mathbb{R}^d}\left\{\chi_{\mathcal{B}'_{\tau'}}(u) + \|x - x'\|_2^2/2\right\}, \quad \text{s.t.} \ \ u = \mathbf{\Phi}x, \tag{47}$$

where $\mathcal{B}'_\tau := \{z \mid \|y - z\|_2^2 < \tau\}$ and $\tau' = 2\tau\sigma^2$. Minimisation problem (47) can be solved by a variety of different optimisation methods, e.g. by the alternating direction method of multipliers (ADMM) and primal-dual algorithms (see, e.g., Parikh and Boyd 2013 and references therein for further details). In the following we present detailed procedures for using the ADMM and primal-dual algorithms to solve problem (47).

5.2.1 Computation using ADMM method

Firstly, the augmented Lagrangian of the minimisation problem (47) can be represented as

$$\begin{aligned}\Lambda(x, u, z) := \chi_{\mathcal{B}'_{\tau'}}(u) + \frac{1}{2}\|x - x'\|_2^2 + \beta z^\dagger(u - \mathbf{\Phi}x)\\ + \frac{\beta}{2}\|u - \mathbf{\Phi}x\|_2^2,\end{aligned} \tag{48}$$

for dual variable z and penalty parameter $\beta > 0$. Starting from an initialisation $x^{(0)}, z^{(0)}$, the augmented Lagrangian of (48) can be minimised with respect to variables u and x alternatively, while updating the dual value z using the dual ascent method to ensure the constraint $u = \mathbf{\Phi}x$ is satisfied for the final solution, i.e.

$$u^{(i)} = \operatorname*{argmin}_{u\in\mathbb{C}^m}\Lambda(x^{(i)}, u, z^{(i)}), \tag{49}$$

$$x^{(i+1)} = \operatorname*{argmin}_{x\in\mathbb{R}^d}\Lambda(x, u^{(i)}, z^{(i)}), \tag{50}$$

$$z^{(i+1)} = z^{(i)} + u^{(i)} - \mathbf{\Phi}x^{(i+1)}, \tag{51}$$

which can be rewritten as the following explicit iterative scheme

$$u^{(i)} = \operatorname*{argmin}_{u \in \mathbb{C}^M} \left\{ \chi_{\mathcal{B}'_{\tau'}}(u) + \frac{\beta}{2} \|u - \mathbf{\Phi} x^{(i)} + z^{(i)}\|_2^2 \right\}, \quad (52)$$

$$x^{(i+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^D} \left\{ \frac{1}{2} \|x - x'\|_2^2 + \frac{\beta}{2} \|u^{(i)} - \mathbf{\Phi} x + z^{(i)}\|_2^2 \right\}, \quad (53)$$

$$z^{(i+1)} = z^{(i)} + u^{(i)} - \mathbf{\Phi} x^{(i+1)}. \quad (54)$$

The solution to problem (52) has a closed-form expression since it is the projection onto a scaled and shifted ℓ_2 ball, i.e.,

$$u^{(i)} = \begin{cases} \mathbf{\Phi} x^{(i)} - z^{(i)}, & \text{if } \mathbf{\Phi} x^{(i)} - z^{(i)} \in \mathcal{B}'_{\tau'}, \\ \frac{\mathbf{\Phi} x^{(i)} - z^{(i)} - Y}{\|\mathbf{\Phi} x^{(i)} - z^{(i)} - Y\|_2} \sqrt{2\tau\sigma^2} + Y, & \text{otherwise.} \end{cases} \quad (55)$$

Problem (53) is differentiable and so can be solved by gradient descent. It is straightforward to show that this problem is equivalent to solving the linear system w.r.t. x

$$(\beta \mathbf{\Phi}^\dagger \mathbf{\Phi} + I) x = x' + \beta \mathbf{\Phi}^\dagger (u^{(i)} + z^{(i)}), \quad (56)$$

which can be solved by using iterative methods, with $(\beta \mathbf{\Phi}^\dagger \mathbf{\Phi} + I)$ positive definite.

The pseudo code to compute the proximal operator, $\text{prox}_{\chi_{\mathcal{B}_\tau}}(x)$, using ADMM is summarised in Algorithm 5. Various stopping criteria can be considered, such as a maximum iteration number or the relative error of solutions at two consecutive iterations, i.e., $\|x^{(i+1)} - x^{(i)}\|_2 / \|x^{(i)}\|_2$.

Algorithm 5 ADMM for proximal operator associated with the likelihood

```
Initialization: $x^{(0)}, z^{(0)}$.
Input: $x, L^*$
Output: $x^*$ (the value of $\text{prox}_{\chi_{\mathcal{B}_\tau}}(x)$).
Compute $\tau = -\log L^*$, and form $\chi_{\mathcal{B}_\tau}$.
for $i = 0, \ldots,$ until the stopping criterion reached
    - Compute $u^{(i)}$ by (55);
    - Compute $x^{(i+1)}$ by solving (56);
    - Update $z^{(i+1)}$ by (54).
end for
Set $x^* = x^{(i+1)}$.
```

5.2.2 Computation using primal-dual method

Alternatively, problem (45) can be solved using a primal-dual method. Note that the problem can be rewritten as

$$\min_{x \in \mathbb{R}^d} \left\{ \chi_{\mathcal{B}'_{\tau'}}(\mathbf{\Phi} x) + \|x - x'\|_2^2 / 2 \right\}, \quad (57)$$

which is equivalent to the saddle-point problem

$$\min_{x \in \mathbb{R}^d} \max_{z \in \mathbb{C}^K} \left\{ z^\dagger \mathbf{\Phi} x - \chi^*_{\mathcal{B}'_{\tau'}}(z) + \|x - x'\|_2^2 / 2 \right\}, \quad (58)$$

where $\chi^*_{\mathcal{B}'_{\tau'}}$ is the convex conjugate of $\chi_{\mathcal{B}'_{\tau'}}$. The saddle-point problem (58) can be solved by alternatively optimising with respect to the primal variable x and the dual variable z. Considering a proximal forward-background step for each alternate optimisation, first for the dual variable z followed by the primal variable x, leads to the following iterative scheme

$$z^{(i+1)} = \text{prox}_{\chi^*_{\mathcal{B}'_{\tau'}}}(z^{(i)} + \delta_1 \mathbf{\Phi} \bar{x}^{(i)}), \quad (59)$$

$$x^{(i+1)} = \text{prox}_h(x^{(i)} - \delta_2 \mathbf{\Phi}^\dagger z^{(i+1)}), \quad (60)$$

$$\bar{x}^{(i+1)} = x^{(i+1)} + \delta_3 (x^{(i+1)} - x^{(i)}), \quad (61)$$

where $h(x) = \|x - x'\|_2^2 / 2$, and δ_k, for $k = 1, 2, 3$, are algorithm step size parameters. We next consider how to solve problem (59) and (60) explicitly.

Problem (59) can be solved by

$$\begin{aligned} z^{(i+1)} &= \text{prox}_{\chi^*_{\mathcal{B}'_{\tau'}}}(z^{(i)} + \delta_1 \mathbf{\Phi} \bar{x}^{(i)}) \\ &= z^{(i)} + \delta_1 \mathbf{\Phi} \bar{x}^{(i)} - \text{prox}_{\chi_{\mathcal{B}'_{\tau'}}}(z^{(i)} + \delta_1 \mathbf{\Phi} \bar{x}^{(i)}), \end{aligned} \quad (62)$$

where we have noted the relationship between the proximal operator of the convex conjugate of a function given by (9). Since $\mathcal{B}'_{\tau'}$ is an ℓ_2 ball, the proximal operator in (62) has the closed-form expression

$$\text{prox}_{\chi_{\mathcal{B}'_{\tau'}}}(z) = \text{proj}_{\mathcal{B}'_{\tau'}}(z) = \begin{cases} z, & \text{if } z \in \mathcal{B}'_{\tau'}, \\ \frac{z - y}{\|z - y\|_2} \sqrt{2\tau\sigma^2} + y, & \text{otherwise.} \end{cases} \quad (63)$$

Problem (60) is to solve

$$x^{(i+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \|x - x'\|_2^2 + \|x - (x^{(i)} - \delta_2 \mathbf{\Phi}^\dagger z^{(i+1)})\|_2^2 \right\}, \quad (64)$$

which involves a differentiable objective function and so can be solved analytically, yielding the closed-form solution

$$x^{(i+1)} = (x' + x^{(i)} - \delta_2 \mathbf{\Phi}^\dagger z^{(i+1)}) / 2. \quad (65)$$

The pseudo code to compute the proximal operator, $\text{prox}^\lambda_{\chi_{\mathcal{B}_\tau}}(x)$, using the primal-dual method is summarised in Algorithm 6. The same stopping criterion as for ADMM in Algorithm 5 can also be used for Algorithm 6.

Note that the main difference between the primal-dual method and ADMM is that the primal-dual method does not need to solve the linear system in (56). Therefore, the primal-dual method is typically more efficient computationally and is the approach used in the numerical experiments that follow. However, there are specific problems for which the linear system in (56) admits a computationally efficient solution and where the ADMM method might be more appropriate.

Algorithm 6 Primal-dual method for proximal operator associated with the likelihood

```
**Initialization:** $x^{(0)}, \bar{x}^{(0)}, z^{(0)}$.
**Input:** $x, L^*$
**Output:** $x^*$ (the value of $\text{prox}_{\chi_{\mathcal{B}_\tau}}(x)$).
Compute $\tau = -\log L^*$, and form $\chi_{\mathcal{B}_\tau}$.
for $i = 0, \ldots$, until the stopping criterion reached
    - Compute $z^{(i+1)}$ by (62);
    - Compute $x^{(i+1)}$ by solving (65);
    - Update $\bar{x}^{(i+1)}$ by (61).
end for
Set $x^* = x^{(i+1)}$.
```

5.3 Explicit iterative formula for drawing samples

We are now in a position to outline the explicit iterative formulas to draw samples for a variety of common priors using our proximal nested sampling method.

The explicit representations of the iterative equations (36) (for differentiable $f(x)$) and (38) (for non-differentiable $f(x)$), which are used in Algorithm 2 to draw an individual sample from the prior under the hard likelihood constraint, for uniform, Gaussian and Laplacian priors, i.e. $f(x)$ constant, $f(x) = \mu\|\boldsymbol{\Psi}^\dagger x\|_2^2$ and $f(x) = \mu\|\boldsymbol{\Psi}^\dagger x\|_1$, respectively, are

$$x^{(k+1)} = (1 - \frac{\delta}{2\lambda})x^{(k)} + \frac{\delta}{2\lambda}x^{*(k)} + \sqrt{\delta}w^{(k+1)}, \quad (66)$$

$$x^{(k+1)} = (1 - \frac{\delta}{2\lambda} - \delta\mu)x^{(k)} + \frac{\delta}{2\lambda}x^{*(k)} + \sqrt{\delta}w^{(k+1)}, \quad (67)$$

$$x^{(k+1)} = (1 - \frac{\delta}{2\lambda})x^{(k)} + \frac{\delta}{2\lambda}\boldsymbol{\Psi}\big(\text{soft}_{\lambda\mu}(\boldsymbol{\Psi}^\dagger x^{(k)}) - \boldsymbol{\Psi}^\dagger x^{(k)}\big) + \frac{\delta}{2\lambda}x^{*(k)} + \sqrt{\delta}w^{(k+1)}, \quad (68)$$

where $x^{*(k)} = \text{prox}_{\chi_{\mathcal{B}_\tau}}(x^{(k)})$ is obtained using Algorithm 5 or 6.

Correspondingly, the explicit representations of equation (41), which is used in Algorithm 3 to draw N_{live} initial live samples from the prior distribution $\pi(x)$ in the prior space Ω, are, respectively,

$$x^{(k+1)} = x^{(k)} + \sqrt{\delta}w^{(k+1)}, \quad (69)$$

$$x^{(k+1)} = (1 - \delta\mu)x^{(k)} + \sqrt{\delta}w^{(k+1)}, \quad (70)$$

$$x^{(k+1)} = x^{(k)} + \frac{\delta}{2\lambda}\boldsymbol{\Psi}\big(\text{soft}_{\lambda\mu}(\boldsymbol{\Psi}^\dagger x^{(k)}) - \boldsymbol{\Psi}^\dagger x^{(k)}\big) + \sqrt{\delta}w^{(k+1)}. \quad (71)$$

We conclude this section with a brief discussion of the types of priors that the proposed proximal nested sampling method supports. While any prior that is log-concave could be addressed by using proximal nested sampling, we only recommend using the method for priors with proximal operators that are easy to evaluate or to approximate numerically. This is the case for many models used in applied high-dimensional statistics, where inference is often conducted by using convex optimisation algorithms that also require computing proximal operators. For more details about how to evaluate proximal operators, their properties, and lists of functions with known mappings please see Bauschke and Combettes (2011), Combettes and Pesquet (2011) and Parikh and Boyd (2013, Ch. 6). A library with MATLAB implementations of frequently used proximity mappings is also available online[4].

Moreover, since the proposed proximal nested sampling approach was specifically designed for models that are log-concave and with Bayesian imaging applications in mind, we anticipate that it will be mostly used with informative priors designed to regularise and stabilise high-dimensional estimation problems. As explained in Llorente et al. (2022), the marginal likelihood can be very sensitive to the choice of the prior. Therefore, it is important that the parameters of the prior are chosen carefully. In particular, we expect that proximal nested sampling will be used in combination with empirical Bayesian strategies that automatically adjust the parameters of the prior by maximum marginal likelihood estimation (see e.g. Vidal et al. 2020).

Furthermore, high-dimensional Bayesian models that are log-concave often result from a careful trade-off between modelling accuracy and computational tractability, and thus they are inherently misspecified (e.g., in the case of Bayesian imaging applications, one would not expect the prior to define a realistic generative model). Consequently, when using proximal nested sampling in this context one is inherently operating in an $\mathcal{M}$-open Bayesian modelling paradigm, where none of the models under consideration are formally "true". We refer the reader to Llorente et al. (2022) for more details about performing model selection in this context, as well as for details about prior sensitivity, objectivity, and the use of data-driven priors in Bayesian model selection.

[4] https://github.com/cvxgrp/proximal.

6 Numerical experiments

In this section we validate our proposed proximal nested sampling method and demonstrate its utility on a range of illustrative problems.

We first validate our method on a problem with a Gaussian likelihood and Gaussian prior where the value of the marginal likelihood (Bayesian evidence) can be computed analytically. The dimensions of the problem considered range from low to very high, i.e. 2 to 10^6 dimensions.

Following on from this, we demonstrate the effectiveness of the proximal nested sampling method by applying it to two canonical imaging inverse problems, namely image denoising and image reconstruction. In particular, we demonstrate the use of proximal nested sampling for the principled Bayesian model selection of the sparsifying dictionary, the regularisation parameter (i.e. the μ parameter of the prior) and the appropriate measurement operator when it may be misspecified. Furthermore, as mentioned already, as a by-product the samples obtained by nested sampling approaches can also be used to perform posterior inferences. This is critical in imaging problems in order to recover point estimates, e.g. restored images. Moreover, alternative forms of uncertainty quantification can also be considered from other posterior inferences, e.g. variance estimates and posterior credible regions (see, e.g., Cai et al. 2018).

6.1 Implementation and computational resources

To perform the numerical experiments presented subsequently, the proximal nested sampling algorithms developed in this article were implemented in MATLAB.[5] The numerical experiments performed to compute the marginal likelihood for low-dimensional problems (i.e., dimensions less than 200) were run on a Macbook laptop with an i7 Intel CPU and memory of 16 GB. A high-performance workstation, with 24 CPU cores, x86 64 architecture and 256 GB memory, was used for high-dimensional problems.

6.2 Validation in high dimensions

We first consider the validation of the proximal nested sampling method. For ease of validation, we consider the prior potential $f(x) = \mu\|\boldsymbol{\Psi}^\dagger x\|_2^2$, with $\mu = 1/2$, $\boldsymbol{\Psi} = I$, and the likelihood potential $g(x) = \|y - \boldsymbol{\Phi}x\|_2^2/2\sigma^2$, with $\sigma = 1$, $\Phi = I$. For this setting, we have a closed-form solution of the marginal likelihood value (see Appendix for further details).

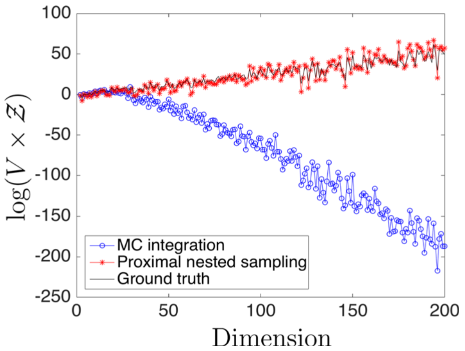


Fig. 1 Validation of our proximal nested sampling technique (for dimensions 2–200) to compute the marginal likelihood (Bayesian evidence) for a scenario where a closed-form solution is accessible. The logarithm of the unnormalised prior volume (V) times the marginal likelihood value ($\mathcal{Z}$) is plotted against the dimensions of the problem considered. The blue-circle line, red-asterisk line and the black-solid line show the results of MC integration, proximal nested sampling and the ground truth, respectively. We can clearly see that the results computed by proximal nested sampling agree with the ground truth well, whereas the result computed by MC integration with 10^5 samples can only achieve acceptable results when the dimension is below ~ 20. The computation time for the problem with dimension 200 is approximately one minute.

Test data $y \in \mathbb{R}^d$ are generated by

$$y = x + w, \tag{72}$$

where x is an d-dimensional vector of uniformly distributed random numbers in $[0, 1]^d$, and w is an d-dimensional vector of normally distributed random numbers. Note that the underlying model used to generate the mock data does not match the prior π used here, but that is fine for validation of the calculation of the marginal likelihood. Also, in imaging setting the prior is never perfectly specified. In the following, we consider increasing dimensions from $d = 2$ to $d = 10^6$. We separate the test into three parts: i) small models of dimension from $d = 2$ to $d = 200$, ii) moderately large models of dimension from $d = 2$ to $d = 10^5$, and iii) high dimensional models with $d = 10^6$.

We first test our method for low-dimensional models (i.e., $d < 200$). For our proximal nested sampling method, we use $N_{\text{live}} = 2 \times 10^2$ live samples and $N = 3 \times 10^3$ dead samples, with a thinning factor of 10. We also compare our result with vanilla Monte Carlo (MC) integration where a uniform prior with integrand $f \cdot g$ is utilised, with the number of samples set to 10^5. Figure 1 presents the results. Our proximal nested sampling method agrees well with the ground truth, whereas simple MC integration can only achieve acceptable results when the dimension is small, say $d < 20$. The computation time for the problem with dimension 200 is approximately one minute.

We now test our proximal nested sampling method for high-dimensional cases. Results for dimensions of y up to

[5] A Python version of the `proxnest` code implementing the proximal nested sampling framework proposed in this article has since been developed and is available at https://github.com/astro-informatics/proxnest.

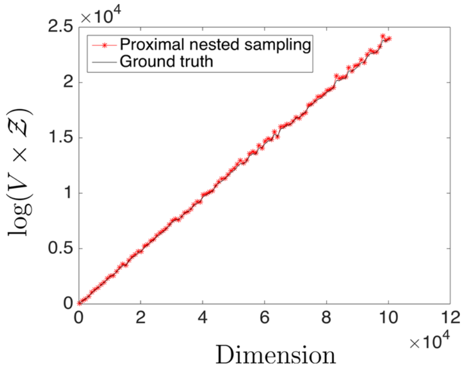


Fig. 2 Validation of our proximal nested sampling technique (for dimensions up to 10^5) to compute the marginal likelihood (Bayesian evidence) for a scenario where a closed-form solution is accessible. The logarithm of the unnormalised prior volume (V) times the marginal likelihood value ($\mathcal{Z}$) is plotted against the dimensions of the problem considered. The red-asterisk line and the black-solid line show the results of proximal nested sampling and the ground truth, respectively. We can clearly see that the results computed by proximal nested sampling agrees with the ground truth well. The computation time for the problem with dimension 10^5 is approximately 10 minutes

10^5 are given in Fig. 2, where we set the number of the live samples $N_{\text{live}} = 10^3$ and the number of dead samples $N = 10^4$, with thinning factor 10 (we do not consider direct MC integration any further since it is already shown to fail for dimensions above ~ 20). These results again show that our proximal nested sampling method can achieve results in close agreement with the ground truth. The computation time for the problem with dimension 10^5 is approximately 10 minutes.

Finally, we consider dimension 10^6 as an example to show that our proximal nested sampling method can be pushed to dimensions much higher than 10^5. With the same parameters as that used for dimension 10^5, ten runs were performed for a 10^6 dimensional setting of the same problem. The logarithm of the ground truth value was calculated to be 2.3850×10^5. The mean of ten runs of proximal nested sampling was computed be to 2.3851×10^5, with standard deviation 0.0002×10^5. The result computed by proximal nested sampling is in excellent agreement with the ground truth. The computation time for each run of the problem with dimension 10^6 is approximately 30 minutes.

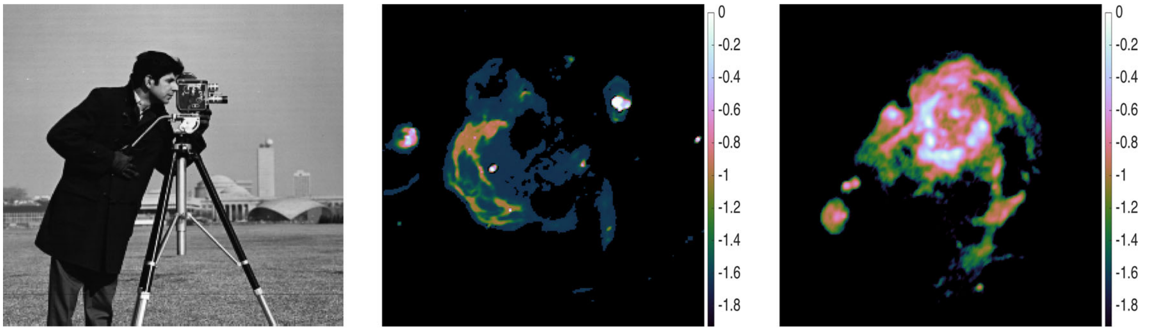
6.3 Model selection in image processing

We now illustrate the application of proximal nested sampling for Bayesian model selection in imaging problems. In particular, we focus on two canonical problems, image denoising and image reconstruction, with different likelihoods and priors. We emphasise that Bayesian model selection for these imaging problems is not well addressed by existing techniques due to the high dimensions considered (i.e., higher than 10^5) and the use of general log-concave priors (e.g., like the sparsity promoting Laplace-type priors that include ℓ_1 terms).

The three images in Fig. 3 are used in the experiments that follow: Cameraman image, the W28 supernova remnant, and the HI region of the M31 galaxy, all with size of 256×256 pixels and with intensities in the range $[0, 255]$. Sparsity-promoting priors (which are not smooth) and Gaussian likelihoods are consider in the following experiments, formed as $f(x) = \mu\|\boldsymbol{\Psi}^\dagger x\|_1$ and $g(x) = \|y - \boldsymbol{\Phi}x\|_2^2/2\sigma^2$, respectively, where μ, $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ are set to different forms for model selection purposes.

6.3.1 Prior model selection in image denoising: dictionary selection

For a standard denoising problem we apply proximal nested sampling to select the dictionary $\boldsymbol{\Psi}$ used for the sparsifying transform. The noisy image y is generated by $y = x + w$, where x is the ground truth clean image and w



(a) Cameraman (b) W28 (c) M31

Fig. 3 Images used to showcase the use of proximal nested sampling for Bayesian model selection in high-dimensional image processing problems. Panel (**a**): Cameraman grey-scale image; Panels (**b**)–(**c**): W28 and M31 radio galaxies normalised to [0, 1] and then shown in `log10` scale (i.e. the numeric labels on the colour bar are the logarithms of the image intensity), respectively

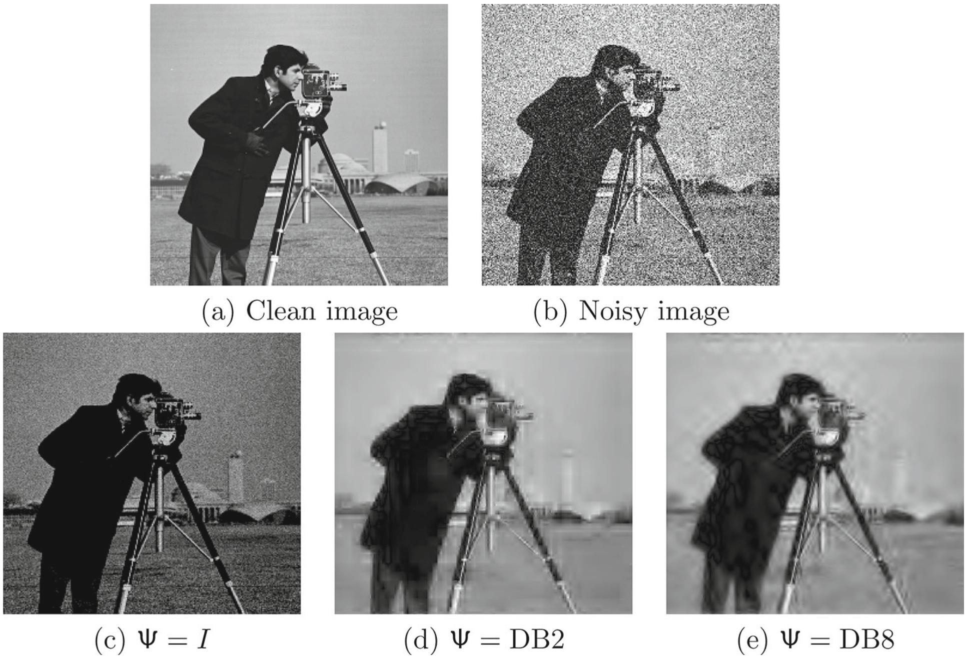


Fig. 4 Dictionary selection for an image denoising problem solved by proximal nested sampling (test image is cameraman). First row shows the clean image and noisy image. Second row shows the posterior mean images recovered by proximal nested sampling for priors with (sparsifying) transforms $\mathbf{\Psi} = I$, DB2 and DB8, respectively, where the log-prior reads $f(x) = \mu\|\mathbf{\Psi}^\dagger x\|_1$. By eye, both DB2 and DB8 wavelets provide superior reconstruction fidelity compared to $\mathbf{\Psi} = I$. The model $\mathbf{\Psi} = \text{DB2}$ may also be judged to provide slightly superior performance to $\mathbf{\Psi} = \text{DB8}$

Table 1 Marginal likelihood (Bayesian evidence) values computed by proximal nested sampling for Bayesian model selection of the sparsifying dictionary for an image denoising problem (see Fig. 4 for corresponding reconstructed images)

Prior	$\log \mathcal{Z}$	RMSE
$\mathbf{\Psi} = I$	$-6.54 \times 10^4 \pm 0.08$	41.07
$\mathbf{\Psi} = \text{DB2}$	$-3.06 \times 10^4 \pm 0.09$	14.29
$\mathbf{\Psi} = \text{DB8}$	$-3.09 \times 10^4 \pm 0.09$	14.51

Sparsity-promoting (non-differentiable) priors are considered with (sparsifying) transforms $\mathbf{\Psi} = I$, DB2 and DB8. Comparing models, Bayesian model selection afforded by proximal nested sampling suggests the model with the DB2 dictionary is superior, followed by DB8, both of which are far superior to the case where $\mathbf{\Psi} = I$, which agrees with the RMSE (root mean square error) values and assessment performed by eye, which require the ground truth to be known

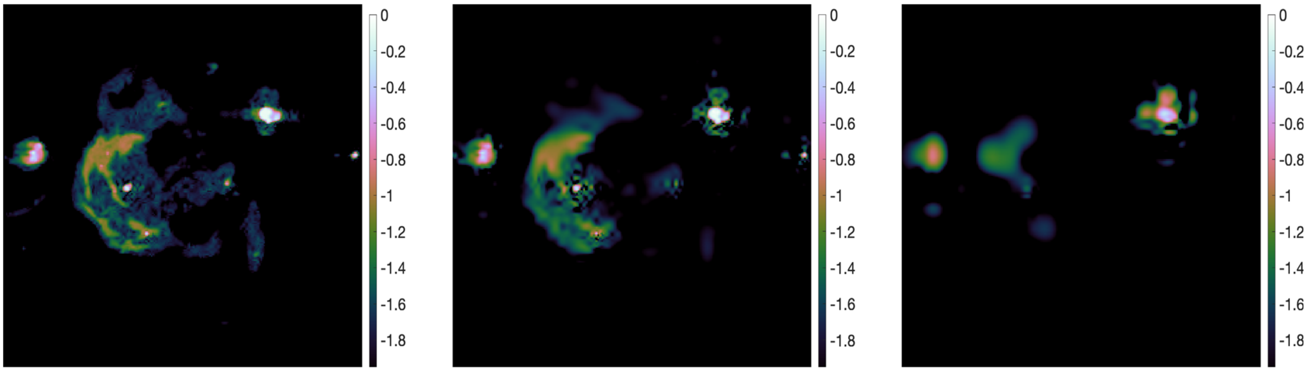
is Gaussian noise with zero mean and standard deviation $\sigma = \|x\|_\infty 10^{-\text{SNR}/20}$, where $\|\cdot\|_\infty$ is the infinity norm, and the input signal-to-noise ratio (SNR) is set to 20. Set $\mathbf{\Phi} = I$ in the likelihood $g(x) = \|y - \mathbf{\Phi}x\|_2^2/2\sigma^2$ (i.e., $g(x) = \|y - x\|_2^2/2\sigma^2$). We then investigate the influence of different choices for $\mathbf{\Psi}$ in the prior term $f(x) = \mu\|\mathbf{\Psi}^\dagger x\|_1$, with $\mu = 10^5$. Specifically, three forms of $\mathbf{\Psi}$ are considered, namely the identity (I), Daubechies 2 wavelets (DB2), and Daubechies 8 wavelets (DB8). For the proximal nested sampling method, the number of the live samples N_{live} and dead samples N is respectively set to 2×10^3, and 4×10^4 with thinning factor 10^2, which is sufficient to ensure convergence.

Figure 4 presents the posterior means recovered (i.e. the reconstructed images) for the three dictionaries considered, i.e. for $\mathbf{\Psi} = \{I, \text{DB2}, \text{DB8}\}$. It is clear that the reconstructed images corresponding to $\mathbf{\Psi} = \text{DB2}$ and DB8 are significantly better than that for $\mathbf{\Psi} = I$. Moreover, while the difference between the reconstructed images of the models for $\mathbf{\Psi} = \text{DB2}$ and DB8 is small, by eye the image recovered with DB2 may be judged slightly superior.

Table 1 presents the calculated marginal likelihood values[6] for the different sparsifying transforms $\mathbf{\Psi}$ selected for the prior. The root mean square error (RMSE) is also given, where the RMSE gauges the difference between the posterior mean image and the ground truth image. Note that the RMSE cannot normally be computed in practical problems since the ground truth is not known. Since for these experiments we know the ground truth the RMSE is a useful measure for comparison purposes.

Table 1 shows that the model with $\mathbf{\Psi} = I$ possesses the smallest marginal likelihood value. This implies that for this denoising problem the model with $\mathbf{\Psi} = I$ is inferior to models where $\mathbf{\Psi}$ is set to DB2 and DB8. Moreover, the marginal likelihood difference between models where $\mathbf{\Psi}$ is set to DB2 or DB8 is not dramatic, nevertheless this implies that DB2 is preferred. These finding inferred by Bayesian model selection agree with the RSME values computed for each model, where the model with $\mathbf{\Phi} = \text{DB2}$ is slightly preferred over DB8, and both models with DB2 and DB8 are highly preferred over the model with $\mathbf{\Phi} = I$ (recall that in practice it is not possible to compute the RMSE since it requires knowledge of the underlying ground truth). Furthermore, the

[6] The value of the log marginal likelihoods computed is low (in other words, its absolute value is very high) since the problems we consider are extremely high-dimensional.



$\mu = 10^6$ $\mu = 10^7$ $\mu = 10^8$

Fig. 5 Regularisation parameter selection for an image reconstruction problem solved by proximal nested sampling (test image is W28 radio galaxy). Images from left to right are the posterior mean images recovered by proximal nested sampling for μ in the prior definition set to 10^6, 10^7 and 10^8, respectively. The data y are generated by measuring 30% of noisy Fourier coefficients of the test image. On close inspection it may be noticed that reconstruction for model with $\mu = 10^6$ is superior to the one with $\mu = 10^7$, which is superior to the one with $\mu = 10^8$

model preferences inferred by proximal nested sampling also agree with the assessment of reconstructed image quality by-eye discussed above. The results obtained are consistent with common knowledge that it is typically more effective to denoise a natural image using a prior that promotes sparsity in some (sparsifying) transform domain (e.g. a wavelet domain) rather than in the image domain itself. The computation time for the problem with $\boldsymbol{\Psi} = I$ is approximately 10 minutes, and for the problem with $\boldsymbol{\Psi}$ = DB2 or DB8 is approximately 60 minutes.

In high-dimensional settings note that Bayes factors can be very large due to the concentration of probability in high-dimensions, hence it is not meaningful to consider traditional scales for assessing model comparisons such as the Jeffery's scale (Nesseris and García-Bellido 2013). Instead, we recommend comparing marginal likelihood values directly.

Table 2 Marginal likelihood (Bayesian evidence) values computed by proximal nested sampling for Bayesian model selection of the regularisation parameter μ for an image reconstruction problem (see Fig. 5 for corresponding reconstructed images)

μ	$\log \mathcal{Z}$	RMSE
10^6	$-2.61 \times 10^4 \pm 0.09$	1.82
10^7	$-5.39 \times 10^4 \pm 0.09$	2.81
10^8	$-2.90 \times 10^5 \pm 0.09$	6.70

Prior definition with μ set to 10^6, 10^7 and 10^8, respectively, are considered. Comparing models, Bayesian model selection afforded by proximal nested sampling suggests the model with $\mu = 10^6$ is superior to the one with $\mu = 10^7$, which is superior to the one with $\mu = 10^8$, which agrees with the RMSE (root mean square error) values and assessment performed by eye, which require the ground truth to be known

6.3.2 Prior model selection in image reconstruction: regularisation parameter selection

We now apply proximal nested sampling to a standard reconstruction problem and, firstly, consider the selection of the regularisation parameter μ defining the width of the prior. It is typically very challenging to optimally set the regularisation parameter μ, which controls the strength of prior knowledge and plays a key role in reconstruction quality. Consider noisy observations (noisy measurements)

$$y = \boldsymbol{\Phi} x + w, \tag{73}$$

where w again denotes Gaussian noise with zero mean and $\sigma = \|x\|_\infty 10^{-\mathrm{SNR}/20}$ (standard deviation), with SNR set to 30, and m and d are respectively the dimension of y and image x. Consider the prior $f(x) = \mu \|\boldsymbol{\Psi}^\dagger x\|_1$, with $\boldsymbol{\Psi}$ = DB8, and likelihood $g(x) = \|y - \boldsymbol{\Phi} x\|_2^2 / 2\sigma^2$. For the reconstruction scenario, $\boldsymbol{\Phi}$ represents the sensing (measurement) operator. In particular, we consider a measurement model comprising incomplete Fourier measurements (common in radio interferometric and magnetic resonance imaging) defined by the sensing operator $\boldsymbol{\Phi} = \boldsymbol{M}\boldsymbol{F}$, constructed from the Fourier transform $\boldsymbol{F}$ followed by a selection mask $\boldsymbol{M}$ which is generated randomly through the variable density sampling profile (Puy et al. 2011). We consider the scenario where only 30% of Fourier coefficients are measured, i.e. $m = 0.3d$. Note that different forms of the mask $\boldsymbol{M}$ result in different sensing operators $\boldsymbol{\Phi}$.

Figure 5 presents the posterior means recovered by proximal nested sampling (i.e. the reconstructed images) for models with μ set to 10^6, 10^7 and 10^8. It is difficult to assess the effectiveness of different regularisation parameters by eye, but on close inspection it may be noticed that the model with $\mu = 10^6$ is superior to the one with $\mu = 10^7$, which is superior to the one with $\mu = 10^8$. The computation time for each problem is approximately 150 minutes.

Table 2 presents the marginal likelihood and RMSE values computed for the models with different regularisation parameters μ. The computed marginal likelihood for the model with $\mu = 10^6$ is larger that the value for the model with $\mu = 10^7$, which is larger than the model with $\mu = 10^8$, suggesting the model with $\mu = 10^6$ is preferred. The computed marginal likelihoods are consistent with the model preferences obtained by comparing the RMSE of each model and by visual inspection. Recall that both RMSE and visual comparisons can only be performed here where the ground truth is available and cannot be used for model comparison in practice. In summary, this example demonstrates that our proximal nested sampling method is capable of selecting superior regularisation parameters for models stemming from high-dimensional inverse problems.

6.3.3 Measurement model selection in image reconstruction

We now apply proximal nested sampling to the same reconstruction problem considered above (i.e. image reconstruction from noisy and incomplete Fourier measurements) but focus on the problem of misspecification of the measurement model $\boldsymbol{\Phi}$. Noisy observations Y are generated by (73), measuring 10% of Fourier coefficients, i.e. with $m = 0.1d$.

We use the ground truth model $\boldsymbol{M}_{\text{truth}}$ to simulate observation data y. However, when solving the resulting inverse problem we consider a number of different measurement models, not only the ground truth model $\boldsymbol{M}_{\text{truth}}$ but also misspecified models $\boldsymbol{M}_{\gamma}$, where $\gamma > 0$ encodes the level of misspecification.

The method by which the model is misspecified in motivated by radio interferometric imaging. In radio interferometry, the coordinates of the Fourier coefficients acquired by the telescope are measured in units of (radio) wavelength. If the wavelength at which observations are made is misspecified, the coordinates of the Fourier coefficients acquired will be scaled. We model precisely this type of misspecified model here to represent the case where the instrument wavelength is not calibrated accurately.

An incorrectly specified wavelength then simply acts to modify the mask of the ground truth measurement model $\boldsymbol{M}_{\text{truth}}$. The misspecified model corresponding to mask $\mathbf{M}_{\gamma}$, for misspecification parameter γ, is generated by extending every measured position in $\mathbf{M}_{\text{truth}}$ radially. Specifically, every measured position is extended radially along the line connecting it to the origin to a length of γd_j, $j \in \Omega_{\text{mask}}$, where γ is the misspecification scaling factor, d_j is the distance from the original measured position j to the origin in $\mathbf{M}_{\text{truth}}$, and Ω_{mask} is the set which contains all the measured positions. It is worth mentioning that the larger the scaling factor γ, the larger the distortion of $\boldsymbol{M}_{\gamma}$ from the ground truth $\boldsymbol{M}_{\text{truth}}$. Note also that $\gamma = 0$ corresponds to a correctly specified model, i.e. $\boldsymbol{M}_{\gamma=0} = \boldsymbol{M}_{\text{truth}}$.

For proximal nested sampling, the number of the live samples N_{live} and dead samples N is respectively set to 2×10^3 and 3×10^4 with thinning factor 10^2, which is sufficient to ensure convergence. Regularisation parameter $\mu = 10^8$ is used for these experiments.

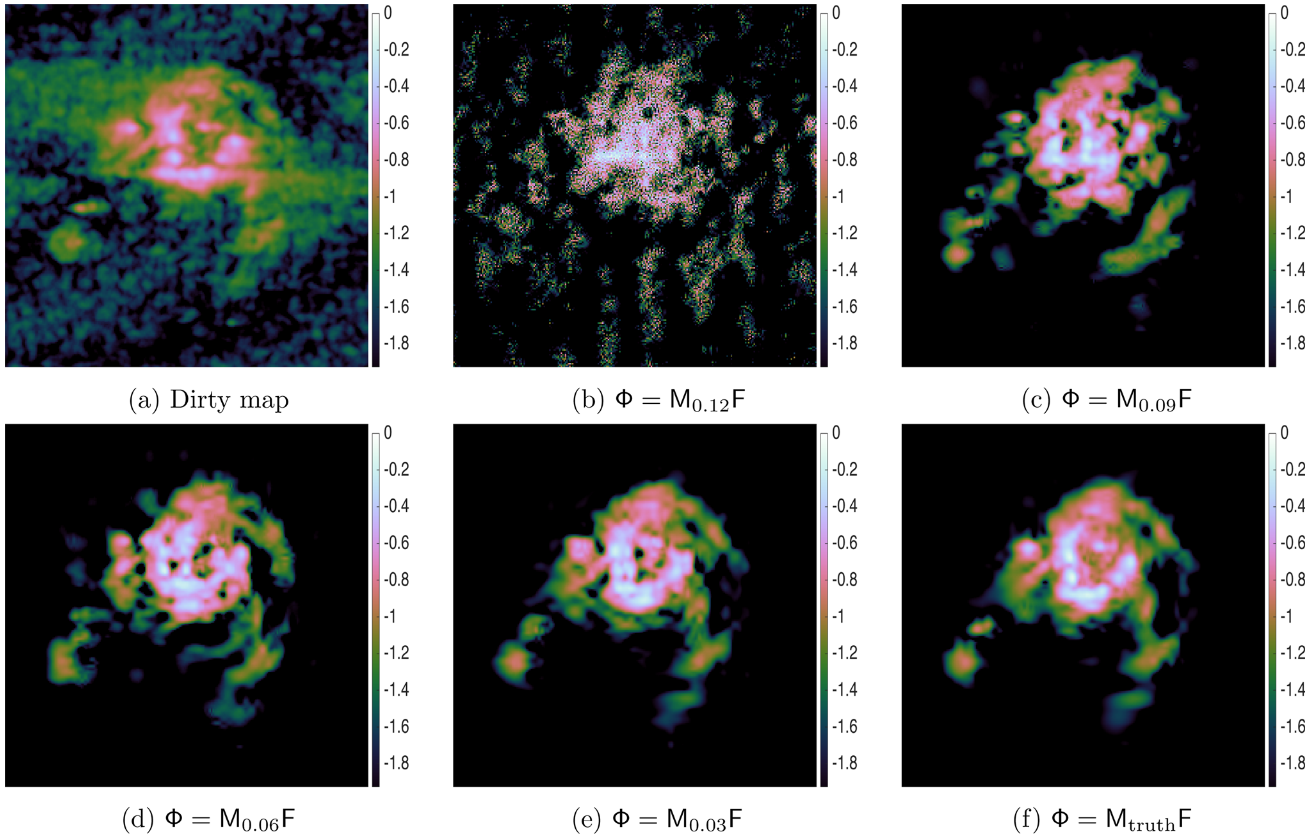
Figure 6 presents the posterior means recovered (i.e. the reconstruction images) for models with $\boldsymbol{\Phi}_{\gamma} = \mathbf{M}_{\gamma}\mathbf{F}$ and $\boldsymbol{\Phi} = \mathbf{M}_{\text{truth}}\mathbf{F}$,. Here misspecified models $\boldsymbol{M}_{0.12}$, $\boldsymbol{M}_{0.09}$, $\boldsymbol{M}_{0.06}$ and $\boldsymbol{M}_{0.03}$ are generated for misspecification scaling factors γ with values of 0.12, 0.09, 0.06 and 0.03, respectively. It is apparent by eye that the posterior mean image recovered with the ground truth model is the best and that the quality of the recovered posterior mean image degrades as the size of the misspecification scale parameter γ increases. The computation time for each problem is approximately 150 minutes.

Table 3 presents the marginal likelihood and RMSE values computed for the different models considered. The computed marginal likelihood is largest when the correct ground truth model is adopted in the likelihood. As the misspecification parameter γ is increased (corresponding to greater misspecification and less accurate models), the corresponding computed marginal likelihood values monotonically decrease (become more negative). For Bayesian model comparison, the model with the lowest misspecification parameter γ is always preferred. The computed marginal likelihoods are consistent with the model preferences obtained by comparing the RMSE of each model and by visual inspection (although recall that such tests cannot be used for model comparison in practice when the ground truth is not known).

7 Conclusions

Nested sampling provides an efficient computational framework to estimate the marginal likelihood (Bayesian evidence) for Bayesian model selection. It effectively re-parameterises the marginal likelihood into a one-dimensional integral of the likelihood with respect to the enclosed prior volume. The challenge of nested sampling is to sample from the prior distribution subject to a hard likelihood constraint. A variety of successful techniques have been developed to perform such sampling in low and moderate dimensional problems. However, existing approaches are not directly useful for imaging applications because they scale poorly to large problems and struggle to support models that are not smooth.

In this article we presented the proximal nested sampling method that is specifically designed for Bayesian models that are log-concave, potentially very high-dimensional ($d = 10^6$



(a) Dirty map

(b) $\Phi = \mathsf{M}_{0.12}\mathsf{F}$

(c) $\Phi = \mathsf{M}_{0.09}\mathsf{F}$

(d) $\Phi = \mathsf{M}_{0.06}\mathsf{F}$

(e) $\Phi = \mathsf{M}_{0.03}\mathsf{F}$

(f) $\Phi = \mathsf{M}_{\mathrm{truth}}\mathsf{F}$

Fig. 6 Measurement model misspecification for an image reconstruction problem solved by proximal nested sampling (test image is M31 radio galaxy). Panel (**a**): dirty (back-projected) image $\mathbf{\Phi}^{\dagger} Y$; Panels (**b**)–(**f**): posterior mean images recovered by proximal nested sampling for misspecified models $\boldsymbol{M}_{\gamma}$, where increasing $\gamma > 0$ corresponds to increasing levels of misspecification (and $\gamma = 0$ corresponds to the ground truth model). It is apparent by eye that the posterior mean image recovered with the ground truth model is the best and that the quality of the recovered posterior mean image degrades as the size of the misspecification scale parameter γ increases

Table 3 Marginal likelihood (Bayesian evidence values computed by proximal nested sampling for Bayesian model selection for measurement model misspecification for an image reconstruction problem (see Fig. 6 for corresponding reconstructed images)

Likelihood	$\log \mathcal{Z}$	RMSE
$\mathbf{\Phi} = \boldsymbol{M}_{\mathrm{truth}}\boldsymbol{F}$	$-4.47 \times 10^3 \pm 0.08$	3.40
$\mathbf{\Phi} = \boldsymbol{M}_{0.03}\boldsymbol{F}$	$-4.88 \times 10^3 \pm 0.08$	7.85
$\mathbf{\Phi} = \boldsymbol{M}_{0.06}\boldsymbol{F}$	$-5.63 \times 10^3 \pm 0.08$	12.01
$\mathbf{\Phi} = \boldsymbol{M}_{0.09}\boldsymbol{F}$	$-9.21 \times 10^3 \pm 0.07$	15.71
$\mathbf{\Phi} = \boldsymbol{M}_{0.12}\boldsymbol{F}$	$-1.44 \times 10^4 \pm 0.08$	18.08

Misspecified models are denoted $\boldsymbol{M}_{\gamma}$, where increasing $\gamma > 0$ corresponds to increasing levels of misspecification (and $\gamma = 0$ corresponds to the ground truth model). Comparing models, Bayesian model selection afforded by proximal nested sampling suggests the model with the lowest misspecification parameter γ is always preferred, which also agrees with the RMSE (root mean square error) values and assessment performed by eye, which require the ground truth to be known

and beyond), and potentially not smooth. This is achieved by exploiting tools from proximal calculus and Moreau-Yosida regularisation to efficiently sample from the prior subject to the hard likelihood constraint through a proximal MCMC approach. The resulting Markov chain iterations combine a gradient step that approximates a Langevin SDE that scales efficiently to large problems, with a projection term that acts to push the Markov chain back into the likelihood constraint set if it wanders outside of it, and a Metropolis-Hastings correction step to ensure the hard likelihood constraint is satisfied.

The proposed proximal nested sampling framework was implemented and validated on a Gaussian model for which the marginal likelihood could be calculated in closed-form, showing excellent agreement between values computed analytical and by proximal nested sampling, even in very high dimensions. The use of proximal nested sampling for principled Bayesian model selection was then showcased on a variety of imaging problems with non-smooth sparsity-promoting prior distributions. In particular, model selection problems were considered related to dictionary selection, and selection of the appropriate measurement model when it may be misspecified.

Proximal nested sampling allows Bayesian model selection to be performed at a much higher dimension than that was previously possible, while also supporting non-smooth priors that are widely used in imaging. It is our hope that proximal nested sampling will thus find widespread use for high-dimensional Bayesian model selection, particularly in the imaging sciences.

Important perspectives for future work include: a detailed theoretical analysis of the convergence properties of proximal nested sampling; an extension to (biased) accelerated proximal methods (Vargas et al. 2020); and an analysis of the properties of marginal maximum likelihood estimation for the class of models considered in this paper, such as estimator consistency for model selection in an $\mathcal{M}$-closed setting and concentration in an $\mathcal{M}$-open setting (Llorente et al. 2022). Moreover, it would be interesting to apply proximal nested sampling to other types of models, such as models with likelihood-based priors (Llorente et al. 2022), which can be handled straightforwardly by proximal nested sampling when the likelihood is log-concave. It would also be interesting to modify proximal nested sampling to tackle high-dimensional models that are multi-modal, particularly models with data-driven priors encoded by neural networks (see e.g. Mukherjee et al. 2022, Section 5).

Acknowledgements This work was supported by the Leverhulme Trust and by EPSRC grants EP/T007346/1 and EP/W007673/1. The authors would like to thank the editor and two anonymous reviewers for their valuable suggestions to improve the manuscript. The authors are also grateful to Abdul-Lateef Haji-Ali for helpful comments. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Appendix A

The volume of the prior $f(x) = \mu\|\mathbf{\Psi}^{\dagger}x\|_2^2$ with $\mathbf{\Psi} = I$ is

$$\begin{aligned} V &= \int_{-\infty}^{\infty} \exp\left(-\mu\|x\|_2^2\right) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}2\mu x^{\top}x\right) \mathrm{d}x \\ &=\sqrt{\frac{(2\pi)^d}{(2\mu)^d}}. \end{aligned} \tag{A1}$$

For the prior $f(x) = \mu\|\mathbf{\Psi}^{\dagger}x\|_2^2$ with $\mathbf{\Psi} = I$ and the likelihood $g(x) = \|y - \mathbf{\Phi}x\|_2^2/2\sigma^2$ with $\mathbf{\Phi} = I$, the Bayesian evidence value has the following closed-form representation:

$$\begin{aligned} &\frac{1}{V}\int_{-\infty}^{\infty} \exp(-\mu\|x\|_2^2)\exp(-\|y-x\|_2^2/2\sigma^2)\mathrm{d}x \\ &= \frac{1}{V}\int_{-\infty}^{\infty} \exp\left(-\mu\|x\|_2^2 - \|y-x\|_2^2/2\sigma^2\right)\mathrm{d}x \\ &= \frac{1}{V}\int_{-\infty}^{\infty} \exp\left(-(\mu+1/2\sigma^2)x^{\top}x + y^{\top}x/\sigma^2 - y^{\top}y/2\sigma^2\right)\mathrm{d}x \\ &= \frac{1}{V}\exp\left(-\frac{y^{\top}y}{2\sigma^2}\right)\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(2\mu+1/\sigma^2)x^{\top}x + y^{\top}x/\sigma^2\right)\mathrm{d}x \\ &= \frac{1}{V}\sqrt{\frac{(2\pi)^d}{(2\mu+1/\sigma^2)^d}}\exp\left(-\frac{y^{\top}y}{2\sigma^2}\right)\exp\left(\frac{1}{2}\frac{1}{2\mu+1/\sigma^2}\frac{y^{\top}y}{\sigma^4}\right), \end{aligned} \tag{A2}$$

whose logarithmic value is

$$\begin{aligned} &\log\sqrt{\frac{(2\pi)^d}{(2\mu+1/\sigma^2)^d}} + \left(-\frac{y^{\top}y}{2\sigma^2}\right) \\ &+ \left(\frac{1}{2}\frac{1}{2\mu+1/\sigma^2}\frac{y^{\top}y}{\sigma^4}\right) - \log V. \end{aligned} \tag{A3}$$

References

Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer-Verlag, New York (2011). https://link.springer.com/book/10.1007/978-1-4419-9467-7

Betancourt, M.: Nested sampling with constrained Hamiltonian Monte Carlo. AIP Conference Proceedings **1305**, 165 (2011). https://doi.org/10.1063/1.3573613

Brewer, B.J., Pártay, L.B., Csányi, G.: Diffusive nested sampling. Stat. Comput. **21**, 649–656 (2011)

Brosse, N., Durmus, A., Éric Moulines, et al.: Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In: Kale, S., Shamir, O. (eds) Proceedings of the 2017 Conference on Learning Theory, Proceedings of Machine Learning Research, vol 65. PMLR, Amsterdam, Netherlands, pp. 319–342 (2017)

Cai, X., Pereyra, M., McEwen, J.D.: Uncertainty quantification for radio interferometric imaging I: proximal-MCMC methods. Mon. Not. R. Astron. Soc. (MNRAS) **480**(3), 4154–4169 (2018)

Cai, X., Pratley, L., McEwen, J.D.: Online radio interferometric imaging: assimilating and discarding visibilities on arrival. Mon. Not. R. Astron. Soc. (MNRAS) **485**(4), 4559–4572 (2019)

Casella, G., Berger, R.L.: Statistical Inference. Duxbury - Thomson Learning, Boston (2002). https://books.google.co.uk/books/about/Statistical_Inference.html?id=ZpkPPwAACAAJ&redir_esc=y

Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**, 1313–1321 (1995)

Chib, S., Jeliazkov, I.: Marginal likelihood from the Metropolis-Hastings output. J. Am. Stat. Assoc. **96**, 270–281 (2001)

Chopin, N., Robert, C.P.: Properties of nested sampling. Biometrika **97**(3), 741–755 (2010)

Clyde, M.A., Berger, J.O., Bullard, F., et al.: Current challenges in Bayesian model choice. In: Statistical Challenges in Modern Astronomy IV ASP Conference Series, vol. **371**, pp. 224–240 (2007)

Combettes, P., Pesquet, J.C.: Proximal Splitting Methods in Signal Processing. Springer, New York (2011)

Durmus, A., Moulines, E., Pereyra, M.: Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. SIAM J. Imaging Sci. **1**(1), 473–506 (2018)

Feroz, F., Skilling, J.: Exploring multi-modal distributions with nested sampling. In: AIP Conference Proceedings, vol. **1553**,pp. 106–113 (2013)

Feroz, F., Hobson, M.P.: Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis. Mon. Not. R. Astron. Soc. (MNRAS) **384**(2), 449–463 (2008)

Feroz, F., Hobson, M.P., Bridges, M.: MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. Mon. Not. R. Astron. Soc. (MNRAS) **398**(4), 1601–1614 (2009)

Friel, N., Wyse, J.: Estimating the evidence - a review. Stat. Neerl. **66**(3), 288–308 (2012)

Green, P.J.: Reversible jump markov chain monte carlo computation and bayesian model determinatio. Biometrika **82**(4), 711–732 (1995)

Green, P.J., Łatuszyński, K., Pereyra, M., et al.: Bayesian computation: a summary of the current state, and samples backwards and forwards. Stat. Comput. **25**(4), 835–862 (2015)

Handley, W.J., Hobson, M.P., Lasenby, A.N.: POLYCHORD: nested sampling for cosmology. Mon. Not. R. Astron. Soc. Lett. **450**, L61–L65 (2015)

Harroue, B.: Approche bayésienne pour la sélection de modèles : application á la restauration d'image. PhD thesis, http://www.theses.fr/2020BORD0127 (2020)

Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New-York (2005)

Kamary, K., Mengersen, K., Robert, C.P., et al.: Testing hypotheses via a mixture estimation model. arXiv: 1412.2044 (2018)

Llorente, F., Martino, L., Curbelo, E., et al.: On the safe use of prior densities for Bayesian model selection. arXiv:2206.05210v1 (2022)

Llorente, F., Martino, L., Delgado, D., et al.: Marginal likelihood computation for model selection and hypothesis testing: an extensive review. arXiv: 2005.08334 (2020)

Lucka, F.: Fast gibbs sampling for high-dimensional bayesian inversion. Inverse Probl. **32**(11), 115019 (2016)

Lunz, S., Hauptmann, A., Tarvainen, T., et al.: On learned operator correction in inverse problems. SIAM J. Imaging Sci. **14**(1), 92–127 (2021)

Martino, L., Elvira, V., et al.: Layered adaptive importance sampling. Stat. Comput. **27**, 599–623 (2017)

McEwen, J.D., Wallis, C.G.R., Price, M.A. et al.: Machine learning assisted Bayesian model comparison: the learnt harmonic mean estimator. Stat. Comput. arXiv: 2111.12720 (2022)

Melidonis, S., Dobson, P., Altmann, Y., et al.: Efficient Bayesian computation for low-photon imaging problems. arXiv: 2206.05350 (2022)

Mukherjee, S., Hauptmann, A., Öktem, O., et al.: Learned reconstruction methods with convergence guarantees. arXiv: 2206.05431 (2022)

Mukherjee, P., Parkinson, D., Liddle, A.R.: A nested sampling algorithm for cosmological model selection. Astrophys. J. **638**, L51–L54 (2006)

Neal, R.: Annealed importance sampling. Stat. Comput. **11**, 125–139 (2001)

Nesseris, S., García-Bellido, J.: Is the Jeffreys' scale a reliable tool for Bayesian model comparison in cosmology? J. Cosmol. Astropart. Phys. **2013**, 036 (2013)

Newton, M.A., Raftery, A.E.: Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. **56**, 3–48 (1994)

O'Ruanaidh, J., Fitzgerald, W.J.: Numerical Bayesian Methods Applied to Signal Processing. Springer-Verlag, New York (1996)

Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**, 123–231 (2013)

Pereyra, M., McLaughlin, S.:Comparing bayesian models in the absence of ground truth. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 528–532 (2016)

Pereyra, M.: Proximal Markov chain Monte Carlo algorithms. Stat. Comput. **26**, 745–760 (2016)

Pereyra, M., Schniter, P., Chouzenoux, E., et al.: A survey of stochastic simulation and optimization methods in signal processing. IEEE J. Sel. Top. Signal Process. **10**(2), 224–241 (2016)

Puy, G., Vandergheynst, P., Wiaux, Y.: On variable density compressive sampling. IEEE Signal Process. Lett. **18**, 595–598 (2011)

Robert, C.P.: The Bayesian Choice. Springer-Verlag, New York (2007)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer-Verlag, New York (2004)

Sivia, D., Skilling, J.: Data Analysis: A Bayesian Tutorial. Oxford Science Publications, Oxford (2006)

Skilling, J.: Nested sampling for general Bayesian computation. Bayesian Anal. **1**, 833–859 (2006)

Tierney, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**, 82–86 (1986)

Trotta, R.: Applications of Bayesian model selection to cosmological parameters. Mon. Not. R. Astron. Soc. (MNRAS) **378**, 72–82 (2007)

Vargas, L., Pereyra, M., Zygalakis, K.C.: Accelerating proximal markov chain monte carlo by using an explicit stabilised method. SIAM J. Imaging Sci., in press, arXiv: 1908.08845 (2020)

Vidal, A.F., Bortoli, V.D., Pereyra, M., et al.: Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: an empirical bayesian approach part i: methodology and experiments. SIAM J. Imaging Sci. **13**(4), 1945–1989 (2020). https://doi.org/10.1137/20m1339829

Zhou, Q., Yu, T., Zhang, X., et al.: Bayesian inference and uncertainty quantification for medical image reconstruction with poisson data. SIAM J. Imaging Sci. **13**(1), 29–52 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.