A Bayesian interpretation of the L-curve

Jérôme Antoni, Jérôme Idier, Sébastien Bourguignon

A Bayesian interpretation of the L-curve

Jérôme Antoni[1], Jérôme Idier[2], and Sébastien Bourguignon[2]

[1]Univ Lyon, INSA Lyon, LVA, EA677, 69621 Villeurbanne, France
[2]LS2N, Ecole Centrale de Nantes, CNRS, 1 rue de la Noë, Nantes 44321, France

June 16, 2023

Abstract

The L-curve is a popular heuristic to tune Tikhonov regularization in linear inverse problems. This paper shows how it naturally arises when the problem is solved from a Bayesian perspective. Specifically, it establishes that the L-curve is a graphical way of searching for the maximum a posteriori solution after marginalization over the priors. The framework is general enough to enclose the classical (linear, logarithmic and square-root) L-curves as particular cases and to allow the design of new L-curves. It also explicitly accounts for the dimensions of the inverse problem (number of observations versus number of unknowns) in regularization. Elaborating on this framework, new criteria for locating the corner of the L-curve are discovered, such as the "minimum speed on the curve" and the "maximum angular speed", and conditions are established for their equivalence with the maximum curvature and the marginalized maximum a posteriori. All results are supported by numerical experiments. Experiments also show that the Bayesian L-curve rooted on appropriate priors can succeed in inverse problems where the classical L-curve is prone to fail.

1 Introduction

Inverse problems, as being devoted to discovering the origins of observed phenomena, are ubiquitous in science and engineering. This paper is concerned with discrete linear inverse problems, where the mapping between the causes and the observations is expressed by a matrix. In many instances, inverse problems are difficult to solve since the observations are only partially available, and *ad hoc* strategies such as regularization are required to enforce a unique solution. More specifically, of concern is the solution $\mathbf{X}$ to the set of linear equations

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_P] \in \mathbb{R}^{M\times P}$ is as matrix of P independent observations $\mathbf{y}_i$ of dimension M, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_P] \in \mathbb{R}^{K\times P}$ is a matrix of P explanatory variables $\mathbf{x}_i$ of dimension K, $\mathbf{N} \in \mathbb{R}^{M\times P}$ is a matrix that collects errors, and $\mathbf{A} \in \mathbb{R}^{M\times K}$ is a discrete operator. The problem is said ill-posed when a solution does not exist (e.g. $M > K$), the solution is not unique (e.g. $M < K$), or the operator $\mathbf{A}$ is badly-conditioned, i.e. the matrix $\mathbf{A}^\top\mathbf{A}$ (with $\mathbf{A}^\top$ the transpose of $\mathbf{A}$) has a large condition number, so that the presence of even small errors $\mathbf{N}$ cannot be neglected when attempting to recover an estimate of $\mathbf{X}$ from the observations $\mathbf{Y}$. The standard approach is to replace an ill-posed problem

by one less sensitive to errors, such as the Tikhonov-regularized problem [52, 14]:

$$\text{P1:} \quad \min_{\mathbf{X}} \left(T(\mathbf{X}) + \lambda \cdot U(\mathbf{X}) \right), \; 0 \leq \lambda < \infty \tag{2}$$

where

$$T(\mathbf{X}) = ||\mathbf{Y} - \mathbf{A}\mathbf{X}||_F^2 \quad \text{and} \quad U(\mathbf{X}) = ||\mathbf{X}||_F^2 \tag{3}$$

stand for the squared Frobenius norms of the residuals $\mathbf{Y} - \mathbf{A}\mathbf{X}$ and of the solution $\mathbf{X}$, respectively, T and U are two scalar "potential" functions of $\mathbf{X}$ from $\mathbb{R}^{K \times P}$ to $\mathbb{R}^+$, and $\lambda \in \mathbb{R}^+$ stands for the regularization parameter [1]. The regularized solution then takes the form

$$\mathbf{X}(\lambda) = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^\top \mathbf{Y}, \tag{4}$$

where the notation $\mathbf{X}(\lambda)$ makes it explicit that it is a function of the regularization parameter λ. A difficult question is then how to properly choose the value of the latter. Several strategies have been proposed to do so [14, 7], based either on prior knowledge of the norm of the errors $\mathbf{N}$, if available (e.g. the discrepancy principle [41]), or solely on the observations $\mathbf{Y}$ (e.g. cross-validation [50, 16]). The L-curve pertains to the latter category. It is a simple and elegant heuristic, initially introduced by Lawson and Hanson [33], thoroughly studied by Hansen (see e.g. [19, 22]), and widely used in practical applications [21].

The L-curve is a plot of the norm of the residuals with respect to the norm of the solution as a function of the regularization parameter. Specifically, upon introducing some monotonically increasing *scale function* f, it is the curve γ defined by the parametric equation

$$\gamma : \mathbb{R}^+ \to \mathbb{R}^2; \lambda \mapsto \gamma = \begin{cases} \zeta(\lambda) = f(T(\mathbf{X}(\lambda))) \\ \eta(\lambda) = f(U(\mathbf{X}(\lambda))). \end{cases} \tag{5}$$

with $\mathbf{X}(\lambda)$ as defined in Eq. (4).

It can be shown that γ is a decreasing curve and that any matrix $\mathbf{X}$ is mapped to a point $(f(T(\mathbf{X})); f(U(\mathbf{X})))$ in the region delimited above or on the curve [11]. In addition, the branch $\zeta(\lambda)$ of the curve increases with λ, while the branch $\eta(\lambda)$ decreases. This tends to convey an L-shape to the curve, where the horizontal branch of the "L" corresponds to a regime where the fitting error T dominates for large values of λ and where the vertical branch corresponds to a regime where the solution tends to become unstable – i.e. large values of its norm U – for small values of λ. As a consequence, the vertex of the "L" materializes a point of equilibrium between two antagonist regimes and a natural principle is to select the corner of the "L" as the location of the optimal value of the regularization parameter.

At this stage, two choices are left to the user when constructing and analyzing the L-curve: first, the selection of the scaling function f and, second, the criterion to locate the corner of the curve. It transpires from the literature that the choice of the scaling function is often made arbitrarily, even though Reginska early warned that such a choice strongly influences the tendency of the L-curve to bend around a corner (section 3 will introduce a principled way to choose f). Three typical examples of scale functions are $f = id$, $f = \sqrt{}$ and $f = \ln$, which produce the linear, square-root and logarithmic L-curves (the last two are hereafter denoted $\gamma_{\ln}$ and $\gamma_{\sqrt{}}$) [33, 45]. The logarithmic

[1] A more general form of problem P1 is $\hat{\mathbf{X}} = \text{Arg}\left(\min_{\mathbf{X}} T(\mathbf{X}) + \lambda \|\mathbf{L}\mathbf{X}\|_F^2\right)$, with $\mathbf{L}$ some well-conditioned matrix [19]. One can often recast the problem by a suitable change of variable so that $\mathbf{L} = \mathbf{I}$.

scaling is strongly advocated in Ref. [22] and is indeed most often encountered in the literature; it provides scale-invariance of the L-curve, meaning that the shape of the latter does not depend on the system of units of the measurements $\mathbf{Y}$ and of the variables $\mathbf{X}$. For this reason, it will be referred to from now on as the "reference L-curve". Strategies for the location of the corner of the L-curve have been discussed in the literature much more than the scaling function (see e.g. [14, 10, 29, 11, 13] and references therein). The reference algorithm is to search for the point of maximum curvature of γ [22], as measured by

$$\kappa = \frac{\zeta'\eta'' - \zeta''\eta'}{(\zeta'^2 + \eta'^2)^{3/2}}. \tag{6}$$

Various implementations and variants have been proposed based on this idea [10, 13, 11, 36, 12]. One rationale beyond the maximum curvature criterion is its invariance with respect to re-scaling of the regularization parameter λ. Interestingly, Ref. [2] proves that a criterion for selecting the regularization parameter is invariant under i) scale transformation of the data and ii) re-scaling of the regularization parameter if and only if it is a function of κ and ζ'/η' only, with $f = \ln$. However, the scale invariance of λ (ii) is a questionable property that will be reappraised in section 3 of this paper.

Although theoretical studies have provided insightful justifications of the L-curve in terms of its properties (e.g. [19, 22]), it fundamentally remains a heuristic criterion. Reginska attempted an algebraic explanation of the L-curve by recasting it into an optimization problem [45], where from the onset the optimal regularization parameter is searched as the minimum of the cost function

$$\Psi(\lambda) = T(\mathbf{X}(\lambda))U(\mathbf{X}(\lambda))^{\mu},\ 0 < \mu < \infty. \tag{7}$$

Reginska then proved that the minimum of $\Psi(\lambda)$ is attained where the L-curve is tangent with the line of slope $-1/\mu$. Figuring out a perfect L-shape curve $\gamma(\lambda)$ and an inclined straight line with (non-zero and finite), negative slope, the point of tangency must happen at the corner of the "L". Reginska's principle has been rediscovered several times in the literature, most often in the special case $\mu = 1$, the reason why it is referred to under different names such as the minimum product criterion [9, 29], the zero-crossing method [28, 27], the modified L-curve [34, 44], or multiplicative regularization [48, 15, 1, 49, 4, 5, 6]. Interestingly, with $\mu = 1$, the minimization of cost function (7) is equivalent to searching for the minimum of the L-curve rotated by $\pi/2$ (counterclockwise) in the (ζ, η) plane, as proved in [17].

The L-curve is an *ad hoc* strategy to tune Tikhonov regularization that proceeds from a purely deterministic approach. At first glance, it seems disconnected from other authoritative approaches used for regularization, such as the Bayesian framework. The aim of this paper is to bridge this gap and to provide a Bayesian interpretation of the L-curve. For this purpose, section 2 first establishes that the marginalized maximum a posteriori (MMAP) estimate of $\mathbf{X}$ leads to a Tikhonov-regularized solution. Section 3 then shows that the corresponding optimization problem is solved graphically by searching for the corner of a "Bayesian" L-curve. The conditions for the existence of the corner and the equivalence with the maximum curvature criterion are investigated in details. Based on these findings, section 3 proceeds with the introduction of new criteria for locating the corner of the L-curve. Next, section 4 investigates several particular cases of the L-curve induced by different priors used in the MMAP solution, thus offering an answer to the choice of the scaling function f. In particular, the gamma prior is found an important case, which produces the logarithmic scaling $f = \ln$ and provides a geometrical definition of the reference L-curve $\gamma_{\ln}$ in terms of the Legendre transform of the MMAP

cost function. Finally, section 5 discusses some possible extensions of the L-curve offered by the Bayesian framework, and section 6 illustrates many of the ideas by means of numerical experiments.

2 Bayesian interpretations of Tikhonov regularization

Probabilistic approaches, and in particular the Bayesian one, provide a different but nonetheless valuable treatment of regularization. Bayesian regularization has been the object of several seminal research works, such as [39, 35, 40, 47, 23], yet as far the authors know, no direct connection has ever been drawn with the L-curve. In this regard, the present section starts by resuming the Bayesian set-up, which naturally leads to Tikhonov regularization. The link with the L-curve will be further developed in section 3.

In the Bayesian setting, the explanatory variables $\mathbf{x}_i$ and the errors $\mathbf{n}_i$ are both seen as random vectors. Let us take them mutually independent and identically distributed, with zero-mean and diagonal covariance matrices $\tau_X^{-1}\mathbf{I}$ and $\tau_N^{-1}\mathbf{I}$ (this can always be forced by a suitable change of variables), where τ_X and τ_N stand for the precisions (the inverses of variances) of $\mathbf{x}_i$ and $\mathbf{n}_i$. In general, the precisions τ_N and τ_X are unknown, thus themselves considered as random variables, with prior probability densities ϕ_N and ϕ_X. Let $p(\mathbf{X}, \tau_X, \tau_N|\mathbf{Y})$ denote the joint probability density of $(\mathbf{X}, \tau_X, \tau_N)$ given the observations $\mathbf{Y}$ and take the Gaussians

$$p(\mathbf{Y}|\mathbf{X}, \tau_N) = \left(\frac{\tau_N}{2\pi}\right)^{\frac{MP}{2}} \exp\left(-\tau_N T(\mathbf{X})/2\right) \quad \text{and} \quad p(\mathbf{X}|\tau_X) = \left(\frac{\tau_X}{2\pi}\right)^{\frac{KP}{2}} \exp\left(-\tau_X U(\mathbf{X})/2\right) \tag{8}$$

for the probability density of $\mathbf{Y}|\mathbf{X}$ and the prior probability density of $\mathbf{X}$ (with a slight abuse of notation since $T(\mathbf{X})$ must also be a function of $\mathbf{Y}$ in the above equation). Strategies can now be worked out that lead to Tikhonov regularized solutions, i.e. solutions that pertain to the manifold

$$\mathcal{R}_\lambda = \{\mathbf{X} : \mathbf{X}(\lambda) = (\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\top\mathbf{Y}); \lambda \in \mathbb{R}^+\} \tag{9}$$

introduced by Eq. (4). Two of them are discussed in the next section. The first one is rooted on the joint maximum a posteriori (JMAP), $p(\mathbf{X}, \tau_X, \tau_N|\mathbf{Y})$, and, as already proved in the literature [24], establishes a connection between the Bayesian framework and Reginska's minimum product principle in a very special case. The second solution elaborates a new idea, which is the core of the present paper. It provides a geometrical interpretation of the L-curve as a graphical way to construct the Tikhonov-regularized solution that maximizes the marginalized posterior density $p(\mathbf{X}|\mathbf{Y})$.

2.1 JMAP estimate

One possibility is to search for the JMAP estimate of the set of variables $(\mathbf{X}, \tau_X, \tau_N)$. This was for instance explored in references [25, 26, 55, 24] and is briefly resumed here for the sake of completeness. The joint MAP estimate reads

$$\begin{aligned}
(\hat{\mathbf{X}}, \hat{\tau}_N, \hat{\tau}_X) &= \operatorname{Arg}\max_{\mathbf{X},\tau_N,\tau_X} p(\mathbf{X}, \tau_X, \tau_N|\mathbf{Y}) \\
&= \operatorname{Arg}\max_{\mathbf{X},\tau_N,\tau_X} p(\mathbf{Y}|\mathbf{X}, \tau_N)\phi_N(\tau_N)p(\mathbf{X}|\tau_X)\phi_X(\tau_X) \\
&= \operatorname{Arg}\min_{\mathbf{X},\tau_N,\tau_X} \left(\tau_N T(\mathbf{X}) + \tau_X U(\mathbf{X}) - MP\ln\tau_N - KP\ln\tau_X - 2\ln\phi_N(\tau_N) - 2\ln\phi_X(\tau_X)\right),
\end{aligned}$$

where the second line results from application of Bayes rule. Assuming that the prior probability densities ϕ_N and ϕ_X are differentiable, the stationary point of the above minimization problem readily gives $\hat{\mathbf{X}} \in \mathcal{R}_\lambda$ in the form of a Tikhonov-regularized solution, with regularization parameter

$$\lambda = \frac{\hat{\tau}_X}{\hat{\tau}_N}, \tag{10}$$

where $\hat{\tau}_X$ and $\hat{\tau}_N$ are solutions of

$$\begin{cases} T(\mathbf{X}(\lambda)) &= MP/\hat{\tau}_N + 2\phi'_N(\hat{\tau}_N)/\phi_N(\hat{\tau}_N) \\ U(\mathbf{X}(\lambda)) &= KP/\hat{\tau}_X + 2\phi'_X(\hat{\tau}_X)/\phi_X(\hat{\tau}_X). \end{cases} \tag{11}$$

To see where the JMAP solution is located on an L-curve, let us *arbitrarily* define the latter by Eq. (5) with the logarithmic scaling $f = \ln$, i.e. with $\zeta(\lambda) = \ln T(\mathbf{X}(\lambda))$ and $\eta(\lambda) = \ln U(\mathbf{X}(\lambda))$. As a general property, it holds that the potential functions follow the differential equation $T' = -\lambda U'$ where the prime denotes the derivative with respect to λ (see property (ii) of Proposition A.2). Therefore, one has $\eta'/\zeta' = (U'/U)/(T'/T) = -(T/U)/\lambda$. Next, replacing T and U by their expressions given by Eq. (11), the regularization parameter $\lambda(\hat{\tau}_X, \hat{\tau}_N) = \hat{\tau}_X/\hat{\tau}_N$ is found where the L-curve has the tangent

$$\frac{\eta'(\lambda(\hat{\tau}_X, \hat{\tau}_N))}{\zeta'(\lambda(\hat{\tau}_X, \hat{\tau}_N))} = -\frac{MP + 2\hat{\tau}_N\phi'_N(\hat{\tau}_N)/\phi_N(\hat{\tau}_N)}{KP + 2\hat{\tau}_X\phi'_X(\hat{\tau}_X)/\phi_X(\hat{\tau}_X)}. \tag{12}$$

In general, this is an implicit equation in $\hat{\tau}_X$ and $\hat{\tau}_N$. A particularly simple case is when the right-hand side is a constant, such that the equation depends on the ratio $\lambda = \hat{\tau}_X/\hat{\tau}_N$ only. This happens when $\phi_X(t) \propto t^{-\alpha_X}$ and $\phi_N(t) \propto t^{-\alpha_N}$, with $\alpha_X, \alpha_N \geq 0$, which are recognized as "improper priors"[2], thus yielding

$$\frac{\eta'(\lambda)}{\zeta'(\lambda)} = -\frac{MP - 2\alpha_N}{KP - 2\alpha_X} \tag{13}$$

(Ito et al. further made the choice $\alpha_X = 0$ in Ref. [24]). Therefore, the JMAP solution ultimately corresponds to the point in the plane (ζ, η) where the L-curve $\gamma_{\ln}$ is tangent with the straight line of slope $-(MP - 2\alpha_N)/(KP - 2\alpha_X)$. From an algebraic point of view, Eq. (13) is also recognized as the optimality condition $\Psi' = 0$ of Reginska's criterion (7) with $\mu = (KP - 2\alpha_X)/(MP - 2\alpha_N)$. In conclusion, for a particular choice of the scaling function and with improper priors on the precisions, the JMAP estimate of $\mathbf{X}$ is equivalent to Tikhonov regularization, with the regularization parameter selected on the L-curve from the minimum product principle. This provides an indirect connection between the Bayesian framework and the L-curve, yet in the limits of a very special case.

Another strategy that yields a more direct and general, yet unexplored connection, is now investigated.

2.2 MMAP estimate

The idea is to treat the unknown precisions τ_X and τ_N as nuisance parameters and to marginalize them, i.e. to integrate them out of $p(\mathbf{X}, \tau_X, \tau_N|\mathbf{Y})$. Hence, the optimization problem of interest

[2] In the Bayesian framework, improper priors are non-integrable probability densities, yet accepted as long as they return valid posterior probability densities [46].

becomes

$$\text{P2:} \quad \hat{\mathbf{X}} = \operatorname{Arg}\max_{\mathbf{X}} p(\mathbf{X}|\mathbf{Y}) = \operatorname{Arg}\max_{\mathbf{X}} p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) \tag{14}$$
$$= \operatorname{Arg}\max_{\mathbf{X}} \int_{\mathbb{R}^+} p(\mathbf{Y}|\mathbf{X},\tau)\phi_N(\tau)d\tau \int_{\mathbb{R}^+} p(\mathbf{X}|\tau)\phi_X(\tau)d\tau.$$

There is in general no closed-form solution to problem P2[3]. However, several analytical results can be obtained when $\mathbf{X}$ and $\mathbf{N}$ are distributed according to a generalized Gaussian. From now on, the Gaussian case will be considered only, as it contains the essence of the idea (the generalized Gaussian case will be briefly addressed in subsection 5.1). In this regard, a few preliminary results must first be introduced. Let us define the Φ-transform of a function ϕ:

$$\Phi(s;n) = \int_{\mathbb{R}^+} e^{-s\tau}\tau^n\phi(\tau)d\tau, \quad n \geq 0, \quad s \in \mathbb{R}^+. \tag{15}$$

Note that $\Phi(0;n)$ reduces to the Mellin transform of $\phi(t)$, while $\Phi(s;0)$ is the Laplace transform of $\phi(t)$ on the real positive line; for n an integer, $\Phi(s;n)$ is $(-1)^n$ times the n-th derivative of the Laplace transform evaluated on the real positive line. The domain of definition of the Φ-transform considered in this work is $s \in]0;\infty[$, which holds for all causal probability density functions $\phi \in \mathcal{L}_1(\mathbb{R}^+)$ and also for improper probability density functions that decrease no faster than τ^{-n}, for instance $\phi(\tau) \propto \tau^{-\alpha}$ with $\alpha \leq n$. The Φ-transform has properties that will turn out useful in what follows:

Proposition 2.1.

1. $\left(-\frac{d}{ds}\right)^k \Phi(s;n) = \Phi(s;n+k)$, $n \geq 0$, $k \in \mathbb{N}$,
2. *If* $\phi(\tau) \geq 0$, *then* $\Phi(s;n) \geq 0$.

Let us now denote $\Phi_N(s;n)$ and $\Phi_X(s;n)$ the Φ-transforms of the prior densities of the precisions, $\phi_N(t)$ and $\phi_X(t)$, respectively. The following result then holds.

Proposition 2.2. *The solution to problem P2 is returned by*

$$\hat{\mathbf{X}} = \operatorname{Arg}\max_{\mathbf{X}} \Phi_N(T(\mathbf{X})/2; MP/2)\Phi_X(U(\mathbf{X})/2; KP/2) \tag{16a}$$
$$= \left(\mathbf{A}^\top\mathbf{A} + \lambda(\hat{\mathbf{X}})\mathbf{I}\right)^{-1}\mathbf{A}^\top\mathbf{Y} \tag{16b}$$

where

$$\lambda(\hat{\mathbf{X}}) = \frac{\Phi_N(T(\hat{\mathbf{X}})/2; MP/2+1)}{\Phi_N(T(\hat{\mathbf{X}})/2; MP/2)} \frac{\Phi_X(U(\hat{\mathbf{X}})/2; KP/2)}{\Phi_X(U(\hat{\mathbf{X}})/2; KP/2+1)} \in \mathbb{R}^+. \tag{17}$$

This last quantity is to be interpreted as the ratio of the conditional means of the precision τ_X and τ_N given $\hat{\mathbf{X}}$ (actually a "noise-to-signal'" ratio), i.e.

$$\lambda(\hat{\mathbf{X}}) = \frac{\mathbb{E}\{\tau_X|U(\hat{\mathbf{X}})\}}{\mathbb{E}\{\tau_N|T(\hat{\mathbf{X}})\}}. \tag{18}$$

[3]It is noteworthy that an approximation to problem P2 is given by the "empirical" Bayesian method. It consists of expressing $p(\mathbf{X}|\mathbf{Y}) = \iint p(\mathbf{X}|\mathbf{Y},\tau_X,\tau_N)p(\tau_X,\tau_N|\mathbf{Y})d\tau_X d\tau_N$, where the approximation $p(\tau_X,\tau_N|\mathbf{Y}) \simeq \delta(\tau_X-\tau_X^\circ)\delta(\tau_N-\tau_N^\circ)$ is made, with τ_X° and τ_N° the MAP estimates in $p(\tau_X,\tau_N|\mathbf{Y})$. Therefore, $p(\mathbf{X}|\mathbf{Y}) \simeq p(\mathbf{X}|\mathbf{Y},\tau_X^\circ,\tau_N^\circ) \propto p(\mathbf{Y}|\mathbf{X},\tau_N^\circ)p(\mathbf{X}|\tau_X^\circ)$, which yields a solution $\mathbf{X}(\lambda) \in \mathcal{R}_\lambda$ with $\lambda = \tau_X^\circ/\tau_N^\circ$. This was for instance investigated in Refs. [3, 43].

It is emphasized that Eq. (16b) is an implicit equation in $\hat{\mathbf{X}}$, yet the important finding is that its solution once again takes the form of Tikhonov regularization. Therefore, the solution to problem P2 is the same as that of problem

$$\text{P3:} \begin{cases} \min_\lambda \left(-\ln \Phi_N(T(\mathbf{X}(\lambda))/2; MP/2) - \ln \Phi_X(U(\mathbf{X}(\lambda))/2; KP/2)\right) \\ \text{s.t.} \quad \mathbf{X}(\lambda) \in \mathcal{R}_\lambda. \end{cases} \tag{19}$$

Similarly, Eq. (18) is an implicit equation in λ, so that the same solution also solves

$$\text{P3':} \begin{cases} \lambda = \mathbb{E}\{\tau_X | U(\mathbf{X}(\lambda))\} / \mathbb{E}\{\tau_N | T(\mathbf{X}(\lambda))\} \\ \text{s.t.} \quad \mathbf{X}(\lambda) \in \mathcal{R}_\lambda. \end{cases} \tag{20}$$

This last equation is particularly insightful: similar to the JMAP estimate discussed in section 2.1 (see Eq. (10)), it highlights the interpretation of the regularization parameter as a noise-to-signal ratio, here defined as the ratio of the expected values of the precisions τ_X and τ_N conditional to the potentials U and T. From now on, the next section will analyze in details how problem P3 is related to the L-curve.

3 Equivalence with the L-curve

It is proved in this section that the L-curve defined by Eq. (5) with the general setting $f = -\ln \Phi$ offers a graphical way to find the MMAP solution.

3.1 L-curve associated with MMAP

3.1.1 Construction of a Bayesian L-curve

For reasons to become clear later, a reparamatrization of the problem is first made. The regularization parameter λ is seen as a bijective, positive, monotonically increasing and differentiable function

$$\lambda : \mathbb{R} \to \mathbb{R}^+; t \mapsto \lambda(t),\ \lambda' \geq 0 \tag{21}$$

of a *curve parameter* t. The inverse function, $t(\lambda) : \lambda \mapsto t$, will be referred to as the *regularization scale*. For instance, it is often found convenient to vary λ on a logarithmic scale, which means that $\lambda(t) = e^t$. Coming back to Eq. (19), the optimal regularization parameter is then returned for that value of $\lambda^* = \lambda(t^*)$ where t^* minimizes the cost function

$$J_1(t) = \zeta(t) + \eta(t) \tag{22}$$

with

$$\begin{cases} \zeta(t) = -\ln \Phi_N(T(\mathbf{X}(\lambda(t)))/2; MP/2) \\ \eta(t) = -\ln \Phi_X(U(\mathbf{X}(\lambda(t)))/2; KP/2), \quad \mathbf{X}(\lambda) \in \mathcal{R}_\lambda. \end{cases} \tag{23}$$

Let us now introduce a candidate for the L-curve. Consider the plane $\mathcal{P}$ parameterized by the set of coordinates $(\zeta; \eta) \doteq (-\ln \Phi_N(T/2; MP/2); -\ln \Phi_X(U/2; KP/2))$ and spanned by all possible

values of the potential functions U and T. Thus, the parametric equation

$$\gamma : \mathbb{R}^+ \to \mathbb{R}^2; t \mapsto \gamma(t) = \begin{pmatrix} \zeta(t) \\ \eta(t) \end{pmatrix}, \tag{24}$$

with $\zeta(t)$ and $\eta(t)$ as given in (23), defines a smooth curve in $\mathcal{P}$ with curve parameter t.

That γ actually defines an L-curve is proved by the following proposition.

Proposition 3.1. *The branches ζ and η are monotonically increasing and decreasing functions of t, respectively, i.e. $\zeta'(t) \geq 0$ and $\eta'(t) \leq 0$. Therefore, γ is a monotonically decreasing curve in the plane $\mathcal{P}$.*

Proof. According to property (i) of Proposition A.1, $\ln \Phi_N(T/2; MP/2)$ and $\ln \Phi_X(U/2; KP/2)$ are differentiable functions of T and U with negative slopes. Besides, it was assumed that $\lambda' \geq 0$. Using the law of composition of functions and property (i) of Proposition A.2 then proves that $\zeta' \geq 0$ and $\eta' \leq 0$. Therefore, $d\eta/d\zeta \leq 0$, and γ is monotonically decreasing. □

It is remarked in light of definition (5) that Eqs. (23) and (24) offspring a large family of L-curves rooted on the Φ-transform and, at the same time, offer a principled way to choose the scaling function $f = -\ln \Phi$ through the selection of the prior distribution of the precisions. This differs from the JMAP solution of section 2.1, which was limited to the arbitrary choice $f = \ln$. How particular cases, such as $f = id$, $f = \sqrt{}$ and $f = \ln$, can be recovered from $-\ln \Phi$, will be discussed in section 4.

3.1.2 Recovery of the reference L-curve

At this stage, there are at least two ways to recover the reference L-curve $\gamma_{\ln}$ from the Bayesian framework.

One is to search for the limit of the Bayesian L-curve when the amount of data P grows to infinity. As $n \to \infty$, the Φ-transform behave as $\Phi(s; n) \to n!\phi(n/s)/s^{n+1}$ (see Appendix C). According to Eq. (23), this defines the L-curve with branches

$$\gamma_{P\to\infty} : t \mapsto \begin{cases} \zeta(t) = -\ln T(\mathbf{X}(\lambda(t))) + c_N \\ \eta(t) = -\ln U(\mathbf{X}(\lambda(t))) + c_X, \end{cases} \tag{25}$$

where $c_N = \ln \phi_X(MP/T) + \ln(MP/2)!$ and $c_X = \phi_X(KP/U) + \ln(KP/2)!$ are two additive constants that can be ignored as they correspond to an inconsequential translation of the curve in plane $\mathcal{P}$ – or equivalently to an arbitrary choice of the origin of the reference frame (ζ, η). This comes with the regularization parameter

$$\lambda(\mathbf{X}) = \frac{T(\mathbf{X})}{U(\mathbf{X})} \frac{KP + 2}{MP + 2}, \tag{26}$$

which involves the ratio of $T(\mathbf{X})/(MP + 2)$ and $U(\mathbf{X})/(KP + 2)$, two asymptotically consistent estimators of the variance of $\mathbf{N}$ and of $\mathbf{X}$, respectively.

It is expected that all Bayesian L-curves will tend to these asymptotic results when the quantity of data becomes large, provided that the probability densities ϕ_N and ϕ_X are continuous and do not depend on P. Clearly, the same results hold when the prior probability densities vanish – i.e. they tend to flat functions – given a fixed value of P.

Another way to recover the reference L-curve $\gamma_{\ln}$ is from the Bayesian L-curve designed with a Jeffrey's prior on the precisions, $\phi(\tau) \propto \tau^{-1}$. As discussed in the Bayesian literature, this is a

non-informative and improper prior, yet its Φ-transform exists. According to Eq. (15), $\Phi(s;n) \propto (n-1)!s^{-n}$, and thus $-\ln\Phi(s;n) = n\ln s - \ln(n-1)!$; the L-curve defined by Eqs. (5) and (23) then becomes

$$\gamma_{\text{Jeffrey}} : t \mapsto \begin{cases} \zeta(t) = (MP/2)\ln T(\mathbf{X}(\lambda(t))) + C_N \\ \eta(t) = (KP/2)\ln U(\mathbf{X}(\lambda(t))) + C_X, \end{cases} \tag{27}$$

with $C_N = -\ln(MP/2-1)!$ and $C_X = -\ln(KP/2-1)!$ two additive constants. When $M = P$, the scale factor $MP/2 = KP/2$ corresponds to an homothety of the L-curve, which can be removed together with the constants to recover the reference L-curve $\gamma_{\ln}$. (The justification for the presence of two different scaling factors $MP/2$ and $KP/2$ on the branches ζ and η when $M \neq P$ will be discussed in section 4.)

The associated optimal regularization parameter is

$$\lambda(\mathbf{X}) = \frac{T(\mathbf{X})}{U(\mathbf{X})}\frac{KP}{MP}, \tag{28}$$

which is asymptotically consistent with Eq. (26) when $P \to \infty$.

3.1.3 Corner of the Bayesian L-curve

If the connection between the cost function J_1 and the curve γ is obvious, there still remains to prove that the minimum of J_1 is found at the corner of γ. The rest of the section progressively establishes this property by answering the related questions:

- where is the minimum of J_1 located on γ?
- what is the condition for the existence of a minimum of J_1?
- what is the condition for the minimum of J_1 to coincide with a point of maximum curvature of γ?

The answer to the first question is provided by the following result.

Proposition 3.2. *The optimal regularization parameter $\lambda^* = \lambda(t^*)$ – with t^* which minimizes $J_1(t)$ – is located where the curve $\gamma(t)$ is tangent with the straight line of slope -1.*

Proof. The optimality condition on J_1 reads $dJ_1(t)/dt = 0 \Leftrightarrow d\eta(t)/d\zeta(t) = -1$. □

The above Proposition 3.2 generalizes Reginska's result (Theorem 1 in [45]) to the family of scale functions $f = -\ln\Phi$ defined in Eq. (23). It is a necessary condition for the minimum of the cost function J_1 to coincide with the corner of the curve γ, because *if* the latter has a marked L-shape, then its vertex must be the point of tangency with the line inclined at $-\pi/4$. This is schematically illustrated in Fig. 1.

The condition for λ^* to exist is now discussed.

Proposition 3.3. *The second derivative of the cost function J_1 at the optimum t^* is*

$$J_1''(t^*) = -\zeta'(t^*)^2\left(CV_N(\lambda(t^*)) + CV_X(\lambda(t^*)) - \frac{\lambda'(t^*)}{\lambda(t^*)\zeta'(t^*)}\right) \tag{29a}$$

$$= -\eta'(t^*)^2\left(CV_N(\lambda(t^*)) + CV_X(\lambda(t^*)) + \frac{\lambda'(t^*)}{\lambda(t^*)\eta'(t^*)}\right) \tag{29b}$$

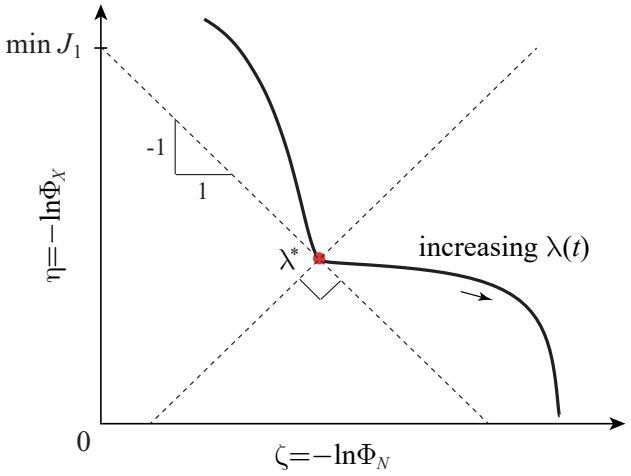

Figure 1: Construction of a Bayesian L-curve $\gamma : t \mapsto \{\zeta(\lambda(t)); \eta(\lambda(t))\}$. The point of tangency with the line of slope -1 corresponds to the optimal regularization parameter λ^* returned by the minimum of J_1. The minimum value of J_1 is given by the intercept of the tangent with axis η. Also shown is the bisector axis (perpendicular to the tangent at λ^*) mentioned in Proposition 3.6: having the L-curve locally symmetric – on the third order in t – around it will guaranty that the minimizer of J_1 coincides with the maximizer of the curvature.

where $CV_N(\lambda) = \mathbb{V}\{\tau_N|\mathbf{X}(\lambda)\}/\mathbb{E}\{\tau_N|\mathbf{X}(\lambda)\}^2$ *and* $CV_X(\lambda) = \mathbb{V}\{\tau_X|\mathbf{X}(\lambda)\}/\mathbb{E}\{\tau_X|\mathbf{X}(\lambda)\}^2$ *are the squared coefficients of variation of the precisions* τ_N *and* τ_X*, respectively.*

The two terms CV_N and CV_X are non-negative by definition, whereas $\zeta' \geq 0$ and $\eta' \leq 0$ according to Proposition A.2. Therefore, a sufficient condition for the second derivative to be positive (a minimum of J_1) is

$$\zeta'^* \frac{\lambda^*}{\lambda'^*} \leq \frac{1}{\mathrm{CV}_N^* + \mathrm{CV}_X^*} \tag{30}$$

or, equivalently,

$$-\eta'^* \frac{\lambda^*}{\lambda'^*} \leq \frac{1}{\mathrm{CV}_N^* + \mathrm{CV}_X^*}, \tag{31}$$

where, for simplicity, the superscript * means that a quantity is evaluated at t^*. Since the regularization scale $t \mapsto \lambda$ is arbitrary, the choice $\lambda = e^t$ (thus implying $\lambda' = \lambda$) is now made as it leads to an insightful result:

Proposition 3.4. *The cost function* $J_1(t)$ *is minimum at* t^* *if the rate of change of the branches* ζ *and* $-\eta$ *with respect to a logarithmic variation of the regularization parameter* λ *is upper bounded by*

$$\zeta'^* = -\eta'^* \leq \frac{1}{CV_N^* + CV_X^*}, \tag{32}$$

a quantity that reflects the well-posedness of the inverse problem.

The fact that the upper bound in the above proposition can be interpreted as a measure of well-posedness of the inverse problem is because it is inversely proportional to $CV_N + CV_X$, the uncertainty induced by the ignorance of the precisions τ_N and τ_X. This implies that a well-posed problem, i.e. with $CV_N + CV_X$ small, will more likely fulfill inequality (32). Alternatively, it results from Proposition 3.4 that ζ' may be interpreted as a measure of the sensitivity of the solution to a deficit of the information required to solve the inverse problem.

It now remains to establish under which conditions the minimum of the cost function J_1 coincides with the corner of the L-curve, as defined by its maximum curvature. Let us introduce the curvature $\kappa(t)$ of the L-curve, as defined in Eq. (6), as a function of the curve parameter t (note that, by construction, the curvature is a metric independent of the choice of the curve parameter) [22]:

Proposition 3.5. *At t^*, the second derivative of the cost function J_1 is related to the curvature κ of the L-curve as*

$$J_1''(t^*) = 2\sqrt{2}\zeta'(t^*)^2\kappa(t^*). \tag{33}$$

Proof. From Proposition 3.2, $\eta = -\zeta$. Substituting this equality for η into (6) directly yields (33). □

Proposition 3.5 states that $J_1''(t^*)$ and $\kappa(t^*)$ have the same signs. Therefore, to a minimum of J_1 ($J_1'' \geq 0$) corresponds a region where the L-curve is locally convex ($\kappa \geq 0$). Together with Proposition 3.2, it establishes a one-to-one correspondence between the minima of the cost function J_1 and the corners of the L-curve, as defined by the loci where the curve is tangent to the line of slope -1 without crossing it.

However, in general, the minimum of the cost function J_1 will not necessarily coincide with a stationary point of the curvature κ. For this to happen, a simple sufficient condition is found.

Proposition 3.6. *A sufficient condition for the minimum of J_1 to correspond to a stationary point of κ at t^* is*

$$\begin{cases} \zeta'(t^*) = -\eta'(t^*) & \text{(34a)} \\ \zeta''(t^*) = \eta''(t^*) \geq 0 & \text{(34b)} \\ \zeta'''(t^*) = -\eta'''(t^*) & \text{(34c)} \end{cases}$$

or, equivalently, that the curve γ is locally symmetric around the axis of slope +1 passing through the point with curve parameter t^*.

Proof. A stationary point of criterion (6) is found where $\kappa' = 0$, which gives

$$2(\zeta''' + \eta''')\zeta' = 3(\zeta''^2 - \eta''^2). \tag{35}$$

A solution to this equation is obtained when the two branches ζ and η are related by conditions (34a)-(34c). The first condition is the same as in Proposition 3.2 and the second condition one guaranties that $J_1''(t) \geq 0$, thus proving that an extremum of κ coincide with a minimum of J_1. The set of Eqs. (34a)-(34c) yields

$$\begin{cases} \zeta(t^* + \Delta t) &= \zeta(t^*) + b\Delta t + c\Delta t^2 + d\Delta t^3 + O(\Delta t^4) \\ \eta(t^* + \Delta t) &= \eta(t^*) - b\Delta t + c\Delta t^2 - d\Delta t^3 + O(\Delta t^4), \end{cases} \tag{36}$$

for some parameters $b \in \mathbb{R}^+$ (so that $\zeta'^* \geq 0$ and $\eta'^* \leq 0$ as requested by Proposition 3.1), $c \in \mathbb{R}^+$ (so that $\zeta''^* \geq 0$), and $d \in \mathbb{R}$. Therefore,

$$\begin{cases} \zeta(t^* + \Delta t) - \zeta(t^*) &= \eta(t^* - \Delta t) - \eta(t^*) + O(\Delta t^4) \\ \eta(t^* + \Delta t) - \eta(t^*) &= \zeta(t^* - \Delta t) - \zeta(t^*) + O(\Delta t^4), \end{cases} \tag{37}$$

which expresses the local symmetry of the curve γ around its bisector axis, i.e. the straight line of slope +1 passing through the point $(\zeta(t^*), \eta(t^*))$. This is illustrated in Fig. 1. □

Two remarks come with Proposition 3.6. First, conditions (34a)-(34c) depend on the existence of a regularization scale $t(\lambda)$ where these equations hold true. For instance, this might be the case for $t = \ln\lambda$, but not for $t = \lambda$, which illustrates the importance of properly selecting the curve parameter associated with a given L-curve. Second, conditions (34a)-(34c) alone do not guarantee that a minimum of J_1 actually corresponds to a *maximum* of κ (it may be a minimum), although it always locates a "convex corner" since $\kappa(t^*) \geq 0$ in virtue of Proposition 3.5. Instances where $\kappa(t^*)$ is a minimum would probably reflect situations where the use of the maximum curvature criterion is anyway troublesome for finding the optimal regularization parameter.

3.2 Link with the EM algorithm

The implicit nature of the solution given in Proposition 2.2 suggests an iterative resolution, where Eq. (16b) at the j-th iteration is evaluated with the estimate of the regularization parameter in Eq. (18) given at iteration $j-1$. When compared to a basic-search method, such an algorithm may be beneficial if the evaluation of the inverse operator (16b) is costly.

This is actually recognized as the maximization (M) and expectation (E) steps of the EM algorithm:

M step:

$$\mathbf{X}^{(j)} = \mathrm{Arg}\max_{\mathbf{X}} \iint_{\mathbb{R}^{2+}} \ln\left(p(\mathbf{Y}|\mathbf{X},\tau)p(\mathbf{X}|u)\right) p(\tau,u|\mathbf{Y},\mathbf{X}^{(j-1)})d\tau du \tag{38}$$

$$= \left(\mathbf{A}^\top\mathbf{A} + \frac{\mathbb{E}\{\tau_X|U(\mathbf{X}^{(j-1)})\}}{\mathbb{E}\{\tau_N|T(\mathbf{X}^{(j-1)})\}}\mathbf{I}\right)^{-1}\mathbf{A}^\top\mathbf{Y} \tag{39}$$

$$\tag{40}$$

E step:

$$\mathbb{E}\{\tau_N|T(\mathbf{X}^{(j)})\} = \iint_{\mathbb{R}^+} up(\tau,u|\mathbf{Y},\mathbf{X}^{(j)})d\tau du = -\frac{d}{ds}\ln\Phi_N(s;MP/2)\bigg|_{s=T(\mathbf{X}^{(j)})/2} \tag{41}$$

$$\mathbb{E}\{\tau_X|U(\mathbf{X}^{(j)})\} = \iint_{\mathbb{R}^+} \tau p(\tau,u|\mathbf{Y},\mathbf{X}^{(j)})d\tau du = -\frac{d}{ds}\ln\Phi_X(s;KP/2)\bigg|_{s=U(\mathbf{X}^{(j)})/2} \tag{42}$$

where $\mathbf{X}^{(j)}$ denotes the estimate of $\mathbf{X}$ at the j-th iteration, $p(\tau_N,\tau_X|\mathbf{Y},\mathbf{X}^{(j-1)})$ is the probability density of (τ_N,τ_X) conditioned on $(\mathbf{Y},\mathbf{X}^{(j-1)})$, and $\mathbb{E}\{\tau_X|U(\mathbf{X}^{(j-1)})\}$ and $\mathbb{E}\{\tau_N|T(\mathbf{X}^{(j-1)})\}$ are the expected values of τ_X and τ_N conditional to $(\mathbf{X}^{(j-1)},\mathbf{Y})$ as defined in Proposition 2.2 (see also Eq. (75)).

The above algorithm is very similar in its principle to the fixed point algorithm introduced in Refs. [8, 55, 24], yet it fundamentally differs in the way the current value of λ is computed at each iteration (the E step), since the latter references do not follow the MMAP paradigm. In addition, being an EM algorithm, the convergence is guaranteed by the general results established in Ref. [53].

An important quantity coming with the EM algorithm is the "missing information" matrix, which measures the loss of information – specifically, the deficit in the Fisher information matrix – in estimating $\mathbf{X}$ without knowing the regularization parameter as compared to the ideal case when it is known. As such, it also controls the rate of convergence of the EM algorithm [37]. In the present case, the missing information matrix related to the estimation of a column $\mathbf{x}_i$ of $\mathbf{X}$ is

$$\mathbf{I}_m = \frac{\partial\lambda}{\partial\mathbf{x}_i}\frac{\partial\lambda}{\partial\mathbf{x}_i^\top}\left(\frac{\zeta'^*}{\lambda'^*}\right)^2(CV_N^* + CV_X^*)\,, \tag{43}$$

where $\partial\lambda/\partial\mathbf{x}_i$ stands for the gradient of λ with respect to $\mathbf{x}_i$, CV_N^* and CV_X^* are the squared coefficients of variation introduced in Proposition 3.2, and all quantities are evaluated at the optimum $\lambda = \lambda^*$ (see

proof in Appendix G). This result involves the same critical factors $(\zeta'^*)^2$, and $(CV_N^* + CV_X^*)$ as the second derivative of the cost function J_1 given by Eq. (29a). Clearly, the smaller ζ'^2, CV_N and CV_X, the faster the rate of convergence of the EM algorithm and the higher the precision of the estimate of λ returned by the minimum of J_1.

3.3 On variant criteria

The previous sections have evidenced that the use of the L-curve is intimately bounded up with a criterion to locate its corner. This section opens the discussion to other criteria than those already introduced and attempts to connect them.

3.3.1 The minimum speed of the L-curve

The maximum curvature criteria is a heuristic, which does not proceed from an optimality principle. It accounts for the shape of the curve from a geometrical point of view, independently of its parameterization. This means it is not affected by how fast a point moves along the curve. Although this is the property originally sought by the concept, it may miss relevant characteristics of the problem at hand. Experimental results often show that for a particular regularization scale – typically a logarithmic scale $t = \ln \lambda$ – the speed of the curve tends to be minimum in the region of the corner, as reflected by a higher density of evaluation points (see for instance Fig. 1 of Ref. [12], Figs. 5 and 6 of Ref. [42], Figs. 1, 3 and 13 of Ref. [31] and Fig. 2 of this paper). The corresponding cost function – the square of the *speed of the L-curve* – reads

$$J_2(t) = \zeta'(t)^2 + \eta'(t)^2. \tag{44}$$

That searching for the minimum of $J_2(t)$ is a sensible criteria should not come with too much surprise, as Property 3.3 and Eq. (43) have already shown the importance of keeping the magnitude of both ζ'^2 and η'^2 small in order for J_1 to be minimum[4]. A more formal justification is as follows. Let us introduce the (squared) Euclidian distance, $d^2(P_2, P_1) = (\zeta_2 - \zeta_1)^2 + (\eta_2 - \eta_1)^2$, between two points $P_1(\zeta_1; \eta_1)$ and $P_2(\zeta_2; \eta_2)$ in plane $\mathcal{P}$. Substituting expressions (23) for ζ and η, this is also the (squared) Riemannian distance, $\ln^2(\Phi_N(T_2/2)/\Phi_N(T_1/2)) + \ln^2(\Phi_X(U_2/2)/\Phi_X(U_1/2))$, between the positive forms $(\Phi_N(T_2/2); \Phi_X(U_2/2))$ and $(\Phi_N(T_1/2); \Phi_X(U_1/2))$ [38] – a scale invariant distance that depends only on the shape of Φ_N and Φ_X. Now let $T_t = T(\mathbf{X}(\lambda(t)))$ and $U_t = U(\mathbf{X}(\lambda(t)))$ for notational simplicity. One would expect an "optimal" value of $\lambda(t)$ to lie in a region where a small perturbation Δt implies a minimal modification of the positive form $(\Phi_N(T_t/2); \Phi_X(U_t/2))$, that is a minimum of

$$d^2(P_{t+\Delta t}, P_t) = (\zeta(t + \Delta t) - \zeta(t))^2 + (\eta(t + \Delta t) - \eta(t))^2 \simeq \left(\eta'(t)^2 + \zeta'(t)^2\right) \Delta t^2 = J_2(t) \Delta t^2, \tag{45}$$

which is proportional to $J_2(t)$.

The following proposition holds.

[4]Incidentally, Ref. [54] introduced a "f-slope" criterion, which exactly amounts to minimizing η'^2 with respect to $t = \ln \lambda$ in a predefined interval in the case $-\ln \Phi_\tau(s) = \sqrt{s}$. The criterion is shown to have remarkable performance, although it does not depend at all on the branch ζ; the reason of the latter observation might be that the f-slope method preselects a stable region where $\zeta'^2 \simeq \eta'^2$. It is noteworthy that the f-slope criterion can be recovered as a particular case of the present Bayesian framework by setting a flat prior on τ_N (for instance $\beta_N/\alpha_N \to \infty$ in Eq. (51)) so that ζ'^2 will vanish in comparison to η'^2

Proposition 3.7. *Under the conditions* (34a)-(34c) *of Proposition 3.6,*

$$\operatorname{Arg\,min}_t J_2(t) = \operatorname{Arg\,min}_t J_1(t). \tag{46}$$

This validates the relevance of criterion J_2, as under certain circumstances it will spot the same optimal regularization parameter as the cost function J_1 and, in turn according to Proposition 3.6, as the curvature criterion (6).

Interestingly, the curvature κ and the squared speed J_2 are related to the acceleration of a point moving on the curve γ. Let $\mathbf{T}(t)$ denote the unitary *tangent vector* on $\gamma(t)$ in plane $\mathcal{P}$ and $\mathbf{N}(t)$ the unitary *normal vector*, i.e. the vector in plane $\mathcal{P}$ orthogonal to $\mathbf{T}(t)$ and pointing inside the curve ($\{\mathbf{T}, \mathbf{N}\}$ forms the so-called Frenet frame). The acceleration on $\gamma(t)$ is then expressed as

$$\mathbf{\Gamma}(t) = \frac{d}{dt}\sqrt{J_2(t)}\mathbf{T}(t) + J_2(t)\kappa(t)\mathbf{N}(t). \tag{47}$$

At point $t^\bullet$ where the speed $\sqrt{J_2}$ is minimum, the acceleration $\mathbf{\Gamma}(t^\bullet) = J_2(t^\bullet)\kappa(t^\bullet)\mathbf{N}(t^\bullet)$ is then purely centripetal (i.e. oriented along $\mathbf{N}$, towards the centre of the circle around which the curve is rolled up). Under the conditions of Propositions 3.6 and 3.7, the norm $\|\mathbf{\Gamma}(t)\|$ of the acceleration then finds a stationary point at $t^* = t^\bullet$. These ideas are illustrated in Fig. 2.

Propositions 3.6 and 3.7 together open the way to the definition of many other plausible criteria for finding the corner of the L-curve. Indeed, if t^* is a stationary point of $\kappa(t)$, $J_1(t)$, $J_2(t)$, then it is also a stationary point for any combination of the form $g_0(\kappa(t))g_1(J_1(t)) + g_2(J_1(t))g_3(J_2(t)) + g_4(\kappa(t))g_5(J_2(t))$, with g_i, $i = 0, ..., 5$ some continuous and monotonous functions. The next subsection gives one such example.

3.3.2 The maximum angular speed

Inspection of Fig. 2 shows that the corner of the L-curve is also the locus where the normal vector $\mathbf{N}$ experiences the fastest rate of rotation. For a perfect L-shaped curve, the rotation would be counterclockwise from $\theta = 0$ to $\theta = \pi/2$, where θ denotes the angle between $\mathbf{N}$ and the horizontal axis of the frame (ζ, η). Therefore, a sensible criterion is to look for the maximum of the *angular speed*

$$J_3(t) = \frac{d\theta(t)}{dt} = \frac{\zeta'(t)\eta''(t) - \zeta''(t)\eta'(t)}{\zeta'(t)^2 + \eta'(t)^2}. \tag{48}$$

It is recognized that $J_3(t) = \kappa(t)\sqrt{J_2(t)}$ ($g_0 = g_1 = g_2 = g_3 = 0$, $g_4 = id$ and $g_5 = \sqrt{}$), which means that J_3 has the same stationary point as κ and J_2 under the conditions of Propositions 3.6 and 3.7. The relationship between J_3 and the curvature κ is insightful: the maximum curvature criterion searches for the fastest rate of rotation of the normal vector $\mathbf{N}$ when the curve is travelled at constant speed, i.e. when the derivative of θ is taken with respect to the arc-length $s(t) = \int_{u_0}^{t} \sqrt{J_2(u)}du$ instead of the curve parameter t. If the speed on the curve matters, then J_3 should be preferred to κ.

4 Particular cases

This section investigates the particular forms taken by the Bayesian L-curve and the issuing properties when it is assigned some specific priors ϕ_N and ϕ_X. Known particular cases are recovered and new formulations are elicited.

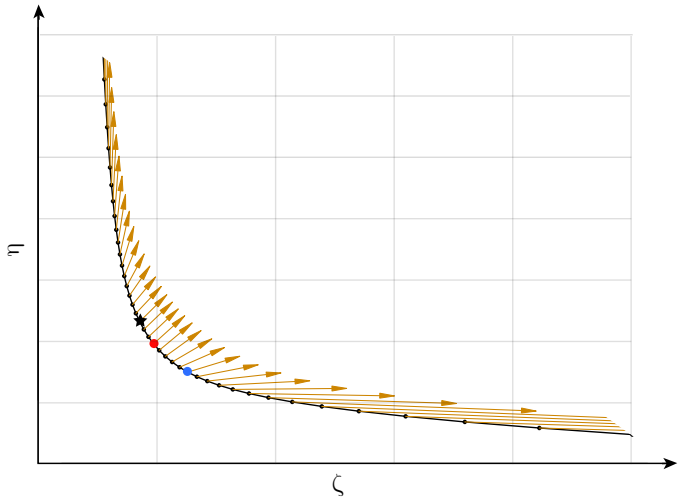

Figure 2: Example of an acceleration vector field, $\mathbf{\Gamma}(t)$, on the L-curve. The arrow feet (black points) are equispaced on the regularization scale $t = \ln \lambda$; the arrow lengths are proportional to $\|\mathbf{\Gamma}(t)\|$. The minima of the cost functions $J_1(t)$ and $J_2(t)$ are marked by the red and blue bullets, respectively, and the maximum of the curvature κ by the black star. It is noteworthy that the acceleration is normal to the curve (minimum of $J_2(t)$) in the neighborhood of the maximum curvature. (The curve was simulated with $M = 15$, $K = 13$, $P = 50$, $\tau_X = 1$, $\tau_N = 10^6$, and $\eta = 0.1$ according to the method described in Appendix H.)

4.1 L-curve with gamma priors

The case with gamma priors plays a special role because it provides a generalization of the reference L-curve defined with logarithmic scales.

4.1.1 Basic derivation

Let first note that the Φ-transform of the gamma probability density

$$\phi(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \tag{49}$$

with shape and rate parameters $\alpha > 0, \beta > 0$ is

$$\Phi(s; n) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + n)}{(\beta + s)^{\alpha+n}}. \tag{50}$$

Therefore, plugging gamma priors $\phi_N(\tau)$ and $\phi_X(\tau)$ with hyperparameters (α_N, β_N) and (α_X, β_X) in Eq. (23) produces the curve

$$\gamma_\Gamma : t \mapsto \begin{cases} \zeta(t) &= (\alpha_N + MP/2) \ln(\beta_N + T(\mathbf{X}(\lambda(t)))/2) \\ \eta(t) &= (\alpha_X + KP/2) \ln(\beta_X + U(\mathbf{X}(\lambda(t)))/2) \end{cases} \tag{51}$$

associated with the optimal regularization parameter

$$\lambda^* = \frac{\mathbb{E}\{\tau_X|U^*\}}{\mathbb{E}\{\tau_N|T^*\}} = \left(\frac{2\alpha_X + KP}{2\beta_X + U(\mathbf{X}(\lambda^*))}\right)\left(\frac{2\beta_N + T(\mathbf{X}(\lambda^*))}{2\alpha_N + MP}\right). \tag{52}$$

Remark: L-curve with Jeffrey's priors
Jeffrey's prior arrives as a particular case of the gamma prior when both the shape and the rate parameters collapse ($\alpha \to 0$ and $\beta \to 0$). One then finds the L-curve (27) already introduced in section 3.

4.1.2 Comparison with the reference L-curve

When the rate parameters vanish ($\beta_N = \beta_X \to 0$) and ($\alpha_N + MP/2$) = ($\alpha_X + KP/2$) = c, the reference L-curve $\gamma_{\ln}$ is then exactly recovered up to the global scale factor c. For arbitrary values of $\alpha_N + MP/2$ and $\alpha_X + KP/2$ (but still vanishing rate parameters), this L-curve is implicitly the one used in Reginska's criterion. To see this, let us introduce the change of variables

$$\begin{cases} \zeta_G(t) &= \zeta(t)/(\alpha_N + MP/2) - \ln 2 \quad &= \ln(2\beta_N + T(\mathbf{X}(\lambda(t)))) \\ \eta_G(t) &= \eta(t)/(\alpha_X + KP/2)) - \ln 2 \quad &= \ln(2\beta_X + U(\mathbf{X}(\lambda(t)))) \end{cases} \tag{53}$$

and the corresponding curve $\gamma_\Gamma : t \mapsto (\zeta_G(t); \eta_G(t))$. This change of variable requires Proposition 3.2 to be reformulated:

Corollary 4.1. *The optimal regularization parameter λ^* that minimizes the cost function J_1 is located where the curve γ_Γ is tangent with the straight line of slope $-(\alpha_N + MP/2)/(\alpha_X + KP/2)$.*

The above result turns out identical to Reginska's criterion with $\mu = (\alpha_X + KP/2)/(\alpha_N + MP/2)$ in Eq. (7) and $\beta_N = \beta_N \to 0$ in Eq. (53); it is also identical to the JMAP estimate discussed in subsection 2.1 (see Eq. (13)).

It is important to emphasize that, in general, the L-curve γ_Γ defined by Eq. (51) and the reference L-curve $\gamma_{\ln}$ do not share the same shape and the same corner. First, the reference L-curve allows the branches $\ln U$ and $\ln T$ to diverge to $-\infty$ when $\lambda \to \infty$ and $\lambda \to 0$, respectively, which tends to destroy global convexity by bending the tip of the vertical branch of the "L" towards the left and the end of its horizontal branch downwards (thus producing two inverted "L"s). By adding non-zero constants β_N and β_X into the logarithms, the branches $\ln(\beta_N + T/2)$ and $\ln(\beta_X + U/2)$ can no longer diverge when $T \to 0$ or $U \to 0$ (see Fig. 4). Second, the introduction of weights $\alpha_N + MP/2$ and $\alpha_X + KP/2$ balances the relative importance of the branches $\ln(\beta_N + T/2)$ and $\ln(\beta_X + U/2)$ depending on the dimensions of the problem. Specifically, more weight is given to $\ln(\beta_N + T/2)$ when the dimension of the measurements M is large as compared to the number of unknowns K, which tends to move the corner eastward in plane $\mathcal{P}$, thus favoring smaller values of λ^* than with the reference L-curve; in other words, less regularization is required when the inverse problem becomes well-posed. Alternatively, more weight is given to $\ln(\beta_X + U/2)$ when the number of unknowns K becomes large as compared to the measurements M, which tends to move the corner upward, thus favoring larger values of λ^* than with the reference L-curve when the problem becomes ill-posed. The dependence of the regularization parameter on the problem dimensions does not exist with the reference L-curve, neither with Reginska's cost function, where the slope $-1/\mu$ is *arbitrarily* set to a constant value (most often $\mu = 1$ in the literature). As far as the authors know, that the location of the corner of the L-curve should depend on the problem dimensions has never been recognized before.

	quantile	$\log_{10}\lambda_{SE}$	$\log_{10}\lambda_\kappa^{BL}$	$\log_{10}\lambda_{J_1}^{BL}$	$\log_{10}\lambda_\kappa$
$M/K=1$	Q_{25}	-8.06	-7.90	-8.55	-7.90
	Q_{50}	-8.00	-7.86	-8.51	-7.86
	Q_{75}	-7.95	-7.74	-8.40	-7.74
$M/K=10$	Q_{25}	-8.05	-8.05	-8.00	-7.08
	Q_{50}	-8.00	-8.02	-7.98	-7.06
	Q_{75}	-7.94	-8.00	-7.96	-7.03
$M/K=50$	Q_{25}	-8.05	-8.01	-7.96	-6.34
	Q_{50}	-8.00	-7.99	-7.95	-6.31
	Q_{75}	-7.95	-7.97	-7.93	-6.29
$M/K=100$	Q_{25}	-8.05	-8.01	-7.96	-6.00
	Q_{50}	-8.00	-7.99	-7.94	-5.98
	Q_{75}	-7.94	-7.97	-7.93	-5.95

Table 1: Quantiles of estimates of the regularization parameter: λ_{SE} stands the reference that minimizes the square error $\|\mathbf{X}-\mathbf{X}(\lambda)\|_F^2$, λ_κ^{BL} for the maximum curvature estimate of the Bayesian L-curve, $\lambda_{J_1}^{BL}$ for the minimum of J_1, and λ_κ for the maximum curvature estimate of the reference L-curve.

Example

These observations are briefly illustrated by means of a numerical example. A discrete inverse problem is simulated according to the method of Appendix H with $K=10$, $\tau_X=1$, $\tau_N=10^4$, $P=100$, $\eta=10$, and $M=rK$ with $r=1,10,50,100$. In this particular case, the normalized mean-square error (NMSE) reads

$$\frac{\mathbb{E}\|\mathbf{X}-\mathbf{X}(\lambda)\|_F^2}{\mathbb{E}\|\mathbf{X}\|_F^2}=\frac{1}{K}\sum_{k=1}^{K}\frac{(\tau_X/\tau_N)s_k^2+\lambda^2}{(s_k^2+\lambda^2)^2}, \tag{54}$$

where $\{s_k; k=1,...,K\}$ are the singular values of matrix $\mathbf{A}$ (see Appendix H). It is remarked that the NMSE is theoretically independent of M, thus implying that the optimal regularization parameter should also be independent of M in this case. For each value of the ratio r, the reference L-curve is computed together with the "Bayesian L-curve" (51) rooted on gamma priors with non-informative hyperparameters $\alpha_N=\alpha_X=0.1$ and $\beta_N=\beta_X=10^{-16}$. The optimal regularization parameter is estimated from the maximum curvature for the two L-curves and also from the minimum of the cost function J_1 for the gamma L-curve. This is repeated for 10^4 independent random draws of $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{N}$. Table 1 reports the quantiles of the estimated regularization parameters and Fig. 3 displays the boxplots of the normalized square error (NSE) $\|\mathbf{X}-\mathbf{X}(\lambda)\|_F^2/\|\mathbf{X}\|_F^2$. It is seen that both the maximum curvature estimate and the minimizer of J_1 are statistically very close to the reference λ_{SE} that minimizes the NSE on the gamma L-curve, whatever the value of M; on the contrary, the maximum curvature estimate of the reference L-curve significantly over-regularizes the problem when M increases, thus inflating the normalized mean-square error (see Fig. 3).

4.1.3 Sufficient condition for the existence of a minimum of J_1

The condition of existence of a minimum of the cost function J_1 in the case of gamma priors is given by Proposition 3.4. Using Proposition A.1, it is readily found that $\mathrm{CV}_N=1/(\alpha_N+MP/2)$ and $\mathrm{CV}_X=1/(\alpha_X+KP/2)$. Hence, condition (32) becomes

$$\frac{d}{dt}\ln(\beta_N+T(\mathbf{X}(\lambda(t)))/2)\bigg|_{t=t^*}\leq\frac{1}{\left(1+\frac{\alpha_N+MP/2}{\alpha_X+KP/2}\right)} \tag{55}$$

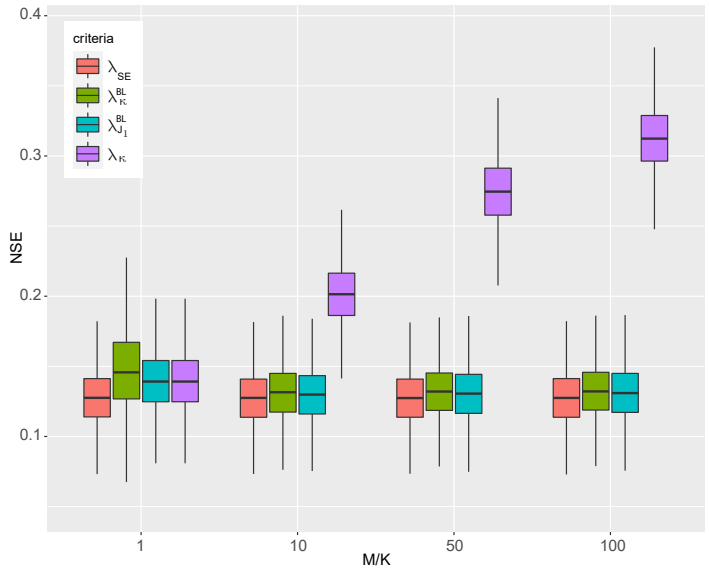

Figure 3: Boxplots (marking the lower, middle, and upper quartiles and whiskers set to four times the interquartile range) of the normalized square error (estimated on 10^4 runs) for increasing values of the ratio M/K: λ_{MSE}, λ_κ^{BL}, $\lambda_{J_1}^{BL}$, and λ_κ are as defined in Table 1.

or, equivalently,

$$-\frac{d}{dt}\ln(\beta_X + U(\mathbf{X}(\lambda(t)))/2)\Big|_{t=t^*} \leq \frac{1}{\left(1+\frac{\alpha_X+KP/2}{\alpha_N+MP/2}\right)}. \tag{56}$$

This indicates that the rate of variation of the branches $\zeta(t)$ and $-\eta(t)$ should be upper bounded. Equation (55) is relevant for high values of λ^* (towards the horizontal branch of the L) where $\ln(\beta_N + T/2)$ is likely to grow fast, while at the same time $\ln(\beta_X + U/2)$ is likely to reach an asymptote. Inversely, Eq. (56) is relevant for small values of λ^* (towards the vertical branch of the L), where the behaviors of $\ln(\beta_N+T/2)$ and $\ln(\beta_X+U/2)$ are reversed. The two equations elicit the importance of the hyperparameters β_N and β_X to control the variations of $\ln(\beta_N+T(\mathbf{X}(\lambda))/2)$ and $\ln(\beta_X+U(\mathbf{X}(\lambda))/2)$ – it always holds that $d\ln T/dt \geq d\ln(\beta_N + T/2)/dt$ and $-d\ln U/dt \geq -d\ln(\beta_X + U/2)/dt$ – and of α_N and α_X to control the upper-bounds, and therefore to meet the condition of convexity of the cost function J_1. This is illustrated in Fig. 4.

4.1.4 A geometrical interpretation

As explained in this section, the gamma priors with vanishing rate parameters provide a geometrical interpretation to the reference L-curve as well as a justification to the cost function J_3 introduced in subsection 3.3.2. To start with, consider the reparametrization (53) and denote as M the point in plane $\mathcal{P}$ whose position is given by the coordinates $(\ln T(\mathbf{X}(\lambda^*)); \ln U(\mathbf{X}(\lambda^*)))$, with λ^* the optimal regularization parameter that minimizes the cost function J_1 for a particular pair of hyperparameters (α_N, α_X). According to Corollary 4.1, the position of M is where the curve γ intersects the straight line of slope $p = -(\alpha_N + MP/2)/(\alpha_X + KP/2)$. Specifically, one has $\eta = g(p) + p \cdot \zeta$, where $g(p)$ is the Legendre transform of the function $\zeta \mapsto \eta$. When the hyperparameters (α_N, α_X) are continuously varied, the trajectory of point M then follows the curve γ_Γ, which is interpreted as the envelope of the family of straight lines parametrized by different values of p. This is illustrated in Fig. 5.

Based on this interpretation, a natural idea is to select, among all optimal solutions parametrized

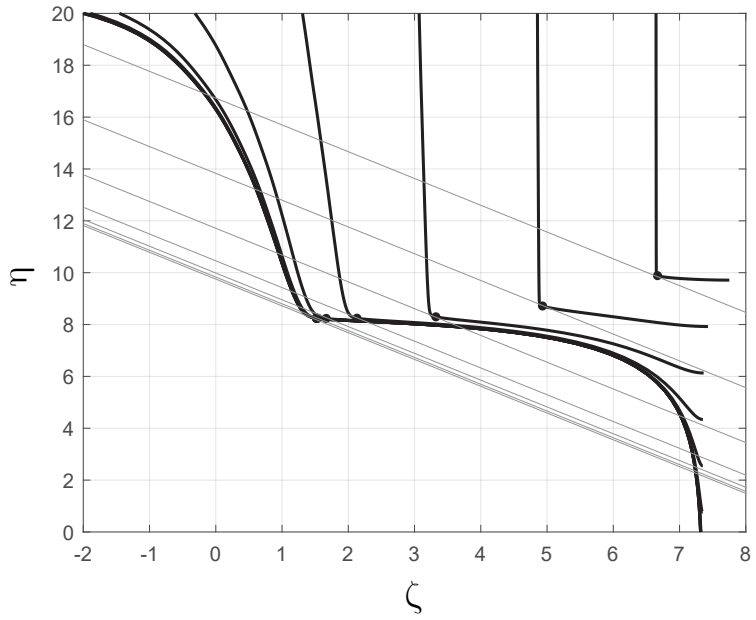

Figure 4: Example of "Bayesian L-curves" $\lambda \mapsto \{\zeta(\lambda) = (\alpha_N + MP/2)\ln(\beta_N + T(\mathbf{X}(\lambda))/2); \eta(\lambda) = (\alpha_X + KP/2)\ln(\beta_X + U(\mathbf{X}(\lambda))/2)\}$ and associated tangents at λ^* in the gamma family of priors, for different values of the ratio $\beta_N/\alpha_N = \beta_X/\alpha_X$. The L-shape becomes more and more pronounced and its corner acute as the ratio increases, i.e. as the prior distributions become informative. (The curve was simulated with $M = 20$, $K = 61$, $P = 5$, $\tau_X = 1$, $\tau_N = 100$, and $\eta = 0.5$ according to the method described in Appendix H.)

by p, the regularization parameter that is the least sensitive to a change in the hyperparameters. This amounts to searching for the minimum of the variation of t due to a perturbation in the angle $\theta = \arctan(p)$ along the trajectory γ_Γ, i.e.

$$\operatorname{Arg\,min}_t \frac{dt}{d\theta} = \operatorname{Arg\,max}_t \frac{d\theta}{dt} = \operatorname{Arg\,max}_t J_3(t), \tag{57}$$

thus recovering the maximum angular speed criterion (48).

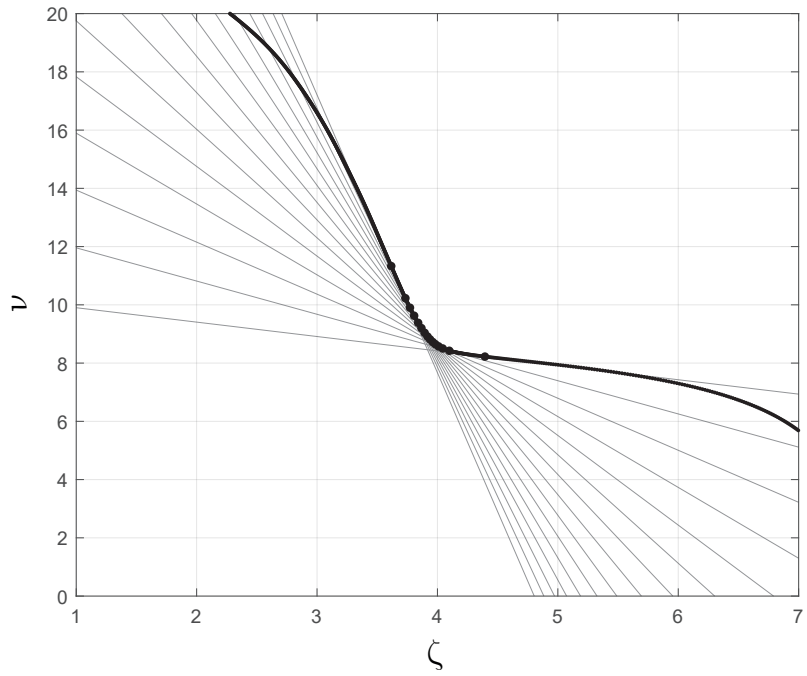

Figure 5: Interpretation of the convex part of the reference L-curve $\lambda \mapsto \{\zeta(\lambda) = \ln(T(\mathbf{X}(\lambda)))\,; \eta(\lambda) = \ln(U(\mathbf{X}(\lambda)))\}$ as the envelope of the straight lines parameterized by slope $p = -(\alpha_N + MP/2)/(\alpha_X + KP/2)$ in the family of gamma priors and ordinate given by the Legendre transform of function $\zeta \mapsto \eta$. The corner of the curve is where the position of the regularization parameter least depends on the hyperparametrization of the gamma priors: this corresponds to the smallest step-size between points of tangency – marked by black bullets – with the straight lines whose slopes are incremented by regular angles $\Delta\theta = \Delta \arctan(p)$. (The curve was simulated as in Fig. 4.)

4.2 L-curve with inverse gamma priors

The Φ-transform of the inverse gamma density

$$\phi(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{-(\alpha+1)} e^{-\beta/\tau} \tag{58}$$

with shape and rate parameters $\alpha > 0, \beta > 0$ is

$$\Phi(s; n) = \frac{2\beta^{(\alpha+n)/2}}{\pi^n \Gamma(\alpha)} s^{(a-n)/2} K_{n-a}(2\sqrt{bs}), \tag{59}$$

with K_{n-a} the modified Bessel function of the second-kind and order $n - a$. Taking inverse gamma priors for the precisions τ_N and τ_X with hyperparameters (α_N, β_N) and (α_X, β_N), Eq. (23) then yields the curve

$$\gamma_{\Gamma^{-1}} : t \mapsto \left\{ \begin{array}{ll} \zeta(t) & = \frac{(MP/2-\alpha_N)}{2} \ln(T(\mathbf{X}(\lambda(t)))/2) - \ln K_{MP/2-\alpha_N}(\sqrt{2\beta_N U(\mathbf{X}(\lambda(t)))}) \\ \eta(t) & = \frac{(KP/2-\alpha_X)}{2} \ln(U(\mathbf{X}(\lambda(t)))/2) - \ln K_{KP/2-\alpha_X}(\sqrt{2\beta_X T(\mathbf{X}(\lambda(t)))}). \end{array} \right. \tag{60}$$

The associated optimal regularization parameter is $\lambda^* = \mathbb{E}\{\tau_X|U^*\}/\mathbb{E}\{\tau_N|T^*\}$, where

$$\left\{\begin{array}{ll} \mathbb{E}\{\tau_N|T\} & = -\frac{d}{ds}\ln\Phi_N(s;MP/2)|_{s=T/2} \\ \mathbb{E}\{\tau_X|U\} & = -\frac{d}{ds}\ln\Phi_X(s;KP/2)|_{s=U/2} \end{array}\right. \begin{array}{l} = \frac{(MP/2-\alpha_N)}{T} - \frac{\sqrt{\beta_N}}{\sqrt{T}}\frac{K'_{MP/2-\alpha_N}(\sqrt{2\beta_N T})}{K_{MP/2-\alpha_N}(\sqrt{2\beta_N T})} \\ = \frac{(KP/2-\alpha_X)}{U} - \frac{\sqrt{\beta_X}}{\sqrt{U}}\frac{K'_{KP/2-\alpha_X}(\sqrt{2\beta_X U})}{K_{KP/2-\alpha_X}(\sqrt{2\beta_X U})}. \end{array} \tag{61}$$

The asymptotic forms of the above equations provide some insight. Since $K'_{n-a}(2\sqrt{bs})/K_{n-a}(2\sqrt{bs}) \to -(n-a)/(2\sqrt{bs})$ as $s \to \infty$, it comes that $\mathbb{E}\{\tau_N|T\} \to MP/T$ and $\mathbb{E}\{\tau_X|U\} \to KP/U$ as $P \to \infty$ provided that α_N, α_X, β_N and β_X do not depend on P. This correctly returns the asymptotic expressions of the regularization parameter established in Eqs. (26) and (27).

One consequence of the inverse gamma priors is to provide a Bayesian interpretation to the L-curve with square root scaling. To see this, let $\alpha_N = (MP+1)/2$, $\beta_N = (MP)^2/2T_0$, $\alpha_X = (KP+1)/2$, $\beta_X = (KP)^2/2U_0$, where T_0 and U_0 stand for guess values of U^* and T^*. This way, the prior expected value of τ_N (resp. τ_X) is $\mathbb{E}\{\tau_N\} \sim MP/T_0$ (resp. $\mathbb{E}\{\tau_X\} \sim KP/U_0$). Then, using the property $K_{1/2}(z) = \exp(-z)\sqrt{\pi/2z}$, it can be shown that

$$\gamma_{\surd} : t \mapsto \left\{\begin{array}{ll} \zeta(t) & = MP\sqrt{T(\mathbf{X}(\lambda(t)))/T_0} \\ \eta(t) & = KP\sqrt{U(\mathbf{X}(\lambda(t)))/U_0}. \end{array}\right. \tag{62}$$

This is recognized equal to the square-root L-curve defined with scaling function $f = \surd$ in Eq. (5). The associated optimal regularization parameter is

$$\lambda^* = \frac{\mathbb{E}\{\tau_X|U^*\}}{\mathbb{E}\{\tau_N|T^*\}} = \left(\frac{KP}{\sqrt{U_0 U^*}}\right)\left(\frac{\sqrt{T_0 T^*}}{MP}\right) \tag{63}$$

and is located where the curve $\gamma_{\surd}$ is tangent with the straight line of slope $-(M\sqrt{U_0})/(K\sqrt{T_0})$.

It is remarked that this construction actually requires the initial knowledge of the signal-to-noise ratio U_0/T_0. If not available, one could still use the square root L-curve (62), yet with a criterion that does not depend on the prior hyperparameters, such as κ, J_2, or J_3. Another strategy is to replace the ratio $(T_0/MP)/(U_0/KP)$ in Eq. (63) by a current estimated of λ, thus leading to the iterations

$$\lambda^{k+1} = \sqrt{\lambda^k \frac{T(\mathbf{X}(\lambda^k))/MP}{U(\mathbf{X}(\lambda^k))/KP}}. \tag{64}$$

It is readily checked that this algorithm converges to $\lambda^* = (T(\mathbf{X}(\lambda^*))/MP)/(U(\mathbf{X}(\lambda^*))/KP)$ provided that it is initialized sufficiently close to λ^*. The regularization parameter thus estimated corresponds to a non-informative prior, as $(T_0/MP)/(U_0/KP)$ is no longer fixed but is free to evolve.

4.3 Minimum conditional-variance priors

The question naturally arises as whether there exist priors for the precisions τ_N and τ_X that maximize the sharpness of the J_1 cost function or, equivalently, the ability to locate the corner of the L-curve. Referring back to condition (30), this amounts to minimizing the squared coefficients of variation CV_N and CV_X, that is to select priors as narrow as possible. The extreme case is obtained when the conditional variances $\mathbb{V}\{\tau_N|T\}$ and $\mathbb{V}\{\tau_X|U\}$ are null, which according to Proposition A.1 is $(\ln\Phi)'' = 0$. The solution to this differential equation is $\Phi(s;n) = C_1 e^{-C_2 s}$ for some constants $C_1 \in \mathbb{R}^+$ and $C_2 \in \mathbb{R}$. One sees that this is the Φ-transform of $\phi(\tau) = \delta(\tau - C_2)$. Therefore, without

surprise, this is recognized as the ideal case where τ_N and τ_X are known beforehand, in which case $\lambda^* = \tau_X/\tau_N$. The corresponding L-curve is $\gamma_{lin} : t \mapsto \{\zeta = \tau_N T; \eta = \tau_X U\}$, recognized as a scaled version of the linear L-curve (the construction of an L-curve is still possible, although futile since the value of λ is now known!).

4.4 Other curves?

A last question to address is whether other types of L-curves can be similarly interpreted within the present Bayesian framework. This question is equivalent to as whether a given scaling function f in Eq. (5) is a Φ-transform.

Proposition 4.1. *Let $f(s) : \mathbb{R}^+ \to \mathbb{R}$ denote a scaling function and $\Phi(s;n)$ the Φ-transform of some probability density $\phi(\tau)$ defined on $\mathbb{R}^+$. A sufficient and necessary condition for having $f(s) = -\ln \Phi(s;n) + C$, with C a constant, is that $(-d/ds)^k \exp(-f(s)) \geq 0$, $k \in \mathbb{N}$ and $\mathbb{E}\{\tau^n\} = \int_{\mathbb{R}^+} \tau^n \phi(\tau) d\tau < \infty$.*

Proof. The necessary condition readily follows from differentiating Eq. (15) wherein $\phi(\tau) \geq 0$. The sufficient condition is a consequence of Bernstein's theorem, which states that a function $F(s)$ has the property $(-d/ds)^k F(s)) \geq 0$, $k \in \mathbb{N}$, if and only if $F(s) = \int_{\mathbb{R}^+} e^{-s\tau} d\alpha(\tau)$ where $\alpha(\tau)$ is bounded and non-increasing (see Ref. [51][chap. 12]). Here $d\alpha(\tau) = \tau^n \phi(\tau) d\tau$. □

Proposition 4.1 clearly applies to $f(s) = \ln(s)$, $f(s) = \sqrt{s}$, and $f(s) = s$, the three particular cases already recovered above. It applies, more generally, to any monomial function $f(s) = Bs^{1/n}$, with $n > 0$ and B a positive constant. An important conclusion is that $f(s)$ must be a crescent function that does not grow faster than s. The proposition is also useful for excluding some other cases. For instance, the so-called U-curve [31, 32], which consists in finding λ that minimizes the cost function

$$J_U(\lambda) = \frac{1}{U(\mathbf{X}(\lambda))} + \frac{1}{T(\mathbf{X}(\lambda))}, \tag{65}$$

corresponds to $f(s) = \exp(-1/s)$; this expression does not satisfy the necessary condition of the proposition.

5 Extensions

This section addresses two possible extensions of the Bayesian L-curve beyond the Gaussian assumptions introduced so far for the priors of $\mathbf{Y}|\mathbf{X}$ and $\mathbf{X}$. The first one considers the case of generalized Gaussians priors and the second one the case of complex Gaussians.

5.1 The generalized Gaussian case

So far, the paper has considered that both $\mathbf{Y}|\mathbf{X}$ and $\mathbf{X}$ are Gaussian distributed, undoubtedly addressing a working assumption of practical importance. As mentioned in section 2.2, the same methodology applies to other distributions, and in particular when $\mathbf{Y}|\mathbf{X}$ and/or $\mathbf{X}$ are distributed according to generalized Gaussians. The case is now briefly investigated where $\mathbf{X}$ a priori follows the generalized Gaussian $\mathcal{GG}(\mathbf{X}; a_X, b_X)$ with shape and rate parameters $a_X > 0$ and $b_X > 0$, that is

$$p(\mathbf{X}) = \prod_{i=1}^{P} \prod_{k=1}^{K} \frac{b_X}{2a_X \Gamma(1/b_X)} e^{-\left(\frac{|x_{ki}|}{a_X}\right)^{b_X}} = \left(\frac{b_X}{2\Gamma(1/b_X)}\right)^{KP} \frac{e^{-a_X^{-b_X} \sum_{i=1}^{P} \sum_{k=1}^{K} |x_{ki}|^{b_X}}}{a_X^{KP}}. \tag{66}$$

Setting $\tau_X = b_X a_X^{-b_X}$ and $s = \sum_{i=1}^{P}\sum_{k=1}^{K}|x_{ki}|^b \doteq ||\mathbf{X}||_{b_X}^{b_X}$ and assigning it an arbitrary prior probability density $\phi_X(\tau)$, one has the Φ-transform

$$\int_{\mathbb{R}^+} p(\mathbf{X}|\tau)\phi_X(\tau)d\tau = \Phi_X(\|\mathbf{X}\|_{b_X}^{b_X}/b_X; KP/b). \tag{67}$$

The solution to problem P2 is then the solution $\hat{\mathbf{X}}$ of the equation

$$\mathbf{A}^\top\mathbf{A}\mathbf{X} + \lambda(\mathbf{X})\nabla_{\mathbf{X}}\|\mathbf{X}\|_{b_X}^{b_X} = \mathbf{A}^\top\mathbf{Y}, \tag{68}$$

where matrix $\nabla_{\mathbf{X}} = [\partial/\partial\mathbf{x}_1, \ldots, \partial/\partial\mathbf{x}_P]$ contains in its *i-th* column the gradient with respect to $\mathbf{x}_i$, the i-th column of $\mathbf{X}$, with regularization parameter

$$\lambda(\mathbf{X}) \;=\; \frac{\mathbb{E}\{\tau_N|T(\mathbf{X})\}}{\mathbb{E}\{\tau_X|\|\mathbf{X}\|_{b_X}^{b_X}\}} \quad \text{with} \quad \mathbb{E}\{\tau_X|\|\mathbf{X}\|_{b_X}^{b_X}\} = \frac{\Phi_X(\|\mathbf{X}(\lambda(t))\|_{b_X}^{b_X}/b_X; KP/b_X+1)}{\Phi_X(\|\mathbf{X}(\lambda(t))\|_{b_X}^{b_X}/b_X; KP/b_X)}. \tag{69}$$

Apart from the resolution of Eq. (68), which no longer has a closed-form solution (e.g. as Eq. (16b)), all results remain unchanged, yet with the potential functions substituted by the norm $\|\mathbf{X}\|_{b_X}^{b_X}$ and KP reduced by the factor b_X. For instance, with a gamma prior on τ_X, the corresponding branch of the L-curve is

$$\eta(t) = (\alpha_X + KP/b)\ln(\beta_X + \|\mathbf{X}(\lambda(t))\|_{b_X}^{b_X}/b_X). \tag{70}$$

The use of such norms in regularization was for instance explored in Ref. [6].

5.2 Complex-valued data

The introduced framework is also readily extended to complex valued-data $\mathbf{Y} \in \mathbb{C}^{M\times P}$, $\mathbf{X} \in \mathbb{C}^{K\times P}$, $\mathbf{A} \in \mathbb{C}^{M\times K}$. In this case, the probability density of $\mathbf{Y}|\mathbf{X}$ and the prior probability densities of $\mathbf{X}$ are taken as the complex Gaussians

$$p(\mathbf{Y}|\mathbf{X},\tau_N) = \left(\frac{\tau_N}{\pi}\right)^{MP}\exp\left(-\tau_N T(\mathbf{X})\right) \quad \text{and} \quad p(\mathbf{X}|\tau_X) = \left(\frac{\tau_X}{\pi}\right)^{KP}\exp\left(-\tau_X U(\mathbf{X})\right) \tag{71}$$

(with a slight abuse of notation since $T(\mathbf{X})$ must also be a function of $\mathbf{Y}$ in the above equation). One has the Φ-transforms $\int_{\mathbb{R}^+} p(\mathbf{N}|\tau)\phi_N(\tau)d\tau = \Phi_N(T(\mathbf{X}); MP)$ and $\int_{\mathbb{R}^+} p(\mathbf{X}|\tau)\phi_X(\tau)d\tau = \Phi_X(U(\mathbf{X}); KP)$, so that all results remain unchanged, yet without the one-half factor on T, U, MP and KP. The solution to Problem 2 is

$$\hat{\mathbf{X}} \;=\; \left(\mathbf{A}^{\mathsf{H}}\mathbf{A} + \lambda(\hat{\mathbf{X}})\mathbf{I}\right)^{-1}\mathbf{A}^{\mathsf{H}}\mathbf{Y} \tag{72}$$

where $^{\mathsf{H}}$ stands for the transpose conjugate operator and with the regularization parameter returned by

$$\lambda(\mathbf{X}) \;=\; \frac{\Phi_N(T(\mathbf{X}); MP+1)}{\Phi_N(T(\mathbf{X}); MP)}\frac{\Phi_X(U(\mathbf{X}); KP)}{\Phi_X(U(\mathbf{X}); KP+1)}. \tag{73}$$

For instance, the L-curve with gamma priors is then defined as

$$\gamma_\Gamma : t \mapsto \left\{\begin{array}{ll}\zeta(t) & = (\alpha_N + MP)\ln(\beta_N + T(\mathbf{X}(\lambda(t))))\\ \eta(t) & = (\alpha_X + KP)\ln(\beta_X + U(\mathbf{X}(\lambda(t)))).\end{array}\right. \tag{74}$$

6 Numerical experiments

This last section illustrates the ideas elaborated in the paper by means of numerical experiments. The aim is to apply the Bayesian L-curve to some of the test problems collected in Ref. [20] and to investigate the performance of the four cost functions κ, J_1, J_2, and J_3. The so-called test problems "`baart`", "`shaw`", and "`ilaplace`" are selected, as they are representative of the results also obtained on other problems. For the three problems, a square matrix $\mathbf{A}$ is used with $M = K = 20$ and one set of observations ($P = 1$). Undetermined configurations with ($M = K/2$ and $M = K/3$) were also tested, yet the results are not reported here, as they happened to show very similar trends as for the determined case. Additive white Gaussian noise was added with noise-to-signal ratios (NSR) varying from -20dB to -2dB; this was repeated for 300 independent random draws in order to conduct Monte Carlo analyses. Results for the gamma L-curve only are reported here, with non-informative hyperparameters $\alpha_N = \alpha_X = 0.1$ and $\beta_N = \beta_X = 10^{-16}$. Intensive experiments showed that there were no major differences with the inverse-gamma and square-root (used with the update rule (64)) L-curves, as long as the respective priors were taken flat enough. In all experiments, the curve parameter $t = \exp(\lambda)$ was used. Finally, since the test problems come with the true data $\mathbf{X}$, the reference regularization parameter λ_{opt} was computed as the minimizer of the square error $J_{SE}(\lambda) = \|\mathbf{X}(\lambda) - \mathbf{X}\|_F^2$.

Figures 6, 8, and 10 display the cost functions J_{SE}, κ, J_1 and J_2 for the three test problems, and Figs. 7, 9, and 11 display the corresponding boxplots for the estimated regularization parameters (now including J_3) and the normalized square errors $\|\mathbf{X}(\lambda) - \mathbf{X}\|_F^2 / \|\mathbf{X}\|_F^2$. The following observations are in order.

Overall, all criteria demonstrate very similar trends. There seems to be a slight tendency of the cost function J_2 to return a regularization parameter with a larger variance than for the other cost functions, yet it does not necessarily result in a higher or a more dispersed NMSE. While the cost functions κ, J_2 and J_3 – which all involve derivatives – have a tendency to fluctuate with respect to λ, the cost function J_1 evidences a smoother evolution, close to a convex curve. The similarity of the curve J_1 with the square error J_{SE} is striking.

Two regimes of noise are to be distinguished. For a significant amount of noise (say NSR above -10dB), all cost functions return estimates of the regularization parameters that are statistically very similar, and actually very close to the reference λ_{opt} (their boxplots most often overlap). Inspection of the L-curves actually reveals that in noisy cases the "L" shape is generally well-marked, with a clear corner. These are instances where Propositions 3.6 and 3.7 apply (the L-curve is locally symmetric around its bisector axis), thus implying that all estimates are equal.

For a low amount of noise (say NSR below -10dB), the test problems can be classified according to two behaviors. In a first category, all estimates of the regularization parameters decrease proportionally with the NSR together with λ_{opt} (see Figs. 7 and 9) and the NMSE also decreases with the noise level. A second category is where the reference λ_{opt} is lower-bounded, but the estimates of the regularization parameters keep on decreasing below it (see Fig. 11). As a consequence, the NSME no longer decreases monotonically when the noise level goes to zero. This behavior has been described in Ref. [18] as an instance where the L-curve may fail. Inspection of the latter actually reveals that its vertical branch tends to shrink, and thus its corner to disappear.

The situation can be fixed by making the prior more informative. In order to avoid the regularization parameter to collapse, one can increase the value of the hyperparameter β_N to enforce smaller values of the precision τ_N, compliant with a small NSR. This can also be understood as forcing the

vertical branch of the L-curve to protrude. This is illustrated here for the test problem "`ilaplace`". The hyperparameter β_N is set to 10^{-3} (this setting was found robust enough, in the sense that it could be changed by at least two orders of magnitude without significantly altering the results). Figure 12 shows that the estimated regularization parameters now closely follows λ_{opt} for small noise levels, thus considerably reducing the NMSE.

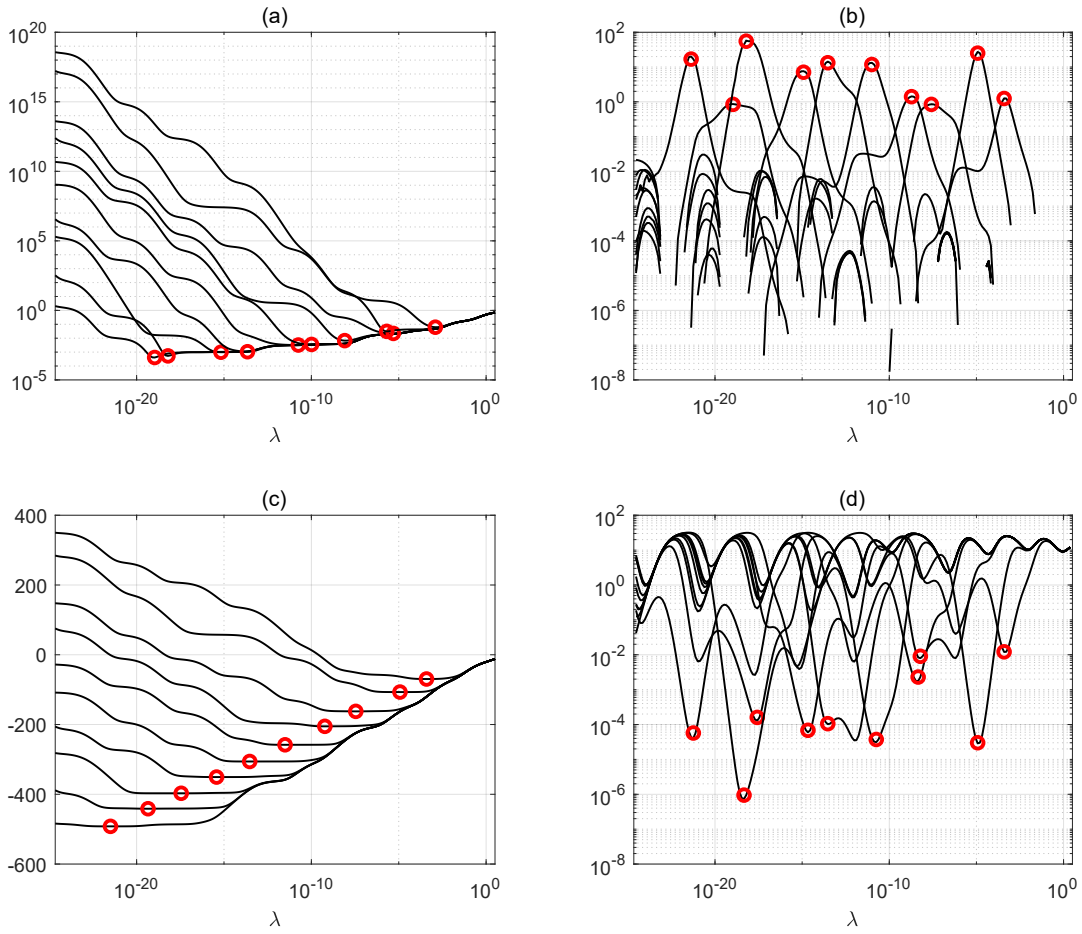

Figure 6: Test problem `baart` for SNRs ranging from -20dB to -2dB. a) square error J_{SE}, b) curvature κ, c) cost function J_1, and d) cost function J_2 with respect to regularization parameter λ. Optima in each case are marked by red circles.

7 Conclusion

The aim of the paper was to provide a Bayesian interpretation of the L-curve used to solve linear inverse problems. When introducing prior distributions and treating their precisions (inverses of the variances) as nuisance parameters, the MMAP solution naturally offsprings an L-curve together with a criterion for locating its corner. The solution is in some respects similar to Reginska's minimum product criterion, but contrary to the latter it is not empirical, and it comes with a rich framework. It is general, in the sense that different priors will generate different types of L-curves. The classical linear, logarithmic and square-root L-curves are recovered as particular cases, and the reference (logarithmic) L-curve is also the asymptote of the "Bayesian L-curve" when the effect of the priors vanishes. Among other benefits of the framework, the Bayesian L-curve explicitly accounts for the dimensions

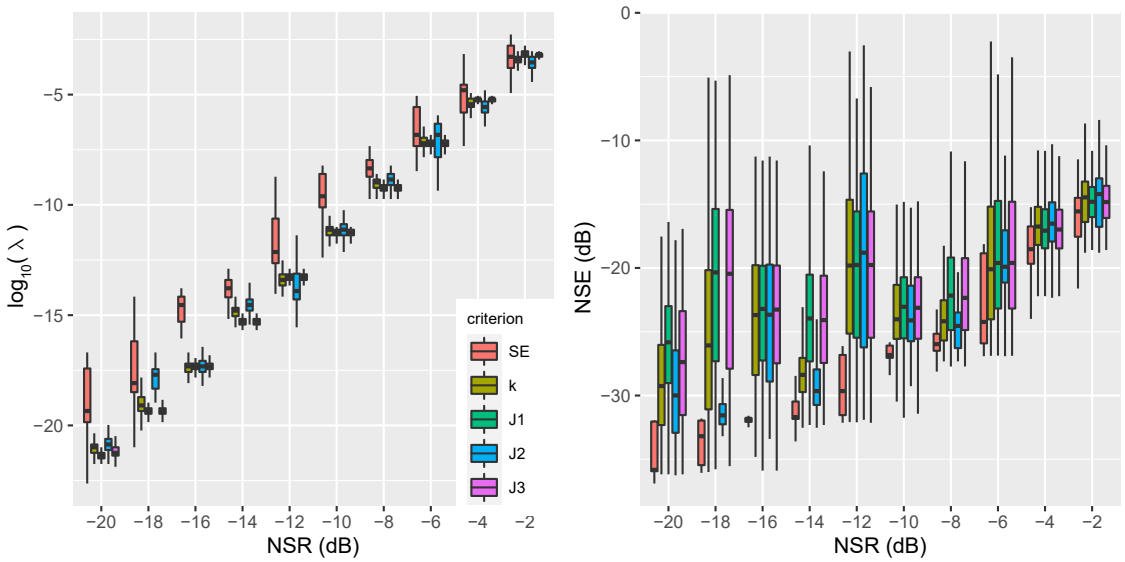

Figure 7: Test problem `baart`. Boxplots of estimated regularization parameters (left) and of corresponding normalized square error (right).

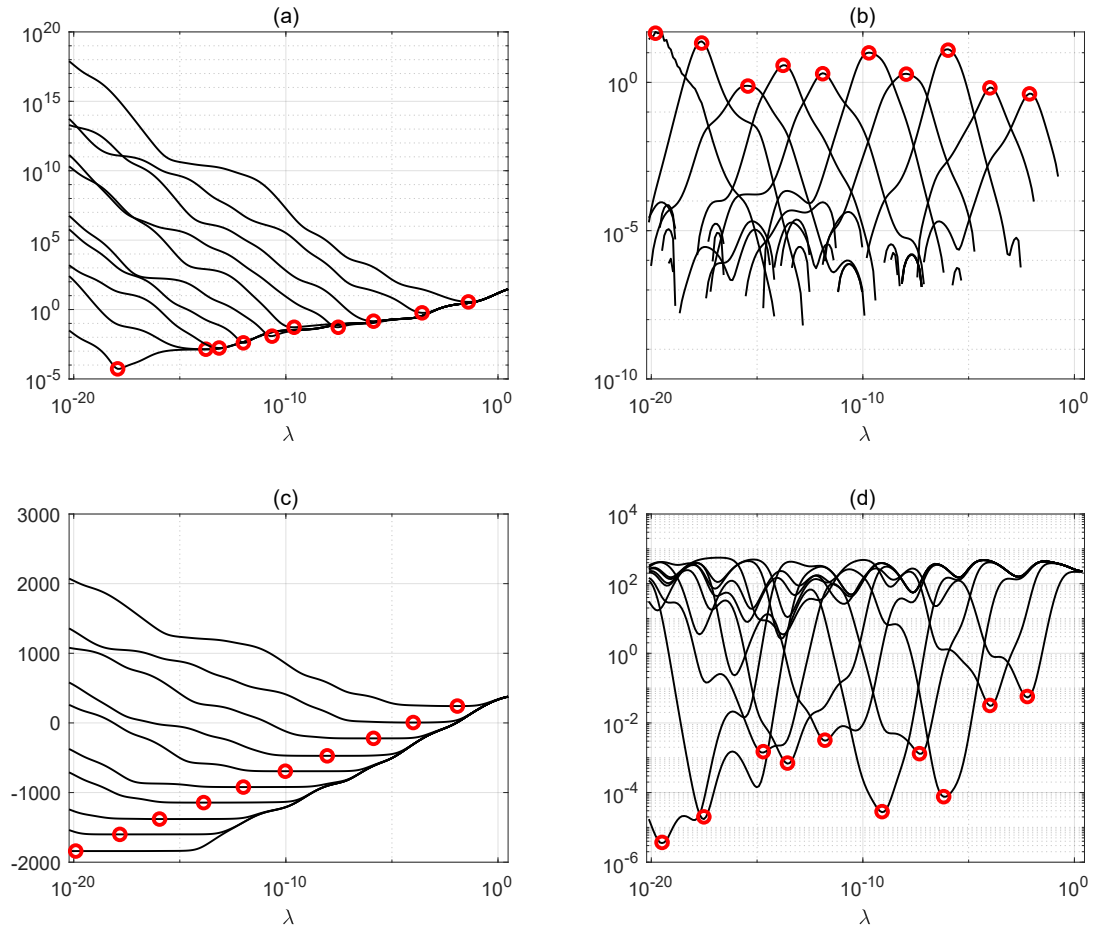

Figure 8: Test problem `shaw` for SNRs ranging from -20dB to -2dB. a) square error J_{SE}, b) curvature κ, c) cost function J_1, and d) cost function J_2 with respect to regularization parameter λ. Optima in each case are marked by red circles.

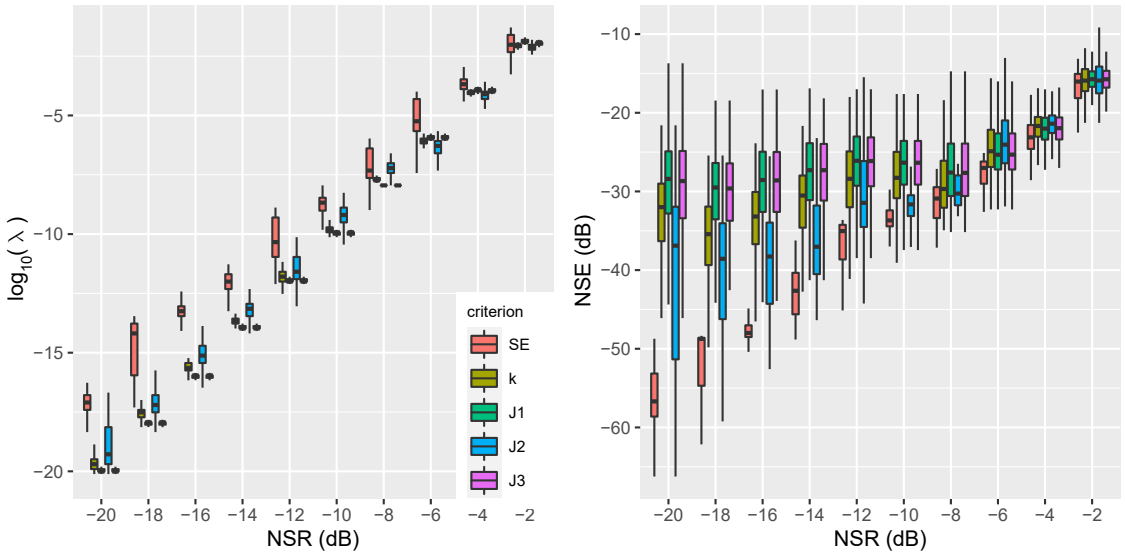

Figure 9: Test problem `shaw`. Boxplots of estimated regularization parameters (left) and of corresponding normalized square error (right).

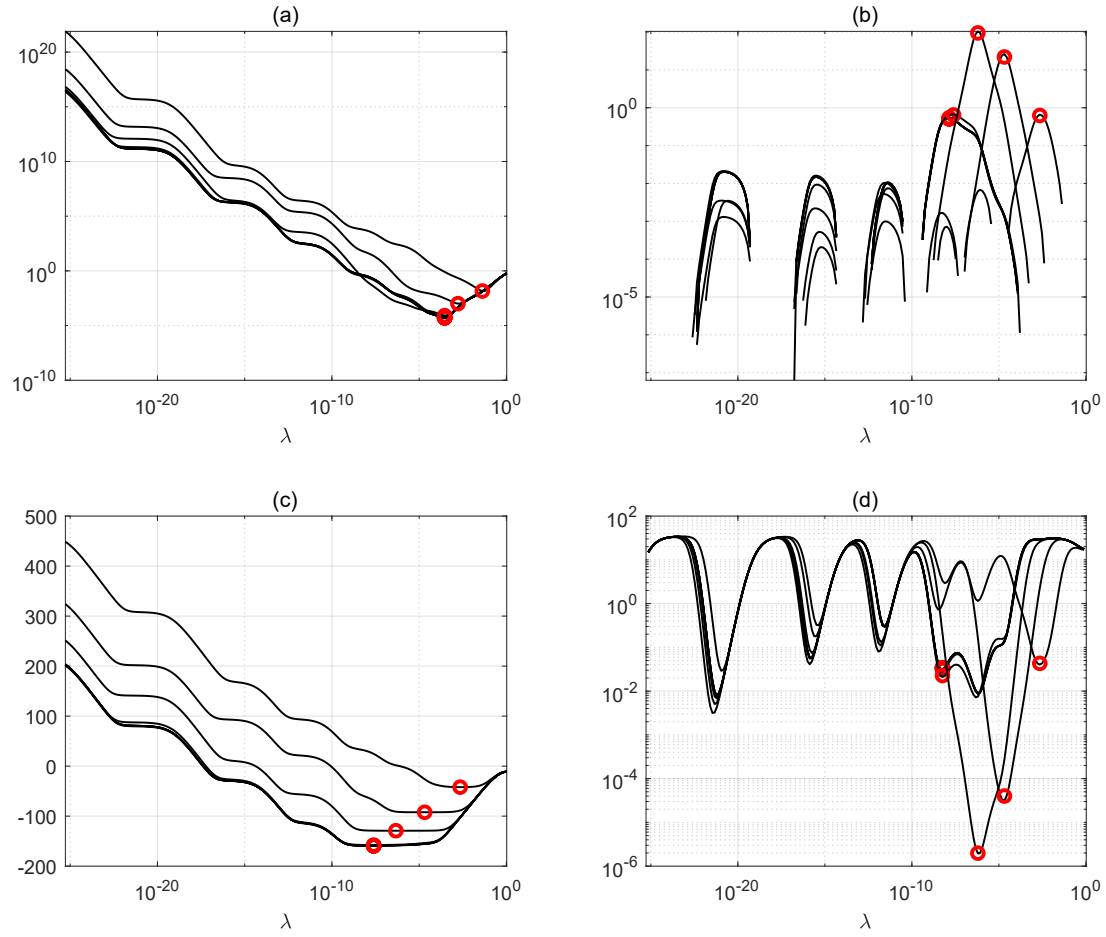

Figure 10: Test problem `ilaplace` for SNRs ranging from -20dB to -2dB. a) square error J_{SE}, b) curvature κ, c) cost function J_1, and d) cost function J_2 with respect to regularization parameter λ. Optima in each case are marked by red circles.

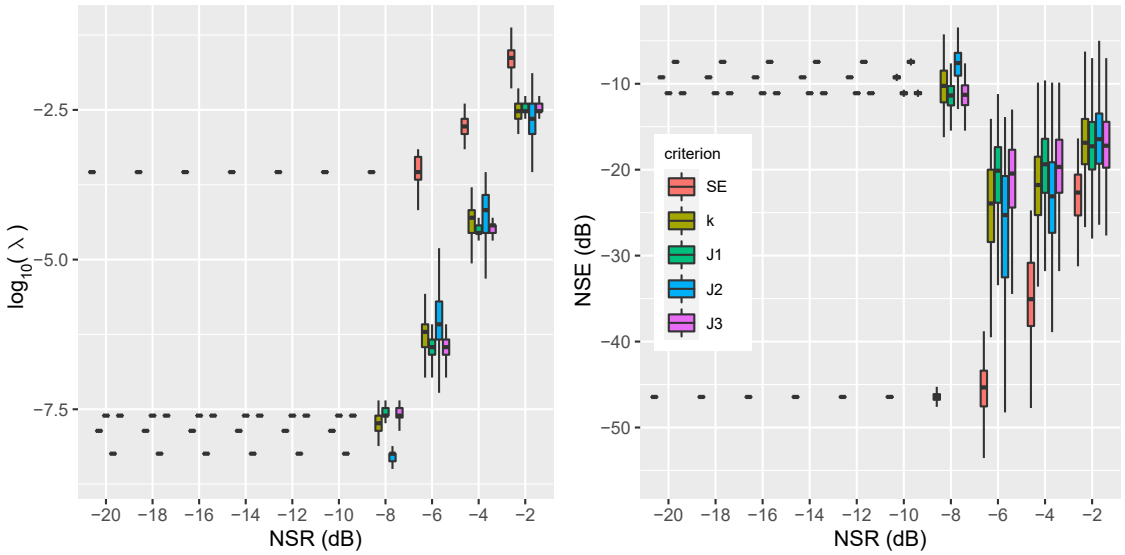

Figure 11: Test problem `ilaplace`. Boxplots of estimated regularization parameters (left) and of corresponding normalized square error (right).

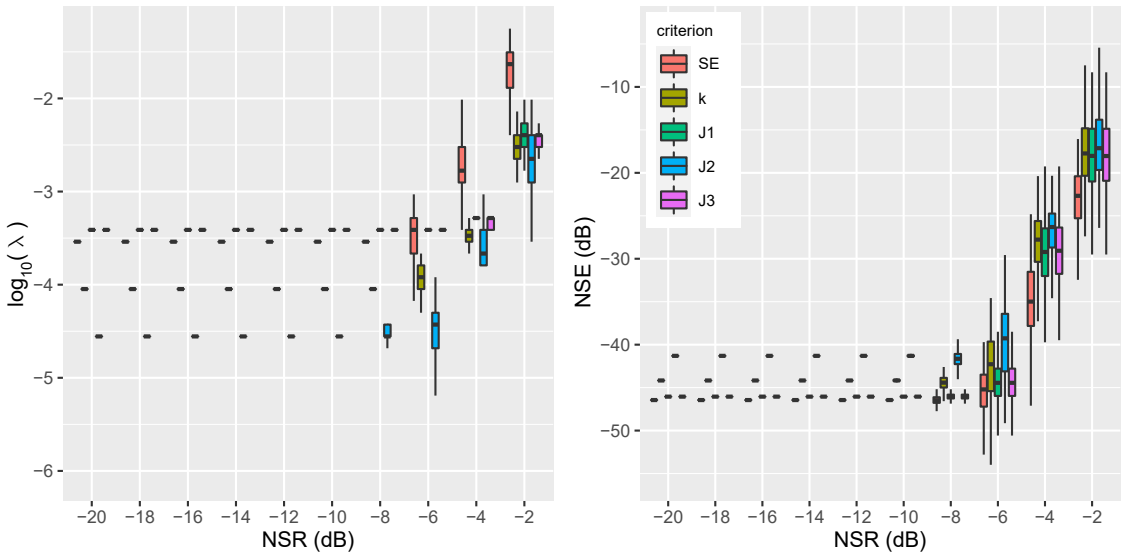

Figure 12: Test problem `ilaplace` with $\beta_N = 10^{-3}$ (as opposed to $\beta_N = 10^{-6}$ in Fig. 11). Boxplots of estimated regularization parameters (left) and of corresponding normalized square error (right).

of the inverse problem (i.e. number of observations versus number of unknowns) in its regularization. Analysis of the L-curve properties have shown that several competing criteria may be devised to locate its corner. To the authors' knowledge, the "minimum speed on the curve" and the "maximum angular speed" have been introduced here for the first time. Conditions have been established under which the criteria – including the maximum curvature – all return the same solution. This was validated by numerical experiments. The equivalence of the criteria happens to depend on the curve parameter – a monotonic function of the regularization parameter – whose principled selection is a matter of future research. Numerical experiments have also shown that a proper selection of the priors lead to a Bayesian L-curve that can succeed where the reference L-curve is known to fail. Further research is needed to provide guidelines in this direction.

Acknowledgment

This work was performed on occasion of a delegation offered by the CNRS (French National Research Center) to the first author and in the framework of the LABEX CeLyA (ANR-10-LABX-0060) of Université de Lyon, within the program « Investissements d'Avenir » (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

Appendices

A Preliminary results

One will need the following results.

Proposition A.1. *Let $\Phi_\tau(s;n)$ be the Φ-transform of the probability density $\phi_\tau(t)$ of a random variable τ. It then holds that*

i) $-\frac{d}{ds}\ln\Phi_\tau(s;n) = \mathbb{E}\{\tau|s\}$

ii) $\frac{d^2}{ds^2}\ln\Phi_\tau(s;n) = \mathbb{V}\{\tau|s\} = -\frac{d}{ds}\mathbb{E}\{\tau|s\}$

where $\mathbb{E}\{\tau|s\}$ and $\mathbb{V}\{\tau|s\}$ stands for the expected value and variance of τ conditioned on s.

Proof.

$$\frac{d}{ds}\ln\Phi_\tau(s;n) = \frac{\Phi'_\tau(s;n)}{\Phi_\tau(s;n)} = -\frac{\int_{\mathbb{R}^+}\tau e^{-s\tau}\tau^n\phi_\tau(\tau)d\tau}{\int_{\mathbb{R}^+}e^{-s\tau}\tau^n\phi_\tau(\tau)d\tau} = -\mathbb{E}\{\tau|s\}. \tag{75}$$

Similarly,

$$\frac{d^2}{ds^2}\ln\Phi_\tau(s;n) = \frac{\Phi''_\tau(s;n)}{\Phi_\tau(s;n)} - \left(\frac{\phi'_\tau(s)}{\phi_\tau(s)}\right)^2 = \mathbb{E}\{\tau^2|s\} - \mathbb{E}\{\tau|s\}^2 = \mathbb{V}\{\tau|s\}. \tag{76}$$

□

Proposition A.2.

i) The potentials $T(\mathbf{X}(\lambda))$ and $U(\mathbf{X}(\lambda))$ are monotonically increasing and decreasing functions of the regularization parameter, i.e. $\frac{d}{d\lambda}T(\mathbf{X}(\lambda)) \geq 0$ and $\frac{d}{d\lambda}U(\mathbf{X}(\lambda)) \leq 0$.

ii) The rate of change of $T(\mathbf{X}(\lambda))$ and $-U(\mathbf{X}(\lambda))$ are in the proportion of the regularization parameter, i.e. $dT(\mathbf{X}(\lambda)) = -\lambda \cdot dU(\mathbf{X}(\lambda))$.

iii) The second-order differentials of $T(\mathbf{X}(\lambda))$ and $U(\mathbf{X}(\lambda))$ are related as

$$d^2T(\mathbf{X}(\lambda)) + \lambda \cdot d^2U(\mathbf{X}(\lambda)) = -d\lambda \cdot dU(\mathbf{X}(\lambda)) = \frac{d\lambda}{\lambda}dT(\mathbf{X}(\lambda)). \tag{77}$$

The proof of (i) and (ii) is for instance found in lemma 2 of Reginska [45]. Property (iii) immediately follows from (ii).

B Proof of Proposition 2.2

The maximum of $\Phi_N(T(\mathbf{X});M)\Phi_X(U(\mathbf{X});K)$ is found by equating its gradient with respect to $\mathbf{x}_i$ to zero, for each i. This yields

$$\nabla_{\mathbf{x}_i}\Phi_N(T(\mathbf{X})/2;M/2)\Phi_X(U(\mathbf{X})/2;K/2)+\Phi_N(T(\mathbf{X})/2;M/2)\nabla_{\mathbf{x}_i}\Phi_X(U(\mathbf{X})/2;K/2)=0. \tag{78}$$

Next, differentiating under the integral sign,

$$\begin{aligned}&\Phi_X(U(\mathbf{X})/2;K/2)\int\mathbf{A}^\top(\mathbf{y}_i-\mathbf{A}\mathbf{x}_i)\tau p(\mathbf{Y}|\mathbf{X},\tau)d\tau-\Phi_N(T(\mathbf{X})/2;M/2)\int\mathbf{x}_i\tau p(\mathbf{X}|\tau)d\tau=0\\ \Leftrightarrow\quad&\mathbf{A}^\top(\mathbf{y}_i-\mathbf{A}\mathbf{x}_i)\Phi_N(T(\mathbf{X})/2;M/2+1)\Phi_X(U(\mathbf{X})/2;K/2)-\Phi_N(T(\mathbf{X})/2;M/2)\Phi_X(U(\mathbf{X})/2;K/2+1)=0,\end{aligned}$$

from which (16b) immediately follows.

C An asymptotic result of the Φ-transform

Let $s\sim O(n)$. If $\phi(\tau)$ is a continuous function independent of n, it holds that

$$\lim_{n\to\infty}\left|\Phi(s;n)-n!\phi(n/s)s^{-(n+1)}\right|=0. \tag{79}$$

Proof. The sketch of the proof is as follows. As $n\to\infty$, the function $\exp(-s\tau)\tau^n$ becomes more and more peaked around $\tau=n/s$. Hence, in the integral (15), $\phi(\tau)$ can be considered almost constant in the neighborhood of $\tau=n/s$, which gives

$$\Phi(n;s)\simeq\phi(n/s)\int_{\mathbb{R}^+}e^{-s\tau}\tau^n d\tau=\phi(n/s)\frac{n!}{s^{n+1}}. \tag{80}$$

□

D Proof of Proposition 3.3

The first derivative of the branch ζ is

$$\zeta'=-\frac{dT}{dt}\frac{d}{dT}\ln\Phi_N=T'\mathbb{E}\{\tau_N|T\}. \tag{81}$$

It is next noted that

$$\frac{d}{dt}\mathbb{E}\{\tau_N|T\}=\frac{dT}{dt}\frac{d}{dT}\mathbb{E}\{\tau_N|T\}=-T'\mathbb{V}\{\tau_N|T\},$$

where the last equality results from (ii) of Proposition A.1. Therefore, the second derivative of the branch ζ is $\zeta''=T''\mathbb{E}\{\tau_N|T\}-T'^2\mathbb{V}\{\tau_N|T\}$. It is similarly found that $\eta'=U'\mathbb{E}\{\tau_X|U\}$ and $\eta''=U''\mathbb{E}\{\tau_X|U\}-U'^2\mathbb{V}\{\tau_X|U\}$. Adding the two quantities,

$$J_1''=\zeta''+\eta''=T''\mathbb{E}\{\tau_N|T\}+U''\mathbb{E}\{\tau_X|U\}-T'^2\mathbb{V}\{\tau_N|T\}-U'^2\mathbb{V}\{\tau_X|U\}. \tag{82}$$

The next step is to express J_1'' in terms of T and its derivative only. Using the relationship $\mathbb{E}\{\tau_X|U\} = \lambda\mathbb{E}\{\tau_N|T\}$ (Proposition 2.2) and properties (ii) and (iii) of Proposition A.2,

$$\begin{aligned} J_1'' &= \mathbb{E}\{\tau_N|T\}(T'' + \lambda U'') - T'^2\mathbb{V}\{\tau_N|T\} - U'^2\mathbb{V}\{\tau_X|U\} && (83)\\ &= \mathbb{E}\{\tau_N|T\}T'(\lambda/\lambda') - T'^2\mathbb{V}\{\tau_N|T\} - (T'^2/\lambda^2)\mathbb{V}\{\tau_X|U\} && (84)\end{aligned}$$

where $\mathbb{E}\{\tau_X|U\} = \lambda\mathbb{E}\{\tau_N|T\}$ (Eq. (18) Proposition 2.2) was used in the second line and properties (ii) and (iii) Proposition A.2 in the third line. Taking $(\mathbb{E}\{\tau_N|T\}T')^2$ as a factor and using Eq. (18) one again, one arrives at

$$J_1'' = (\mathbb{E}\{\tau_N|T\}T')^2\left(\frac{\lambda'}{\lambda\mathbb{E}\{\tau_N|T\}T'} - CV_N - CV_X\right) \tag{85}$$

with $CV_N = \mathbb{V}\{\tau_N|T\}/\mathbb{E}\{\tau_N|T\}^2$ and $CV_X = \mathbb{V}\{\tau_X|U\}/\mathbb{E}\{\tau_X|U\}^2$. Setting $\mathbb{E}\{\tau_N|T\}T' = \zeta'$ as given by Eq. (81) then yields Eq. (29a). Similarly, expressing J_1'' in terms of U and its derivative only gives Eq. (29b).

E Proof of Proposition 3.7

It is readily verified that, under the conditions (34a)-(34c), $J_2'(t^*) = 0$ and $J_2''(t^*) = 4\zeta''(t^*)^2 \geq 0$.

F Condition for $\kappa'' \geq 0$

Straightforward calculation gives the condition

$$4\zeta'\zeta''\zeta''' + 3\zeta''^3 \geq \zeta'^2(\zeta'''' + \eta'''') \tag{86}$$

at $t = t^*$.

G Missing information matrix

For the sake of simplicity, the proof is given here as if only one column, say $\mathbf{x}$, of matrix $\mathbf{X}$ was to be recovered from the EM algorithm of section 3.2. The missing information matrix defined as $\mathbf{I}_m = \mathbf{I}_c - \mathbf{I}_i$, where $\mathbf{I}_c = -\mathbb{E}_{\tau_X,\tau_N}\{\nabla^2_{\mathbf{x}} \ln p(\mathbf{X}|\mathbf{Y},\tau_X,\tau_N)\}$ and $\mathbf{I}_i = \nabla^2_{\mathbf{x}} \ln p(\mathbf{X}|\mathbf{Y})$ are the "complete-data" and "incomplete-data" observed information matrices [37], $\nabla^2_{\mathbf{x}} = \partial^2/\partial\mathbf{x}\partial\mathbf{x}^\top$ stands for the second derivative of a function with respect to vectors $\mathbf{x}$ and $\mathbf{x}^\top$, and $\mathbb{E}_{\tau_X,\tau_N}$ stands for the expected value with respect to τ_X and τ_N conditioned on $\mathbf{Y}$. Calculations give

$$\begin{aligned} \mathbf{I}_c =& -\mathbb{E}_{\tau_X,\tau_N}\{\nabla^2_{\mathbf{x}}(\ln p(\mathbf{Y}|\mathbf{X},\tau_N) + \ln p(\mathbf{X}|\tau_X))\} = -\mathbb{E}_{\tau_X,\tau_N}\{\nabla^2_{\mathbf{x}}(\tau_N T(\mathbf{X}) + \tau_X U(\mathbf{X}))\} && (87)\\ =& \mathbb{E}\{\tau_N|T\}\nabla^2_{\mathbf{x}}T(\mathbf{X}) + \mathbb{E}\{\tau_X|U\}\nabla^2_{\mathbf{x}}U(\mathbf{X}) && (88)\end{aligned}$$

and

$$\begin{aligned} \mathbf{I}_i =& -\nabla^2_{\mathbf{x}} \ln(\Phi_N(T(\mathbf{X}))\Phi_X(U(\mathbf{X}))) && (89)\\ =& -\nabla^2_{\mathbf{x}}T(\mathbf{X}) \cdot \frac{d}{dT}\ln\Phi_N(T(\mathbf{X})) - \nabla_{\mathbf{x}}T(\mathbf{X})\nabla_{\mathbf{X}^\top}T(\mathbf{X}) \cdot \frac{d^2}{dT^2}\ln\Phi_N(T(\mathbf{X})) && (90)\\ & -\nabla^2_{\mathbf{x}}U(\mathbf{X}) \cdot \frac{d}{dU}\ln\Phi_X(U(\mathbf{X})) - \nabla_{\mathbf{x}}U(\mathbf{X})\nabla_{\mathbf{X}^\top}U(\mathbf{X}) \cdot \frac{d^2}{dU^2}\ln\Phi_X(U(\mathbf{X})), && (91)\end{aligned}$$

where $\nabla_{\mathbf{x}} = \partial/\partial\mathbf{x}$ is the gradient of a function with respect to $\mathbf{x}$. Using properties (i) and (ii) of Proposition A.1, this is equal to

$$\begin{aligned}\mathbf{I}_i &= \mathbb{E}\{\tau_X|T\}\nabla_{\mathbf{x}}^2 T(\mathbf{X}) - \mathbb{V}(\tau_N|T)\nabla_{\mathbf{x}} T(\mathbf{X})\nabla_{\mathbf{x}^\top} T(\mathbf{X}) + \mathbb{E}\{\tau_N|U\}\nabla_{\mathbf{x}}^2 U(\mathbf{X}) - \mathbb{V}(\tau_X|U)\nabla_{\mathbf{x}} U(\mathbf{X})\nabla_{\mathbf{x}^\top} U(\mathbf{X}) \\ &= \mathbf{I}_c - \mathbb{V}(\tau_N|T)\nabla_{\mathbf{x}} T(\mathbf{X})\nabla_{\mathbf{x}^\top} T(\mathbf{X}) - \mathbb{V}(\tau_X|U)\nabla_{\mathbf{x}} U(\mathbf{X})\nabla_{\mathbf{x}^\top} U(\mathbf{X}).\end{aligned} \tag{92}$$

(93)

Therefore,

$$\mathbf{I}_m = \mathbb{V}(\tau_N|T)\nabla_{\mathbf{x}} T(\mathbf{X})\nabla_{\mathbf{x}^\top} T(\mathbf{X}) + \mathbb{V}(\tau_X|U)\nabla_{\mathbf{x}} U(\mathbf{X})\nabla_{\mathbf{x}^\top} U(\mathbf{X}). \tag{94}$$

Now, using the chain rule $\nabla_{\mathbf{x}} = \frac{\partial\lambda}{\partial\mathbf{x}}\frac{\partial t}{\partial\lambda}\frac{\partial}{\partial t}$, the latter equation becomes

$$\mathbf{I}_m = \frac{\partial\lambda}{\partial\mathbf{x}}\frac{\partial\lambda}{\partial\mathbf{x}^\top}\frac{1}{\lambda'^2}\left(\mathbb{V}(\tau_N|T)\left(\frac{dT}{dt}\right)^2 + \mathbb{V}(\tau_X|U)\left(\frac{dU}{dt}\right)^2\right). \tag{95}$$

Finally, let us observe that $T' = \zeta'/\mathbb{E}\{\tau_N|T\}$ from Eq. (81). Therefore, in combination with $T' + \lambda U' = 0$ and $\lambda = \mathbb{E}\{\tau_X|U\}/\mathbb{E}\{\tau_N|T\}$, it also comes $U' = -\zeta'/\mathbb{E}\{\tau_X|U\}$. Substituting T' and U' for these expressions then yields Eq. (43).

H Simulation of the L-curve

The L-curves illustrated in Figs. 4 and 5 were simulated based on model (1) with $\mathbf{A}$ expressed by its singular value decomposition, $\mathbf{A} = \mathbf{WSV}$, where $\mathbf{W} \in \mathbb{C}^{M\times M}$ and $\mathbf{V} \in \mathbb{C}^{K\times K}$ are random unitary matrices distributed according to a Bingham distribution [30] and $\mathbf{S}$ has non-zero elements $s_k = \exp(-\eta(k-1)/(\min(M,K)-1))$, $k = 1, ..., \min(M,K)$, $\eta > 0$ only on its main diagonal. The elements of the additive errors $\mathbf{N}$ and of the dependent variable $\mathbf{X}$ were generated as independent and identically distributed Gaussian variables with variances $1/\tau_N$ and $1/\tau_X$, respectively.

Assuming $M \geq K$, the Tikhonov solution reads

$$\mathbf{X}(\lambda) = \mathbf{VS}^2(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^\top\mathbf{X} + \mathbf{VS}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{N} \tag{96}$$

and the mean-square error is

$$\begin{aligned}\mathbb{E}\|\mathbf{X} - \mathbf{X}(\lambda)\|_F^2 &= \mathbb{E}\|\mathbf{V}\lambda(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^\top\mathbf{X}\|_F^2 + \mathbb{E}\|\mathbf{VS}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{N}\|_F^2 \\ &= \lambda^2\tau_X^{-1}\mathrm{trace}\{(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\} + \tau_N^{-1}\mathrm{trace}\{\mathbf{S}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\} \\ &= \sum_{k=1}^{K}\frac{\lambda^2\tau_X^{-1} + s_k^2\tau_N^{-1}}{(s_k^2 + \lambda)^2}.\end{aligned} \tag{97}$$

References

[1] A. Abubaker and P. M. van den Berg. Total variation as a multiplicative constraint for solving inverse problems. *IEEE Transactions on Image Processing*, 10(9):1384–1392, 2001.

[2] U. A. S. Aguirre, M. Ceberio, and V. Kreinovich. Why curvature in L-curve: Combining soft constraints. In M. Ceberio and V. Kreinovich, editors, *Constraint Programming and Decision Making*, pages 175–179. Springer International Publishing, Cham, 2014.

[3] J. Antoni. A Bayesian approach to sound source reconstruction: Optimal basis, regularization, and focusing. *The Journal of the Acoustical Society of America*, 131(4):2873–2890, apr 2012.

[4] M. Aucejo and O. D. Smet. A multiplicative regularization for force reconstruction. *Mechanical Systems and Signal Processing*, 85:730–745, feb 2017.

[5] M. Aucejo and O. D. Smet. An iterated multiplicative regularization for force reconstruction problems. *Journal of Sound and Vibration*, 437:16–28, dec 2018.

[6] M. Aucejo and O. D. Smet. A generalized multiplicative regularization for input estimation. *Mechanical Systems and Signal Processing*, 157:107637, aug 2021.

[7] F. Bauer and M. A. Lukas. Comparing parameter choice methods for regularization of ill-posed problems. *Mathematics and Computers in Simulation*, 81(9):1795–1841, may 2011.

[8] F. S. V. Bazán. Fixed-point iterations in determining the Tikhonov regularization parameter. *Inverse Problems*, 24(3):035001, apr 2008.

[9] F. S. Bazán, J. Francisco, K. H. Leem, and G. Pelekanos. A maximum product criterion as a Tikhonov parameter choice rule for Kirsch's factorization method. *Journal of Computational and Applied Mathematics*, 236(17):4264–4275, nov 2012.

[10] D. Calvetti, G. H. Golub, and L. Reichel. Estimation of the L-curve via Lanczos bidiagonalization. *Bit Numerical Mathematics*, 39(4):603–619, 1999.

[11] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari. Tikhonov regularization and the L-curve for large discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, 123(1-2):423–446, nov 2000.

[12] D. Calvetti, L. Reichel, and A. Shuibi. L-curve and curvature bounds for Tikhonov regularization. *Numerical Algorithms*, 35(2-4):301–314, apr 2004.

[13] A. Cultrera and L. Callegaro. A simple algorithm to find the L-curve corner in the regularisation of ill-posed inverse problems. *IOP SciNotes*, 1(2):025004, aug 2020.

[14] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Springer Netherlands, July 1996.

[15] M. S. Gockenbach and E. Gorgin. On the convergence of a heuristic parameter choice rule for Tikhonov regularization. *SIAM Journal on Scientific Computing*, 40(4):A2694–A2719, jan 2018.

[16] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, may 1979.

[17] M. Gulliksson and P.-Å. Wedin. Optimization tools for Tikhonov regularization of nonlinear equations using the L-curve and its dual. In *Methods and Applications of Inversion*, pages 155–170. Springer-Verlag, 2000.

[18] M. Hanke. Limitations of the L-curve method in ill-posed problems. *BIT Numerical Mathematics*, 36(2):287–301, jun 1996.

[19] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, dec 1992.

[20] P. C. Hansen. REGULARIZATION TOOLS: A matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6(1):1–35, mar 1994.

[21] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. Society for Industrial and Applied Mathematics, jan 1998.

[22] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, nov 1993.

[23] J. Idier. *Bayesian Approach to Inverse Problems*. ISTE LTD, June 2008.

[24] K. Ito, B. Jin, and J. Zou. A new choice rule for regularization parameters in Tikhonov regularization. *Applicable Analysis*, 90(10):1521–1544, oct 2011.

[25] B. Jin and J. Zou. Augmented Tikhonov regularization. *Inverse Problems*, 25(2):025001, dec 2008.

[26] B. Jin and J. Zou. A Bayesian inference approach to the ill-posed cauchy problem of steady-state heat conduction. *International Journal for Numerical Methods in Engineering*, 76(4):521–544, oct 2008.

[27] P. R. Johnston and R. M. Gulrajani. An analysis of the zero-crossing method for choosing regularization parameters. *SIAM Journal on Scientific Computing*, 24(2):428–442, jan 2002.

[28] P. Johnston and R. Gulrajani. A new method for regularization parameter determination in the inverse problem of electrocardiography. *IEEE Transactions on Biomedical Engineering*, 44(1):19–39, 1997.

[29] P. Johnston and R. Gulrajani. Selecting the corner in the L-curve approach to Tikhonov regularization. *IEEE Transactions on Biomedical Engineering*, 47(9):1293–1296, 2000.

[30] J. T. Kent, A. M. Ganeiber, and K. V. Mardia. A new method to simulate the Bingham and related distributions in directional data analysis with applications, 2013.

[31] D. Krawczyk-Stańdo and M. Rudnicki. Regularization parameter selection in discrete ill-posed problems — the use of the U-curve. *International Journal of Applied Mathematics and Computer Science*, 17(2):157–164, jun 2007.

[32] D. Krawczyk-Stańdo and M. Rudnicki. The use of L-curve and U-curve in inverse electromagnetic modelling. In *Studies in Computational Intelligence*, pages 73–82. Springer Berlin Heidelberg, 2008.

[33] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, jan 1995.

[34] S. Lu and P. Mathé. Heuristic parameter selection based on functional minimization: Optimality and model function approach. *Mathematics of Computation*, 82(283):1609–1630, feb 2013.

[35] D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, jul 1999.

[36] P. J. Mc Carthy. Direct analytic model of the L-curve for Tikhonov regularization parameter selection. *Inverse Problems*, 19(3):643–663, apr 2003.

[37] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions, 2E.* John Wiley & Sons, Inc., feb 2008.

[38] M. Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, jan 2005.

[39] A. Mohammad-Djafari. A full Bayesian approach for inverse problems. In *Maximum Entropy and Bayesian Methods*, pages 135–144. Springer Netherlands, 1996.

[40] R. Molina, A. Katsaggelos, and J. Mateos. Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Transactions on Image Processing*, 8(2):231–246, 1999.

[41] V. A. Morozov. On the solution of functional equations by the method of regularization. *Doklady Mathematics*, 7:414–417, 1966.

[42] K. Okamoto and B. Q. Li. Optimal numerical methods for choosing an optimal regularization parameter. *Numerical Heat Transfer, Part B: Fundamentals*, 51(6):515–533, apr 2007.

[43] A. Pereira, J. Antoni, and Q. Leclère. Empirical bayesian regularization of the inverse acoustic problem. *Applied Acoustics*, 97:11–29, oct 2015.

[44] T. Raus and U. Hämarik. Heuristic parameter choice in Tikhonov method from minimizers of the quasi-optimality function. In *Trends in Mathematics*, pages 227–244. Springer International Publishing, 2018.

[45] T. Regińska. A regularization parameter in discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 17(3):740–749, may 1996.

[46] C. P. Robert. *The Bayesian choice: a decision-theoretic motivation.* Springer-Verlag, 1994.

[47] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, may 2010.

[48] P. M. van den Berg, A. L. van Broekhoven, and A. Abubakar. Extended contrast source inversion. *Inverse Problems*, 15(5):1325–1344, oct 1999.

[49] P. M. van den Berg, A. Abubakar, and J. T. Fokkema. Multiplicative regularization for contrast profile inversion. *Radio Science*, 38(2):n/a–n/a, apr 2003.

[50] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667, sep 1977.

[51] D. V. Widder. *Laplace transform (PMS-6).* Princeton university press, 2015.

[52] R. A. Willoughby. Solutions of ill-posed problems (A. N. Tikhonov and V. Y. Arsenin). *SIAM Review*, 21(2):266–267, apr 1979.

[53] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1), mar 1983.

[54] L. Wu. A parameter choice method for Tikhonov regularization. *Electronic Transactions on Numerical Analysis*, pages 107–128, 2003.

[55] L. Yan, F.-L. Yang, and C.-L. Fu. A new numerical method for the inverse source problem from a Bayesian perspective. *International Journal for Numerical Methods in Engineering*, 85(11):1460–1474, sep 2010.