論文 / 著書情報
Article / Book Information

Title	Accelerating cross-validation with total variation and its application to super-resolution imaging
Authors	Tomoyuki Obuchi, Shiro Ikeda, Kazunori Akiyama, Yoshiyuki Kabashima
Citation	PLoS ONE, 12, 12, e0188012(1-14)
Pub. date	2017, 12
Creative Commons	Information is in the article.

PLOS ONE

RESEARCH ARTICLE

Accelerating cross-validation with total variation and its application to super-resolution imaging

Tomoyuki Obuchi[1]*, Shiro Ikeda[2], Kazunori Akiyama[3,4,5], Yoshiyuki Kabashima[1]

1 Department of Mathematical and Computing Science/Tokyo Institute of Technology, Yokohama 226-8502, Japan, **2** The Institute of Statistical Mathematics, Tachikawa, Tokyo, 190-8562, Japan, **3** Haystack Observatory/Massachusetts Institute of Technology, Westford, MA, 01886, United States of America, **4** National Astronomy Observatory of Japan, Osawa 2-21-1, Mitaka, Tokyo 181-8588, Japan, **5** Black Hole Initiative, Harvard University, Cambridge, MA, 02138, United States of America

* obuchi@c.titech.ac.jp

Citation: Obuchi T, Ikeda S, Akiyama K, Kabashima Y (2017) Accelerating cross-validation with total variation and its application to super-resolution imaging. PLoS ONE 12(12): e0188012. https://doi.org/10.1371/journal.pone.0188012

Editor: Yuanquan Wang, Beijing University of Technology, CHINA

Received: December 10, 2016

Accepted: October 30, 2017

Published: December 7, 2017

Data Availability Statement: All image data used in the paper are available from http://vlbiimaging.csail.mit.edu/imagingchallenge.

Funding: This work was supported by Japan Society for the Promotion of Science 25120008 to Prof. Shiro Ikeda; 26870185 to Dr. Tomoyuki Obuchi; 25120013 to Prof. Yoshiyuki Kabashima; 17H00764 to Dr. Tomoyuki Obuchi and Prof. Yoshiyuki Kabashima; Research Abroad Program to Dr. Kazunori Akiyama; and National Science Foundation AST-1614868 to Dr. Kazunori Akiyama.

Abstract

We develop an approximation formula for the cross-validation error (CVE) of a sparse linear regression penalized by ℓ_1-norm and total variation terms, which is based on a perturbative expansion utilizing the largeness of both the data dimensionality and the model. The developed formula allows us to reduce the necessary computational cost of the CVE evaluation significantly. The practicality of the formula is tested through application to simulated black-hole image reconstruction on the event-horizon scale with super resolution. The results demonstrate that our approximation reproduces the CVE values obtained via literally conducted cross-validation with reasonably good precision.

1 Introduction

At present, in many practical situations of science and technology, large high-dimensional observational datasets are created and accumulated on a continuous basis. An essential difficulty concerning the treatment of such high-dimensional data is the extraction of meaningful information. Sparse modeling [1, 2] is a promising framework for overcoming this difficulty, which has recently been utilized in many disciplines [3, 4]. In this framework, a statistical or machine-learning model with a large number of parameters (explanatory variables) is fitted to the data, in conjunction with a certain sparsity-inducing penalty. This penalty should be appropriately chosen with consideration of the processed data. One representative penalty is the ℓ_1 regularization, which retains certain preferred properties, such as the statistical model convexity [5, 6]. A similar penalty that has received more recent focus is the so-called "total variation (TV)" [7–9], which can be regarded as the ℓ_1 regularization imposed on the difference between neighboring explanatory variables. The TV yields "continuity" of the neighboring variables, which is suitable for the processing of certain datasets expected to have such continuity, such as natural images [4, 7–9].

Another common difficulty associated with the use of statistical models is model selection. In the context of image processing using the ℓ_1 and TV regularizations, this difficulty appears

Competing interests: The authors have declared that no competing interests exist.

during the selection of appropriate regularization weights. A practical framework to select these weights, which is applicable to general situations, is cross-validation (CV). CV provides an estimator of the statistical-model generalization error, *i.e.*, the CV error (CVE), using the data under control, and the minimum CVE obtained when sweeping the weights yields the optimal weight values. This versatile framework is, however, computationally demanding for large datasets/models, and this problem frequently becomes a bottleneck affecting model selection. Thus, reducing the CVE computational cost could have a significant impact on a broad range of sparse modeling applications in various disciplines.

Considering these circumstances, in this paper, we provide a CVE approximation formula for a statistical model of linear regression penalized by the ℓ_1 and TV terms, to efficiently reduce the computational cost. Note that the formula for the case penalized by the ℓ_1 term alone has already been proposed in [10], and the formula presented herein is a generalization of it. Below, we show the formula derivation and perform a demonstration in the context of super-resolution imaging. The processed images employed in this study are reconstructed from simulated observations of black holes on the event-horizon scale for the Event Horizon Telescope (EHT, see [11–13]) full array. Note that our formula will be applied to actual EHT observations to be conducted after April 2017.

2 Problem setting

Let us suppose that our measurement is a linear process, and denote the measurement result as $\boldsymbol{y} \in \mathbb{R}^M$ and the measurement matrix as $A = \{A_{\mu i}\}_{\mu=1,\cdots,M;\ i=1,\cdots N} \in \mathbb{R}^{M\times N}$. The explanatory variables, corresponding to the images that will be examined in the later demonstration, are denoted by $\boldsymbol{x} \in \mathbb{R}^N$. The quality of the fit to the data is described by the residual sum of squares (RSS), *i.e.*, $\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}, A) = \frac{1}{2}||\boldsymbol{y} - A\boldsymbol{x}||_2^2$. In addition, we consider the following penalty consisting of ℓ_1 and TV terms:

$$R(\boldsymbol{x}; \lambda_{\ell_1}, \lambda_T) = \lambda_{\ell_1}||\boldsymbol{x}||_1 + \lambda_T T(\boldsymbol{x}), \tag{1}$$

where the $T(\boldsymbol{x})$ term corresponds to the TV and is expressed as

$$T(\boldsymbol{x}) = \sum_i \sqrt{\sum_{j\in\partial i}(x_j - x_i)^2} \equiv \sum_i t_i, \tag{2}$$

and ∂i denotes the neighboring variables of the ith variable. There is some variation in the definition of "neighbors"; here, we follow the standard approach [7–9]. That is, $\boldsymbol{x}$ is assumed to be a two-dimensional image and the neighbors of the ith pixel correspond to the right and down pixels. However, the bottom row (the rightmost column) of the image is exceptional, as the neighbor of each pixel in that row (column) corresponds to the right (down) pixel only. Note that the developed approximation formula presented below is independent of this specific choice of neighbors and can be applied to general cases.

For this setup, we consider the following linear regression problem with the penalty given in Eq (1)

$$\hat{\boldsymbol{x}}(\lambda_{\ell_1}, \lambda_T) = \arg\min_{\boldsymbol{x}}\{\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}, A) + R(\boldsymbol{x}; \lambda_{\ell_1}, \lambda_T)\}, \tag{3}$$

where $\arg\min_u\{f(u)\}$ generally represents the argument that minimizes an arbitrary function

$f(u)$. Further, we consider the leave-one-out (LOO) CV of Eq (3) in the form

$$\hat{\boldsymbol{x}}^{\backslash\mu}(\lambda_{\ell_1},\lambda_T) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\sum_{\nu(\neq\mu)} (y_\nu - A_{\nu i}x_i)^2 + R \right\}$$
$$\equiv \arg\min_{\boldsymbol{x}} \{\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}^{\backslash\mu}, A^{\backslash\mu}) + R(\boldsymbol{x};\lambda_{\ell_1},\lambda_T)\}. \quad (4)$$

Note that the system without the μth row of $\boldsymbol{y}$ and A is referred to as the "μth LOO system" hereafter. In this procedure, the CVE, *i.e.*, the generalization error estimator, is

$$\mathcal{E}_{\mathrm{LOO}}(\lambda_{\ell_1},\lambda_T) = \frac{1}{2}\sum_{\mu=1}^{M} (y_\mu - \boldsymbol{a}_\mu^{\mathrm{T}}\hat{\boldsymbol{x}}^{\backslash\mu}(\lambda_{\ell_1},\lambda_T))^2, \quad (5)$$

where $\boldsymbol{a}_\mu^\top = (A_{\mu 1},\cdots,A_{\mu N})$ is the μth row vector of A. We term this simply the "LOO error (LOOE)."

Computing the LOOE requires solution of Eq (4) M times, by definition, which is computationally expensive. Therefore, the purpose of this paper is to avoid this computational expense by deriving an approximation formula of Eq (5).

3 Approximation formula for softened system

When M is sufficiently large, *i.e.*, the number of observations is large enourgh, the difference between the LOO solution $\hat{\boldsymbol{x}}^{\backslash\mu}$ and the full solution $\hat{\boldsymbol{x}}$ is expected to be small. This intuition naturally motivates us to conduct a perturbation connecting these two solutions. To conduct this perturbation, we "soften" the penalty by introducing a small cutoff $\delta(>0)$ in the TV, having the form

$$R \to R^\delta(\boldsymbol{x};\lambda_{\ell_1},\lambda_T) = \lambda_{\ell_1}\sum_i ||\boldsymbol{x}|| + \lambda_T T^\delta(\boldsymbol{x}), \quad (6)$$

where

$$T^\delta = \sum_i \sqrt{\sum_{j\in\partial i}(x_j - x_i)^2 + \delta^2} \equiv \sum_i t_i^\delta. \quad (7)$$

An approximation formula in the presence of ℓ_1 regularization with smooth cost functions has already been proposed in [10]. We employ that formula here and take the limit $\delta \to 0$.

To state the approximation formula, we begin by defining "active" and "killed" variables. Owing to the ℓ_1 term, some variables are set to zero in $\hat{\boldsymbol{x}}$; we refer to these variables as "killed variables." The remaining finite variables are termed "active variables." We denote the index sets of the active and killed variables by S_A and S_K, respectively. The active (killed) components of a vector $\boldsymbol{x}$ are formally expressed as $\boldsymbol{x}_{S_A}(\boldsymbol{x}_{S_K})$. For any matrix X, we use double subscripts in the same manner. For example, for an $N\times N$ matrix, a submatrix having row and column components of S_A and S_K, respectively, is denoted by $X_{S_A\,SK}$.

The approximation formula can be derived through the following two steps. Note that, in this derivation, a crucial assumption is that the sets of active and killed variables are common among the full and LOO systems. This assumption may not hold exactly in practice, but the resultant formula is asymptotically exact in the large-N limit [10].

The first step is to compute the values of the active variables and their response to small perturbation. The active variables are determined by the extremization condition of the softened

cost function with respect to the active variables, such that

$$\frac{\partial(\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}^{\setminus\mu}, A^{\setminus\mu}) + R^{\delta}(\boldsymbol{x}; \lambda_{\ell_1}, \lambda_T))}{\partial \boldsymbol{x}_{S_A}} = \boldsymbol{0} \Rightarrow (\hat{\boldsymbol{x}}^{\delta\setminus\mu})_{S_A}. \tag{8}$$

The focus here is the response of this solution when a small perturbation $-\boldsymbol{h}\cdot\boldsymbol{x}$ is incorporated into the cost function. A simple computation demonstrates that the active–active components of the response function, $(\chi^{\delta\setminus\mu})_{S_A S_A} = \frac{\partial}{\partial \boldsymbol{h}_{S_A}} (\hat{\boldsymbol{x}}^{\delta\setminus\mu})_{S_A}|_{\boldsymbol{h}=0}$, are equivalent to the inverse of the cost-function Hessian

$$(\chi^{\delta\setminus\mu})_{S_A S_A} = (H^{\delta\setminus\mu}_{S_A S_A})^{-1}, \tag{9}$$

$$H^{\delta\setminus\mu} = \partial^2_{\boldsymbol{x}}(\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}^{\setminus\mu}, A^{\setminus\mu}) + R^{\delta}(\boldsymbol{x}; \lambda_{\ell_1}, \lambda_T)) = G^{\setminus\mu} + \partial^2_{\boldsymbol{x}} R^{\delta}(\boldsymbol{x}; \lambda_{\ell_1}, \lambda_T), \tag{10}$$

where $\partial^2_{\boldsymbol{x}}$ denotes the Hessian operator $\partial^2_{\boldsymbol{x}} \equiv \left(\frac{\partial^2}{\partial x_i \partial x_j}\right)$ and $G^{\setminus\mu}$ is the Gram matrix of $A^{\setminus\mu}$, *i.e.*, $G^{\setminus\mu} \equiv (A^{\setminus\mu})^{\top} A^{\setminus\mu}$. The other components of the response function are identically zero, from the stability assumption of S_K and because the killed variables are zero, with $\hat{\boldsymbol{x}}_{S_K} = \hat{\boldsymbol{x}}^{\setminus\mu}_{S_K} = \boldsymbol{0}$.

In the second step, we connect the full solution to the LOO solution, through the above perturbation with an appropriate $\boldsymbol{h}$. To specify the perturbation, we assume that the difference $\boldsymbol{d}^{\delta} = \hat{\boldsymbol{x}}^{\delta} - \hat{\boldsymbol{x}}^{\delta\setminus\mu}$ is small and expand the RSS of the full system with respect to $\boldsymbol{d}^{\delta}$ as follows:

$$\mathcal{E}(\hat{\boldsymbol{x}}^{\delta\setminus\mu}|\boldsymbol{y}, A) \approx \mathcal{E}(\hat{\boldsymbol{x}}^{\delta}|\boldsymbol{y}, A) - \sum_{\mu=1}^{M} (y_\mu - \boldsymbol{a}_\mu^{\top}\hat{\boldsymbol{x}}^{\delta})\boldsymbol{a}_\mu^{\top}\boldsymbol{d}^{\delta}. \tag{11}$$

This equation implies that the perturbation between the full and LOO systems can be expressed as $\boldsymbol{h}_\mu = (y_\mu - \boldsymbol{a}_\mu^{\top}\hat{\boldsymbol{x}}^{\delta})\boldsymbol{a}_\mu$. Hence, we obtain

$$\hat{\boldsymbol{x}}^{\delta} \approx \hat{\boldsymbol{x}}^{\delta\setminus\mu} + \chi^{\delta\setminus\mu}\boldsymbol{h}_\mu = \hat{\boldsymbol{x}}^{\delta\setminus\mu} + (y_\mu - \boldsymbol{a}_\mu^{\top}\hat{\boldsymbol{x}}^{\delta})\chi^{\delta\setminus\mu}\boldsymbol{a}_\mu. \tag{12}$$

The Hessian of the full system has a simple relationship with the LOO Hessian, such that

$$H^{\delta} \equiv G^{\setminus\mu} + (\boldsymbol{a}_\mu\boldsymbol{a}_\mu^{\top}) + \partial^2_{\boldsymbol{x}} R^{\delta}(\hat{\boldsymbol{x}}^{\delta}) \approx H^{\delta\setminus\mu} + (\boldsymbol{a}_\mu\boldsymbol{a}_\mu^{\top}), \tag{13}$$

where the approximation at the righthand side comes from replacing $\hat{\boldsymbol{x}}^{\delta}$ with $\hat{\boldsymbol{x}}^{\delta\setminus\mu}$ in the argument of $R^{\delta}(\boldsymbol{x})$. Inserting Eqs (12 and 13) in conjunction with $\chi^{\delta}_{S_A S_A} = (H^{\delta}_{S_A S_A})^{-1}$ into Eq (5) and using the Sherman-Morrison formula for matrix inversion, we find

$$\mathcal{E}_{\mathrm{LOO}}(\lambda_{\ell_1}, \lambda_T) \approx \frac{1}{2}\sum_{\mu=1}^{M} \frac{(y_\mu - \boldsymbol{a}_\mu^{\top}\hat{\boldsymbol{x}}^{\delta})^2}{(1 - \boldsymbol{a}^{\top}_{\mu S_A}(\chi^{\delta})_{S_A S_A}\boldsymbol{a}_{\mu S_A})^2}. \tag{14}$$

According to Eq (14), we can compute the LOOE only from the full solution $\hat{\boldsymbol{x}}^{\delta}$, without actually performing CV, which facilitates considerable reduction of the computational cost.

4 Handling a singularity

Let us generalize Eq (14) to the limit $\delta \to 0$, where the penalty contains another singular term in addition to the ℓ_1 term. This TV singularity tends to "lock" some of the neighboring variables, *i.e.*, $x_j = x_i$ ($\forall j \in \partial i$), which corresponds to $t_i = 0$ in Eq (2). If two different vanishing TV terms, t_i and t_j, share a common variable x_r, all the variables in those TV terms take the same value $x_k = x_r$ ($\forall k \in (\{i\} \cup \partial i \cup \{j\} \cup \partial j)$). In this manner, the active variables are separated into several "locked" clusters, with all the variables inside a cluster having an identical value. This

implies that the variable response to a perturbation, $\chi = \lim_{\delta\to 0}\chi^{\delta}$, should have the same value for all variables in a cluster and may, therefore, be merged. Below, we demonstrate this behavior for the $\delta \to 0$ limit. For the derivation, we assume that the clusters are common to both the full and LOO systems, similar to the assumption for S_A and S_K. For convenience, we index the clusters by $\alpha, \beta \in C$ and denote the number of clusters by $|C|$; the index set of variables in a cluster α is represented by S_α and the total set of indices in all clusters is denoted by $S_C \equiv \cup_\alpha S_\alpha$. Hereafter, we concentrate on the active variable space only and omit the killed variable space. The complement set of S_C, *i.e.*, the set of isolated variables that do not belong to any cluster, is denoted by S_I and, thus, $S_A = S_I \cup S_C$.

Two crucial observations for the derivation are the "scale separation" and the presence of the "zero mode." For vanishing TV terms, a natural scaling to satisfy $\lim_{\delta\to 0} t_i^{\delta} = t_i = 0$ is $|\hat{x}_j^{\delta} - \hat{x}_i^{\delta}| \propto \delta\ (\forall j \in \partial i)$. Once this scaling is assumed, we realize that the components of the Hessian that are directly related to the clusters diverge. Let us define by $\hat{S}_\alpha$ the set of TV terms corresponding to cluster α, *i.e.*, $\hat{S}_\alpha = \{i|(\{i\} \cup \partial i) \subset S_\alpha\}$. Hence, by construction and for all $\alpha \in C$, all components of $D_\alpha^{\delta} \equiv \lambda_T(\partial_{\boldsymbol{x}}^2 \sum_{i\in\hat{S}_\alpha} t_i^{\delta})_{S_\alpha S_\alpha}$ are scaled as $1/\delta$ and, thus, diverge as $\delta \to 0$. The remaining terms are retained as $O(1)$. According to this "scale separation," we decompose the Hessian as $H^{\delta} = D^{\delta} + F^{\delta}$, where D^{δ} is the direct sum of the diverging components in the naively extended space; $D^{\delta} = \oplus_\alpha D_\alpha^{\delta}$; and F^{δ} consists of the remaining $O(1)$ terms. This decomposition can be schematically expressed as

$$H^{\delta} = D^{\delta} + F^{\delta} = \left(\begin{array}{ccc|c} D_1^{\delta} & 0 & 0 & \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & D_{|C|}^{\delta} & \\ \hline & 0 & & 0 \end{array}\right) + \left(\begin{array}{c|c} F_{S_C S_C}^{\delta} & F_{S_C S_I}^{\delta} \\ \hline F_{S_I S_C}^{\delta} & F_{S_I S_I}^{\delta} \end{array}\right). \tag{15}$$

We denote the basis of the current expression by $\{\boldsymbol{e}_i\}_{i\in S_A}$, with $(\boldsymbol{e}_i)_j = \delta_{ij}$, and move to another basis that diagonalizes $D_{S_C S_C}^{\delta}$. Each D_α^{δ} has a "zero mode," and its normalized eigenvector is given by $\boldsymbol{z}_\alpha = (z_{i\alpha})$, where $z_{i\alpha} = 1/\sqrt{|S_\alpha|}$ for $i \in S_\alpha$ and 0 otherwise, in the full space. This behavior originates from the symmetry, such that the $\{t_i^{\delta}\}_{i\in\hat{S}_\alpha}$ are invariant under a uniform shift in the S_α sub-space, *i.e.*, $x_j \to x_j + \Delta\ (\forall j \in S_\alpha)$ for $\forall\Delta \in \mathbb{R}$. This invariance can also be directly seen from a property of the Hessian, *i.e.*, $\frac{\partial^2}{\partial x_i^2} t_i^{\delta} + \sum_{j\in\partial i} \frac{\partial}{\partial x_i \partial x_j} t_i^{\delta} = 0$.

In addition, we represent the set of normalized eigenvectors of all the other modes of D_α^{δ}, which have eigenvalues $\lambda_{\alpha a}$ that are proportional to $1/\delta$ and positively divergent, as $\{\boldsymbol{u}_{\alpha a}\}_{a=1}^{|S_\alpha|-1}$. Then, $\{\{\{\boldsymbol{u}_{\alpha a}\}_a, \boldsymbol{z}_\alpha\}_\alpha\}$ diagonalizes $D_{S_C S_C}^{\delta}$ and $\{\{\{\boldsymbol{u}_{\alpha a}\}_a, \boldsymbol{z}_\alpha\}_\alpha, \{\boldsymbol{e}_i\}_{i\in S_I}\}$ constitutes an orthonormal basis of the full space. Corresponding to this variable change, we denote $\hat{\boldsymbol{S}}_Z$, $\hat{\boldsymbol{S}}_{I+Z}$, and $\hat{\boldsymbol{S}}_{C-Z}$ as the index set of variables in the space spanned by $\{\boldsymbol{z}_\alpha\}_\alpha$, $\{\{\boldsymbol{z}_\alpha\}_\alpha, \{\boldsymbol{e}_i\}_{i\in S_I}$, and $\{\boldsymbol{u}_{\alpha a}\}_{\alpha,a}$, respectively. In the new expression, we can rewrite $H^{\delta} = D^{\delta} + F^{\delta}$ as

$$H^{\delta} = \left(\begin{array}{c|c} D_{\tilde{\boldsymbol{s}}_{C-Z}\tilde{\boldsymbol{s}}_{C-Z}}^{\delta} & 0 \\ \hline 0 & 0 \end{array}\right) + \left(\begin{array}{c|c} F_{\tilde{\boldsymbol{s}}_{C-Z}\tilde{\boldsymbol{s}}_{C-Z}}^{\delta} & F_{\tilde{\boldsymbol{s}}_{C-Z}\tilde{\boldsymbol{s}}_{I+Z}}^{\delta} \\ \hline F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{C-Z}}^{\delta} & F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}}^{\delta} \end{array}\right), \tag{16}$$

where $D_{\tilde{\boldsymbol{s}}_{C-Z}\tilde{\boldsymbol{s}}_{C-Z}}^{\delta} = \mathrm{diag}(\{\lambda_{\alpha a}\}_{\alpha,a})$. Because of the divergence of $D_{\tilde{\boldsymbol{s}}_{C-Z}\tilde{\boldsymbol{s}}_{C-Z}}^{\delta}$, only $F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}}^{\delta}$ is

relevant for the evaluation of $(H^{\delta})^{-1}$. These considerations yield the explicit formula of χ as

$$(H^{\delta})^{-1} = \left(\begin{array}{c|c} (D^{\delta}_{\tilde{\boldsymbol{s}}_{C-Z}\tilde{\boldsymbol{s}}_{C-Z}})^{-1} & 0 \\ \hline 0 & (F^{\delta}_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}})^{-1} \end{array}\right) + O(\delta) \to \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & (F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}})^{-1} \end{array}\right) = \chi, \tag{17}$$

where $F = \lim_{\delta\to 0} F^{\delta}$.

By construction, in the reduced space to span $(\{\boldsymbol{z}_{\alpha}\}_{\alpha}, \{\boldsymbol{e}_i\}_{i\in S_I})$, $F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}}$ can be expressed as

$$F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}} = \sum_{\alpha,\beta}(F_{\alpha\beta}\boldsymbol{z}_{\alpha}\boldsymbol{z}_{\beta}^{\top} + F_{\beta\alpha}\boldsymbol{z}_{\beta}\boldsymbol{z}_{\alpha}^{\top}) + \sum_{\alpha}\sum_{i\in S_I}(F_{\alpha i}\boldsymbol{z}_{\alpha}\boldsymbol{e}_i^{\top} + F_{i\alpha}\boldsymbol{e}_i\boldsymbol{z}_{\alpha}^{\top}) + \sum_{i,j\in S_I}F_{ij}\boldsymbol{e}_i\boldsymbol{e}_j^{\top}. \tag{18}$$

As the non-zero components of the zero mode $\boldsymbol{z}_{\alpha}$ are identically given as $1/\sqrt{|S_{\alpha}|}$, all these coefficients can be easily expressed by the original coefficients F_{ij}, as

$$F_{\alpha\beta} = \boldsymbol{z}_{\alpha}^{\top}F\boldsymbol{z}_{\beta} = \frac{1}{\sqrt{|S_{\alpha}||S_{\beta}|}}\sum_{i\in S_{\alpha}, j\in S_{\beta}}F_{ij}, \tag{19a}$$

$$F_{\alpha i} = \boldsymbol{z}_{\alpha}^{\top}F\boldsymbol{e}_i = \frac{1}{\sqrt{|S_{\alpha}|}}\sum_{j\in S_{\alpha}}F_{ji}, \tag{19b}$$

and $F_{i\alpha} = F_{\alpha i}$ by the symmetry. Now, all the components are explicitly specified. The form of χ in the original basis $\{\boldsymbol{e}_i\}_{i\in S_A}$ can be accordingly assessed by moving back from the basis $\{\{\{\boldsymbol{u}_{\alpha a}\}_a, \boldsymbol{z}_{\alpha}\}_{\alpha}, \{\boldsymbol{e}_i\}_{i\in S_I}\}$, which completes the computation.

Some additional consideration of the above computation demonstrates that we can shorten some steps and obtain a more interpretable result. We introduce a $|\tilde{\boldsymbol{S}}_{I+Z}| \times |\tilde{\boldsymbol{S}}_{I+Z}|$ matrix $\bar{F}$ as

$$\bar{F}_{\alpha\beta} = \sqrt{|S_{\alpha}||S_{\beta}|}F_{\alpha\beta}, \ \bar{F}_{\alpha i} = \sqrt{|S_{\alpha}|}F_{\alpha i}, \tag{20}$$

with the remaining components being identical to those of $F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}}$, *i.e.*, $\bar{F}_{S_IS_I} = F_{S_IS_I}$. Eqs (19) and (20) indicate that $\bar{F}$ is simply the matrix summing the rows and columns in each cluster to a row and a column. It is natural that $\bar{F}$ has a direct connection to χ, because the locked variables in a cluster should exhibit the same response against perturbation. In fact, the response function χ in the original basis is expressed using $\bar{F}$ as

$$\begin{aligned}\chi = &\sum_{i,j\in S_I}\bar{F}^{-1}_{ij}(\boldsymbol{e}_i\boldsymbol{e}_j^{\top} + \boldsymbol{e}_j\boldsymbol{e}_i^{\top}) + \sum_{\alpha,\beta}\bar{F}^{-1}_{\alpha\beta}\sum_{i\in S_{\alpha}}\sum_{j\in S_{\beta}}\boldsymbol{e}_i\boldsymbol{e}_j^{\top} \\ &+ \sum_{\alpha}\left(\sum_{i\in S_{\alpha}}\sum_{j\in S_I}\bar{F}^{-1}_{\alpha j}\boldsymbol{e}_i\boldsymbol{e}_j^{\top} + \sum_{i\in S_I}\sum_{j\in S_{\alpha}}\bar{F}^{-1}_{i\alpha}\boldsymbol{e}_i\boldsymbol{e}_j^{\top}\right).\end{aligned} \tag{21}$$

This can be directly shown from Eqs (17 and 19), using the relation $F_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}} = P\bar{F}_{\tilde{\boldsymbol{s}}_{I+Z}\tilde{\boldsymbol{s}}_{I+Z}}P$ with $P = \mathrm{diag}(\{\{1\}_{i\in S_I}, \{\sqrt{|S_{\alpha}|}^{-1}\}_{\alpha}\})$, and the blockwise matrix inversion formula. Eqs (14) and (21) constitute the main result of this paper.

5 Algorithmic implementation

5.1 Numerical stability and the softening constant δ

For handling the singularity of the cost-function Hessian, we have introduced the softening constant δ in the TV and finally taken the $\delta \to 0$ limit. In practical implementations, however, we should keep δ small but finite. To see the reason, it is sufficient to see a simple example with

just three variables $\{x_i\}_{i=1}^3$. The softened TV is defined as

$$T^\delta(\boldsymbol{x}) = \sqrt{(x_2 - x_1)^2 + (x_3 - x_1)^2 + \delta^2} = \sqrt{p^2 + q^2 + \delta^2}, \tag{22}$$

where $p = x_2 - x_1$, $q = x_3 - x_1$ are introduced. The corresponding gradient and Hessian are

$$\frac{\partial T^\delta}{\partial \boldsymbol{x}} = \frac{1}{(p^2 + q^2 + \delta^2)^{1/2}} \begin{pmatrix} -p - q \\ p \\ q \end{pmatrix}, \tag{23}$$

$$\partial_{\boldsymbol{x}}^2 T^\delta = \frac{1}{(p^2 + q^2 + \delta^2)^{3/2}} \begin{pmatrix} (p-q)^2 + 2\delta^2 & pq - q^2 - \delta^2 & pq - p^2 - \delta^2 \\ pq - q^2 - \delta^2 & q^2 + \delta^2 & -pq \\ pq - p^2 - \delta^2 & -pq & p^2 + \delta^2 \end{pmatrix}. \tag{24}$$

The zero point of the gradient is given by $p = q = 0$ irrespectively of the δ value. Inserting this into the Hessian, we get one zero mode proportional to $(1, 1, 1)^\top$ and two finite modes whose eigenvalues are $(3/\delta, 1/\delta)$ being divergent in the $\delta \to 0$ limit. This exactly matches with the assumptions of the approximation formula.

On the other hand, if we first take the limit $\delta \to 0$ before taking the zero gradient limit $p, q \to 0$, we see that two zero modes appear: One is proportional to $(1, 1, 1)^\top$ and the other is to $(p + q, q - 2p, p - 2q)^\top$. This is a bad news because the second zero mode, which remains even in the limit $p, q \to 0$, is never taken into account when deriving the approximation formula: The derivation essentially depends on how the zero mode behaves and our formula loses its justification if such unexpected zero modes exist.

These considerations manifest that the two limits, $\lim_{\delta \to 0}$ and $\lim_{p,q \to 0}$, are not exchangeable in the TV Hessian. The derivation of our approximation formula assumes $\lim_{\delta \to 0} \lim_{p,q \to 0}$ and thus the algorithmic implementation should reflect this limit in a certain way. A simple way is to keep δ small but finite, which is actually a common technique to enhance the numerical stability when using the TV [14]. The choice of the amplitude of δ is related to the numerical precision when solving the optimization problem (3). A practical choice is stated in the next subsection.

5.2 Procedures

Here, we state the procedures for implementation of Eqs (14 and 21) in a numerical computation. Suppose that we have an algorithm to solve Eq (3) and to provide the solution $\hat{\boldsymbol{x}}$ given $\boldsymbol{y}$, A, λ_{ℓ_1}, and λ_T. Using this solution and introducing a finite δ in the Hessian by the reason discussed above, we can assess the LOOE through the following steps:

1. The sets of active and killed variables, S_A and S_K, are specified from $\hat{\boldsymbol{x}}$.
2. The values of all TV terms $\{t_i^\delta(\hat{\boldsymbol{x}})\}_{i=1}^N$ are computed.
3. All clusters C and the index sets belonging to the clusters $\{S_\alpha\}_{\alpha \in C}$ are enumerated from $\{t_i^\delta(\hat{\boldsymbol{x}})\}_{i=1}^N$, as well as the one of isolated variables, S_I.
4. The total variation from which the vanishing TV terms are removed is denoted by $\tilde{\boldsymbol{T}}^\delta(\hat{\boldsymbol{x}})$, and the regular part of the Hessian is computed as $F = G + \lambda_T \partial_{\boldsymbol{x}}^2 \tilde{\boldsymbol{T}}^\delta(\hat{\boldsymbol{x}})$.

5. A new index set $S_R = \{\{\alpha\}_{\alpha\in C}, S_I\}$ is defined.
6. On S_R, the merged Hessian $\bar{F}$ is constructed from F, as $\bar{F}_{S_IS_I} = F_{S_IS_I}$, $\bar{F}_{\alpha\beta} = \sum_{i\in S_\alpha, i\in S_\beta} F_{ij}$, $\bar{F}_{\alpha S_I} = \sum_{i\in S_\alpha} F_{iS_I}$, and $\bar{F}_{S_I\alpha} = \sum_{i\in S_\alpha} F_{S_Ii}$. Similarly, the merged measurement matrix $\bar{A}$ is defined as $\bar{A}_{\mu S_I} = A_{\mu S_I}$, $\bar{A}_{\mu\alpha} = \sum_{i\in S_\alpha} A_{\mu i}$.
7. Using $\bar{F}$ and $\bar{A}$, the LOOE factor in Eq (14) is computed as $1 - \boldsymbol{a}^T_{\mu S_A}\chi_{S_AS_A}\boldsymbol{a}_{\mu S_A} = 1 - \bar{\boldsymbol{a}}^T_{\mu S_R}(\bar{F}_{S_RS_R}\backslash\bar{\boldsymbol{a}}_{\mu S_R})$, where $\bar{\boldsymbol{a}}^T_\mu$ is the μth row vector of $\bar{A}$ and $\boldsymbol{x} = A\backslash\boldsymbol{b}$ is the solution of the linear equation $A\boldsymbol{x} = \boldsymbol{b}$.
8. Using the LOOE factor and $\hat{\boldsymbol{x}}$, the LOOE is evaluated from Eq (14).

At step 7, we take the left division $\bar{F}_{S_RS_R}\backslash\bar{\boldsymbol{a}}_{\mu S_R}$ instead of the inverse $\chi = \bar{F}^{-1}$ for numerical stability. The cluster enumeration at step 3 involves a delicate point in the definition of C and $\{S_\alpha\}_{\alpha\in C}$. Because of the limited precision in the numerics, the TV term $|\hat{x}_j - \hat{x}_i|$ $(j \in \partial i)$ never exactly vanishes; therefore, we need a certain threshold to extract the cluster structure from the TV terms. Here, we introduce the threshold θ and enumerate the clusters as follows:

3-1. If $t_i^\delta(\hat{\boldsymbol{x}}) \le \delta + \theta$, the variables in $\{i\} \cup \partial i$ are regarded as "linked." All the links are enumerated by testing $t_i^\delta(\hat{\boldsymbol{x}}) \le \delta + \theta$ for all $i = 1, \cdots, N$. The set of links is denoted by L, and the index set of all variables in L is denoted by S_L.

3-2. An empty set $C = \phi$ is prepared and the cluster index $\alpha = 1$ is defined.

3-3. The following steps are repeatedly implemented while L is non-empty:

(i). Two empty sets, $S_{\text{tmp}} = \phi$ and $S_{\text{cluster}} = \phi$, are prepared;

(ii). One link is selected and removed from L. The variable indices in the link are entered into S_{tmp};

(iii). The following steps are repeatedly implemented while S_{tmp} is non-empty:

a. One index i in S_{tmp} is selected and moved from S_{tmp} to S_{cluster};

b. If the above chosen index i exists in S_L, all the links to i are removed from L, and S_L is updated accordingly. The variables linked to i are entered into S_{tmp};

c. $S_{\text{tmp}} \leftarrow S_{\text{tmp}} - S_{\text{cluster}}$.

(iv). The variables in S_{cluster} constitute a cluster. $S_\alpha = S_{\text{cluster}}$ is defined and α is entered into C;

(v). $\alpha \leftarrow \alpha + 1$.

3-4. If $S_\alpha \cap S_K \neq \phi$, α is removed from C. This is checked for all $\alpha \in C$.

3-5. C, $\{S_\alpha\}_{\alpha\in C}$, and $S_I = S_A - \cup_{\alpha\in C} S_\alpha$ are returned.

The entire procedure presented above implements Eqs (14 and 21).

A debatable point would be the values of θ and δ. In most of iterative algorithms as the one in [8, 9], there is an inevitable finite error of the TV term even when it should vanish. Let us express the "scale" of this error as $t_i(\hat{\boldsymbol{x}}) \approx \theta_{\text{num}} > 0$. By construction, the threshold θ is related to this numerical error and it is appropriate to choose $\theta \approx \theta_{\text{num}}$; the softening constant δ should be sufficiently larger than θ_{num} because it does implement the assumed order of two

limits, $\lim_{\delta\to 0}\lim_{p,q\to 0}$, in derivation of the approximation formula. Overall, the relation

$$\theta \approx \theta_{\mathrm{num}} \ll \delta \tag{25}$$

must be satisfied. We have numerically checked how strict this principled relation is, and found that the approximation result is not sensitive to the choice of θ as long as it is sufficiently smaller than δ. Although a little more delicate points are involved in the choice of δ, we have also found that in a wide range of δ the approximation result is stable and the cost-function Hessian is safely invertible. Based on these observations, in the application of our formula below, the default values are set to be $\delta = 10^{-4}$ and $\theta = 10^{-12}$. They are chosen according to our datasets and experimental setup: The maximum value of the non-softened TV terms is scaled as $\max_i t_i(\hat{\boldsymbol{x}}) \gtrsim 10^{-4}$ and the numerical precision is about $\theta_{\mathrm{num}} \approx 10^{-12}$; the former value is reflected to δ and the latter one is used in θ. Coincidently, this default value of δ accords with the one in [14]. The examination result of the sensitivity to δ and θ will be reported below.

Another noteworthy point is that these procedures can be easily extended to other variants of the TV. For example, for the so-called anisotropic TV [9], $T_{\mathrm{ani}} = \sum_i \sum_{j\in\partial i}|x_j - x_i|$, we set $F = G$ in step 4 and modify the definition of the link in step 3-1 accordingly, so as to render our formula applicable. In the case of the square TV, $T_{\mathrm{sq}} = \sum_i \sum_{j\in\partial i}(x_j - x_i)^2 \equiv (1/2)\boldsymbol{x}^\top J\,\boldsymbol{x}$, the formula can be significantly simpler, because this TV has no sparsifying effect and the formula of the simple ℓ_1 case can be employed. We can employ Eq (14) with $\chi_{S_A S_A} = (G_{S_A S_A} + \lambda_T J_{S_A S_A})^{-1}$ directly, without the need for cluster enumeration.

6 Application to super-resolution imaging

To test the usefulness of the developed formula, let us apply the derived expression to the super-resolution reconstruction of astronomical images. A number of recent studies have demonstrated that sparse modeling is an effective means of reconstructing astronomical images obtained through radio interferometric observations [15–18]. In particular, the capability of high-fidelity imaging in super-resolution regimes has been shown, which renders this technique a useful choice for the imaging of black holes with the EHT [17–21]. We adopt the same problem setting as [17, 20] and demonstrate the efficacy of our approximation formula through comparison with the literally conducted 10-fold CV result. Here, x_i denotes the ith pixel value and A is (part of) the Fourier matrix. The dataset $\boldsymbol{y}$ is generated through the linear process

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{\xi} \tag{26}$$

where $\boldsymbol{\xi}$ is a noise vector and $\boldsymbol{x}_0$ is the simulated image, which we infer given $\boldsymbol{y}$ and A.

In this work, we use data for simulated EHT observations based on three different astronomical images, which are available as sample data for the EHT Imaging Challenge. Our datasets 1, 2, and 3 correspond to the sample datasets 1, 2, and 5, respectively, available from [22] at July 2017. The images are reconstructed with $N = 10000 = 100 \times 100$ pixels and with 160, 250, and 100 μ as fields of view, which are identical to the original images of Datasets 1, 2, and 3 from the EHT Imaging Challenge, respectively. We test four values for each λ_{ℓ_1} and λ_T: $\lambda_{\ell_1} \in (M/2) \times \{1, 10, 100, 1000\}$ and $\lambda_T \in (M/8) \times \{1, 10, 100, 1000\}$. M is 1910, 1910, and 2786, for Datasets 1–3, respectively. Later, we also use different size data from the same datasets, for checking the size dependence of the result.

Table 1 shows the mean CVE values for the three datasets, determined by the 10-fold CV and by our approximation formula for varying λ_T. λ_{ℓ_1} is fixed to the optimal value, which is coincidently common for all datasets and satisfies $2\lambda_{\ell_1}/M = 1$. It is clear that the approximate CVE values accord well with the 10-fold results, even on the error-bar scale, demonstrating

Table 1. CVE values determined by 10-fold CV and our approximation formula against λ_T. λ_{ℓ_1} is fixed to the optimal value ($2\lambda_{\ell_1}/M = 1$, coincidentally common to all cases). The number in brackets denotes the error bar to the last digits. The optimal values are bolded. The tuning constants δ and θ are set to be $\delta = 10^{-4}$ and $\theta = 10^{-12}$, respectively.

	$8\lambda_T/M$	1	10	100	1000
Dataset 1	10-fold	1.101(47)	**1.090(44)**	1.091(44)	1.455(108)
	Approx.	1.087(35)	**1.080(35)**	1.082(35)	1.385(49)
Dataset 2	10-fold	1.368(91)	**1.260(55)**	1.286(65)	2.843(234)
	Approx.	1.180(37)	**1.157(36)**	1.210(37)	2.669(108)
Dataset 3	10-fold	1.026(18)	**1.020(19)**	1.020(22)	1.235(52)
	Approx.	1.028(26)	**1.018(26)**	1.020(27)	1.226(40)

https://doi.org/10.1371/journal.pone.0188012.t001

that our approximation formula works very well. Note that the error bar for the approximation is given by the standard deviation of the M terms in Eq (14) divided by $\sqrt{M-1}$.

To directly observe the reconstruction quality, in Fig 1 we display the images at all investigated parameters and the reconstructed image at the optimal λ_{ℓ_1} and λ_T for Dataset 3, as well as the associated errors plotted against λ_{ℓ_1} and λ_T in Fig 2. Again, we can see the proposed method approximates the 10-fold result well, and the reconstructed image reasonably resembles the original. The RSS is monotonic with respect to the changes of λ_{l_1} and λ_T but the approximate LOOE is not, which implies that the LOOE factor computed through Eq (21) appropriately reflects the effect of the penalty terms.

Next, we check the sensitivity of the approximate result to the tuning constants δ and θ. In Fig 3, the approximate LOOEs at the optimal λ_{ℓ_1} are plotted against λ_T when changing δ (left) and θ (right). This indicates that the approximate LOOEs are stable against the change of both δ and θ. Hence, we may choose these values rather arbitrarily. This is a good news because tuning them makes the problem more numerically amenable: Enlarging δ makes the computation

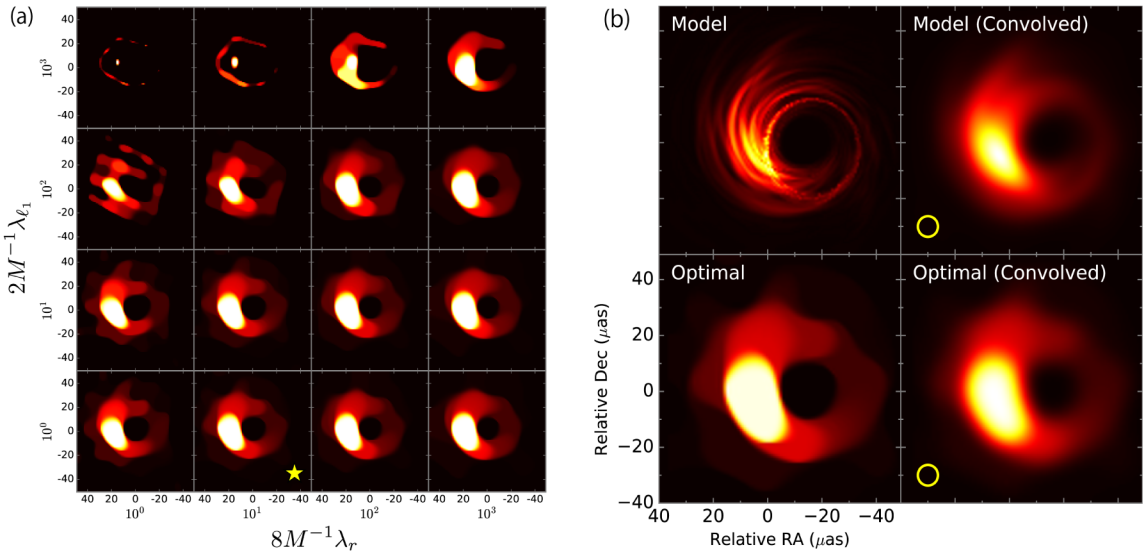

Fig 1. Super-resolution imaging results for Dataset 3 based on model image of supermassive black hole at center of nearby elliptical galaxy, M87. (a) Images for all investigated parameters; the star-marked panel is obtained at the optimum parameters. (b) Original images (top) and reconstructed images (bottom) at optimal parameters $((2\lambda_{\ell_1}, 8\lambda_T)/M = (1, 10))$. The images are convolved with a circular Gaussian beam on the right-hand side, the full width at half maximum (FWHM) of which is 25% of the nominal angular resolution of the EHT and corresponds to the diameters of the yellow circles. This coincides with the optimal resolution minimizing the mean square error between them.

https://doi.org/10.1371/journal.pone.0188012.g001

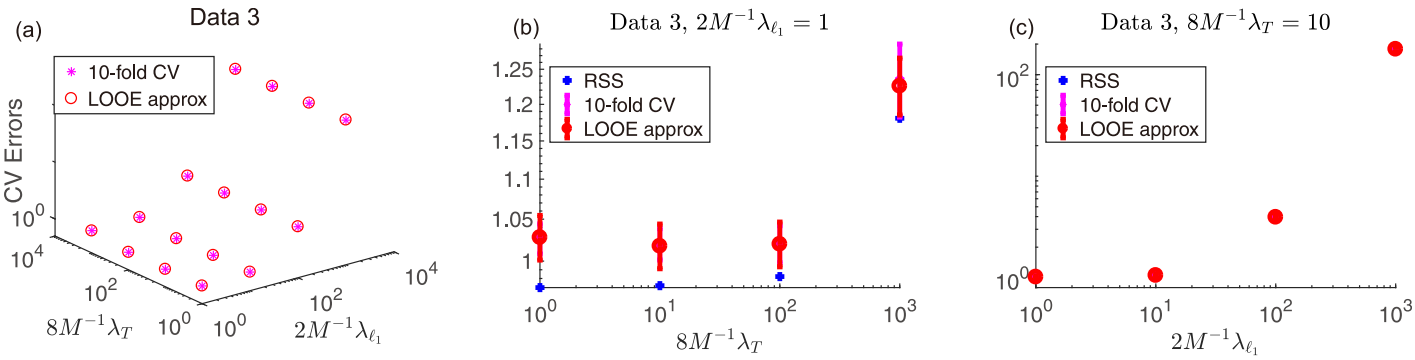

Fig 2. (a) 3D plot of mean CVEs against λ_{ℓ_1} and λ_T without error bars. (b) Plot of mean CVEs and RSS against λ_T at the optimal value of λ_{ℓ_1}, $2\lambda_{\ell_1}/M = 1$. (c) Plot of mean CVEs and RSS against λ_{ℓ_1} at the optimal value of λ_T, $8\lambda_T/M = 10$. For (c), the RSS is overlapped with the CVEs in the symbol size. In all the cases, the agreement between the approximate LOOE and the 10-fold CVE is fairly good. The tuning constants δ and θ are set to be $\delta = 10^{-4}$ and $\theta = 10^{-12}$, respectively.

https://doi.org/10.1371/journal.pone.0188012.g002

of the Hessian inversion more numerically stable; increasing θ lowers the effective degrees of freedom. The second property associated with θ is really beneficial when treating a large-size dataset, because it can downsize the Hessian and reduce the cost for computing its matrix inversion. In Table 2, the values of the effective degrees of freedom are given when changing θ. The reduction of the degree of freedom at large (yet small enough compared to $\delta = 10^{-4}$) θ is significant, which encourages us to apply the proposed formula to larger-size datasets.

Finally, let us see the data-size dependence of the approximation accuracy and of the computational cost for solving Eq (3) and for obtaining the approximate LOOE from the solution. The data analyzed here is an identical simulated image of black hole expressed with different number of pixels. When solving Eq (3), we used Intel(R) Core(TM) i7-5820K CPU of 3.30GHz with 6 cores for $N = 50^2 = 2500$ and Intel(R) Xeon(R) CPU E5-2699 v3 of 2.30GHz with 36 cores for $N = 100^2$ and 150^2, and employed an algorithm called "MFISTA" proposed in [8, 9]. Meanwhile, we used a laptop of a 1.7 GHz Intel Core i7 with two CPUs for evaluating

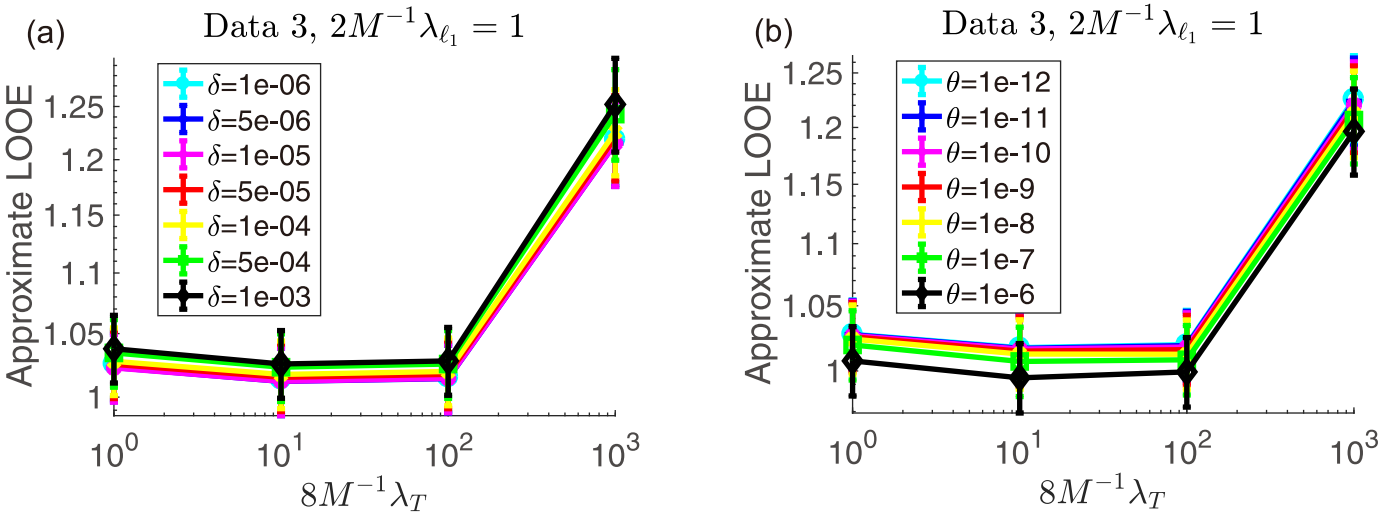

Fig 3. Comparative plots of mean approximate LOOEs against λ_T at $2M^{-1}\lambda_{\ell_1} = 1$ when (a) δ changes as 10^{-6}–10^{-3} with fixed $\theta = 10^{-12}$; (b) θ changes as 10^{-12}–10^{-6} with fixed $\delta = 10^{-4}$. They show that the LOOE curves are rather stable against the choice of the tuning constants.

https://doi.org/10.1371/journal.pone.0188012.g003

Table 2. The effective degrees of freedom $|\tilde{S}_{I+Z}|$, the number of clusters + the number of isolated variables, against θ for Dataset 3 at $\delta = 10^{-4}$ and the optimal parameters $(2\lambda_{\ell_1}, 8\lambda_T)/M = (1, 10)$.

θ	1e-12	1e-11	1e-10	1e-09	1e-08	1e-07	1e-06
$\lvert\tilde{S}_{I+Z}\rvert$	5733	5524	5243	4814	4112	2922	1408

https://doi.org/10.1371/journal.pone.0188012.t002

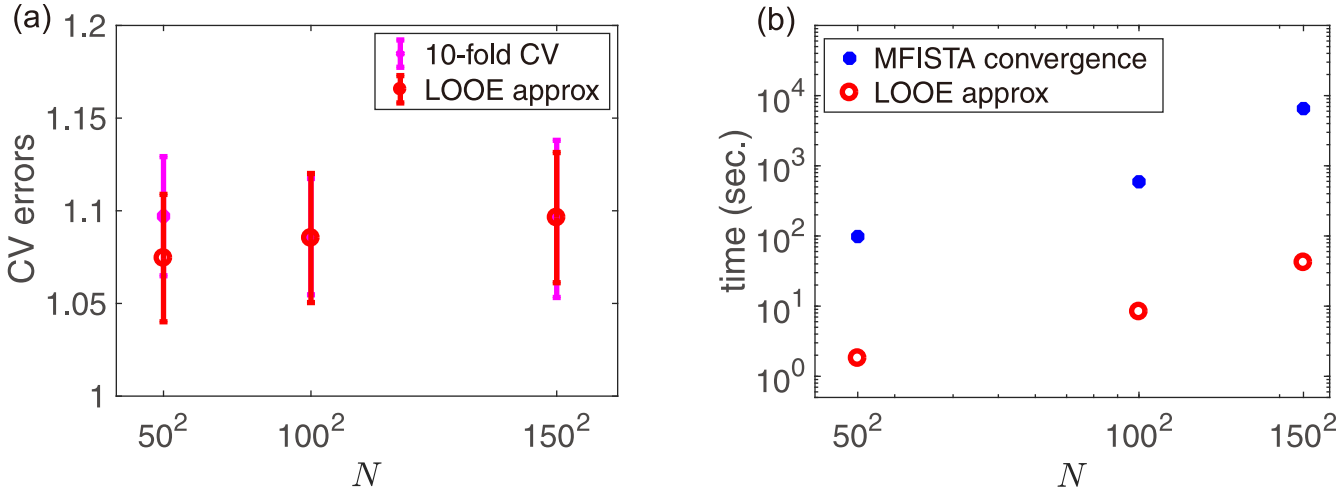

Fig 4. (a) Plot of mean CVEs at optimal parameters of different sizes. (b) Log-log plot of the computational times for solving the optimization problem (3) and for obtaining the approximate value of CVE against the size of datasets.

https://doi.org/10.1371/journal.pone.0188012.g004

the approximate LOOE. Hence the comparison is not fair and unfavorable to the approximation formula. The left panel indicates that the approximation accuracy becomes better for larger sizes. This is reasonable because the perturbation we have employed should have better accuracy as the model and data become larger, though the accuracy at $N = 50^2 = 2500$ is already good. The right panel clearly shows the advantage of the developed formula: The actual computational time of the approximate LOOE is significantly shorter than that of the algorithm convergence for solving Eq (3) in the investigated range of system sizes, even under the unfair comparison mentioned above. However, this advantage will be less prominent if the model becomes very large: Our approximation formula needs the Hessian inversion whose computational cost is scaled as $O((|C| + |S_I|)^3) \approx O(N^3)$, while MFISTA requires the cost of $O(N^2)$ as long as the number of steps to convergence is constant against N. The crossover size at which these two computational costs become comparable is roughly estimated as $N_\times \approx 10^6$, though such crossover tendency cannot be seen yet from Fig 4. For such large systems, a new fundamental solution should be tailored to resolve the computational-cost problem, though tuning θ to a large value in the present method can still be a good first aide.

7 Conclusion

In this paper, we have developed an approximation formula for the CVE of a sparse linear regression penalized by ℓ_1 and TV terms, and demonstrated its usefulness in the reconstruction of simulated black hole images. Our derivation is based on the perturbation assuming the small difference between the full and leave-one-out solutions. This assumption will not be fulfilled for some specific cases, i.e. when the measurement matrix is sparse. However, for most of dense measurement matrices, such as the Fourier matrix discussed in this paper, our

assumption will be reasonably satisfied. Hence we expect the range of application of our formula is wide enough and we would like encourage the readers to use this formula in their own work. It is also straightforward to generalize the developed formula to other types of TV, and two examples of the generalization for the anisotropic and square TVs have been explained.

The key concept of our formula, perturbation between the LOO and full systems, is very general and can be applied to more general statistical models and inference frameworks [23]. The development of practical formulas for those cases will facilitate higher levels of modeling and computation.

Acknowledgments

We would like to express our sincere gratitude to Mareki Honma and Fumie Tazaki for their helpful discussions. We thank Katherine L. Bouman for preparing the EHT Imaging Challenge website [22, 24]. We also thank Andrew Chael and Lindy Blackburn for writing a simulation software to produce sample data sets [25].

Author Contributions

Conceptualization: Yoshiyuki Kabashima.

Data curation: Shiro Ikeda, Kazunori Akiyama.

Formal analysis: Tomoyuki Obuchi, Yoshiyuki Kabashima.

Funding acquisition: Tomoyuki Obuchi, Shiro Ikeda, Kazunori Akiyama, Yoshiyuki Kabashima.

Investigation: Tomoyuki Obuchi.

Methodology: Tomoyuki Obuchi, Yoshiyuki Kabashima.

Project administration: Shiro Ikeda, Yoshiyuki Kabashima.

Software: Shiro Ikeda.

Validation: Tomoyuki Obuchi.

Visualization: Tomoyuki Obuchi, Kazunori Akiyama.

Writing – original draft: Tomoyuki Obuchi.

Writing – review & editing: Tomoyuki Obuchi.

References

1. Rish I, Grabarnik G. Sparse Modeling: Theory, Algorithms, and Applications. CRC Press; 2014.
2. Hastie T, Tibshirani R, Wainwright M. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press; 2015.
3. http://sparse-modeling.jp/index_e.html
4. Mairal J, Bach F, Ponce J. Sparse modeling for image and vision processing. Available from: arXiv:1411.3230v2.
5. Tibshirani R. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58, 267 (1996).
6. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann. Stat.*, 32, 407 (2004). https://doi.org/10.1214/009053604000000067
7. Rudin L I, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. *Physica D* 60, 259 (1992). https://doi.org/10.1016/0167-2789(92)90242-F
8. Chambolle A. An algorithm for total variation minimization and applications. *J. Math. Imaging Vision*, 20, 89 (2004). https://doi.org/10.1023/B:JMIV.0000011321.19549.88

9. Beck A, Teboulle M, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process.*, 18, 2419 (2009). https://doi.org/10.1109/TIP.2009.2028250 PMID: 19635705
10. Obuchi T, Kabashima Y. Cross validation in LASSO and its acceleration. *J. Stat. Mech.*, 053304 (2016). https://doi.org/10.1088/1742-5468/2016/05/053304
11. http://www.eventhorizontelescope.org
12. Asada K, Kino M, Honma M, Hirota T, Lu R.-S, Inoue M. White Paper on East Asian Vision for mm/submm VLBI: Toward Black Hole Astrophysics down to Angular Resolution of 1 R_S, arXiv:1705.04776
13. Akiyama K, Lu R, Fish V L, Doeleman S S, Broderick A E, Dexter J, et al. 230 GHz VLBI observations of M87: Event-horizon-scale structure during an enhanced very-high-energy γ-ray state in 2012. *Astrophys. J.*, 807, 150 (2015). https://doi.org/10.1088/0004-637X/807/2/150
14. Chan T F, Osher S, Shen J. The Digital TV Filter and Nonlinear Denoising, *IEEE Trans. Image Process.*, 10, 231 (2001). https://doi.org/10.1109/83.902288 PMID: 18249614
15. Wiaux Y, Jacques L, Puy G, Scaife A M M, Vandergheynst P. Compressed sensing imaging techniques for radio interferometry. *Mon. Not. R. Astron. Soc.*, 395, 1733 (2009). https://doi.org/10.1111/j.1365-2966.2009.14665.x
16. Li F, Cornwell T J, de Hoog F. The application of compressive sampling to radio astronomy I. Deconvolution. *A&A*, 528, A31 (2011).
17. Honma M, Akiyama K, Uemura M, Ikeda S. Super-resolution imaging with radio interferometry using sparse modeling. *Publ. Astron. Soc. Japan*, 66, 1 (2014). https://doi.org/10.1093/pasj/psu070
18. Honma M, Akiyama K, Tazaki F, Kuramochi K, Ikeda S, Hada K et al. Imaging black holes with sparse modeling. *JPCS*, 699, 012006 (2016).
19. Ikeda S, Tazaki F, Akiyama K, Hada K. PRECL: A new method for interferometry imaging from closure phase. *Publ. Astron. Soc. Japan*, 68, 45 (2016). https://doi.org/10.1093/pasj/psw042
20. Akiyama K, Ikeda S, Pleau M, Fish V, Tazaki F, Kuramochi K et al. Superresolution Full-polarimetric Imaging for Radio Interferometry with Sparse Modeling. The Astronomical Journal, 153, 1 (2017). https://doi.org/10.3847/1538-3881/aa6302
21. Akiyama K, Kuramochi K, Ikeda S, Fish V, Tazaki F, Honma M et al. Imaging the Schwarzschild-radius-scale Structure of M87 with the Event Horizon Telescope Using Sparse Modeling. The Astrophysical Journal, 838, 1 (2017). https://doi.org/10.3847/1538-4357/aa6305
22. http://vlbiimaging.csail.mit.edu/imagingchallenge
23. Kabashima Y, Obuchi T, Uemura M, Approximate cross–validation formula for Bayesian linear regression. Available from arXiv:1610.07733.
24. Bouman K L, Johnson M D, Zoran D, Fish V L, Doeleman S S, Freeman W T. Computational Imaging for VLBI Image Reconstruction. The IEEE Conference on Computer Vision and Pattern Recognition, 913 (2016).
25. Chael A A, Johnson M D, Narayan R, Doeleman S S, Wardle J F C, Bouman K L. High-resolution Linear Polarimetric Imaging for the Event Horizon Telescope. The Astrophysical Journal, 829, 15 (2016). https://doi.org/10.3847/0004-637X/829/1/11