ON THE DEGREES OF FREEDOM IN TOTAL VARIATION MINIMIZATION

Feng Xue[⋆] *and Thierry Blu*[†⋆]

[⋆]National Key Laboratory of Science and Technology on Test Physics and Numerical Mathematics,
Beijing, 100076, China
[†]Dept. of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

ABSTRACT

In the theory of linear models, the degrees of freedom (DOF) of an estimator play a pivotal role in risk estimation, as it quantifies the complexity of a statistical modeling procedure. Considering the total-variation (TV) regularization, we present a theoretical study of the DOF in Stein's unbiased risk estimate (SURE), under a very mild assumption. First, from the duality perspective, we give an analytic expression of the exact TV solution, with identification of its support. The closed-form expression of the DOF is derived based on the Karush-Kuhn-Tucker (KKT) conditions. It is also shown that the DOF is upper bounded by the nullity of a sub-analysis-matrix. The theoretical analysis is finally validated by the numerical tests on image recovery.

Index Terms— Degrees of freedom (DOF), Stein's unbiased risk estimate (SURE), total variation (TV) regularization, duality.

1. INTRODUCTION

Consider a classic linear model in finite-dimensional Hilbert space [1]:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{b} \tag{1}$$

with the fixed design matrix $\mathbf{A} : \mathbb{R}^N \mapsto \mathbb{R}^M$ and the observed data $\mathbf{y} \in \mathbb{R}^M$, where $\mathbf{b} \in \mathbb{R}^M$ denotes the measurement or modelling error. The goal of regression is to design an estimator of $\mathbf{x}_0 \in \mathbb{R}^N$, that frequently arises in fields such as statistical inference, machine learning, signal processing, imaging sciences and other inverse problems [2].

In regularized regression, total variation (TV) and its nonlocal versions have been a popular choice of regularizer during the past two decades, especially in image processing [3, 4, 5, 6]. In this paper, we restrict ourselves to a standard TV-regularized least square regression [4]:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}} \frac{1}{2}\big\|\mathbf{y} - \mathbf{A}\mathbf{x}\big\|_2^2 + \lambda \cdot g(\mathbf{D}\mathbf{x}) \tag{2}$$

where $\mathbf{D} = [\mathbf{D}_1^\top \ \mathbf{D}_2^\top]^\top : \mathbb{R}^N \mapsto \mathbb{R}^N \times \mathbb{R}^N : \mathbf{x} \mapsto (\mathbf{d}_1, \mathbf{d}_2)$ (where $\mathbf{d}_1 = \mathbf{D}_1\mathbf{x}$ and $\mathbf{d}_2 = \mathbf{D}_2\mathbf{x}$) is a 2-D first-order difference operator, consisting of both horizontal and vertical directions $\mathbf{D}_1 : \mathbb{R}^N \mapsto \mathbb{R}^N$ and $\mathbf{D}_2 : \mathbb{R}^N \mapsto \mathbb{R}^N$[1]. The *isotropic* total variation function is defined as $g : \mathbb{R}^N \times \mathbb{R}^N \mapsto \mathbb{R} : (\mathbf{d}_1, \mathbf{d}_2) \mapsto \sum_{n=1}^N \sqrt{(\mathbf{d}_1)_n^2 + (\mathbf{d}_2)_n^2}$.

Here, we regard the TV regularization (2) as a regression procedure or fitting model. Let $\boldsymbol{\mu}_0 = \mathbf{A}\mathbf{x}_0$. Denote any estimate of $\boldsymbol{\mu}_0$ by $\widehat{\boldsymbol{\mu}} = \mathbf{A}\widehat{\mathbf{x}}$, where $\widehat{\mathbf{x}}$ is any estimate of $\mathbf{x}_0$[2]. The notion of the degrees of freedom (DOF) was proposed in [8, 9], which is used to quantify the complexity of a statistical modeling procedure $\mathcal{M} : \mathbf{y} \mapsto \widehat{\boldsymbol{\mu}}(\mathbf{y})$. From the viewpoint of regression (in classical statistics), the DOF is the number of linearly independent parameters (variables). Here, we use a more general definition of DOF of an adaptively fitted model (see [10, Eq.(3.60)]), which implies that: the harder that we fit to the data, the larger DOF we will have (i.e. the more parameters we need to use in the fitted model). Refer to [10, Chapter 3] for more elaborations of the effective DOF of a model.

Assuming that the error $\mathbf{b}$ in (1) follows normal distribution, i.e. $\mathbf{b} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_M)$, the well-known *Stein's lemma* asserts that the DOF, defined as [10, Eq.(3.60)], can be unbiasedly estimated by [11]:

$$\widehat{\mathrm{df}} = \mathrm{Tr}\big(\mathbf{J}_\mathbf{y}(\widehat{\boldsymbol{\mu}})\big) = \mathrm{Tr}\big(\mathbf{A}\mathbf{J}_\mathbf{y}(\widehat{\mathbf{x}})\big) \tag{3}$$

where $\mathbf{J}_\mathbf{y}(\widehat{\boldsymbol{\mu}})$ and $\mathbf{J}_\mathbf{y}(\widehat{\mathbf{x}})$ denote the Jacobian matrices of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\mathbf{x}}$ w.r.t. $\mathbf{y}$, respectively. Tr denotes the matrix trace. Then, the *Stein's unbiased risk estimate* (SURE), given by:

$$\mathrm{SURE} = \frac{1}{M}\big\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\big\|_2^2 + \frac{2\sigma^2}{M}\underbrace{\mathrm{Tr}\big(\mathbf{A}\mathbf{J}_\mathbf{y}(\widehat{\mathbf{x}})\big)}_{\widehat{\mathrm{df}}} - \sigma^2 \tag{4}$$

is an unbiased estimate of the *mean squared error* (MSE) [12, 13, 14], i.e. $\mathbb{E}\big\{\mathrm{SURE}\big\} = \frac{1}{M}\mathbb{E}\big\{\big\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\big\|_2^2\big\}$.

SURE provides a principled and efficient way for optimization in various applications, see [15, 16, 17, 18, 19, 20, 21] for example. The main difficulty of SURE-based methods lies in the evaluation of DOF. For the TV regularization,

*This work was supported by a grant CUHK14210617 from the Hong Kong research grant council.

[1] In the 1-D case, the TV regularization reduces to a generalized lasso problem, where the analysis operator is $\mathbf{D}$. The DOF of the lasso problem has been extensively studied before, e.g.,[7]. Here, we consider 2-D TV case, where $\mathbf{x}$ is a 2-D signal (e.g., an image), for more applications to image processing.

[2] We distinguish the notation $\widehat{\mathbf{x}}$ from $\mathbf{x}^\star$, in a sense that $\widehat{\mathbf{x}}$ denotes any estimate of $\mathbf{x}_0$, while $\mathbf{x}^\star$ refers to the exact TV solution to (2) only.

the DOF can be practically computed by recursively differentiating the sequence of iterates $\mathbf{x}^{(i+1)} := \mathbf{f}(\mathbf{x}^{(i)})$ (i denotes the iteration number of some optimization algorithms), that converges to a solution of the TV minimization problem (2), e.g. [22, 23, 24]. In the aspect of theoretical analysis, the properties of DOF for the (generalized) lasso solution have been investigated in [7, 25, 14, 26, 27, 2]. However, to the best of our knowledge, the theoretical study of the DOF of TV solution (2) has not been performed before.

In this paper, we explore the theoretical properties of the TV solution $\mathbf{x}^\star$ in (2) and its DOF, and provide an upper bound of the DOF. This study may pave a way for more efficient evaluation of the SURE for the TV minimizer, and can be extended to more general (convex) regularized M-estimators, typically associated with non-smooth regularizers.

2. THE EXACT SOLUTION TO TV MINIMIZATION

Before analyzing the DOF of the TV solution, it is necessary to first find the exact expression of the TV solution to (2). We remind readers that all the proofs of lemmas and theorems in Sections 2 and 3 are left to Section 5.

2.1. Notations and definitions

The set $\mathcal{I}$ is the *D-support* of the solution $\mathbf{x}^\star \in \mathbb{R}^N$ in (2), if $\mathcal{I} = \{i : (\mathbf{D}\mathbf{x}^\star)_i \neq 0\} \subseteq \{1, 2, ..., N\}$. Its *D-cosupport*, denoted by $\mathcal{J}$, is $\mathcal{J} = \{j : (\mathbf{D}\mathbf{x}^\star)_j = 0\}$. Also, $\mathcal{J} = \mathcal{I}^c$ (c denotes the complement). $|\mathcal{J}|$ is the cardinality of the set $\mathcal{J}$ [28].

$\mathbf{x}_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}|}$ and $\mathbf{x}_\mathcal{J} \in \mathbb{R}^{|\mathcal{J}|}$ extract the elements indexed by $\mathcal{I}$ and $\mathcal{J}$ from $\mathbf{x}$. $(\mathbf{D}_1)_\mathcal{J} \in \mathbb{R}^{|\mathcal{J}| \times N}$ denotes a sub-matrix of $\mathbf{D}_1$, whose *rows* are indexed by $\mathcal{J}$ in $\mathbf{D}_1$; $(\mathbf{D}_2)_\mathcal{J} \in \mathbb{R}^{|\mathcal{J}| \times N}$ is defined similarly. $\mathbf{D}_\mathcal{J} := [(\mathbf{D}_1)_\mathcal{J}^\top \ (\mathbf{D}_2)_\mathcal{J}^\top]^\top \in \mathbb{R}^{2|\mathcal{J}| \times N}$, and $\mathbf{D}_\mathcal{I} := [(\mathbf{D}_1)_\mathcal{I}^\top \ (\mathbf{D}_2)_\mathcal{I}^\top]^\top \in \mathbb{R}^{2|\mathcal{I}| \times N}$ [26]. In addition, $\mathbf{I}_N$ denotes a $N \times N$ identity matrix.

Throughout this paper, we always assume that the assumption $\mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{D}_\mathcal{J}) = \{\mathbf{0}\}$ holds, where $\mathcal{N}$ denotes the null space of a matrix [28, 2]. Notice that this assumption is very mild in many interesting applications [14, 28].

2.2. Exact solution to the TV minimization

Due to the non-smoothness of the TV term, it is almost intractable to exactly solve (2) directly from the primal form. The duality is able to find the exact solution to TV minimization [29]. This was initiated by a seminal work of [30], where the dual formulation of TV denoising (i.e. $\mathbf{A} = \mathbf{I}_N$, ROF model [3]) is proposed. For the invertible $\mathbf{A}$, the dual was proposed in [31]. We now extend the dual form to general $\mathbf{A}$. This new result is given by the following lemma.

Lemma 2.1 *The dual to the primal problem* (2) *is given by:*

$$\min_{\mathbf{q}_\mathcal{I}} \left\| \mathbf{A}\boldsymbol{\Gamma}_\mathcal{J}\mathbf{D}_\mathcal{I}^\top \mathbf{q}_\mathcal{I} - \lambda^{-1}\mathbf{y} \right\|_2^2 \quad \text{s.t.} \quad \max_{n=1,...,|\mathcal{I}|} \left\| (\mathbf{q}_\mathcal{I})_n \right\|_2 \leq 1 \tag{5}$$

where $\boldsymbol{\Gamma}_\mathcal{J} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top$, *the columns of* $\boldsymbol{\Gamma} \in \mathbb{R}^{N \times s}$ *form an orthonomal basis of the null space of* $\mathbf{D}_\mathcal{J}$ *(where* $s := \text{nullity}(\mathbf{D}_\mathcal{J})$*), i.e.* $\mathbf{D}_\mathcal{J}\boldsymbol{\Gamma} = \mathbf{0}_{\mathbf{2}|\mathcal{J}|\times\mathbf{s}}$. $\mathbf{D}_\mathcal{I}$ *is defined similarly to* $\mathbf{D}_\mathcal{J}$. $\mathbf{q}_\mathcal{I} := \begin{bmatrix} \mathbf{q}_{1,\mathcal{I}}^\top & \mathbf{q}_{2,\mathcal{I}}^\top \end{bmatrix}^\top \in \mathbb{R}^{2|\mathcal{I}|}$, *and* $(\mathbf{q}_\mathcal{I})_n := \begin{bmatrix} (\mathbf{q}_{1,\mathcal{I}})_n & (\mathbf{q}_{2,\mathcal{I}})_n \end{bmatrix}^\top \in \mathbb{R}^2$.

Remark 1: it is easy to check that $\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma}$ is invertible (such that $\boldsymbol{\Gamma}_\mathcal{J}$ is well-defined), by the assumption $\mathcal{N}(\mathbf{A}) \cap \mathcal{N}(\mathbf{D}_\mathcal{J}) = \{\mathbf{0}\}$.

By solving the above dual problem (5), and the connection between the primal variable $\mathbf{x}^\star$ and the dual $\mathbf{q}_\mathcal{I}^\star$ (see the proof below), the TV solution $\mathbf{x}^\star$ can be expressed as follows.

Lemma 2.2 *The exact solution to the total variation regularization is given by:*

$$\mathbf{x}^\star = \left(\mathbf{I}_N + \lambda \cdot \boldsymbol{\Gamma}_\mathcal{J}\mathbf{D}_\mathcal{I}^\top \overline{\mathbf{M}}^{-1}\mathbf{D}_\mathcal{I}\right)^{-1}\boldsymbol{\Gamma}_\mathcal{J}\mathbf{A}^\top\mathbf{y} \tag{6}$$

where $\overline{\mathbf{M}} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \in \mathbb{R}^{2|\mathcal{I}| \times 2|\mathcal{I}|}$, $\mathbf{M} = \text{diag}(\mu_n)$ *with* $\mu_n = \sqrt{(\mathbf{d}_{1,\mathcal{I}}^\star)_n^2 + (\mathbf{d}_{2,\mathcal{I}}^\star)_n^2}$, *and* $\mathbf{d}_{1,\mathcal{I}}^\star = \mathbf{D}_{1,\mathcal{I}}\mathbf{x}^\star$ *and* $\mathbf{d}_{2,\mathcal{I}}^\star = \mathbf{D}_{2,\mathcal{I}}\mathbf{x}^\star$.

Remark 2: Lemma 2.2 gives the expression of *exact* TV solution, which is obtained by duality, following the *rigorous* isotropic TV definition (given in (2)), rather than a smooth approximation of TV, as in [4, 32].

Remark 3: The exact solution $\mathbf{x}^\star$ is fully characterized in the support domain $\mathcal{I}$, while restricted to satisfy $\mathbf{D}_\mathcal{J}\mathbf{x}^\star = \mathbf{0}$. This is similar to the notion of active sub-manifold in [2].

3. MAIN RESULTS

3.1. Exact DOF for TV regularization

The main results of this paper are obtained from the above lemmas. Theorem 3.1 gives the exact DOF expression for the general TV regularization.

Theorem 3.1 *The exact DOF for TV regularization is:*

$$\widehat{\text{df}} = \text{Tr}\left((\mathbf{I}_N + \lambda\boldsymbol{\Gamma}_\mathcal{J}\mathbf{D}_\mathcal{I}^\top\mathbf{S}^\star\mathbf{D}_\mathcal{I})^{-1}\boldsymbol{\Gamma}_\mathcal{J}\mathbf{A}^\top\mathbf{A}\right) \tag{7}$$

where $\boldsymbol{\Gamma}_\mathcal{J}$ *is defined in Lemma 2.1, the block-diagonal matrix* $\mathbf{S}^\star = \begin{bmatrix} \mathbf{G}^\star & \mathbf{T}^\star \\ \mathbf{T}^\star & \mathbf{R}^\star \end{bmatrix} \in \mathbb{R}^{2|\mathcal{I}| \times 2|\mathcal{I}|}$, $\mathbf{G}^\star = \text{diag}\left((\mathbf{d}_{2,\mathcal{I}}^\star)_n^2 \cdot \mu_n^{-3}\right)$, $\mathbf{R}^\star = \text{diag}\left((\mathbf{d}_{1,\mathcal{I}}^\star)_n^2 \cdot \mu_n^{-3}\right)$, *and* $\mathbf{T}^\star = \text{diag}\left((\mathbf{d}_{1,\mathcal{I}}^\star)_n \cdot (\mathbf{d}_{2,\mathcal{I}}^\star)_n \cdot \mu_n^{-3}\right)$, μ_n *is given in* (6).

Remark 4: This main result is limited to the TV minimizer $\mathbf{x}^\star$ only. It is not applicable for other estimates $\widehat{\mathbf{x}}$, since this result is obtained based on the KKT optimality conditions (see the proofs for more details), which are not satisfied by other estimates $\widehat{\mathbf{x}}$.

An upper bound of the DOF naturally follows from Theorem 3.1.

Corollary 3.1 *The exact DOF, given by* (7), *satisfies:*

$$\widehat{\mathrm{df}} \leq \mathrm{nullity}(\mathbf{D}_{\mathcal{J}}) \tag{8}$$

Remark 5: Recall that $\widehat{\mathrm{df}} = \mathrm{nullity}(\mathbf{D}_{\mathcal{J}})$ for the generalized lasso solution with full-column rank $\mathbf{A}$ and arbitrary $\mathbf{D}$ [26]. This corollary states that *the DOF of the TV minimizer is no greater than that of the generalized lasso solution.*

3.2. A simple case: TV denoising

For the TV denoising solution (corresponding to $\mathbf{A} = \mathbf{I}_N$ in (2)), the DOF can be easily obtained from Theorem 3.1.

Proposition 3.1 *The exact DOF for TV denoiser (with* $\mathbf{A} = \mathbf{I}_N$ *in* (2)*) is:*

$$\widehat{\mathrm{df}} = \mathrm{Tr}\big((\mathbf{I}_s + \lambda \mathbf{\Gamma}^\top \mathbf{D}_{\mathcal{I}}^\top \mathbf{S}^\star \mathbf{D}_{\mathcal{I}} \mathbf{\Gamma})^{-1}\big) \tag{9}$$

where s, $\mathbf{D}_{\mathcal{I}}$ *and* $\mathbf{\Gamma}$ *are defined in Lemma 2.1,* $\mathbf{S}^\star$ *is given in Theorem 3.1.*

Remark 6: it is more straightforward to obtain the upper bound (8) from (9). Indeed,

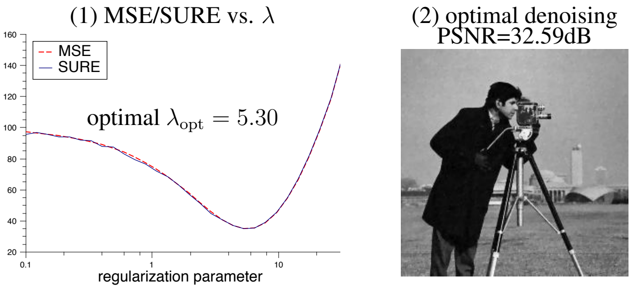
$$\widehat{\mathrm{df}} = \mathrm{Tr}\big((\mathbf{I}_s + \lambda \mathbf{\Gamma}^\top \mathbf{D}_{\mathcal{I}}^\top \mathbf{S}^\star \mathbf{D}_{\mathcal{I}} \mathbf{\Gamma})^{-1}\big) \leq \mathrm{Tr}(\mathbf{I}_s) = s$$

4. NUMERICAL RESULTS AND DISCUSSIONS

In this part, we verify the theoretical results and illustrate the accuracy of the proposed DOF estimator (and its associated SURE) on a parameter selection problem in the context of TV-based image recovery. In particular, SURE provides an automatic and objective way to select the regularization parameter λ in (2), such that the restored image achieves minimum MSE and best visual quality.

4.1. Image denoising

We consider *Cameraman* as a test image, corrupted by Gaussian noise with the variance $\sigma^2 = 100$. Fig.1-(1) depicts the SURE, with the DOF computed by (9), as a function of regularization parameter λ. We can see that the SURE, evaluated by (9), is always very close to the true MSE (defined immediately after Eq.(4)) for all values of λ, and the optimal value is $\lambda = 5.30$, indicated by the minimum SURE. The optimally denoised image is shown in Fig.1-(2) for visual inspection of the denoising quality.



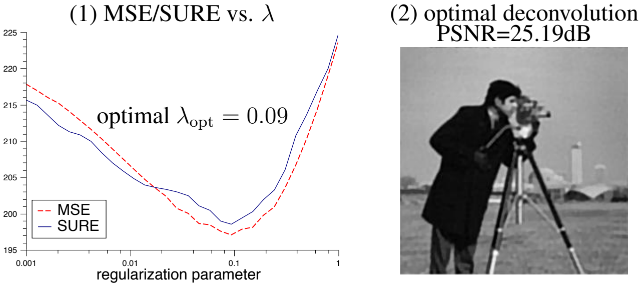
(1) MSE/SURE vs. λ

(2) optimal denoising PSNR=32.59dB

Fig. 1. The SURE-based optimization of λ (denoising).

4.2. Image deconvolution

The image *Cameraman* is blurred by a 2-D Gaussian kernel (with variance 4.0), and subsequently contaminated by Gaussian noise with $\sigma^2 = 5$. The DOF of SURE is computed by (7). Fig.2-(1) presents the variation of MSE/SURE w.r.t. the regularization parameter λ. The best restored image with the optimal $\lambda = 0.09$ is shown in Fig.2-(2).



(1) MSE/SURE vs. λ

(2) optimal deconvolution PSNR=25.19dB

Fig. 2. The SURE-based optimization of λ (deconvolution).

5. PROOFS

5.1. Proof of Lemma 2.1

Before obtaining the dual, we need to first consider primal-dual formulation, summarized in the following lemma.

Lemma 5.1 *The primal solution* $\mathbf{x}^\star$ *and the dual* $\mathbf{q}_{\mathcal{I}}^\star$ *are connected via:*

$$\mathbf{x}^\star = \mathbf{\Gamma}_{\mathcal{J}}\big(\mathbf{A}^\top \mathbf{y} - \lambda \cdot \mathbf{D}_{\mathcal{I}}^\top \mathbf{q}_{\mathcal{I}}^\star\big) \tag{10}$$

where $\mathbf{\Gamma}_{\mathcal{J}}$ *is defined in Lemma 2.1.*

Proof The primal-dual formulation of (2) is given by [31, 33]:

$$\min_{\mathbf{x}} \max_{\|q_n\|_2 \leq 1} \frac{1}{2}\big\|\mathbf{A}\mathbf{x} - \mathbf{y}\big\|_2^2 + \lambda \cdot \mathbf{x}^\top \mathbf{D}^\top \mathbf{q}, \ \forall n \tag{11}$$

where the dual variable is $\mathbf{q} = [\mathbf{q}_1^\top \ \mathbf{q}_2^\top]^\top \in \mathbb{R}^N \times \mathbb{R}^N$, and $\|q_n\|_2 := \sqrt{(\mathbf{q}_1)_n^2 + (\mathbf{q}_2)_n^2}, \forall n$. Denote the solutions by $\mathbf{x}^\star$ and $\mathbf{q}^\star$, the first-order optimality condition yields $\mathbf{A}^\top \mathbf{A}\mathbf{x}^\star = \mathbf{A}^\top \mathbf{y} - \lambda \cdot \mathbf{D}^\top \mathbf{q}^\star$.

Let $\mathcal{J}$ be the D-support of $\mathbf{x}^\star$, then $\mathbf{D}_{\mathcal{J}}\mathbf{x}^\star = \mathbf{0}$. We express $\mathbf{x}^\star$ as: $\mathbf{x}^\star = \mathbf{\Gamma}\xi^\star$, where the columns of $\mathbf{\Gamma} \in \mathbb{R}^{N \times s}$ form an orthonormal basis of the null space of $\mathbf{D}_{\mathcal{J}}$, s is the nullity of $\mathbf{D}_{\mathcal{J}}$, i.e. $s := \mathrm{nullity}(\mathbf{D}_{\mathcal{J}})$. Simple algebra leads to $\xi^\star = (\mathbf{\Gamma}^\top \mathbf{A}^\top \mathbf{A}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^\top\big(\mathbf{A}^\top \mathbf{y} - \lambda \cdot \mathbf{D}_{\mathcal{I}}^\top \mathbf{q}_{\mathcal{I}}^\star\big)$, where $\mathbf{q}_{\mathcal{I}}^\star$ collects the elements of $\mathbf{q}^\star$, indexed by $\mathcal{I}$. Substituting $\xi^\star$ into $\mathbf{x}^\star = \mathbf{\Gamma}\xi^\star$ completes the proof. ■

Now, we are ready to prove Lemma 2.1.

Proof of Lemma 2.1: Putting (10) into (11), using the basic facts that $\mathbf{\Gamma}_{\mathcal{J}}^\top = \mathbf{\Gamma}_{\mathcal{J}}$, $\mathbf{\Gamma}_{\mathcal{J}}\mathbf{A}^\top \mathbf{A}\mathbf{\Gamma}_{\mathcal{J}} = \mathbf{\Gamma}_{\mathcal{J}}$ and $(\mathbf{I} - \mathbf{A}\mathbf{\Gamma}_{\mathcal{J}}\mathbf{A}^\top)^2 = \mathbf{I} - \mathbf{A}\mathbf{\Gamma}_{\mathcal{J}}\mathbf{A}^\top$, we obtain:

$$\mathcal{L}(\mathbf{x}^\star, \mathbf{q}) = -\frac{\lambda^2}{2}\big\|\mathbf{A}\mathbf{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top \mathbf{q}_{\mathcal{I}} - \lambda^{-1}\mathbf{y}\big\|_2^2 + \mathbf{y}^\top\Big(\mathbf{I} - \frac{1}{2}\mathbf{A}\mathbf{\Gamma}_{\mathcal{J}}\mathbf{A}^\top\Big)\mathbf{y}$$

Maximizing $\mathcal{L}(\mathbf{x}^\star, \mathbf{q})$ is equivalent to (5). ■

5.2. Proof of Lemma 2.2

Proof First, we need to solve the dual problem (5). Recalling that $\mathbf{q} = [\mathbf{q}_1^\top \ \ \mathbf{q}_2^\top]^\top$, by Lagrangian, we obtain:

$$\min_{\mathbf{q}} \left\|\mathbf{A}\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top \mathbf{q}_{\mathcal{I}} - \lambda^{-1}\mathbf{y}\right\|_2^2 + \lambda^{-1} \cdot \sum_{n=1}^{|\mathcal{I}|} \mu_n \big((\mathbf{q}_{1,\mathcal{I}})_n^2 + (\mathbf{q}_{2,\mathcal{I}})_n^2 - 1\big)$$

i.e.

$$\left\|\mathbf{A}\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top \mathbf{q}_{\mathcal{I}} - \lambda^{-1}\mathbf{y}\right\|_2^2 + \lambda^{-1}\big(\mathbf{q}_{1,\mathcal{I}}^\top \mathbf{M}\mathbf{q}_{1,\mathcal{I}} + \mathbf{q}_{2,\mathcal{I}}^\top \mathbf{M}\mathbf{q}_{2,\mathcal{I}} - \mathrm{Tr}(\mathbf{M})\big)$$

where $\mathbf{M} = \mathrm{diag}(\mu_n) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$.

Differentiating w.r.t. $\mathbf{q}_1$ and $\mathbf{q}_2$, and combining with (10), we obtain $\mathbf{M}\mathbf{q}_{1,\mathcal{I}}^\star = \mathbf{D}_{1,\mathcal{I}}\mathbf{x}^\star := \mathbf{d}_{1,\mathcal{I}}^\star$ and $\mathbf{M}\mathbf{q}_{2,\mathcal{I}}^\star = \mathbf{D}_{2,\mathcal{I}}\mathbf{x}^\star := \mathbf{d}_{2,\mathcal{I}}^\star$. By the KKT conditions [34], we claim that:

- if $(\mathbf{q}_{1,\mathcal{I}}^\star)_n^2 + (\mathbf{q}_{2,\mathcal{I}}^\star)_n^2 - 1 = 0$, then $\mu_n \neq 0$. Thus, $(\mathbf{q}_{1,\mathcal{I}}^\star)_n = \mu_n^{-1}(\mathbf{d}_{1,\mathcal{I}}^\star)_n$ and $(\mathbf{q}_{2,\mathcal{I}}^\star)_n = \mu_n^{-1}(\mathbf{d}_{2,\mathcal{I}}^\star)_n$. From the constraint, we obtain $\mu_n = \sqrt{(\mathbf{d}_{1,\mathcal{I}}^\star)_n^2 + (\mathbf{d}_{2,\mathcal{I}}^\star)_n^2}$. This exactly satisfies $\mu_n \neq 0$, by the definition of the support $\mathcal{I}$.

- If $(\mathbf{q}_{1,\mathcal{I}}^\star)_n^2 + (\mathbf{q}_{2,\mathcal{I}}^\star)_n^2 - 1 \neq 0$, then $\mu_n = 0$. It yields that $(\mathbf{d}_{1,\mathcal{I}}^\star)_n = (\mathbf{d}_{2,\mathcal{I}}^\star)_n = 0$, which contradicts the definition of support $\mathcal{I}$. Thus, this case does not happen.

The solution can be rewritten as $\mathbf{q}_{\mathcal{I}}^\star = \overline{\mathbf{M}}^{-1}\mathbf{D}_{\mathcal{I}}\mathbf{x}^\star$, where $\overline{\mathbf{M}}$ is defined in (6). Substituting it into (10) leads to (6), after some algebra rearrangements. ■

5.3. Proof of Theorem 3.1

Proof First, we need to find $\mathbf{J}_{\mathbf{y}}(\mathbf{x}^\star)$. Rewriting (6) as $(\mathbf{I}_N + \lambda \cdot \boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top \overline{\mathbf{M}}^{-1}\mathbf{D}_{\mathcal{I}})\mathbf{x}^\star = \boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{A}^\top\mathbf{y}$, and taking differentiation w.r.t. $\mathbf{y}$ on both sides, we obtain:

$$\mathbf{J}_{\mathbf{y}}(\mathbf{x}^\star) + \lambda\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{J}_{\mathbf{y}}\big(\mathbf{D}_{\mathcal{I}}^\top\overline{\mathbf{M}}^{-1}\mathbf{D}_{\mathcal{I}}\mathbf{x}^\star\big) = \boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{A}^\top$$

After many algebra steps of vector calculus (omitted here to save page space), we have:

$$\big(\mathbf{I} + \lambda\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}\big)\mathbf{J}_{\mathbf{y}}(\mathbf{x}^\star) = \boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{A}^\top$$

where $\mathbf{S}^\star$ is defined in (7).

Finally, by (3), the exact DOF becomes:

$$\begin{aligned}
\widehat{\mathrm{df}} &= \mathrm{Tr}\big(\mathbf{A}\mathbf{J}_{\mathbf{y}}(\mathbf{x}^\star)\big) \\
&= \mathrm{Tr}\big(\mathbf{A}(\mathbf{I}_N + \lambda\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}})^{-1}\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{A}^\top\big) \\
&= \mathrm{Tr}\big((\mathbf{I}_N + \lambda\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}})^{-1}\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{A}^\top\mathbf{A}\big)
\end{aligned}$$

The proof is completed. ■

5.4. Proof of Corollary 3.1

Proof Putting $\boldsymbol{\Gamma}_{\mathcal{J}} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top$ into (7) yields:

$$\widehat{\mathrm{df}} = \mathrm{Tr}\big(\mathbf{B}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\big)$$

where $\mathbf{B} = \mathbf{I}_N + \lambda\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}$. By matrix inversion lemma, $\mathbf{B}^{-1}$ becomes:

$$\mathbf{B}^{-1} = \mathbf{I}_N - \boldsymbol{\Gamma}\underbrace{\big(\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}\boldsymbol{\Gamma} + \lambda^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma}\big)^{-1}}_{\mathbf{T}}\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}$$

Thus, we have:

$$\begin{aligned}
\widehat{\mathrm{df}} &= \mathrm{Tr}\big((\mathbf{I}_N - \boldsymbol{\Gamma}\mathbf{T}\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}})\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\big) \\
&= \mathrm{Tr}\Big(\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\Big) \\
&\quad - \mathrm{Tr}\Big(\boldsymbol{\Gamma}\mathbf{T}\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\Big) \\
&= \mathrm{Tr}\Big(\underbrace{(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma}}_{\mathbf{I}_s}\Big) \\
&\quad - \mathrm{Tr}\Big(\mathbf{T}\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}\boldsymbol{\Gamma}\underbrace{(\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{A}^\top\mathbf{A}\boldsymbol{\Gamma}}_{\mathbf{I}_s}\Big) \\
&= \underbrace{\mathrm{nullity}(\mathbf{D}_{\mathcal{J}})}_{s} - \underbrace{\mathrm{Tr}(\mathbf{T}\boldsymbol{\Gamma}^\top\mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}\boldsymbol{\Gamma})}_{\geq 0} \\
&\leq \mathrm{nullity}(\mathbf{D}_{\mathcal{J}})
\end{aligned}$$

The proof is completed. ■

5.5. Proof of Proposition 3.1

Proof For the TV denoising problem, $\mathbf{A} = \mathbf{I}_N$. Thus, $\boldsymbol{\Gamma}_{\mathcal{J}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$, since $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \mathbf{I}_s$. From Theorem 3.1, simply denoting $\mathbf{B} = \mathbf{D}_{\mathcal{I}}^\top\mathbf{S}^\star\mathbf{D}_{\mathcal{I}}$, the DOF becomes:

$$\begin{aligned}
\widehat{\mathrm{df}} &= \mathrm{Tr}\big((\mathbf{I}_N + \lambda\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{B})^{-1}\boldsymbol{\Gamma}_{\mathcal{J}}\mathbf{A}^\top\mathbf{A}\big) \\
&= \mathrm{Tr}\big((\mathbf{I}_N + \lambda\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\big) \\
&= \mathrm{Tr}\Big(\big(\mathbf{I}_N - \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{B}\boldsymbol{\Gamma} + \lambda^{-1}\mathbf{I}_s)^{-1}\boldsymbol{\Gamma}^\top\mathbf{B}\big)\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\Big) \\
&= \mathrm{Tr}\Big(\boldsymbol{\Gamma}\big(\mathbf{I}_s - (\boldsymbol{\Gamma}^\top\mathbf{B}\boldsymbol{\Gamma} + \lambda^{-1}\mathbf{I}_s)^{-1}\boldsymbol{\Gamma}^\top\mathbf{B}\boldsymbol{\Gamma}\big)\boldsymbol{\Gamma}^\top\Big) \\
&= \mathrm{Tr}\big(\mathbf{I}_s - (\boldsymbol{\Gamma}^\top\mathbf{B}\boldsymbol{\Gamma} + \lambda^{-1}\mathbf{I}_s)^{-1}\boldsymbol{\Gamma}^\top\mathbf{B}\boldsymbol{\Gamma}\big) \\
&= \mathrm{Tr}\big((\lambda\boldsymbol{\Gamma}^\top\mathbf{B}\boldsymbol{\Gamma} + \mathbf{I}_s)^{-1}\big)
\end{aligned}$$

where the third line comes from matrix inversion lemma. ■

6. CONCLUSIONS

In this paper, we presented a theoretical analysis for the degrees of freedom of the total variation solution. This was achieved through the concept of duality, identification of *D*-support $\mathcal{I}$ and KKT conditions. The result paved a way to derive the SURE. The simulations confirm our theoretical findings and show that our risk estimator provides a viable way for automatic choice of the regularization parameter.

In principle, our analysis can be generalized to more general convex regularized M-estimators. Extending our results to the non-convex case would also be very interesting. This would, however, require more sophisticated techniques in variational and non-smooth analysis. The above problems will be left to future work.

7. REFERENCES

[1] C.R. Rao, H. Toutenburg, H.C. Shalabh, and C. Heumann, *Linear Models and Generalizations: Least Squares and Alternatives*, 3rd edition, Springer, 2008.

[2] S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal, "The degrees of freedom of partly smooth regularizers," *Ann. Inst. Stat. Math.*, vol. 69, pp. 791–832, 2017.

[3] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.

[4] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 460–489, 2005.

[5] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *SIAM Multiscale Modeling and Simulation*, vol. 7, no. 3, pp. 1005–1028, 2008.

[6] Z. Li, F Malgouyres, and T. Zeng, "Regularized non-local total variation and application in image restoration," *Journal of Mathematical Imaging & Vision*, vol. 59, no. 3, pp. 296–317, 2017.

[7] Hui Zou, Trevor Hastie, and Robert Tibshirani, "On the "degrees of freedom" of the lasso," *The Annals of Statistics*, vol. 35, no. 5, pp. 2173–2192, 2007.

[8] B. Efron, "How biased is the apparent error rate of a prediction rule?," *J. Amer. Statist. Assoc.*, vol. 81, no. 394, pp. 461–470, 1986.

[9] J. Ye, "On measuring and correcting the effects of data mining and model selection," *J. Amer. Statist. Assoc.*, vol. 93, pp. 120–131, 1998.

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, vol. 2, Springer, 2009.

[11] Charles M Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, pp. 1135–1151, 1981.

[12] R. Giryes, M. Elad, and Y.C. Eldar, "The projected GSURE for automatic parameter tuning in iterative shrinkage methods," *Applied and Computational Harmonic Analysis*, vol. 30, no. 3, pp. 407–422, 2010.

[13] Y.C. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 471–481, 2009.

[14] S. Vaiter, C. Deledalle, G. Peyré, J. Fadili, and C. Dossal, "Local behavior of sparse analysis regularization: Applications to risk estimation," *Applied and Computational Harmonic Analysis*, vol. 35, no. 3, pp. 433–451, 2013.

[15] T. Blu and F. Luisier, "The SURE-LET approach to image denoising," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2778–2786, 2007.

[16] C. A. Deledalle, V. Duval, and J. Salmon, "Non-local methods with shape-adaptive patches (NLM-SAP)," *Journal of Mathematical Imaging and Vision*, vol. 43, no. 2, pp. 103–120, 2012.

[17] F. Xue, F. Luisier, and T. Blu, "Multi-Wiener SURE-LET deconvolution," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1954–1968, 2013.

[18] Magnus O. Ulfarsson and Victor Solo, "Selecting the number of principal components with SURE," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 239–243, 2015.

[19] S. K. Yadav, R. Sinha, and P. K. Bora, "An efficient SVD shrinkage for rank estimation," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2406–2410, 2015.

[20] Danning Li and Hui Zou, "SURE information criteria for large covariance matrix estimation and their asymptotic properties," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2153–2169, 2016.

[21] Christopher A. Metzler, Ali Mousavi, Reinhard Heckel, and Richard Baraniuk, "Unsupervised learning with Stein's unbiased risk estimator," in *Proc. of International Biomedical and Astronomical Signal Processing (BASP) Frontiers Workshop,* 2019.

[22] S. Ramani, Zhihao Liu, J. Rosen, J. Nielsen, and J.A. Fessler, "Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3659–3672, 2012.

[23] Feng Xue, Thierry Blu, Jiaqi Liu, and Xia Ai, "Recursive evaluation of SURE for total variation denoising," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, Alberta, Canada, 2018, pp. 1338–1342.

[24] F. Xue, J. Liu, and X. Ai, "Recursive SURE for image recovery via total variation minimization," *Signal, Image and Video Processing*, vol. 13, no. 4, pp. 795–803, 2019.

[25] Ryan J Tibshirani and Jonathan Taylor, "The solution path of the generalized Lasso," *The Annals of Statistics*, vol. 39, pp. 1335–1371, 2011.

[26] Ryan J Tibshirani and Jonathan Taylor, "Degrees of freedom in lasso problems," *The Annals of Statistics*, vol. 40, no. 2, pp. 1198–1232, 2012.

[27] Charles Dossal, Maher Kachour, Jalal Fadili, Gabriel Peyré, and Christophe Chesneau, "The degrees of freedom of the lasso for general design matrix," *Staistica Sinica*, vol. 23, no. 2, pp. 809–828, 2013.

[28] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili, "Robust sparse analysis regularization," *IEEE Trans. Information Theory*, vol. 59, no. 4, pp. 2001–2016, 2013.

[29] N Komodakis and J.C. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, 2015.

[30] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, pp. 89–97, 2004.

[31] Mingqiang Zhu and Tony F. Chan, "An efficient primal-dual hybrid gradient algorithm for total variation image restoration," CAM Report 08-34, UCLA, Los Angeles, CA, 2008.

[32] C. R. Vogel and M. E. Oman, "Iterative methods for total variation denoising," *SIAM J. Sci. Comput.*, vol. 17, no. 1, pp. 227–238, 1996.

[33] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[34] Stephen P. Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.