

χ^2 TEST FOR TOTAL VARIATION REGULARIZATION PARAMETER SELECTION

J. MEAD*

Department of Mathematics
Boise State University
Boise, ID , USA

ABSTRACT. Total Variation (TV) is an effective method of removing noise in digital image processing while preserving edges. The scaling or regularization parameter in the TV process defines the amount of denoising, with a value of zero giving a result equivalent to the input signal. The discrepancy principle is a classical method for regularization parameter selection whereby data is fit to a specified tolerance. The tolerance is often identified based on the fact that the least squares data fit is known to follow a χ^2 distribution. However, this approach fails when the number of parameters is greater than or equal to the number of data. Typically, heuristics are employed to identify the tolerance in the discrepancy principle and this leads to oversmoothing. In this work we identify a χ^2 test for TV regularization parameter selection assuming the blurring matrix is full rank. In particular, we prove that the degrees of freedom in the TV regularized residual is the number of data and this is used to identify the appropriate tolerance. The importance of this work lies in the fact that the χ^2 test introduced here for TV automates the choice of regularization parameter selection and can straightforwardly be incorporated into any TV algorithm. Results are given for three test images and compared to results using the discrepancy principle and MAP estimates.

1. Introduction. Most image processing problems take a two-dimensional signal as input and produce an improved image that comes in the form of a set of parameters related to the image. Since the input signal typically contains noise or blurring, a regularization term is included in the inversion. A regularization parameter that controls the amount of regularization is necessary for the inversion. Unfortunately, the choice of regularization parameter is often overlooked and an inversion is run many times with different parameter values until a clear image is produced. This process can be improved by solving a sequence of regularized problems iteratively with the regularization parameter decreasing to zero. In this case the initial inverse estimate is updated at each iterate so that regularization is no longer necessary at the end of the iterative procedure. Alternatively, we suggest solving a single regularized problem with an optimally chosen regularization parameter. Solving a single regularized linear problem rather than iterating a sequence of regularized linear problems is advantageous for nonlinear or computationally large problems.

The difficulty with choosing the optimal regularization parameter is that essentially the regularization term contains prior information about the unknown parameters and there is potential to add bias to the solution. This is addressed in Bayesian methods by identifying appropriate prior probability models [39]. Bayesian methods find a range of parameter sets in the form posterior

2010 *Mathematics Subject Classification.* 65F22, 62J07, 94A08.

Key words and phrases. regularization parameter, total variation, chi-squared test, discrepancy principle, regression.

This work was funded by National Science Foundation, DMS10431047.

distributions. Rather than estimate a posterior distribution for the parameters, we instead consider the inverse problem as a statement about the prior and posterior means of the parameters and input signal, and use standard deviations to weight them.

If the error in data and initial parameter estimates are assumed to be normal, the maximum a posteriori probability (MAP) estimate is found solving a least squares problem with Tikhonov regularization. Choice of regularization parameter for Tikhonov regularization is well studied and common methods include the Discrepancy Principle, L-curve and generalized cross-validation (GCV) [21]. The validity of all of these approaches relies on knowing the optimal solution of the quadratic least squares functional, which is not possible for the TV functional. For example, in the L-curve approach, the norm of the data residual vs. the solution can be plotted, but with the TV functional there is no guarantee that the curve will be L-shaped with a corner containing the optimal regularization parameter value.

Various authors have been able to exploit quadratic aspects of TV regularization to find regularization parameters. In [28, 38] the choice of scalar regularization parameter is made with the discrepancy principle applied to the data residual. The discrepancy principle relies on estimates of the data error or the standard deviation of data noise. In [42] they assume the noise is from a Poisson or multiplicative Gamma distribution and use this property to find the expected value of the data error. The discrepancy principle can lead to oversmoothing when data error is not known because it is difficult to estimate the degrees of freedom in the residual. In [48] the discrepancy principle is used with correct, equivalent degrees of freedom [45]. The equivalent degrees of freedom can be found for a quadratic functional and while TV is not quadratic, such a functional arises in their particular iterative primal-dual proximal point algorithm for TV optimization.

In [30] the Unbiased Predictive Risk Estimator (UPRE) is extended to TV regularization where they approximate the TV term with a quadratic functional. Again the quadratic approximation of the TV functional simplifies the degrees of freedom choice. As an alternative to approximating the TV functional with a quadratic functional, SURE methods use Stein's estimate for the correct degrees of freedom in the predictive risk estimator. Stein's degrees of freedom estimate involves gradients which in [9] they calculate with finite differences. Alternatively, in [4] the SURE-LET algorithm assumes the optimal parameter estimates are a linear combination of denoising functions. This linear approximation allows for simpler estimates of the gradient calculation in the degrees of freedom estimate.

The discrepancy principle and predictive risk methods rely on variance estimators of the data and empirically give a Bayesian framework, at least for the observations. In [11] they use TV regularization with the Richardson–Lucy algorithm, which takes into consideration the Poisson nature of the noise, for 3D imaging of biological specimens. However, the TV regularization parameter is chosen in an ad hoc manner. Alternatively in [25] they exploit GCV to determine how to iteratively update the regularization parameter in a variable splitting algorithm. They use GCV to determine the regularization parameter for the quadratic part of the split and then step through regularization parameter values to identify the best choice in the TV minimization part of the split.

Another approach to TV regularization parameter selection is bi-level optimization. Bi-level optimization is similar to the discrepancy and predictive risk methods because there are two levels of optimization. In the discrepancy and predictive risk methods, the low level optimization is the regularized inverse problem that finds parameter estimates, and the high-level optimization problem finds the regularization parameter(s) according to a specified loss function. In the discrepancy principle the loss function is mean-squared error while in UPRE and SURE methods the loss function is predictive risk. Bi-level optimization in [10, 27] also solves the regularized inverse problem in the low level optimization but they set up the high-level loss function as training problem that finds regularization parameters which minimizes the difference between the estimated parameters and the training set value for the parameters. Another bi-level optimization method [22, 23] solves the Fenchel-predual problem in a low-level optimization because it involves minimizing a smooth objective function. The high-level problem optimizes a loss function so that weighted data residual lies within upper and lower bounds of the data noise variance, which is similar to a discrepancy type principle.

Here we derive an approach that does not rely on a quadratic interpretation of TV regularization nor does it require training data. We identify correct degrees of freedom for predictive risk but do not need to estimate degrees of freedom using Stein's theorem because we use the regularized residual. With the regularized residual no estimation procedure is necessary for the

degrees of freedom as long as the smoothing operator \mathbf{A} is full rank. Therefore the regularization parameter selection algorithm proposed here using a χ^2 test on the regularized residual can be used with any TV optimization algorithm.

In Section 2 we first explain why the TV functional is a χ^2 random variable. Then in Section 3 we derive the χ^2 test for TV regularization using the TV regularized residual. Knowing the correct degrees of freedom is crucial to the method and we use recent results on the degrees of freedom of the Lasso method [13, 44] to derive the degrees of freedom with a full rank blurring matrix. In Section 4 we show results from three different images, two types of noise, three different noise levels and three TV algorithms: **deconvtv** [7], **FTVd** [47] and **tvdeconv** [15]. Successful image reconstruction is measured through Improved Signal to Noise Ration (ISNR). Given the exact image, the regularization parameter that gives the highest ISNR is calculated and compared to the ISNR with regularization parameter selection algorithms for a MAP estimate, the discrepancy principle and the χ^2 test.

2. Least Squares Estimators. Regularization parameter selection is well studied for Tikhonov regularization [21] in linear problems,. We will briefly review it in a framework that introduces the χ^2 test strategy proposed here for TV and explains how the Discrepancy principle can be viewed as a χ^2 test.

Linear inverse problems are often stated as

$$\mathbf{d} = \mathbf{Ax} + \boldsymbol{\epsilon} \quad (1)$$

where for two dimensional $m \times n$ greyscale image reconstruction problems $\mathbf{d} \in \mathbb{R}^{mn}$ is a vector of noisy observations, $\mathbf{A} \in \mathbb{R}^{mn \times mn}$ is a known model, and $\mathbf{x} \in \mathbb{R}^{mn}$ contains the unknown information about the true image that has m columns and n rows. The vector $\boldsymbol{\epsilon}$ represents noise in the observations and it is assumed it has zero mean with variance σ^2 . In image processing, these vectors and matrix are specified by concatenating in a column- or row- wise fashion pixel values for the image. The matrix \mathbf{A} can be used to represent a shift in the image, in which case the image is noisy, or \mathbf{A} may represent a blur filter induced by motion in the image. In either case, a_{ij} acts on the value of the image pixel at the point indexed by (i, j) . For example an image represented by 256×256 pixels, has $m = n = 256$.

A least squares fit of the unknowns to the observation solves the minimization problem

$$\min_{\mathbf{x}} \|\mathbf{d} - \mathbf{Ax}\|_2^2. \quad (2)$$

Typically this optimization problem is ill-posed so a regularization term, or constraints, are added to it. The focus of this work is when the regularization term is a TV functional, but first we discuss least squares regularization which takes the form:

$$\hat{\mathbf{x}}(\alpha) = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{d} - \mathbf{Ax}\|_2^2 + \alpha^2 \|\mathbf{L}(\mathbf{x} - \mathbf{x}_p)\|_2^2. \quad (3)$$

The matrix \mathbf{L} is a $q \times mn$ shaping matrix with no restriction on q , while \mathbf{x}_p is an initial guess that is often taken to be zero. If α^2 is chosen large, the regularization term dominates and more weight will be placed on the second misfit, hence the estimate $\hat{\mathbf{x}}$ will be "close" to \mathbf{x}_p . If \mathbf{L} is chosen to be the identity, there can be significant smoothing of the solution which may not be desirable in imaging applications. If \mathbf{L} is diagonal or dense, non-smooth least squares estimates result [35], and applying \mathbf{L} has the effect of applying a spatially dependent regularization parameter [12]. This regularization process can also be viewed as constrained linear inversion [8] or ridge regression [31].

In Bayesian inference the unknown parameters \mathbf{x} are treated as random variables. If \mathbf{x}_p is the mean of \mathbf{x} and $\alpha^2 \mathbf{L}^T \mathbf{L}$ is the inverse error covariance, the optimal parameters found by Tikhonov regularization can be viewed as the maximum a posteriori estimate when error in the initial estimate and data are normally distributed. Using Bayes theorem, the optimal parameters are described by the posterior distribution of \mathbf{x} given by

$$p(\mathbf{x}|\mathbf{d}) = p(\mathbf{d}|\mathbf{x})p(\mathbf{x}).$$

In many applications it is often the case that the error distributions are not normal. Reasonable estimates can still be found with Tikhonov regularization if the errors are not normal as long as bias in the data and initial guess are removed, and the misfits are weighted by their inverse covariances. In practice this can be done with M-estimators which use initial estimates of the parameters, and data weighted with a diagonal matrix containing inverse variances of their errors. The difficulty then lies in estimating variances for errors in initial parameter estimates. Estimating this prior

information is the focus of many Bayesian methods, for example in [3] they use a hierarchical Bayesian model for TV in blind deconvolution.

Estimation of initial parameter errors with a scalar can be viewed as equivalent to regularization methods that focus on choice of parameter α^2 . The following χ^2 tests are often used to find α .

Proposition 1. Suppose that $(\mathbf{d} - \mathbf{Ax})_i \sim \mathcal{N}(0, \sigma^2)$ $i = 1, \dots, mn$ and $\mathbf{L}(\mathbf{x} - \mathbf{x}_p)_i \sim \mathcal{N}(0, \alpha^{-1})$ $i = 1, \dots, q$, with all elements independent, then

$$\mathcal{J}_{tik}(\mathbf{x}) = \frac{\|\mathbf{d} - \mathbf{Ax}\|_2^2}{\sigma^2} + \alpha^2 \|\mathbf{L}(\mathbf{x} - \mathbf{x}_p)\|_2^2 \sim \chi_{mn+q}^2.$$

Proof. It is well known that the first term in the sum is distributed as χ_{mn}^2 [5], since it is the sum of squares of mn normally distributed random variables. Similarly the second term is distributed as χ_q^2 , and since they are independent, their sum is χ_{mn+q}^2 . \square

The Discrepancy Principle can be viewed as a χ^2 test using just the data misfit, i.e. a value of α^2 is found so that the optimal estimate $\hat{\mathbf{x}}$ defined by (3) satisfies

$$\mathcal{J}_d(\hat{\mathbf{x}}) = \frac{\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}\|_2^2}{\sigma^2} = \tau mn \quad (4)$$

where τ is a predetermined real number. If $\hat{\mathbf{x}}$ is independent of \mathbf{d} then $\tau = 1$ satisfies the χ^2 test, however this is not the case because $\hat{\mathbf{x}}$ is the least squares solution and depends on \mathbf{d} . When there is the same number of data as unknowns we cannot appropriately reduce the degrees of freedom in the χ^2 test because there no χ^2 distribution with fewer than one degrees of freedom. In this case we can use the regularized residual as a χ^2 test [32, 33] for choice of α^2 . It is based on the following theorem.

Theorem 1. Suppose that $(\mathbf{d} - \mathbf{Ax})_i$ $i = 1, \dots, mn$ and $\mathbf{L}(\mathbf{x} - \mathbf{x}_p)_i$ $i = 1, \dots, q$ are independent, not necessarily normally distributed, but have mean zero and standard deviation σ and α^{-1} , respectively. Then for large mn

$$\mathcal{J}_{tik}(\hat{\mathbf{x}}) = \frac{\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}\|_2^2}{\sigma^2} + \alpha^2 \|\mathbf{L}(\hat{\mathbf{x}} - \mathbf{x}_p)\|_2^2 \sim \chi_q^2.$$

Proof. Shown in [33]. \square

Theorem 1 shows that the regularized residual is a χ^2 random variable and it gives the corresponding degrees of freedom, q . This suggests the following χ^2 test for choice of α^2 in Tikhonov regularization:

$$\frac{\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\alpha)\|_2^2}{\sigma^2} + \alpha^2 \|\mathbf{L}(\hat{\mathbf{x}}(\alpha) - \mathbf{x}_p)\|_2^2 \approx q. \quad (5)$$

A more general formula for the degrees of freedom is given in [14]

$$df(\mathbf{A}\hat{\mathbf{x}}) = \sum_{i=1}^{mn} \text{cov}((\mathbf{A}\hat{\mathbf{x}})_i, d_i)/\sigma^2. \quad (6)$$

In particular for Tikhonov regularization the degrees of freedom is $df(\mathbf{A}\hat{\mathbf{x}}(\alpha)) = \text{tr}(\mathbf{N}(\alpha))$ where $\mathbf{N}(\alpha) = \mathbf{A}(\mathbf{A}^T \mathbf{A} + \alpha \mathbf{I})^{-1} \mathbf{A}^T$. This gives the χ^2 test for the residual in the discrepancy principle:

$$\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\alpha)\|_2^2/\sigma^2 \approx mn - \text{tr}(\mathbf{N}(\alpha))$$

where $mn - \text{tr}(\mathbf{N}(\alpha))$ is the “effective degrees of freedom”.

The χ^2 test for regularization parameter selection is similar to finding a parameter that minimizes an estimate of the predictive risk $\mathbb{E}(\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}\|_2^2)$. The unbiased estimate of the mean squared error (MSE) is

$$\mathbb{E}(\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\alpha)\|_2^2) = -2\sigma^2 df(\mathbf{A}\hat{\mathbf{x}}) + mn\sigma^2.$$

For Tikhonov regularization the degrees of freedom are known (i.e. it is the effective degrees of freedom), and the Unbiased Predictive Risk Estimator method (UPRE) finds

$$\alpha \in \underset{\alpha}{\operatorname{argmin}} \left\{ \|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\alpha)\|_2^2/\sigma^2 + 2\text{tr}(\mathbf{N}(\alpha)) - mn \right\}.$$

3. Least Squares with Total Variation Regularization. Ill-conditioning of the image reconstruction problem defined by (2) is often alleviated by TV regularization because TV results in reconstructions that maintain sharp edges in the image. With this type of regularization the prior information is a gradient vector containing the finite difference approximation to partial derivatives with appropriate boundary conditions. Here we use one sided differences at the boundaries and the anisotropic discrete total variation functional denoted as

$$TV(\mathbf{x}) = \|\mathbf{D}\mathbf{x}\|_1$$

where \mathbf{D} is a $mn \times mn$ matrix containing the stacking of the difference approximations. The TV image reconstruction problem is typically written as

$$\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \frac{\lambda}{2} \|\mathbf{d} - \mathbf{A}\mathbf{x}\|_2^2 + TV(\mathbf{x}). \quad (7)$$

We can still view the Discrepancy Principle for TV as a χ^2 test since it only acts on the residual. However, we must determine the reduced degrees of freedom. Since there is no closed form expression for $\hat{\mathbf{x}}$ for nonlinear smoothers like TV, it is more difficult to find the effective degrees of freedom than it is with Tikhonov regularization. In addition, based on the discussion in Section 2, it may appear that the degrees of freedom are effectively zero for square matrices \mathbf{A} as is the case with the Tikhonov solution. However, we do not observe this in practice.

In cases where the reconstruction operator \mathbf{A} is not linear, or regularization operators other than quadratic Tikhonov are used, Stein's Lemmas [Stein 1981] is often used to determine the degrees of freedom in the predictive risk estimator. In particular, Stein shows that

$$df(\mathbf{A}\hat{\mathbf{x}}) = \sum_{i=1}^{mn} \frac{\partial(\mathbf{A}\hat{\mathbf{x}})_i}{\partial d_i} / \sigma^2.$$

Using this estimate of degrees of freedom gives Stein's Unbiased Risk Estimation (SURE) methods. The challenge with SURE methods is in estimating the gradient. In [9] they use finite differences to estimate the gradient while in [4] they assume the optimal estimate $\hat{\mathbf{x}}$ is a linear combination of denoising functions which simplifies the calculations.

Unbiased degrees of freedom estimates for TV are obtained in [13] using Stein's lemma, while in [44] they derived them from Karush-Kuhn-Tucker (KKT) conditions. Here's we use the estimate from [44]:

$$df(\mathbf{A}\hat{\mathbf{x}}(\lambda)) = \mathbb{E}[dim(\mathbf{A}(null(\mathbf{D}_{-\mathcal{A}})))], \quad \mathcal{A} = \{i : (\mathbf{D}\hat{\mathbf{x}}(\lambda))_i \neq 0\}, \quad (8)$$

where \mathcal{A} is the active set of the TV, or Lasso, estimate and consists of indices of $\mathbf{D}\hat{\mathbf{x}}$ that are nonzero. In addition, $\mathbf{D}_{-\mathcal{A}}$ is the derivative matrix \mathbf{D} after removing rows that are indexed by \mathcal{A} and $null(\mathbf{D}_{-\mathcal{A}})$ is its nullspace. The expected value of the dimension of the linear subspace $\mathbf{A}(null(\mathbf{D}_{-\mathcal{A}}))$ gives the degrees of freedom.

The degrees of freedom estimate in (8) may be used in a Discrepancy principle or SURE method for TV to get the correct degrees of freedom, however, it can be difficult to compute. As an alternative to just using the predictive risk estimator, i.e. the data residual, here we use the degrees of freedom estimate (8) for the TV regularized residual $\frac{\lambda}{2} \|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\lambda)\|_2^2 + TV(\hat{\mathbf{x}}(\lambda))$. In particular when \mathbf{A} is full rank, we have identified a χ^2 test on the regularized residual that does not require direct calculation of (8). This χ^2 test will be developed in Section 3.1.

3.1. TV Functional and Laplacian distribution. TV can be viewed as a sparsity prior in the gradient domain that conforms to a Laplacian distribution with zero mean. We assume that pixel intensities are independent, which may not be accurate, but is a good first approximation to reality. An important consideration in modeling the TV functional as differentially Laplacian is the assumption that coefficients sum to zero, which is the case with pixel intensities.

Define $y_i = x_{i+1} - x_i$, and assume it follows a Laplacian distribution with probability density function

$$f(y_i|\theta, \beta) = \frac{1}{2\beta} e^{-|y_i - \theta|/\beta}$$

where θ is the mean of y_i and $2\beta^2$ is the variance. In [17] they suggest that the standard deviation can be estimated by the observed pixel differences. If we also assume the data are estimates of the mean of $\mathbf{A}\mathbf{x}$, and the errors follow a normal distribution, then the MAP estimate is

$$\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \frac{\|\mathbf{d} - \mathbf{A}\mathbf{x}\|_2^2}{2\sigma^2} + \frac{\|\mathbf{y}\|_1}{\beta}. \quad (9)$$

Lemma 2. *TV regularization parameter λ defined by (7) gives the MAP estimate when $\lambda = \beta/\sigma^2$.*

Lemma 2 was observed in [17] and is the basis for a regularization parameter selection method given in Algorithm 1.

Algorithm 1 TV as MAP estimate; Laplace hyperparameter from pixel differences.

- 1: input \mathbf{d} , σ (data and its std. dev.)
 - 2: $\mathbf{D} = \text{diff}(\mathbf{d})$; %The difference of adjacent elements in \mathbf{d} .
 - 3: $\beta = \text{std}(\mathbf{D})/\sqrt{2}$; %The standard deviation of elements in array \mathbf{D}
 - 4: $\lambda = \beta/\sigma^2$
-

3.2. χ^2 test for TV regularization parameter selection. We now develop a χ^2 test for TV regularization which intuitively may not be apparent because the TV term is in L_1 , not L_2 . However, Laplace random variables can form a χ^2 distribution as described in the next two Propositions.

Proposition 2. *If $y_i \sim \text{Laplace}(\theta, \beta)$ then*

$$\frac{2|y_i - \theta|}{\beta} \sim \chi^2_2.$$

Proof. Laplace random variables are equivalent to the difference of two exponential random variables, i.e. for random variable Y with same density as y_i

$$Y = \theta + \beta(X_1 - X_2)$$

where X_1, X_2 are exponentially distributed. This can be seen by manipulating the appropriate characteristic functions [26].

Now the probability density function of a χ^2 random variable h with 2 degrees of freedom is $f(h) = 1/2e^{-h/2}$, for $h \geq 0$. This is also the exponential distribution with parameter 1/2 and we can therefore write

$$Y = \theta + \frac{\beta}{2}(H_1 - H_2)$$

where $H_1, H_2 \sim \chi^2_2$ and i.i.d. [26]. \square

Proposition 3. *If $y_i \sim \text{Laplace}(\theta, \beta)$ $i = 1, \dots, n$, all independent, then*

$$\sum_{i=1}^n \frac{2|y_i - \theta|}{\beta} \sim \chi^2_{2n}.$$

Proof. The sum of n independent χ^2 random variables is also χ^2 with degrees of freedom equal to the sum of degrees of freedom of each term. \square

Theorem 3. *Suppose that $\epsilon_i \sim \mathcal{N}(0, \sigma)$ and $y_i \sim \text{Laplace}(\theta, \beta)$ with $\epsilon = [\epsilon_1, \dots, \epsilon_m]^T$ and $\mathbf{y} = [y_1, \dots, y_n]^T$, each containing independent elements, then*

$$\frac{\|\epsilon\|_2^2}{\sigma^2} + \frac{2\|\mathbf{y}\|_1}{\beta} \sim \chi^2_{m+2n}. \quad (10)$$

Proof. As described in Proposition 1 the first term in the sum is distributed as χ^2_m while the second term is χ^2_{2n} by Theorem 3. Since they are assumed independent their sum is χ^2_{m+2n} . \square

Theorem 3 is illustrated in Figure (1) where histograms of random variables sampled 5000 times are plotted together with theoretical probability distributions. The ϵ_i came from a normal distribution with mean 0 and $\sigma = 0.07$, while y_i came from a Laplace distribution with $\theta = 0$ and $\beta = 100\sigma^2$. In addition there were 300 elements in ϵ and 1000 in \mathbf{y} . In this idealized example, the histograms give excellent agreement with the theoretical probability distributions.

Equation (10) gives a χ^2 test for TV, however, if $\epsilon = \mathbf{d} - \mathbf{A}\hat{\mathbf{x}}$ and $\mathbf{y} = TV(\hat{\mathbf{x}})$, then the terms in the sum are not independent and the degrees of freedom must be reduced. In particular

$$\mathcal{J}_{tv}(\hat{\mathbf{x}}) = \frac{\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\lambda)\|_2^2}{\sigma^2} + \frac{2\|\mathbf{D}\hat{\mathbf{x}}(\lambda)\|_1}{\beta} \sim \chi^2_{mn - df(\mathbf{A}\hat{\mathbf{x}}(\lambda)) + df(\mathbf{D}\hat{\mathbf{x}}(\lambda))}. \quad (11)$$

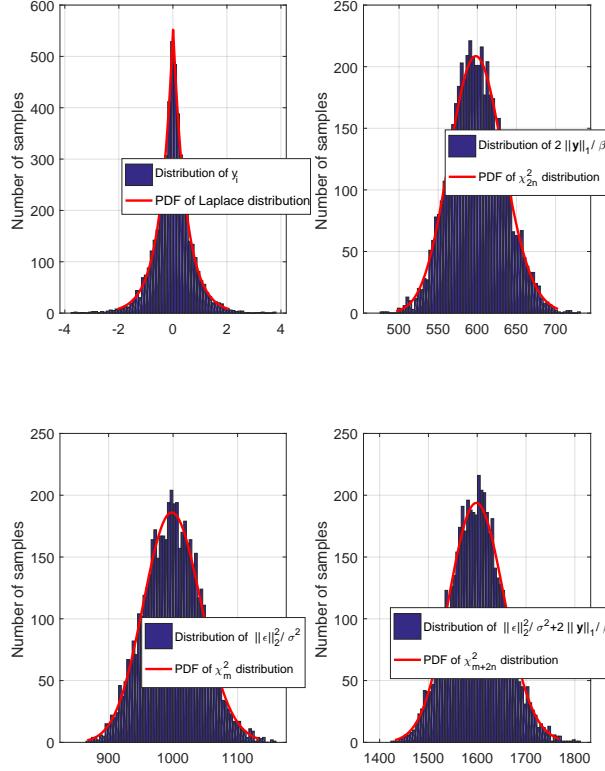


FIGURE 1. Example of Theorem 3

Theorem 4. Suppose that $(\mathbf{d} - \mathbf{Ax})_i \sim \mathcal{N}(0, \sigma)$ and $(\mathbf{Dx})_i \sim \text{Laplace}(\theta, \beta)$ with \mathbf{A} and \mathbf{D} full rank. Then

$$\frac{\|\mathbf{d} - \mathbf{Ax}(\lambda)\|_2^2}{\sigma^2} + \frac{2\|\mathbf{Dx}(\lambda)\|_1}{\beta} \sim \chi_{mn}^2.$$

Proof. It was shown in [44] that

$$df(\hat{\mathbf{x}}(\lambda)) = \mathbb{E}[\text{nullity}(\mathbf{D}_{-\mathcal{A}})],$$

i.e. the degrees of freedom in the TV estimate is the expected value of the dimension of the nullspace of $\mathbf{D}_{-\mathcal{A}}$. They also show that when \mathbf{A} and \mathbf{D} are full rank

$$df(\mathbf{Ax}(\lambda)) = df(\mathbf{Dx}(\lambda)) = df(\hat{\mathbf{x}}(\lambda)).$$

This results in a regularized residual with degrees of freedom

$$mn - df(\mathbf{Ax}(\lambda)) + df(\mathbf{Dx}(\lambda)) = mn.$$

□

Note that Theorem 4 shows that when both \mathbf{A} and \mathbf{D} are full rank, we do not need to calculate $df(\mathbf{Ax}(\lambda))$ when using the regularized residual. This gives in the following χ^2 test for choice of λ in TV regularization:

$$\|\mathbf{d} - \mathbf{Ax}(\lambda)\|_2^2 + \frac{2}{\lambda} \text{TV}(\hat{\mathbf{x}}(\lambda)) \approx mn\sigma^2. \quad (12)$$

Proposition 4. If $\hat{\mathbf{x}}$ has a bounded gradient then

$$f(\lambda) = \|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\lambda)\|_2^2 + \frac{2}{\lambda} \text{TV}(\hat{\mathbf{x}}(\lambda))$$

decreases with increasing λ .

Proof. As λ increases, the data residual decreases since more weight is placed on the data fit in the optimization problem. In addition, since $\text{TV}(\hat{\mathbf{x}})$ remains bounded, the second term in the sum will approach zero for large λ . \square

Remark 1. When $\hat{\mathbf{x}}$ has a bounded gradient it may be the case that $f(\lambda) > mn\sigma^2$ for all λ . Given a large enough value of σ there will be a λ that passes the χ^2 test which implies the data are not as accurate as indicated by the standard deviation estimate σ . Similarly it could be that $f(\lambda) < mn\sigma^2$ for all λ whereby a smaller value of σ is needed for there to exist a root, meaning the data are more accurate than anticipated. If it is possible to find a λ so that $f(\lambda) = mn\sigma^2$ there is a unique λ that passes the χ^2 test.

Remark 2. When $\hat{\mathbf{x}}$ does not have a bounded gradient there may not be a unique λ for which $f(\lambda) = 0$ or $f(\lambda)$ may have multiple extremum. For $\lambda \approx 0$ TV regularization gives $\hat{\mathbf{x}}$ with $\text{TV}(\hat{\mathbf{x}})$ small. As λ increases $\|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\lambda)\|_2^2$ will decrease but $\text{TV}(\hat{\mathbf{x}})$ may increase. A smooth or piecewise constant image results in a small value of $\text{TV}(\hat{\mathbf{x}})$ while large values of $\text{TV}(\hat{\mathbf{x}})$ indicate a highly textured image. This prior understanding of the image can be used to choose a non-unique λ . In particular if the image is assumed to be smooth the best choice of λ is the smallest value for which $f(\lambda) = 0$.

The proposed method for TV regularization parameter selection is given in Algorithm 2. To find the optimum we must solve the TV problem multiple times with different λ to determine which satisfies the χ^2 test. We also give an algorithm for the discrepancy principle in Algorithm 3 which similarly solves the TV problem multiple times with different λ , and automatically computes the regularization parameter as is done in [28]. In both cases there is an outer iteration that finds the regularization parameter according to a discrepancy or χ^2 principle, and each also contains an inner iteration that find the TV parameter estimates. In this work we have described why the discrepancy principle has the incorrect degrees of freedom in Algorithm 3, and developed a χ^2 test for the regularization residual that contains the correct degrees of freedom for TV in Algorithm 2. In both cases we tested the algorithms with inner iterations that find the TV parameter estimates using three publicly available TV codes: **deconvtv** [7], **FTVd** [47] and **tvdeconv** [15]. The first two codes use an augmented Lagrangian method while the third uses the Split Bregman method. Numerical results from Algorithms 2 and 3 are presented in Section 4 along with results from the MAP estimate and the estimate that gives maximum ISNR.

Algorithm 2 TV χ^2 test.

- 1: input \mathbf{d} , \mathbf{A} , σ
 - 2: $\lambda \in \operatorname{argmin}_{\lambda} \left\{ \|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\lambda)\|_2^2 + \frac{2}{\lambda} \text{TV}(\hat{\mathbf{x}}(\lambda)) - mn\sigma^2 \right\}$
where $\hat{\mathbf{x}}(\lambda) \in \operatorname{argmin}_{\mathbf{x}} \left\{ \frac{\lambda}{2} \|\mathbf{d} - \mathbf{A}\mathbf{x}\|_2^2 + \text{TV}(\mathbf{x}) \right\}$
-

Algorithm 3 TV Discrepancy principle.

- 1: input \mathbf{d} , \mathbf{A} , σ
 - 2: $\lambda \in \operatorname{argmin}_{\lambda} \left\{ \|\mathbf{d} - \mathbf{A}\hat{\mathbf{x}}(\lambda)\|_2^2 - mn\sigma^2 \right\}$
where $\hat{\mathbf{x}}(\lambda) \in \operatorname{argmin}_{\mathbf{x}} \left\{ \frac{\lambda}{2} \|\mathbf{d} - \mathbf{A}\mathbf{x}\|_2^2 + \text{TV}(\mathbf{x}) \right\}$
-

4. Numerical Results.

4.1. Test images. The images on which we chose to test the Algorithms are gray-scale with intensity values normalized to the range $[0, 1]$. They were artificially blurred with a discrete linear blur filter \mathbf{A} that is a convolution of the image with a point spread function. We blurred them with Gaussian and uniform lowpass filters with variance 9 and a 15×15 grid respectively. Both were implemented with the MATLAB command *fspecial*. In each case the noise was quantified using the Blurred Signal to Noise Ratio (BSNR):

$$\text{BSNR} = 20 \log_{10} \frac{\|\mathbf{Ax}\|_2}{\|\boldsymbol{\epsilon}\|_2}.$$

We compared BSNRs of 20, 30 and 40 dB, where the lower levels represent more noise.

The three images we chose to test the algorithms were the Camerman image, an artificial MRI, and a Mountain image as seen in Figures 2-4. The blurred images are in the top row of the Figures with BSNR 20, 30 and 40 ranging from left to right. The Camerman image with Gaussian noise is shown in Figure 2 while the MRI and Mountain images are shown with uniform noise in Figures 3-4. In all experiments the standard deviation used to generate noise in the images was also used to reconstruct the images. Although this may not be realistic, it allows us to determine the effectiveness of Algorithms 1-3 in the ideal case.

4.2. ISNR. The optimal value of λ was identified to be that which gives the best Improved Signal-to-Noise Ratio (ISNR) defined by:

$$\text{ISNR} = 20 \log_{10} \frac{\|\mathbf{d} - \mathbf{x}\|_2}{\|\hat{\mathbf{x}}(\lambda) - \mathbf{x}\|_2}. \quad (13)$$

The ISNR is an estimate of the perceptual quality of an image with larger values indicating better quality. The value of λ that gives the largest ISNR is also the value that gives the smallest mean square error (MSE).

Since we have the exact image, we are able to determine the optimal value of λ that gives the highest ISNR value. This is illustrated in Figures 5 and 6. In all Figures the graph associated with the right vertical axis is the ISNR value for a range of λ . There is a clear maximum ISNR value in all cases and the value of λ where this occurs is the λ used to reconstruct the images labeled ‘‘Maximum ISNR’’. The resulting ISNR from reconstructed images using this λ is also reported in Tables 1-2.

Figures 5 and 6 also plot on the left vertical axis is the value the regularized residual in (12), or cost function, for all possible λ along with the targeted value of the cost function which is $m n \sigma^2$. Where the cost function and the targeted value intersect will be the value of λ that satisfies the TV χ^2 test, and the value of λ chosen for the χ^2 method. Results from reconstructed images using this λ are reported in Tables 1-6.

4.3. Reconstructed Images. The images were reconstructed using regularization parameters found by Algorithms 1-3 and the regularization parameter that gives the largest ISNR. These regularization parameters were used to reconstruct the images in the bottom three rows of Figures 2-4. The left column shows results from BSNR of 20, the middle BSNR of 30 and the right BSNR of 40. In all cases the MAP reconstruction looks the worst, while the χ^2 reconstruction has a higher contrast than the reconstruction that produces the best possible ISNR. The reconstructed image using the discrepancy principle looks similar to the χ^2 reconstruction, even though the discrepancy principle has the incorrect degrees of freedom, so the reconstructed image is not shown here.

The resulting ISNR values from the reconstructions are given in Tables 1 and 2. Algorithm 1, using the MAP estimate as suggested in [17], relies on estimating the Laplace distribution parameter β with the data, and in all cases this gives the lowest ISNR value. We also note that the resulting ISNR values when using the discrepancy principle are consistently less than when using the TV χ^2 test. This is to be expected because the discrepancy principle does not have the correct degrees of freedom. However, the resulting ISNR values for the χ^2 test and the discrepancy principle are very close in value. This indicates that $2TV(\hat{\mathbf{x}})/\lambda$ is small and does not significantly change the value of the data residual.

The peak signal to noise ratio (PSNR) was also calculated to measure how close the reconstructed images are to the true image. A higher PSNR value indicates a better reconstruction and these values are given in Tables 3 and 4. The regularization parameter that gives the highest PSNR is the one that also gives the maximum ISNR. The PSNR values resulting from the regularization parameter found with the χ^2 test are consistently larger than discrepancy principle.



FIGURE 2. Gaussian filter with variance 9. BSNR 20 (Left), BSNR 30 (Middle), BSNR 40 (Right).

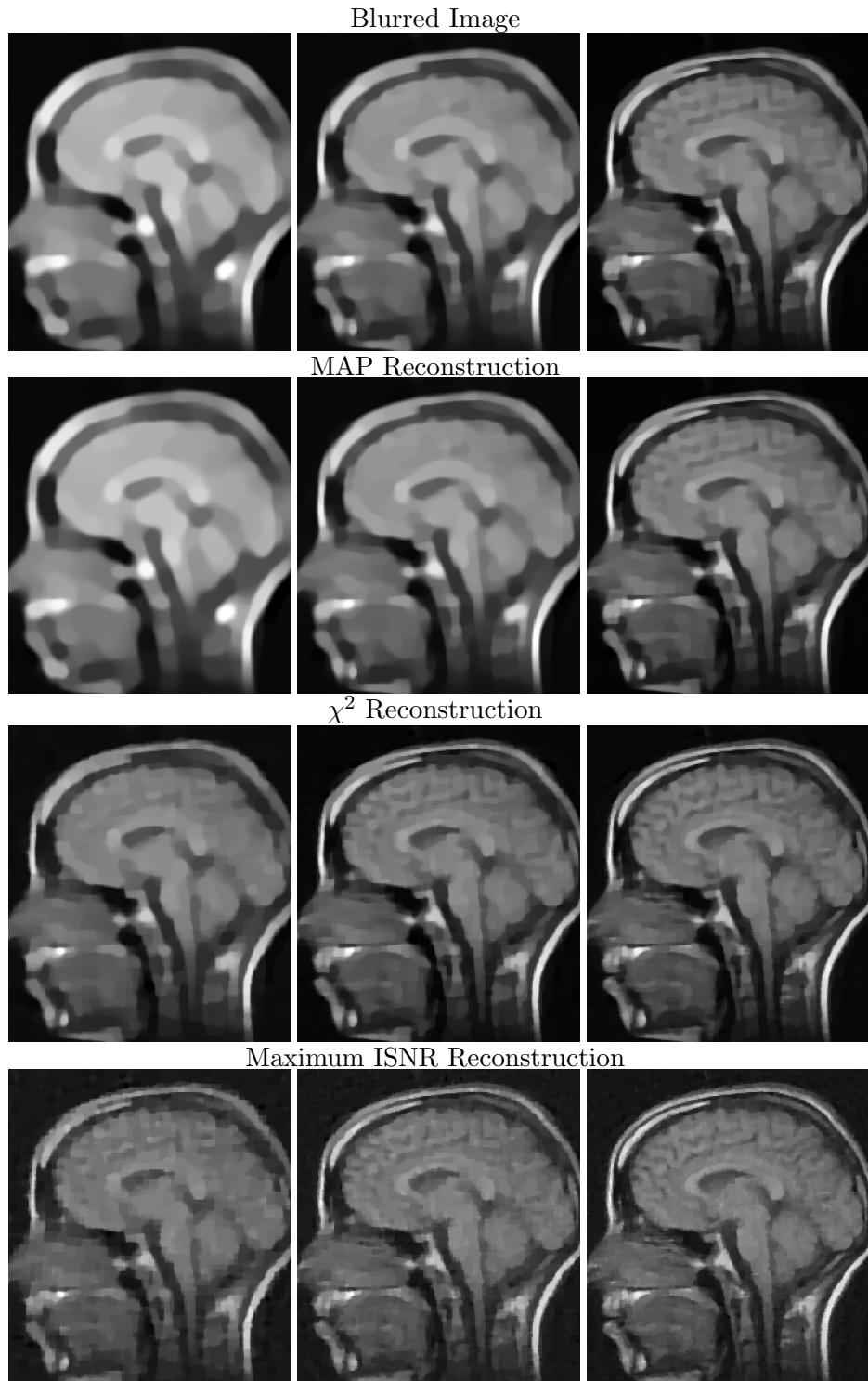


FIGURE 3. Uniform 15 x 15 filter. BSNR 20 (Left), BSNR 30 (Middle), BSNR 40 (Right).

Blurred Image



MAP Reconstruction

 χ^2 Reconstruction

Maximum ISNR Reconstruction



FIGURE 4. Uniform 15 x 15 filter. BSNR 20 (Left), BSNR 30 (Middle), BSNR 40 (Right).

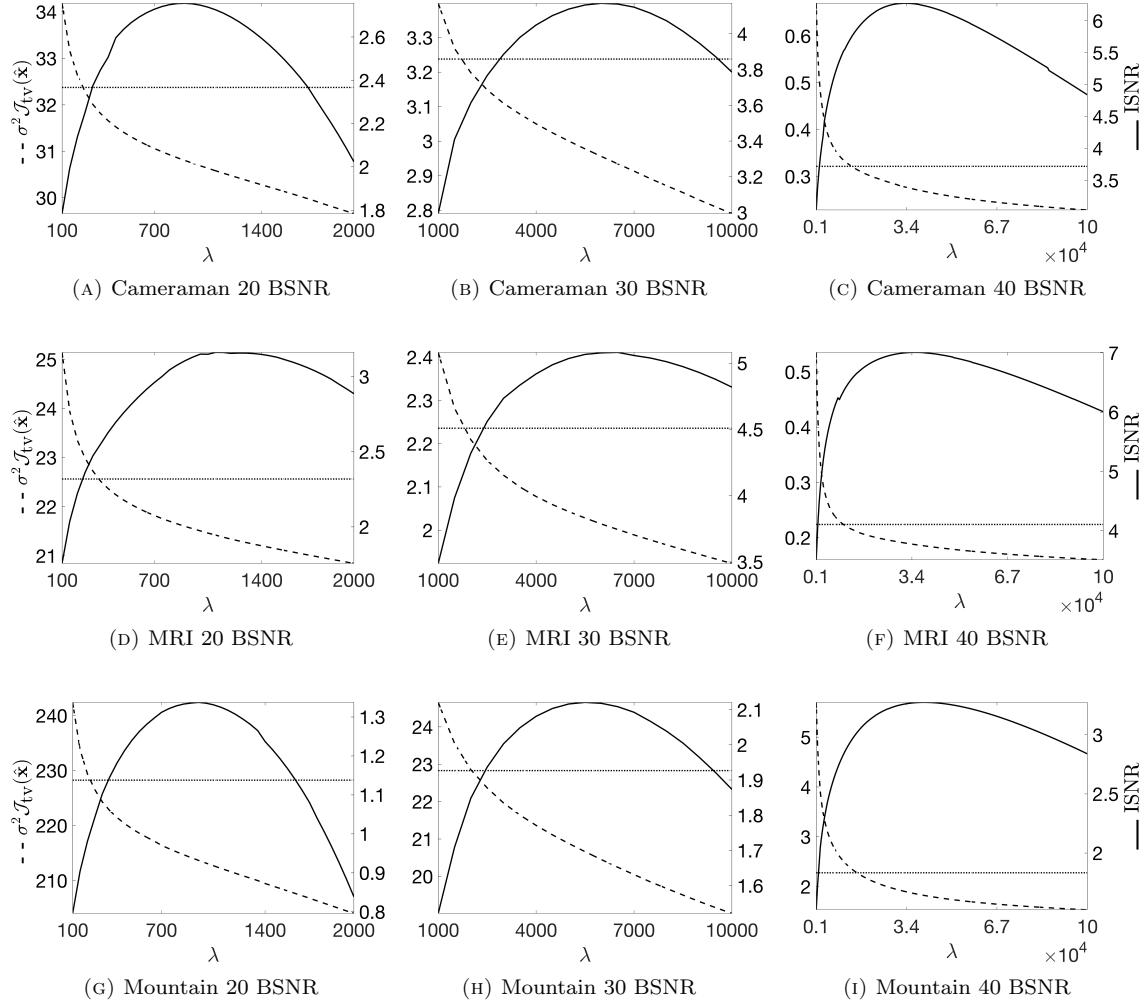


FIGURE 5. Left vertical axis: χ^2 value (--) and target value $mno\sigma^2$ (..). Right vertical axis: ISNR (—). Gaussian filter with variance 9.

The PSNR values for both methods are very similar and have values close to the value when a regularization parameter with maximum ISNR is used.

The structural similarity indices (SSIM) that result when using regularization parameters from each of the methods are given in Tables 5 and 6. It measures the perceived difference between the true and reconstructed images with possible values between 0 (no correlation in the images) and 1. With this metric the regularization parameter that gives maximum ISNR doesn't always result in the highest SSIM. For the cameraman and MRI images with BSNR of 20 and 40 there are instances where the χ^2 test results in a slightly larger SSIM. This difference is reflected in Figures 2-3 where we see more contrast in the figures reconstructed with regularization parameters found by the χ^2 test, than with the regularization parameters that give the best ISNR.

TV is often the preferred method to recover images with fine structures, details, and textures. The mountain image has these properties and we see in Tables 5-6 that the SSIM is consistently smaller for all methods as compared to the other images. The SSIM that results from the mountain

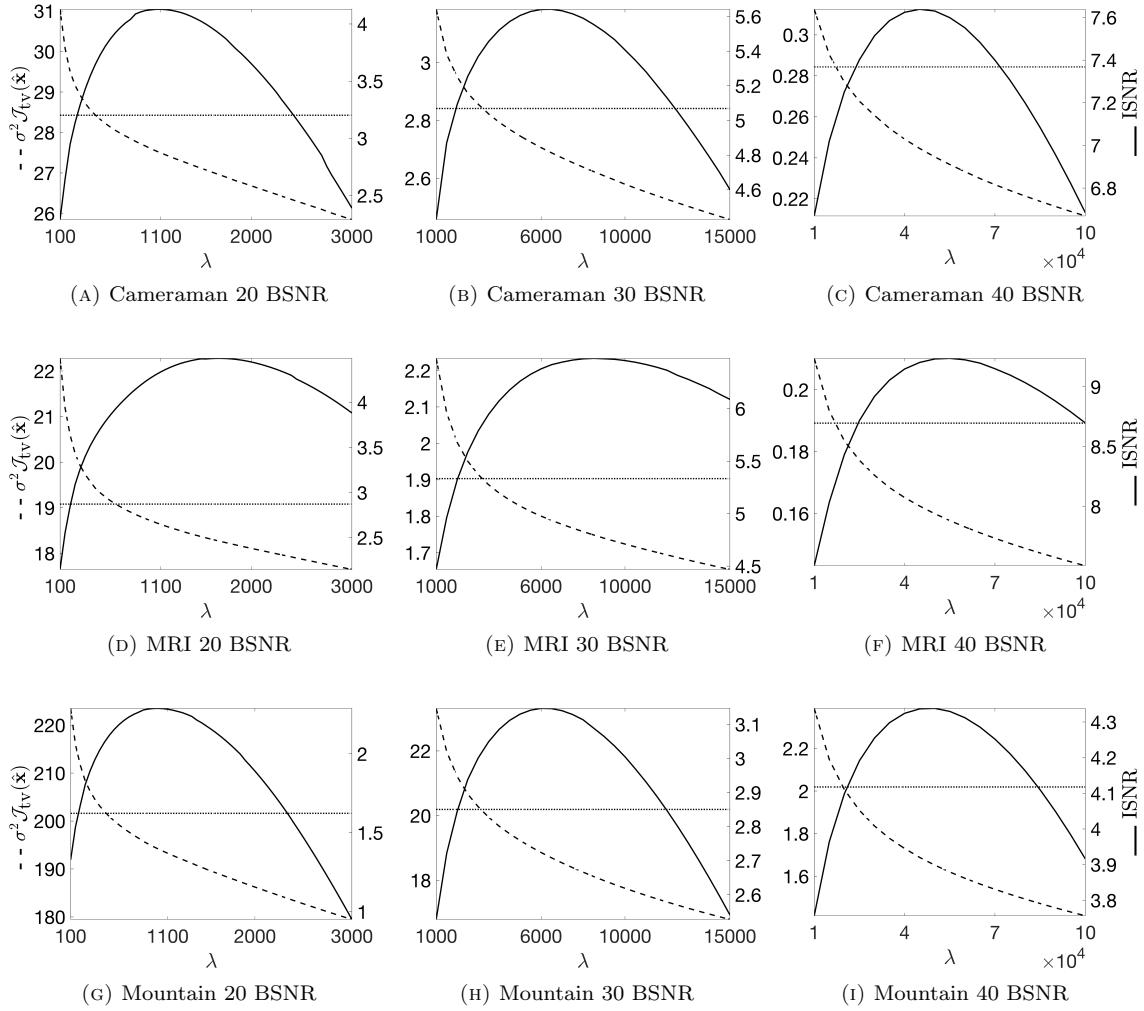


FIGURE 6. Left vertical axis: χ^2 value (--) and target value $m n \sigma^2$ (..). Right vertical axis: ISNR (—). Uniform filter 15 x 15

image for all BSNR values is largest when using the maximum ISNR method, while the SSIM is smaller when using discrepancy principle than when using the χ^2 test. However, values of $TV(\hat{x})$ are not large enough to significantly differentiate the discrepancy principle from the χ^2 test.

5. Summary and Conclusions. The effectiveness of TV relies on appropriate choice of regularization parameter λ as defined by (7). If λ is too small the image will not appropriately fit the data while if it is too large, the image will be under smoothed. A suitable choice of λ is often found by trying different values until one is found that results in a clear image. This is not the most efficient way of choosing the parameter, and may not be an option if the TV optimization problem is part of a greater iteration scheme, such as that for nonlinear problems. In addition, when the exact image is unknown, it may be the case that even though the image is clear, it does not correctly identify key features. Here we propose an automated method for regularization parameter selection that utilizes statistical properties of TV and can be viewed as finding a value of λ that passes a χ^2 test.

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum
Cameraman ($m = n = 256$)				
40	3.8419	5.5893	5.6180	6.2762
30	2.1214	3.4184	3.4507	4.2052
20	1.4033	2.1465	2.1675	2.7272
MRI ($m = n = 256$)				
40	5.2304	6.2520	6.3032	7.0209
30	2.5435	4.2183	4.2656	5.0079
20	1.4242	2.5236	2.5483	3.2903
Mountain ($m = 480, n = 640$)				
40	2.0275	3.0206	3.0242	3.2735
30	0.9674	1.8473	1.8526	2.1171
20	0.5623	1.0222	1.0265	1.3306

TABLE 1. ISNR: Gaussian blur with variance 9.

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum
Cameraman ($m = n = 256$)				
40	5.0019	7.0914	7.1123	7.6329
30	2.8671	5.3398	5.3556	5.6426
20	1.8228	3.6031	3.6241	4.0441
MRI ($m = n = 256$)				
40	5.6696	8.1718	8.2201	9.2978
30	3.2113	5.9225	5.9510	6.5944
20	1.7017	3.8260	3.8474	4.5641
Mountain ($m = 480, n = 640$)				
40	2.8357	4.0904	4.0938	4.3440
30	1.6074	2.9915	2.9945	3.1432
20	0.9594	2.1004	2.1049	2.2803

TABLE 2. ISNR: Uniform 15 x 15

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum ISNR
Cameraman ($m = n = 256$)				
40.0	25.3775	27.1479	27.1772	27.8249
30.0	23.6516	24.9772	25.0077	25.6775
20.0	22.6682	23.3500	23.3718	24.0393
MRI ($m = n = 256$)				
40.0	27.7080	28.8271	28.8750	29.5010
30.0	24.9866	26.7687	26.8176	27.5807
20.0	23.6173	24.6765	24.7007	25.4519
Mountain ($m = 480, n = 640$)				
40.0	19.0529	20.0588	20.0624	20.3040
30.0	17.9755	18.8666	18.8720	19.1485
20.0	17.4244	17.8787	17.8830	18.1806

TABLE 3. PSNR: Gaussian blur with variance 9.

The TV estimate is the MAP estimate if the regularization term is viewed as prior information about pixel differences from a Laplace distribution. This prior information assumption is common in the statistical modeling of natural images [41]. In this work we exploit this assumption to show that the objective function for TV regularization follows a χ^2 distribution. The discrepancy principle similarly exploits χ^2 properties of the least squares data residual. However, when data are used to find a regularization parameter, the degrees of freedom in the χ^2 test must be reduced, otherwise the image will be over-smoothed. For Tikhonov regularization the analytical expression for the minimum of the objective function can be used to determine the equivalent degrees of

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum ISNR
Camerman ($m = n = 256$)				
40.0	24.1719 (9.84)	26.3170 (1.84)	26.3371 (1.77)	26.8111
30.0	22.0192 (11.10)	24.4882 (1.14)	24.5037 (1.08)	24.7709
20.0	20.8288 (9.70)	22.5266 (2.34)	22.5493 (2.24)	23.0660
MRI ($m = n = 256$)				
40.0	25.0072 (12.82)	27.5739 (3.87)	27.6225 (3.70)	28.6832
30.0	22.5294 (13.29)	25.1393 (3.24)	25.1710 (3.12)	25.9823
20.0	20.9540 (11.69)	23.1799 (2.31)	23.1974 (2.23)	23.7269
Mountain ($m = 480, n = 640$)				
40.0	18.3899 (7.56)	19.6504 (1.22)	19.6538 (1.20)	19.8931
30.0	17.1524 (8.24)	18.5660 (0.68)	18.5688 (0.66)	18.6924
20.0	16.4198 (7.48)	17.5595 (1.06)	17.5633 (1.04)	17.7470

TABLE 4. PSNR: Uniform 15 x 15. The value in parentheses is the percent relative difference from the PSNR obtained with the λ that achieves maximum ISNR.

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum ISNR
Camerman ($m = n = 256$)				
40.0	0.7956	0.8402	0.8408	0.8379
30.0	0.7393	0.7809	0.7818	0.7855
20.0	0.7024	0.7271	0.7278	0.7327
MRI ($m = n = 256$)				
40.0	0.8496	0.8717	0.8725	0.8767
30.0	0.7846	0.8279	0.8288	0.8332
20.0	0.7246	0.7686	0.7693	0.7607
Mountain ($m = 480, n = 640$)				
40.0	0.5301	0.6197	0.6201	0.6343
30.0	0.4440	0.5156	0.5160	0.5377
20.0	0.3868	0.4348	0.4352	0.4526

TABLE 5. SSIM: Gaussian blur with variance 9.

BSNR	MAP estimate	Discrepancy	χ^2 test	Maximum ISNR
Camerman ($m = n = 256$)				
40.0	0.7571	0.8281	0.8285	0.8158
30.0	0.6766	0.7660	0.7662	0.7628
20.0	0.6300	0.6931	0.6938	0.6984
MRI ($m = n = 256$)				
40.0	0.7683	0.8341	0.8349	0.8270
30.0	0.6732	0.7694	0.7702	0.7553
20.0	0.6037	0.6934	0.6938	0.6706
Mountain ($m = 480, n = 640$)				
40.0	0.4609	0.5834	0.5837	0.5965
30.0	0.3602	0.4801	0.4804	0.4906
20.0	0.3060	0.3901	0.3904	0.4003

TABLE 6. SSIM: Uniform 15 x 15

freedom [46]. Since there is no such analytical expression for the TV minimizer it has been an open problem to determine the appropriate degrees of freedom for the TV χ^2 test. Here we use recent results concerning the degrees of freedom for the lasso fit [44] to derive a χ^2 test with appropriately reduced degrees of freedom for TV regularization parameter selection.

Results were shown for three different images, two types of noise, three noise levels and three TV algorithms: **deconvtv** [7], **FTVd** [47] and **tvdeconv** [15]. In each case the χ^2 test for regularization parameter selection, Algorithm 2, performed best. In addition, since the exact images are known, we also computed regularization parameters that gave optimal ISNR and showed that the χ^2 test gives near optimal parameters.

In all test cases Algorithm 1, which uses a rough estimate of the hyperparameter in TV as a MAP estimate, gave the worse results. Algorithm 3 uses the discrepancy principle with incorrect degrees of freedom and it consistently performed worse than the χ^2 test in Algorithm 2. However, the improvement was not significant because $TV(\hat{\mathbf{x}})$ is typically small compared to the L_2 norm of the data residual. While the discrepancy principle with incorrect degrees of freedom gave similar results to the χ^2 test we do not recommend using the discrepancy principle in this manner because images with large $TV(\hat{\mathbf{x}})$ will be over-smoothed. In addition, there is no significant increase in computational cost to use the correct χ^2 test over the discrepancy principle since the only difference in the methods is evaluation of the TV regularization operator at the parameter estimate.

6. Acknowledgments. I would like to thank the authors of **deconvtv**, **FTVd** and **tvdeconv** for making their source codes freely available. In addition, I am thankful to Prof. Renaut and Anna Nelson for stimulating discussions that led to this work.

REFERENCES

- [1] A. Ali, R. Tibshirani, The Generalized Lasso Problem and Uniqueness, submitted.
- [2] R.C. Aster, B. Borchers and C. Thurber, Parameter Estimation and Inverse Problems, Academic Press, 2005.
- [3] S.D. Dabacan, R. Molina, and A.K. Katsaggelos, Variational Bayesian Blind Deconvolution Using a Total Variation Prior, *IEEE Trans. Imag. Proc.*, **18**, 12-26, 2009.
- [4] T. Blu and F. Luisier, The SURE-LET Approach to Image Denoising, *IEEE Trans. Imag. Proc.*, **16**, No. 11, 2778-2786, 2007.
- [5] G. Casella and R. Berger, Statistical Inference, Duxbury, 2001.
- [6] J.M. Bioucas-Dias, M.A.T. Figueiredo and J.P. Oliveira, Adaptive total variation image deconvolution: A majorization-minimization approach, Proc. EUSIPCO, Florence, Italy, 2006.
- [7] S.H. Chan, R. Khoshabeh, K.B. Gibson, P.E. Gill and T.Q. Nguyen, An augmented Lagrangian method for total variation video restoration, *IEEE Trans. Image Process.*, **20**, 3097-3111, 2011.
- [8] J., Dahl, P.C. Hansen, S.H. Jensen and T.L. Jensen, Algorithms and software for total variation image reconstruction via first-order methods, *Numer. Alg.*, **53**, 67-92, 2010.
- [9] C.A. Deledalle, S. Vaiter, J. Fadili and G. Peyré, Stein Unbiased GrAdient estimator of the Risk (SUGAR) for Multiple Parameter Selection, *SIAM J. Imag. Sci.* , **7**, No. 4, 2448-2487, 2014.
- [10] J.C. De los Reyes and C.B. Schönlieb, Image denoising: Learning the noise model via non-smooth PDE-constrained optimization, *Inv. Prob. Imag.* , **7**, No. 4, 1183-1214, 2013.
- [11] N. Dey, L. Blanc-Feraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin and J. Zerubia, “Richardson–Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution”, *Microsc. Res. Tech.* , 69: 260-266, 2006.
- [12] Y. Dong, Kintermüller and M.M. Rincon-Camacho, Automated Regularization Parameter Selection in Multi-Scale Total Variation Model for Image Restoration, *J. Math Imaging Vis.*, **40**, 82-104, 2011.
- [13] C. Dossal, M. Kachour, M. Fadili, G. Peyré and C. Chesneau, The Degrees of Freedom of the Lasso for General Design Matrix, *Statistica Sinica*, **23**, 809-828, 2013.
- [14] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression, *Ann. Statist.*, **32** , 407-499, 2004.
- [15] P. Getreuer, Rudin-Osher-Fatemi Total Variation Denoising using Split Bregman, *Image Processing On Line*. DOI: 10.5201/ipol.2012.g-tvd, 2012.
- [16] P. Getreuer, Total Variation Deconvolution using Split Bregman, *Image Processing On Line*. DOI: 10.5201/ipol.2012.g-tvdc, 2012.
- [17] M.L., Green, *Statistics of Images, the TV Algorithm of Rudin-Osher-Fatemi for Image Denoising and an Improved Denoising Algorithm*, UCLA CAM 02-55, 2002.
- [18] T. Goldstein and S.J. Osher, *The Split Bregman Method for L^1 Regularized Problems*, UCLA CAM Report 08-29, 2008.
- [19] P. Hall and D.M. Titterington, Common Structure of Techniques for Choosing Smoothing Parameters in Regression Problems, *J. Royal Stat Soc B*, **49**, 184-198, 1987.
- [20] P. C. Hansen, , J. G. Nagy and D. P. O’Leary, Deblurring Images, Matrices, Spectra and Filtering, SIAM, Philadelphia, 2006.
- [21] P.C. Hansen, Regularization Tools Version 4.0 for Matlab 7.3, *Numerical Algorithms*, **46**, 189-194, 2007.
- [22] M. Hintermüller and C.N. Rautenberg, Optimal Selection of the Regularization Function in a Weighted Total Variation Model. Part I: Modelling and Theory, *J. Math. Imag. and Vision*, Vol. 59, 3, 498–514 , 2017.
- [23] M. Hintermüller, C.N. Rautenberg, T. Wu and A. Langer, Optimal Selection of the Regularization Function in a Weighted Total Variation Model. Part II: Algorithm, Its Analysis and Numerical Tests, *J. Math. Imag. and Vision*, Vol. 59, 3, 515–533 , 2017.
- [24] L. Huang and D. Mumfor, Statistics of Natural Images and Models, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 541-547, 1999.
- [25] H. Liao, , L. Fang and M.K. Ng, Selection of regularization parameter in total variation image restoration, *J. Opt. Soc. Am. A*, **26**, 2311-2320, 2009.
- [26] S. Kotz, T. Kozubowski and K. Podgorski, The Laplace Distribution and Generalizations, Birkhäuser, 2001.

- [27] K. Kunisch and T. Pock, A Bilevel Optimization Approach for Parameter Learning in Variational Models, *SIAM J. Imag. Sci.*, Vol. 6, No. 22, 938–983, 2013.
- [28] A. Langer, Automated Parameter Selection for Total Variation Minimization in Image Restoration, *J. Math. Imag. and Vision*, **57**, 239–268, 2017.
- [29] J. Lee and P.K. Kitanidis, Bayesian inversion with total variation prior for discrete geologic structure identification, *Water Resour. Res.*, **49**, 7658-7669, 2013.
- [30] Y. Lin, B. Wohlberg and G. Hongbin , UPRE method for total variation parameter selection, *Signal Processing*, **90**, 2546-2551, 2010.
- [31] D. W. Marquardt, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, **12**, 591-612, 1970.
- [32] J. Mead, Parameter estimation: A new approach to weighting a priori information, *J. Inv. Ill-posed Problems*, **16**, 175-194, 2008.
- [33] J. Mead and R. A. Renaut, A Newton root-finding algorithm for estimating the regularization parameter for solving ill-conditioned least squares problems, *Inverse Problems*, **25**, No. 2, 2008.
- [34] J. Mead and R. A. Renaut, Least squares problems with inequality constraints as quadratic constraints, *Linear Algebra and its Applications*, **432**, 1936-1949, 2010.
- [35] J. Mead, Discontinuous parameter estimates with least squares estimators, *Applied Mathematics and Computation*, **219** , 5210-5223, 2013.
- [36] J. Mead and C. Hammerquist, Chi-squared tests for choice of regularization in nonlinear inverse problems, *SIMAX*, **34**, 1213-1230, 2013.
- [37] Morozov, Methods for Solving Incorrectly Posed Problems, Springer Verlag, 1984.
- [38] M. K. Ng, P. Weiss and X. Yuan, “Solving Constrained Total-variation Image Restoration and Reconstruction Problems via Alternating Direction Methods”, *SIAM J. Sci. Comput.* 32:5, 2710-2736, 2010.
- [39] J.P. Oliveira, J.M Bioucas-Dias and M.A.T. Figueiredo, Adaptive total variation image de-blurring: A majorization-minimization approach, *Signal Processing*, **89**, 1683-1693, 2006.
- [40] L.I. Rudin, S. Osher and E. Fatemi , Nonlinear total variation based noise removal algorithms, *Physica D*, **60**, 259-268, 1992.
- [41] A. Srivastava, A. Lee, , E. Simoncelli and S.-C. Zhu, On advances in statistical modeling of natural images, *J. Mathematical Imaging and Vision*, **18**, 17-33, 2003.
- [42] T. Teuber, G. Steidl and R. H. Chan, “Minimization and parameter estimation for seminorm regularization models with I-divergence constraints”, *Inverse Problems*, **29**, 3, 2013.
- [43] R.J. Tibshirani and J. Taylor, The solution path of the generalized lasso, *The Annals of Stat.*, **33**, 1335-1371, 2011.
- [44] R.J. Tibshirani and J. Taylor, Degrees of Freedom in Lasson Problems, *The Annals of Stat.*, **40**, 1198-1232, 2012.
- [45] D.M. Titterington, Choosing the regularization parameter in image restoration, *Lecture Notes-Monograph Series, JSTOR*, **20**, 392-402, 1991.
- [46] G. Wahba, Bayesian confidence intervals for the cross-validated smoothing spline, *J. R. Statist. Soc. B*, **45**, 133-150, 1983.
- [47] Y. Wang, J. Yang, W. Yin and Y. Zhang, A New Alternating Minimization Algorithm for Total Variation Image Reconstruction, *SIAM J. Imag. Sci.*, **1**, 248-272, 2008.
- [48] Y-W. Wen and R.H. Chan, Parameter Selection for Total-Variation-Based Image Restoration Using Discrepancy Principle, *IEEE Trans. Imag. Proc.*, **21**, 1770-1781, 2012.

E-mail address: jmead@boisestate.edu