Proceedings of the Project Review, Geo-Mathematical Imaging Group (Purdue University, West Lafayette IN), Vol. 1 (2011) pp. 1-20.

DATA ANALYSIS TOOLS FOR UNCERTAINTY QUANTIFICATION OF INVERSE PROBLEMS

LUIS TENORIO*, FREDRIK ANDERSSON†, MAARTEN V. DE HOOP‡, AND PING MA§

Abstract. We present exploratory data analysis methods to assess inversion estimates using examples based on ℓ^2- and ℓ^1-regularization. These methods can be used to reveal the presence of systematic errors such as bias and discretization effects, or to validate assumptions made on the statistical model used in the analysis. The methods include: bounds on the performance of randomized estimators of a large matrix, confidence intervals and bounds for the bias, resampling methods for model validation, and construction of training sets of functions with controlled local regularity.

1. Introduction. The analysis of physical problems based on experimental data commonly relies on idealized mathematical and statistical models. We consider the particular framework of inverse problems where the data are modeled as a vector $\boldsymbol{y}$ related to an unknown function f via an equation of the form $\boldsymbol{y} = A[f] + \boldsymbol{\varepsilon}$, where A is a linear operator defined on a linear space and $\boldsymbol{\varepsilon}$ is a random noise vector. We assume that the problem is ill-posed; A does not have a bounded inverse. Since the measurements as well as the assumptions on A and $\boldsymbol{\varepsilon}$ are subject to error, any solution of the inverse problem must include some assessment of its associated uncertainty. This uncertainty is usually quantified using statistical methods that lead to formal statements of uncertainty However, these methods rely on the validity of the modeling assumptions.

For example, consider the typical Bayesian framework for uncertainty quantification of inverse problems that can be summarized as follows: Select a probability distribution for the noise and a prior distribution for the unknown function; model all the available information using random variables with some probability distributions; use Bayes's theorem to define the posterior distribution of the function given the data and Markov chain Monte Carlo methods to sample from it. Uncertainty quantification is then based on the variability under the posterior distribution. Since this procedure provides a statement of uncertainty even when the choice of distributions is inadequate and inconsistent with the observations, it is important to validate all the assumptions.

Similarly, there is a variety of regularization techniques that can be used to find solutions of ill-posed inverse problems. Some of these techniques have known rates of convergence and other good asymptotic properties provided some regularity conditions are satisfied (e.g., [7, 15]). But in any practical problem we should always question if is there any reason not to trust an estimate $\widehat{f}$ of the function f given the fixed sample size and modeling assumptions. That is, we should determine if the features we see in the estimate can be explained by systematic errors or random variability.

To avoid drawing false conclusions, any uncertainty quantification should include a validation of the modeling assumptions that may point to, for example, problems with the calibration of the instruments or the presence of systematic errors (e.g., unmodeled physics). One of our goals is to call attention to the question of model validation for ill-posed inverse problems. We provide validation methods in the framework of ℓ^2- and ℓ^1-regularization.

As it is often done in applications, we assume that the inverse problem has been discretized. This assumption is especially convenient in the statistical analysis as it frees us from having to define probability measures on function spaces. The data are modeled as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{f}$ is the discretized function to be recovered, $\boldsymbol{A}$ is a possibly ill-conditioned matrix, and $\boldsymbol{\varepsilon}$ is the noise vector modeled as random with zero-mean and covariance matrix $\sigma^2\boldsymbol{I}$; assumptions that

*Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO, USA, ltenorio@mines.edu

†Centre for Mathematical Sciences, Lund University, Lund, Sweden, fa@maths.lth.se

‡Center for Computational and Applied Mathematics, Purdue University, West Lafayette, IN, USA, mdehoop@math.purdue.edu

§Department of Statistics, University of Illinois, Urbana-Champaign, IL, USA, pingma@illinois.edu

have to be validated. For example, the mean of the errors will not be zero if there is a significant discretization effect. This can be checked using, for example, the confidence intervals for the bias defined in Section 2, or the resampling methods we present in Section 3.

Since model validation techniques depend on the chosen regularization method, to present concrete examples we have chosen two different types of inversion estimates that are often used in practice and that we shall call ℓ^2- and ℓ^1-regularization. The former is based on the idea of finding a 'smooth' function that is consistent with the data. This can be done via a discrete Tikhonov approach where the estimate $\widehat{\boldsymbol{f}}_{\ell^2}$ is chosen so as to balance an ℓ^2-data-misfit with an ℓ^2-smoothness penalty:

$$\widehat{\boldsymbol{f}}_{\ell^2} = \arg\min_{\widetilde{\boldsymbol{f}}} \| \boldsymbol{y} - \boldsymbol{A}\widetilde{\boldsymbol{f}}\|_2^2 + \lambda^2 \| \boldsymbol{D}\widetilde{\boldsymbol{f}}\|_2^2, \tag{1.2}$$

where $\boldsymbol{D}$ is a finite-difference operator of some order and λ is a regularization parameter. In Section 3 we also consider a Bayesian framework where $\widehat{\boldsymbol{f}}_{\ell^2}$ is a posterior mean. The second approach exploits sparse representations of $\boldsymbol{f}$ to regularize the problem: $\boldsymbol{f}$ is assumed to have a representation $\boldsymbol{f} = \boldsymbol{W}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of coefficients in the sparse representation. An ℓ^1-norm penalty is used to promote sparsity of the vector of coefficients (see, e.g., [1, 21, 30]). If $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{W}$, then the ℓ^1-estimate of $\boldsymbol{f}$ is $\widehat{\boldsymbol{f}}_{\ell^1} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}$, where

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \| \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} \|_2^2 + \lambda \| \widetilde{\boldsymbol{\beta}} \|_1. \tag{1.3}$$

A variety of methods have been developed for model validation and uncertainty quantification for classic linear regression (e.g., well-posed linear inverse problems [2, 6]) and nonparametric regression (e.g., [20]) but these questions are more difficult in the framework of ill-posed inverse problems. One of the difficulties is that in applications such as 2D and 3D geophysical inversions the inverse problems are large-scale and routine tasks such as computing the trace of a matrix or extracting its diagonal entries become computationally demanding. Another complication is that regularized estimates may be subject to significant bias (more so than in nonparametric regression); a component of the error that is difficult to assess for it depends on the object to be recovered.

The paper is organized as follows. In Section 2 we present some auxiliary tools that are useful in the analysis of large-scale problems. We summarize robust recursive estimates of location and scale that are used to analyze simulation results and study the stability of inversion estimates. We also review the derivation of randomized estimators of the trace of a large matrix. The use of such estimators requires the selection of the number of random realizations to be averaged in the approximation. We provide bounds (based on the behavior of the eigenvalues) for the relative variance and concentration of the trace estimator that help answer this question. In Section 3 we present confidence intervals for the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}$ as an estimator of $\boldsymbol{A}\boldsymbol{f}$ which helps us check if $\widehat{\boldsymbol{f}}$, even if biased, is still consistent with the data. To validate the estimate, we also present examples of frequentist and Bayesian resampling approaches to create synthetic predictions that are compared to the observations. Up to Section 3, we do model validation by checking consistency with the data. In Section 4 we consider methods to assess characetristics of the estimate $\widehat{\boldsymbol{f}}$ itself, which requires more assumptions on $\boldsymbol{f}$. We derive bounds that provide information about the geometry of the bias of ℓ^2- or linearized ℓ^1-estimates. We also discuss applications of wavelet characterizations of function regularity to generate training sets of functions with controlled regularity. These functions can be used to assess the relative errors one may expect in estimates of $\boldsymbol{f}$. The paper concludes with a summary in Section 5.

2. Auxiliary tools for large-scale simulations.

2.1. Exploring stability. Most of the methods we describe are based on simulations. At each simulation run j we obtain an estimate $\widehat{\boldsymbol{f}}_j$ and functions thereof. If these quantities were one-dimensional, one could store thousands of them and then study their simulation distributions. But this cannot be easily done with large-scale problems such as 2D or 3D data. In this case we may

be limited to estimating properties of their distributions that can be computed recursively. Clear choices with well known recursive formulas are the mean and standard deviation, which provide location and scale summaries of the distributions. To check for stability, asymmetry or the presence of outliers, it is advisable to also compute more robust measures of location and scale. Here we consider the median and median absolute deviation from the median (MAD) for which there are recursive approximations. The method we use is based on the algorithms described in [9, 16], which have been shown to have similar asymptotic properties to those of the non-recursive sample median and MAD. For completeness, the procedure for a sequence of $N \times N$ images $\boldsymbol{T}_i$ is described in Algorithm 1.

Algorithm 1 Recursive estimates of the median and MAD for a sequence $\boldsymbol{T}_i$ of $N \times N$ images. For simplicity some of the operations are written as MATLAB commands.

Initialize: $b = 0.2$, $g_1 = \text{median} = g_2 = \texttt{zeros}(N, N)$
Initialize: $s_1 = s_2 = \text{mad} = \texttt{ones}(N, N) * 10^{-5}$
for $k = 1 : \text{sims}$ **do**
  $s_1 = (k-1) * s_1/k + |\, \boldsymbol{T}_k - s_1 \,|/k$
  $\ell_1 = \big(|\text{median} - \boldsymbol{T}_k| < s_1/k^b \big)$
  $g_1 = (1 - 1/k) * g_1 + k^{b-1}. * \ell_1./s_1$
  $c = 0.1 * k^{-1/4}$
  $a_1 = \texttt{max}(c./s_1,\, g_1)$
  $\text{median} = \text{median} + \texttt{sign}(\boldsymbol{T}_k - \text{median})./(k * a_1)$
  $z = |\, \boldsymbol{T}_k - \text{median} \,|$
  $s_2 = (k-1) * s_2/k + |\, z - s_2 \,|/k$
  $g_2 = (1 - 1/k) * g_2 + k^{b-1}. * \ell_2./s_2$
  $a_2 = \texttt{max}(c./s_2,\, g_2)$
  $\text{mad} = \text{mad} + \texttt{sign}(z - \text{mad})./(k * a_2)$
end for

As an illustration we compare the stability of some ℓ^2- and ℓ^1-regularized estimates. To compute $\widehat{\boldsymbol{f}}_{\ell^2}$ given the $n \times 1$ data vector $\boldsymbol{y}$, we use the value of λ that minimizes the generalized cross-validation (GCV) function [28, 41]: $\text{GCV}(\lambda) = \| \boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{f}}_\lambda \|^2 / [\, n - \text{tr}\, \boldsymbol{H}(\lambda) \,]^2$, where $\boldsymbol{H}(\lambda) = \boldsymbol{A}\,(\boldsymbol{A}^t\boldsymbol{A} + \lambda^2 \boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{A}^t$. Since the nonlinearity introduced by the ℓ^1-penalty in (1.3) does not lead to a simple formulation of a GCV-type function, we use Morozov's discrepancy principle for which there are some efficient implementations [40]. The idea is to choose λ so as to minimize the ℓ^1-norm of $\boldsymbol{\beta}$ subject to fitting the data to within the noise level:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \|\widetilde{\boldsymbol{\beta}}\|_1 \quad \text{s.t.} \ \ \|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|^2 \le n\sigma^2. \tag{2.1}$$

The tolerance $n\sigma^2$ is what is used most often as it is the expected value of the data-misfit norm for iid Gaussian errors: $\mathbb{E}\,\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \mathbb{E}\,\|\boldsymbol{\varepsilon}\|^2 = n\,\sigma^2$. Since σ^2 is unknown, we use a data-based estimate obtained as follows: σ is the standard deviation of the variability of y_i about $(\boldsymbol{A}\boldsymbol{f})_i$. If $\boldsymbol{A}\boldsymbol{f}$ is the discretization of a function that is expected to be smoother than that corresponding to $\boldsymbol{f}$, then it makes sense to use the Tikhonov estimate (1.2) with $\boldsymbol{A} = \boldsymbol{I}$. That is, the data are modeled as noisy direct observations of $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f}$ and $\boldsymbol{D}$ acts on $\boldsymbol{\mu}$. The estimate of σ^2 is then

$$\widehat{\sigma}^2 = \frac{\| \boldsymbol{y} - \widehat{\boldsymbol{\mu}}_\lambda \|^2}{n - \text{tr}\,[\,(\boldsymbol{I} + \lambda^2\boldsymbol{D}^t\boldsymbol{D})^{-1}\,]}. \tag{2.2}$$

The degrees of freedom of $\widehat{\boldsymbol{\mu}}_\lambda$ (as defined in [37]) is $\text{tr}\,[\,(\boldsymbol{I} + \lambda^2\boldsymbol{D}^t\boldsymbol{D})^{-1}\,]$ so $\widehat{\sigma}^2$ is the residual sum of squares normalized by an 'effective number' of observations; the same formula as in linear regression. The estimate $\widehat{\sigma}^2$ is used in place of σ^2 in (2.1). Of course, in some application $\boldsymbol{A}\boldsymbol{f}$ may not be smooth. In these case there are other methods to estimate σ depending on the characteristics of $\boldsymbol{A}\boldsymbol{f}$. See, for example, [11, 17, 20, 27, 33].

In the following example we compare the stability of ℓ^2- and ℓ^1-estimates. The point is not to compare the two regularization methods but to illustrate the information that can be obtained by comparing robust to non-robust simulation results.

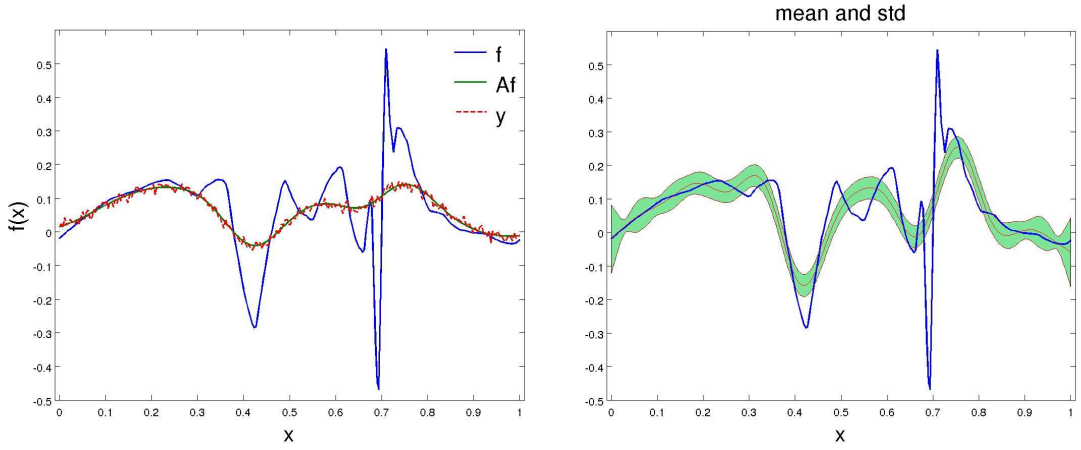

FIG. 1. *Left:* $\boldsymbol{f}$, $\boldsymbol{Af}$ *and a sample* $\boldsymbol{y}$. *Right: Simulation results. The boundary of the band is the mean and mean +/− 2(std) over the simulations of* $\widehat{\boldsymbol{f}}_{\ell^2}$ *with* λ *fixed.*

EXAMPLE 1. We simulate one-dimensional data using the model (1.1) with the matrix $\boldsymbol{A}$ of a Gaussian kernel $\propto e^{-(x-y)^2/2\gamma^2}$ ($\gamma = 0.08$) and a vector $\boldsymbol{f}$ of discretized values of a function f with a sparse wavelet representation in the Daubechies-4 basis. The left panel in Figure 1 shows $\boldsymbol{f}$, $\boldsymbol{Af}$ and a realization of $\boldsymbol{y}$. The right panel shows the mean and mean ± 2(standard deviation) over 20,000 estimates $\widehat{\boldsymbol{f}}_{\ell^2}$ based on different noise realizations with λ fixed and equal to the mean value of the GCV selections over the simulations. The results seem reasonable but they are misleading because in practice λ is estimated from the data and this selection may introduce important variability. This is illustrated in Figure 2. The left panel shows the results when λ is selected by GCV in every simulation. The panel on the right shows the median and median ± 2(MAD/0.675) (recall that MAD/0.675 is a robust estimate of the standard deviation of a Gaussian distribution). The drastic difference in variability observed in the two plots indicates the occurrence of outliers introduced by the selection of λ (this behavior can be modeled with Gaussian mixtures as explained in Section 4). For some data realizations the GCV selects too small a value of λ that leads to under-smoothed estimates. A corrected version of the GCV where the denominator is replaced by $n - \gamma \operatorname{tr} \boldsymbol{H}(\lambda)$, with γ a constant slightly greater than one, is sometimes used to help reduce this problem. However, for the case of indirect observations with an ill-posed problem the selection of γ is not easy; it depends on how ill-posed the operator is. For direct observations the value $\gamma \approx 1.2$ has been suggested [13, 28, 31]. We use this value to estimate σ^2 using (2.2). Figure 3 shows the version of Figure 2 corresponding to $\widehat{\boldsymbol{f}}_{\ell^1}$. It is clear that the ℓ^1-estimates are more stable, which may not be surprising given that it includes two regularizations: one based on sparsity and the other on the stopping of the iterations in the ℓ^1-code. The results with the recursive mean and MAD (not shown) are almost identical to those with their non-recursive counterparts.

2.2. Approximating the trace of a large matrix. The computation of the trace of a large matrix is often required, for example, for evaluating the GCV function, computing mean-squared errors, and approximating confidence intervals or the objective function of A-experimental designs for inverse problems. Stochastic trace estimators [5, 19, 23] can be used to approximate the trace of a large matrix. We start with a brief derivation and then proceed to determine bounds for their variance and concentration around their mean in terms of the behavior of the matrix eigenvalues. Other bounds and estimators are discussed in [3].

If $\boldsymbol{U}$ is an $n \times 1$ random vector with covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{H}$ is any fixed symmetric $n \times n$ matrix, then $\mathbb{E}(\boldsymbol{U}^t \boldsymbol{H} \boldsymbol{U}) = \operatorname{tr}(\boldsymbol{\Sigma} \boldsymbol{H})$. The variance of this quadratic form is [36, p.35]: $\mathbb{V}\mathrm{ar}(\boldsymbol{U}^t \boldsymbol{H} \boldsymbol{U}) = 2\operatorname{tr}(\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{D}_H \boldsymbol{D}_4 \boldsymbol{\Sigma})$, where $\mathbb{E}(U_i^4) = \beta_i \sigma_i^4$ for some $\beta_i > 0$, $\sigma_i^2 =$

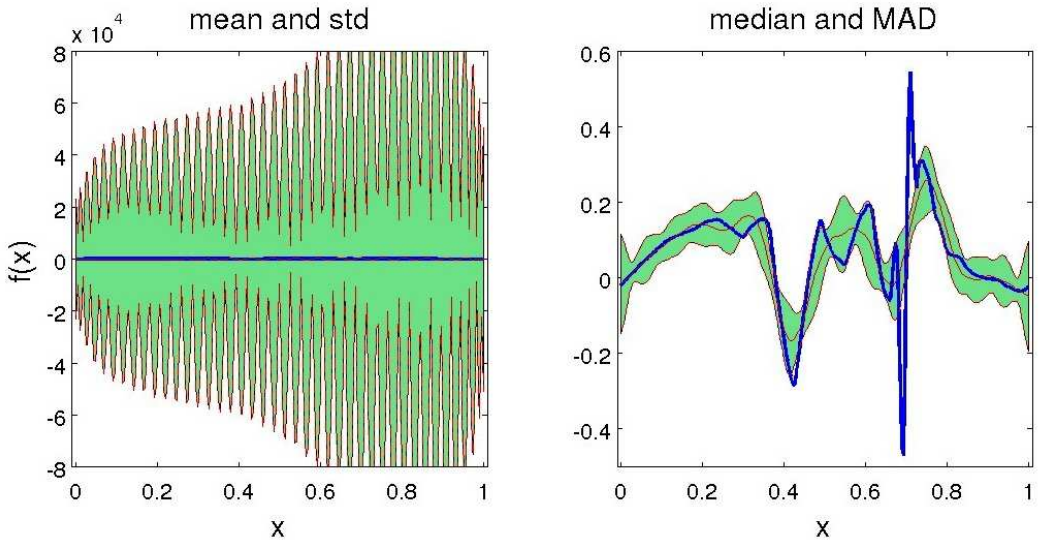

FIG. 2. *Left: Bands for $\widehat{\boldsymbol{f}}_{\ell^2}$ with λ selected by GCV in every simulation. Right: The band is defined by the the median $\pm$ 2(MAD/0.675). The blue line is $\boldsymbol{f}$.*

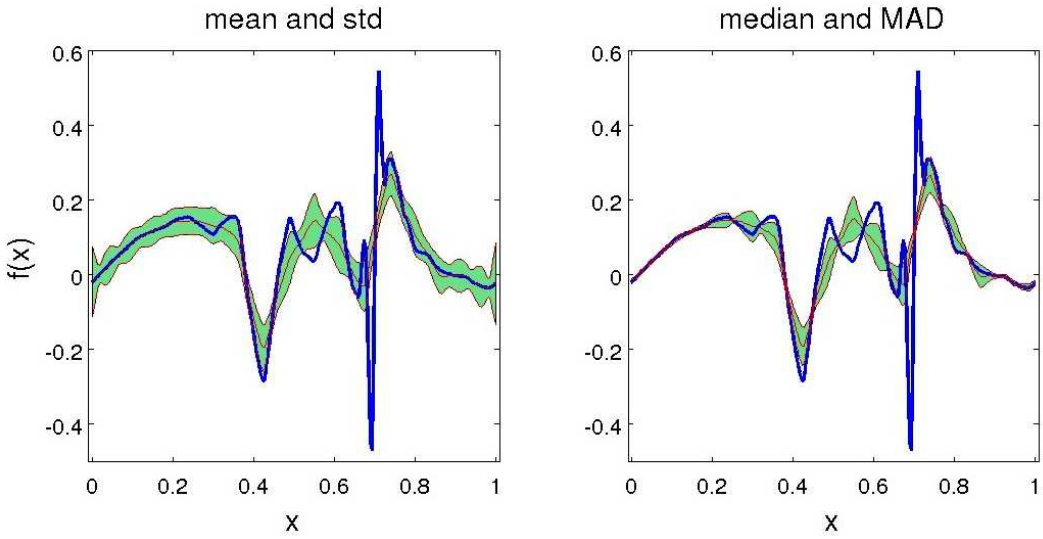

FIG. 3. *Same as Figure 2 but for $\widehat{\boldsymbol{f}}_{\ell^1}$ obtained by solving (2.1).*

$\mathbb{V}\mathrm{ar}(U_i)$, $\boldsymbol{D}_H = \mathrm{Diag}(\boldsymbol{H})$ and $\boldsymbol{D}_4 = \mathrm{Diag}\{\beta_1 - 3, ..., \beta_n - 3\}$. For example, if the U_i are independent and zero mean with $\mathbb{V}\mathrm{ar}(U_i) = \sigma^2$ and $\mathbb{E}(\,U^4\,) = \beta\,\sigma^4$, then $\mathbb{E}\,(\,\boldsymbol{U}^t\boldsymbol{H}\boldsymbol{U}\,) = \sigma^2\mathrm{tr}(\,\boldsymbol{H}\,)$ and

$$\mathbb{V}\mathrm{ar}(\,\boldsymbol{U}^t\boldsymbol{H}\boldsymbol{U}\,) = 2\,\sigma^4\mathrm{tr}(\,\boldsymbol{H}^2\,) + \sigma^4(\beta - 3)\,\|\,\boldsymbol{h}\,\|^2, \tag{2.3}$$

where $\boldsymbol{h}$ is the vector of diagonal entries of $\boldsymbol{H}$. An unbiased estimate of $\mathrm{tr}(\boldsymbol{H})$ requires $\sigma = 1$, and since $\beta \geq 0$, (2.3) implies that the smallest variance is obtained with the smallest β. But, if X is a zero-mean, unit-variance random variable with finite fourth moment, then Jensen's inequality implies $1 = (\,\mathbb{E}(\,X^2\,)\,)^2 \leq \mathbb{E}(\,X^4\,)$. Hence $\beta = 1$ is the smallest possible value of $\mathbb{E}(\,X^4\,)$, which is achieved with X uniform on $\{-1, 1\}$. In this case (2.3) reduces to $\mathbb{V}\mathrm{ar}(\,\boldsymbol{U}^t\boldsymbol{H}\boldsymbol{U}\,) = 2[\,\mathrm{tr}(\boldsymbol{H}^2) - \|\,\boldsymbol{h}\,\|^2\,]$. This proves part (i) of the following proposition.

PROPOSITION 2.1. Let $\boldsymbol{H}$ be a symmetric $n \times n$ matrix with diagonal $\boldsymbol{h}$ and $\boldsymbol{U}_1$,...,$\boldsymbol{U}_m$ be independent $n \times 1$ vectors, each with independent entries uniform on $\{-1, 1\}$. For a fixed integer m define the trace estimator $\widehat{T}_m(\boldsymbol{H}) = (1/m) \sum_{i=1}^m \boldsymbol{U}_i^t \boldsymbol{H} \boldsymbol{U}_i$. Then:

(i) $\mathbb{E}(\widehat{T}_m(\boldsymbol{H})) = \mathrm{tr}(\boldsymbol{H})$ and $\mathbb{V}\mathrm{ar}(\widehat{T}_m(\boldsymbol{H})) = \frac{2}{m}[\mathrm{tr}(\boldsymbol{H}^2) - \|\boldsymbol{h}\|^2]$.

(ii) If $\boldsymbol{H}$ is also non-negative definite with eigenvalues α_i and $\mathrm{tr}(\boldsymbol{H}) > 0$, then the relative variance $V_r(\widehat{T}_m)$ of $\widehat{T}_m$ is

$$V_r(\widehat{T}_m) = \frac{\mathbb{V}\mathrm{ar}(\widehat{T}_m(\boldsymbol{H}))}{\mathrm{tr}(\boldsymbol{H})^2} \leq \frac{2}{mn} \frac{S_\alpha^2}{\bar{\alpha}^2},$$

where $\bar{\alpha} = (1/n) \sum_i \alpha_i$ and $S_\alpha^2 = (1/n) \sum_i (\alpha_i - \bar{\alpha})^2$.

(iii) For any $t > 0$,

$$\mathbb{P}(\widehat{T}_m(\boldsymbol{H}) \geq \mathrm{tr}(\boldsymbol{H})(1+t)) \leq \mathrm{e}^{-m t^2 / 4 V_r(\widehat{T}_1)} \tag{2.4}$$

$$\mathbb{P}(\widehat{T}_m(\boldsymbol{H}) \leq \mathrm{tr}(\boldsymbol{H})(1-t)) \leq \mathrm{e}^{-m t^2 / 4 V_r(\widehat{T}_1)}. \tag{2.5}$$

Proof: The proof of (ii) is straightforward: Since the minimum of $\|\boldsymbol{h}\|^2$ subject to $\sum h_i = \mathrm{tr}(\boldsymbol{H})$ is $\mathrm{tr}(\boldsymbol{H})^2/n$, it follows that

$$\frac{\mathrm{tr}(\boldsymbol{H}^2) - \|\boldsymbol{h}\|^2}{\mathrm{tr}(\boldsymbol{H})^2} = \frac{n S_\alpha^2 + n\bar{\alpha}^2 - \|\boldsymbol{h}\|^2}{\mathrm{tr}(\boldsymbol{H})^2} \leq \frac{n S_\alpha^2 + n\bar{\alpha}^2 - \mathrm{tr}(\boldsymbol{H})^2/n}{\mathrm{tr}(\boldsymbol{H})^2} = \frac{S_\alpha^2}{n\,\bar{\alpha}^2}.$$

(iii) We first prove (2.4) for $m = 1$ using a martingale concentration inequality that can be found in [29]. Define the increasing sequence of σ-algebras $\mathcal{F}_k = \sigma(U_1, ..., U_k)$ for $k = 1, ..., n$. Then,

$$\mathbb{E}(\boldsymbol{U}^t \boldsymbol{H} \boldsymbol{U} \,|\, \mathcal{F}_k) = \sum_{i=1}^k H_{i,i} U_i^2 + \sum_{i=k+1}^n H_{i,i} + 2 \sum_{i=1}^{k-1} \sum_{j>i}^k H_{i,j} U_i U_j,$$

from which one can easily verify the equality

$$d_k = \mathbb{E}(\boldsymbol{U}^t \boldsymbol{H} \boldsymbol{U} \,|\, \mathcal{F}_k) - \mathbb{E}(\boldsymbol{U}^t \boldsymbol{H} \boldsymbol{U} \,|\, \mathcal{F}_{k-1}) = 2\, U_k \sum_{i=1}^{k-1} H_{i,k} U_i,$$

for $k = 2, ..., n$ and $d_1 = 0$. It follows that each d_k is bounded: $|d_k| \leq B_k = 2 \sum_{i=1}^{k-1} |H_{i,k}|$, and

$$B_k^2 = 4 \sum_{i,j}^{k-1} |H_{i,k}|\,|H_{k,j}| \leq 4 (\sum_{i,j}^{n} |H_{i,k}|\,|H_{k,j}| - H_{k,k}^2),$$

which implies $\sum_k B_k^2 \leq 4(\mathrm{tr}(\boldsymbol{H}^2) - \|\boldsymbol{h}\|^2) = 2\,\mathbb{V}\mathrm{ar}(\widehat{T}_1(\boldsymbol{H}))$. It now follows from Lemma 4.1 in [29, p.68] that

$$\mathbb{P}(\widehat{T}_1(\boldsymbol{H}) \geq \mathrm{tr}(\boldsymbol{H})(1+t)) \leq \mathrm{e}^{-\mathrm{tr}(\boldsymbol{H})^2 t^2 / 4 \mathbb{V}\mathrm{ar}(\widehat{T}_1(\boldsymbol{H}))} = \mathrm{e}^{-t^2 / 4 V_r(\widehat{T}_1)}.$$

To prove the result for $m > 1$ note that we can write $\widehat{T}_m(\boldsymbol{H}) = \widetilde{\boldsymbol{U}}^t \widetilde{\boldsymbol{H}} \widetilde{\boldsymbol{U}}$, where $\widetilde{\boldsymbol{U}} = (\boldsymbol{U}_1^t \cdots \boldsymbol{U}_m^t)^t$ and $\widetilde{\boldsymbol{H}} = \mathrm{Diag}(\boldsymbol{H}, ..., \boldsymbol{H})/m$. Set $\widetilde{\boldsymbol{h}} = \mathrm{Diag}(\widetilde{\boldsymbol{H}})$. Then, $\mathrm{tr}(\widetilde{\boldsymbol{H}}) = \mathrm{tr}(\boldsymbol{H})$, $\mathrm{tr}(\widetilde{\boldsymbol{H}}^2) = \mathrm{tr}(\boldsymbol{H})/m$ and $\|\widetilde{\boldsymbol{h}}\|^2 = \|\boldsymbol{h}\|^2/m$ and therefore the result follows from the case $m = 1$. The proof of (2.5) follows similarly using the left-tail concentration inequality in [29, p.68]. □

Part (ii) shows that the relative variance of the trace estimate is small if the scatter of the eigenvalues is small compared to their mean. For example if $\boldsymbol{H} = a\boldsymbol{I}$ then the variance bound is zero as it should. Also, the bound is small if $\boldsymbol{H}$ is diagonal with small variation in the diagonal entries. The bounds may serve as a guide to choose m.

EXAMPLE 2. The standard least-squares estimate of $\boldsymbol{x}$ based on the $n\times 1$ data vector $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}+\boldsymbol{\varepsilon}$ with $\boldsymbol{A}$ of full column rank leads to the projection matrix $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t$, which projects $\boldsymbol{y}$ orthogonally onto the k-dimensional subspace spanned by the columns of $\boldsymbol{A}$. This matrix is used, for example, for model validation and construction of confidence intervals. Since $\boldsymbol{H}$ is symmetric and idempotent, it has k eigenvalues equal to one and the rest equal to zero. In this case the relative variance and bound (2.1) become

$$V_r(\widehat{T}_m) = \frac{2}{mk}\left(1 - \frac{\|\boldsymbol{h}\|^2}{k}\right) \le \frac{2}{mk}\left(1 - \frac{k}{n}\right).$$

For example, for the variance to be less than ν, we need $m \ge 2(1/k - 1/n)/\nu$. The right-tail probability bound is

$$\mathbb{P}(\,\widehat{T}_m(\boldsymbol{H}) \ge \mathrm{tr}(\boldsymbol{H})\,(1+t)\,) \le \mathrm{e}^{-mk^2t^2/8(k-\|\boldsymbol{h}\|^2)} \le \mathrm{e}^{-mkt^2/8}.$$

Note that this standard least-squares approach is often used after an ℓ^1-regularization is employed for variable selection to reduce dimensionality. In this case k is the number of variables kept (i.e., target sparsity). The performance of such procedure is considered in Example 8.

EXAMPLE 3. Consider now the matrix $\boldsymbol{H}$ from Example 1 with $n = 2^{10}$ and $\mathrm{tr}(\boldsymbol{H}) = 20.5$; $\boldsymbol{H}$ is not a projection matrix but all of its eigenvalues are bounded by one. The true relative variance in this case is $V_r(\widehat{T}_1) = 0.0917$ while the bound from Proposition 2.1 is 0.0919. With $m = 20$ the variance bound is decreased to 0.005. The deviation probabilities that $\widehat{T}_1(\boldsymbol{H})$ differs from $\mathrm{tr}(\boldsymbol{H})$ by ± 2.5 are $P(\,\widehat{T}_1(\boldsymbol{H}) \ge 1.12\,\mathrm{tr}(\boldsymbol{H})\,) = 0.322$ and $P(\,\widehat{T}_1(\boldsymbol{H}) \le 0.88\,\mathrm{tr}(\boldsymbol{H})\,) = 0.372$. This time (2.4) provides the very conservative bound 0.962.

3. Checking the model fit. Suppose $\widehat{\boldsymbol{f}}$ is a 'good' estimate of $\boldsymbol{f}$, the noise $\boldsymbol{\varepsilon}$ is truly $N(0,\sigma^2\boldsymbol{I})$ and $\widehat{\sigma}^2$ is a reasonable estimate of σ^2. It then stands to reason that simulating data via $\boldsymbol{A}\widehat{\boldsymbol{f}} + \widehat{\boldsymbol{\varepsilon}}$ with $\widehat{\boldsymbol{\varepsilon}} \sim N(0,\widehat{\sigma}^2\boldsymbol{I})$ should produce synthetic data consistent with the observations $\boldsymbol{y}$. This may happen even if $\widehat{\boldsymbol{f}}$ has a significant bias as long as the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}$ is small. We check the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}$ by constructing confidence intervals.

3.1. Confidence intervals for the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}$. For a fixed λ, the estimate $\widehat{\boldsymbol{f}}_{\ell^2}$ can be written explicitly: $\widehat{\boldsymbol{f}}_{\ell^2} = \boldsymbol{G}(\lambda)^{-1}\boldsymbol{A}^t\boldsymbol{y}$, where

$$\boldsymbol{G}(\lambda) = \boldsymbol{A}^t\boldsymbol{A} + \lambda^2\boldsymbol{D}^t\boldsymbol{D}. \tag{3.1}$$

Since we will simulate data using $\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}$ as a proxy for $\boldsymbol{A}\boldsymbol{f}$, we need to check the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}$ given by $\mathrm{Bias}(\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}) = \boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2})$. This bias is easier to assess than that of $\widehat{\boldsymbol{f}}_{\ell^2}$ because the data are direct observations of $\boldsymbol{A}\boldsymbol{f}$. We employ a method similar to that used for constructing prediction intervals in regression [12], where estimates are unbiased. We construct confidence intervals for $\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}})$ and $\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}}_{(-i)})$, where $\widehat{\boldsymbol{f}}_{(-i)}$ is a leave-one-out estimate of $\boldsymbol{f}$ defined as follows: Let $\boldsymbol{y}_{(-i)}$ and $\boldsymbol{A}_{(-i)}$ be, respectively, the vector $\boldsymbol{y}$ and matrix $\boldsymbol{A}$ without the ith row. Then, it is easy to see that $\widehat{\boldsymbol{f}}_{(-i)} = (\boldsymbol{A}_{(-i)}^t\boldsymbol{A}_{(-i)} + \lambda^2\boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{A}_{(-i)}^t\boldsymbol{y}_{(-i)}$. Define $\widehat{y}_{(-i),i}$ to be the ith entry of the prediction $\boldsymbol{A}\widehat{\boldsymbol{f}}_{(-i)}$: $\widehat{y}_{(-i),i} = \boldsymbol{e}_i^t\boldsymbol{A}\widehat{\boldsymbol{f}}_{(-i)}$ ($\{\boldsymbol{e}_i\}$ is the standard basis of $\mathbb{R}^n$). Hence,

$$\mathbb{E}(\,\widehat{y}_i - y_i\,) = \boldsymbol{e}_i^t\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}}), \quad \mathbb{E}(\,\widehat{y}_{(-i),i} - y_i\,) = \boldsymbol{e}_i^t\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}}_{(-i)}).$$

Using a standard rank-one update [12], we obtain $\widehat{y}_{(-i),i} - y_i = (\widehat{y}_i - y_i)/(1 - H_{ii})$, which yields

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\,\widehat{y}_i - y_i\,) &= \sigma^2\left(\,(1-H_{ii})^2 + (H^2)_{ii} - (H_{ii})^2\,\right)\\ \mathbb{V}\mathrm{ar}(\,\widehat{y}_{(-i),i} - y_i\,) &= \sigma^2\left(1 + \frac{(H^2)_{ii} - (H_{ii})^2}{(1-H_{ii})^2}\right).\end{aligned}$$

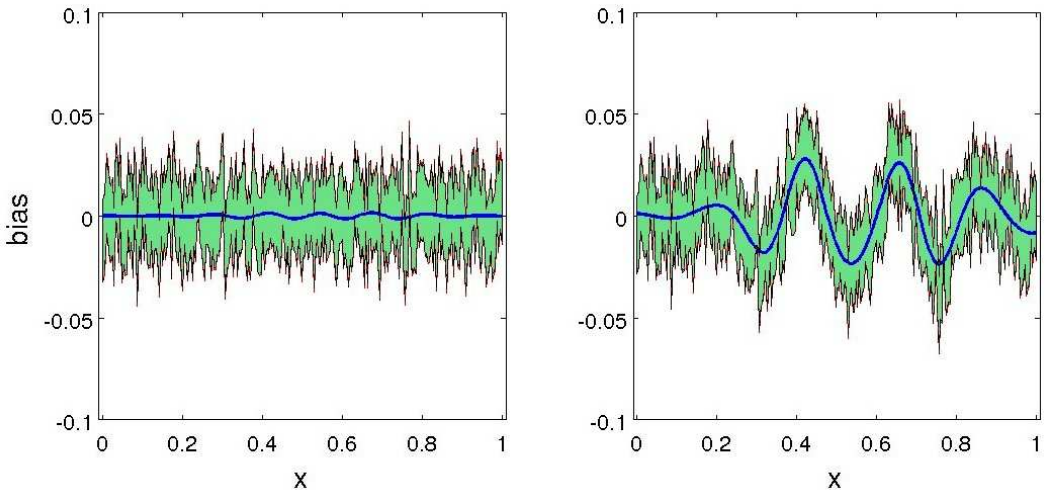

FIG. 4. *Each plot shows 95% confidence intervals (3.2) (indistinguishable from (3.3)) drawn as a continuous green band for the data in Figure 1. The blue lines are the bias of* $\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}$. *The estimates on the right panel was obtained using a large* λ *to introduce a bias.*

If σ is known, Gaussian approximations of $1-\alpha$ confidence intervals for $\boldsymbol{e}_i^t\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}})$ and $\boldsymbol{e}_i^t\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}}_{(-i)})$, are, respectively,

$$\widehat{y}_i - y_i \pm z_{\alpha/2}\,\sigma\sqrt{(1-H_{ii})^2 + (H^2)_{ii} - (H_{ii})^2} \tag{3.2}$$

$$\frac{\widehat{y}_i - y_i}{1-H_{ii}} \pm z_{\alpha/2}\,\sigma\sqrt{1 + \frac{(H^2)_{ii} - (H_{ii})^2}{(1-H_{ii})^2}} \tag{3.3}$$

($z_{\alpha/2}$ is defined by $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0,1)$). If σ is unknown we use $\widehat{\sigma}$ instead. Approximations of these intervals defined in terms of traces (to allow the use randomized trace estimators) are obtained by replacing H_{ii} and $(\boldsymbol{H}^2)_{ii}$ with $\mathrm{tr}(\boldsymbol{H})/n$; a similar thing is done to obtain the GCV from the cross-validation function [41]:

$$\widehat{y}_i - y_i \pm z_{\alpha/2}\,\widehat{\sigma}\,\sqrt{\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H})/n} \tag{3.4}$$

$$\frac{\widehat{y}_i - y_i}{\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H})/n} \pm z_{\alpha/2}\,\widehat{\sigma}\,\frac{1}{\sqrt{\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H})/n}}. \tag{3.5}$$

If the ith observation is not too influential, then the intervals for $\boldsymbol{e}_i^t\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}}_{(-i)})$ and $\boldsymbol{e}_i^t\boldsymbol{A}\mathrm{Bias}(\widehat{\boldsymbol{f}})$ should be similar. If the intervals do not include zero it may be an indication that the bias of $\boldsymbol{e}_i^t\boldsymbol{A}\widehat{\boldsymbol{f}}$ is significant. To check for the influence that the different y_i may have on the fit, we may plot $|\widehat{y}_{(-i)} - \widehat{y}_i| = H_{ii}\,|\widehat{y}_i - y_i|/(1-H_{ii})$ normalized by its standard error.

In the derivation of the intervals, we assumed that λ was fixed but in practice we have found that the variability introduced by the selection of λ (e.g., by GCV) does not have much of an effect on the coverage of these intervals. This is shown in the next example.

EXAMPLE 4. We construct 95% confidence intervals for the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}$ in Example 1. The left panel in Figure 4 shows confidence intervals (3.2) and (3.3) (for a single realization of $\boldsymbol{y}$) drawn as a band with red boundaries for (3.3) and black boundaries for (3.2) (almost identical). For reference, the correct bias is shown in blue. Here λ is chosen by GCV and $\widehat{\sigma}$ is used in place of σ. The right panel shows confidence intervals when λ is set to a large value (20 times the GCV selected value) to increase the bias. The left panel does not show evidence of bias while the right one does. To check the behavior of the intervals for different realizations of $\boldsymbol{y}$, we use simulations

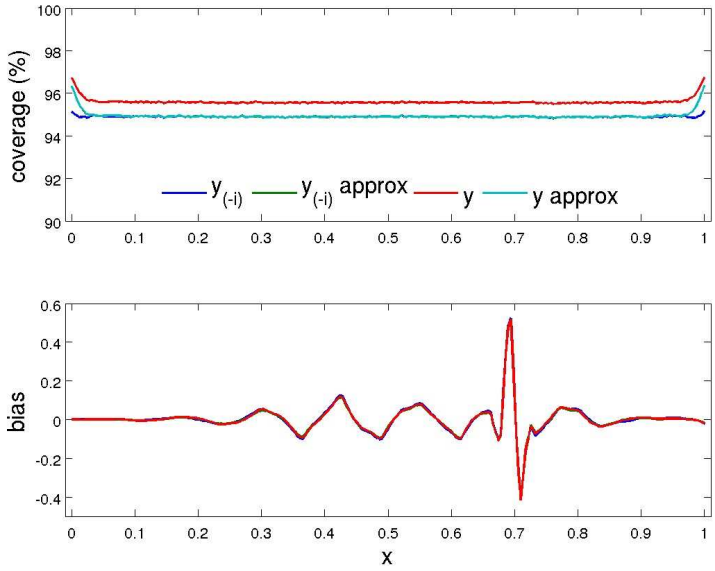

FIG. 5. *Top: pointwise coverage of the 95% confidence intervals (3.2) and (3.4) (labeled y), and (3.3) and (3.5) (labeled $y_{(-i)}$). Bottom: the blue, green and red lines (almost identical) depict, respectively, the median bias of $\widehat{\boldsymbol{f}}_{\ell^2}$, and the mean and median of $B(\widehat{\lambda})$ (4.3).*

to estimate the pointwise coverage of the intervals (3.2) through (3.5). The results are shown in Figure 5. The intervals (3.3) have the correct coverage almost everywhere while (3.2) are slightly more conservative across the function; a result of including y_i in the data to predict itself. The approximate intervals (3.4) and (3.5) have essentially the same coverage, close to the target except at the boundaries where they are slightly conservative.

3.2. Comparing predictions to observations. Another way to explore model fits is by comparing characteristics of simulated data $\boldsymbol{A}\widehat{\boldsymbol{f}} + \widehat{\boldsymbol{\varepsilon}}$ with those of the original observations $\boldsymbol{y}$. We illustrate this using parametric and nonparametric bootstrap methods [14]. For the parametric approach we assume that $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and $\widehat{\boldsymbol{\varepsilon}}$ is sampled randomly from $N(\boldsymbol{0}, \widehat{\sigma}^2 \boldsymbol{I})$. This method can be used with linear or nonlinear estimates (e.g., $\widehat{\boldsymbol{f}}_{\ell^2}$ or $\widehat{\boldsymbol{f}}_{\ell^1}$). In the nonparametric approach $\widehat{\boldsymbol{\varepsilon}}$ is randomly sampled with replacement from the (corrected) residuals of the fit. The corrections are needed because even though one would expect the residuals to behave approximately like the true unobservable errors $\boldsymbol{\varepsilon}$ if the inversion estimate is reasonable, this is only true up to bias and variance residual corrections. For example, the vector of residuals for $\widehat{\boldsymbol{f}}_{\ell^2}$ is $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2} = (\boldsymbol{I} - \boldsymbol{H}(\lambda))\,\boldsymbol{y}$ and for a fixed λ, $\mathbb{E}\,\boldsymbol{r} = -\lambda^2 \boldsymbol{A}\boldsymbol{G}(\lambda)^{-1}\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{f}$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{r}) = \sigma^2\,(\boldsymbol{I} - \boldsymbol{H}(\lambda))^2$. Hence, even with λ fixed, the residuals are, unlike the true errors, correlated, of variance different from σ^2 and not centered at zero. If the bias is small the residuals should be approximately centered at zero but may be correlated even when the fit is reasonably good. One could use the covariance matrix of the residuals to approximately correct their correlation but it would be computationally expensive. As it is customary in regression analysis, we only correct the residuals for heteroscedasticity.

To simplify the correlation structure of the residuals as well as the computation of their corrections, the nonparametric simulations will be based on the residuals of the ℓ^2-fit to $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{f}$. That is, the same fit used to determine $\widehat{\sigma}$ (2.2). The corrected residuals are then

$$\boldsymbol{r}_c = [\,\mathrm{Diag}(\boldsymbol{I} - (\boldsymbol{I} + \widehat{\lambda}^2 \boldsymbol{D}^t\boldsymbol{D})^{-1})\,]^{-1}\boldsymbol{r}. \tag{3.6}$$

We use the following resampling procedures to generate B parametric/nonparametric bootstrap samples $\{\boldsymbol{y}_i^*\}$: For each $i = 1, ..., B$, (parametric) generate noise $\boldsymbol{\varepsilon}_i^*$ from the distribution $N(\boldsymbol{0}, \widehat{\sigma}^2 \boldsymbol{I})$, or (nonparametric) by resampling from the residuals $\boldsymbol{r}_c$ (3.6). The ith synthetic data vector is $\boldsymbol{y}_i^* = \boldsymbol{A}\widehat{\boldsymbol{f}} + \boldsymbol{\varepsilon}_i^*$. It is useful to compare the results of parametric and nonparametric simulations; for example, a large discrepancy may indicate problems with the presumed distribution of the noise.

If the model fit is reasonable, the characteristics of the simulated data $\boldsymbol{y}^*$ should be similar to those of the original observations. Similarly, if the hypotheses on the noise are correct and the fit is good, then the characteristics of the corrected residuals $\boldsymbol{r}_c$ should be consistent with those of an iid Gaussian sample. To make these comparisons, we define statistics to capture different distributional characteristics. The choice depends on the particular problem, as an example we use the following test statistics:

$T_1(\boldsymbol{y}) = \min(\boldsymbol{y})$	$T_6(\boldsymbol{y}) = \text{MAD}(\boldsymbol{y})$	$T_5(\boldsymbol{y}) = \text{sample median } (\boldsymbol{y})$
$T_2(\boldsymbol{y}) = \max(\boldsymbol{y})$	$T_7(\boldsymbol{y}) = \text{1st sample quartile } (\boldsymbol{y})$	
$T_3(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y} - \text{mean}(\boldsymbol{y}))^3$	$T_8(\boldsymbol{y}) = \text{3rd sample quartile } (\boldsymbol{y})$	
$T_4(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y} - \text{mean}(\boldsymbol{y}))^4$	$T_9(\boldsymbol{y}) = \#$ runs above/below $T_5(\boldsymbol{y})$	

Note that $\boldsymbol{y}$ is a sample from a multivariate distribution, its entries y_i are not iid as they have different means. Thus the statistics T_j have to be interpreted as different functions of $\boldsymbol{y}$ not as characteristics of marginal distributions. One reason for doing this is that there is only one sample $\boldsymbol{y}$ so it is difficult to assess multivariate properties. In some cases it may be reasonable to use test statistics in different regions of the domain of the data, or at different scales via a wavelet transform as we do in Example 5.

To check if the observed values of the test statistics are consistent with those from the simulations, we use the p-values $P_j = \mathbb{P}(\,|T_j(\boldsymbol{y}^*)| \geq |T_j^o|\,)$, where $T_j^o = T_j(\boldsymbol{y})$. If the fit is reasonable and the simulations are consistent with the process that generated the data, then the p-values should be neither too close to zero nor too close to one. But, as we know from classic hypothesis testing, the decision of what is too small or too large a p-value is subjective. Data analysis is not an exact science. We consider the validation process more as a way to get insight into possible sources of problems with the results than as a process to obtain formal statements of model fit.

We use the bootstrap samples to construct confidence intervals for the P_j. To obtain more reliable estimates of p-values close to the boundaries, we use the method of Agresti-Coull (e.g., [8]): Let B be the number of bootstrap samples and $c_j = \#\{|T_j(\boldsymbol{y}^*)| \geq |T_j^o|\}/B$. Define $B' = B + z_{\alpha/2}^2$ and $\widehat{P}_j = (B/B')\, c_j + (1/2B)\, z_{\alpha/2}^2$. An approximate $1-\alpha$ confidence interval for P_j is

$$\widehat{P}_j \pm z_{\alpha/2}\sqrt{\widehat{P}_j(1-\widehat{P}_j)/B'}. \tag{3.7}$$

A similar method can be used with the ℓ^1-estimate: The parametric bootstrap is exactly the same and the nonparametric bootstrap can still be based on resampling from the corrected residuals (3.6). This time, however, there is no simple formula for the correction factors of the ℓ^1-residuals. Although some corrections can be derived by linearization (see Section 4); one can also use the residuals of the parametric bootstrap to derive approximate corrections.

EXAMPLE 5. We return to Example 1. One realization of $\boldsymbol{y}$ is generated in three different ways: $\boldsymbol{y}^{(1)}$ satisfies all the model assumptions with Gaussian noise and $\widehat{\boldsymbol{f}}_{\ell^2}$ is obtained using a GCV selection of λ; $\boldsymbol{y}^{(2)}$ uses a right-skewed noise distribution (a χ_3^2-distribution centered and normalized to unit-variance) and $\widehat{\boldsymbol{f}}_{\ell^2}$ with λ selected by GCV; $\boldsymbol{y}^{(3)}$ is like $\boldsymbol{y}^{(1)}$ but with $\widehat{\boldsymbol{f}}_{\ell^2}$ computed using a large λ, as in Example 4, to introduce a bias. Figure 6 shows the results. Each panel shows three sets of 95% confidence intervals (which happen to be as small as the symbols) for each statistic T_j. The blue circles and red squares correspond, respectively, to the parametric and nonparametric bootstrap simulations. The black triangles compare the statistics of the corrected residuals to those of a sample of iid Gaussian variables $N(0, \widehat{\sigma}^2)$. The results for $\boldsymbol{y}^{(1)}$ show that the p-values of T_2 and T_9 for the nonparametric bootstrap are higher than those for the parametric resampling. This indicates that the maximum is not as large as it should be for a Gaussian and that the residuals are

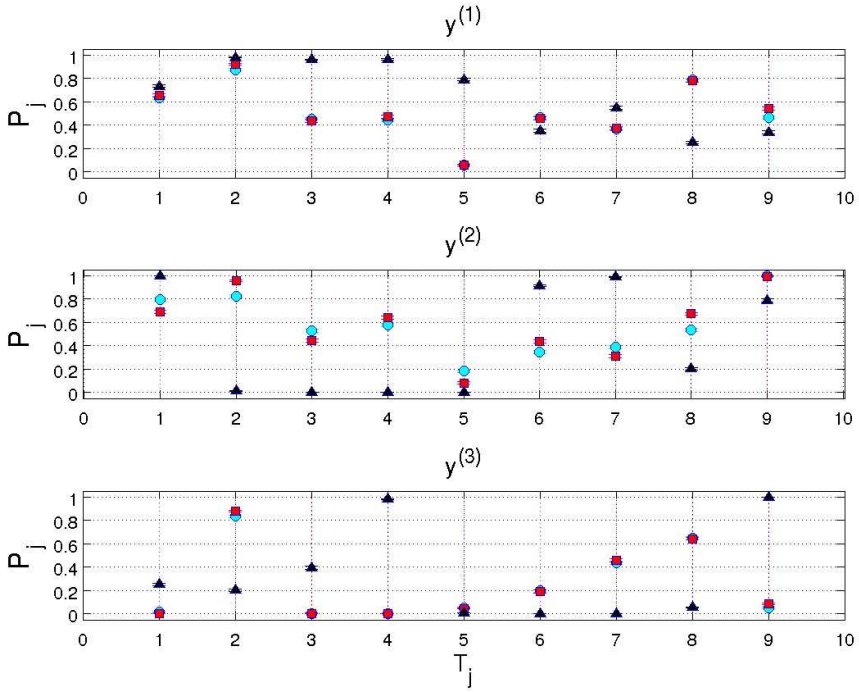

FIG. 6. *95% confidence intervals (3.7) of $T_1, ..., T_9$ for $\widehat{\boldsymbol{f}}_{\ell^2}$. The blue circles and red squares are, respectively, for the parametric and nonparametric resamplings $\boldsymbol{y}^*$. The black triangles serve to compare the corrected residuals to an iid Gaussian sample.*

correlated —this was expected given the discussion on the residuals above— but otherwise there do not seem to be serious problems. The plot for $\boldsymbol{y}^{(2)}$ shows large differences between the parametric and nonparametric p-values as well as in the statistics of the residuals; a warning that the assumed Gaussian distribution of the errors is not consistent with the residuals (because the noise is right-skewed). In addition, we see some other very small or large p-values that correctly warn us that the simulations may not be consistent with the obervations. This is also clear in the results for $\boldsymbol{y}^{(3)}$.

To check if the problem in $\boldsymbol{y}^{(3)}$ is with the noise distribution or a systematic bias, Figure 7 shows the results for $\boldsymbol{y}^{(3)}$ in two ways. The first panel reconstructs $\boldsymbol{y}^{(3)}$ using only its finest wavelet coefficients in a Symmlet-4 wavelet representation. The finest-scale coefficients are expected to be dominated by noise and since the effect of the large λ should be mostly on the bias, the panel does not show unreasonable p-values. The second panel shows the reconstruction using all but the finest wavelet coefficients, which reduces the effect of white noise. This time we see suspicious p-values, indicating a possible bias problem or the effects of coherent noise. Although not included here, the corresponding plots for $\widehat{\boldsymbol{f}}_{\ell^1}$ show similar results.

3.3. Bayesian model validation. We consider the ℓ^2-estimate (1.2), which can be derived in a Bayesian framework assuming $\boldsymbol{f}$ is random with a Gaussian smoothness prior. Define the following hierarchical sequence of distributions:

$$\boldsymbol{y}\,|\,\boldsymbol{f},\sigma,\gamma \;\sim\; N(\boldsymbol{A}\boldsymbol{f},\sigma^2\boldsymbol{I}), \quad \boldsymbol{f}|\,\gamma \;\propto\; \exp\left(-\boldsymbol{f}^t\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{f}/2\gamma^2\right), \quad (\sigma,\gamma) \;\sim\; \pi.$$

In its simplest formulation, σ and γ are assumed to be known leading to a Gaussian posterior with the following mean and covariance matrix:

$$\mathbb{E}(\boldsymbol{f}|\,\boldsymbol{y}) = \left(\boldsymbol{A}^t\boldsymbol{A} + \frac{\sigma^2}{\gamma^2}\,\boldsymbol{D}^t\boldsymbol{D}\right)^{-1}\boldsymbol{A}^t\boldsymbol{y}, \quad \mathbb{V}\mathrm{ar}(\boldsymbol{f}|\,\boldsymbol{y}) = \left(\frac{1}{\sigma^2}\,\boldsymbol{A}^t\boldsymbol{A} + \frac{1}{\gamma^2}\,\boldsymbol{D}^t\boldsymbol{D}\right)^{-1}.$$

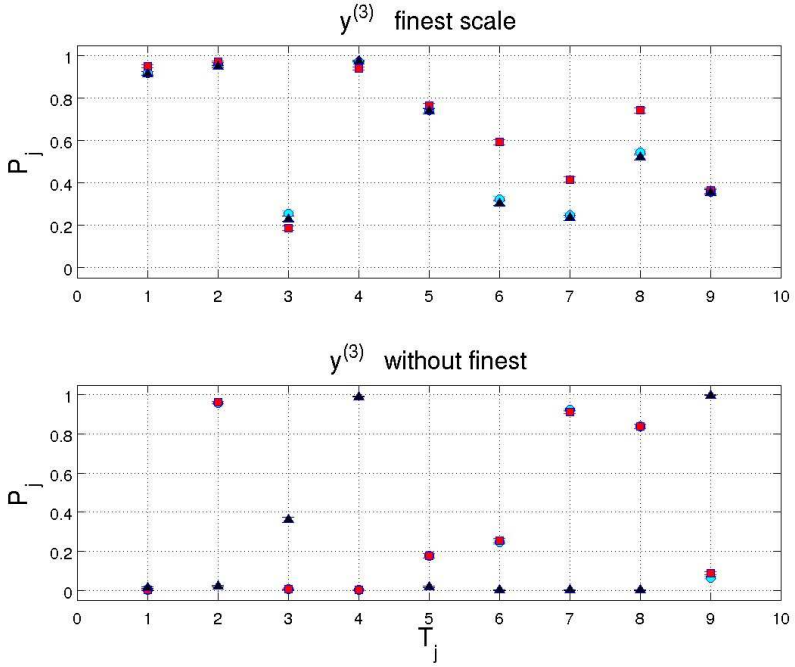

FIG. 7. *p-values for $\boldsymbol{y}^{(3)}$ in Figure 6 using data reconstructed with only the finest wavelet scales or without them.*

Hence, the posterior mean and mode coincide with $\widehat{\boldsymbol{f}}_{\ell^2}$ if $\gamma = \sigma/\lambda$.

To check if the modeling assumptions are reasonable, we can sample $\boldsymbol{f}^*$ and σ^* from the posterior distribution of $(\boldsymbol{f}, \sigma)$ given $\boldsymbol{y}$ to obtain simulated data $\boldsymbol{y}^* = \boldsymbol{A}\boldsymbol{f}^* + \sigma^*\boldsymbol{\varepsilon}^*$ with $\boldsymbol{\varepsilon}^* \sim N(\boldsymbol{0}, \boldsymbol{I})$. The consistency of $\boldsymbol{y}$ with the simulated $\boldsymbol{y}^*$ can then be explored through test statistics as in Section 3.2 (e.g., [18]). In some cases one may also be able to check the model through the marginal distribution of $\boldsymbol{y}$. For example, assuming again that λ and σ are known and $\gamma = \sigma/\lambda$, then one can study the distribution of $\boldsymbol{y} - \mathbb{E}(\boldsymbol{A}\boldsymbol{f}|\,\boldsymbol{y})$. If the model is reasonably correct this vector should be approximately zero-mean Gaussian with covariance matrix

$$\mathbb{V}\mathrm{ar}(\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{H})\left(\gamma^2\boldsymbol{A}(\boldsymbol{D}^t\boldsymbol{D})^{-1}\boldsymbol{A}^t + \sigma^2\boldsymbol{I}\right)(\boldsymbol{I} - \boldsymbol{H}). \tag{3.8}$$

The marginal distribution cannot be used when $\boldsymbol{D}^t\boldsymbol{D}$ is not full rank (e.g., the prior is improper) even when the posterior is proper.

EXAMPLE 6. We use the estimates $\widehat{\lambda}$ and $\widehat{\sigma}^2$ obtained as in Example 1 and assume them to be the fixed true values in the priors (a common practice). This time a slightly different definition of $\boldsymbol{D}$ is used to make $\boldsymbol{D}^t\boldsymbol{D}$ nonsingular; two rows are added to force zero boundary conditions. The data $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}$ and $\boldsymbol{y}^{(3)}$ are defined as in Example 5. Figure 8 shows the results. The blue circles correspond to the p-values of samples from the posterior and the red squares to samples from the zero-mean Gaussian with covariance matrix (3.8). The results for $\boldsymbol{y}^{(1)}$ are consistent with those in Example 5. In the panel for $\boldsymbol{y}^{(2)}$ we see that the data seem to be consistent with the samples from the prior but not with those from the marginal distribution. For example the value of T_1 seems to be smaller while that of T_2 seems to be larger than expected under the marginal. This is a correct warning as the noise distribution was right-skewed. The bias introduced in $\boldsymbol{y}^{(3)}$ leads to unusual p-values for samples from the posterior and the marginal.

4. Validating the estimate $\widehat{\boldsymbol{f}}$.

4.1. ℓ^2-regularization. For the linear (linearized) case one can derive explicit formulas for the bias and variance that may help determine the nature of the uncertainties of the full nonlinear

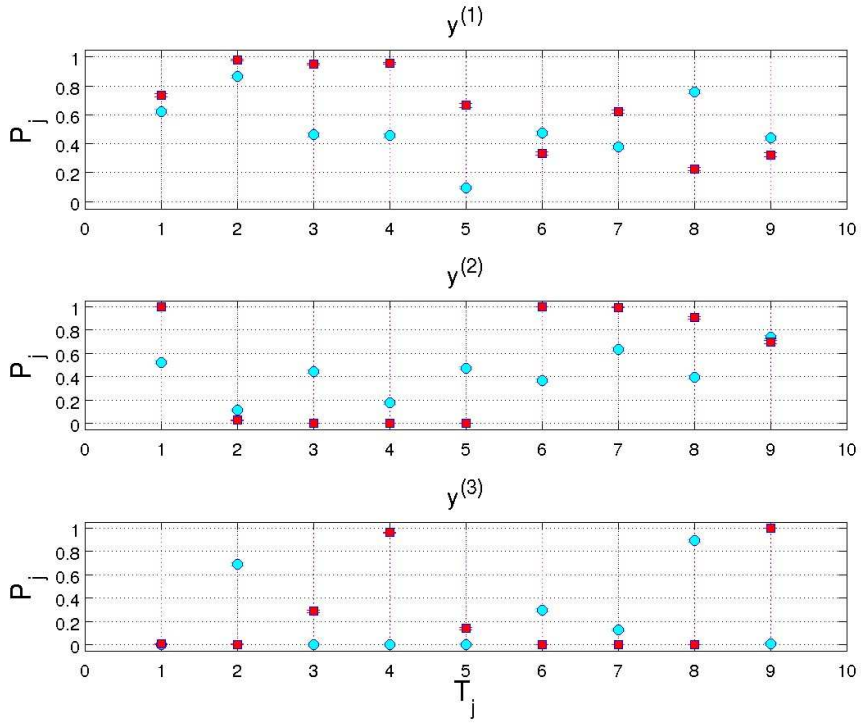

FIG. 8. *p-values for $\widehat{\boldsymbol{f}}_{\ell^2}$ as a posterior mean. The blue circles and red squares are, respectively, for samples drawn from the posterior and the marginal distributions.*

problem. For example, consider the variance and bias of $\widehat{\boldsymbol{f}}_{\ell^2}$: For a fixed $\lambda > 0$ we have:

$$\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{f}}_{\ell^2} \mid \lambda) = \sigma^2 \boldsymbol{G}(\lambda)^{-1} \boldsymbol{A}^t \boldsymbol{A}\, \boldsymbol{G}(\lambda)^{-1} \tag{4.1}$$

$$\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2} \mid \lambda) = \mathbb{E}(\widehat{\boldsymbol{f}}_{\ell^2} \mid \lambda) - \boldsymbol{f} = B(\lambda), \tag{4.2}$$

where $\boldsymbol{G}(\lambda)$ is defined in (3.1) and

$$B(\lambda) = -\lambda^2 \boldsymbol{G}(\lambda)^{-1}\, \boldsymbol{D}^t \boldsymbol{D} \boldsymbol{f}. \tag{4.3}$$

We also use the median bias defined as $\mathrm{Bias}_{\mathrm{M}}(\widehat{\boldsymbol{f}}_{\ell^2}) = \mathbb{M}(\widehat{\boldsymbol{f}}_{\ell^2}) - \boldsymbol{f}$, where $\mathbb{M}(\boldsymbol{X}) = (\mathrm{median}(X_j))$.

Since $\widehat{\boldsymbol{f}}$ and $\widehat{\lambda}$ are correlated, (4.1) may not be valid when $\widehat{\lambda}$ is used in place of λ; even if it is approximately valid, with large-scale problems it is unlikely that one can compute the full covariance matrix; instead we may attempt to estimate only the variances. One possible solution is to use a resampling approach as in Section 3.2 to produce a sequence of estimates $\{\widehat{\boldsymbol{f}}_i^*\}$ whose sample variances are used to estimate the true variances, which include the variability introduced by the selection of λ.

Assessing the bias is clearly more difficult because $\boldsymbol{f}$ is unknown. If Eqn.(4.2) were valid when the data are used to select λ, then it could yield information about the relative bias (this is illustrated in Example 7). Figure 5 shows the mean and median of $B(\widehat{\lambda})$ where $\widehat{\lambda}$ is the value obtained with GCV in each simulation. The plot shows that $\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2})$ and $\mathbb{E}[B(\widehat{\lambda})]$ are quite different (Figure 2 shows that the former is of order 10^4) but it also shows that $\mathrm{Bias}_{\mathrm{M}}(\widehat{\boldsymbol{f}}_{\ell^2}) \approx \mathbb{M}[B(\widehat{\lambda})] \approx \mathbb{E}[B(\widehat{\lambda})]$. A study of the simulation results led to the following explanation of this: The conditional distribution of $\widehat{\boldsymbol{f}}_{\ell^2}$ given $\widehat{\lambda}$ can be modeled as a two-component Gaussian mixture where about 99% of the realizations are from the Gaussian distribution with bias and variance given by (4.2) and (4.1), respectively, with

$\widehat{\lambda}$ in place of λ. The advantage of having the result $\mathrm{Bias}_{\mathrm{M}}(\,\widehat{\boldsymbol{f}}_{\ell^2}\,) \approx \mathbb{M}[\,\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2}\,|\,\widehat{\lambda})\,] \approx \mathbb{M}[\,B(\widehat{\lambda})\,]$ is that —unlike $\mathrm{Bias}_{\mathrm{M}}(\,\widehat{\boldsymbol{f}}_{\ell^2}\,)$— $B(\widehat{\lambda})$ provides an explicit formula that may yield geometric information about the bias. If the approximation did not hold, it would still be natural to ask what information regarding the distribution of $\widehat{\boldsymbol{f}}_{\ell^2}$ is provided by the median of the conditional bias $\mathbb{M}[\,\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2}\,|\,\widehat{\lambda})\,]$. The following general result relates the median of the conditional mean to the quartiles of the marginal distribution.

PROPOSITION 4.1. Let X and Y be random variables. Assume the conditional distribution $F_{X|Y}$ of X given Y has finite expectation. Let $Q_{\scriptscriptstyle X}(p)$ for $0 < p < 1$ denote the pth quantile of X. Fix $-1/2 < \beta < 1/2$. Then:

(i) If $\mathbb{P}[\,X \geq \mathbb{E}(X|Y)\,|Y\,] \geq 1/2 + \beta$, then $\mathbb{M}(\,\mathbb{E}(X|Y)\,) \leq Q_{\scriptscriptstyle X}(3/4 - \beta/2)$.

(ii) If $\mathbb{P}[\,X \leq E(X|Y)\,|Y\,] \geq 1/2 + \beta$, then $Q_{\scriptscriptstyle X}(1/4 + \beta/2) \leq \mathbb{M}(\,\mathbb{E}(X|Y)\,)$.

(iii) If $F_{X|Y}$ is symmetric, then $Q_{\scriptscriptstyle X}(1/4) \leq \mathbb{M}(\,\mathbb{E}(X|Y)\,) = \mathbb{M}(\,\mathbb{M}(X|Y)\,) \leq Q_{\scriptscriptstyle X}(3/4)$.

Proof:(i) The proof follows easily by conditioning on Y:

$$\begin{aligned}\mathbb{P}[\,X \geq \mathbb{M}(\,\mathbb{E}(X|Y)\,)\,] &\geq \mathbb{E}\left[\,I_{X \geq \mathbb{E}(X|Y)}\, I_{\mathbb{E}(X|Y) \geq \mathbb{M}(\,\mathbb{E}(X|Y)\,)}\,\right] \\ &= \mathbb{E}\left[\,\mathbb{P}(X \geq \mathbb{E}(X|Y)\,|\,Y)\, I_{\mathbb{E}(X|Y) \geq \mathbb{M}(\,\mathbb{E}(X|Y)\,)}\,\right] \\ &\geq (1/2 + \beta)\,\mathbb{P}[\,\mathbb{E}(X|Y) \geq \mathbb{M}(\,\mathbb{E}(X|Y)\,)\,] \geq 1/4 + \beta/2.\end{aligned}$$

Therefore $\mathbb{M}(\,\mathbb{E}(X|Y)\,) \leq Q_{\scriptscriptstyle X}(3/4 - \beta/2)$ by the definition of $Q_{\scriptscriptstyle X}(3/4 - \beta/2)$. The proof of (ii) is analogous and (iii) follows from (i) and (ii). □

For example, if the conditional distribution of each component $\widehat{\boldsymbol{f}}_{\ell^2,i}$ is approximately symmetric about its mean (as in the simulations for Figure 5), then it follows from Proposition 4.1 that $\mathbb{M}[\,\mathbb{E}(\widehat{\boldsymbol{f}}_{\ell^2,i}\,|\,\widehat{\lambda}\,)\,]$ is between the first and third quartiles of the distribution of $\widehat{\boldsymbol{f}}_{\ell^2,i}$ and therefore

$$Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(1/4) - \boldsymbol{f} \leq \mathbb{M}(\,\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2}\,|\,\widehat{\lambda})\,) \leq Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(3/4) - \boldsymbol{f}.$$

In particular, if $Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(1/4) \leq \boldsymbol{f} \leq Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(3/4)$, then the median bias provides a lower bound for the inter-quartile range of $\widehat{\boldsymbol{f}}_{\ell^2}$:

$$|\,\mathbb{M}(\,\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2}\,|\,\widehat{\lambda}))\,| \leq Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(3/4) - Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(1/4) = \mathrm{IQR}(\widehat{\boldsymbol{f}}_{\ell^2}).$$

On the other hand, if $\boldsymbol{f} = Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(1/4) - \delta$ or $\boldsymbol{f} = Q_{\widehat{\boldsymbol{f}}_{\ell^2}}(3/4) + \delta$ for some $\delta > 0$, then

$$\delta \leq |\,\mathbb{M}(\,\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2}\,|\,\widehat{\lambda}))\,| \leq \mathrm{IQR}(\widehat{\boldsymbol{f}}_{\ell^2}) + \delta,$$

so that $|\,\mathbb{M}(\,\mathrm{Bias}(\widehat{\boldsymbol{f}}_{\ell^2}\,|\,\widehat{\lambda}))\,|$ provides an upper bound for how far below/above the first/third quartiles the true $\boldsymbol{f}$ may be.

We now use $B(\widehat{\lambda})$ to obtain geometric information about the relative bias. Note that $B(\lambda)$ is a discrete version of the Backus-Gilbert averaging kernel [4, 34] and the bounds are based on different characteristics of this kernel. By Hölder's inequality

$$|\,B(\widehat{\lambda})\,| \leq \widehat{\lambda}^2\, \|\boldsymbol{D}\boldsymbol{G}(\widehat{\lambda})^{-1}\boldsymbol{e}_i\|_p\, \|\boldsymbol{D}\boldsymbol{f}\|_q$$

for $p > 0$, $1/p + 1/q = 1$ or, in terms of $\boldsymbol{\beta}$,

$$|\,B(\widehat{\lambda})\,| \leq \widehat{\lambda}^2\, \|\boldsymbol{W}\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{G}(\widehat{\lambda})^{-1}\boldsymbol{e}_i\|_p\, \|\boldsymbol{\beta}\|_q.$$

These bounds can be used in several ways. For example, plots of $U_{p,i}(\widehat{\lambda}) = \|\boldsymbol{D}\boldsymbol{G}(\widehat{\lambda})^{-1}\boldsymbol{e}_i\|_p$ and $U^w_{p,i}(\widehat{\lambda}) = \|\boldsymbol{W}\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{G}(\widehat{\lambda})^{-1}\boldsymbol{e}_i\|_p$ as functions of x_i provide complementary bounds on the relative bias of the Tikhonov estimate. They do not provide bias bounds but they do show regions in x space where the bias may be large or where it is expected to be small. On the other hand, plots of

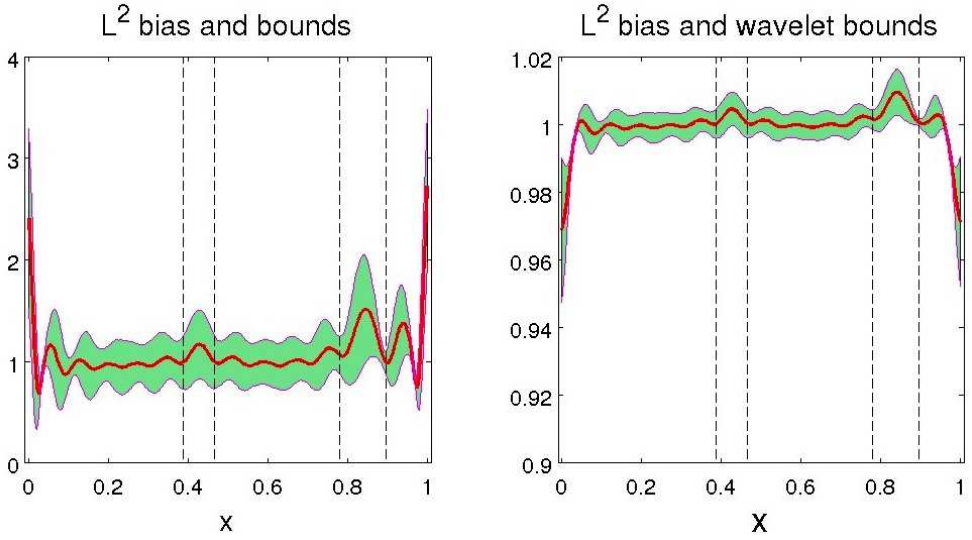

FIG. 9. *Median and median ± 2(MAD/0.675) of $U_{2,i}$ (left) and $U^w_{2,i}$ (right) normalized to median one.*

$\widehat{\boldsymbol{f}}_{\ell^2}/U_{p,i}(\widehat{\lambda})$ and $\widehat{\boldsymbol{f}}_{\ell^2}/U^w_{p,i}(\widehat{\lambda})$ as functions of x_i provide information as to the value of $\|\boldsymbol{D}\boldsymbol{f}\|_q$ or $\|\boldsymbol{\beta}\|_q$ required for the value of $(\widehat{\boldsymbol{f}}_{\ell^2})_i$ not to be dominated by the bias. For example, one may ask what values of $\|\boldsymbol{D}\boldsymbol{f}\|_q$ or $\|\boldsymbol{\beta}\|_q$ would be required for the bias bound in a particular region to be below the noise level. One may then decide if such size reasonable for these norms and thus whether or not the bias seems acceptable.

EXAMPLE 7. We continue the 1D simulations described in Example 1. This time we introduce an additional artifact; data corresponding to values of x between 0.4-0.47 and 0.78-0.9 are discarded. Figure 9 shows median and median ± 2(MAD/0.675) of $U_{2,i}(\widehat{\lambda})$ (left) and $U^w_{2,i}(\widehat{\lambda})$ (right) computed over 20,000 simulations. To compare the two plots the results have been normalized to median one. We see that the scatter of $U^w_{2,i}(\widehat{\lambda})$ around its median is a factor of ten smaller than that of $U_{2,i}(\widehat{\lambda})$; that is, the ℓ^1-bias bound is less variable. In both plots we see wider bias bounds in the regions where the data were discarded.

4.2. ℓ^1-regularization. There are no closed formulas for the bias and variance of $\widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{f}}_{\ell^1}$ (even for a fixed σ) but some approximations can be made: Using the dual formulation of the ℓ^1-optimization problem, [35] have shown that any solution $\widehat{\boldsymbol{\beta}}$ has to satisfy the equation $\boldsymbol{X}\boldsymbol{y} = (\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{R})\,\widehat{\boldsymbol{\beta}}$, where $\boldsymbol{R} = \boldsymbol{X}^t\boldsymbol{r}\boldsymbol{r}^t\boldsymbol{X}/\|\widehat{\boldsymbol{\beta}}\|_1\|\boldsymbol{X}\boldsymbol{r}\|_\infty$, with $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. The approximation suggested by [35] consists of using $\boldsymbol{A} = \boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{R}$ as fixed at the value achieved with $\widehat{\boldsymbol{\beta}}$. This leads to approximations similar to those in (4.1) and (4.2), $\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) \approx \sigma^2\boldsymbol{A}^{-1}\boldsymbol{X}^t\boldsymbol{X}\,\boldsymbol{A}^{-1}$ and $\mathrm{Bias}(\widehat{\boldsymbol{\beta}}) \approx -\boldsymbol{A}^{-1}\boldsymbol{R}\boldsymbol{\beta}.$, which can be used as we did with $\widehat{\boldsymbol{f}}_{\ell^2}$.

An alternative to linearization is to try to avoid the bias of ℓ^1-estimates by only using ℓ^1 to select the variables: If the unknown $\boldsymbol{f}$ has a sparse representation in the columns of $\boldsymbol{W}$, it then seems reasonable to use the ℓ^1-regularization to find the sparse representation followed by standard least-squares on the smaller set of variables. This would seem to solve the problem of bias and uncertainty quantification for ℓ^1- and ℓ^2-regularization but this is not quite clear. The problem is accounting for the variability in the determination of the sparse representation. As the following example shows, once this variability is taken into account, it is not really true that it is better to use $\widehat{\boldsymbol{f}}_{\ell^1}$ than to obtain an estimate $\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}$ using the two-stage procedure: ℓ^1 to reduce dimensionality followed by ordinary least-squares on the reduced problem.

EXAMPLE 8. The top row in Figure 10 shows a 2D image of the *Sigsbee* synthetic seismic

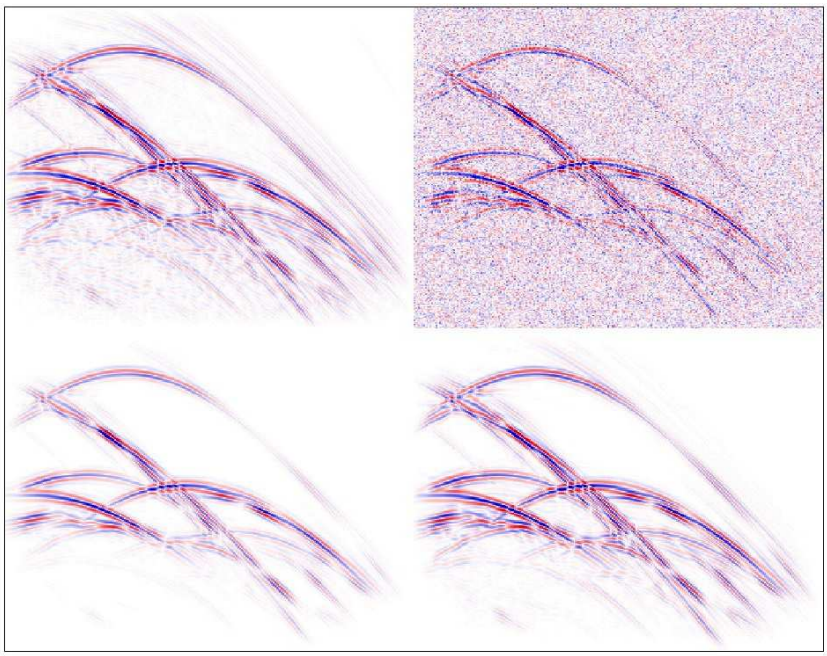

FIG. 10. *Top: the* Sigsbee *image (left) and a noisy sample (right). Bottom: simulation means of* $\widehat{\boldsymbol{f}}_{\ell^1}$ *(left) and* $\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}$ *(right).*

dataset (`http://www.delphi.tudelft.nl/SMAART/sigsbee2a.htm`) and an example of its noisy observations. The data are modeled after the geologic structure in the Sigsbee escarpment in the Gulf of Mexico. This type of wavefield has a sparse representation in a frame of Gaussian wave packets [1]. The bottom row in Figure 10 shows the simulation means of $\widehat{\boldsymbol{f}}_{\ell^1}$ (left) and $\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}$ (right). It seems that on the average $\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}$ provides a better estimate but this is not obvious without considering the uncertainties. We use simulations to compare their MSE. The left panel on Figure 11 shows the ratio $\log_{10}(\mathrm{MSE}(\widehat{\boldsymbol{f}}_{\ell^1})/\mathrm{MSE}(\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}))$ at each location. The MSE of $\widehat{\boldsymbol{f}}_{\ell^1}$ is almost two orders of magnitude worse along the stronger fronts. However, the MSE of $\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}$ can be up to three orders of magnitude worse where there is no structure. If the goal is to find the main fronts, then $\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}$ may be a better choice but it is more likely than $\widehat{\boldsymbol{f}}_{\ell^1}$ to show spurious small features. This can be also be seen in the right panel of Figure 11; it shows boxplots of the relative MSE for large and small non-zero values in the original image.

4.3. Generating plausible *f*. The frequentist resampling procedures we have used were based on creating synthetic noise samples using the fixed estimate $\widehat{\boldsymbol{f}}$ as a proxy for the unknown $\boldsymbol{f}$, which could be misleading if $\widehat{\boldsymbol{f}}$ happens to be very different from $\boldsymbol{f}$. We now use known functions so that a comparison between estimate and truth is possible. The idea is to use functions that are in some way consistent with the unknown $\boldsymbol{f}$. To obtain such functions Bayesians have the option of sampling from the prior or posterior distribution of $\boldsymbol{f}$. In the frequentist framawork, one possibility is to set up an optimization problem to search for vectors $\boldsymbol{f}^*$ that lead to fitted values consistent with the data. For example, start by defining a confidence region $C_\alpha \subset \mathbb{R}^n$ such that $\mathbb{P}[\,\boldsymbol{\varepsilon} = \boldsymbol{y} - A\boldsymbol{f} \in C_\alpha\,] \geq 1-\alpha$ and let $R_\alpha = \{\,\boldsymbol{f}^* : \boldsymbol{y} - \boldsymbol{A}\boldsymbol{f}^* \in C_\alpha\,\}$. One can then pose an optimization problem to solve for the largest MSE of estimates of $\boldsymbol{f}^* \in R_\alpha$ (see, for example, [38]). We consider a more computationally tractable problem based on wavelet characterizations of regularity.

Under some regularity conditions, membership of a function in, for example, a Sobolev, Besov or L^p space can be determined from the behavior of its wavelet coefficients as a function of scale

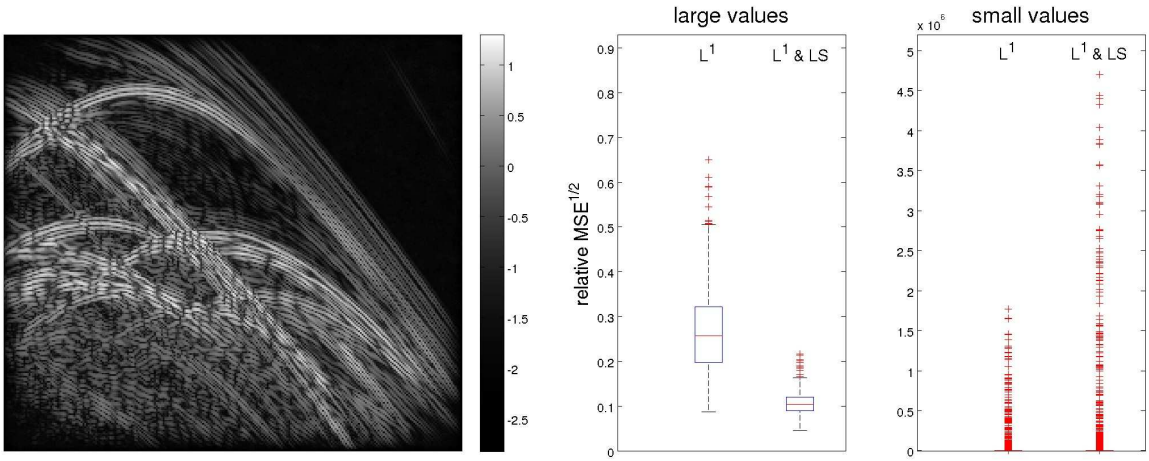

FIG. 11. *Left: simulation estimate of* $\log_{10}(\mathrm{MSE}(\widehat{\boldsymbol{f}}_{\ell^1})/\mathrm{MSE}(\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}))$. *Right: boxplots for the relative values of* $(\mathrm{MSE}(\widehat{\boldsymbol{f}}_{\ell^1}))^{1/2}$ *and* $(\mathrm{MSE}(\widehat{\boldsymbol{f}}^{\,\mathrm{ls}}_{\ell^1}))^{1/2}$ *for large and small values of the* Sigsbee *image.*

[22, 32]. Hence by modifying the wavelet coefficients without changing such behavior one obtains another function with the same global regularity. Wavelet characterizations of local regularity can be used in a similar way to define new functions with controlled local regularity. We provide an illustration of this approach.

We return to the original undiscretized function f. Assume the goal is to assess the error in estimates of $f(x_0)$ for a fixed x_0. Suppose we could generate functions $f_1, ..., f_n$ with the same 'local regularity' of f at x_0. One could then generate noise samples as before to create synthetic data $A[\boldsymbol{f}_j] + \boldsymbol{\varepsilon}^*$. The errors of the estimates of $f_j(x_0)$ may provide useful information about the type of errors one can expect for functions with the same local regularity.

There are different ways to define local regularity; as an example we consider pointwise Hölder regularity, which is defined as follows: A locally bounded function f on an interval I is said to be pointwise Hölder $\alpha > 0$ at $x_0 \in I$ (we write $f \in C^\alpha(x_0)$) if there is a constant $C > 0$ and a polynomial P_{x_o} of degree less than α such that $|\, f(x) - P_{x_o}(x)\,| \le C|\, x - x_0\,|^\alpha$ for all x in a neighborhood of x_0. Let $c_{j,k}$ be the wavelet coefficients of f with respect to an orthonormal wavelet basis. For each integer $j \ge 0$ there is a dyadic interval $Q(j,k_j) = [k_j/2^j, (k_j+1)/2^j)$ that contains x_0. Let $Q_j(x_0)$ be the interval obtained by attaching one dyadic interval at each end of $Q(j,k_j)$. The wavelet leaders of f at x_0 are defined as $d_j(x_0) = \sup 2^{j'/2}\,|c_{j',k'}|$, where the supremum is over all dyadic intervals $Q(j',k') \subset Q_j(x_0)$ for $j \ge 0$. Roughly speaking, membership in $C^\alpha(x_0)$ is equivalent to a decay $d_j(x_0) \sim C\,2^{-\alpha j}$ as $j \to \infty$. For the purpose of our applications we do not need a precise statement of this result which can be found in [24, 25]. The dependence of the wavelet leaders on the wavelet coefficients suggests a simple way to perturb the wavelet coefficients while controlling the Hölder regularity: Define a function $f^* = R(f; \boldsymbol{u}, \alpha, x_0)$ with the same wavelet coefficients as f except in a neighborhood of x_0, where the coefficients are instead $2^{\beta j} u_{i,j} c_{j,k}$ with $|u_{j,k}| \le 1$ and a chosen β. This is done in the following example using discrete wavelet transforms on the vectors of discretized values of the functions.

EXAMPLE 9. The left panel in Figure 12 shows two functions $\boldsymbol{f}^*$ obtained from $\boldsymbol{f}$ using $u_{i,j} = 1$, and $\beta = 0.2$ or $\beta = -0.7$, where x_0 is chosen to be the location of the highest peak in $\boldsymbol{f}$. We see that β may be used to change the sharpness of the function at x_0. It then makes sense to generate synthetic data: $\boldsymbol{y}^*_i = \boldsymbol{A}\, R(\widehat{\boldsymbol{f}}; \boldsymbol{u}_i, \alpha, x_0) + \boldsymbol{\varepsilon}^*_i$, where the entries of the vector $\boldsymbol{u}_i$ are iid uniform on $[-1, 1]$ and α is chosen depending on how one wants to change the regularity at x_0. For example, suppose we want to determine the relative MSE we may expect for the nonlinear estimate $\widehat{\boldsymbol{f}}_{\ell^1}$ at x_0.

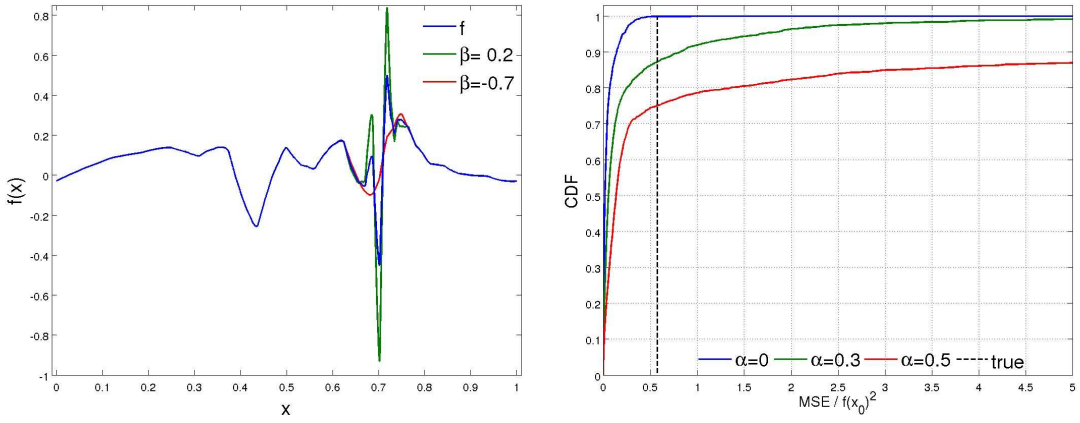

FIG. 12. *Left:* $\boldsymbol{f}$ *from Figure 1 (blue) and two examples obtained by transforming the wavelet coefficients* $c_{j,k}$ *of* $\boldsymbol{f}$ *in a neighborhood of* $x_0 = 0.72$. *Right: empirical cumulative distributions of the relative errors* $|\widehat{\boldsymbol{f}^*}_{\ell^1}(x_0) - \boldsymbol{f}^*(x_0)|^2/\boldsymbol{f}^*(x_0)^2$ *for the random perturbations of* $\widehat{\boldsymbol{f}}_{\ell^1}$. *The black line marks the relative MSE of* $\widehat{\boldsymbol{f}}_{\ell^1}$.

Since each new simulation $\boldsymbol{f}_i^* = R(\widehat{\boldsymbol{f}}; \boldsymbol{u}_i, \alpha, x_0)$ creates a function with different values at x_0, we use the relative errors $|\widehat{\boldsymbol{f}^*_{\ell^1}}(x_0) - \boldsymbol{f}^*(x_0)|^2/\boldsymbol{f}^*(x_0)^2$. The right panel in Figure 12 shows the empirical distribution function of these errors for three different values of α and, for reference, the true relative MSE of $\widehat{\boldsymbol{f}}_{\ell^1}$ at x_0 —that is, $\text{MSE}(\widehat{\boldsymbol{f}}_{\ell^1})/\boldsymbol{f}(x_0)^2 = 0.57$. If we think of the functions $\boldsymbol{f}^*$ as controls, the figure shows that without changing the sharpness at x_0 ($\beta = 0$) almost all the controls lead to relative errors less that 0.5 and 10% of them have associated relative errors greater than 0.14. With $\beta = 0.3$ and $\beta = 0.5$, respectively, we see that about 15% and 28% of the controls lead to relative errors greater that 0.4. It seems reasonable to question the significance of our estimate of $f(x_0)$ if we had found that, say, 50% of the $\beta = 0$ controls have relative errors above 1. One can also determine values of β required to have 50% of the controls with relative errors above 1 and then decide, if possible, if such values of β are reasonable based on prior information about the model and the effect of ℓ^1 estimators on local regularity.

5. Summary. We have presented examples of exploratory tools to study inversion estimates based on ℓ^2- and ℓ^1-regularizations. For ℓ^2-estimates we took advantage of explicit formulas for bias and variance for fixed values of the regularization parameter to derive approximate confidence intervals for the bias of $\boldsymbol{A}\widehat{\boldsymbol{f}}_{\ell^2}$. This is possible because the data are noisy observations of $\boldsymbol{Af}$. A similar approach can be used with linear approximations of ℓ^1 (or other nonlinear estimates)' which of course depend of how good the approximations are. Bias effects observed in the plots of these confidence intervals may reveal problems with the assumptions on the noise or the choice of forward operator, or may point to the presence of discretization effects.

We have also provided examples of resampling methods to check different aspects of $\boldsymbol{A}\widehat{\boldsymbol{f}}$ as an estimate of $\boldsymbol{Af}$. The basic idea is that if the modeling assumptions are reasonable and the estimate $\boldsymbol{A}\widehat{\boldsymbol{f}}$ is good, then we should be able to create synthetic data $\widehat{\boldsymbol{y}}$ whose statistics are similar to those of the observation vector $\boldsymbol{y}$. In the Bayesian framework the comparisons are made to, for example, the statistics of the posterior predictive distribution. The choice of statistics depends on the particular application and the type of plausible anomalies one could expect. Clearly the more statistics we check the more likely we are to incorrectly detect a problem; this is the usual multiple testing problem. We have not worried about this as our goal has been exploratory data analysis but (under appropriate conditions) one should be able to make multiple test corrections.

Even if $\boldsymbol{A}\widehat{\boldsymbol{f}}$ is a good estimate of $\boldsymbol{Af}$, $\widehat{\boldsymbol{f}}$ may be a poor estimate of $\boldsymbol{f}$. An assessment of $\widehat{\boldsymbol{f}}$ requires more prior information about $\boldsymbol{f}$ in addition to the data $\boldsymbol{y}$. Again, the explicit formula for

the bias of ℓ^2-estimates was used to provide bias bounds that may give useful information about the relative bias and its spatial dependence. On the other hand, if a collection of plausible $\boldsymbol{f}$ were available (e.g., a training data set), one could repeat the resampling approach to generate synthetic data for functions in the collection. The relative errors of the estimates would yield information on the type of errors we could have in the actual estimate $\widehat{\boldsymbol{f}}$. We have presented a way to obtain a training set of functions with controlled regularity at a point. The idea is based on a wavelet characterization of pointwise Hölder regularity. Characterizations of other types of local regularity can be used to capture different local behavior [25, 26]. As we noted before, there are also wavelet characterizations of function spaces that can be used to create collections of functions with controlled global regularity.

As we have explained, one way to validate the model in the Bayesian framework is by sampling from the posterior predictive distribution. In this framework it is also important to study the sensitivity of the results to the choice of prior distributions; the results should not be driven by the priors. Methods for sensitivity analysis are discussed in, for example, [10, 18], but there is still a need to further develop computationally efficient methods for sensitivity analyzes for large-scale inverse problems. In addition, although not every Bayesian will agree, it is important to understand the frequentist behavior of Bayesian procedures as it frees us from the potential subjectivity of the prior (e.g., [10, 39]) and provides a calibration of the procedures based on repeated sampling that is easy to interpret. But here again the computational cost can be quite high, especially for high-dimensional priors where a sensitivity analysis is more important.

In closing, we note that the validity of formal uncertainty statements —such as statistical significance, p-values, confidence regions or characteristics of posterior distributions— hinges on the assumptions made on the mathematical and statistical models used for the inference. It is important to check that such assumptions are consistent with the process that generated the actual observations. Yet, this task can be computationally expensive and may require tedious hours of systematic error checking and sensitivity analyzes. Furthermore, such hard work may not lead to reassuring formal statements regarding the validity of the assumptions for such statements would require yet more assumptions. Some subjective decisions about model validity may have to be made. Such is the nature of data analysis.

Acknowledgements. This work was partially supported by NSF (DMS 0724715, 0914987, 0723759, 0800631) and the members, ConocoPhillips, ExxonMobil, PGS, Statoil and Total of the Geo-Mathematical Imaging Group. FA was supported by the SRC (2008-23883-61232-34) and the Swedish Foundation for International Cooperation in Research and Higher Education (YR2010-7033). The examples made use of Stanford's *WaveLab*, P.C. Hansen's *Regularization Tools*, and M. Friedlander's *SPGL1*

REFERENCES

[1] F. Andersson, M. Carlsson & L. Tenorio. On the representation of functions with Gaussian wave packets. Submitted.
[2] A. Atkinson & M. Riani. *Robust Diagnostic Regression Analysis*. Springer, 2000.
[3] H. Avron & S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. Pre-print: `http://www.cs.tau.ac.il/~haima/trace-v0.pdf`
[4] G. Backus & F. Gilbert. Uniqueness in the inversion of inaccurate gross earth data. *Philos. Trans. Roy. Soc. London, A*, 266:123–192, 1970.
[5] Z. Bai, M. Fahey & G. Golub. Some large-scale matrix computation problems. *J. Computational & Applied Math*, 74:71–89, 1996.
[6] D.A. Belsey, E. Kuh & R.E. Welsh. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, 2004.
[7] N. Bissantz, T. Hohage, A. Munk & F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45:2610–2636. 2007.
[8] L.D. Brown, T. Cai, & A. DasGupta. Interval estimation from a binomial proportion. *Statistical Science*, 16:101–133, 2001.
[9] M.A. Cameron & T.R. Turner. Recursive location and scale estimators. *Commun. Statist.–Theory Meth.*, 22:2503–2515, 1993.

[10] J.B. Carlin & T.A. Louis. *Bayesian Methods for Data Analysis (3rd Edn.)*. Chapman & Hall, 2008.
[11] C.K. Carter & G.K. Eagleson. A comparison of variance estimators in nonparametric regression. *Journal of the Royal Statistical Society, B*, 54:773–780, 1992.
[12] R. Christensen. *Plane Answers to Complex Questions. The Theory of Linear Models*. Springer, 1987.
[13] D.J. Cummins, T.G. Filloon & D. Nychka. Confidence intervals for nonparametric curve estimates: toward more uniform pointwise coverage. *J. Amer. Stat. Assoc.*, 96:233–246, 2001.
[14] B. Efron & R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.
[15] H. Engl, M. Hanke & A. Neubauer. *Regularization of Inverse Problems*. Kluwer, Dordrecht, 1996.
[16] J.E. Englund, U. Holst, U. & D. Ruppert. Recursive M-estimators of location and scale for dependent sequences. *Scandinavian Journal of Statistics*, 15:147–159, 2001.
[17] N.P. Galatsanos & A.K. Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Trans. Image. Proc.*, 1:322–336, 1992.
[18] A. Gelman, J.B. Carlin, H.S. Stern & D. B. Rubin. *Bayesian Data Analysis (2nd Edn.)*. Chapman & Hall, 2003.
[19] G. Golub & U. Von Matt. Tikhonov regularization for large-scale problems. *Technical report SCCM 4-79*, 1997.
[20] J.D. Hart. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, 1997.
[21] F.J. Herrmann, P.P. Moghaddam & C.C. Stolk. Sparsity- and continuity-promoting seismic image recovery with curvelet frames. *Appl. Comp. Harm. Anal.*, 24:150–173, 2008.
[22] E. Hernández & G. Weiss. *A First Course on Wavelets*. CRC Press, 1996.
[23] M.F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *J. Commun. Statist. Simul.*, 19, 1990.
[24] S. Jaffard. Pointwise smoothness, two-microlocalisation and wavelet coefficients. *Publications Mathématiques*, 35:155–168, 1991.
[25] S. Jaffard. Pointwise regularity criteria. *C. R. Acad. Sci. Paris, Ser I*, 339:757–762, 2004.
[26] S. Jaffard. Wavelet techniques for pointwise regularity. *Ann. Fac. Sci. Toulouse Math.*, 1:3–33, 2006.
[27] P.J. Green & B.W. Silverman. *Nonparametric Regression and Generalized Linear Models: a roughness penalty approach*. Chapman & Hall, 1993.
[28] C. Gu. *Smoothing Spline ANOVA Models*. Springer, 2002.
[29] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, Monograph 89, 2001.
[30] I. Loris, G. Nolet, I. Daubechies & F.A. Dahlen. Tomographic inversion using ℓ_1-norm regularization of wavelet coefficients. *Geophys. J. Int.*, 170:359–370, 2007.
[31] M.A. Lukas. Strong robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, 25:34006, 2008.
[32] Y. Meyer. *Wavelets and Operators*. Cambridge University Press, 1992.
[33] D. Nychka. The average posterior variance of a smoothing spline and a consistent estimate of the average squared error. *Annals of Statistics*, 18:415–428, 1990.
[34] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:502–527, 1986.
[35] M. Osborne, B. Presnell & B. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9:319–337, 2000.
[36] C.R. Rao & J. Kleffe. *Estimation of Variance Components and Applications*. North-Holland, 1988.
[37] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 9:1135–1151, 1981.
[38] P.B. Stark. Inference in infinite dimensional inverse problems: discretization and duality. *Journal of Geophysical Research*, 97:14055–14082, 1992.
[39] P.B. Stark & L. Tenorio. A primer of frequentist and Bayesian inference in inverse problems. In *Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty*, pp.9–32, Wiley, 2010.
[40] E. van den Berg & M.P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31:890–912, 2008.
[41] G. Wahba. *Spline Models for Observational Data*. Regional Conference Series in Applied Mathematics, Vol. 59, SIAM, 1990.