Travel time tomography with adaptive dictionaries

Michael J. Bianco, Student Member, IEEE, Peter Gerstoft, Member, IEEE,

Abstract—We develop a 2D travel time tomography method which regularizes the inversion by modeling groups of slowness pixels from discrete slowness maps, called patches, as sparse linear combinations of atoms from a dictionary. We propose to use dictionary learning during the inversion to adapt dictionaries to specific slowness maps. This patch regularization, called the local model, is integrated into the overall slowness map, called the global model. The local model considers small-scale variations using a sparsity constraint and the global model considers larger-scale features constrained using $\ell_2$ regularization. This strategy in a locally-sparse travel time tomography (LST) approach enables simultaneous modeling of smooth and discontinuous slowness features. This is in contrast to conventional tomography methods, which constrain models to be exclusively smooth or discontinuous. We develop a maximum a posteriori formulation for LST and exploit the sparsity of slowness patches using dictionary learning. The LST approach compares favorably with smoothness and total variation regularization methods on densely, but irregularly sampled synthetic slowness maps.

Index Terms—Dictionary learning, machine learning, inverse problems, geophysics, seismology, sparse modeling

I. INTRODUCTION

Travel time tomography methods estimate Earth slowness structure, which contains smooth and discontinuous features at multiple spatial scales, from travel times of seismic waves between recording stations. The estimation of slowness (inverse of speed) models from travel times is often formulated as a discrete linear inverse problem, where the perturbations in travel time relative to a reference are used to infer the unknown structure [1], [2]. Such problems are ill-posed, with irregular ray coverage of environments, and require regularization to obtain physically plausible solutions.

We propose a 2D travel time tomography method which regularizes the inversion by assuming small groups of slowness pixels from a discrete slowness map, called patches, are well approximated by sparse linear combinations of atoms from a dictionary. In this sparse model [3], [4], the atoms represent elemental slowness patches and can be generic dictionaries, e.g. wavelets, or adapted to specific data by dictionary learning [5], [6]. This patch regularization, called the local model, is integrated into the overall slowness map, called the global model. Whereas the local model considers small-scale variations using a sparsity constraint, the global model considers larger-scale features which are constrained using $\ell_2$ regularization.

This local-global modeling strategy with dictionary learning has been successful in image denoising [7], and inpainting [9], where natural image content is recovered from noisy or incomplete data. We use this strategy to recover true slowness fields from travel time tomography by simultaneously modeling smooth and discontinuous slowness features. This gives an improvement over conventional methods with global damping and smoothness regularization [2], [10] and $\ell_2$ pixel level regularization, e.g. total variation (TV) regularization [11], [12], which regularize tomography by encouraging smooth or discontinuous slownesses. Relative to existing tomography methods based on wavelets [13], [14] and sparse dictionaries [15]–[18], our formulation of the tomography problem permits the adaptation of the sparse dictionaries to travel time data and ray sampling by dictionary learning, for which many methods exist [3], [5], [6], [8], [19]. Sparse reconstruction performance is often improved using adaptive dictionaries, which represent well specific data, relative to generic dictionaries, which achieve acceptable performance for many tasks [4].

Sparse modeling assumes signals can be reconstructed to acceptable accuracy using a small (sparse) number of vectors, called atoms, from a potentially large set or dictionary of atoms. The parsimony of sparse representations [4] often provides better regularization than, for example, traditional $\ell_2$ model damping [2]. Early sparse approaches were developed in seismic deconvolution [20], [21]. This philosophy has since become ubiquitous in signal processing for image and video denoising [3], [7], [8] and inpainting [9], and medical imaging [22], [23], to name a few examples. Recent works in acoustics and seismics have utilized sparse modeling, e.g. beamforming [24], [25], matched field processing [26], [27], estimation of ocean acoustic properties [28]–[33]. Dictionary learning has been used to denoise seismic [34], [35] and ocean acoustic [28] recordings, to regularize full waveform inversion [36], [37], and to regularize ocean sound speed profile inversion [29], [30].

Inspired by image denoising [7], we develop a sparse and adaptive 2D travel time tomography method, which we refer to as locally-sparse travel time tomography (LST). Whereas in [7], the image pixel values are directly observed, in LST the pixel values are inferred from measurements [23]. This necessitates an extra term to fit slowness pixels to travel time observations. We develop a maximum a posteriori (MAP) formulation for LST and use the iterative thresholding and signed K-means (ITKM) [6] dictionary learning algorithm to design adaptive dictionaries. This improves slowness models over generic dictionaries. We demonstrate the performance of LST for 2D surface wave tomography with synthetic slowness maps and travel time data. The LST results compare favorably with two competing methods: a smoothing and damping approach [38] referred to as conventional tomography, and TV regularization [11].
II. LST MODEL FORMULATION

In developing the LST method, we consider the case of 2D travel time tomography, where slowness of the medium varies only in two dimensions. In seismic tomography, surface wave tomography is one case where this assumption is valid [39]. The sensing configuration for such a scenario is shown in Fig. 1. We discretize a 2D slowness map as a $W_1 \times W_2$ pixel image, where the pixels have constant slowness. An array of sensors in the 2D map measure waves propagating across the array. From these observations, wave travel times between the sensors, $t' \in \mathbb{R}^M$, are obtained. We assume $t'$ given and disregard refraction of the waves, yielding a ‘straight-ray’ formulation of the problem. The tomography problem is to estimate the slowness pixels (see Fig. 1) from $t'$.

In the following, we develop separately two slowness models, deemed the $\text{global}$ and $\text{local}$ models, which will be related in Section III, and discuss dictionaries for sparse modeling. The global model considers the larger scale or global features and relates travel times to slowness. The local model considers smaller scale, localized features with sparse modeling.

A. Global model and travel times

In the global model, slowness pixels (see Fig. 1) are represented by the vector $s' = s_g + s_0 \in \mathbb{R}^N$, where $s_0$ is reference slowness and $s_g$ is perturbations from the reference, here referred to as the global slowness, with $N = W_1 W_2$. Similarly, the travel times of the $M$ rays are given as $t' = t + t_0$, where $t$ is the travel time perturbation and $t_0$ is the reference travel time. The tomography matrix $A \in \mathbb{R}^{M \times N}$ gives the discrete path lengths of $M$ straight-rays through $N$ pixels (see Fig. 1).

Thus $t$ and $s_g$ are related by the linear measurement model

$$t = As_g + \epsilon,$$  \hspace{1cm} (1)

where $\epsilon \in \mathbb{R}^M$ is Gaussian noise $\mathcal{N}(0, \sigma^2 I)$, with mean 0 and covariance $\sigma^2 I$. We estimate the perturbations, with $s_0$ and $t_0 = As_0$, known. We call (1) the $\text{global model}$, as it captures the large-scale features that span the slowness map and generates $t$.

B. Local model and sparsity

In the local model, slowness pixels (see Fig. 1) are represented by the vector $s' = s_s + s_0 \in \mathbb{R}^N$, where the sparse slowness $s_s$ is perturbations from the reference $s_0$. The slowness $s_s$ is an auxiliary latent variable that is introduced to capture local slowness behavior, and is instrumental in the estimation procedure proposed in Sec. III. In Sec. III-A, $s_s$ is precisely related to $s_g$ in a Bayesian hierarchy.

Formulating the local model, we assume that patches, or $\sqrt{n} \times \sqrt{n}$ groups of pixels from $s_s$ (see Fig. 1) are well approximated by a sparse linear combination of atoms from a dictionary $D \in \mathbb{R}^{n \times Q}$ of Q atoms. The patches are selected from $s_s$ by the binary matrix $R_i \in \{0,1\}^{n \times N}$. Hence the slownesses in patch $i$ are $R_i s_s$. The sparse model is

$$R_i s_s \approx Dx_i$$  \hspace{1cm} (2)

where $|\cdot|$ is cardinality, $x_i \in \mathbb{R}^n$ is the sparse coefficients, and $T \ll n$ is the number of non-zero coefficients. $Dx_i$ is referred to as the patch slowness. We call (2) the $\text{local model}$, as it models smaller scale, localized features contained by patches.

Each slowness patch $R_i s_s$ is indexed by the row $w_1$ and column $w_2$ of its top-left pixel in the 2D image as $(w_1, w_2)$. We consider all overlapping patches, with $w_1$ and $w_2$ differing from their neighbor by $\pm 1$ (stride of one). Further, the patches wrap-around the edges of the image [23], [40]. Thus, for a $N = W_1 \times W_2$ pixel image, there are $N^2$ patches, and the number of patches per pixel is $n$. Wrapping the patches helps capture local features at the edges of the slowness map, if there is sufficient ray sampling (see Sec. V). Without patch wrapping, pixels are modeled by as few as one patch, increasing to $n$ patches at $\sqrt{n}$ pixels from the map edge.

The atoms in $D$ are considered “elemental patches”, where only a small number of atoms are necessary to adequately approximate $R_i s_s$. Atoms can be generic functions, e.g. wavelets or the discrete cosine transform (DCT), or learned from the data (see Sec. III-C). An example of DCT atoms are shown in Fig. 2. Adaptive dictionaries, which are designed from specific instances of data using dictionary learning algorithms, often achieve greater reconstruction accuracy over generic dictionaries. Examples of dictionaries learned from synthetic travel time data (from slowness maps in Fig. 3) are shown in Fig. 4. Relative to generic dictionaries, learned dictionaries can represent the smooth and discontinuous seismic features encountered in real inversion scenarios.

III. LST MAP OBJECTIVE AND EVALUATION

We now derive a Bayesian MAP objective for the LST variables and an algorithm for its evaluation, with the ultimate goal of estimating the sparse slowness $s_s$ from (2). Assuming the travel times $t$, tomography matrix $A$, and dictionary $D$ known, the solution to the objective gives MAP estimates of the global slowness $s_g$ (from (1)), $s_s$, and the coefficients $X = [x_1, ..., x_I] \in \mathbb{R}^{Q \times I}$ describing all patches $I$ (from (2)). Since we use a non-Bayesian dictionary learning algorithm (ITK [6]), dictionary learning is added after the MAP derivation in Sec. III-C.
conditioned only on the conditional probabilities from (6), we obtain the covariance of the patch slownesses. Taking the logarithm of \( \ln p(X) = \sum_i \ln p(x_i) \).  

(8)

From (5), and (7), and (8) we obtain

\[
\ln p(s_g, s_b, X|t) \propto \ln \left\{ p(t|s_g)p(s_b|s_g)p(s_b|X)p(X) \right\} 
\]

\[
\propto -(t - A{s_g})^T \Sigma^{-1}_e (t - A{s_g}) - (s_b - s_g)^T \Sigma^{-1}_b (s_b - s_g) 
\]

\[
- \sum_i \left\{ (Dx_i - R_is_b)^T \Sigma^{-1}_{p,i} (Dx_i - R_is_b) + 2 \ln p(x_i) \right\} 
\]

(9)

Assuming the coefficients \( x_i \) sparse, we approximate \( \ln p(x_i) \) with the \( \ell_0 \) pseudo-norm \( ||x_i||_0 \), which counts the number of non-zero coefficients [3]. We further assume the number of non-zero coefficients \( T \) is the same for each patch. This gives the MAP objective as

\[
\max \left\{ \ln p(s_g, s_b, X|t) \right\} = \min \left\{ -\ln p(s_g, s_b, X|t) \right\} 
\]

\[
\propto \min \left\{ (t - A{s_g})^T \Sigma^{-1}_e (t - A{s_g}) + (s_b - s_g)^T \Sigma^{-1}_b (s_b - s_g) 
\]

\[+ \sum_i (Dx_i - R_is_b)^T \Sigma^{-1}_{p,i} (Dx_i - R_is_b) \right\} \sum_i ||x_i||_0 = T \forall i. \]

(10)

Further simplifying, we assume the errors are Gaussian iid. Thus, \( \Sigma_e = \sigma^2_e I \), \( \Sigma_b = \sigma^2_b I \), and \( \Sigma_{p,i} = \sigma^2_{p,i} I \), where \( I \) is the identity matrix. The LST MAP objective is thus

\[
\{\hat{s}_g, \hat{s}_b, \hat{X}\} = \arg \min_{s_g, s_b, X} \left\{ \frac{1}{\sigma^2_e} \|t - A{s_g}\|^2 
\]

\[+ \frac{1}{\sigma^2_b} \|s_b - s_g\|^2 + \sum_i \frac{1}{\sigma^2_{p,i}} \|Dx_i - R_is_b\|^2 \right\} \sum_i ||x_i||_0 = T \forall i, \]

subject to \( ||x_i||_0 = T \forall i \), where \( \{\hat{s}_g, \hat{s}_b, \hat{X}\} \) are the estimates of the LST variables.

B. Solving for the MAP estimate

We find the MAP estimates \( \{\hat{s}_g, \hat{s}_b, \hat{X}\} \) solving (11) via an alternating minimization algorithm [7], [23]. This strategy divides the solution of (11) into three subproblems: 1) the global problem, corresponding to the global model (1), which estimates \( \hat{s}_g \), 2) the local problem, corresponding to the local model (2) which estimates \( \hat{X} \); and 3) an averaging procedure which estimates \( \hat{s}_b \). The global problem is least squares with \( \ell_2 \) regularization. The global result is substituted into the sparse model (2) to obtain local structure by solving for the patches, and averaging their estimates. Global \( \ell_2 \) regularization has been used with TV regularization in seismic tomography [11], which was an extension of [41]. This modified TV approach is adapted to travel time tomography as a competing method in Section IV-B. LST combines a global \( \ell_2 \)-norm constraint with local regularization of patches, following [3]. A major distinction between LST and image denoising [7] is that the 2D image is inferred from travel time measurements, and not directly observed, similar to [23].
Fig. 3: Synthetic slowness maps and ray sampling. Slowness $s'$ for (a) checkerboard map and (b) smooth-discontinuous map (sinusoidal variations with discontinuity). Both maps are $W_1 = W_2 = 100$ pixels (1 km/pixel). (c) 64 stations (red X’s), giving in 2016 straight ray (surface wave) paths through synthetic images. (d) Density of ray sampling, in $\log_{10}$ rays per pixel.

The global problem is written from (11) as

$$\tilde{s}_g = \arg\min_{s_g} \frac{1}{\sigma_g^2} \|t - As_g\|_2^2 + \frac{1}{\sigma_p^2} \|s_g - s_s\|_2^2$$

where $\lambda_1 = (\sigma_p/\sigma_g)^2$ is a regularization parameter.

The local problem is written from (11), with each patch solved from the global estimate $s_s = \tilde{s}_s$ from (12) (decoupling the local and global problems), giving

$$\tilde{x}_i = \arg\min_{x_i} \|Dx_i - R_is_s\|_2^2$$

subject to $\|x_i\|_0 = T$. (13)

C. LST algorithm with dictionary learning

The expressions (12), (13), and (16) are solved iteratively, giving the LST algorithm, in Table I, as a MAP estimate of a slowness image with local sparsity constraints and a known dictionary $D$. Further implementation details are given in Sec. III-D. Dictionary learning is added to the algorithm in the solution to the local problem (13), by optimizing $D$:

$$\tilde{D} = \arg\min_D \{ \min_{x_i} \|Dx_i - R_is_g\|_2^2 \}$$

subject to $\|x_i\|_0 = T \forall i$. (17)

The dictionary learning problem (17) is here solved using the ITKM algorithm, Table II (for details, see App. A).

The ITKM solves this bilinear optimization problem (17) by alternately solving for the sparse coefficients $\tilde{x}_i$ with $D$ fixed, using thresholding [3], and solving for $\tilde{D}$ with $\tilde{x}_i$ fixed using a ‘signed’ K-means objective. In the ITKM iterations, the columns of $D$, $d_q$, are constrained to have unit norm, to prevent scaling ambiguity. For fixed sparsity $T$, the ITKM is computationally more efficient and has better guarantees of dictionary recovery than the K-SVD [6] [5].

To illustrate the content of learned dictionaries, the atoms learned during LST are shown for the checkerboard (Fig. 4(a)) and smooth-discontinuous map (Fig. 4(b)). The atoms from checkerboard (Fig. 4(a)) contain sharp edges, which correspond to shifted sharp edges from the checkerboard pattern. The atoms from smooth-discontinuous map (Fig. 4(b)) contain both smooth and discontinuous features. The smooth atoms correspond to the sinusoidal variations, whereas the atoms with sharp edges correspond to shifted features the fault region. These features give the shift-invariance property to the sparse, representation, which will be discussed in Sec. III-E. Since learned dictionaries are adapted to specific data, they better model specific data with a minimal number of atoms than prescribed dictionaries, such as Haar wavelets or DCT. Methods using dictionary learning have obtained superior performance over prescribed dictionaries in e.g. image denoising and inpainting [4].
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCI.2018.2862644, IEEE Transactions on Computational Imaging

A sparse linear combinations of atoms. This property has been

patches of diverse image content are well approximated with

neuroscience, e.g. [3], [8], [19], [45], which have shown that

Such patterns are the atoms in the dictionary

combinations of a small set of elemental patches, or patterns.

seismic image patches can be represented as sparse linear

regularized tomography is obtained under the hypothesis that

case when

obtain reasonable run times (see Sec. V-B3). In the special

complexity to be dominated by LSQR. In our simulations we

Table II). For large slowness maps, we expect the LST

algorithm [43]. For LST with dictionary learning, first the

are obtained using the orthogonal matching pursuit (OMP)

in the local estimates (13), (17)

i. Setting

j. Obtain

The complexity of each LST iteration is determined primar-

Table II: ITKM algorithm

Given: \( j, \hat{s}^j, D^0 = D^{j-1} \in \mathbb{R}^{n \times Q}, T, \) and \( h = 1 \)

Repeat until convergence:

1. Find dictionary indices per (27)

with \( \lambda_i \) such that \( \| d_h^{i-1} y_i \|_2 = 1 \).

2. Update dictionary per (32) using coefficient indices \( K \)

Out: \( D^j = D^h \)

TABLE I: LST algorithm

Given: \( t \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, s^j_\delta = 0 \in \mathbb{R}^n, D'' = \) Haar wavelet,

DCT (or) noise \( \Lambda'(0, 1) \in \mathbb{R}^{n \times Q}, \lambda_1, \lambda_2, T, \) and \( j = 1 \)

Repeat until convergence:

1. Global estimate: solve (18) using LSQR [42].

\[
\hat{s}_\delta = \arg \min_{s_\delta} \| t - A(s_\delta + s_\delta^{-1}) \|_2^2 + \lambda_1 \| s_\delta \|_2^2,
\]

with the substitution \( s_\delta = s_\delta - s_\alpha \), giving \( \hat{s}_\delta = \hat{s}_\delta + s_\alpha \).

The sparse coefficients \( x_i \) in the local estimates (13), (17) are obtained using the orthogonal matching pursuit (OMP) algorithm [43]. For LST with dictionary learning, first the dictionary is obtained from ITKM, then (13) is solved with the same sparsity level \( T \) as ITKM. Before solving (13), (17), the slowness patches \( \{ R_i \hat{s}_R \} \) are centered [4] – i.e. the mean of the pixels in each patch is subtracted. The mean of patch \( i \) is \( T_i = \frac{1}{n^2} R_i \hat{s}_R \). Hence, \( R_i \hat{s}_R \approx D x_i + 1 \sigma_i \).

The complexity of each LST iteration is determined primarily by LSQR computation in the global problem, \( O(2MN) \), and in the local problem by ITKM \( O(k n Q N) \) and OMP \( O(T n Q N) \), where \( k \) is the number of ITKM iterations (see Table II). For large slowness maps, we expect the LST complexity to be dominated by LSQR. In our simulations we obtain reasonable run times (see Sec. V-B3). In the special case when \( T = 1 \), the solution is not combinatorial, and the dictionary learning problem is equivalent to gain-shape vector quantization [5], [44].

D. Implementation and complexity

In the following, we give the implementation details and complexity of the LST algorithm (see Table I). Since \( A \) is sparse, the global estimate (12) is solved using the sparse least squares program LSQR [42]. For LSQR, (12) is rewritten as

with the substitution \( s_\delta = s_\delta - s_\alpha \), giving \( \hat{s}_\delta = \hat{s}_\delta + s_\alpha \).

Table II: ITKM algorithm

Given: \( j, \hat{s}^j, D^0 = D^{j-1} \in \mathbb{R}^{n \times Q}, T, \) and \( h = 1 \)

Repeat until convergence:

1. Find dictionary indices per (27)

with \( \lambda_i \) such that \( \| d_h^{i-1} y_i \|_2 = 1 \).

2. Update dictionary per (32) using coefficient indices \( K \)

Out: \( D^j = D^h \)

Fig. 4: Dictionary atoms learned during LST with ITKM, with \( n = 100 \) and \( Q = 150 \) for (a) checkerboard map (Fig. 3(a)) with \( T = 1 \) and (b) smooth-discontinuous map (Fig. 3(b)) with \( T = 2 \). Atoms adjusted to full grayscale range for display.

E. Advantages

Improved inversion performance over conventional and TV regularized tomography is obtained under the hypothesis that seismic image patches can be represented as sparse linear combinations of a small set of elemental patches, or patterns. Such patterns are the atoms in the dictionary \( D \). This hypothesis follows numerous works in image processing and neuroscience, e.g. [3], [8], [19], [45], which have shown that patches of diverse image content are well approximated with a sparse linear combinations of atoms. This property has been

exploited for signal denoising and inpainting [3], [8] and classification [9], [46].

Further, sparse dictionaries trained on overlapping patches possess the shift-invariance property, whereby features such as edges are recovered regardless of where they are located in an image [3]. LST enables finer resolution as permitted by the atoms in the dictionary and can exploit shift invariance. Slowness features in Fig. 3(a–b) are shifted such that only small cells of constant slowness may be used with conventional tomography (which necessitates damping, Sec. IV-A) to
illustrate this effect.

IV. COMPETING METHODS

A. Conventional tomography

We illustrate conventional tomography with a Bayesian approach [38], which regularizes the inversion with a global smoothing (non-diagonal) covariance. Considering the measurements (1), the MAP estimate of the slowness is

$$\hat{s}_{g} = (A^T A + \eta \Sigma_L^{-1})^{-1} A^T t,$$

where $\eta = (\sigma_s / \sigma_c)^2$ is a regularization parameter, $\sigma_c$ is the conventional slowness variance, and

$$\Sigma_L(i, j) = \exp(-D_{i,j}/L).$$

(20)

Here $D_{i,j}$ is the distance between cells $i$ and $j$, and $L$ is the smoothness length scale [10], [38].

B. Total variation regularization

We implement the modified TV regularization method [11], [41]. TV regularization penalizes the gradient between pixels, enforcing piecewise constant models [47], hence we might expect TV regularization to preserve well discontinuous or constant features. The TV method is adapted to the travel time tomography problem, giving the objective

$$\{\hat{s}_{g}; \hat{s}_{TV}\} = \arg \min_{s_g, s_{TV}} \left\{ \frac{1}{\sigma_c} \|t - As_g\|^2_2 + \frac{1}{\sigma_g} \|s_g - s_{TV}\|^2_2 + \frac{1}{\sigma_T} \|s_{TV}\|^2_2 \right\},$$

where $\|s_{TV}\|^2_2$ is the TV regularizer, which penalizes the gradient, and $s_{TV} \in \mathbb{R}^N$ is the TV estimate of the slowness. Similar to LST, TV (21) is solved by decoupling the problem into two subproblems: 1) damped least squares and 2) TV. The least squares problem is

$$\hat{s}_g = \arg \min_{s_g} \frac{1}{\sigma_c} \|t - As_g\|^2_2 + \frac{1}{\sigma_g} \|s_g - s_{TV}\|^2_2,$$

$$\hat{s}_{TV} = \arg \min_{s_{TV}} \frac{1}{\sigma_T} \|s_{TV}\|^2_2 + \lambda_1 \|s_g - s_{TV}\|^2_2,$$

where $\lambda_1$ is related to global LST problem (see (12)) and $s_{TV}$ is initialized to the reference slowness. The TV problem is

$$\hat{s}_{TV} = \arg \min_{s_{TV}} \frac{1}{\sigma_T} \|s_{TV}\|^2_2 + \lambda TV \|s_{TV}\|^2_2,$$

$$\lambda TV = \frac{2}{\sigma_T^2},$$

(21) is solved by alternately solving (22) and (23), like LST, as an alternating minimization algorithm. In this paper (22) is solved using LSQR [42] and (23) is solved using the TV algorithm of Chambolle [47], [48]. We set the gradient step size $\alpha = 0.25$, which is optimal for convergence and stability of the algorithm [47]. The stopping tolerance is set to $1e-2$.

V. SIMULATIONS

We demonstrate the performance of LST (Sec. III, Table I), using both dictionary learning and prescribed dictionaries, relative to conventional (Sec. IV-A) and TV (Sec. IV-B) tomography on synthetic slowness maps (e.g. Fig. 3(a,b)). The recovered slownesses from the methods are plotted in Figs. 5–11, with their performance summarized in Table III. The convergence of LST and its sensitivity to the sparsity level $T$ are shown in Fig. 12. For LST without dictionary learning, the dictionary $D$ is either the overcomplete Haar wavelet dictionary or the DCT (see Fig. 2). LST with prescribed dictionaries performs similarly to previous works which use wavelet transforms with a sparsity constraint on the coefficients [15]–[18]. Though LST with prescribed dictionaries is not competing with these methods, as only two resolutions are considered (global and local). Relative to conventional and
TV regularization, good performance can be obtained without dictionary learning and slowness is better recovered when dictionary learning is used.

Experiments are conducted using simulated seismic data from two synthetic 2D seismic slowness maps (Fig. 3(a,b)) with dimensions $W_1 = W_2 = 100$ pixels (km), as well as variations of these maps. The boxcar checkerboard in Fig. 3(a) demonstrates the recovery of discontinuous seismic slownesses. While the checkerboard slowness is quite unrealistic, it is commonly used as a benchmark for seismic tomography methods. The smooth-discontinuous map Fig. 3(b), is more realistic and illustrates fault-like discontinuities in an otherwise smoothly varying (sinusoidal) slowness map, as used in [15]. These examples illustrate the modeling flexibility of the LST algorithm. We also generate a variety of synthetic slowness maps, by altering the size of the checkerboard squares and the width and horizontal location of the discontinuity, in the checkerboard and smooth-discontinuous slowness maps, respectively. For more details, see Sec. V-A2. The travel times from the synthetic slowness maps are generated by the global algorithm. We also generate a variety of synthetic slowness maps, smoothly varying (sinusoidal) slowness map, as used in [15].

These examples illustrate the modeling flexibility of the LST algorithm. We also generate a variety of synthetic slowness maps, by altering the size of the checkerboard squares and the width and horizontal location of the discontinuity, in the checkerboard and smooth-discontinuous slowness maps, respectively. For more details, see Sec. V-A2. The travel times from the synthetic slowness maps are generated by the global algorithm. We also generate a variety of synthetic slowness maps, smoothly varying (sinusoidal) slowness map, as used in [15].

The slowness map pixels are 1 km square and sampled by $M = 2016$ straight-rays between 64 sensors, shown in Fig. 3(c). The travel times $t$ are calculated by numerically integrating along these ray paths. The 2D valid-region for slowness map estimates using conventional and TV tomography is bounded by the outermost pixels along the ray paths (see Fig. 3(c)). The valid-region for LST is obtained by a dilation operation [49] on the conventional/TV valid-region, effectively padding it by 5 pixels. The conventional/TV tomography valid region is used for error calculations for all methods.

To avoid overfitting during dictionary learning (Table II), we exclude patches from training if more than 10% of the pixels are not sampled by ray paths. This heuristic works well and we have not investigated dictionary learning from incomplete information. The RMSE (ms/km) of the estimates is given by

$$RMSE = \sqrt{\frac{1}{NP} \sum_{n} \sum_{p} (s'_{n,p} - \hat{s}_{est,n,p}^c)^2},$$

where $s'_{est,n,p}$ is $\hat{s}^c$, $\hat{s}^g$, or $\hat{s}^T_V$ for the nth pixel location, and the p-th trial. For $\sigma_s = 0$, $P = 1$. The RMSE is printed on the 1D/2D slowness estimates in Figs. 5–12.

**A. Without travel time error**

We first simulate travel times without errors ($\sigma_s = 0$) to obtain best-case results for the tomography methods. See Figs. 5 and 6 for results from LST, conventional, and TV tomography.

1) Inversion parameters for $\sigma_s = 0$: The LST tuning parameters are $\lambda_1, \lambda_2, n, Q$, and $T$. The sensitivity of the LST solutions with dictionary learning to $\lambda_1$ and $\lambda_2$ are shown in Fig. 7(c,f) for nominal values of $n = 100, Q = 150$, $T = 1$ for the checkerboard and $T = 2$ for the smooth-discontinuous (see Fig. 12(b)). With the prior $\sigma_s = 0$ s, we expect the best value of $\lambda_1 = 0$ km$^2$ (from (12)). From (12), (14) $\sigma^2_s$ is proportional to the variance of the true slowness. For the checkerboard (smooth-discontinuous) is $\sigma_s = 0.10 (0.05)$ s/km. We assume the slowness patches are well approximated by the sparse model (13), and expect $\sigma^2_s \ll \sigma^2_g$. Hence we expect the best value of $\lambda_2$ (from (14)) to be small. It is shown for both the checkerboard (Fig. 7(c)) and smooth-discontinuous maps (Fig. 7(f)) that the best RMSE for the LST with dictionary learning is obtained when $\lambda_1 = 0$ km$^2$ and $\lambda_2 = 0$, though the LST exhibits low sensitivity to these values and recovers well the true slowness for a large range of values. We show the effect of varying the sparsity level $T$ on LST RMSE performance, relative to the true slowness per (24), for the nominal LST parameters in Fig. 12(b). The values used for $T$ were chosen based on the minimum error for

![Slowness Distribution](image_url)
these curves, though LST performance exceeds conventional and TV performance for a wide range of $T$.

For the LST with the Haar wavelet and DCT dictionaries (both $Q = 169$, $n = 64$, since Haar wavelet dimensions power of 2), the best performance by minimum RMSE was achieved with $\lambda_1 = 0$ km$^2$, $\lambda_2 = 0$, and $T = 5$ for the checkerboard and $T = 2$ for smooth-discontinuous maps.

For conventional tomography, there are several methods for estimating the best values of the regularization parameters $L$ and $\eta$, but the methods not always reliable [2], [50]. To find the best parameters, we systematically varied the values of $L$ and $\eta$ (see Fig. 7(a,d)). The minimum RMSE for conventional tomography was obtained by $L = 10$ km and $\eta = 0.1$ km$^2$ for both the checkerboard and smooth-discontinuous maps.

Similarly, for TV tomography, the best values of the tuning parameters, $\lambda_1$ and $\lambda_{TV}$, were obtained by systematically varying their values (see Fig. 7(b,e)). The minimum RMSE for TV tomography was obtained by $\lambda_1 = 1$ km$^2$ and $\lambda_{TV} = 0.01$ s for both the checkerboard and smooth-discontinuous maps.

2) Results for $\sigma_0 = 0$: While the discontinuous shapes in the Haar dictionary are similar to the discontinuous content of the checkerboard image, the local features in the higher order Haar wavelets overfit the rays where the ray sampling density is poor (see Fig. 3(d)). The performance of the Haar wavelets is better for the smooth-discontinuous slowness map (Fig. 6(g–i)) than for the checkerboard (Fig. 5(e,f)). As shown in Fig. 6(g–i), the Haar wavelets add false high frequency structure to the slowness reconstruction but the trends in the smooth-discontinuous features are well preserved. The inversion performance of the DCT transform (Fig. 5(g,h) and Fig. 6(j–l)) is better than the Haar wavelets for both cases, but matches less closely the discontinuous slowness features, as the DCT atoms are smooth. The smoothness of the DCT atoms better preserve the smooth slowness structure.

The LST with dictionary learning (Fig. 5(i,j) and Fig. 6(m–o)) achieves the best overall fit to the true slowness, recovering nearly exactly the true slownesses. As in the other cases, the performance degrades near the edges of the ray sampling, where the ray density is low, but high resolution is maintained across a large part of the sampling region. The RMSE of the LST with the Haar wavelet dictionary for the checkerboard (Table III) is greater than for conventional tomography, although a better qualitative fit to the true slowness is ob-

---

**Fig. 7:** Conventional, TV, and LST tomography results for different values of regularization parameters for (a–c) checkerboard (Fig. 3(a)) and (d–f) smooth-discontinuous (Fig. 3(b)) maps, without travel time error. (a,d) Conventional $\hat{s}'_T$, effect of $L$ and $\eta$. (b,e) TV regularization $\hat{s}'_{TV}$, effect of $\lambda_1$ and $\lambda_{TV}$. (c,f) LST $\hat{s}'_L$, effect of $\lambda_1$ and $\lambda_2$. RMSE (ms/km, (24)), is printed on 2D slownesses.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCI.2018.2862644, IEEE Transactions on Computational Imaging.

Fig. 8: Conventional, TV, and LST tomography results for checkerboard map (Fig. 3(a)) with 100 realizations of Gaussian travel time error (STD 2% mean travel time): 1D slice of inversion for one noise realization against true slowness, 1D slice of mean from over all noise realizations with STD of inversion for one noise realization against true slowness, 1D slice of RMSE (ms/km) against travel time error (STD 2% mean travel time): 1D slice of

B. With travel time error

We also test the performance of the tomography methods with travel time errors. We simulate Gaussian travel time errors with \( \sigma_t = 0.02 \bar{t} \), or the uncertainty is 2% of the mean travel time, which is similar to the model implemented in [14]. For each true slowness map and method, we run the inversions for 100 realizations of noise \( \mathcal{N}(0, \sigma_t) \) (also the random initialization of \( \mathbf{D} \) also changes 100 times for LST) and summarize the statistics of the results. The noise simulation results for conventional, TV, and LST tomography are in Figs. 8 and 9. The RMSE for both approaches, calculated by (24) with \( P = 100 \), are in Table III.

1) Inversion parameters for \( \sigma_t = 0.02 \bar{t} \), the sensitivity of the LST solutions with dictionary learning to \( \lambda_1 \) and \( \lambda_2 \) are shown in Fig. 10(c,f) for nominal values of \( n = 100, Q = 150 \) and \( T = 2 \) (per Fig. 12(b)) for both the checkerboard and smooth-discontinuous maps. With the prior \( \sigma_t = 0.02 \bar{t} \), we expect for the checkerboard map (\( \sigma_r = 0.27 \bar{s}, \sigma_g = 0.10 \text{s/km} \)) the best value of \( \lambda_1 \approx 7.5 \text{ km}^2 \), and for the smooth-discontinuous (\( \sigma_r = 0.28 \bar{s}, \sigma_g = 0.05 \text{s/km} \)) the best value of \( \lambda_1 \approx 28.3 \text{ km}^2 \) (from (12)). We use \( \lambda_1 = 2 \text{ km}^2 \) for the checkerboard (Fig. 8(g–i)) and \( \lambda_1 = 10 \text{ km}^2 \) for the smooth-discontinuous map (Fig. 9(i–l)), and achieve lower RMSE than \( \lambda_1 = 1 \text{ km}^2 \) (see Fig. 10(c,f)). Although we expect the true values \( \sigma_g \) to decrease over the LST iterations, prior values of \( \sigma_g \) proportional to the variance of the true slowness work well. It is further shown in Fig. 10 that, as in the noise-free case (Sec. V-A1), the LST recovers well the true slowness for a large range of values.

For the LST with Haar wavelet and DCT dictionaries, the best performance by minimum RMSE was achieved with \( \lambda_1 = 5 \text{ km}^2, \lambda_2 = 0, \) and \( T = 5 \) for the checkerboard and \( T = 2 \) for smooth-discontinuous maps. For conventional tomography, the best values by minimum RMSE were \( L = 6 \text{ km} \) and \( \eta = 10 \text{ km}^2 \) for the checkerboard and \( L = 12 \text{ km} \) and \( \eta = 10 \text{ km}^2 \) for the smooth-discontinuous slowness maps Fig. 10(a,d). For TV tomography, the best values by minimum RMSE were \( \lambda_1 = 5 \text{ km}^2 \) and \( \lambda_{TV} = 0.02 \text{s} \) for both the checkerboard and smooth-discontinuous slowness maps Fig. 10(b,e).

Considering again the influence of the choice of \( \lambda_1 \) and \( \lambda_2 \) on the LST with dictionary learning in Fig. 10(c,f), since
Fig. 9: Conventional, TV, and LST tomography results for smooth-discontinuous map (Fig. 3(b)) with 100 realizations of Gaussian travel time error (STD 2% mean travel time). Same format as Fig. 8, except vertical 1D slice of inversion is included, and only one LST case (i–l) \( \hat{s}_i \) with dictionary learning. RMSE (ms/km, (24)), is printed on 1D errors.

\( n = 100 \) the case \( \lambda_2 = 100 \) gives equal weight to the global \( \hat{s}_{g,n} \) and patch slowness \( ns_p \) in (16). The LST obtains results similar to the best conventional estimates (see Fig. 10) for the checkerboard with \( \lambda_1 = 1 \) and \( \lambda_2 = 100 \) and for the smooth-discontinuous map with \( \lambda_1 = 10 \) and \( \lambda_2 = 500 \), though these parameter choices are suboptimal. Although in this case, the sparsity regularization of the patches has an effect similar to the conventional damping and smoothing regularization from (19), there is no direct relationship between the regularization in (19) and the sparsity and dictionary learning in (13).

2) Results for \( \sigma_\epsilon = 0.02 \bar{t} \): The LST with dictionary learning (Fig. 8(g–i) and Fig. 9(i–l)) achieves the best overall fit to the true slowness, relative to conventional tomography (Fig. 8(a–c) and Fig. 9(a–d)) and TV tomography (Fig. 8(d–f) and Fig. 9(e–h)) as evidenced by qualitative fit and RMSE (Table III). In the case of the checkerboard, a higher value \( \lambda_1 = 7 \) km\(^2\) is tested and shown in Fig. 8(j–l). Because of the increased damping, the standard deviation (STD) of the estimate (Fig. 8(k)) is less than the case \( \lambda_1 = 2 \) km\(^2\) (Fig. 8(h)), and fit to true profile is improved.

We also simulate inversion for a variety of checkerboard and smooth-discontinuous slowness maps with different geometries (per Sec. V-A2), with travel time error. The results of these tests are summarized in Table III and Fig. 11. Inversions with 10 realizations of Gaussian travel time error (\( \sigma_\epsilon = 0.02 \bar{t} \)) were performed using conventional, TV, and LST (with dictionary learning) tomography using the nominal parameters for the slowness maps in Fig. 3(a,b). Fig. 11(a,b) and Fig. 11(c,d) show the results for the varied checkerboard and smooth-discontinuous maps. LST obtains lower RMSE than TV or conventional for all simulations for both varied checkerboard and smooth-discontinuous maps, as shown in Fig. 11(a,c), a better subjective fit to the true slowness is also observed (Fig. 11(b,d)).

3) Convergence and run time: The algorithms were coded in Matlab. The LST algorithm (Table I) used 100 iterations for all cases and the ITKM (Table II) used 50 iterations. The TV algorithm (Sec. IV-B) also used 100 iterations for all cases. The inversions in Fig. 5 and Fig. 6 with (without) dictionary learning took 4 min (3 min) on a Macbook Pro 2.5 GHz Intel Core i7. For the same examples, conventional tomography (Sec. IV-A) took 20 s and TV tomography took 3 s.

In Fig. 12(a), it is shown that the LST RMSE travel time error (from \( s_e \)) decreased over the iterations and converged within at most 50 iterations. For the cases with travel time error, the RMSE approaches or falls only slightly below \( \sigma_\epsilon \). Hence, the travel time data was not overfit.

VI. CONCLUSIONS

We have derived a method for travel time tomography which incorporates a local sparse prior on patches of the slowness image, which we refer to as the LST algorithm. The LST can use predefined or learned dictionaries, though learned dictionaries gives improved performance. Relative to the conventional and TV tomography methods presented, the LST is less sensitive the regularization parameters. LST with sparse prior and dictionary learning can solve for both smooth and discontinuous slowness features in an image.

We considered the case of 2D surface wave tomography and using LST, for the dense sampling configuration and
A. ITKM algorithm details

The ITKM dictionary learning algorithm [6] (see Table II) is derived from a ‘signed’ K-means objective. In signed K-means, T-sparse coefficients $C = [c_1, ..., c_I] \in \mathbb{R}^{|T| \times I}$ with $c^I_i \in \{-1, 1\}$ are assigned to training examples $\{y_i, ..., y_I\} \in \mathbb{R}^n \times I$. The training examples are obtained from patch $i$ as $y_i = R_i s_i$, and centered. The minimization problem is

$$\begin{align*}
\{C, D\} &= \arg \min_D \sum_i \arg \min_T \|y_i - Dc_i\|_2^2 \\
&= \arg \min_D \sum_i \arg \min_{T, c_i^T = \pm 1} \|y_i - \sum_{t} d_t c^T_t y_i\|_2^2, 
\end{align*}$$

(25)

where $c^I_i$ is a non-zero coefficient and $d_t$ is the corresponding dictionary atom. Expanding (25) and requiring $\|d_i\|_2 = 1$,

$$\begin{align*}
\min_D \sum_i \min_{T, c_i^T = \pm 1} \left\{ \|y_i\|_2^2 - 2 \sum_{t} c^T_t y_i + B \right\} \\
= \|Y\|_F^2 + B - 2 \max_D \sum_{|K|=|T|} \max_{i} \|d_i\|_2 \|
\end{align*}$$

(26)

where $\cdot$ is the Frobenius norm, $B$ is a constant, and $K$ is the set of $T$ dictionary indices having the largest absolute inner product $\text{abs}(d_i y_i)$, by is by thresholding [3]

$$K(D, y_i) = \arg \max_{|K|=|T|} \|d_i y_i\|_1.$$  

(27)

From (26), the dictionary learning objective is

$$\max_D \sum_{|K|=|T|} \max_{i} \|D y_i\|_1, 
$$

(28)

which finds $D$ that maximizes the absolute norm of the $T$-largest responses from $K$. The ITKM solves (28) as a two-step algorithm. First, $K$ is obtained from (27). Then, the dictionary atoms are updated per

$$\max_{d_i} \sum_{i \in K(D, y_i)} \text{abs}(d_i y_i).$$

(29)
(29) is solved using Lagrange multipliers, with the constraint $\|d_i\|_2 = 1$. The Lagrangian function is

$$\Phi(d_l, \lambda) = \sum_{i \in K(D, y_i)} \|d_i^T y_i\|_1 - \lambda(d_l^T d_l - 1),$$

(30)

with $\lambda$ the Lagrange multiplier. Differentiating (30) gives

$$\frac{d\Phi}{dd_l} = \sum_{i \in K(D, y_i)} \text{sign}(d_l^T y_i) y_i - 2\lambda d_l,$$

(31)

The stationary point of (30), per (31), gives the update for $d_l$

$$d_l^{\text{new}} = \lambda_l \sum_{i \in K(D^{old}, y_i)} \text{sign}(d_l^{old} d_l^T y_i) y_i,$$

(32)

where $\lambda_l = 1/(2\lambda)$. The complexity of each ITKM iteration is dominated by matrix multiplication, $O(nQI)$, which is much less than K-SVD [5] which for each iteration calculates the SVD of a $n \times I$ matrix $Q$ times.

**REFERENCES**


Fig. 12: (a) LST algorithm travel time RMSE convergence vs. iteration (Table I) and (b) slowness RMSE vs. the sparsity level $T$ with and without travel time error, with dictionary learning. Results shown with and without travel time error, corresponding to the checkerboard (Fig. 5(i), 8(g)) and smooth-discontinuous (Fig. 6(m), 9(i)) slowness maps.

1983.


Michael J. Bianco received his B.Sc. in Aeronautical and Astronautical Engineering from the Purdue University College of Engineering in 2007 and his M.Sc. from the University of California San Diego Scripps Institution of Oceanography in 2015. From 2008–2012 he was an engineer with Rocketdyne, Los Angeles, USA. He is currently pursuing his Ph.D. at the University of California San Diego in the Marine Physical Laboratory. His research interests include signal processing and machine learning, with application to acoustic and seismic inverse problems.

Peter Gerstoft received the Ph.D. from the Technical University of Denmark, Lyngby, Denmark, in 1986. From 1987-1992 he was with Odegaard and Danneskiold-Samsøe, Copenhagen, Denmark. From 1992-1997 he was at Nato Undersea Research Centre, La Spezia, Italy. Since 1997, he has been with the Marine Physical Laboratory, University of California, San Diego. His research interests include modeling and inversion of acoustic, elastic and electromagnetic signals. Dr. Gerstoft is a Fellow of Acoustical Society of America and an elected member of the International Union of Radio Science, Commission F.