UNIVERSITY OF CALIFORNIA SAN DIEGO

Machine learning and sparse modeling for geophysical inverse problems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Oceanography

by

Michael J. Bianco

Committee in charge:

Peter Gerstoft, Chair
Bruce Cornuelle
William Hodgkiss
William Kuperman
Bhaskar Rao
Peter Shearer

2018

The dissertation of Michael J. Bianco is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California San Diego

2018

DEDICATION

To all my family and friends, whom I have had the luck and joy to know. To my parents, Jane and Michael A. Bianco, who always encouraged me to pursue my dreams.

EPIGRAPH

We are all agreed that your theory is crazy. The question that divides us is whether it is crazy enough to have a chance of being correct.

—Niels Bohr

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Peter Gerstoft, for his generous support and enthusiasm for our research. He has always made himself available for discussion, scientific or otherwise, is always fast with feedback, and has fostered a positive and productive lab environment. I will always be indebted to him for the opportunities I have had as a member of NoiseLab. I would also like to acknowledge my colleagues in NoiseLab, with whom I have shared many enjoyable years.

I would like to thank my committee members who have provided much helpful discussion. They have often helped me to see research problems from different perspectives, which has improved my research.

I would like to thank my family and friends, who over the years have given me so much love and support. I feel so fortunate to have such great people in my life.

Finally, I would like to acknowledge my bandmates in the WonderDogs, with whom I have shared many good times and funky grooves.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of the Acoustical Society of America 2016. Bianco, Michael; Gerstoft, Peter, Acoustical Society of America, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in the Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and Signal Processing 2017. Bianco, Michael; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in the Journal of the Acoustical Society of America 2017. Bianco, Michael; Gerstoft, Peter, Acoustical Society of America, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in the Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and

Signal Processing 2018. Bianco, Michael; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of the material as it appears in the Institute of Electrical and Electronics Engineers Transactions on Computational Imaging 2018. Bianco, Michael J.; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 7, in part, is a reprint of the material soon to be submitted for publication. Bianco, Michael J.; Gerstoft, Peter; Olsen, Kim B.; Lin, Fan-Chi. The dissertation author was the primary investigator and author of this paper.

VITA

2007	B. S. in Aeronautical and Astronautical Engineering, Purdue University, West Lafayette, Indiana
2007-2012	Aerospace engineer at Rocketdyne, Canoga Park, Los Angeles, California
2012-2018	Graduate Student Researcher, University of California San Diego, La Jolla, California
2015	M. S. in Oceanography, University of California San Diego
2018	Ph. D. in Oceanography, University of California San Diego

PUBLICATIONS

M.J. Bianco, P. Gerstoft, K.B. Olsen, and F.-C. Lin "High-resolution seismic tomography of Long Beach, CA using machine learning: delineating aquifers," 2018 (in preparation).

Q. Kong, D.T. Trugman, Z.E. Ross, **M.J. Bianco**, B. Meade, and P. Gerstoft "Machine learning in seismology - Turning data into insights," *Seismological Research Letters*, 2018 (accepted).

M.J. Bianco and P. Gerstoft, "Travel time tomography with adaptive dictionaries," *IEEE Transactions on Computational Imaging*, 2018, DOI: 10.1109/TCI.2018.2862644.

M. Wagner, **M. Bianco**, S. Nannuru, and P. Gerstoft, "Array shape calibration using low rank projections," *Proc. of 52nd IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2018.

P. Gerstoft, C.F. Mecklenbräuker, W. Seong, and **M. Bianco**, "Introduction to compressive sensing in acoustics," *Journal of the Acoustical Society of America*, 143, pp. 3731–3736, 2018.

M. Bianco and P. Gerstoft, "Adaptive travel time tomography with local sparsity," *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

M. Bianco and P. Gerstoft, "Dictionary learning of sound speed profiles," *Journal of the Acoustical Society of America*, 141(3), pp. 1749–1758, 2017.

M. Bianco and P. Gerstoft, "Regularization of geophysical inversion using dictionary learning," *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

M. Bianco and P. Gerstoft, "Compressive acoustic sound speed profile estimation," *Journal of the Acoustical Society of America*, 139(3), pp. EL90–94, 2016.

ABSTRACT OF THE DISSERTATION

Machine learning and sparse modeling for geophysical inverse problems

by

Michael J. Bianco

Doctor of Philosophy in Oceanography

University of California San Diego, 2018

Peter Gerstoft, Chair

In ocean acoustics and seismology, the Earth's subsurface is imaged using acoustic and seismic waves. As they propagate through the ocean and solid earth, these waves obtain geophysical information. This information is recovered via optimization procedures which fit physical models to wavefield observations from sensor arrays. The estimation of geophysical model parameters from the observations, typically referred to as *inverse* problems, are challenging due to many issues, e.g. noisy and incomplete observations, as well as non-linear forward models.

In this dissertation, geophysical inversion methods are developed based on sparse modeling and dictionary learning, an unsupervised machine learning method. These techniques employ more sophisticated model priors and latent representations than conventional methods, and obtain

state-of-the-art performance in a variety of signal processing tasks. Sparse modeling assumes that signals can be reconstructed to acceptable accuracy using a small (sparse) number of vectors, called *atoms*, from a larger set of atoms, or *dictionary*. Sparsifying dictioniaries can be designed from generic functions such as wavelets, or can be learned directly from the data via dictionary learning. Provided sufficient signal examples exist, dictionary learning can learn sparsifying dictionaries which obtain better performance than generic dictionaries. Conventional methods, rely on smoothness and second order statistics (e.g. empirical orthogonal functions (EOFs)) to estimate geophysical structure. In contrast, sparse methods potentially permit the recovery of true smooth and discontinuous geophysical structures.

Ocean acoustic sound speed profile (SSP) estimation requires the inversion of acoustic fields using limited observations. A specific case of sparse modeling, called compressive sensing (CS) asserts that certain underdetermined problems can be solved in high resolution, provided their solutions are sparse. CS is used to estimate SSPs in a range-independent shallow ocean by inverting a non-linear acoustic propagation model. It is shown that SSPs can be estimated using CS to resolve fine-scale structure.

To provide constraints on their inversion, ocean sound speed profiles (SSPs) are modeled often using empirical orthogonal functions (EOFs). However, this regularization, which uses the leading order EOFs with a minimum-energy constraint on their coefficients, often yields low resolution SSP estimates. It is shown that dictionary learning, a form of unsupervised machine learning, can improve SSP resolution by generating a dictionary of shape functions for sparse modeling that optimally compress SSPs; both minimizing the reconstruction error and the number of coefficients. These learned dictionaries (LDs) are not constrained to be orthogonal and thus, fit the given signals such that each signal example is approximated using few LD entries. LDs describing SSP observations from the High Frequency '97 experiment and the South China Sea are generated using the K-SVD algorithm. These LDs better explain SSP variability and require fewer coefficients than EOFs, describing much of the variability with one coefficient. Thus, LDs

improve the resolution of SSP estimates with negligible computational burden.

A 2D travel time tomography method is developed based on sparse modeling and dictionary learning. The method regularizes the inversion by modeling groups of slowness pixels from discrete slowness maps, called patches, as sparse linear combinations of atoms from a dictionary. Dictionary learning is used in the inversion method 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 ℓ_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.

Finally, the LST travel time tomography method is used to obtain high-resolution subsurface geophysical structure in Long Beach, CA, from seismic noise recorded on a "large-N" array with 5200 geophones ($\sim$13.5 million travel times). LST exploits the dense sampling obtained by ambient noise processing on large arrays by learning a dictionary of local, or small-scale, geophysical features directly from the data. Using LST, a high-resolution 1 Hz Rayleigh wave phase speed map of Long Beach is obtained. Among the geophysical features shown in the map, the important Silverado aquifer is well isolated relative to previous surface wave tomography studies. The results show promise for LST in obtaining detailed geophysical structure in travel time tomography studies.

Chapter 1

Introduction

1.1 Background

Physical information about solid earth and oceanic structure has many different purposes, such as seismic hazard assessment [1,2], estimation of ocean heating from climate change [3], and acoustic communication [4]. However, it is generally too expensive and difficult to obtain directly large scale estimates of ocean temperature distributions or geology. Challenges for obtaining these measurements include obvious physical limitations, e.g. the impracticality of excavation or sampling every point in the ocean, and the high costs of existing sensor arrays and deployment. Hence, geoscientists and oceanographers rely on remote or indirect measurements for estimating geophysical parameters. The problems of estimating such parameters from observations, generally referred to as *inverse problems*, pose their own challenges which must be addressed to obtain useful models.

In ocean acoustics and seismology, the Earth's subsurface is imaged using acoustic and seismic waves. These waves are generated by anthropogenic sources such as explosions or airguns (or even traffic noise [5, 6]), and natural sources such as earthquakes and ocean-atmospheric coupling [7]. In their transmission, which can extend across ranges of thousands

of kilometers, seismic and acoustic waves obtain information about the transmission medium. When these wavefields are observed on arrays of sensors, this information can be extracted. Generally, observations are obtained as either travel times between the array elements, full seismic waveforms, or acoustic modes. The inverse problem is then to find a physical model that best fits the observations. A basic inverse problem statement is

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_{n=1}^{N} p(y_i - G(\mathbf{x})) \tag{1.1}$$

where p is a penalty (e.g. least squares), y_n is the data (N points), G is the forward model, $\mathbf{x} \in \mathbb{R}^M$ are M model parameters, and $\widehat{\mathbf{x}}$ is the model parameter estimate. In travel time tomography, y_n is a travel time observation, and G is the ray propagation relating y_i to slownesses $\mathbf{x}$.

Typically, inverse problems in geoscience (e.g. Eq. 1.1) are difficult to solve due to limited and imperfect observations, and the non-linearity of the forward operator G. Thus, solving Eq. 1.1 directly will almost certainly give poor results. These problems require careful regularization to obtain physically plausible results. Much research in geophysics has been devoted to inverse problem theory [8–11]. Conventional regularization techniques in geophysics include damping and smoothing [8, 11] and total variation (TV) [12, 13], and characterize environment variability using only second-order statistics, e.g. empirical orthogonal functions (EOFs) [14,15]. These methods regularize Eq. 1.1 by penalizing rough or piecewise-discontinuous features in $\mathbf{x}$. Damping is accomplished via ℓ_2-regularization (a.k.a. Tikhonov regularization) [11], and TV-regularization [16] penalizes the gradient between pixels.

While the smoothing and orthogonality constraints in conventional regularization are convenient, they potentially over-simplify the parameterization and mask true geophysical features, which are smooth and discontinuous at multiple spatial scales [1, 11]. As a result, the parameter estimates may not converge to their true values, and geophysical features such as faults and fine-scale ocean sound speed structures and are potentially lost. A large body of inverse problem

theory has been developed in related signal processing fields, including image processing and medical imaging [17,18]. Although these fields are not geophysical, their strategies are relevant to geophysical inverse problems. Important relelvant developments have included sparse modeling, a modern interpretation of parsimony, and machine learning. These modern signal processing paradigms have the potential to transform geophysical inverse theory.

This dissertation concerns the development of sparse modeling (a.k.a. compressive sensing (CS) in acoustics) [17–20] and machine learning-based methods [21, 22] for improving geophysical parameter estimation in ocean acoustics and seismology. Sparse modeling and machine learning techniques represent a modern paradigm for signal processing. These techniques, which employ more sophisticated model priors and latent representations than conventional methods, obtain state-of-the-art performance in a variety of signal processing tasks. Sparse modeling assumes that signals can be reconstructed to acceptable accuracy using a small (sparse) number of vectors, called *atoms*, from a potentially large set of atoms, called a *dictionary*. Early sparse approaches were developed in seismic deconvolution [23, 24]. This philosophy has since become ubiquitous in signal processing for image restoration tasks such as denoising and inpainting [18], and medical imaging [25]. Recent works in ocean acoustics and seismics have utilized sparse modeling, e.g. beamforming [26, 27], matched field processing [28], estimation of ocean acoustic properties [29–34], and seismic tomography [35–38]. In Chapter 2 of this dissertation [31], a CS-based approach to ocean SSP inversion is developed.

Machine learning in the context of this dissertation provides a means of learning an optimal sparsifying dictionary directly from data examples. An important problem in sparse modeling is finding the best dictionary for sparsely representing specific signals. Such dictionaries can, for instance, be composed of wavelet functions, or the discrete cosine transform (DCT). These predefined or "generic" dictionaries perform well for many signals. Some works in seismic tomography have used sparse modeling, e.g. [36, 39]. However, using a form of unsupervised machine learning, called *dictionary learning*, optimal dictionaries can be learned directly from

specific data [18]. It has been shown that learned dictionaries outperform generic dictionaries when sufficient signal examples are available. Dictionary learning-based techniques have recently obtained compelling results in ocean acoustics [31–33, 40] and seismology [41–43]. Dictionary learning has been used to denoise seismic [39,44] and ocean acoustic [29] recordings, to regularize full waveform inversion [41, 45], and to regularize ocean SSP inversion [30, 32]. In Chapters 3 and 4 of this dissertation [32, 33], dictionary learning is applied to ocean SSP data.

Machine learning has found many useful applications in geosciences, including seismic waveform classification [46], event localization [45, 47, 48], earthquake prediction [49], and earthquake early warning [50]. Part of the success of these methods arises from the existence of large amounts of training and ground-truth data. However, in geophysical tomography, little-to-no training or ground-truth exists. This issue has driven development of ML-based methods that do not depend on large volumes of training data [41, 43, 45] or use synthetic data for training [51, 52]. Methods that do not require explicit training data are based on adaptive, dictionary learning frameworks [17, 18]. In these adaptive approaches, data observations themselves are used for training with unsupervised learning. Recent results have been obtained in full waveform inversion in a geophysical exploration context [41, 45]. In Chapters 5 and 6 of this dissertation, [33, 43] an adaptive approach to travel time tomography, called locally sparse travel time tomography (LST) is developed. LST is a machine learning-based travel time tomography method which requires no training data. Relative to previous travel time tomography methods which enforce exclusively smooth and discontinuous models, including conventional straight ray [53] and eikonal tomography [54], the sparse model in LST permits both smooth and discontinuous geophysical features. In Chapter 7 of this dissertation, the LST method use to perform ambient noise tomography (ANT) using data from a large seismic array.

1.2 Objectives

The objectives of this dissertation are listed below:

1. Determine whether CS can be use to recover a true geoacoustic model better than conventional methods. CS, a special case of sparse modeling, asserts that certain underdetermined problems can be solved in high resolution, provided their solutions are sparse. CS theory is applied to the ocean SSP estimation in a shallow water environment to evaluate its applicability to geophysical inversion. Acoustic data is simulated using a modal acoustic forward model with SSPs from the SWellEx-96 acoustics experiment. The response of the non-linear forward model is linearized and the SSPs are estimated assuming they are sparse in a half-sinusoidal dictionary and in EOF representations.

2. Assess the compression and inversion performance of dictionary learning-based ocean SSP representations relative to conventional, EOF-based representations. A major challenge in sparse modeling is the design of sparsifying dictionaries for signal representation. Such dictionaries can be learned from specific signal examples using dictionary learning. Dictionaries of ocean SSPs from the High Frequency (HF)-97 experiment [55, 56] and internal waves studies from the South China Sea (SCS) [57] are obtained with dictionary learning to see if improved inversion results over EOFs can be obtained. The SSP compression performance of the learned dictionaries is also compared with EOFs, for environmental communication.

3. Develop a sparse modeling-based theory for travel time tomography, which learns the dictionary from the travel time data, and compare its performance with conventional tomography methods. Dictionary learning-based techniques have achieved state-of-the-art performance in adaptive image denoising and medical imaging, using corrupted signal examples themselves as training data (no ground-truth data available). A dictionary learning-based approach to 2D travel time tomography, formulated for ANT, is developed and

tested on synthetic tomographic data. The performance of the approach is compared with conventional (damping and smoothing) and TV-regularized tomography techniques.

4. Apply the sparse modeling-based travel time tomography method to data from a large, dense seismic array and compare the results with competing approaches. The dictionary learning-based 2D tomography approach is demonstrated on seismic data obtained from the very dense Long Beach array in 2011. The high-resolution features obtained are assessed against those obtained via a competing state-of-the-art seismic tomography method: eikonal tomography. The geophysical features visible in the LST phase speed map are supported by prior geological studies and simulation based on the studies.

1.3 Basic concepts

1.3.1 Sparse modeling

Sparse models represent a modern interpretation of parsimony in statistics and model selection [17, 18, 58–60], and are capable of representing diverse signal features, including localized features like bumps and edges. There has been an explosion of sparse modeling literature in the last two decades, first applied in geophysics [23, 24], and popularized in statistics by the LASSO estimator [61]. This popularity of sparse modeling has also been driven by the development of CS [19, 20], which showed that under certain conditions the frequency resolution of recovered signals can exceed the limit established by the Shannon-Nyquist sampling theorem. The applications of sparse modeling are very broad. Some examples from signal processing include image denoising [17, 62, 63], inpainting [64], and medical imaging [25, 65]. Recent works in acoustics and seismology have utilized sparse modeling, e.g. beamforming [26, 27], matched field processing [28, 66], estimation of ocean acoustic properties [29–34]. In this dissertation sparse modeling is used to model ocean SSPs, and solid earth geophysical features. Dictionary

learning [17, 18] is also used in these works, and is discussed in the following section.

In the following, some basic concepts of sparse modeling and their relation to the work in this dissertation are discussed. A linear measurement model of an arbitrary signal $\mathbf{x}$ gives observations $\mathbf{y} = [y_1, ..., y_N]^{\mathrm{T}} \in \mathbb{R}^N$ as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon} \tag{1.2}$$

where $\mathbf{A} \in \mathbb{R}^{N \times M}$ is the measurement matrix relating data $\mathbf{y} = [y_1, ..., y_N]^{\mathrm{T}}$ to model coefficients $\mathbf{x}$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ is Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$. Given signal observations $\mathbf{y}$, the model coefficients $\mathbf{x}$ can be estimated using Eq. 1.1, which corresponds to the maximum likelihood estimator when the penalty p is least-squares. However, as often the case in geophysics, maximum likelihood will give erratic results due the ill-posedness of the problem and the often ill-conditioned nature of $\mathbf{A}$. One classic regularization technique is ℓ_2 regularization (which is a special case of Tikhonov regularization [11]). The ℓ_2-regularized estimator of $\mathbf{x}$ is

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda_1 \|\mathbf{x}\|_2, \tag{1.3}$$

where λ_1 is a regularization parameter. The ℓ_2-norm penalizes large coefficient values in $\mathbf{x}$. Eq. 1.3 can be understood as a Bayesian *maximum a posteriori* (MAP) estimator of $\mathbf{x}$. In this interpretation, the ℓ_2-norm is derived from a Gaussian *prior* distribution with zero mean and uniform, diagonal covariance. Eq. 1.3 has an analytic solution.

It has been observed however, that many real-world data sets have supergaussian (e.g. Laplacian) distributions, which have 'heavy tails' and large mass near zero [67]. Coefficients drawn from such distributions are considered sparse, since they tend to be zero value, or if non-zero, have large values. Sample Gaussian and Laplacian distributions are plotted in Fig. 7.1. If the true model is sparse, as in the case of compressive beamforming [26, 68], where there are often only few active directions of arrival, the true DOAs can be better recovered using sparse

modeling. A basic sparse model is given formally as

$$\mathbf{y} \approx \mathbf{A}\mathbf{x} \text{ and } |\mathbf{x} \neq 0| \ll M, \tag{1.4}$$

where $|\cdot|$ denotes cardinality. An objective for sparse coefficients $\widehat{\mathbf{x}}$ is

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda_2 \|\mathbf{x}\|_0, \tag{1.5}$$

where λ_2 is a regularization parameter, and $\|\cdot\|_0$ is the ℓ_0 pseudo-norm, which counts the number of non-zero coefficients. In this way, the sparse objective Eq. 1.4 explicitly encourages parsimonious solutions to the likelihood problem. However, Eq. 1.4 is a non-convex, NP-hard combinatorial problem, which in many cases is prohibitively expensive to solve. Approximate solution methods, such as the greedy orthogonal matching pursuit (OMP) method [69], can make solution of Eq. 1.4 reasonable. Given sufficiently sparse signals, OMP provides an adequate approximation to the true, combinatorial solution.

A sparse objective to Eq. 1.2 can also be written using the ℓ_1-norm, which is a convex relaxation of the ℓ_0-norm

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda_3 \|\mathbf{x}\|_1, \tag{1.6}$$

where λ_3 is a regularization parameter. This formulation is also known as the LASSO [61]. The ℓ_1-norm, from the Laplacian prior, encourages estimates $\widehat{\mathbf{x}}$ which have few non-zero entries with large magnitude (relative to the ℓ_2 which has many small entries) due the probability mass distribution of the Laplacian. Under certain conditions, Eq. 1.4 is equivalent to Eq. 1.5 [20], and the ℓ_1-norm provides an adequate approximation to the ℓ_0-norm regularizer and can solve (potentially large) problems in polynomial time. Other means of enforcing sparsity include the popular sparse bayesian learning (SBL) approach, which is also quite fast [59]. SBL uses the Type-II likelihood to optimize the hyperparameters of a Gaussian prior. It is less sensitive to the

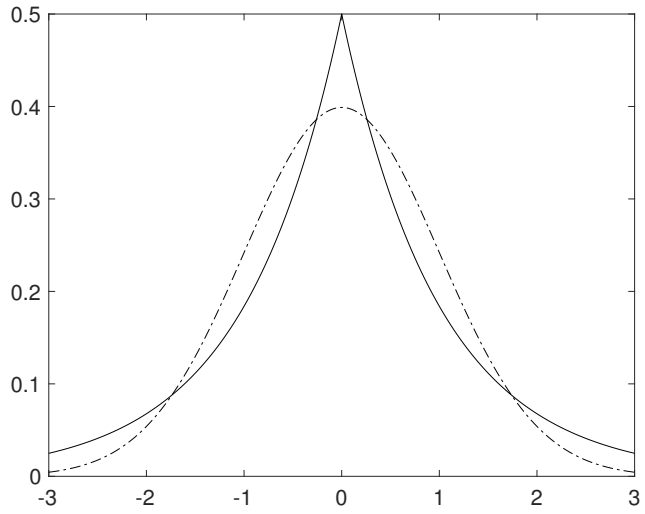


Figure 1.1: Probability density functions of Gaussian (dashed line) and Laplacian (solid line) distributions. Densities have zero mean.

sensing matrix or dictionary coherence of $\mathbf{A}$ than ℓ_1 methods.

Sparse distributions can be obtained by projecting signals onto sets, or *dictionaries* of shape functions, such as wavelets [70]. For example, it has been observed that the projection of small rectangular groups of pixels from images, called *patches*, of natural scenes onto wavelet bases have sparse distributions [71,72]. Here the dictionary is defined as $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_Q] \in \mathbb{R}^{N \times Q}$, which represents N-dimensional signals, and is comprised of Q vectors, called *atoms*. Typically, the atoms are normalized such that $\|\mathbf{d}_q\|_2 = 1$. Signals $\mathbf{x}$ are approximated as a sparse linear combination of atoms from $\mathbf{D}$. This is stated formally as

$$\mathbf{x} \approx \mathbf{D}\alpha \text{ and } |\alpha \neq 0| \ll Q, \tag{1.7}$$

where $\alpha \in \mathbb{R}^Q$ are the coefficients. A sparse MAP solution to Eq. 1.2, with Eq. 1.7, is

$$\widehat{\alpha} = \arg\min_{\alpha} \|\mathbf{AD}\alpha - \mathbf{y}\|_2^2 + \lambda_4 \|\alpha\|_1, \tag{1.8}$$

where λ_4 is a regularization parameter. Under certain conditions on the sensing matrix $\mathbf{A}$, the true signal α (and thereby $\mathbf{x}$) can be well recovered [73]. This concept is the cornerstone of CS, and is applied in this dissertation to ocean SSP profile estimation from few acoustic measurements [31].

Although the ℓ_0-norm is non-convex, for sufficiently sparse signals greedy solution methods (e.g. OMP) provide good approximations to the true, combinatorial solution. Further, the ℓ_0-norm regularizer provides a means of explicitly defining the sparsity of the desired solution, which is useful for adaptive signal processing frameworks such as adaptive denoising [17]. Eq. 1.5 can be rewritten as

$$\widehat{\alpha} = \arg\min_{\alpha} \|\mathbf{AD}\alpha - \mathbf{y}\|_2 \text{ subject to } \|\alpha\|_0 \leq T, \tag{1.9}$$

where T is the number of non-zero coefficients in the sparse solution. This sparse objective formulation is employed in Chapters 5–7, in an adaptive travel time tomography formulation.

1.3.2 Dictionary learning

Many signals, including natural images [74–76], and seismic profiles [23] are well approximated using a small (sparse) number of vectors from a dictionary $\mathbf{D}$, provided a dictionary exists under which their representation is sparse. The atoms in $\mathbf{D}$ are considered "elemental signals", where only a small number of atoms are necessary to adequately approximate $\mathbf{x}$. The challenge addressed here is the design of optimal sparsifying dictionaries for given classes of signals.

Dictionary atoms can be designed ad-hoc using generic functions, e.g. wavelets or the discrete cosine transform (DCT). Given enough signal examples for training, atoms can can also be learned directly from data using dictionary learning [17, 77–82]. Sparse reconstruction performance in a variety of problems is often improved using learned dictionaries, which represent well specific data, relative to generic dictionaries, which achieve acceptable performance for many tasks [18]. Applications where dictionary learning has achieved state-of-the-art results

include image compression, denoising, and inpainting [17, 62, 74, 75], and medical imaging [25]. Dictionary learning-based techniques have recently obtained compelling results in ocean acoustics [29–32, 40] and seismology [39, 41, 41–45].

Early dictionary learning techniques were based on neural networks [72] and vector quantization [76], e.g. [77, 79]. Many popular dictionary learning approaches now exist [17, 18, 83, 84]. Given I data examples $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_I] \in \mathbb{R}^{N \times I}$, the dictionary learning problem can be written as

$$\{\widehat{\mathbf{D}}, \widehat{\mathbf{A}}\} = \arg\min_{\mathbf{D}} \left\{ \arg\min_{\alpha_i} \|\mathbf{D}\alpha_i - \mathbf{x}_i\|_2 \text{ subject to } \|\alpha_i\|_0 \leq \ T \ \forall \ i \right\}, \tag{1.10}$$

where $\mathbf{A} = [\alpha_1, ..., \alpha_I] \in \mathbb{R}^{Q \times I}$ are the coefficients corresponding to the I examples, and $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{A}}$ are the dictionary and coefficient estimates. Eq. 1.10 is bilinear optimization problem, which in clustering-based dictionary learning is solved using a simple, two-step alternating minimization procedure:

1. Sparse coding: Given dictionary $\mathbf{D}$, solve for up to T non-zero coefficients in α_i corresponding to examples $\mathbf{x}_i$ for all i

2. Dictionary update: Given coefficients $\mathbf{A}$, solve for $\mathbf{D}$ which minimizes reconstruction error for $\mathbf{X}$, with $\|\mathbf{d}_q\|_2 = 1$.

Dictionary atom normalization ($\|\mathbf{d}_q\|_2 = 1$) is typical, as it removes scaling ambiguity from the coefficient and atom estimates. The sparse coding stage can be accomplished using any number of sparse solvers (e.g. OMP). Since $\mathbf{D}$ is initially unknown, it must be initialized. Acceptable initializations include Gaussian noise or generic dictionaries (e.g. DCT).

One clustering-based dictionary learning algorithm is the classic K-SVD algorithm [79]. The K-SVD performs the dictionary update step (2, from above) by solving for atoms based on the residual error obtained from excluding that atoms from the reconstruction of examples $\mathbf{X}$ (for

more details, see Chapter 4). The K-SVD algorithm is given in Table 1.1. The K-SVD is used in Chapters 3 and 4 to compress ocean SSPs. In Chapters 5–7, the iterative thresholding and K-means dictionary learning algorithm [83] is used to obtain sparse, adaptive representations of geophysical features.

Table 1.1: The K-SVD Algorithm (Ref. [79])

	Given: $\mathbf{X} \in \mathbb{R}^{N \times I}$, $\mathbf{D}^0 \in \mathbb{R}^{N \times Q}$, $T \in \mathbb{N}$, and $k = 0$
	Repeat until convergence:
1.	Sparse coding
	for $i = 1 : I$
	solve using any sparse solver
a:	$\widehat{\alpha}_i = \underset{\alpha_i \in \mathbb{R}^N}{\arg\min} \lVert \mathbf{x}_i - \mathbf{D}^k \alpha_i \rVert_2$ subject to $\lVert \alpha_i \rVert_0 \leq T$
	end
b:	$\mathbf{A} = [\widehat{\alpha}_1, ..., \widehat{\alpha}_I]$
2.	Dictionary update
	for $j = 1 : Q$
a:	compute reconstruction error $\mathbf{E}_j$ as
	$\mathbf{E}_j = \mathbf{X} - \sum_{p \neq j} \mathbf{d}_n^k \alpha_T^p$
b:	obtain $\mathbf{E}_j^R$, α_R^j corresponding to nonzero α_T^j
c:	apply SVD to $\mathbf{E}_j^R$
	$\mathbf{E}_j^R = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$
d:	update $\mathbf{d}_j^k$: $\mathbf{d}_j^k = \mathbf{U}(:, 1)$
e:	update α_R^j: $\alpha_R^j = \mathbf{V}(:, 1)\mathbf{S}(1, 1)$
	end
f:	$\mathbf{D}^{k+1} = \mathbf{D}^k$
	$k = k + 1$

1.3.3 Ambient noise processing and tomography

In ANT, seismic noise generated by ocean-atmospheric interactions [7] or anthropogenic sources [5, 6], is cross-correlated between seismic sensors over periods of days to months to obtain travel times between sensors. The calculated travel times are used to estimate phase speed structure [85, 86]. The number travel times in ANT can be very large, and the coverage of a region

dense, as the number of cross-correlations is $N(N-1)/2$ (with N the number of sensors). Yet, the estimation of high-resolution phase speed structure with ANT remains an ill-posed inverse problem due to many factors, including irregular sensor distributions, phase ambiguities in the noise cross-correlations (especially for high-frequency surface waves), and non-isotropic noise distributions [54]. The success of ANT, critical for improving our understanding of Earth's structure and characterizing seismic hazard [1, 2], are reliant on signal processing theory to obtain physically plausible results. In this dissertation (Chapter 7), a the machine learning-based tomography method called LST (Chapter 6) [43], is demonstrated on data obtained from the "large-N" Long Beach array, deployed in Long Beach, CA, USA in 2011. It is shown that LST can improve ANT results over previous methods.

1.4 Scope of the dissertation

Chapter 2 [31] investigates the application of CS theory to geophysical inversion. Ocean SSPs are sparsely modeled using dictionaries of sound speed perturbations from a reference sound speed. The perturbations are estimated from an acoustic forward model using CS in a 160m depth shallow water environment. Acoustic data is simulated using the Kraken modal acoustic forward model [87] with SSPs from the SWellEx-96 acoustics experiment. For use with CS theory, the response of the non-linear acoustic forward model is linearized. The SSP dictionaries are comprised of half-sinusoial shape functions and EOFs from the SWellEx-96 acoustics experiment.

Chapters 3 [33] investigates the use of dictionary learning for ocean SSP parameterization. Dictionaries of ocean SSPs are learned from the HF-97 experiment [55, 56]. The compression and inversion capabilities of learned dictionaries are compared with EOFs estimated from the same data.

Chapter 4 [32] extends the investigation of Chapter 2, using dictionary learning for

ocean SSP parameterization with data from both the from the HF-97 experiment [55, 56] and internal waves studies from the SCS [57]. The compression and inversion capabilities of learned dictionaries are compared with EOFs estimated from the same data and the theory is further developed.

Chapter 5 [42] investigates the formulation of a machine-learning (dictionary learning) based travel time tomography method, called LST. A hierarchal modeling strategy is developed, which constrains large-scale or *global* slowness features with ℓ_2 regularization, and smaller-scale or *local* slowness features using sparse modeling and dictionary learning. The local features are rectangular groups of neighboring pixels called *patches*. The patches are represented using few (sparse) atoms from the dictionary. Further, the dictionary is trained on patches from the ℓ_2-regularized global slowness map. LST is formulated as a Bayesian MAP estimator, and is solved iteratively to obtain the final slowness map. The performance of LST is evaluated against conventional tomography on synthetic slownesses.

Chapter 6 [43] further formalizes the theoretical developement from Chapter 5 of the LST tomography method. The performance of LST is further compared with TV-regularization and a greater variety of synthetic slownesses.

Chapter 7 performs ANT with the LST method [43] using seismic data obtained from the very dense Long Beach array in 2011. The high-resolution geophysical features obtained are compared with those obtained from a competing state-of-the-art seismic tomography method: eikonal tomography. The geophysical features visible in the LST phase speed map are supported by prior geological studies and simulations based on the studies. It is found that the Silverado aquifer, which is an important source of fresh water in the Long Beach region, is better localized relative to previous travel time tomography studies.

1.5 Bibliography

[1] N. Rawlinson, S. Pozgay, and S. Fishwick, "Seismic tomography: a window into deep earth," *Phys. Earth Plan. Inter.*, vol. 178, no. 3, pp. 101–135, 2010.

[2] D. C. Bowden, V. C. Tsai, and F. C. Lin, "Site amplification, attenuation, and scattering from noise correlation amplitudes across a dense array in Long Beach, CA," *Geo. Res. Let.*, vol. 42, no. 5, pp. 1360–1367, 2015.

[3] K. F. Woolfe, S. Lani, K. G. Sabra, and W. A. Kuperman, "Monitoring deep-ocean temperatures using acoustic ambient noise," *Geophys. Res. Lett.*, vol. 42, no. 8, pp. 2878–2884, 2015.

[4] A. C. Singer, J. K. Nelson, and S. S. Kozat, "Signal processing for underwater acoustic communications," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 90–96, 2009.

[5] N. Riahi and P. Gerstoft, "The seismic traffic footprint: Tracking trains, aircraft, and cars seismically," *Geophys. Res. Letters*, vol. 42, no. 8, pp. 2674–2681, 2015.

[6] J. P. Chang, S. A. L. de Ridder, and B. L. Biondi, "High-frequency Rayleigh-wave tomography using traffic noise from Long Beach, California," *Geophys.*, vol. 81, no. 2, pp. B43–B53, 2016.

[7] J. Traer, P. Gerstoft, P. D. Bromirski, and P. M. Shearer, "Microseisms and hum from ocean surface gravity waves," *J. Geophys Res. Solid Earth*, vol. 117, no. B11, 2012.

[8] A. Tarantola, *Inverse problem theory*. Elsevier Sci. Pub. Co., Inc., 1987.

[9] F. A. Dahlen and J. Tromp, *Theoretical global seismology*. Princeton university press, 1998.

[10] C. D. Rodgers, *Inverse methods for atmospheric sounding: theory and practice*. World Sci. Pub. Co., 2000.

[11] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter estimation and inverse problems*, 2nd ed. Elsevier, San Diego, 2013.

[12] Y. Lin and L. Huang, "Quantifying subsurface geophysical properties changes using double-difference seismic-waveform inversion with a modified total-variation regularization scheme," *Geo. J. Int.*, vol. 203, no. 3, pp. 2125–2149, 2015.

[13] X. Zhang and J. Zhang, "Model regularization for seismic traveltime tomography with an edge-preserving smoothing operator," *J. Appl. Geo.*, vol. 138, pp. 143–153, 2017.

[14] L. R. LeBlanc and F. H. Middleton, "An underwater acoustic sound velocity data model," *J. Acoust. Soc. Am.*, vol. 67, no. 6, pp. 2055–2062, 1980.

[15] M. I. Taroudakis and J. S. Papadakis, "A modal inversion scheme for ocean acoustic tomography," *J. Comput. Acoust.*, vol. 1, no. 4, pp. 395–421, 1993.

[16] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging and Vision*, vol. 20, no. 1–2, pp. 89–97, 2004.

[17] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[18] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014.

[19] E. Candés, "Compressive sampling," *Proc. Internat. Cong. Math.*, vol. 3, pp. 1433–1452, 2006.

[20] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[21] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[22] K. Murphy, *Machine learning: a probabilistic perspective*, 1st ed. MIT Press, Cambridge, MA USA, 2012.

[23] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1–norm," *Geophys.*, vol. 44, no. 1, pp. 39–52, 1979.

[24] N. R. Chapman and I. Barrodale, "Deconvolution of marine seismic data using the l1 norm," *Geophys. J. Int.*, vol. 72, no. 1, pp. 93–100, 1983.

[25] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1028–1041, 2011.

[26] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 260–271, 2014.

[27] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," *IEEE Sig. Proc. Let.*, vol. 23, no. 10, pp. 1469–1473, 2016.

[28] K. L. Gemba, W. S. Hodgkiss, and P. Gerstoft, "Adaptive and compressive matched field processing," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 92–103, 2017.

[29] M. Taroudakis and C. Smaragdakis, "De-noising procedures for inverting underwater acoustic signals in applications of acoustical oceanography," *EuroNoise 2015 Maastricht*, pp. 1393–1398, 2015.

[30] T. Wang and W. Xu, "Sparsity-based approach for ocean acoustic tomography using learned dictionaries," *OCEANS 2016 Shanghai IEEE*, pp. 1–6, 2016.

[31] M. Bianco and P. Gerstoft, "Compressive acoustic sound speed profile estimation," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. EL90–EL94, 2016.

[32] ——, "Dictionary learning of sound speed profiles," *J. Acoust. Soc. Am*, vol. 141, no. 3, pp. 1749–1758, 2017.

[33] ——, "Regularization of geophysical inversion using dictionary learning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[34] Y. Choo and W. Seong, "Compressive sound speed profile inversion using beamforming results," *Remote Sens.*, vol. 10, no. 5, p. 704, 2018.

[35] I. Loris, G. Nolet, I. Daubechies, and F. A. Dahlen, "Tomographic inversion using ℓ_1–norm regularization of wavelet coefficients," *Geo. J. Int.*, vol. 170, no. 1, pp. 359–370, 2007.

[36] I. Loris, H. Douma, G. Nolet, I. Daubechies, and C. Regone, "Nonlinear regularization techniques for seismic tomography," *J. Comput. Phys.*, vol. 229, no. 3, pp. 890–905, 2010.

[37] J. Charléty, S. Voronin, G. Nolet, I. Loris, F. J. Simons, K. Sigloch, and I. C. Daubechies, "Global seismic tomography with sparsity constraints: comparison with smoothing and damping regularization," *J. Geophys. Research: Solid Earth*, vol. 118, no. 9, pp. 4887–4899, 2013.

[38] H. Fang and H. Zhang, "Wavelet-based double-difference seismic tomography with sparsity regularization," *Geo. J. Int.*, vol. 199, no. 2, pp. 944–955, 2014.

[39] S. Beckouche and J. Ma, "Simultaneous dictionary learning and denoising for seismic data," *Geophys.*, vol. 79, no. 3, pp. 27–31, 2014.

[40] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," *J. Acoust. Soc. Am.*, vol. 142, no. 3, pp. 1176–1188, 2017.

[41] L. Zhu, E. Liu, and J. H. McClellan, "Sparse-promoting full waveform inversion based on online orthonormal dictionary learning," *Geophys.*, vol. 82, no. 2, pp. R87–R107, 2017.

[42] M. Bianco and P. Gerstoft, "Adaptive travel time tomography with local sparsity," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[43] ——, "Travel time tomography with adaptive dictionaries," *IEEE Trans. Comput. Imag.*, 2018, DOI: 10.1109/TCI.2018.2862644.

[44] Y. Chen, "Fast dictionary learning for noise attenuation of multidimensional seismic data," *Geo. J. Int.*, vol. 209, no. 1, pp. 21–31, 2017.

[45] A. Li, Z. Peng, D. Hollis, L. Zhu, and J. H. McClellan, "High-resolution seismic event detection using local similarity for Large-N arrays," *Nature Scientific Reports*, vol. 8, no. 1, p. 1646, 2018.

[46] Z. E. Ross, M.-A. Meier, and E. Hauksson, "P wave arrival picking and first-motion polarity determination with deep learning," *Journal of Geophysical Research: Solid Earth*, vol. 123, pp. 5120–5129, 2018.

[47] C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," *Sci. Advances*, vol. 1, no. 11, p. e1501057, 2015.

[48] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *Sci. Advances*, vol. 4, no. 2, p. e1700578, 2018.

[49] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, "Machine learning predicts laboratory earthquakes," *Geophys. Res. Letters*, vol. 44, no. 18, pp. 9276–9282, 2017.

[50] Q. Kong, R. M. Allen, L. Schreier, and Y. W. Kwon, "MyShake: A smartphone seismic network for earthquake early warning and beyond," *Sci. Advances*, vol. 2, no. 2, p. e1501055, 2016.

[51] S. Gupta, K. Kothari, and M. V. de Hoop, "Deep mesh projectors for inverse problems," *ArXiv e-prints*, May 2018, https://arxiv.org/abs/1805.11718.

[52] M. Araya-Polo, J. Jennings, A. Adler, and T. Dahlke, "Deep-learning tomography," *The Leading Edge*, vol. 37, no. 1, pp. 58–66, 2018.

[53] M. P. Barmin, M. H. Ritzwoller, and A. L. Levshin, "A fast and reliable method for surface wave tomography," *Pure and Appl. Geo.*, vol. 158, no. 8, pp. 1351–1375, 2001.

[54] F. C. Lin, D. Li, R. W. Clayton, and D. Hollis, "High-resolution 3D shallow crustal structure in long beach, california: Application of ambient noise tomography on a dense seismic array," *Geophysics*, vol. 78, no. 4, pp. Q45–Q56, 2013.

[55] N. Carbone and W. S. Hodgkiss, "Effects of tidally driven temperature fluctuations on shallow-water acoustic communications at 18khz," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 84–94, 2000.

[56] W. S. Hodgkiss, W. A. Kuperman, and D. E. Ensberg, "Channel impulse response fluctuations at 6 kHz in shallow water," in *Impact of Littoral Environmental Variability of Acoustic Predictions and Sonar Performance*. Springer, 2002.

[57] C. T. Liu, R. Pinkel, M. K. Hsu, J. M. Klymak, H. W. Chen, and C. Villanoy, "Nonlinear internal waves from the luzon strait," *Eos Trans. AGU*, vol. 87, no. 42, pp. 449–451, 2006.

[58] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[59] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[60] P. Gerstoft, C. F. Mecklenbräuker, W. Seong, and M. Bianco, "Introduction to compressive sensing in acoustics," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3731–3736, 2018.

[61] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Series B*, pp. 267–288, 1996.

[62] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[63] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *ACM Proc. 26th Inter. Conf. Mach. Learn.*, pp. 689–696, 2009.

[64] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 4, pp. 791–804, 2012.

[65] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: the application of compressed sensing for rapid MR imaging," *Mag. Res. Medic.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[66] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multi-frequency sparse Bayesian learning for robust matched field processing," *J. Acoust. Soc. Am.*, vol. 141, no. 5, pp. 3411–3420, 2017.

[67] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Intersci., 2001.

[68] P. Gerstoft, A. Xenaki, and C. F. Mecklenbräuker, "Multiple and single snapshot compressive beamforming," *J. Acoust. Soc. Am.*, vol. 138, no. 4, pp. 2003–2014, 2015.

[69] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *27th. Annual Asilomar Conference on Signals, Systems and Computers, IEEE Proc.*, pp. 40–44, 1993.

[70] S. Mallat, *A wavelet tour of signal processing*, 2nd ed. Elsevier, San Diego, CA, 1999.

[71] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. on Info. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[72] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by v1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[73] E. J. Candes, Y. C. Eldar, D. Needell, and R. Paige, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmonic Anal.*, vol. 31, no. 1, pp. 59–73, 2011.

[74] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Sci. Bus. Media, 2009.

[75] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Trans. Cons. Elec.*, vol. 46, pp. 1103–1127, 2000.

[76] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Kluwer Academic, Norwell, MA, 1991.

[77] K. Engan, S. O. Aase, and J. H. H. y, "Multi-frame compression: theory and design," *Signal Processing*, vol. 80, pp. 2121–2140, 2000.

[78] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003.

[79] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, pp. 4311–4322, 2006.

[80] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[81] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.

[82] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd," *Appl. and Comp. Harm. Anal.*, vol. 37, no. 3, pp. 464–491, 2014.

[83] ——, "Local identification of overcomplete dictionaries," *J. Mach. Learn. Res.*, no. 16, pp. 1211–1242, 2015.

[84] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3180–3193, 2016.

[85] N. M. Shapiro, M. Campillo, L. Stehly, and M. H. Ritzwoller, "High-resolution surface-wave tomography from ambient seismic noise," *Science*, vol. 307, no. 5715, pp. 1615–1618, 2005.

[86] G. D. Bensen, M. H. Ritzwoller, M. P. Barmin, A. L. Levshin, F. C. Lin, M. Moschetti, N. M. Shapiro, and Y. Yang, "Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements," *Geo. J. Int.*, vol. 169, no. 3, pp. 1239–1260, 2007.

[87] M. B. Porter, "The kraken normal mode program," Naval Research Laboratory, Washington D.C. No. NRL/MR/5120-92-6920, Tech. Rep., 1992.

Chapter 2

Compressive acoustic sound speed profile estimation

2.1 Introduction

Inversion of ocean parameters using acoustic propagation models requires simultaneous optimization of water column sound speed profile (SSP) and sediment properties using limited observations. Such inverse problems are ill-posed, and require regularization to ensure physically realistic solution [1]. The water column SSP estimation problem has been regularized traditionally by minimizing the energy of the solution to a least-squares cost function, which requires undersampling of complex SSP structure or explaining the structure using few shape functions [2–4]. This reduction in resolution causes SSP uncertainty, especially when internal waves or currents generate strong, temporally varying SSP anomalies [5–7]. This uncertainty can severely effect the accuracy of inversion for other parameters [5, 7].

We show that SSPs in range-independent shallow ocean environments can be resolved using compressive sensing (CS) [8, 9]. CS asserts that parameters can be recovered robustly for certain highly underdetermined linear problems via sparse regularization of a least-squares cost

function, provided that the solutions are sparse, i.e. few non-zero parameters (out of many) explain the observations. Recent applications of CS in ocean acoustics have demonstrated performance improvements to coherent passive fathometry [10] and beamforming [11–13] under sparsity assumptions. Here, the inversion for ocean acoustic SSPs is formulated as an underdetermined linear problem where SSP perturbations are parameterized in a sparse domain using shape functions.

Pressure observations from a vertical line array (VLA) of hydrophones in a shallow ocean are forward-modeled using normal modes. The non-linear response of the forward model to SSP is linearized using a first order Taylor expansion. The linearized sensitivities of few observations are calculated for many shape functions, which parameterize SSPs in a sparse domain. Thus, SSP estimation is expressed as an underdetermined linear problem and solved using convex optimization of a least squares cost function with an ℓ_1 sparsity constraint [8, 9].

2.2 Compressive estimation of water column SSP

A K-point discretized ocean SSP, $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^K$, is modeled as

$$\mathbf{c}(\mathbf{x}) = \mathbf{c}_0 + \mathbf{Q}\mathbf{x}, \tag{2.1}$$

where $\mathbf{c}_0 \in \mathbb{R}^K$ is a reference SSP (e.g. derived from historical statistics or CTD casts), $\mathbf{Q} = [\mathbf{q}_1, ...\mathbf{q}_N] \in \mathbb{R}^{K\times N}$ is a dictionary of N unique shape functions $\mathbf{q}_n$ discretized into K points, and $\mathbf{x} \in \mathbb{R}^N$ is the coefficients. The SSP is modeled by shape functions [1] that describe SSP perturbations with few non-zero coefficients in $\mathbf{x}$. SSP perturbations that are isolated to specific depths of the water column are here modeled using half-sinusoidal shape functions. Larger SSP variation is modeled using EOFs [4, 6, 7].

The pressure $\mathbf{p}_{obs} \in \mathbb{C}^M$ received at an M element VLA is modeled as

$$\mathbf{p}_{obs} = \mathbf{g}(\mathbf{x}) + \mathbf{n}, \tag{2.2}$$

where $\mathbf{g}(\mathbf{x}) \in \mathbb{C}^M$ is the normal mode propagation given $\mathbf{c}(\mathbf{x})$ and $\mathbf{n} \in \mathbb{C}^M$ is Gaussian white noise. Here, the source is known and included in the $\mathbf{g}(\mathbf{x})$ term.

Assuming the perturbations to the reference SSP are small, the non-linear response of $\mathbf{g}(\mathbf{x})$ to SSP perturbations is linearized using the first order Taylor expansion

$$\mathbf{g}(\mathbf{x}) \approx \mathbf{g}(\mathbf{0}) + \left.\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{0}} \mathbf{x} = \mathbf{g}(\mathbf{0}) + \mathbf{D}\mathbf{x}. \tag{2.3}$$

The matrix $\mathbf{D} = [\mathbf{d}_1, ...\mathbf{d}_N] \in \mathbb{C}^{M \times N}$ contains the derivatives of the M pressure observations relative to the N shape functions in $\mathbf{Q}$. The columns $\mathbf{d}_n$ are calculated using two-sided finite differences, by perturbing the reference profile by a fraction of $\mathbf{q}_n$.

Provided the columns of $\mathbf{D}$ are sufficiently incoherent, a sparse estimate of $\mathbf{x}$ is found using ℓ_1-norm convex optimization [8, 9]. The sparse solution to Eq. (2) is formulated as

$$\widehat{\mathbf{x}}_{\ell_1}(\mu) = \underset{\mathbf{x} \in \mathbb{R}^N}{\arg\min} \| \frac{\mathbf{g}(\mathbf{0}) + \mathbf{D}\mathbf{x}}{|\mathbf{p}_{obs}|} - \widetilde{\mathbf{p}}_{obs} \|_2^2 + \mu \|\mathbf{x}\|_1, \tag{2.4}$$

where (assuming a signal is present) Eq. (2) is scaled by $|\mathbf{p}_{obs}|$ as

$$\widetilde{\mathbf{p}}_{obs} = \frac{\mathbf{p}_{obs}}{|\mathbf{p}_{obs}|} = \frac{\mathbf{g}(\mathbf{0}) + \mathbf{D}\mathbf{x} + \mathbf{n}}{|\mathbf{p}_{obs}|}, \tag{2.5}$$

$\widehat{\mathbf{x}}_{\ell_1}$ is the sparse estimate of the SSP perturbation coefficients, and μ is the regularization parameter which controls the relative importance of sparsity (ℓ_1-norm regularization) and measurement fit (ℓ_2-norm). Eq. (2) is scaled to improve numerical stability, by reducing the difference in magnitude between the ℓ_1-norm of $\mathbf{x}$ and the ℓ_2 cost function.

The sparse solution $\widehat{\mathbf{x}}_{\ell_1}$ selects the coefficients in $\mathbf{x}$ that best explain the observations, but their values are biased [14]. The coefficients are optimized using least-squares criteria by solving the overdetermined problem [12, 13]

$$\widehat{\mathbf{x}}_{CS} = \mathbf{D}_A^{+}[\mathbf{p}_{obs} - \mathbf{g}(\mathbf{0})], \tag{2.6}$$

where $\mathbf{D}_A$ contains only the active columns of $\mathbf{D}$, corresponding to non-zero elements in $\widehat{\mathbf{x}}_{\ell_1}$, and $\mathbf{D}_A^{+}$ is its Moore-Penrose pseudoinverse. Here, $\widehat{\mathbf{x}}_{CS}$ is the optimal compressive sensing solution to Eq. (2).

The non-sparse (minimum energy) estimate $\widehat{\mathbf{x}}_{\ell_2}$ of $\mathbf{x}$ is written as

$$\widehat{\mathbf{x}}_{\ell_2}(\mu) = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min} \| \frac{\mathbf{g}(\mathbf{0}) + \mathbf{D}\mathbf{x}}{|\mathbf{p}_{obs}|} - \widetilde{\mathbf{p}}_{obs} \|_2^2 + \lambda \|\mathbf{x}\|_2^2, \tag{2.7}$$

where λ controls the relative importance of solution energy (ℓ_2-norm regularization) and measurement fit (ℓ_2-norm).

2.3 Simulation and results

The acoustic field in a 160 m constant-depth ocean was simulated using the Kraken normal mode model [15]. The field, generated by an 100 Hz acoustic source at 30 m depth, was sampled at 2 km range by $M = 10$ evenly spaced VLA elements spanning 10 to 150 m depth. The reference SSP was the mean SSP from the SWellEx-96 experiment. The bottom sound speed was 1800 m/s, density 2.0 g/cm^3, and attenuation 0.1 dB/λ. A diagram of the environment is shown in Fig. 1(a).

The sparse problems were solved using the CVX toolbox, which specifies and solves convex optimization problems [16]. Sparsity of $\widehat{\mathbf{x}}_{\ell_1}$, Eq. (4), was enhanced using reweighted ℓ_1-norm minimization method; for details see Refs. [12] and [17]. For comparison, the minimum

energy estimate $\widehat{\mathbf{x}}_{\ell_2}$ of the parameters, Eq. (7), was solved with $\lambda = 5 \times 10^{-5}$.

2.3.1 Compressive estimation of half-sinusoidal SSP perturbations

To simulate SSP anomalies confined to narrow depth ranges, two half-sinusoidal perturbations [Fig. 1(b)] were added to the reference SSP at depths of 63.3 and 126.7 m with magnitudes of ± 2 m/s and 20 m widths. Eq. (3) was formulated with $\mathbf{Q}$ containing $N = 100$ half-sinusoid shape functions having the same widths as the perturbations.

Fig. 2(a)–(b) shows the sparse solution $\widehat{\mathbf{x}}_{CS}$, Eq. (6) after solving (4)–(5), for the half-sinusoidal perturbations with 30 dB SNR for $\mu = 5 \times 10^{-4}$. It can be seen that for one realization of Gaussian white noise, the sparse magnitudes and locations of the perturbations were estimated accurately. With 10 observations and 100 potential shape functions, this problem was underdetermined by a factor of 10 and CS still gave an accurate result. The minimum energy solution $\widehat{\mathbf{x}}_{\ell_2}$, Eq. (7), shown in Fig. 2(c)–(d), did not provide the true parameters. Instead the solution was non-sparse, having many small perturbations.

2.3.2 Compressive estimation of SSP using EOFs

A set of EOFs was calculated using 26 CTD casts spanning depths $\geq$160m from the SWellEx-96 experiment, resulting in a dictionary $\mathbf{Q}$ with $N = 26$ EOFs. The first three EOFs are shown in Fig. 1(c). Synthetic SSPs and SWellEx-96 profiles were estimated using CS.

The robustness of the CS EOF inversion was tested by finding sparse estimates $\widehat{\mathbf{x}}_{CS}$ for 1000 synthetic SSPs, having randomly selected active EOF components with coefficients of ± 6 m/s. The active EOF components were selected randomly from a half-Gaussian distribution, with the peak located at the first EOF, to simulate the relative importance of the EOFs. Each of the synthetic SSPs were compressively inverted without observation noise with $\mu = 5 \times 10^{-5}$ (corresponding to 3 sparse EOF coefficients for most cases). The SSP estimation error (standard

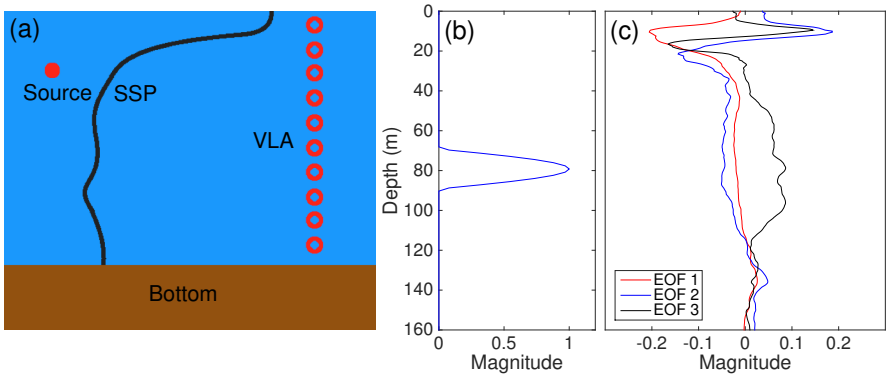


Figure 2.1: (Color online) (a) Ocean environment and measurement configuration. (b) Half-sinusoid shape function. (c) First 3 EOFs derived from the SWellEx-96 data.

deviation, STD) versus depth in Fig. 3(a) shows that SSP error increases toward shallow depths. This is likely due to downward refraction of the acoustic waves by the warmer surface water, which makes the inversion insensitive to near-surface variability. For deeper SSP variability, 1 STD of error is within ± 0.2 m/s. In Fig. 3(b)–(c) a CS EOF inversion result is shown for one random trial. Considering a SSP from the SWellEx-96 experiment, Fig. 3(d)–(e) shows CS inversion of 3 EOF components with $\mu = 1.5 \times 10^{-4}$.

Fig. 4(a)–(b) shows the CS estimation of the 3 EOF SSP used in Fig. 3 with a 30 dB SNR for $\mu = 2 \times 10^{-4}$ (μ increases due to noise). With one realization of Gaussian white noise, the EOF components are estimated. As shown in Fig. 4(c)–(d) the $\widehat{\mathbf{x}}_{\ell_2}$ solution was non-sparse and provided inaccurate estimates of the true (sparse) parameters.

2.4 Conclusion

A method for compressive inversion of ocean acoustic SSPs was developed and demonstrated. With medium SNR, *a priori* knowledge of the ocean sound speed statistics, and a dictionary of shape functions that sparsely represent the SSPs, fine-scale SSP structure is well estimated using CS. Robust recovery of sparse SSP perturbations was shown using dictionaries containing either half-sinusoidal shape functions or EOFs.

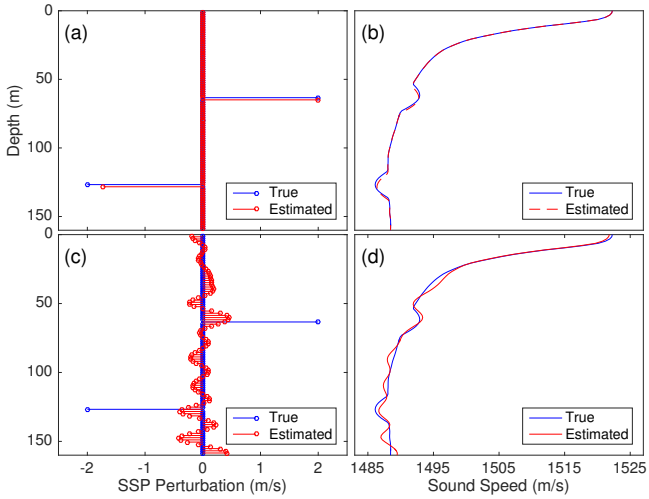


Figure 2.2: (Color online) Sparse $\widehat{\mathbf{x}}_{CS}$ [(a) and (b)] and minimum energy $\widehat{\mathbf{x}}_{\ell_2}$ [(c) and (d)] estimates for half-sinusoidal ocean SSP perturbations from noisy observations (SNR=30 dB). [(a) and (c)] Depth and magnitude of perturbations, and [(b) and (d)] the corresponding SSPs.

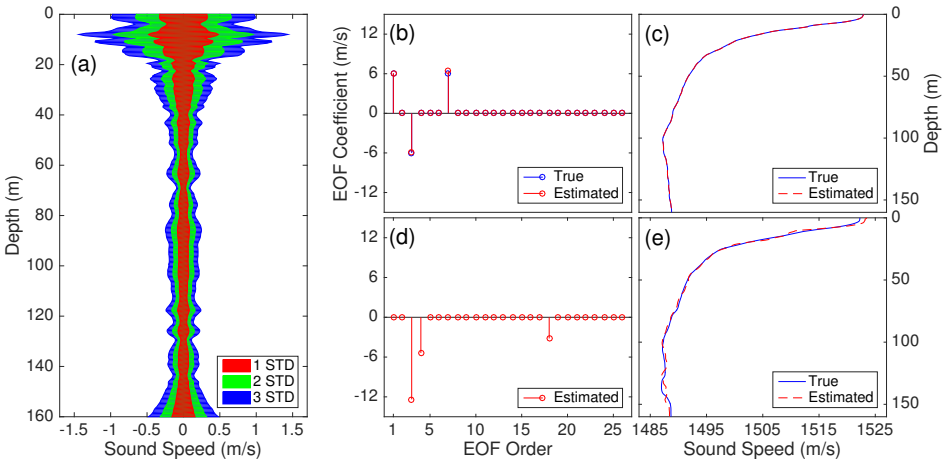


Figure 2.3: (Color online) (a) SSP estimation error versus depth of the sparse solution $\widehat{\mathbf{x}}_{CS}$ (without noise) for 1000 randomly generated synthetic SSPs with 3 EOF components. For one randomly generated SSP, $\widehat{\mathbf{x}}_{CS}$ estimate (b) of the EOF coefficients and (c) the corresponding SSP. For one SWellEx-96 SSP, $\widehat{\mathbf{x}}_{CS}$ estimate for (d) EOF coefficients and (e) corresponding SSP.

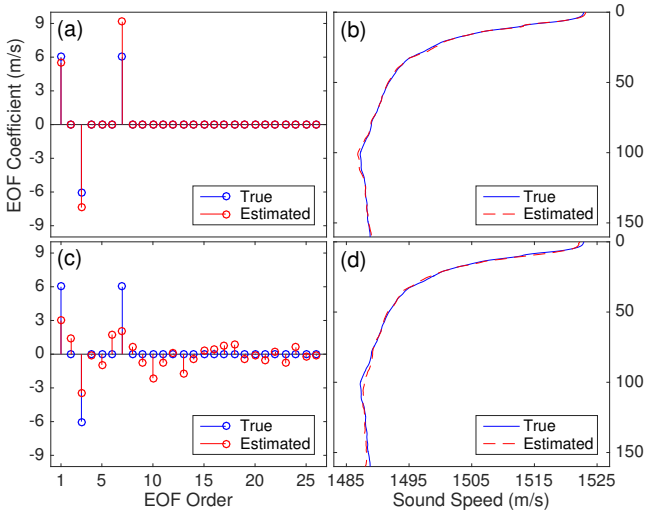


Figure 2.4: (Color online) Sparse $\widehat{\mathbf{x}}_{CS}$ [(a) and (b)] and minimum energy $\widehat{\mathbf{x}}_{\ell_2}$ [(c) and (d)] estimates of 3 SSP EOF coefficients from noisy observations (SNR=30 dB). [(a) and (c)] EOF coefficients and [(b) and (d)] the corresponding SSPs.

2.5 Acknowledgments

This work was supported by the Office of Naval Research, Grant No. N00014-11-1-0439.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of the Acoustical Society of America 2016. Bianco, Michael; Gerstoft, Peter, Acoustical Society of America, 2016. The dissertation author was the primary investigator and author of this paper.

2.6 Bibliography

[1] P. Gerstoft, "Inversion of acoustic data using a combination of genetic algorithms and the gauss–newton approach," *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2181–2190, 1995.

[2] M. I. Taroudakis and J. S. Papadakis, "A modal inversion scheme for ocean acoustic tomography," *J. Comput. Acoust.*, vol. 1, no. 4, pp. 395–421, 1993.

[3] C. Park, W. Seong, P. Gerstoft, and W. S. Hodgkiss, "Geoacoustic inversion using backpropagation," *IEEE Journal of Oceanic Engineering*, vol. 35, no. 4, pp. 722–731, 2010.

[4] B. A. Tan, P. Gerstoft, C. Yardim, and W. S. Hodgkiss, "Broadband synthetic aperture geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 312–322, 2013.

[5] M. Siderius, P. L. Nielsen, J. Sellschopp, M. Snellen, and D. Simons, "Experimental study of geo-acoustic inversion uncertainty due to ocean sound-speed fluctuations," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 769–781, 2001.

[6] Y. T. Lin, C. F. Chen, and J. F. Lynch, "An equivalent transform method for evaluating the effect of water-column mismatch on geoacoustic inversion," *IEEE Journal of Oceanic Engineering*, vol. 31, no. 2, pp. 284–298, 2006.

[7] C.-F. Huang, P. Gerstoft, and W. S. Hodgkiss, "Effect of ocean sound speed uncertainty on matched-field geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. EL162–EL168, 2008.

[8] E. Candés, "Compressive sampling," *Proc. Internat. Cong. Math.*, vol. 3, pp. 1433–1452, 2006.

[9] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[10] C. Yardim, P. Gerstoft, W. S. Hodgkiss, and J. Traer, "Compressive geoacoustic inversion using ambient noise," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1245–1255, 2014.

[11] G. F. Edelmann and C. F. Gaumond, "Beamforming using compressive sensing," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. EL232–EL237, 2011.

[12] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 260–271, 2014.

[13] P. Gerstoft, A. Xenaki, and C. F. Mecklenbräuker, "Multiple and single snapshot compressive beamforming," *J. Acoust. Soc. Am.*, vol. 138, no. 4, pp. 2003–2014, 2015.

[14] M. A. Figueiredo, R. D., Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, 2007.

[15] M. B. Porter, "The kraken normal mode program," Naval Research Laboratory, Washington D.C. No. NRL/MR/5120-92-6920, Tech. Rep., 1992.

[16] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming, version 2.1," 2008.

[17] E. J. Candés, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. App.*, vol. 14, no. 5-6, pp. 877–905, 2008.

Chapter 3

Regularization of geophysical inversion using dictionary learning

3.1 Introduction

Dictionary learning algorithms may provide optimal regularization bases for geophysical inversion. Inversion for geophysical phenomena is typically ill-posed and requires significant regularization to obtain physically plausible solutions and moderate the size of the parameter search [1]. If many representative measurements are available, the dimension of the model typically is reduced using principal component analysis (PCA) [2]. PCA, or empirical orthogonal function (EOF, in the geosciences) analysis, provides a set of orthogonal shape functions which describe the features of largest variance in the data [2, 3]. However, this requirement of orthogonality may limit the regularization effectiveness.

Many signals, including natural images [4, 5], audio [6], and seismic profiles [7] are well approximated using sparsifying dictionaries. Given a signal, a dictionary is defined here as a set of ℓ_2-normalized vectors which describe the signal using few coefficients. The sparse processor is then an ℓ_2-norm cost function with an ℓ_0-norm penalty on the number of non-zero

coefficients [8]. Signal sparsity is exploited for a number of purposes including signal compression and denoising [9]. Here, signal sparsity is exploited for inverse problem regularization of dynamic geophysical phenomena.

Dictionaries of vectors that approximate a given class of signals using few coefficients can be designed with dictionary learning. A popular dictionary learning approach, the K-SVD algorithm [9], based on vector quantization (VQ) [6], finds a dictionary of vectors that optimally partition the data from the training set such that the few dictionary vectors describe each data example. Relative to EOFs, the learned dictionary vectors are not constrained to be orthogonal and potentially provide more optimal signal compression because the vectors are on average, nearer to the signal examples [10]. Dictionary learning has been applied in the geophysics community to improve denoising results in seismics [11] and ocean acoustics [12, 13].

In this paper, dictionaries describing 1D ocean sound speed profile (SSP) data are generated using the K-SVD algorithm and the reconstruction performance is evaluated against EOF methods. In Section 2, the EOF methods and sparse processing are briefly introduced. In Section 3, dictionary learning and the K-SVD algorithm are described. In Section 4, SSP reconstruction results are given using the two methods. It is shown that each vector in the learned dictionaries explains more SSP variability than the leading order EOFs trained on the same data. Further, it is demonstrated that SSPs can be reconstructed up to acceptable error using as few as one non-zero coefficient. This compression can improve the resolution of ocean SSP estimates with negligible computational burden. More details of the approach and further experimental results are available in a forthcoming paper [14].

3.2 EOF and sparse methods

EOF analysis seeks to reduce representation complexity of continuously sampled space-time fields by finding spatial patterns which explain much of the variance of the process. These

spatial patterns or EOFs correspond to the principal components, from principal component analysis (PCA), of the temporally varying field [3]. Here, the field is a collection of ocean SSP anomaly vectors $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \mathbb{R}^{K \times M}$ which are sampled over K discrete points in depth and M instants in time. They are defined as

$$\mathbf{y}_m = \mathbf{c}_m - \bar{\mathbf{c}} \tag{3.1}$$

where $\mathbf{c}_m \in \mathbb{R}^K$ is an observed K-point SSP and $\bar{\mathbf{c}} \in \mathbb{R}^K$ is the mean SSP over M observations.

The singular value decomposition (SVD) [15] finds the EOFs as the eigenvectors of $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ by

$$\mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \mathbf{P}\mathbf{\Lambda}^2\mathbf{P}^{\mathrm{T}}, \tag{3.2}$$

where $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_L] \in \mathbb{R}^{K \times L}$ are L EOFs (eigenvectors, $L = \min(\mathrm{M}, \mathrm{K})$) and $\mathbf{\Lambda}^2 = \mathrm{diag}([\lambda_1^2, ..., \lambda_{\mathrm{L}}^2]) \in \mathbb{R}^{\mathrm{L} \times \mathrm{L}}$ are the total variances of the data along the principal directions defined by the EOFs $\mathbf{p}_l$, with $\lambda_1^2 \geq ... \geq \lambda_L^2$.

Since the leading-order EOFs often explain much of the variance in $\mathbf{Y}$, the representation of $\mathbf{Y}$ can be compressed by retaining only the leading order EOFs $P < K$ for reconstruction of $\mathbf{y}_m$. Each of the SSP anomaly vectors $\mathbf{y}_m$ are approximated as $\mathbf{y}_m = \sum_{p=1}^{P} \mu_p \mathbf{p}_p$, where the EOF coefficients μ_p are solved as $\mu_p = \mathbf{p}_p^{\mathrm{T}} \mathbf{y}_m$. For ocean SSPs, usually no more than 5 EOF coefficients have been used to reconstruct ocean SSPs [16, 17].

A signal $\mathbf{y}_m$, whose model is sparse in the dictionary $\mathbf{Q} = [\mathbf{q}_1, ..., \mathbf{q}_N] \in \mathbb{R}^{K \times N}$, is reconstructed to an acceptable error using few vectors $\mathbf{q}_n$ [8]. The inversion for these sparse coefficients is phrased as an ℓ_2-norm minimization problem with an ℓ_0-norm penalization on the number of non-zero coefficients

$$\widehat{\mathbf{x}}_m = \underset{\mathbf{x}_m \in \mathbb{R}^N}{\arg\min} \|\mathbf{y}_m - \mathbf{Q}\mathbf{x}_m\|_2 \ \text{ subject to } \ \|\mathbf{x}_m\|_0 \leq T, \tag{3.3}$$

where $\mathbf{x}_m \in \mathbb{R}^P$ is the vector of coefficients for $\mathbf{Q}$, $\widehat{\mathbf{x}}_m$ is the sparse estimate of $\mathbf{x}_m$, and T is the number of non-zero coefficients in the solution ($T \ll K$). The sparse reconstruction $\widehat{\mathbf{y}}_m \in \mathbb{R}^K$ of the signal $\mathbf{y}_m$ is then $\widehat{\mathbf{y}}_m = \mathbf{Q}\widehat{\mathbf{x}}_m$. The ℓ_0-norm constraint is non-convex and imposes combinatorial search for the exact solution to (4.6). Here, orthogonal matching pursuit (OMP) [18] is used as the sparse solver.

3.3 Dictionary learning with K-SVD

The K-SVD algorithm [9] is inspired by the iterative K-means for VQ codebook design [6]. The N columns of the dictionary $\mathbf{Q}$, like the entries in VQ codebooks, correspond to partitions in $\mathbb{R}^K$. However, they are constrained to have unit ℓ_2-norm and thus separate the magnitude (coefficients $\mathbf{x}_n$) from the shapes (dictionary entries $\mathbf{q}_n$) for the sparse processing objective (4.6). When $T = 1$, the ℓ_2-norm objective in (4.6) is minimized by the dictionary entry $\mathbf{q}_n$ that has the greatest inner product with example $\mathbf{y}_m$ [8]. Thus for $T = 1$, $[\mathbf{q}_1, ..., \mathbf{q}_N]$ define radial partitions of $\mathbb{R}^K$. This corresponds to a special case of VQ, called gain-shape VQ [6]. For $T = 1$, the sequential updates of the K-SVD provide optimal dictionary updates for gain-shape VQ [6, 9]. Optimal updates to the gain-shape dictionary will, like K-means updates, either improve or leave unchanged the MSE and convergence to a local minimum is guaranteed. For $T > 1$, convergence of the K-SVD updates to a local minimum depends on the accuracy of the sparse-solver used in the sparse coding stage [9].

The dictionary learning objective is

$$\min_{\mathbf{Q}} \Big\{ \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{Q}\mathbf{X}\|_{\mathcal{F}}^2 \ \text{ subject to } \ \forall_m, \|\mathbf{x}_m\|_0 \leq T \Big\}, \tag{3.4}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M]$ is the matrix of coefficient vectors corresponding to examples $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M]$, and $\mathcal{F}$ is the Frobenius norm. The K-means algorithm is generalized to the dictionary learning problem as the two steps: 1) sparse coding and 2) dictionary update.

In the K-SVD algorithm, each iteration i sequentially improves both the entries $\mathbf{q}_n \in \mathbf{Q}^i$ and the coefficients in $\mathbf{x}_m \in \mathbf{X}^i$, without change in support. Expressing the coefficients as row vectors $\mathbf{x}_T^n \in \mathbb{R}^N$ and $\mathbf{x}_T^j \in \mathbb{R}^N$, which relate all examples $\mathbf{Y}$ to $\mathbf{q}_n$ and $\mathbf{q}_j$, respectively, the ℓ_2-penalty from (4.7) is rewritten as

$$\|\mathbf{Y} - \mathbf{QX}\|_{\mathcal{F}}^2 = \left\|\mathbf{Y} - \sum_{n=1}^{N} \mathbf{q}_n \mathbf{x}_T^n\right\|_{\mathcal{F}}^2 = \|\mathbf{E}_j - \mathbf{q}_j \mathbf{x}_T^j\|_{\mathcal{F}}^2, \tag{3.5}$$

where

$$\mathbf{E}_j = \left(\mathbf{Y} - \sum_{n \neq j} \mathbf{q}_n \mathbf{x}_T^n\right). \tag{3.6}$$

Thus, in (4.15) the ℓ_2-penalty is separated into an error term $\mathbf{E}_j = [\mathbf{e}_{j,1}, ..., \mathbf{e}_{j,M}] \in \mathbb{R}^{K \times M}$, which is the error for all examples $\mathbf{Y}$ if $\mathbf{q}_j$ is excluded from their reconstruction, and the product of the excluded entry $\mathbf{q}_j$ and coefficients $\mathbf{x}_T^j \in \mathbb{R}^N$.

An update to the dictionary entry $\mathbf{q}_j$ and coefficients $\mathbf{x}_T^j$ which minimizes (4.15) is found by taking the SVD of $\mathbf{E}_j$, which provides the best rank-1 approximation of $\mathbf{E}_j$. However, many of the entries in $\mathbf{x}_T^j$ are zero (examples which don't use $\mathbf{q}_j$). To update $\mathbf{q}_j$ and $\mathbf{x}_T^j$ with SVD, (4.15) must be restricted to examples $\mathbf{y}_m$ which use $\mathbf{q}_j$

$$\|\mathbf{E}_j^R - \mathbf{q}_j \mathbf{x}_R^j\|_{\mathcal{F}}^2, \tag{3.7}$$

where $\mathbf{E}_j^R$ and $\mathbf{x}_R^j$ are entries in $\mathbf{E}_j$ and $\mathbf{x}_T^j$, respectively, corresponding to examples $\mathbf{y}_m$ which use $\mathbf{q}_j$, and are defined as

$$\mathbf{E}_j^R = \left\{\mathbf{e}_{j,l} \middle| \forall_l,\, x_l^j \neq 0\right\}, \quad \mathbf{x}_R^j = \left\{x_l^j \middle| \,\forall_l,\, x_l^j \neq 0\right\}. \tag{3.8}$$

The K-SVD algorithm is given in Table 4.2.

Table 3.1: The K-SVD Algorithm [9]

Given: $\mathbf{Y} \in \mathbb{R}^{K\times M}$, $\mathbf{Q}^0 \in \mathbb{R}^{K\times N}$, $T \in \mathbb{N}$, and $i = 0$
Repeat until convergence:
1. Sparse coding
for $m = 1 : M$
a: solve (4.6) for $\widehat{\mathbf{x}}_m$ using OMP
b: $\mathbf{X} = [\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_M]$
2. Dictionary update
for $j = 1 : N$
a: compute reconstruction error $\mathbf{E}_j$ from (4.16)
b: obtain $\mathbf{E}_j^R$, $\mathbf{x}_R^j$ corresponding to nonzero $\mathbf{x}_T^j$
c: apply SVD to $\mathbf{E}_j^R$: $\mathbf{E}_j^R = \mathbf{USV}^{\mathrm{T}}$
d: update $\mathbf{q}_j^i = \mathbf{U}(:,1)$, $\mathbf{x}_R^j = \mathbf{V}(:,1)\mathbf{S}(1,1)$
e: $\mathbf{Q}^{i+1} = \mathbf{Q}^i$
$i = i+1$

3.4 Example

We here apply dictionary learning to ocean SSP data from the HF-97 acoustics experiment [19, 20], conducted off the coast of Point Loma, CA. The reconstruction results are compared with EOF methods. $M = 1000$ (15 point) profiles were used for the training set. The SSPs were interpolated to $K = 30$ points using a shape-preserving cubic spline. EOFs were calculated from (4.2) and learned dictionaries were generated with the K-SVD algorithm (Table 4.2). The number of non-zero coefficients solved with OMP for each dictionary was held fixed at exactly T non-zero coefficients. The initial dictionary $\mathbf{Q}^0$ was populated using randomly selected examples from the training sets $\mathbf{Y}$.

The HF-97 learned dictionary, with $N = K$ and $T = 1$, is compared to the EOFs ($K = 30$) in Fig. 4.4. Only the leading order EOFs (Fig. 4.4(a)) are informative of ocean SSP variability whereas all shape functions in the dictionary (Fig. 4.4(b)) are informative (Fig. 4.4(c)–(d)). By relaxing the requirement of orthogonality for the shape functions, the shape functions are adapted to the data distribution and thereby achieve greater compression. The Gram matrix $\mathbf{G}$, which

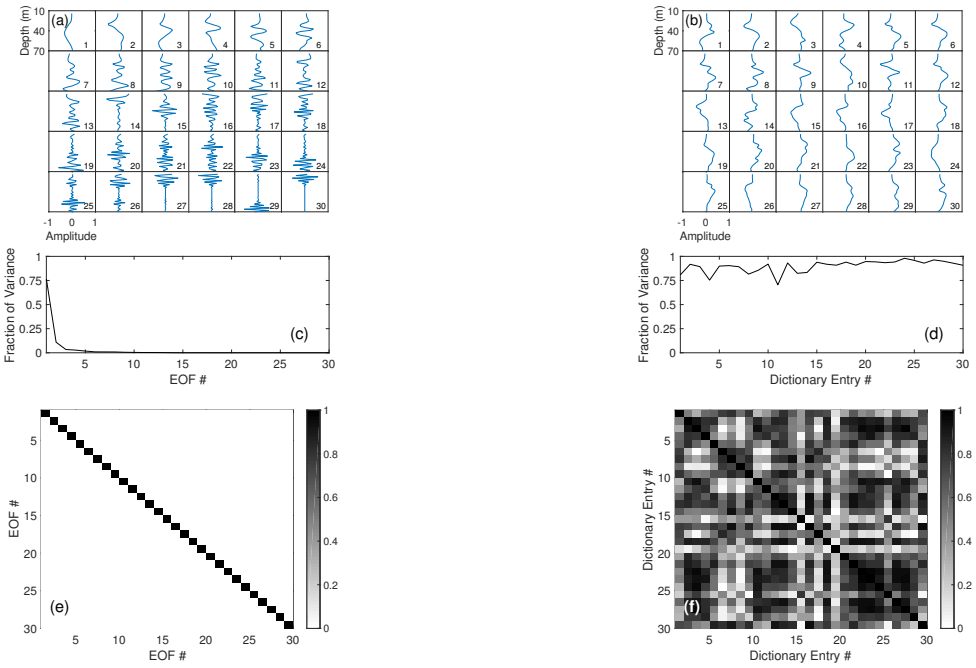


Figure 3.1: (a) EOFs and (b) learned dictionary entries ($N = K$ and $T = 1$, sorted by variance $\sigma^2_{\mathbf{q}_n}$). Fraction of (c) total SSP variance explained by EOFs and (d) SSP variance explained for examples using dictionary entries. Coherence of (e) EOFs and (f) dictionary entries.

gives the coherence of matrix columns, is defined for a matrix $\mathbf{A}$ with unit ℓ_2-norm columns as $\mathbf{G} = |\mathbf{A}^{\mathrm{T}}\mathbf{A}|$. Fig. 4.4(e) shows the shapes in the EOF dictionary are orthogonal, whereas those of the learned dictionary (Fig. 4.4(f)) are not.

3.4.1 Reconstruction of SSP training data

Reconstruction performance of the EOFs and learned dictionaries are evaluated on SSPs within the training set, using a mean error metric. The coefficients for the learned $\mathbf{Q}$ and initial $\mathbf{Q}^0$ dictionaries $\widehat{\mathbf{x}}_m$ are solved from the sparse objective (4.6) using OMP. The least squares (LS) solution for the T leading-order coefficients $\mathbf{x}_L \in \mathbb{R}^T$ from the EOFs $\mathbf{P}$ were solved by

$$\mathbf{x}_L = \mathbf{P}_L^{+}\mathbf{y}_m, \tag{3.9}$$

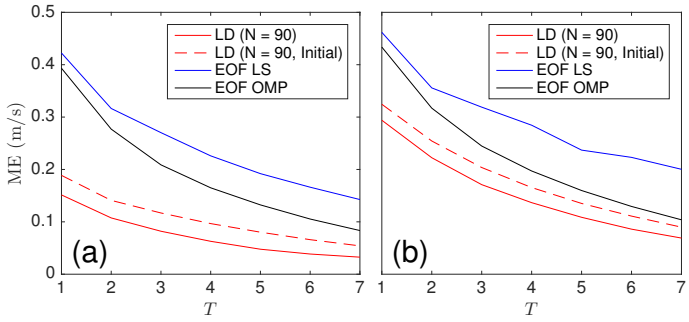


Figure 3.2: (a) Mean reconstruction error (ME) versus T non-zero coefficients for SSPs using EOFs (solved using LS and OMP) and learned dictionary (LD, $N = 90$). (b) Mean reconstruction error ME_{CV} for extra-sample data calculated with K-fold cross validation for $J = 10$ folds.

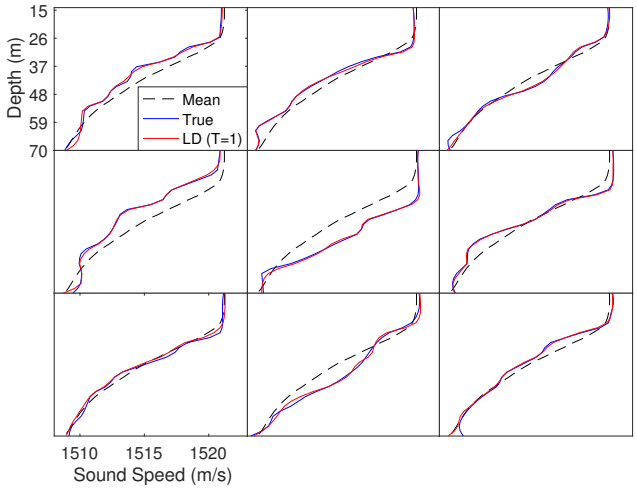


Figure 3.3: True SSP reconstruction of 9 example profiles using one entry ($T = 1$) from learned dictionary (LD, $N = 90$).

where $\mathbf{P}_L$ is the T leading order EOFs from $\mathbf{P}$, and $\mathbf{P}_L^+$ is its pseudoinverse. The best combination of T EOF coefficients was solved from the sparse objective (4.6) using OMP for $\mathbf{Q} = \mathbf{P}$. The mean reconstruction error ME for the training set is

$$\text{ME} = \frac{1}{KM} \|\mathbf{Y} - \widehat{\mathbf{Y}}\|_1. \tag{3.10}$$

The reconstruction error using the EOF dictionary is compared to results from dictionaries $\mathbf{Q}$ with $N = 3K$, using T non-zero coefficients. In Fig. 4.8(a) results are shown for $N = 90$

dictionary entries. Coefficients describing each example $\mathbf{y}_m$, were solved 1) from the learned dictionary $\mathbf{Q}$, 2) from $\mathbf{Q}^0$, the dictionary consisting of N randomly chosen examples from the training set (to illustrate improvements in reconstruction error made in the K-SVD iterations), 3) the leading order EOFs, and 4) the best combination of EOFs. The mean SSP reconstruction error using the dictionaries trained for each sparsity T is less than EOF reconstruction, for either leading order coefficients or best coefficient combination, for all values of T shown. The best combination of EOF coefficients, chosen approximately using OMP, achieves less error than the LS solution to the leading order EOFs, with added cost of search.

Just one learned dictionary entry achieves the same ME as more than 6 leading order EOFs, or greater than 4 EOFs chosen by search (Fig. 4.8(a)). To illustrate the representational power of the learned dictionary entries, both true and reconstructed SSPs are shown in Fig. 4.9. Nine true SSP examples from the training set are reconstructed using one learned dictionary entry. It is shown for each case, that nearly all of the SSP variability is captured using a single entry in $\mathbf{Q}$.

3.4.2 Extra-sample SSP reconstruction

The extra-sample SSP reconstruction performance of learned dictionaries and EOFs is tested using K-fold cross-validation [15]. The SSP data set $\mathbf{Y}$ of M profiles is divided into J subsets with equal numbers of profiles $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_J]$, where the fold $\mathbf{Y}_j \in \mathbb{R}^{K \times (M/J)}$. For each of the J folds: 1) $\mathbf{Y}_j$ is the set extra-sample test cases, and the training set $\mathbf{Y}_{tn}$ is

$$\mathbf{Y}_{tn} = \left\{ \mathbf{Y}_l \middle| \, \forall_{l \neq j} \right\}; \tag{3.11}$$

2) the dictionary $\mathbf{Q}_j$ and EOFs are derived using $\mathbf{Y}_{tn}$; and 3) coefficients estimating test samples $\mathbf{Y}_j$ are solved for $\mathbf{Q}_j$ with sparse processor (4.6), and for EOFs by solving for leading order terms and by solving with sparse processor. The extra-sample error from cross validation $\mathrm{ME}_{\mathrm{CV}}$ for

each method is then

$$\mathrm{ME}_{CV} = \frac{1}{KM} \sum_{j=1}^{J} \|\mathbf{Y}_j - \widehat{\mathbf{Y}}_j\|_1. \tag{3.12}$$

$\mathrm{ME}_{\mathrm{CV}}$ increases over the within-training-set estimates for both the learned and EOF dictionaries, as shown in Fig. 4.8(b) for $J = 10$ folds. The mean reconstruction error using learned dictionaries, as in the within-training-set estimates, is less than the EOF dictionaries. More than 2 EOFs, choosing best combination by search, or more than 3 leading-order EOFs solved with LS, are required to achieve the extra-sample performance as one learned dictionary entry.

3.4.3 Solution space for SSP inversion

Acoustic inversion for ocean SSP is a non-linear problem. One approach is coefficient search using genetic algorithms. [2] Discretizing each coefficient into H values, the number of candidate solutions for T fixed coefficients indices is

$$S_{\text{fixed}} = H^T. \tag{3.13}$$

If the coefficient indices for the solution can vary, as per dictionary learning with learned dictionary $\mathbf{Q} \in \mathbb{R}^{K \times N}$, the number of candidate solutions S_{comb} is

$$S_{\text{comb}} = H^T \frac{N!}{T!(N-T)!}. \tag{3.14}$$

Given the results in the last paragraph of Section 3.4.1, and assuming a typical $H = 100$ point discretization of the coefficients and an unknown SSP similar to the training set, the SSP may be constructed up to acceptable resolution using one entry from the learned dictionary with 10^4 possible solutions. To achieve the similar ME, 6 EOFs are required (10^{12} possible solutions), using fixed indices. The best EOF combination requires 4 EOFs (10^{14} possible solutions).

3.5 Conclusion

Given sufficient training data, dictionary learning generates optimal dictionaries for sparse reconstruction of a given signal class. Since these learned dictionaries are not constrained to be orthogonal, the entries fit the distribution of the data such that signal example is approximated using few dictionary entries. Relative to EOFs, each dictionary entry is informative to the signal variability.

The K-SVD dictionary learning algorithm is applied to ocean SSP data from the HF-97 experiment. The learned dictionaries generated describe ocean SSP variability with high resolution using fewer entries than EOFs. As few as one entry from a learned dictionary describes nearly all the variability in each of the observed ocean SSPs. Provided sufficient SSP training data is available, learned dictionaries can improve SSP inversion resolution. This could provide improvements to geoacoustic inversion [2, 21], matched field processing [22], and underwater communications [19].

3.6 Acknowledgments

This work was supported by the Office of Naval Research, Grant No. N00014-11-1-0439.

Chapter 3, in full, is a reprint of the material as it appears in the Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and Signal Processing 2017. Bianco, Michael; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2017. The dissertation author was the primary investigator and author of this paper.

3.7 Bibliography

[1] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter estimation and inverse problems*, 2nd ed. Elsevier, San Diego, 2013.

[2] P. Gerstoft, "Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 770–782, 1994.

[3] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science: a review," *International Journal of Climatology*, vol. 27, pp. 1119–1152, 2007.

[4] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Sci. Bus. Media, 2009.

[5] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Trans. Cons. Elec.*, vol. 46, pp. 1103–1127, 2000.

[6] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Kluwer Academic, Norwell, MA, 1991.

[7] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1–norm," *Geophys.*, vol. 44, no. 1, pp. 39–52, 1979.

[8] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, pp. 4311–4322, 2006.

[10] K. Engan, S. O. Aase, and J. H. H. y, "Multi-frame compression: theory and design," *Signal Processing*, vol. 80, pp. 2121–2140, 2000.

[11] S. Beckouche and J. Ma, "Simultaneous dictionary learning and denoising for seismic data," *Geophys.*, vol. 79, no. 3, pp. 27–31, 2014.

[12] M. Taroudakis and C. Smaragdakis, "De-noising procedures for inverting underwater acoustic signals in applications of acoustical oceanography," *EuroNoise 2015 Maastricht*, pp. 1393–1398, 2015.

[13] T. Wang and W. Xu, "Sparsity-based approach for ocean acoustic tomography using learned dictionaries," *OCEANS 2016 Shanghai IEEE*, pp. 1–6, 2016.

[14] M. Bianco and P. Gerstoft, "Dictionary learning of sound speed profiles," *J. Acoust. Soc. Am*, vol. 141, no. 3, pp. 1749–1758, 2017.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[16] C.-F. Huang, P. Gerstoft, and W. S. Hodgkiss, "Effect of ocean sound speed uncertainty on matched-field geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. EL162–EL168, 2008.

[17] P. Gerstoft and D. F. Gingras, "Parameter estimation using multifrequency range–dependent acoustic data in shallow water," *J. Acoust. Soc. Am.*, vol. 99, no. 5, pp. 2839–2850, 1996.

[18] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *27th. Annual Asilomar Conference on Signals, Systems and Computers, IEEE Proc.*, pp. 40–44, 1993.

[19] N. Carbone and W. S. Hodgkiss, "Effects of tidally driven temperature fluctuations on shallow-water acoustic communications at 18khz," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 84–94, 2000.

[20] W. S. Hodgkiss, W. A. Kuperman, and D. E. Ensberg, "Channel impulse response fluctuations at 6 kHz in shallow water," in *Impact of Littoral Environmental Variability of Acoustic Predictions and Sonar Performance*. Springer, 2002.

[21] M. Bianco and P. Gerstoft, "Compressive acoustic sound speed profile estimation," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. EL90–EL94, 2016.

[22] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE Journal of Oceanic Engineering*, vol. 18, no. 4, pp. 401–424, 1993.

Chapter 4

Dictionary learning of sound speed profiles

4.1 Introduction

Inversion for ocean sound speed profiles (SSPs) using acoustic data is a non-linear and highly underdetermined problem [1]. To ensure physically realistic solutions while moderating the size of the parameter search, SSP inversion has often been regularized by modeling SSP as the sum of leading order empirical orthogonal functions (EOFs) [2–7]. However, regularization using EOFs often yields low resolution estimates of ocean SSPs, which can be highly variable with fine scale fluctuations. In this paper, it is shown that the resolution of SSP estimates are improved using dictionary learning [8–13], a form of unsupervised machine learning, to generate a dictionary of regularizing shape functions from SSP data for parsimonious representation of SSPs.

Many signals, including natural images [14–16], and seismic profiles [17] are well approximated using sparse (few) coefficients, provided a dictionary of shape functions exist under which their representation is sparse. Given a K-dimensional signal, a dictionary is defined as a set of N, ℓ_2-normalized vectors which describe the signal using few coefficients. The sparse processor is then an ℓ_2-norm cost function with an ℓ_0-norm penalty on the number of non-zero

coefficients. Signal sparsity is exploited for a number of purposes including signal compression and denoising. [9] Applications of compressive sensing, [18] one approximation to the ℓ_0-norm sparse processor, have in ocean acoustics shown improvements in beamforming, [19–22], geoacoustic inversion, [23] and estimation of ocean SSPs [24].

Dictionaries that approximate a given class of signals using few coefficients can be designed using dictionary learning [9]. Dictionaries can be generated ad-hoc from common shape functions such as wavelets or curvelets, however extensive analysis is required to find an optimal set of prescribed shape functions. Dictionary learning proposes a more direct approach: given enough signal examples for a given signal class, learn a dictionary of shape functions that approximate signals within the class using few coefficients. These learned dictionaries (LDs) have improved compression and denoising results for image and video data over ad-hoc dictionaries [9, 11]. Dictionary learning has been applied in the geophysics community to improve denoising results in seismics [25] and ocean acoustics [26, 27], as well as to structural acoustic health monitoring [28].

The K-SVD algorithm [12], a popular dictionary learning method, finds a dictionary of vectors that optimally partition the data from the training set such that the few dictionary vectors describe each data example. Relative to EOFs which are derived using principal component analysis (PCA), [29, 30] these LDs are not constrained to be orthogonal and thus provide potentially better signal compression because the vectors are on average, nearer to the signal examples (see Fig. 4.1) [13].

In this paper, LDs describing 1D ocean SSP data from the HF-97 experiment [31, 32] and from the South China Sea (SCS) [33] are generated using the K-SVD algorithm and the reconstruction performance is evaluated against EOF methods. In Section II, EOFs, sparse reconstruction methods, and compression are introduced. In Section III, the K-SVD dictionary learning algorithm is explained. In Section IV, SSP reconstruction results are given for LDs and EOFs. It is shown that each shape function within the resulting LDs explain more SSP variability

than the leading order EOFs trained on the same data. Further, it is demonstrated that SSPs can be reconstructed up to acceptable error using as few as one non-zero coefficient. This compression can improve the resolution of ocean SSP estimates with negligible computational burden.

Notation: In the following, vectors are represented by bold lower-case letters and matrices by bold uppercase letters. The ℓ_p-norm of the vector $\mathbf{x} \in \mathbb{R}^N$ is defined as $\|\mathbf{x}\|_p = \left(\sum_{n=1}^{N} \left|x_n\right|^p\right)^{1/p}$. Using similar notation, the ℓ_0-norm is defined as $\|\mathbf{x}\|_0 = \sum_{n=1}^{N} \left|x_n\right|^0 = \sum_{n=1}^{N} 1_{|x_n|>0}$. The ℓ_p-norm of the matrix $\mathbf{A} \in \mathbb{R}^{K\times M}$ is defined as $\|\mathbf{A}\|_p = \left(\sum_{m=1}^{M}\sum_{k=1}^{K} \left|a_k^m\right|^p\right)^{1/p}$. The Frobenius norm ($\ell_2$-norm) of the matrix $\mathbf{A}$ is written as $\|\mathbf{A}\|_{\mathcal{F}}$. The hat symbol $\widehat{\ }$ appearing above vectors and matrices indicates approximations to the true signals or coefficients.

4.2 EOFs and compression

4.2.1 EOFs and PCA

Empirical orthogonal function (EOF) analysis seeks to reduce the dimension of continuously sampled space-time fields by finding spatial patterns which explain much of the variance of the process. These spatial patterns or EOFs correspond to the principal components, from principal component analysis (PCA), of the temporally varying field. [29] Here, the field is a collection of zero-mean ocean SSP anomaly vectors $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \mathbb{R}^{K\times M}$, which are sampled over K discrete points in depth and M instants in time. The mean value of the M original observations are subtracted to obtain $\mathbf{Y}$. The variance of the SSP anomaly at each depth sample k, σ_k^2, is defined as

$$\sigma_k^2 = \frac{1}{M}\sum_{m=1}^{M}\left(y_m^k\right)^2 \tag{4.1}$$

where $[y_1^k, ..., y_M^k]$ are the SSP anomaly values at depth sample k for M time samples.

The singular value decomposition (SVD) [34] finds the EOFs as the eigenvectors of $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$

by

$$\mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \mathbf{P}\boldsymbol{\Lambda}^2\mathbf{P}^{\mathrm{T}}, \tag{4.2}$$

where $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_L] \in \mathbb{R}^{K\times L}$ are EOFs (eigenvectors) and $\boldsymbol{\Lambda}^2 = \mathrm{diag}([\lambda_1^2, ..., \lambda_{\mathrm{L}}^2]) \in \mathbb{R}^{\mathrm{L}\times\mathrm{L}}$ are the total variances of the data along the principal directions defined by the EOFs $\mathbf{p}_l$ with

$$\sum_{k=1}^{K} \sigma_k^2 = \frac{1}{M}\mathrm{tr}\big(\Lambda^2\big). \tag{4.3}$$

The EOFs $\mathbf{p}_l$ with $\lambda_1^2 \geq ... \geq \lambda_L^2$ are spatial features of the SSPs which explain the greatest variance of $\mathbf{Y}$. If the number of training vectors $M \geq K$, $L = K$ and $[\mathbf{p}_1, ..., \mathbf{p}_L]$ form a basis in $\mathbb{R}^K$.

4.2.2 SSP reconstruction using EOFs

Since the leading-order EOFs often explain much of the variance in $\mathbf{Y}$, the representation of anomalies $\mathbf{y}_m$ can be compressed by retaining only the leading order EOFs $P < L$

$$\widehat{\mathbf{y}}_m = \mathbf{Q}_P\widehat{\mathbf{x}}_{P,m} \tag{4.4}$$

where $\mathbf{Q}_P \in \mathbb{R}^{K\times P}$ is here the dictionary containing the P leading-order EOFs and $\widehat{\mathbf{x}}_{P,m} \in \mathbb{R}^P$ is the coefficient vector. Since the entries in $\mathbf{Q}_P$ are orthonormal, the coefficients are solved by

$$\widehat{\mathbf{x}}_{P,m} = \mathbf{Q}_P^{\mathrm{T}}\mathbf{y}_m. \tag{4.5}$$

For ocean SSPs, usually no more than $P = 5$ EOF coefficients have been used to reconstruct ocean SSPs [4, 6].

4.2.3 Sparse reconstruction

A signal $\mathbf{y}_m$, whose model is sparse in the dictionary $\mathbf{Q}_N = [\mathbf{q}_1, ..., \mathbf{q}_N] \in \mathbb{R}^{K \times N}$ (N-entry sparsifying dictionary for $\mathbf{Y}$), is reconstructed to acceptable error using $T \ll K$ vectors $\mathbf{q}_n$ [9]. The problem of estimating few coefficients in $\mathbf{x}_m$ for reconstruction of $\mathbf{y}_m$ can be phrased using the canonical sparse processor

$$\widehat{\mathbf{x}}_m = \underset{\mathbf{x}_m \in \mathbb{R}^N}{\arg\min} \|\mathbf{y}_m - \mathbf{Q}\mathbf{x}_m\|_2 \ \text{ subject to } \ \|\mathbf{x}_m\|_0 \leq T. \tag{4.6}$$

The ℓ_0-norm penalizes the number of non-zero coefficients in the solution to a typical ℓ_2-norm cost function. The ℓ_0-norm constraint is non-convex and imposes combinatorial search for the exact solution to Eq. (4.6). Since exhaustive search generally requires a prohibitive number of computations, approximate solution methods such as matching pursuit (MP) and basis pursuit (BP) are preferred [9]. In this paper, orthogonal matching pursuit (OMP) [35] is used as the sparse solver. For small T, OMP achieves similar reconstruction accuracy relative to BP methods, but with much greater speed [9].

It has been shown that non-orthogonal, overcomplete dictionaries $\mathbf{Q}_N$ with $N > K$ (complete, $N = K$) can be designed to minimize both error and number of non-zero coefficients T, and thus provide greater compression over orthogonal dictionaries [9, 13, 16]. While overcomplete dictionaries can be designed by concatenating ortho-bases of wavelets or Fourier shape functions, better compression is often achieved by adapting the dictionary to the data under analysis using dictionary learning techniques [12, 13]. Since Eq. (4.6) promotes sparse solutions, it provides criteria for the design of dictionary $\mathbf{Q}$ for adequate reconstruction of $\mathbf{y}_m$ with a minimum number of non-zero coefficients. Rewriting Eq.(7) with

$$\min_{\mathbf{Q}} \Big\{ \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{Q}\mathbf{X}\|_{\mathcal{F}}^2 \ \text{ subject to } \ \forall_m, \|\mathbf{x}_m\|_0 \leq T \Big\}, \tag{4.7}$$

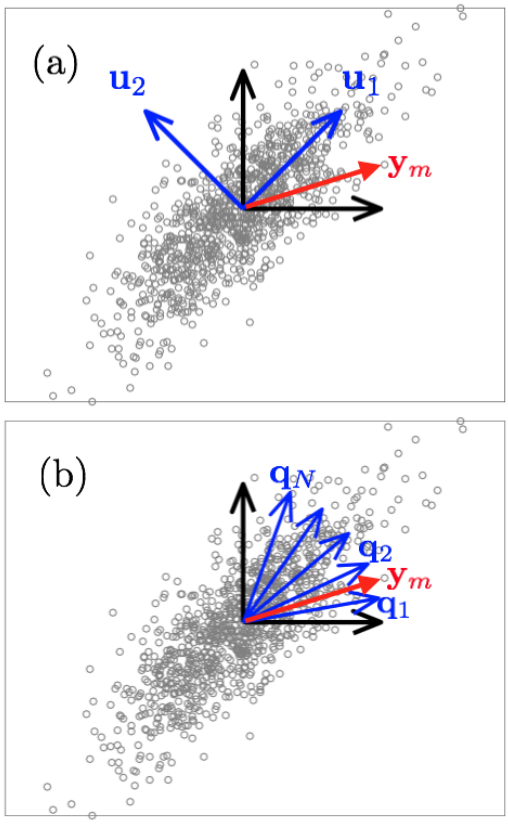


Figure 4.1: (Color online) (a) EOF vectors $[\mathbf{u}_1, \mathbf{u}_2]$ and (b) overcomplete LD vectors $[\mathbf{q}_1, ..., \mathbf{q}_N]$ for arbitrary 2D gaussian distribution relative to arbitrary 2D data observation $\mathbf{y}_m$.

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M]$ is the matrix of coefficient vectors corresponding to examples $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M]$, reconstruction error is minimized relative to the dictionary $\mathbf{Q}$ as well as relative to the sparse coefficients.

In this paper, the K-SVD algorithm, a clustering based dictionary learning method, is used to solve Eq.(4.7). The K-SVD is an adaptation of the K-means algorithm for vector quantization (VQ) codebook design (a.k.a. the generalized Lloyd algorithm) [16]. The learned dictionary (LD) vectors $\mathbf{q}_n$ from this technique partition the feature space of the data rather than $\mathbb{R}^K$, increasing the likelihood that $\mathbf{y}_m$ is as a linear combination of few vectors $\mathbf{q}_n$ in the solution to Eq. (4.6) (see Fig. 4.1). By increasing the number of vectors $N \geq K$ for overcomplete dictionaries, and thus the number of partitions in feature space, the sparsity of the solutions can be increased further [13].

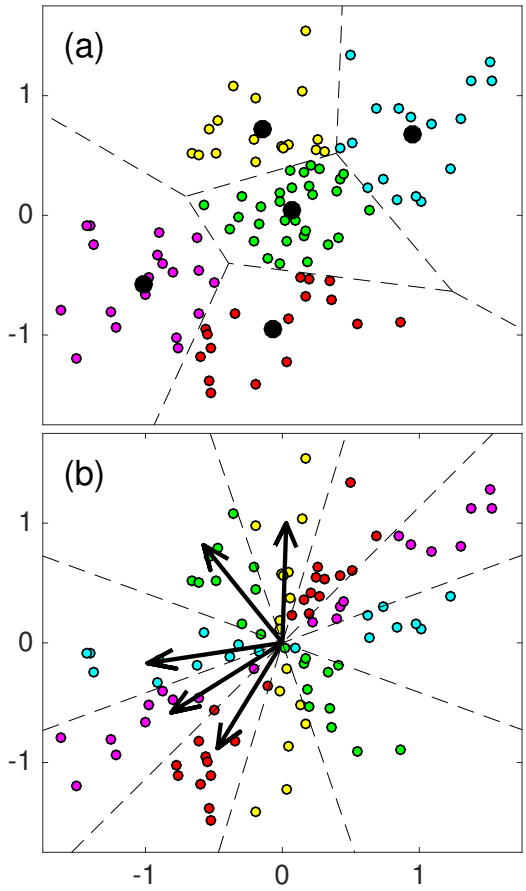


Figure 4.2: (Color online) Partitioning of gaussian random distribution ($\sigma_1 = 0.75, \sigma_2 = 0.5$) using (a) 5 codebook vectors (K-means, VQ) and with (b) 5 dictionary vectors from dictionary learning (K-SVD, $T = 1$).

4.2.4 Vector quantization

Vector quantization (VQ) [16] compresses a class of K–dimensional signals $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \mathbb{R}^{K \times M}$ by optimally mapping $\mathbf{y}_m$ to a set of code vectors $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_N] \in \mathbb{R}^{K \times N}$ for $N < M$, called a codebook. The signals $\mathbf{y}_m$ are then quantized or replaced by the best code vector choice from $\mathbf{C}$ [16]. The mapping that minimizes mean squared error (MSE) in reconstruction

$$\text{MSE}(\mathbf{Y}, \widehat{\mathbf{Y}}) = \frac{1}{\text{N}} \|\mathbf{Y} - \widehat{\mathbf{Y}}\|_{\mathcal{F}}^2, \tag{4.8}$$

where $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1, ..., \widehat{\mathbf{y}}_M]$ is the vector quantized $\mathbf{Y}$, is the assignment of each vector $\mathbf{y}_m$ to the code vectors $\mathbf{c}_n$ based on minimum ℓ_2–distance (nearest neighbor metric). Thus the ℓ_2–distances from the code vectors $\mathbf{c}_n$ define a set of partitions $(R_1, ..., R_N) \in \mathbb{R}^K$ (called Voronoi cells)

$$R_n = \left\{ i \mid \forall_{l \neq n}, \|\mathbf{y}_i - \mathbf{c}_n\|_2 < \|\mathbf{y}_i - \mathbf{c}_l\|_2 \right\}, \tag{4.9}$$

where if $\mathbf{y}_i$ falls within the cell R_n, $\widehat{\mathbf{y}}_i$ is $\mathbf{c}_n$. These cells are shown in Fig. 4.2(a). This is stated formally by defining a selector function S_n as

$$S_n(\mathbf{y}_m) = \begin{cases} 1 & \text{if } \mathbf{y}_m \in R_n \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

The vector quantization step is then

$$\widehat{\mathbf{y}}_m = \sum_{n=1}^{N} S_n(\mathbf{y}_m) \mathbf{c}_n. \tag{4.11}$$

The operations in Eq. (4.9–4.10) are analogous to solving the sparse minimization problem

$$\widehat{\mathbf{x}}_m = \underset{\mathbf{x}_m \in \mathbb{R}^N}{\arg\min} \|\mathbf{y}_m - \mathbf{C}\mathbf{x}_m\|_2 \;\; \text{subject to} \;\; \|\mathbf{x}_m\|_0 = 1, \tag{4.12}$$

where the non-zero coefficients $x_m^n = 1$. In this problem, selection of the coefficient in $\mathbf{x}_m$ corresponds to mapping the observation vector $\mathbf{y}_m$ to $\mathbf{c}_n$, similar to the selector function S_n. The vector quantized $\mathbf{y}_m$ is thus written, alternately from Eq. (4.11), as

$$\widehat{\mathbf{y}}_m = \mathbf{C}\widehat{\mathbf{x}}_m. \tag{4.13}$$

4.2.5 K-means

Given the MSE metric (Eq. (4.8)), VQ codebook vectors $[\mathbf{c}_1, ..., \mathbf{c}_N]$ which correspond to the centroids of the data $\mathbf{Y}$ within $(R_1, ..., R_N)$ minimize the reconstruction error. The assignment of $\mathbf{c}_n$ as the centroid of $\mathbf{y}_j \in R_n$ is

$$\mathbf{c}_n = \frac{1}{|R_n|} \sum_{j \in R_n} \mathbf{y}_j, \tag{4.14}$$

where $|R_n|$ is the number of vectors $\mathbf{y}_j \in R_n$.

The K-means algorithm shown in Table 4.1, iteratively updates $\mathbf{C}$ using the centroid condition Eq. (4.14) and the ℓ_2 nearest–neighbor criteria Eq. (4.9) to optimize the code vectors for VQ. The algorithm requires an initial codebook $\mathbf{C}^0$. For example, $\mathbf{C}^0$ can be N random vectors in $\mathbb{R}^K$ or selected observations from the training set $\mathbf{Y}$. The K-means algorithm is guaranteed to improve or leave unchanged the MSE distortion after each iteration and converges to a local minimum [12, 16].

4.3 Dictionary learning

Two popular algorithms for dictionary learning, the method of optimal directions (MOD) [13] and the K-SVD [12], are inspired by the iterative K-means codebook updates for VQ (Table 4.1). The N columns of the dictionary $\mathbf{Q}$, like the entries in codebook $\mathbf{C}$, correspond to partitions in $\mathbb{R}^K$. However, they are constrained to have unit ℓ_2-norm and thus separate the

Table 4.1: The K-means algorithm (Ref. [16].)

	Given: Training vectors $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \mathbb{R}^{K \times M}$
	Initialize: index $i = 0$, codebook $\mathbf{C}^0 = [\mathbf{c}_1^0, ..., \mathbf{c}_N^0] \in \mathbb{R}^{K \times N}$, MSE^0 solving Eq. (4.8)–(4.11)
I:	Update codebook
	1. Partition $\mathbf{Y}$ into N regions $(R_1, ..., R_N)$ by $R_n = \left\{ i \mid \forall_{l \neq n}, \|\mathbf{y}_i - \mathbf{c}_n^i\|_2 < \|\mathbf{y}_i - \mathbf{c}_l^i\|_2 \right\}$ (Eq. (4.9))
	2. Make code vectors centroids of $\mathbf{y}_j$ in partitions R_n $\mathbf{c}_n^{i+1} = \frac{1}{\lvert R_n^i \rvert} \sum_{j \in R_n^i} \mathbf{y}_j$
II.	Check error
	1. Calculate MSE^{i+1} from updated codebook $\mathbf{C}^{i+1}$
	2. If $\lvert \text{MSE}^{i+1} - \text{MSE}^{i} \rvert < \eta$
	$i = i + 1$, return to I
	else
	end

magnitude (coefficients $\mathbf{x}_n$) from the shapes (dictionary entries $\mathbf{q}_n$) for the sparse processing objective Eq.(4.6). When $T = 1$, the ℓ_2-norm in Eq. (4.6) is minimized by the dictionary entry $\mathbf{q}_n$ that has the greatest inner product with example $\mathbf{y}_m$ [9]. Thus for $T = 1$, $[\mathbf{q}_1, ..., \mathbf{q}_N]$ define radial partitions of $\mathbb{R}^K$. These partitions are shown in Fig. 4.2(b) for a hypothetical 2D ($K = 2$) random data set. This corresponds to a special case of VQ, called gain-shape VQ [16]. However, for sparse processing, only the shapes of the signals are quantized. The gains, which are the coefficients $\mathbf{x}_m$, are solved. For $T > 1$, the sparse solution is analogous to VQ, assigning examples $\mathbf{y}_m$ to dictionary entries in $\mathbf{Q}$ for up to T non-zero coefficients in $\mathbf{x}_m$.

Given these relationships between sparse processing with dictionaries and VQ, the MOD [13] and K-SVD [12] algorithms attempt to generalize the K-means algorithm to optimization of dictionaries for sparse processing for $T \geq 1$. They are two-step algorithms which reflect the two update steps in the K-means codebook optimization: (1) partition data $\mathbf{Y}$ into regions $(R_1, ..., R_N)$ corresponding to $\mathbf{c}_n$ and (2) update $\mathbf{c}_n$ to centroid of examples $\mathbf{y}_m \in R_N$. The K-means algorithm is generalized to the dictionary learning problem Eq.(4.7) as two steps:

1. Sparse coding: Given dictionary $\mathbf{Q}$, solve for up to T non-zero coefficients in $\mathbf{x}_m$ corre-

sponding to examples $\mathbf{y}_m$ for $m = [1, ..., M]$

2. Dictionary update: Given coefficients $\mathbf{X}$, solve for $\mathbf{Q}$ which minimizes reconstruction error for $\mathbf{Y}$.

The sparse coding step (1), which is the same for both MOD and K-SVD, is accomplished using any sparse solution method, including matching pursuit and basis pursuit. The algorithms differ in the dictionary update step.

4.3.1 The K-SVD Algorithm

The K-SVD algorithm is here chosen for its computational efficiency, speed, and convergence to local minima (at least for $T = 1$). The K-SVD algorithm sequentially optimizes the dictionary entries $\mathbf{q}_n$ and coefficients $\mathbf{x}_m$ for each update step using the SVD, and thus also avoids the matrix inverse. For $T = 1$, the sequential updates of the K-SVD provide optimal dictionary updates for gain-shape VQ [12, 16]. Optimal updates to the gain-shape dictionary will, like K-means updates, either improve or leave unchanged the MSE and convergence to a local minimum is guaranteed. For $T > 1$, convergence of the K-SVD updates to a local minimum depends on the accuracy of the sparse-solver used in the sparse coding stage [12].

In the K-SVD algorithm, each dictionary update step i sequentially improves both the entries $\mathbf{q}_n \in \mathbf{Q}^i$ and the coefficients in $\mathbf{x}_m \in \mathbf{X}^i$, without change in support. Expressing the coefficients as row vectors $\mathbf{x}_T^n \in \mathbb{R}^N$ and $\mathbf{x}_T^j \in \mathbb{R}^N$, which relate all examples $\mathbf{Y}$ to $\mathbf{q}_n$ and $\mathbf{q}_j$, respectively, the ℓ_2-penalty from Eq. (4.7) is rewritten as

$$\begin{aligned} \|\mathbf{Y} - \mathbf{QX}\|_{\mathcal{F}}^2 &= \left\| \mathbf{Y} - \sum_{n=1}^{N} \mathbf{q}_n \mathbf{x}_T^n \right\|_{\mathcal{F}}^2 \\ &= \|\mathbf{E}_j - \mathbf{q}_j \mathbf{x}_T^j\|_{\mathcal{F}}^2, \end{aligned} \tag{4.15}$$

where

$$\mathbf{E}_j = \left(\mathbf{Y} - \sum_{n \neq j} \mathbf{q}_n \mathbf{x}_T^n \right). \tag{4.16}$$

Thus, in Eq. (4.15) the ℓ_2-penalty is separated into an error term $\mathbf{E}_j = [\mathbf{e}_{j,1}, ..., \mathbf{e}_{j,M}] \in \mathbb{R}^{K \times M}$, which is the error for all examples $\mathbf{Y}$ if $\mathbf{q}_j$ is excluded from their reconstruction, and the product of the excluded entry $\mathbf{q}_j$ and coefficients $\mathbf{x}_T^j \in \mathbb{R}^N$.

An update to the dictionary entry $\mathbf{q}_j$ and coefficients $\mathbf{x}_T^j$ which minimizes Eq. (4.15) is found by taking the SVD of $\mathbf{E}_j$, which provides the best rank-1 approximation of $\mathbf{E}_j$. However, many of the entries in $\mathbf{x}_T^j$ are zero (corresponding to examples which don't use $\mathbf{q}_j$). To properly update $\mathbf{q}_j$ and $\mathbf{x}_T^j$ with SVD, Eq. (4.15) must be restricted to examples $\mathbf{y}_m$ which use $\mathbf{q}_j$

$$\|\mathbf{E}_j^R - \mathbf{q}_j \mathbf{x}_R^j\|_{\mathcal{F}}^2, \tag{4.17}$$

where $\mathbf{E}_j^R$ and $\mathbf{x}_R^j$ are entries in $\mathbf{E}_j$ and $\mathbf{x}_T^j$, respectively, corresponding to examples $\mathbf{y}_m$ which use $\mathbf{q}_j$, and are defined as

$$\mathbf{E}_j^R = \left\{ \mathbf{e}_{j,l} \middle| \forall_l,\, x_l^j \neq 0 \right\},\ \mathbf{x}_R^j = \left\{ x_l^j \middle|\, \forall_l,\, x_l^j \neq 0 \right\}. \tag{4.18}$$

Thus for each K-SVD iteration, the dictionary entries and coefficients are sequentially updated as the SVD of $\mathbf{E}_j^R = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$. The dictionary entry $\mathbf{q}_j^i$ is updated with the first column in $\mathbf{U}$ and the coefficient vector $\mathbf{x}_R^j$ is updated as the product of the first singular value $\mathbf{S}(1,1)$ with the first column of $\mathbf{V}$. The K-SVD algorithm is given in Table 4.2.

The dictionary $\mathbf{Q}$ is initialized using N randomly selected, ℓ_2-normalized examples from $\mathbf{Y}$. [9, 12] During the iterations, one or more dictionary entries may become unused. If this occurs, the unused entries are replaced using the most poorly represented examples $\mathbf{y}_m$ (ℓ_2-normlized), determined by reconstruction error.

Table 4.2: The K-SVD Algorithm (Ref. [12])

	Given: $\mathbf{Y} \in \mathbb{R}^{K\times M}$, $\mathbf{Q}^0 \in \mathbb{R}^{K\times N}$, $T \in \mathbb{N}$, and $i = 0$
	Repeat until convergence:
1.	Sparse coding
	for $m = 1 : M$
	solve Eq. (4.6) using any sparse solver
a:	$\widehat{\mathbf{x}}_m = \underset{\mathbf{x}_m \in \mathbb{R}^N}{\arg\min} \|\mathbf{y}_m - \mathbf{Q}^i \mathbf{x}_m\|_2$ subject to $\|\mathbf{x}_m\|_0 \leq T$
	end
b:	$\mathbf{X} = [\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_M]$
2.	Dictionary update
	for $j = 1 : N$
a:	compute reconstruction error $\mathbf{E}_j$ as
	$\mathbf{E}_j = \mathbf{Y} - \sum_{n \neq j} \mathbf{q}_n^i \mathbf{x}_T^n$
b:	obtain $\mathbf{E}_j^R$, $\mathbf{x}_R^j$ corresponding to nonzero $\mathbf{x}_T^j$
c:	apply SVD to $\mathbf{E}_j^R$
	$\mathbf{E}_j^R = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$
d:	update $\mathbf{q}_j^i$: $\mathbf{q}_j^i = \mathbf{U}(:,1)$
e:	update $\mathbf{x}_R^j$: $\mathbf{x}_R^j = \mathbf{V}(:,1)\mathbf{S}(1,1)$
	end
f:	$\mathbf{Q}^{i+1} = \mathbf{Q}^i$
	$i = i + 1$

4.4 Experimental results

To demonstrate the usefulness of the dictionary learning approach, we here analyze two data sets: (1) thermistor data from the HF-97 acoustics experiment [31, 32], conducted off the coast of Point Loma, CA and (2) CTD data collected across the Luzon Strait near the South China Sea (SCS) [33]. Training data $\mathbf{Y}$ were derived from the data sets by converting raw thermistor and CTD data to SSPs and subtracting the mean. The HF-97 thermistor data was recorded every 15 s, over a 48 hour period, from 14 to 70 m depth, with 4 m spacing (15 points). The full 11,488 profile data set was down-sampled to $M = 1000$ profiles for the training set, and SSPs were interpolated to $K = 30$ points using a shape-preserving cubic spline. The SCS CTD data was recorded at about 1 m resolution from 116 to 496 m depth (384 points). From the SCS data set,

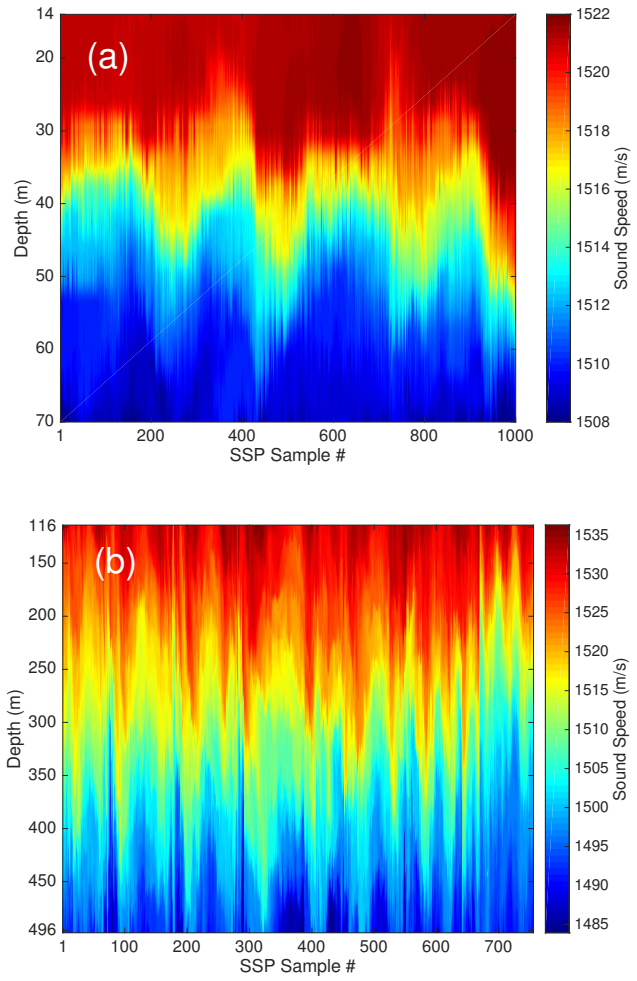


Figure 4.3: (Color online) SSP data from (a) HF-97 experiment and (b) SCS.

$M = 755$ profiles were used as the training set, and the profiles were uniformly down-sampled to $K = 50$ points. The SSP data sets are shown in Fig. 4.3. Both data sets have small and large spatiotemporal variations.

EOFs were calculated from the SVD (Eq. 4.2) and LDs (learned dictionaries) were generated with the K-SVD algorithm (Table 4.2), using OMP for the sparse coding stage. The number of non-zero coefficients solved with OMP for each dictionary was held fixed at exactly T non-zero coefficients. The initial dictionary $\mathbf{Q}^0$ was populated using randomly selected examples from the training sets $\mathbf{Y}$.

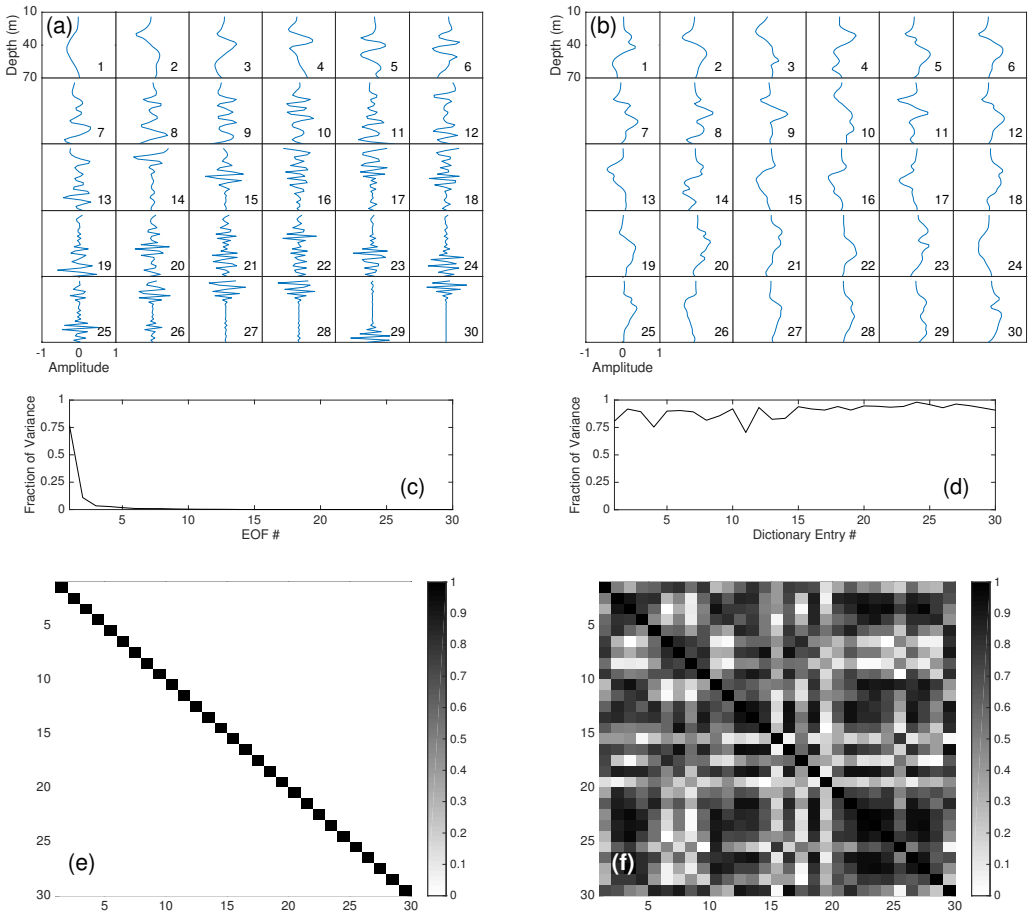


Figure 4.4: (Color online) HF-97: (a) EOFs and (b) LD entries ($N = K$ and $T = 1$, sorted by variance $\sigma^2_{\mathbf{q}_n}$). Fraction of (c) total SSP variance explained by EOFs and (d) SSP variance explained for examples using LD entries. Coherence of (e) EOFs and (f) LD entries.

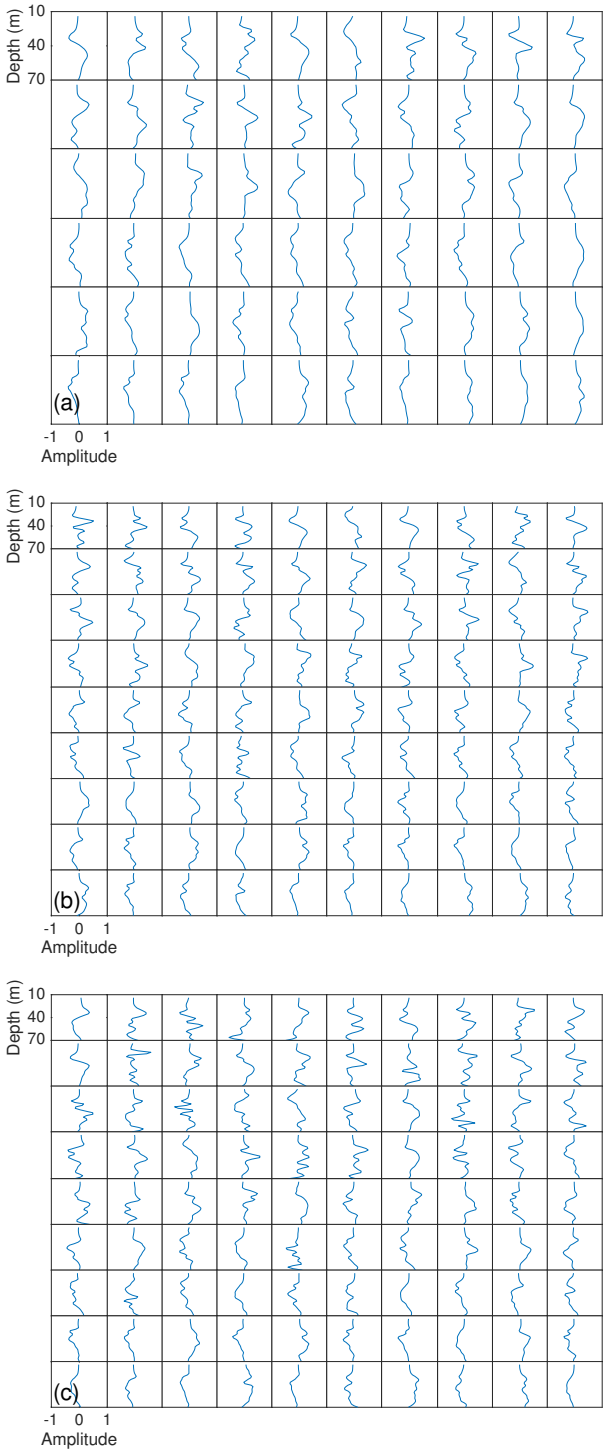


Figure 4.5: (Color online) HF-97: LD entries (a) $N = 60$ and $T = 1$, (a) $N = 90$ and $T = 1$, and (c) $N = 90$ and $T = 5$. Dictionary entries are sorted in descending variance $\sigma^2_{\mathbf{q}_n}$.

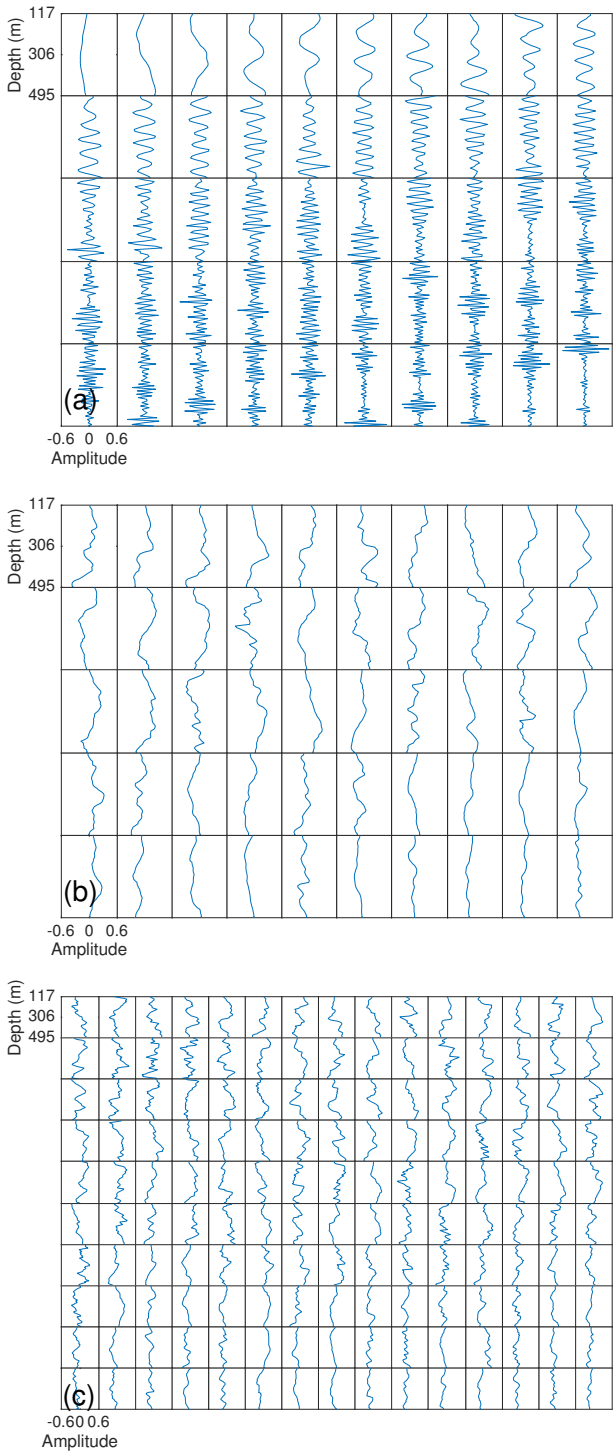


Figure 4.6: (Color online) SCS: EOFs (a) and LD entries; (b) $N = K = 50$ and $T = 1$ and (c) $N = 150$ and $T = 1$. Dictionary entries are sorted in descending variance $\sigma^2_{\mathbf{q}_n}$.

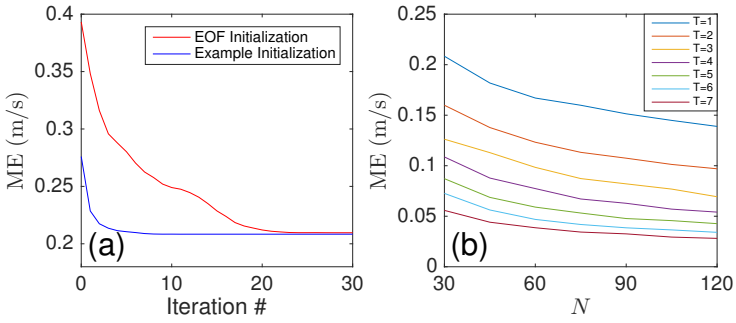


Figure 4.7: (Color online) HF-97: (a) Convergence of LD ($N = 30$, $T = 1$) mean reconstruction error (ME), initialized using EOFs or N randomly selected examples from $\mathbf{Y}$. (b) ME versus non-zero coefficients T and number of dictionary entries N.

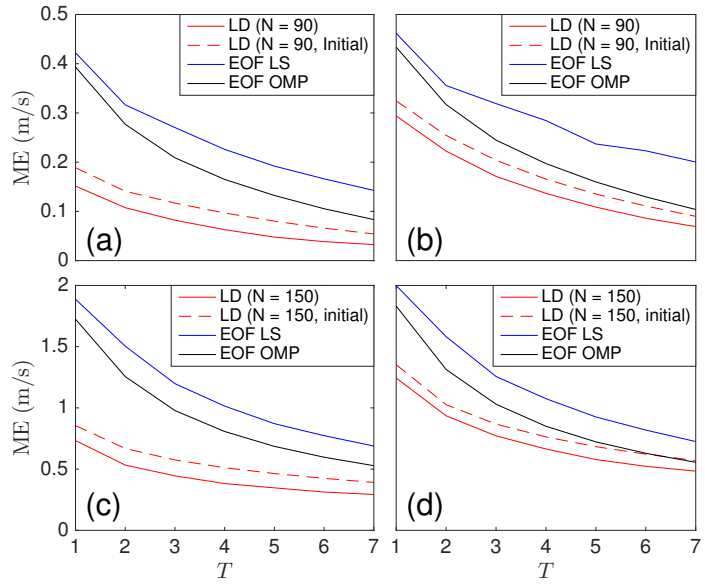


Figure 4.8: (Color online) Mean reconstruction error (ME) versus T using EOFs (solved using LS and OMP) and LDs ($N = 90$ for HF-97 and $N = 150$ for SCS) for (a) HF- 97 and (c) SCS. Mean reconstruction error ME_{CV} for out-of-sample data calculated with K-fold cross validation for J = 10 folds, (b) HF-97 and (d) SCS.

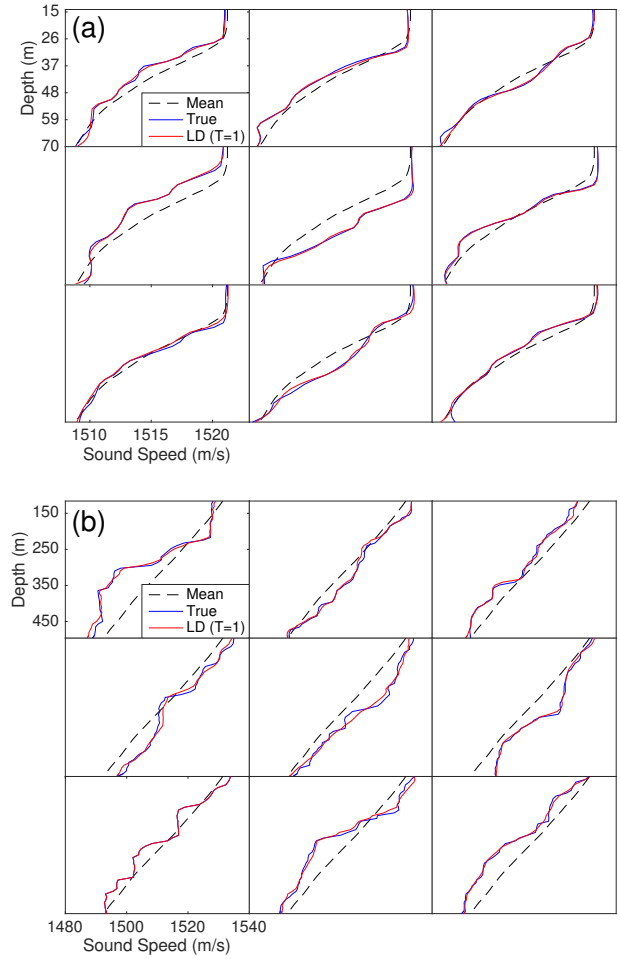


Figure 4.9: (Color online) True SSP reconstruction of 9 example profiles using one coefficient ($T = 1$) from LD for (a) HF-97 ($N = 90$) and (b) SCS ($N = 150$).

4.4.1 Learning SSP dictionaries from data

Here, LDs and EOFs were generated using the full SSP data from HF-97 ($M = 1000$) and SCS ($M = 755$). The EOFs and LDs from HF-97 are shown in Fig. 4.4–4.5 and from the SCS in Fig. 4.6. The HF-97 LD, with $N = K$ and $T = 1$, is compared to the EOFs ($K = 30$) in Fig. 4.4. Only the leading order EOFs (Fig. 4.4(a)) are informative of ocean SSP variability whereas all shape functions in the LD (Fig. 4.4(b)) are informative (Fig. 4.4(c)–(d)). This behavior is also evident for the SCS data set (Fig. 4.6). The EOFs ($K = 50$) calculated from the full training set are shown in Fig. 4.6(a), and the LD entries for $N = 50$ and $T = 1$ sparse coefficient are shown in

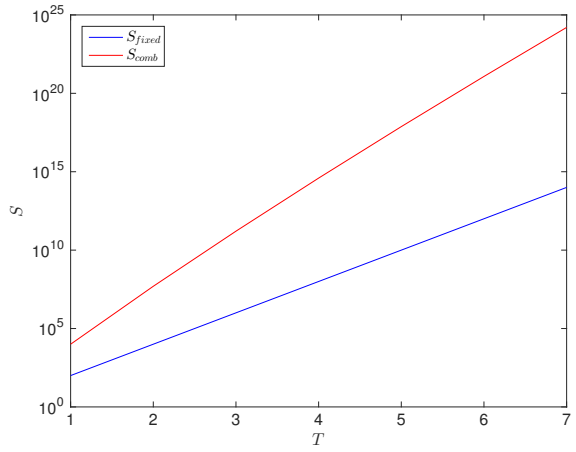


Figure 4.10: (Color online) Number of candidate solutions S for SSP inversion versus T, S_{fixed} using fixed indices and S_{comb} best combination of coefficients. Each coefficient is discretized with $H = 100$ for dictionary $\mathbf{Q} \in \mathbb{R}^{K \times N}$ with $N = 100$.

Fig. 4.6(b). The overcomplete LDs for the HF-97 data shown in Fig. 4.5 and for the SCS data in Fig. 4.6(c).

As illustrated in Fig. 4.1, by relaxing the requirement of orthogonality for the shape functions, the shape functions can be better fit to the data and thereby achieve greater compression. The Gram matrix $\mathbf{G}$, which gives the coherence of matrix columns, is defined for a matrix $\mathbf{A}$ with unit ℓ_2-norm columns as $\mathbf{G} = |\mathbf{A}^{\mathrm{T}}\mathbf{A}|$. The Gram matrix for the EOFs (Fig. 4.4(e)) shows the shapes in the EOF dictionary are orthogonal ($\mathbf{G} = \mathbf{I}$, by definition), whereas those of the LD (Fig. 4.4(f)) are not.

4.4.2 Reconstruction of SSP training data

In this section, EOFs and LDs are trained on the full SSP data sets $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M]$. Reconstruction performance of the EOF and LDs are then evaluated on SSPs within the training set, using a mean error metric.

The coefficients for the learned $\mathbf{Q}$ and initial $\mathbf{Q}^0$ dictionaries $\widehat{\mathbf{x}}_m$ are solved from the sparse objective (Eq. (4.6)) using OMP. The LS solution for the T leading-order coefficients $\mathbf{x}_L \in \mathbb{R}^T$

from the EOFs $\mathbf{P}$ were solved by Eq. (4.5). The best combination of T EOF coefficients was solved from the sparse objective (Eq. (4.6)) using OMP. Given the coefficients $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_m]$ describing examples $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_m]$, the reconstructed examples $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1, ..., \widehat{\mathbf{y}}_m]$ are given by $\widehat{\mathbf{Y}} = \mathbf{Q}\widehat{\mathbf{X}}$. The mean reconstruction error ME for the training set is then

$$\mathrm{ME} = \frac{1}{KM} \|\mathbf{Y} - \widehat{\mathbf{Y}}\|_1. \tag{4.19}$$

We here use the ℓ_1-norm to stress the robustness of the LD reconstruction.

To illustrate the optimality of LDs for SSP compression, the K-SVD algorithm was run using EOFs as the initial dictionary $\mathbf{Q}^0$ for $T = 1$ non-zero coefficient. The convergence of ME for the K-SVD iterations is shown in Fig. 4.7(a). After 30 K-SVD iterations, the mean error of the $M = 1000$ profile training set is decreased by nearly half. The convergence is much faster for $\mathbf{Q}^0$ consisting of randomly selected examples from $\mathbf{Y}$.

For LDs, increasing the number of entries N or increasing the number of sparse coefficients T will always reduce the reconstruction error (N and T are decided with computational considerations). The effect of N and T on the mean reconstruction error for the HF-97 data is shown in Fig. 4.7(b). The errors are calculated for the range $N = K$ to $N = 4K$ and the dictionaries were optimized to use a fixed number non-zero coefficients (T).

The reconstruction error using the EOF dictionary is compared to results from LDs $\mathbf{Q}$ with $N = 3K$, using T non-zero coefficients. In Fig. 4.8[(a) and (c)] results are shown for the HF-97 ($N = 90$) and SCS ($N = 150$) data, respectively. Coefficients describing each example $\mathbf{y}_m$, were solved (1) from the LD $\mathbf{Q}$, (2) from $\mathbf{Q}^0$, the dictionary consisting of N randomly chosen examples from the training set (to illustrate improvements in reconstruction error made in the K-SVD iterations), (3) the leading order EOFs, and (4) the best combination of EOFs. The mean SSP reconstruction error using the LDs trained for each sparsity T is less than EOF reconstruction, for either leading order coefficients or best coefficient combination, for all values of T shown.

The best combination of EOF coefficients, chosen approximately using OMP, achieves less error than the LS solution to the leading order EOFs, with added cost of search.

Just one LD entry achieves the same ME as more than 6 leading order EOF coefficients, or greater than 4 EOF coefficients chosen by search (Fig. 4.8[(a) and (c)]). To illustrate the representational power of the LD entries, both true and reconstructed SSPs are shown in Fig. 4.9(a) for the HF-97 data and in Fig. 4.9(b) for the SCS data. Nine true SSP examples from each training set, for HF-97 (SCS) taken at 100 (80) point intervals from $m = 100$ to 900 (80 to 720), are reconstructed using one LD coefficient. It is shown for each case, that nearly all of the SSP variability is captured using a single LD coefficient.

4.4.3 Cross-validation of SSP reconstruction

The out-of-sample SSP reconstruction performance of LDs and EOFs is tested using K-fold cross-validation [34]. The entire SSP data set $\mathbf{Y}$ of M profiles, for each experiment, is divided into J subsets with equal numbers of profiles $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_J]$, where the fold $\mathbf{Y}_j \in \mathbb{R}^{K \times (M/J)}$. For each of the J folds: (1) $\mathbf{Y}_j$ is the set of out of sample test cases, and the training set $\mathbf{Y}_{tr}$ is

$$\mathbf{Y}_{tr} = \left\{ \mathbf{Y}_l \middle| \, \forall_{l \neq j} \right\}; \tag{4.20}$$

(2) the LD $\mathbf{Q}_j$ and EOFs are derived using $\mathbf{Y}_{tr}$; and (3) coefficients estimating test samples $\mathbf{Y}_j$ are solved for $\mathbf{Q}_j$ with sparse processor Eq. (4.6), and for EOFs by solving for leading order terms and by solving with sparse processor. The out of sample error from cross validation $\mathrm{ME_{CV}}$ for each method is then

$$\mathrm{ME}_{CV} = \frac{1}{KM} \sum_{j=1}^{J} \| \mathbf{Y}_j - \widehat{\mathbf{Y}}_j \|_1. \tag{4.21}$$

The out of sample reconstruction error $\mathrm{ME_{CV}}$ increases over the within-training-set estimates for both the learned and EOF dictionaries, as shown in Fig. 4.8[(b) and (d)] for $J = 10$ folds. The mean reconstruction error using the LDs, as in the within-training-set estimates, is

less than the EOF dictionaries. For both the HF-97 (SCS) data, more than 2 (2) EOF coefficients, choosing best combination by search, or more than 3 (equal to 3) leading-order EOF coefficients solved with LS, are required to achieve the same out of sample performance as one LD entry.

4.4.4 Solution space for SSP inversion

Acoustic inversion for ocean SSP is a non-linear problem. One approach is coefficient search using genetic algorithms [1]. Discretizing each coefficient into H values, the number of candidate solutions for T fixed coefficients indices is

$$S_{\text{fixed}} = H^T. \tag{4.22}$$

If the coefficient indices for the solution can vary, as per dictionary learning with LD $\mathbf{Q} \in \mathbb{R}^{K\times N}$, the number of candidate solutions S_{comb} is

$$S_{\text{comb}} = H^T \frac{N!}{T!(N-T)!}. \tag{4.23}$$

Using a typical $H = 100$ point discretization of the coefficients, the number of possible solutions for fixed and combinatorial dictionary indices are plotted in Fig. 4.10. Assuming an unknown SSP similar to the training set, the SSP may be constructed up to acceptable resolution using one coefficient from the LD (10^4 possible solutions, see Fig. 4.10). To achieve the similar ME, 7 EOFs coefficients are required (10^{14} possible solutions, Fig. 4.10) using fixed indices and the best EOF combination requires 5 EOFs (10^{17} possible solutions, Fig. 4.10).

4.5 Conclusion

Given sufficient training data, dictionary learning generates optimal dictionaries for sparse reconstruction of a given signal class. Since these LDs are not constrained to be orthogonal, the

entries fit the distribution of the data such that signal example is approximated using few LD entries. Relative to EOFs, each LD entry is informative to the signal variability.

The K-SVD dictionary learning algorithm is applied to ocean SSP data from the HF-97 and SCS experiments. It is shown that the LDs generated describe ocean SSP variability with high resolution using fewer coefficients than EOFs. As few as one coefficient from a LD describes nearly all the variability in each of the observed ocean SSPs. This performance gain is achieved by the larger number of informative elements in the LDs over EOF dictionaries. Provided sufficient SSP training data is available, LDs can improve SSP inversion resolution with negligible computational expense. This could provide improvements to geoacoustic inversion [1], matched field processing [36, 37], and underwater communication [31].

4.6 Acknowledgments

The authors would like to thank Dr. Robert Pinkel for the use of the South China Sea CTD data. This work is supported by the Office of Naval Research, Grant No. N00014-11-1-0439.

Chapter 4, in full, is a reprint of the material as it appears in the Journal of the Acoustical Society of America 2017. Bianco, Michael; Gerstoft, Peter, Acoustical Society of America, 2017. The dissertation author was the primary investigator and author of this paper.

4.7 Bibliography

[1] P. Gerstoft, "Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 770–782, 1994.

[2] L. R. LeBlanc and F. H. Middleton, "An underwater acoustic sound velocity data model," *J. Acoust. Soc. Am.*, vol. 67, no. 6, pp. 2055–2062, 1980.

[3] M. I. Taroudakis and J. S. Papadakis, "A modal inversion scheme for ocean acoustic tomography," *J. Comput. Acoust.*, vol. 1, no. 4, pp. 395–421, 1993.

[4] P. Gerstoft and D. F. Gingras, "Parameter estimation using multifrequency range–dependent acoustic data in shallow water," *J. Acoust. Soc. Am.*, vol. 99, no. 5, pp. 2839–2850, 1996.

[5] C. Park, W. Seong, P. Gerstoft, and W. S. Hodgkiss, "Geoacoustic inversion using backpropagation," *IEEE Journal of Oceanic Engineering*, vol. 35, no. 4, pp. 722–731, 2010.

[6] C.-F. Huang, P. Gerstoft, and W. S. Hodgkiss, "Effect of ocean sound speed uncertainty on matched-field geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. EL162–EL168, 2008.

[7] B. A. Tan, P. Gerstoft, C. Yardim, and W. S. Hodgkiss, "Broadband synthetic aperture geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 312–322, 2013.

[8] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[9] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[10] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.

[11] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd," *Appl. and Comp. Harm. Anal.*, vol. 37, no. 3, pp. 464–491, 2014.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, pp. 4311–4322, 2006.

[13] K. Engan, S. O. Aase, and J. H. H. y, "Multi-frame compression: theory and design," *Signal Processing*, vol. 80, pp. 2121–2140, 2000.

[14] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Sci. Bus. Media, 2009.

[15] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Trans. Cons. Elec.*, vol. 46, pp. 1103–1127, 2000.

[16] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Kluwer Academic, Norwell, MA, 1991.

[17] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1–norm," *Geophys.*, vol. 44, no. 1, pp. 39–52, 1979.

[18] E. Candés, "Compressive sampling," *Proc. Internat. Cong. Math.*, vol. 3, pp. 1433–1452, 2006.

[19] G. F. Edelmann and C. F. Gaumond, "Beamforming using compressive sensing," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. EL232–EL237, 2011.

[20] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 260–271, 2014.

[21] P. Gerstoft, A. Xenaki, and C. F. Mecklenbräuker, "Multiple and single snapshot compressive beamforming," *J. Acoust. Soc. Am.*, vol. 138, no. 4, pp. 2003–2014, 2015.

[22] Y. Choo and W. Seong, "Compressive spherical beamforming for localization of incipient tip vortex cavitation," *J. Acoust. Soc. Am.*, vol. 140, no. 6, pp. 4085–4090, 2016.

[23] C. Yardim, P. Gerstoft, W. S. Hodgkiss, and J. Traer, "Compressive geoacoustic inversion using ambient noise," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1245–1255, 2014.

[24] M. Bianco and P. Gerstoft, "Compressive acoustic sound speed profile estimation," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. EL90–EL94, 2016.

[25] S. Beckouche and J. Ma, "Simultaneous dictionary learning and denoising for seismic data," *Geophys.*, vol. 79, no. 3, pp. 27–31, 2014.

[26] M. Taroudakis and C. Smaragdakis, "De-noising procedures for inverting underwater acoustic signals in applications of acoustical oceanography," *EuroNoise 2015 Maastricht*, pp. 1393–1398, 2015.

[27] T. Wang and W. Xu, "Sparsity-based approach for ocean acoustic tomography using learned dictionaries," *OCEANS 2016 Shanghai IEEE*, pp. 1–6, 2016.

[28] K. S. Alguri and J. B. Harley, "Consolidating guided wave simulations and experimental data: a dictionary learning approach," *Proc. SPIE*, vol. 9805, p. 98050Y, 2016.

[29] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science: a review," *International Journal of Climatology*, vol. 27, pp. 1119–1152, 2007.

[30] A. H. Monahan, J. C. Fyfe, M. H. Ambaum, D. B. Stephenson, and G. R. North, "Empirical orthogonal functions: The medium is the message," *J. Climate*, vol. 22, no. 24, pp. 6501–6514, 2009.

[31] N. Carbone and W. S. Hodgkiss, "Effects of tidally driven temperature fluctuations on shallow-water acoustic communications at 18khz," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 84–94, 2000.

[32] W. S. Hodgkiss, W. A. Kuperman, and D. E. Ensberg, "Channel impulse response fluctuations at 6 kHz in shallow water," in *Impact of Littoral Environmental Variability of Acoustic Predictions and Sonar Performance*. Springer, 2002.

[33] C. T. Liu, R. Pinkel, M. K. Hsu, J. M. Klymak, H. W. Chen, and C. Villanoy, "Nonlinear internal waves from the luzon strait," *Eos Trans. AGU*, vol. 87, no. 42, pp. 449–451, 2006.

[34] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[35] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *27th. Annual Asilomar Conference on Signals, Systems and Computers, IEEE Proc.*, pp. 40–44, 1993.

[36] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE Journal of Oceanic Engineering*, vol. 18, no. 4, pp. 401–424, 1993.

[37] C. M. Verlinden, J. Sarkar, W. S. Hodgkiss, W. A. Kuperman, and K. G. Sabra, "Passive acoustic source localization using sources of opportunity," *J. Acoust. Soc. Am. Expr. Let.*, vol. 138, no. 1, pp. EL54–EL59, 2015.

Chapter 5

Adaptive travel time tomography with local sparsity

5.1 Introduction

Travel time tomography methods attempt to estimate complex Earth structure, which contains smooth and discontinuous features at multiple spatial scales, using seismic and acoustic wave travel times between recording stations [1,2]. The inversion of the travel times for a slowness model (inverse of speed) is ill-posed, with often dense but irregular ray coverage of environments. Conventional tomography techniques regularize the inversion by restricting the models to be only smooth or discontinuous, which include 1st and 2nd order Tikhonov regularization [1, 2]. Other regularization methods have been proposed which employ of wavelet functions [3–6], total variation (TV) [7, 8], or adaptive discretizations of slowness [9].

Recent works in acoustics have utilized sparse modeling and compressive sensing (CS) [10–12] to improve performance in beamforming [13, 14] and inversion for ocean acoustic properties [15–18]. Similarly, the more recent wavelet-based methods in seismic tomography, e.g. [4, 6, 19], assume sparse wavelet coefficients. In sparse modeling and CS, inverse problems

are regularized by modeling signals as sparse combinations of vectors or *atoms* from set or *dictionary* of atoms, which can be prescribed or learned [11, 12, 20]. This paradigm is ubiquitous in signal processing for image denoising and inpainting [11, 12], and medical imaging [21, 22], to name a few examples. Learned dictionaries can improve reconstruction performance over prescribed dictionaries and recently, inversion methods with dictionary learning have been developed in the geosciences. Applications include denoising seismic traces [23] and ocean acoustic recordings [24], full waveform inversion [25], and estimation of ocean sound speed profiles [15, 16].

In this paper, we develop a sparse and adaptive approach to 2D travel time tomography, which we refer to as locally-sparse travel time tomography (LST). The LST sparsely models local behaviors of overlapping groups of pixels from a discrete slowness map, called *patches*. Large scale features in the slowness map are constrained using least squares. This approach is similar to works in image denoising [26] and CS magnetic resonance imaging (MRI) [22]. We develop a *maximum a posteriori* (MAP) formulation to the problem and use the iterative thresholding and signed K-means (ITKM) dictionary learning algorithm to improve the slowness models over prescribed dictionaries. We demonstrate the performance of LST considering 2D surface wave tomography with synthetic slowness maps and travel time data. The results are compared with conventional tomography. More details of the approach and further experimental results are available in a forthcoming paper [27].

5.2 Overview of LST

Given travel time perturbations $\mathbf{t} \in \mathbb{R}^M$ from M ray paths through a discrete slowness map (see Fig. 6.1(a)), and tomography matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, LST estimates the *sparse slowness* $\mathbf{s}_\mathrm{s}$. We first estimate the *global slowness* $\mathbf{s}_\mathrm{g}$, and then obtain the *patch slowness* $\mathbf{D}\mathbf{x}_i$ for patch i of $\mathbf{s}_\mathrm{g}$. Here $\mathbf{D} \in \mathbb{R}^{n \times Q}$ is a dictionary of Q atoms, and $\mathbf{x}_i \in \mathbb{R}^n$ is the sparse coefficients with n the

number of pixels in a patch. Finally slownesses $\{\mathbf{D}\widehat{\mathbf{x}}_i \; \forall \; i\}$ are averaged with $\mathbf{s}_\mathrm{g}$ to obtain $\mathbf{s}_\mathrm{s}$.

5.2.1 Global slowness and travel time

We discretize a 2D slowness map as a $W_1 \times W_2$ pixel image, shown in Fig. 6.1(a), where each pixel has constant slowness. The slowness pixels are represented by the vector $\mathbf{s}' = \mathbf{s}_\mathrm{g} + \mathbf{s}_0 \in \mathbb{R}^N$, where $\mathbf{s}_0$ is reference slownesses and $\mathbf{s}_\mathrm{g}$ is perturbations from the reference, with $N = W_1 W_2$. We assume travel time observations $\mathbf{t}' = \mathbf{t} + \mathbf{t}_0$ from M straight ray paths, where $\mathbf{t}_0$ and $\mathbf{t}$ are the reference travel time and perturbations. Since $\mathbf{s}_0$ and $\mathbf{t}_0 = \mathbf{A}\mathbf{s}_0$ are known, we estimate the perturbations

$$\mathbf{t} = \mathbf{A}\mathbf{s}_\mathrm{g} + \boldsymbol{\varepsilon}, \tag{5.1}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ is Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$, with mean $\mathbf{0}$ and covariance $\sigma_\varepsilon^2 \mathbf{I}$, in the travel time observations. We call (S1) the *global model*, as it captures the large-scale features that span the discrete map and generates $\mathbf{t}$. We assume dense ray coverage, and do not explicitly account for varying ray density (see Sec. 6.3).

5.2.2 Local sparse model

Each patch is a $\sqrt{n} \times \sqrt{n}$ group of pixels from $\mathbf{s}_\mathrm{s}$, see Fig. 6.1(a). The patches are selected from $\mathbf{s}_\mathrm{s}$ by the binary matrix $\mathbf{R}_i \in \{0,1\}^{n \times N}$. Hence the slownesses in patch i are $\mathbf{R}_i \mathbf{s}_\mathrm{s}$. Each patch is indexed by the row w_1 and column w_2 of its top-left pixel in the 2D image as $(w_{1,i}, w_{2,i})$. We consider all overlapping patches, with $w_{1,i}$ and $w_{2,i}$ differing from their neighbor by ± 1 (*stride* of one). Thus, for a $W_1 \times W_2$ pixel image, the number of patches is $I = (W_1 - \sqrt{n} + 1)(W_2 - \sqrt{n} + 1)$.

$\mathbf{R}_i \mathbf{s}_\mathrm{s}$ is approximated by sparse combinations atoms from $\mathbf{D}$. The coefficients $\mathbf{x}_i$ are estimated using the ℓ_0 pseudo-norm (see (S5)), which penalizes the number of non-zero coefficients [12]. We call (S5) the local model, as it captures the smaller scale, localized features

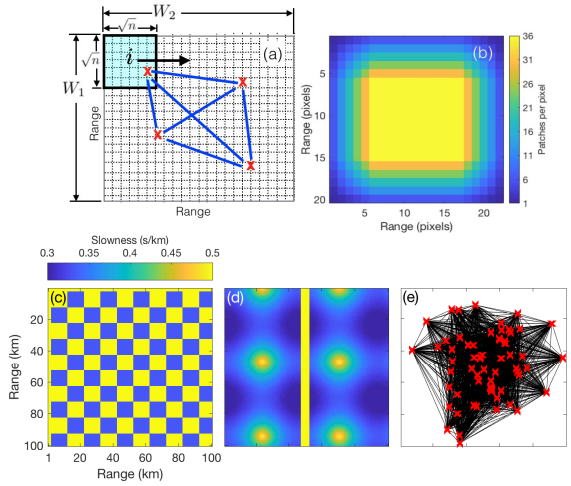


Figure 5.1: (a) 2D slowness patches and slowness map parameters, with (b) example patch distribution. Synthetic slowness $\mathbf{s}'$ for (c) checkerboard map and (d) smooth-discontinuous map ($W_1 = W_2 = 100$ pixels (km)). (e) 2016 straight ray paths (surface wave) from 64 seismic stations (red X's).

contained by patches.

The atoms in $\mathbf{D}$ are considered "elemental patches", where only a small number of atoms from $Q \ll I$ are necessary to adequately approximate $\mathbf{R}_i\mathbf{s}_\mathrm{s}$. Atoms can be prescribed functions, e.g. wavelets or the discrete cosine transform (DCT), or learned from the data (see Sec. 6.3.3).

5.3 Derivation of LST MAP objective

Starting with Bayes' rule, we derive the LST MAP objective for $\mathbf{s}_\mathrm{s}$, incorporating both local sparse prior and global constraints. For the derivation, we assume the dictionary $\mathbf{D}$ and sensing matrix $\mathbf{A}$ known. In Sec. 6.3.3, dictionary learning is included in the algorithm.

The posterior density is formulated as

$$p(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X})\, p(\mathbf{s}_\mathrm{g}|\mathbf{s}_\mathrm{s}, \mathbf{X})\, p(\mathbf{s}_\mathrm{s}|\mathbf{X})\, p(\mathbf{X}), \tag{5.2}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_I] \in \mathbb{R}^{Q \times I}$ are the coefficients describing all patches. If $\mathbf{s}_\mathrm{g}$ ($\mathbf{s}_\mathrm{s}$) is known, so is

$\mathbf{t}$ ($\mathbf{s}_\mathrm{g}$), whereby

$$p\big(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big|\mathbf{t}\big) \propto p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big) p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big) p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big) p\big(\mathbf{X}\big). \tag{5.3}$$

We assume $p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big)$, $p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big)$, and $p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big)$ are Gaussian, which results in a simple LST objective. Hence, for the global model, $p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big) = \mathcal{N}(\mathbf{A}\mathbf{s}_\mathrm{g}, \boldsymbol{\Sigma}_\varepsilon)$ and $p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big) = \mathcal{N}(\mathbf{s}_\mathrm{s}, \boldsymbol{\Sigma}_\mathrm{g})$ where $\boldsymbol{\Sigma}_\varepsilon \in \mathbb{R}^{K\times K}$ is the covariance of ε and $\boldsymbol{\Sigma}_\mathrm{g} \in \mathbb{R}^{N\times N}$ is the covariance of $\mathbf{s}_\mathrm{g}$.

For the local model, the patch slownesses $\{\mathbf{R}_i\mathbf{s}_\mathrm{s} \;\forall\, i\;\}$ are considered independent, giving the local likelihood $p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big)$

$$p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big) = \prod_i p\big(\mathbf{R}_i\mathbf{s}_\mathrm{s}\big|\mathbf{x}_i\big) = \prod_i \mathcal{N}\big(\mathbf{D}\mathbf{x}_i, \boldsymbol{\Sigma}_{p,i}\big) \tag{5.4}$$

where $\boldsymbol{\Sigma}_{p,i} \in \mathbb{R}^{n\times n}$ is the covariance of the patch slownesses for each patch i. Assuming the coefficients $\mathbf{x}_i$ independent and sparse, $\ln p\big(\mathbf{X}\big) = \sum_i \ln p\big(\mathbf{x}_i\big)$, with $\ln p\big(\mathbf{x}_i\big) \propto \|\mathbf{x}_i\|_0$. We further assume the number of non-zero coefficients T is the same for every patch (for which the ℓ_0-norm penalty is well suited), and errors iid with $\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}$, $\boldsymbol{\Sigma}_\mathrm{g} = \sigma_\mathrm{g}^2\mathbf{I}$, and $\boldsymbol{\Sigma}_{p,i} = \sigma_{p,i}^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Hence the MAP estimate $\big\{\widehat{\mathbf{s}}_\mathrm{g}, \widehat{\mathbf{s}}_\mathrm{s}, \widehat{\mathbf{X}}\big\}$ is from (6.5)

$$\begin{aligned}\big\{\widehat{\mathbf{s}}_\mathrm{g},\, \widehat{\mathbf{s}}_\mathrm{s},\, \widehat{\mathbf{X}}\big\} = \underset{\mathbf{s}_\mathrm{g},\, \mathbf{s}_\mathrm{s},\, \mathbf{X}}{\arg\min} \Bigg\{ &\frac{1}{\sigma_\varepsilon^2}\|\mathbf{t} - \mathbf{A}\mathbf{s}_\mathrm{g}\|_2^2 + \frac{1}{\sigma_\mathrm{s}^2}\|\mathbf{s}_\mathrm{g} - \mathbf{s}_\mathrm{s}\|_2^2 \\ &+ \frac{1}{\sigma_{p,i}^2}\sum_i \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}\|_2^2 \Bigg\} \text{ subject to } \|\mathbf{x}_i\|_0 = T \;\forall\, i.\end{aligned} \tag{5.5}$$

5.3.1 Solving for the MAP estimate

We find the MAP estimates $\big\{\widehat{\mathbf{s}}_\mathrm{g}, \widehat{\mathbf{s}}_\mathrm{s}, \widehat{\mathbf{X}}\big\}$ solving (S3) via block-coordinate minimization, similar to [22, 26]. The global objective is written from (S3)

$$\widehat{\mathbf{s}}_\mathrm{g} = \underset{\mathbf{s}_\mathrm{g}}{\arg\min}\; \|\mathbf{t} - \mathbf{A}\mathbf{s}_\mathrm{g}\|_2^2 + \lambda_1\|\mathbf{s}_\mathrm{g} - \mathbf{s}_\mathrm{s}\|_2^2, \tag{5.6}$$

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

The local objective from (S3) for each patch is solved with $\mathbf{s}_s = \widehat{\mathbf{s}}_g$ (decoupling the local and global objectives), giving

$$\widehat{\mathbf{x}}_i = \arg\min_{\mathbf{x}_i} \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\widehat{\mathbf{s}}_g\|_2^2 \text{ subject to } \|\mathbf{x}_i\|_0 = T. \tag{5.7}$$

With $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_I]$ from (S5) and $\widehat{\mathbf{s}}_g$ from (S4), we find $\widehat{\mathbf{s}}_s$ from (S3), assuming $\sigma_{p,i}^2 = \sigma_p^2$

$$\widehat{\mathbf{s}}_s = \arg\min_{\mathbf{s}_s} \lambda_2\|\widehat{\mathbf{s}}_g - \mathbf{s}_s\|_2^2 + \sum_i \|\mathbf{D}\widehat{\mathbf{x}}_i - \mathbf{R}_i\mathbf{s}_s\|_2^2, \tag{5.8}$$

where $\lambda_2 = (\sigma_p/\sigma_g)^2$ is a regularization parameter. The estimate $\widehat{\mathbf{s}}_s$ is obtained analytically from (S7) by

$$\widehat{\mathbf{s}}_s = \left(\lambda_2\mathbf{I} + \sum_i \mathbf{R}_i^T\mathbf{R}_i\right)^{-1}\left(\lambda_2\widehat{\mathbf{s}}_g + \sum_i \mathbf{R}_i^T\mathbf{D}\widehat{\mathbf{x}}_i\right), \tag{5.9}$$

which averages the pixels from the patch estimates $\{\mathbf{D}\widehat{\mathbf{x}}_i \; \forall \; i\}$, with weight given to $\widehat{\mathbf{s}}_g$ by λ_2. From (5.9), the average patch slowness is $\mathbf{s}_p = \left(\sum_i \mathbf{R}_i^T\mathbf{R}_i\right)^{-1}\left(\sum_i \mathbf{R}_i^T\mathbf{D}\widehat{\mathbf{x}}_i\right)$, with $\mathbf{b} = \text{diag}\left(\sum_i \mathbf{R}_i^T\mathbf{R}_i\right) \in \mathbb{Z}^N$ the number of patches per pixel. Hence, (5.9) is expressed as an operation at pixel n by

$$\widehat{s}_{s,n} = \frac{\lambda_2\widehat{s}_{g,n} + b_n s_{p,n}}{\lambda_2 + b_n}. \tag{5.10}$$

5.3.2 LST algorithm with dictionary learning

The results (S4), (S5), and (5.9) give the LST algorithm for estimating $\mathbf{s}_s$, shown in Table 6.1, as a MAP estimate with local sparse priors using a prescribed dictionary $\mathbf{D}$. Dictionary learning via the ITKM [20] is added to the LST in the solution to the local objective (S5). The

Table 5.1: Sparse travel time tomography (LST) algorithm with fixed or adaptive dictionaries

Given: $\mathbf{t} \in \mathbb{R}^M$, $\mathbf{A} \in \mathbb{R}^{M \times N}$, $\mathbf{s}_{\mathrm{s}}^0 = \mathbf{0} \in \mathbb{R}^N$, $\mathbf{D}^0 = \text{Haar}$, DCT (or) noise $\mathcal{N}\big(0,1\big) \in \mathbb{R}^{n \times Q}$, λ_1, λ_2, T, and $j = 1$

Repeat until convergence:

1. Global estimate: solve (S4) using LSQR [28],
$$\widehat{\mathbf{s}}_{\mathrm{g}}^j = \arg\min_{\mathbf{s}_{\mathrm{g}}^j} \|\mathbf{A}\mathbf{s}_{\mathrm{g}}^j - \mathbf{t}\|_2^2 + \lambda_1 \|\mathbf{s}_{\mathrm{g}}^j - \mathbf{s}_{\mathrm{s}}^{j-1}\|_2^2.$$
2. Local estimate
   a: Setting $\mathbf{s}_{\mathrm{s}}^j = \widehat{\mathbf{s}}_{\mathrm{g}}^j$, center patches $\{\mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}^j \; \forall \; i \}$ and
      i. (Dictionary learning) Find $\mathbf{D}^j$ using ITKM [20].
      i. (Prescribed dictionary) Set $\mathbf{D}^j = \mathbf{D}^0$.
      ii. Solve (S5) using OMP,
$$\widehat{\mathbf{x}}_i^j = \arg\min_{\mathbf{x}_i^j} \|\mathbf{D}^j \mathbf{x}_i^j - \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}^j\|_2^2 \;\; \text{subject to } \|\mathbf{x}_i^j\|_0 = T.$$
   b: Obtain $\widehat{\mathbf{s}}_{\mathrm{s}}^j$ by (S8) as
$$\widehat{s}_{\mathrm{s,n}}^j = \frac{\lambda_2 \widehat{s}_{\mathrm{g,n}}^j + b_{\mathrm{n}} s_{p,\mathrm{n}}^j}{\lambda_2 + b_{\mathrm{n}}}$$
   $j = j + 1$

global objective (S4) is solved using the sparse least squares program LSQR [28]. The local objective (S5) is solved using OMP after the slowness patches $\{\mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}} \; \forall \; i \}$ are centered [11].

The complexity of each LST iteration is determined primarily by LSQR computation in the global estimate, $O(2MN)$, and by ITKM $O(knQI)$ and OMP $O(TnQI)$ in the local estimate, where k is the ITKM iterations (see Table 6.1). For large slowness maps, we expect the LST complexity to be dominated by LSQR. In our simulations we obtain reasonable run times (see Sec. 5.4.1).

5.3.3 Conventional tomography

We illustrate conventional tomography with a Bayesian approach [29], which enforces smoothness regularization with a global (non-diagonal) covariance. Considering the measure-

ments (S1), the MAP estimate of the slowness is

$$\widehat{\mathbf{s}}_{\mathrm{g}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \eta\boldsymbol{\Sigma}_L^{-1}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{t}, \tag{5.11}$$

where $\eta = (\sigma_\varepsilon/\sigma_{\mathrm{c}})^2$ is a regularization parameter, σ_{c} is the conventional slowness variance, and smoothness $\boldsymbol{\Sigma}_L(i,j) = \exp\left(-D_{i,j}/L\right)$. Here, $D_{i,j}$ is the distance between cells i and j, and L is the length scale [29, 30].

5.4 Simulation

We demonstrate the performance of LST (Sec. 6.3, Table 6.1) relative to a conventional tomography (Sec. 6.4.1). Experiments are conducted using simulated travel times from two synthetic 2D slowness maps (Fig. 6.1(c,d)) with dimensions $W_1 = W_2 = 100$ pixels (km). The checkerboard pattern (Fig. 6.1(c)) contains only discontinuous slowness whereas the smooth-discontinuous map (Fig. 6.1(d)) contains a fault-like discontinuity in a smooth map. The slowness estimates from LST are plotted as $\widehat{\mathbf{s}}'_{\mathrm{s}} = \widehat{\mathbf{s}}_{\mathrm{s}} + \mathbf{s}_0 \in \mathbb{R}^N$, and for conventional $\widehat{\mathbf{s}}'_{\mathrm{g}} = \widehat{\mathbf{s}}_{\mathrm{g}} + \mathbf{s}_0 \in \mathbb{R}^N$.

The slownesses are sampled by $M = 2016$ straight-rays between 64 seismic stations (see Fig. 6.1(e)). The travel time vector $\mathbf{t}$ is found by integrating along these ray paths. We consider only straight ray propagation, to focus on our proposed inversion approach. The reference slowness is calculated from the mean travel time using the the tomography matrix $\mathbf{A}$. The LST inversion valid-region is obtained with a dilation operation with a patch template along the outermost ray paths. The conventional valid region is the outermost pixels along the ray paths. The conventional valid region is used for error calculations for both methods.

We consider the noise-free case ($\sigma_\varepsilon = 0$). The results of the LST and conventional tomography are shown in Figs. 6.5 and 6.6. RMSE (s/km) of the estimates $\widehat{\mathbf{s}}'_{\mathrm{s}}$ and $\widehat{\mathbf{s}}'_{\mathrm{s}}$ relative to the true slowness $\mathbf{s}'$ is printed on the 2D estimates. We invert using LST with and without dictionary learning. We consider two prescribed dictionaries $\mathbf{D}$, overcomplete Haar wavelet and

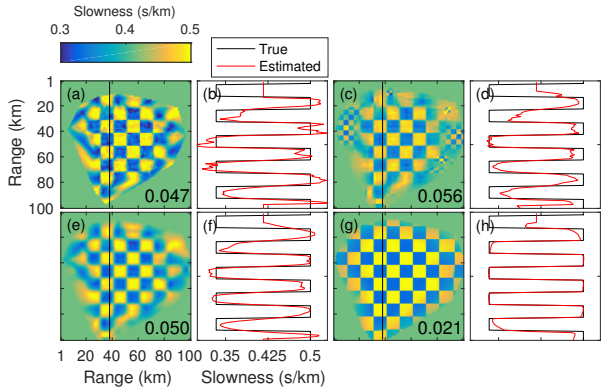


Figure 5.2: LST and conventional tomography for checkerboard map (Fig. 6.1(c)). 2D and 1D (from black line in 2D) slowness estimates against true slowness for: (a,b) conventional $\widehat{\mathbf{s}}'_{\mathrm{g}}$; LST $\widehat{\mathbf{s}}'_{\mathrm{s}}$ with (c,d) Haar dictionary and (e,f) with DCT dictionary $\mathbf{D}$; and (g,h) with dictionary learning. RMSE (s/km) is printed on each 2D image.

DCT dictionaries (both $Q = 169$, $n = 64$, since Haar wavelet dimensions power of 2).

5.4.1 Inversion parameters and results

The regularization parameter values for LST and conventional tomography were selected to minimize RMSE (s/km). For LST, the best parameters were: for both prescribed dictionary and dictionary learning, $\lambda_1 = 0$ km^2 (in (S4)) and $\lambda_2 = 0$ (in (S7)); for prescribed dictionaries $T = 2$ non-zero coefficients in (S5); for dictionary learning, $T = 1$, $n = 100$, and for the checkerboard (smooth-discontinuous) $Q = 166$ ($Q = 268$ atoms). Since for the noise free case $\sigma_\varepsilon = 0$, we expect $\lambda_1 = (\sigma_\varepsilon/\sigma_{\mathrm{g}})^2 = 0$ km^2 to be best. We assume the slowness patches are well approximated by the sparse model (S5), and expect $\sigma_p \ll \sigma_g$. Hence, we expect the best value of $\lambda_2 = (\sigma_p/\sigma_{\mathrm{g}})^2$ in (S7) to be small. For conventional tomography (Sec. 6.4.1), the best parameters were $L = 10$ km and $\eta = 0.1$ km^2 (in (5.11)) for the both the checkerboard and smooth-discontinuous maps, which deviates from the expected value of $\eta = (\sigma_\varepsilon/\sigma_{\mathrm{c}})^2 = 0$.

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 ray sampling where sampling is poor (near the edges of the inversion, Fig. 6.5(c,d)). The

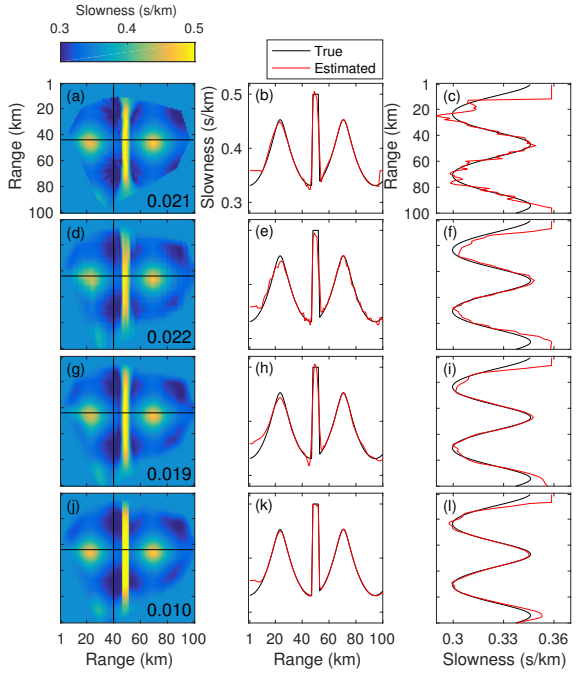


Figure 5.3: LST and conventional tomography for smooth-discontinuous map (Fig. 6.1(d)). 2D and 1D (horizontal and vertical, from black lines in 2D) slowness estimates against true slowness for: (a–c) conventional $\widehat{\mathbf{s}}'_g$; LST $\widehat{\mathbf{s}}'_s$ with (d–f) Haar dictionary and (g–i) with DCT dictionary $\mathbf{D}$; and (j–l) with dictionary learning. RMSE (s/km) is printed on each 2D image.

performance of the Haar wavelets is better for the smooth-discontinuous slowness map (Fig. 6.6(d–f)) than for the checkerboard. As shown in Fig. 6.6(d–f), 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. 6.5(e,f) and Fig. 6.6(g,j)) 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. 6.5(g,h) and Fig. 6.6(j–l)) achieves the best RMSE relative to the true slowness $\mathbf{s}'$. As in the other cases, the performance degrades near the edges of the ray sampling, where the ray coverage is poor, but high resolution is maintained across a large part of the sampling region. The RMSE of the Haar wavelet inversion for the checkerboard is greater than for the conventional method, although resolution is lost in the conventional MAP

inversion near the more densely sampled region of the wave speed maps. The RMSE for the DCT is less than the that of the Haar wavelets and also the conventional MAP inversion. A better qualitative fit to the true slowness is also observed.

The LST algorithm (Table 6.1) used 100 iterations for all cases and the ITKM used 50 iterations. In Matlab, the inversion with dictionary learning took ~ 5 min on a Macbook Pro 2.5 GHz Intel Core i7.

5.5 Conclusions

We derived a travel time tomography method which incorporates a sparse prior on patches of the slowness image, which we refer to as the LST algorithm. The LST uses prescribed or learned dictionaries, though the learned dictionaries improve performance. The local sparse prior and dictionary learning provide an improved slowness model, which is capable of modeling simultaneously smooth discontinuous features.

We considered 2D surface wave tomography, and for densely sampled slowness maps obtained superior results from the LST over conventional tomography. The LST is relevant to other tomography scenarios where slowness structure is irregularly sampled, in for instance ocean [16] and terrestrial [31] acoustics.

5.6 Acknowledgments

This work is supported by the Office of Naval Research, Grant No. N00014-11-1-0439.

Chapter 5, in full, is a reprint of the material as it appears in the Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech, and Signal Processing 2018. Bianco, Michael; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2018. The dissertation author was the primary investigator and author of this paper.

5.7 Bibliography

[1] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter estimation and inverse problems*, 2nd ed. Elsevier, San Diego, 2013.

[2] N. Rawlinson, S. Pozgay, and S. Fishwick, "Seismic tomography: a window into deep earth," *Phys. Earth Plan. Inter.*, vol. 178, no. 3, pp. 101–135, 2010.

[3] L. Y. Chiao and B. Y. Kuo, "Multiscale seismic tomography," *Geo. J. Int.*, vol. 145, no. 2, pp. 517–527, 2001.

[4] I. Loris, G. Nolet, I. Daubechies, and F. A. Dahlen, "Tomographic inversion using ℓ_1–norm regularization of wavelet coefficients," *Geo. J. Int.*, vol. 170, no. 1, pp. 359–370, 2007.

[5] R. Hawkins and M. Sambridge, "Geophysical imaging using trans-dimensional trees," *Geo. J. Int.*, vol. 203, no. 2, pp. 972–1000, 2015.

[6] H. Fang and H. Zhang, "Wavelet-based double-difference seismic tomography with sparsity regularization," *Geo. J. Int.*, vol. 199, no. 2, pp. 944–955, 2014.

[7] Y. Lin and L. Huang, "Quantifying subsurface geophysical properties changes using double-difference seismic-waveform inversion with a modified total-variation regularization scheme," *Geo. J. Int.*, vol. 203, no. 3, pp. 2125–2149, 2015.

[8] X. Zhang and J. Zhang, "Model regularization for seismic traveltime tomography with an edge-preserving smoothing operator," *J. Appl. Geo.*, vol. 138, pp. 143–153, 2017.

[9] T. Bodin, M. Sambridge, and K. Gallagher, "A self-parametrizing partition model approach to tomographic inverse problems," *Inv. Prob.*, vol. 25, no. 5, 2009.

[10] E. Candés, "Compressive sampling," *Proc. Internat. Cong. Math.*, vol. 3, pp. 1433–1452, 2006.

[11] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014.

[12] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[13] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 260–271, 2014.

[14] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," *IEEE Sig. Proc. Let.*, vol. 23, no. 10, pp. 1469–1473, 2016.

[15] T. Wang and W. Xu, "Sparsity-based approach for ocean acoustic tomography using learned dictionaries," *OCEANS 2016 Shanghai IEEE*, pp. 1–6, 2016.

[16] M. Bianco and P. Gerstoft, "Compressive acoustic sound speed profile estimation," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. EL90–EL94, 2016.

[17] ——, "Dictionary learning of sound speed profiles," *J. Acoust. Soc. Am*, vol. 141, no. 3, pp. 1749–1758, 2017.

[18] ——, "Regularization of geophysical inversion using dictionary learning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[19] J. Charléty, S. Voronin, G. Nolet, I. Loris, F. J. Simons, K. Sigloch, and I. C. Daubechies, "Global seismic tomography with sparsity constraints: comparison with smoothing and damping regularization," *J. Geophys. Research: Solid Earth*, vol. 118, no. 9, pp. 4887–4899, 2013.

[20] K. Schnass, "Local identification of overcomplete dictionaries," *J. Mach. Learn. Res.*, no. 16, pp. 1211–1242, 2015.

[21] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: the application of compressed sensing for rapid MR imaging," *Mag. Res. Medic.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[22] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1028–1041, 2011.

[23] Y. Chen, "Fast dictionary learning for noise attenuation of multidimensional seismic data," *Geo. J. Int.*, vol. 209, no. 1, pp. 21–31, 2017.

[24] M. Taroudakis and C. Smaragdakis, "De-noising procedures for inverting underwater acoustic signals in applications of acoustical oceanography," *EuroNoise 2015 Maastricht*, pp. 1393–1398, 2015.

[25] L. Zhu, E. Liu, and J. H. McClellan, "Sparse-promoting full waveform inversion based on online orthonormal dictionary learning," *Geophys.*, vol. 82, no. 2, pp. R87–R107, 2017.

[26] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[27] M. Bianco and P. Gerstoft, "Travel time tomography with adaptive dictionaries," *IEEE Trans. Comput. Imag.*, May 2018, dOI: 10.1109/TCI.2018.2862644.

[28] C. C. Paige and M. A. Saunders, "LSQR: an algorithm for sparse linear equations and sparse least squares," *ACM Trans. Math. Software*, vol. 8, no. 1, pp. 43–71, 1982.

[29] C. D. Rodgers, *Inverse methods for atmospheric sounding: theory and practice*. World Sci. Pub. Co., 2000.

[30] A. Tarantola, *Inverse problem theory*. Elsevier Sci. Pub. Co., Inc., 1987.

[31] R. Rabenstein and P. Annibale, "Acoustic source localization under variable speed of sound conditions," *Wirel. Comm. Mob. Comp.*, 2017.

Chapter 6

Travel time tomography with adaptive dictionaries

6.1 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 ℓ_2 regularization.

This local-global modeling strategy with dictionary learning has been successful in image denoising [3, 7, 8] 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 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 ℓ_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].

6.2 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. 6.1. 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, $\mathbf{t}' \in \mathbb{R}^M$, are obtained. We assume $\mathbf{t}'$ 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. 6.1) from $\mathbf{t}'$.

In the following, we develop separately two slowness models, deemed the *global* and *local* models, which will be related in Section 6.3, and discuss dictionaries for sparse modeling. The global model considers the larger scale or global features and relates travel times to slowness.

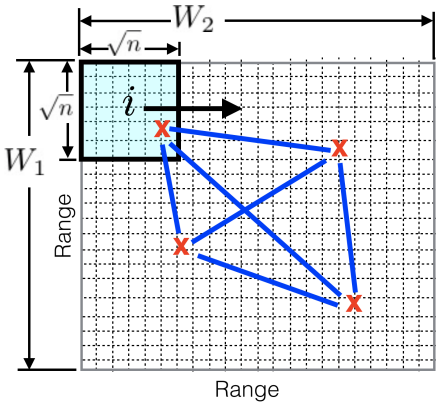


Figure 6.1: 2D slowness image corresponds to slowness map divided into pixels (dashed boxes). The square image patch i contains n pixels. W_1 and W_2 are the vertical and horizontal dimensions of the image in pixels. Sensors are red x's and the ray paths between the sensors are blue lines.

The local model considers smaller scale, localized features with sparse modeling.

6.2.1 Global model and travel times

In the global model, slowness pixels (see Fig. 6.1) are represented by the vector $\mathbf{s}' = \mathbf{s}_g + \mathbf{s}_0 \in \mathbb{R}^N$, where $\mathbf{s}_0$ is reference slownesses and $\mathbf{s}_g$ 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 $\mathbf{t}' = \mathbf{t} + \mathbf{t}_0$, where $\mathbf{t}$ is the travel time perturbation and $\mathbf{t}_0$ is the reference travel time. The tomography matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ gives the discrete path lengths of M straight-rays through N pixels (see Fig. 6.1). Thus $\mathbf{t}$ and $\mathbf{s}_g$ are related by the linear measurement model

$$\mathbf{t} = \mathbf{A}\mathbf{s}_g + \boldsymbol{\varepsilon}, \tag{6.1}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ is Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$, with mean $\mathbf{0}$ and covariance $\sigma_\varepsilon^2 \mathbf{I}$. We estimate the perturbations, with $\mathbf{s}_0$ and $\mathbf{t}_0 = \mathbf{A}\mathbf{s}_0$ known. We call (S1) the *global model*, as it captures the large-scale features that span the slowness map and generates $\mathbf{t}$.

6.2.2 Local model and sparsity

In the local model, slowness pixels (see Fig. 6.1) are represented by the vector $\mathbf{s}' = \mathbf{s}_\mathrm{s} + \mathbf{s}_0 \in \mathbb{R}^N$, where the *sparse slowness* $\mathbf{s}_\mathrm{s}$ is perturbations from the reference $\mathbf{s}_0$. The slowness $\mathbf{s}_\mathrm{s}$ is an auxiliary latent variable that is introduced to capture local slowness behavior, and is instrumental in the estimation procedure proposed in Sec. 6.3. In Sec. 6.3.1, $\mathbf{s}_\mathrm{s}$ is precisely related to $\mathbf{s}_\mathrm{g}$ in a Bayesian hierarchy.

Formulating the local model, we assume that *patches*, or $\sqrt{n} \times \sqrt{n}$ groups of pixels from $\mathbf{s}_\mathrm{s}$ (see Fig. 6.1) are well approximated by a sparse linear combination of atoms from a dictionary $\mathbf{D} \in \mathbb{R}^{n \times Q}$ of Q atoms. The patches are selected from $\mathbf{s}_\mathrm{s}$ by the binary matrix $\mathbf{R}_i \in \{0,1\}^{n \times N}$. Hence the slownesses in patch i are $\mathbf{R}_i \mathbf{s}_\mathrm{s}$. The sparse model is

$$\mathbf{R}_i \mathbf{s}_\mathrm{s} \approx \mathbf{D}\mathbf{x}_i \text{ and } |\mathbf{x}_i \neq 0| = T \;\forall\; i \tag{6.2}$$

where $|\cdot|$ is cardinality, $\mathbf{x}_i \in \mathbb{R}^n$ is the sparse coefficients, and $T \ll n$ is the number of non-zero coefficients. $\mathbf{D}\mathbf{x}_i$ is referred to as the *patch slowness*. We call (S2) the *local model*, as it models smaller scale, localized features contained by patches.

Each slowness patch $\mathbf{R}_i \mathbf{s}_\mathrm{s}$ is indexed by the row w_1 and column w_2 of its top-left pixel in the 2D image as $(w_{1,i}, w_{2,i})$. We consider all overlapping patches, with $w_{1,i}$ and $w_{2,i}$ differing from their neighbor by ± 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 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. 6.5). 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 $\mathbf{D}$ are considered "elemental patches", where only a small number of atoms are necessary to adequately approximate $\mathbf{R}_i \mathbf{s}_\mathrm{s}$. Atoms can be generic functions, e.g. wavelets or the discrete cosine transform (DCT), or learned from the data (see Sec. 6.3.3). An example of DCT

atoms are shown in Fig. 6.2. 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. 6.3) are shown in Fig. 6.4. Relative to generic dictionaries, learned dictionaries can represent the smooth and discontinuous seismic features encountered in real inversion scenarios.

6.3 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 $\mathbf{s}_\mathrm{s}$ (from (S2)). Assuming the travel times $\mathbf{t}$, tomography matrix $\mathbf{A}$, and dictionary $\mathbf{D}$ known, the solution to the objective gives MAP estimates of the global slowness $\mathbf{s}_\mathrm{g}$ (from (S1)), $\mathbf{s}_\mathrm{s}$, and the coefficients $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_I] \in \mathbb{R}^{Q \times I}$ describing all patches I (from (S2)). Since we use a non-Bayesian dictionary learning algorithm (ITKM [6]), dictionary learning is added after the MAP derivation in Sec. 6.3.3.

6.3.1 Derivation of MAP objective

Starting Bayes' rule, we combine the global (S1) and local (S2) models, and formulate the posterior density of the LST variables as

$$p\big(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X} \big| \mathbf{t}\big) \propto p\big(\mathbf{t} \big| \mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big)\, p\big(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big). \tag{6.3}$$

From (S2), $\mathbf{s}_\mathrm{s}$ is conditioned only on $\mathbf{X}$. We further assume the patch coefficients $\mathbf{X}$ independent. Hence, using the chain rule we obtain from (6.3)

$$\begin{aligned} p\big(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big|\mathbf{t}\big) &\propto p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big) p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}, \mathbf{X}\big) p\big(\mathbf{s}_\mathrm{s}, \mathbf{X}\big) \\ &\propto p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big) p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}, \mathbf{X}\big) p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big) p\big(\mathbf{X}\big), \end{aligned} \tag{6.4}$$

From (S1), $\mathbf{t}$ is conditioned only on $\mathbf{s}_\mathrm{g}$, and we assume $\mathbf{s}_\mathrm{g}$ is conditioned only on $\mathbf{s}_\mathrm{s}$. Hence, we obtain from (6.4)

$$p\big(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big|\mathbf{t}\big) \propto p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big) p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big) p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big) p\big(\mathbf{X}\big). \tag{6.5}$$

We approximate $p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big)$, $p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big)$, $p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big)$ as Gaussian, and all patch slownesses from (S2) independent, giving

$$\begin{aligned} p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big) &= \mathcal{N}(\mathbf{A}\mathbf{s}_\mathrm{g}, \boldsymbol{\Sigma}_\varepsilon), \\ p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big) &= \mathcal{N}(\mathbf{s}_\mathrm{s}, \boldsymbol{\Sigma}_\mathrm{g}), \\ p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big) &= \prod_i p\big(\mathbf{R}_i \mathbf{s}_\mathrm{s}\big|\mathbf{x}_i\big) = \prod_i \mathcal{N}\big(\mathbf{D}\mathbf{x}_i, \boldsymbol{\Sigma}_{p,i}\big), \end{aligned} \tag{6.6}$$

where $\boldsymbol{\Sigma}_\varepsilon \in \mathbb{R}^{K\times K}$ is the covariance of the travel time error, $\boldsymbol{\Sigma}_\mathrm{g} \in \mathbb{R}^{N\times N}$ is the covariance of $\mathbf{s}_\mathrm{g}$, and $\boldsymbol{\Sigma}_{p,i} \in \mathbb{R}^{n\times n}$ is the covariance of the patch slownesses. Taking the logarithm of the conditional probabilities from (6.6), we obtain

$$\begin{aligned} \ln p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big) &\propto -\frac{1}{2}(\mathbf{t} - \mathbf{A}\mathbf{s}_\mathrm{g})^\mathrm{T} \boldsymbol{\Sigma}_\varepsilon^{-1} (\mathbf{t} - \mathbf{A}\mathbf{s}_\mathrm{g}), \\ \ln p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big) &\propto -\frac{1}{2}(\mathbf{s}_\mathrm{g} - \mathbf{s}_\mathrm{s})^\mathrm{T} \boldsymbol{\Sigma}_\mathrm{g}^{-1} (\mathbf{s}_\mathrm{g} - \mathbf{s}_\mathrm{s}), \\ \ln p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big) &\propto -\frac{1}{2}\sum_i (\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s})^\mathrm{T} \boldsymbol{\Sigma}_{p,i}^{-1} (\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}). \end{aligned} \tag{6.7}$$

Since neighboring patches may share slowness features, $\mathbf{x}_i$ may be correlated. For

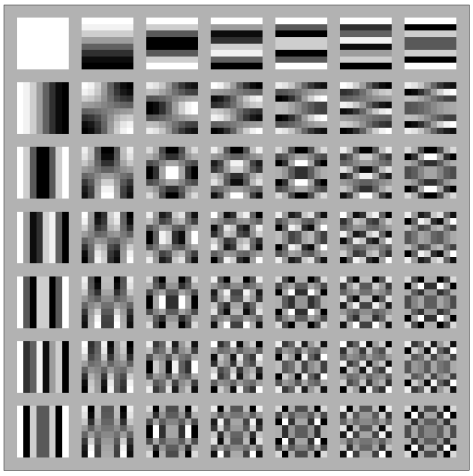


Figure 6.2: Discrete cosine transform (DCT) dictionary atoms. The $Q = 49$ atoms (ordered in a 7×7 grid) each contain $n = 8 \times 8 = 64$ pixels. Atom values stretched to full grayscale range for display.

simplicity, the coefficients $\mathbf{x}_i$ describing the patches are assumed independent,

$$\ln p\big(\mathbf{X}\big) = \sum_i \ln p\big(\mathbf{x}_i\big). \tag{6.8}$$

From (6.5), and (6.7), and (6.8) we obtain

$$\begin{aligned} \ln p\big(\mathbf{s}_\mathrm{g}, \mathbf{s}_\mathrm{s}, \mathbf{X}\big|\mathbf{t}\big) &\propto \ \ln\big\{p\big(\mathbf{t}\big|\mathbf{s}_\mathrm{g}\big)p\big(\mathbf{s}_\mathrm{g}\big|\mathbf{s}_\mathrm{s}\big)p\big(\mathbf{s}_\mathrm{s}\big|\mathbf{X}\big)p\big(\mathbf{X}\big)\big\} \\ &\propto -(\mathbf{t}-\mathbf{A}\mathbf{s}_\mathrm{g})^\mathrm{T}\boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{t}-\mathbf{A}\mathbf{s}_\mathrm{g}) - (\mathbf{s}_\mathrm{g}-\mathbf{s}_s)^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{g}^{-1}(\mathbf{s}_\mathrm{g}-\mathbf{s}_\mathrm{s}) \\ &\quad - \sum_i \Big\{ (\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s})^\mathrm{T}\boldsymbol{\Sigma}_{p,i}^{-1}(\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}) + 2\ln p\big(\mathbf{x}_i\big) \Big\}. \end{aligned} \tag{6.9}$$

Assuming the coefficients $\mathbf{x}_i$ sparse, we approximate $\ln p\big(\mathbf{x}_i\big)$ with the ℓ_0 pseudo-norm $\|\mathbf{x}_i\|_0$, which counts the number of non-zero coefficients [3]. We further assume the number of non-zero

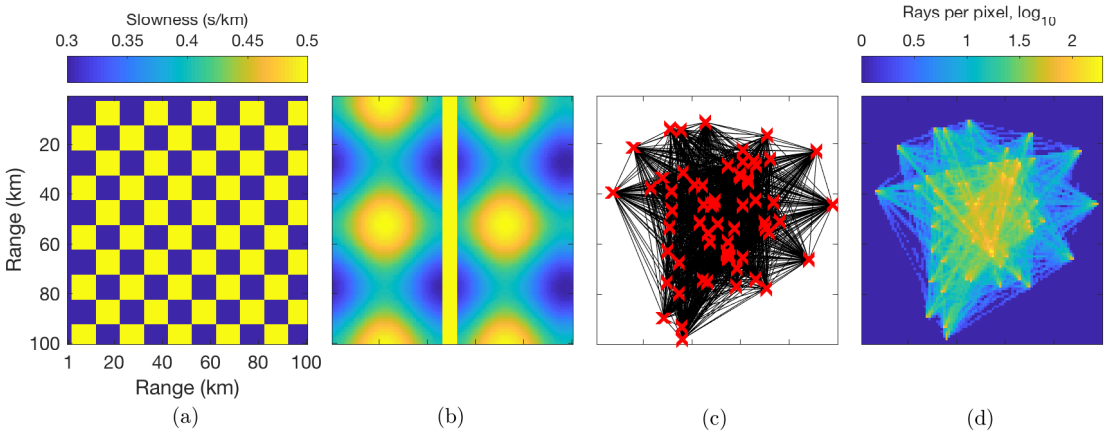


Figure 6.3: Synthetic slowness maps and ray sampling. Slowness $\mathbf{s}'$ for (a) checkerboard map and (b) smooth-discontinuous map (sinusoidal variations with discontinuity). Both maps are $W_1 = W_2 = 100$ pixels (1 km/pixel). (c) 64 stations (red X's), giving in 2016 straight ray (surface wave) paths through synthetic images. (d) Density of ray sampling, in $\log_{10}$ rays per pixel.

coefficients T is the same for every patch. This gives the MAP objective as

$$\begin{aligned}
&\max\left\{\ln p\left(\mathbf{s}_\mathrm{g},\mathbf{s}_\mathrm{s},\mathbf{X}\middle|\mathbf{t}\right)\right\} = \min\left\{-\ln p\left(\mathbf{s}_\mathrm{g},\mathbf{s}_\mathrm{s},\mathbf{X}\middle|\mathbf{t}\right)\right\} \\
&\propto \min\Bigg\{(\mathbf{t}-\mathbf{A}\mathbf{s}_\mathrm{g})^\mathrm{T}\boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{t}-\mathbf{A}\mathbf{s}_\mathrm{g}) + (\mathbf{s}_\mathrm{g}-\mathbf{s}_\mathrm{s})^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{g}^{-1}(\mathbf{s}_\mathrm{g}-\mathbf{s}_\mathrm{s}) \\
&\quad + \sum_i (\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s})^\mathrm{T}\boldsymbol{\Sigma}_{p,i}^{-1}(\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s})\Bigg\} \\
&\qquad \text{subject to } \|\mathbf{x}_i\|_0 = T \;\forall\, i.
\end{aligned} \tag{6.10}$$

Further simplifying, we assume the errors are Gaussian iid. Thus, $\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}$, $\boldsymbol{\Sigma}_\mathrm{g} = \sigma_\mathrm{g}^2\mathbf{I}$, and $\boldsymbol{\Sigma}_{p,i} = \sigma_{p,i}^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The LST MAP objective is thus

$$\begin{aligned}
\left\{\widehat{\mathbf{s}}_\mathrm{g},\widehat{\mathbf{s}}_\mathrm{s},\widehat{\mathbf{X}}\right\} = \underset{\mathbf{s}_\mathrm{g},\mathbf{s}_\mathrm{s},\mathbf{X}}{\arg\min}\Bigg\{&\frac{1}{\sigma_\varepsilon^2}\|\mathbf{t}-\mathbf{A}\mathbf{s}_\mathrm{g}\|_2^2 \\
&+ \frac{1}{\sigma_\mathrm{g}^2}\|\mathbf{s}_\mathrm{g}-\mathbf{s}_\mathrm{s}\|_2^2 + \sum_i \frac{1}{\sigma_{p,i}^2}\|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}\|_2^2\Bigg\} \\
&\text{subject to } \|\mathbf{x}_i\|_0 = T \;\forall\, i,
\end{aligned} \tag{6.11}$$

where $\left\{\widehat{\mathbf{s}}_\mathrm{g},\widehat{\mathbf{s}}_\mathrm{s},\widehat{\mathbf{X}}\right\}$ are the estimates of the LST variables.

6.3.2 Solving for the MAP estimate

We find the MAP estimates $\left\{\widehat{\mathbf{s}}_{\mathrm{g}}, \widehat{\mathbf{s}}_{\mathrm{s}}, \widehat{\mathbf{X}}\right\}$ solving (S3) via an alternating minimization algorithm [7, 23]. This strategy divides the solution of (S3) into three subproblems: 1) the global problem, corresponding to the global model (S1), which estimates $\widehat{\mathbf{s}}_{\mathrm{g}}$; 2) the local problem, corresponding to the local model (S2) which estimates $\widehat{\mathbf{X}}$; and 3) an averaging procedure which estimates $\widehat{\mathbf{s}}_{\mathrm{s}}$. The global problem is least squares with ℓ_2 regularization. The global result is substituted into the sparse model (S2) to obtain local structure by solving for the patches, and averaging their estimates. Global ℓ_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 6.4.2. LST combines a global ℓ_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].

The global problem is written from (S3) as

$$\begin{aligned}\widehat{\mathbf{s}}_{\mathrm{g}} =& \underset{\mathbf{s}_{\mathrm{g}}}{\arg\min}\ \frac{1}{\sigma_{\varepsilon}^2}\|\mathbf{t}-\mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2+\frac{1}{\sigma_{\mathrm{g}}^2}\|\mathbf{s}_{\mathrm{g}}-\mathbf{s}_{\mathrm{s}}\|_2^2 \\ =& \underset{\mathbf{s}_{\mathrm{g}}}{\arg\min}\ \|\mathbf{t}-\mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2+\lambda_1\|\mathbf{s}_{\mathrm{g}}-\mathbf{s}_{\mathrm{s}}\|_2^2,\end{aligned} \tag{6.12}$$

where $\lambda_1 = (\sigma_{\varepsilon}/\sigma_{\mathrm{g}})^2$ is a regularization parameter.

The local problem is written from (S3), with each patch solved from the global estimate $\mathbf{s}_{\mathrm{s}} = \widehat{\mathbf{s}}_{\mathrm{g}}$ from (S4) (decoupling the local and global problems), giving

$$\widehat{\mathbf{x}}_i = \underset{\mathbf{x}_i}{\arg\min}\ \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i\widehat{\mathbf{s}}_{\mathrm{g}}\|_2^2 \ \text{ subject to } \|\mathbf{x}_i\|_0 = T. \tag{6.13}$$

With the coefficients estimate $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_I]$ from (S5) and global slowness $\widehat{\mathbf{s}}_{\mathrm{g}}$ from (S4) we

solve for $\mathbf{s}_\mathrm{s}$ from (S3), assuming constant patch variance $\sigma_{p,i}^2 = \sigma_p^2$,

$$\begin{aligned}\widehat{\mathbf{s}}_\mathrm{s} =& \arg\min_{\mathbf{s}_\mathrm{s}} \frac{1}{\sigma_\mathrm{g}^2}\|\widehat{\mathbf{s}}_\mathrm{g} - \mathbf{s}_\mathrm{s}\|_2^2 + \frac{1}{\sigma_p^2}\sum_i \|\mathbf{D}\widehat{\mathbf{x}}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}\|_2^2 \\ =& \arg\min_{\mathbf{s}_s} \lambda_2\|\widehat{\mathbf{s}}_\mathrm{g} - \mathbf{s}_\mathrm{s}\|_2^2 + \sum_i \|\mathbf{D}\widehat{\mathbf{x}}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}\|_2^2,\end{aligned} \tag{6.14}$$

where $\lambda_2 = (\sigma_p/\sigma_\mathrm{g})^2$ is a regularization parameter. The solution to (S7) is analytic, with $\widehat{\mathbf{s}}_\mathrm{s}$ the stationary point. Differentiating (S7) gives

$$\begin{aligned}&\frac{d}{d\mathbf{s}_\mathrm{s}}\left\{\lambda_2\|\widehat{\mathbf{s}}_\mathrm{g} - \mathbf{s}_\mathrm{s}\|_2^2 + \sum_i \|\mathbf{D}\widehat{\mathbf{x}}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}\|_2^2\right\} \\ &= \lambda_2\big(\mathbf{s}_\mathrm{s} - \widehat{\mathbf{s}}_\mathrm{g}\big) + \sum_i \mathbf{R}_i^\mathrm{T}\big(\mathbf{R}_i\mathbf{s}_\mathrm{s} - \mathbf{D}\widehat{\mathbf{x}}_i\big) \\ &= \big(\lambda_2\mathbf{I} + n\mathbf{I}\big)\mathbf{s}_\mathrm{s} - \lambda_2\widehat{\mathbf{s}}_\mathrm{g} - \mathbf{s}_p,\end{aligned} \tag{6.15}$$

where $n\mathbf{I} = \sum_i \mathbf{R}_i^\mathrm{T}\mathbf{R}_i$ and $\mathbf{s}_p = \frac{1}{n}\sum_i \mathbf{R}_i^\mathrm{T}\mathbf{D}\widehat{\mathbf{x}}_i$. Thus,

$$\widehat{\mathbf{s}}_\mathrm{s} = \frac{\lambda_2\widehat{\mathbf{s}}_\mathrm{g} + n\mathbf{s}_p}{\lambda_2 + n}, \tag{6.16}$$

which gives $\mathbf{s}_\mathrm{s}$ as the weighted average of the patch slownesses $\{\mathbf{D}\widehat{\mathbf{x}}_i \;\forall\; i\}$ and $\widehat{\mathbf{s}}_\mathrm{g}$. When $\lambda_2 \ll n$, $\mathbf{s}_\mathrm{s} \approx \mathbf{s}_p$. When $\lambda_2 = n$, $\mathbf{s}_\mathrm{g}$ and $\mathbf{s}_p$ have equal weight. In image denoising $\lambda_2 = 0$ is typical [4].

6.3.3 LST algorithm with dictionary learning

The expressions (S4), (S5), and (S8) are solved iteratively, giving the LST algorithm, in Table 6.1, as a MAP estimate of a slowness image with local sparsity constraints and a known dictionary $\mathbf{D}$. Further implementation details are given in Sec. 6.3.4. Dictionary learning is added

to the algorithm in the solution to the local problem (S5), by optimizing $\mathbf{D}$:

$$\widehat{\mathbf{D}} = \arg\min_{\mathbf{D}} \Big\{ \min_{\mathbf{x}_i} \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}\|_2^2 \quad \text{subject to } \|\mathbf{x}_i\|_0 = T \;\forall\, i \Big\}. \tag{6.17}$$

The dictionary learning problem (S6) is here solved using the ITKM algorithm, Table 6.2 (for details, see App. 6.8.1).

The ITKM solves this bilinear optimization problem (S6) by alternately solving for the sparse coefficients $\widehat{\mathbf{x}}_i$ with $\mathbf{D}$ fixed, using thresholding [3], and solving for $\widehat{\mathbf{D}}$ with $\widehat{\mathbf{x}}_i$ fixed using a 'signed' K-means objective. In the ITKM iterations, the columns of $\mathbf{D}$, $\mathbf{d}_q$, 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. 6.4(a)) and smooth-discontinuous map (Fig. 6.4(b)). The atoms from checkerboard (Fig. 6.4(a)) contain sharp edges, which correspond to shifted sharp edges from the checkerboard pattern. The atoms from smooth-discontinuous map (Fig. 6.4(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. 6.3.5. 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].

Table 6.1: LST algorithm

Given: $\mathbf{t} \in \mathbb{R}^K$, $\mathbf{A} \in \mathbb{R}^{M \times N}$, $\mathbf{s}_{\mathrm{s}}^0 = \mathbf{0} \in \mathbb{R}^N$, $\mathbf{D}^0 =$ Haar wavelet,
DCT (or) noise $\mathcal{N}\big(0,1\big) \in \mathbb{R}^{n \times Q}$, λ_1, λ_2, T, and $j = 1$

Repeat until convergence:

1. Global estimate: solve (6.18) using LSQR [42],
$\widehat{\mathbf{s}}_{\delta}^{j} = \arg\min_{\mathbf{s}_\delta} \|\mathbf{t} - \mathbf{A}(\mathbf{s}_{\delta}^{j} + \mathbf{s}_{\mathrm{s}}^{j-1})\|_2^2 + \lambda_1 \|\mathbf{s}_\delta\|_2^2$,
$\widehat{\mathbf{s}}_{\mathrm{g}}^{j} = \widehat{\mathbf{s}}_{\delta}^{j} + \mathbf{s}_{\mathrm{s}}^{j-1}$.
2. Local estimate
a: Setting $\mathbf{s}_{\mathrm{s}}^{j} = \widehat{\mathbf{s}}_{\mathrm{g}}^{j}$, center patches $\{\mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}^{j} \; \forall \; i \}$ and
i. (Dictionary learning) Solve (S6) for $\mathbf{D}^j$ using ITKM [6] (Table 6.2),
$\widehat{\mathbf{D}}^{j} = \arg\min_{\mathbf{D}^j} \big\{ \min_{\mathbf{x}_i} \|\mathbf{D}^j \mathbf{x}_i - \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}\|_2^2 \text{ subject to } \|\mathbf{x}_i\|_0 = T \; \forall \; i \big\}$
i. (generic dictionary) Set $\mathbf{D}^j = \mathbf{D}^0$.
ii. Solve (S5) using OMP,
$\widehat{\mathbf{x}}_i^{j} = \arg\min_{\mathbf{x}_i^j} \|\mathbf{D}^j \mathbf{x}_i^j - \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}^{j}\|_2^2 \text{ subject to } \|\mathbf{x}_i^j\|_0 = T \; \forall \; i.$
b: Obtain $\widehat{\mathbf{s}}_{\mathrm{s}}^{j}$ by (S8) $\widehat{\mathbf{s}}_{\mathrm{s}}^{j} = \dfrac{\lambda_2 \widehat{\mathbf{s}}_{\mathrm{g}}^{j} + n \mathbf{s}_{\mathrm{p}}^{j}}{\lambda_2 + n}$
$j = j + 1$

6.3.4 Implementation and complexity

In the following, we give the implementation details and complexity of the LST algorithm (see Table 6.1). Since $\mathbf{A}$ is sparse, the global estimate (S4) is solved using the sparse least squares program LSQR [42]. For LSQR, (S4) is rewritten as

$$\widehat{\mathbf{s}}_{\delta} = \arg\min_{\mathbf{s}_\delta} \|\mathbf{t} - \mathbf{A}(\mathbf{s}_{\delta} + \mathbf{s}_{\mathrm{s}})\|_2^2 + \lambda_1 \|\mathbf{s}_{\delta}\|_2^2, \tag{6.18}$$

with the substitution $\mathbf{s}_{\delta} = \mathbf{s}_{\mathrm{g}} - \mathbf{s}_{\mathrm{s}}$, giving $\widehat{\mathbf{s}}_{\mathrm{g}} = \widehat{\mathbf{s}}_{\delta} + \mathbf{s}_{\mathrm{s}}$. The sparse coefficients $\mathbf{x}_i$ in the local estimates (S5), (S6) are obtained using the orthogonal matching pursuit (OMP) algorithm [43]. For LST with dictionary learning, first the dictionary is obtained from ITKM, then (S5) is solved with the same sparsity level T as ITKM. Before solving (S5), (S6), the slowness patches $\{\mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}} \; \forall \; i \}$ are centered [4] – i.e. the mean of the pixels in each patch is subtracted. The mean of

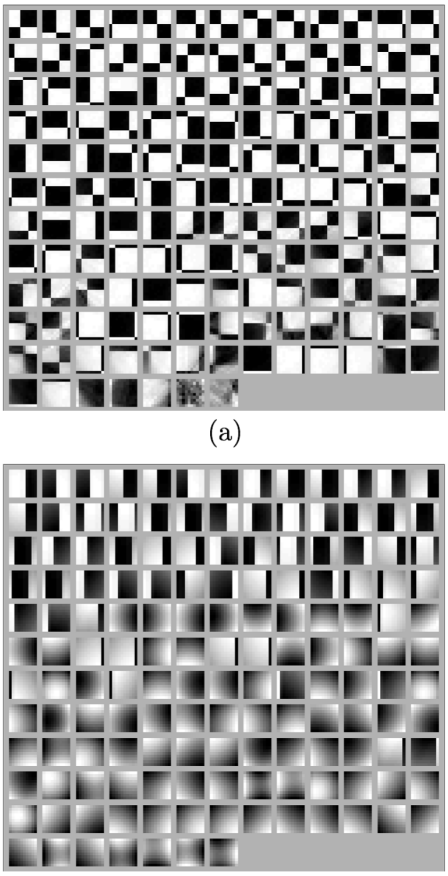


Figure 6.4: Dictionary atoms learned during LST with ITKM, with $n = 100$ and $Q = 150$ for (a) checkerboard map (Fig. 6.3(a)) with $T = 1$ and (b) smooth-discontinuous map (Fig. 6.3(b)) with $T = 2$. Atoms adjusted to full grayscale range for display.

patch i is $\bar{x}_i = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{R}_i\widehat{\mathbf{s}}_{\mathrm{g}}$. Hence, $\mathbf{R}_i\widehat{\mathbf{s}}_{\mathrm{g}} \approx \mathbf{D}\mathbf{x}_i + \mathbf{1}\bar{x}_i$.

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 $O(knQN)$ and OMP $O(TnQN)$, where k is the number of ITKM iterations (see Table 6.2). For large slowness maps, we expect the LST complexity to be dominated by LSQR. In our simulations we obtain reasonable run times (see Sec. 6.5.2). In the special case when $T = 1$, the solution to (S5) is not combinatorial, and the dictionary learning problem is equivalent to gain-shape vector quantization [5, 44].

Table 6.2: ITKM algorithm

Given: j, $\widehat{\mathbf{s}}_{\mathrm{g}}^{j}$, $\mathbf{D}^{0} = \mathbf{D}^{j-1} \in \mathbb{R}^{n \times Q}$, T, and $h = 1$
Repeat until convergence:
1. Find dictionary indices per (6.27)
$K(\mathbf{D}^{h-1}, \mathbf{y}_i) = \max_{\lvert K \rvert = T} \lVert {\mathbf{D}_K^{h-1}}^{\mathrm{T}} \mathbf{y}_i \rVert_1$,
with $\mathbf{y}_i = \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}^{j} - \mathbf{1}\bar{x}_i^{j}$.
2. Update dictionary per (6.32) using coefficient indices K
$\mathbf{d}^{h} = \lambda_l \sum_{i:l \in K(\mathbf{D}^{h-1}, \mathbf{y}_i)} \operatorname{sign}\left({\mathbf{d}_l^{h-1}}^{\mathrm{T}} \mathbf{y}_i\right) \mathbf{y}_i$
with λ_l such that $\lVert \mathbf{d}_l^{h} \rVert_2 = 1$.
$h = h + 1$
Ouput: $\mathbf{D}^{j} = \mathbf{D}^{h}$

6.3.5 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 $\mathbf{D}$. 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. 6.3(a–b) are shifted such that only small cells of constant slowness may be used with conventional tomography (which necessitates damping, Sec. 6.4.1) to illustrate this effect.

6.4 Competing methods

6.4.1 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 (S1), the MAP estimate of the slowness is

$$\widehat{\mathbf{s}}_{\mathrm{g}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \eta\Sigma_L^{-1}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{t}, \tag{6.19}$$

where $\eta = (\sigma_\varepsilon/\sigma_{\mathrm{c}})^2$ is a regularization parameter, σ_{c} is the conventional slowness variance, and

$$\Sigma_L(i,j) = \exp\left(-D_{i,j}/L\right). \tag{6.20}$$

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

6.4.2 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

$$\begin{aligned}\left\{\widehat{\mathbf{s}}_{\mathrm{g}}, \widehat{\mathbf{s}}_{\mathrm{TV}}\right\} = \underset{\mathbf{s}_{\mathrm{g}}, \mathbf{s}_{\mathrm{TV}}}{\arg\min} \Bigg\{ & \frac{1}{\sigma_\varepsilon^2}\|\mathbf{t} - \mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2 \\ & + \frac{1}{\sigma_{\mathrm{g}}^2}\|\mathbf{s}_{\mathrm{g}} - \mathbf{s}_{\mathrm{TV}}\|_2^2 + \mu\|\mathbf{s}_{\mathrm{TV}}\|_{\mathrm{TV}}\Bigg\},\end{aligned} \tag{6.21}$$

where $\|\mathbf{s}_{\mathrm{TV}}\|_{\mathrm{TV}}$ is the TV regularizer, which penalizes the gradient, and $\mathbf{s}_{\mathrm{TV}} \in \mathbb{R}^N$ is the TV estimate of the slowness. Similar to LST, TV (6.21) is solved by decoupling the problem into two

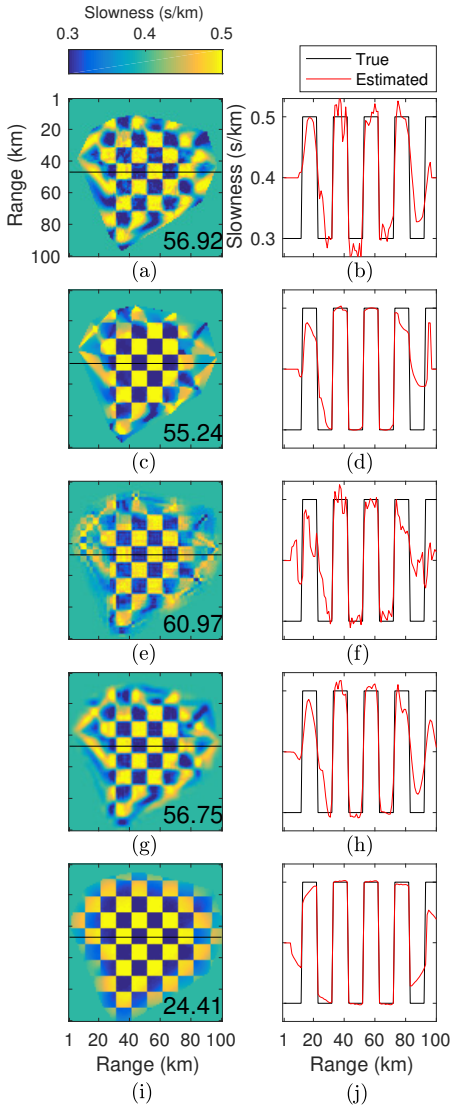


Figure 6.5: Conventional, TV, and LST tomography results without travel time error for checkerboard map (Fig. 6.3(a)): 2D and 1D (from black line in 2D) slowness estimates plotted against true slowness $\mathbf{s}'$. (a,b) conventional, (c,d) TV regularization, $\widehat{\mathbf{s}}'_{\mathrm{TV}}$, and LST with (e,f) Haar wavelet and (g,h) DCT dictionary with $Q = 169$, $T = 5$, $n = 64$; and (g,h) with dictionary learning. RMSE (ms/km, (6.24)), is printed on 2D slownesses.

subproblems: 1) damped least squares and 2) TV. The least squares problem is

$$\begin{aligned}\widehat{\mathbf{s}}_{\mathrm{g}} =& \underset{\mathbf{s}_{\mathrm{g}}}{\arg\min}\ \frac{1}{\sigma_{\varepsilon}^2}\|\mathbf{t}-\mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2+\frac{1}{\sigma_{\mathrm{g}}^2}\|\mathbf{s}_{\mathrm{g}}-\mathbf{s}_{\mathrm{TV}}\|_2^2 \\ =& \underset{\mathbf{s}_{\mathrm{g}}}{\arg\min}\ \|\mathbf{t}-\mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2+\lambda_1\|\mathbf{s}_{\mathrm{g}}-\mathbf{s}_{\mathrm{TV}}\|_2^2,\end{aligned} \tag{6.22}$$

where λ_1 is related to global LST problem (see (S4)) and $\mathbf{s}_{\mathrm{TV}}$ is initialized to the reference slowness. The TV problem is

$$\begin{aligned}\widehat{\mathbf{s}}_{\mathrm{TV}} =& \underset{\mathbf{s}_{\mathrm{TV}}}{\arg\min}\ \frac{1}{\sigma_{\mathrm{g}}^2}\|\mathbf{s}_{\mathrm{g}}-\mathbf{s}_{\mathrm{TV}}\|_2^2+\mu\|\mathbf{s}_{\mathrm{TV}}\|_{\mathrm{TV}} \\ =& \underset{\mathbf{s}_{\mathrm{TV}}}{\arg\min}\ \|\mathbf{s}_{\mathrm{g}}-\mathbf{s}_{\mathrm{TV}}\|_2^2+\lambda_{\mathrm{TV}}\|\mathbf{s}_{\mathrm{TV}}\|_{\mathrm{TV}},\end{aligned} \tag{6.23}$$

with $\lambda_{\mathrm{TV}}=\sigma_{\mathrm{g}}^2\mu$. (6.21) is solved by alternately solving (6.22) and (6.23), like LST, as an alternating minimization algorithm. In this paper (6.22) is solved using LSQR [42] and (6.23) 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.

6.5 Simulations

We demonstrate the performance of LST (Sec. 6.3, Table 6.1), using both dictionary learning and prescribed dictionaries, relative to conventional (Sec. 6.4.1) and TV (Sec. 6.4.2) tomography on synthetic slowness maps (e.g. Fig. 6.3(a,b)). The recovered slownesses from the methods are plotted in Figs. 6.5–6.11, with their performance summarized in Table 6.3. The convergence of LST and its sensitivity to the sparsity level T are shown in Fig. 6.12. For LST without dictionary learning, the dictionary $\mathbf{D}$ is either the overcomplete Haar wavelet dictionary or the DCT (see Fig. 6.2). 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. 6.3(a,b)) with dimensions $W_1 = W_2 = 100$ pixels (km), as well as variations of these maps. The boxcar checkerboard in Fig. 6.3(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. 6.3(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. 6.5.1. The travel times from the synthetic slowness maps are generated by the global model (S1). The slowness estimates from LST are $\widehat{\mathbf{s}}'_{\mathrm{s}} = \widehat{\mathbf{s}}_{\mathrm{s}} + \mathbf{s}_0 \in \mathbb{R}^N$ (from (S8)), for conventional $\widehat{\mathbf{s}}'_{\mathrm{g}} = \widehat{\mathbf{s}}_{\mathrm{g}} + \mathbf{s}_0 \in \mathbb{R}^N$ (from (6.19)), and for TV $\widehat{\mathbf{s}}'_{\mathrm{TV}} = \widehat{\mathbf{s}}_{\mathrm{TV}} + \mathbf{s}_0 \in \mathbb{R}^N$ (from (6.23)).

The slowness map pixels are 1 km square and sampled by $M = 2016$ straight-rays between 64 sensors, shown in Fig. 6.3(c). The travel times $\mathbf{t}$ 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. 6.3(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 6.2), we exclude patches from

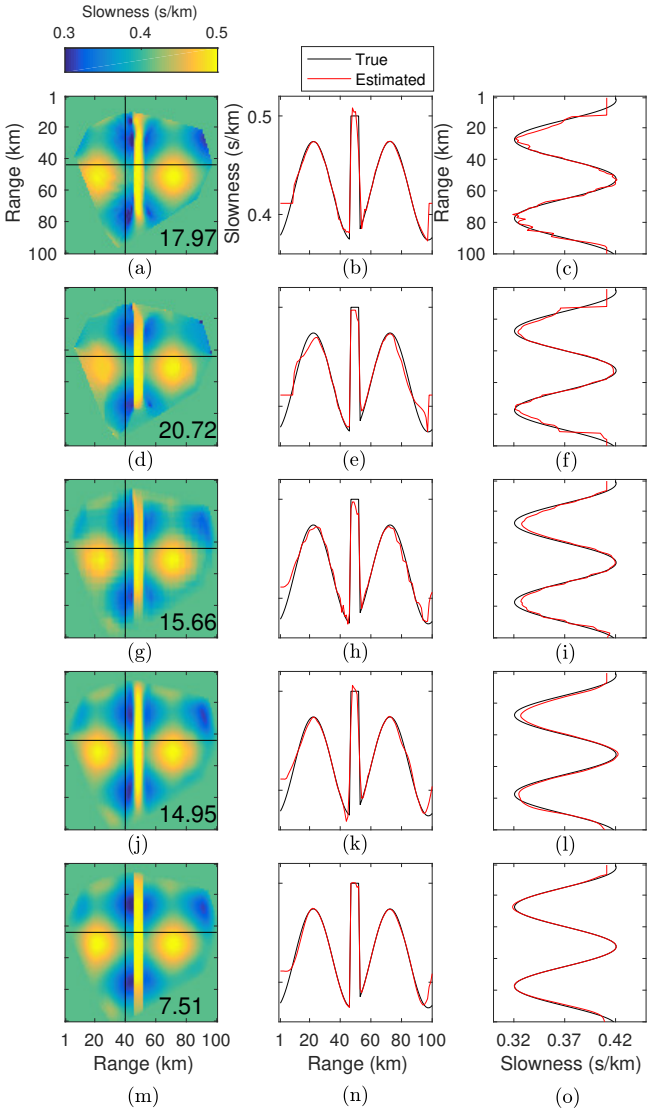


Figure 6.6: Conventional, TV, and LST tomography results without travel time error for smooth-discontinuous map (Fig. 6.3(b)). Same order as Fig. 5, with vertical 1D profile of estimated and true slowness. RMSE (ms/km, (6.24)), is printed on 2D slownesses.

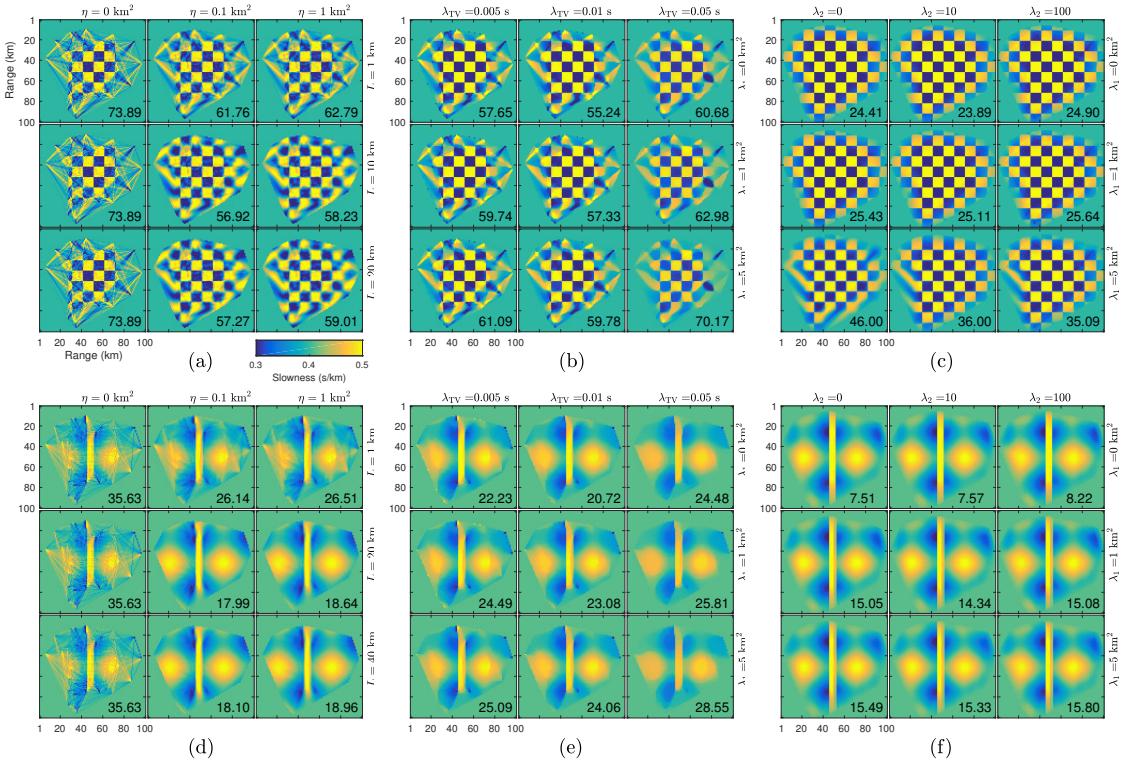


Figure 6.7: Conventional, TV, and LST tomography results for different values of regularization parameters for (a–c) checkerboard (Fig. 6.3(a)) and (d–f) smooth-discontinuous (Fig. 6.3(b)) maps, without travel time error. (a,d) Conventional $\widehat{\mathbf{s}}'_{\mathrm{g}}$, effect of L and η. (b,e) TV regularization $\widehat{\mathbf{s}}'_{\mathrm{TV}}$, effect of λ_1 and λ_{TV}. (c,f) LST $\widehat{\mathbf{s}}'_{\mathrm{s}}$, effect of λ_1 and λ_2. RMSE (ms/km, (6.24)), is printed on 2D slownesses.

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{NP}\sum_{\mathrm{n}}^{N}\sum_{p}^{P}\left(s'_{\mathrm{n}p} - s'_{\mathrm{est.,n}p}\right)^2}, \tag{6.24}$$

where $s'_{\mathrm{est.,n}p}$ is $\widehat{\mathbf{s}}'_{\mathrm{s}}$, $\widehat{\mathbf{s}}'_{\mathrm{g}}$, or $\widehat{\mathbf{s}}'_{\mathrm{TV}}$ for the n-th pixel location, and the p-th trial. For $\sigma_\varepsilon = 0$, $P = 1$. The RMSE is printed on the 1D/2D slowness estimates in Figs. 5–12.

6.5.1 Without travel time error

We first simulate travel times without errors ($\sigma_\varepsilon = 0$) to obtain best-case results for the tomography methods. See Figs. 6.5 and 6.6 for results from LST, conventional, and TV tomography.

Inversion parameters for $\sigma_\varepsilon = 0$

The LST tuning parameters are λ_1, λ_2, n, Q, and T. The sensitivity of the LST solutions with dictionary learning to λ_1 and λ_2 are shown in Fig. 6.7(c,f) for nominal values of $n = 100$, $Q = 150$, $T = 1$ for the checkerboard and $T = 2$ for the smooth-discontinuous (see Fig. 6.12(b)). With the prior $\sigma_\varepsilon = 0$ s, we expect the best value of $\lambda_1 = 0$ km^2 (from (S4)). From (S4), (S7) σ_g^2 is proportional to the variance of the true slowness. For the checkerboard (smooth-discontinuous) is $\sigma_g = 0.10$ (0.05) s/km. We assume the slowness patches are well approximated by the sparse model (S5), and expect $\sigma_p^2 \ll \sigma_g^2$. Hence we expect the best value of λ_2 (from (S7)) to be small. It is shown for both the checkerboard (Fig. 6.7(c)) and smooth-discontinuous maps (Fig. 6.7(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 (6.24), for the nominal LST parameters in Fig. 6.12(b). The values used for T were chosen based on the minimum error for 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 η, but the methods not always reliable [2, 50]. To find the

best parameters, we systematically varied the values of L and η (see Fig. 6.7(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, λ_1 and λ_{TV}, were obtained by systematically varying their values (see Fig. 6.7(b,e)). The minimum RMSE for TV tomography was obtained by $\lambda_1 = 1$ km^2 and $\lambda_{TV} = .01$ s for both the checkerboard and smooth-discontinuous maps.

Results for $\sigma_\varepsilon = 0$

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. 6.3(d)). The performance of the Haar wavelets is better for the smooth-discontinuous slowness map (Fig. 6.6(g–i)) than for the checkerboard (Fig. 6.5(e,f)). As shown in Fig. 6.6(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. 6.5(g,h) and Fig. 6.6(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. 6.5(i,j) and Fig. 6.6(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 6.3) is greater than for conventional tomography, although a better qualitative fit to the true slowness is observed with the LST. Both in the case of the checkerboard and smooth-discontinuous maps, the TV obtained the

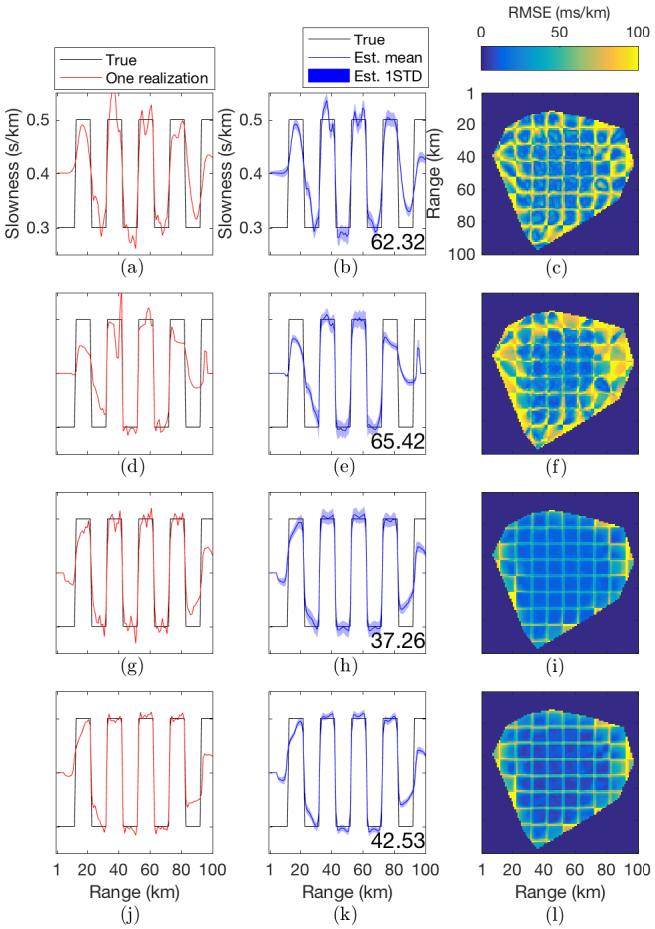


Figure 6.8: Conventional, TV, and LST tomography results for checkerboard map (Fig. 6.3(a)) with 100 realizations of Gaussian travel time error (STD 2% mean travel time): 1D slice of inversion for one noise realization against true slowness, 1D slice of mean from over all noise realizations with STD of estimate against true slowness $\mathbf{s}'$, and 2D RMSE of estimates over noise realizations. (a–c) conventional tomography $\widehat{\mathbf{s}}'_{\mathrm{g}}$, (d–f) TV regularization $\widehat{\mathbf{s}}'_{\mathrm{TV}}$, (g–i) LST $\widehat{\mathbf{s}}'_{\mathrm{s}}$ with dictionary learning with $\lambda_1 = 2$ km^2, and (j–l) $\lambda_1 = 7$ km^2. RMSE (ms/km, (6.24)), is printed on 1D errors.

highest RMSE, though in regions of the map where the ray sampling was dense, the discontinuous and constant features were well recovered (Fig. 6.5(c,d) and Fig. 6.6(d–f)), as expected for TV regularization.

The true variances σ_ε, σ_g, and σ_p in the LST regularization parameters λ_1 and λ_2 provide best LST performance (see Fig. 6.7(c,f)), whereas the true variances for conventional tomography, σ_ε and σ_c, do not correspond to the best solutions (see Fig. 6.7(a,d)). The noise-free prior, $\eta = 0$ km^2, gives erratic inversions, and the conventional tomography is better regularized by $\eta = 0.1$ km^2. The noise-free prior also gives erratic results from TV (see Fig. 6.7(b,e)).

A variety of checkerboard and smooth-discontinuous slowness maps with different geometries were generated to more fully test the tomography methods. The results of these tests are summarized in Table 6.3 (and cases with travel time error are shown in Fig. 6.11). Different checkerboard maps were generated by varying the size horizontal and vertical of the checkerboard boxes from 5 to 20 pixels (256 permutations), and different smooth-discontinuous maps were generated by varying the location of the left edge (from 32 to 62 pixels) and width (from 4 to 10 pixels) of the discontinuity (217 permutations). From each of these sets, 100 slowness maps were randomly chosen for simulation. Inversions without travel time errors were performed using conventional, TV, and LST (with dictionary learning) tomography using the nominal parameters from the aforementioned test cases, corresponding to the slowness maps in Fig. 6.3(a,b). LST obtains lower RMSE than TV or conventional for all simulations for both varied checkerboard and smooth-discontinuous maps.

6.5.2 With travel time error

We also test the performance of the tomography methods with travel time errors. We simulate Gaussian travel time errors with $\sigma_\varepsilon = 0.02\bar{t}$, 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 $\mathcal{N}\left(0, \sigma_\varepsilon\right)$ (also the random initialization of $\mathbf{D}$

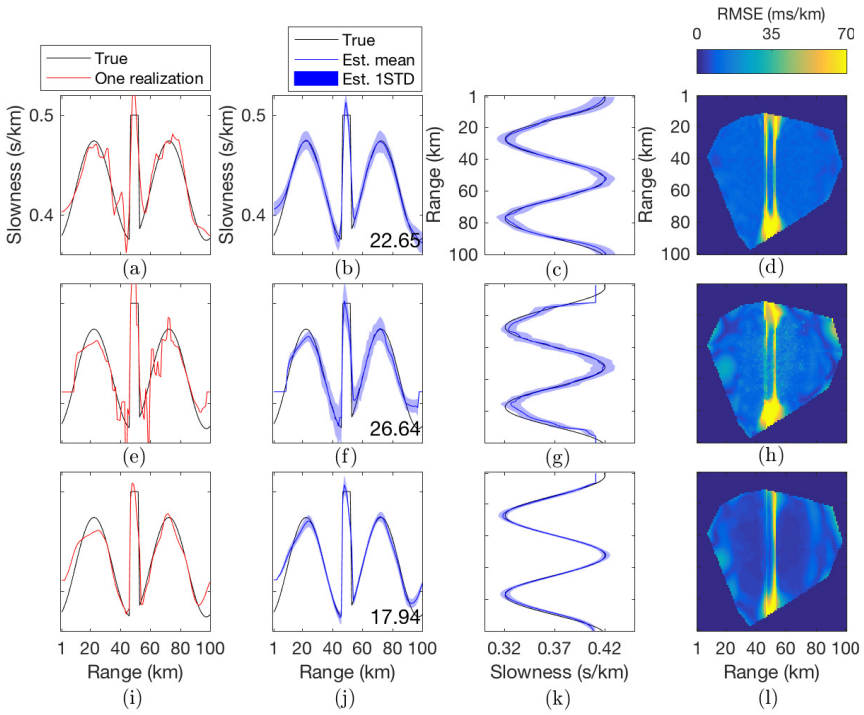


Figure 6.9: Conventional, TV, and LST tomography results for smooth-discontinuous map (Fig. 6.3(b)) with 100 realizations of Gaussian travel time error (STD 2% mean travel time). Same format as Fig. 8, except vertical 1D slice of inversion is included, and only one LST case (i–l) $\widehat{\mathbf{s}}'_{\mathrm{s}}$ with dictionary learning. RMSE (ms/km, (6.24)), is printed on 1D errors.

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. 6.8 and 6.9. The RMSE for both approaches, calculated by (6.24) with $P = 100$, are in Table 6.3.

Inversion parameters for $\sigma_\varepsilon = 0.02\bar{t}$

The sensitivity of the LST solutions with dictionary learning to λ_1 and λ_2 are shown in Fig. 6.10(c,f) for nominal values of $n = 100$, $Q = 150$ and $T = 2$ (per Fig. 6.12(b)) for both the checkerboard and smooth-discontinuous maps. With the prior $\sigma_\varepsilon = 0.02\bar{t}$, we expect for the checkerboard map ($\sigma_\varepsilon = 0.27$ s, $\sigma_g = 0.10$ s/km) the best value of $\lambda_1 \approx 7.5$ km^2, and for the the smooth-discontinuous ($\sigma_\varepsilon = 0.28$ s, $\sigma_g = 0.05$ s/km) the best value of $\lambda_1 \approx 28.3$ km^2 (from (S4)). We use $\lambda_1 = 2$ km^2 for the checkerboard (Fig. 6.8(g–i)) and $\lambda_1 = 10$ km^2 for the smooth-discontinuous map (Fig. 6.9(i–l)), and achieve lower RMSE than $\lambda_1 = 1$ km^2 (see Fig. 6.10(c,f)). Although we expect the true values σ_g to decrease over the LST iterations, prior values of σ_g proportional to the variance of the true slowness work well. It is further shown in Fig. 6.10 that, as in the the noise-free case (Sec. 6.5.1), 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 $\lambda_1 = 5$ 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 $L = 6$ km and $\eta = 10$ km^2 for the checkerboard and $L = 12$ km and $\eta = 10$ km^2 for the smooth-discontinuous slowness maps Fig. 6.10(a,d). For TV tomography, the best values by minimum RMSE were $\lambda_1 = 5$ km^2 and $\lambda_{TV} = 0.02$ s for both the checkerboard and smooth-discontinuous slowness maps Fig. 6.10(b,e).

Considering again the influence of the choice of λ_1 and λ_2 on the LST with dictionary learning in Fig. 6.10(c,f), since $n = 100$ the case $\lambda_2 = 100$ gives equal weight to the global $\widehat{\mathbf{s}}_{g,n}$ and patch slowness $n\mathbf{s}_p$ in (S8). The LST obtains results similar to the best conventional estimates

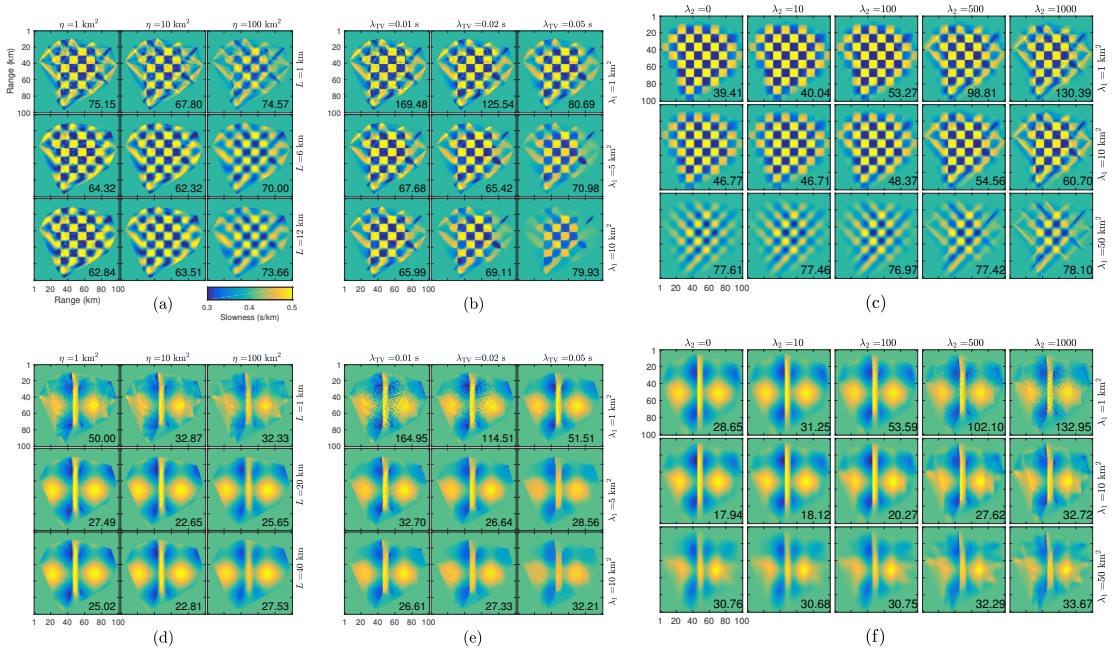


Figure 6.10: Conventional, TV, and LST tomography results for different values of regularization parameters for (a–c) checkerboard (Fig. 6.3(a)) and (d–f) smooth-discontinuous (Fig. 6.3(b)) maps, for 100 realizations of Gaussian travel time error (STD 2% mean travel time). (a,d) Conventional mean $\widehat{\mathbf{s}}'_{\mathrm{g}}$, effect of L and η. (b,e) TV regularization mean $\widehat{\mathbf{s}}'_{\mathrm{TV}}$, effect of λ_1 and λ_{TV}. (c,f) LST mean $\widehat{\mathbf{s}}'_{\mathrm{s}}$, effect of λ_1 and λ_2. RMSE (ms/km, (6.24)), is printed on 2D slownesses.

(see Fig. 6.10) 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 (6.19), there is no direct relationship between the regularization in (6.19) and the sparsity and dictionary learning in (S5).

Results for $\sigma_\varepsilon = 0.02\bar{t}$

The LST with dictionary learning (Fig. 6.8(g–i) and Fig. 6.9(i–l)) achieves the best overall fit to the true slowness, relative to conventional tomography (Fig. 6.8(a–c) and Fig. 6.9(a–d)) and TV tomography (Fig. 6.8(d–f) and Fig. 6.9(e–h)) as evidenced by qualitative fit and RMSE (Table 6.3). In the case of the checkerboard, a higher value $\lambda_1 = 7$ km^2 is tested and shown in

Fig. 6.8(j–l). Because of the increased damping, the standard deviation (STD) of the estimate (Fig. 6.8(k)) is less than the case $\lambda_1 = 2$ km^2 (Fig. 6.8(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. 6.5.1), with travel time error. The results of these tests are summarized in Table 6.3 and Fig. 6.11. Inversions with 10 realizations of Gaussian travel time error ($\sigma_\varepsilon = 0.02\bar{t}$) were performed using conventional, TV, and LST (with dictionary learning) tomography using the nominal parameters for the slowness maps in Fig. 6.3(a,b). Fig. 6.11(a,b) and Fig. 6.11(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. 6.11(a,c), a better subjective fit to the true slowness is also observed (Fig. 6.11(b,d)).

Convergence and run time

The algorithms were coded in Matlab. The LST algorithm (Table 6.1) used 100 iterations for all cases and the ITKM (Table 6.2) used 50 iterations. The TV algorithm (Sec. 6.4.2) also used 100 iterations for all cases. The inversions in Fig. 6.5 and Fig. 6.6 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. 6.4.1) took 20 s and TV tomography took 3 s.

In Fig. 6.12(a), it is shown that the LST RMSE travel time error (from $\mathbf{s}_s$) 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 σ_ε. Hence, the travel time data was not overfit.

6.6 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 synthetic images we used, well recovered true slowness maps with smooth and discontinuous features. The LST is relevant to other tomography scenarios where slowness structure is irregularly sampled, for instance in ocean [51] and terrestrial [52, 53] acoustics.

6.7 Acknowledgments

This work is supported by the Office of Naval Research, Grant No. N00014-18-1-2118.

Chapter 6, in full, is a reprint of the material as it appears in the Institute of Electrical and Electronics Engineers Transactions on Computational Imaging 2018. Bianco, Michael J.; Gerstoft, Peter, Institute of Electrical and Electronics Engineers, 2018. The dissertation author was the primary investigator and author of this paper.

6.8 Appendix

6.8.1 ITKM algorithm details

The ITKM dictionary learning algorithm [6] (see Table 6.2) is derived from a 'signed' K-means objective. In signed K-means, T-sparse coefficients $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_I] \in \mathbb{R}^{Q \times I}$ with $\mathbf{c}_i^T \in \{-1, 1\}$ are assigned to training examples $\{\mathbf{y}_i, ..., \mathbf{y}_I\} \in \mathbb{R}^{n \times I}$. The training examples are obtained

Table 6.3: Conventional, TV, and LST tomography RMSE performance. Bold entries are least error.

	RMSE (ms/km)							
	Checkerboard				Smooth-discon.			
	Nominal(Fig. 6.3(a))		Varied		Nominal(Fig. 6.3(b))		Varied	
Error $\sigma_\varepsilon =$	0^*	$0.02\bar{t}^{**}$	0	$0.02\bar{t}^{\ddagger}$	$0^\dagger$	$0.02\bar{t}^{\dagger\dagger}$	0	$0.02\bar{t}^{\ddagger}$
Conv.	56.92	62.32	59.96	65.77	17.97	22.65	13.98	19.00
TV	55.24	65.42	58.28	67.05	20.72	26.64	17.16	22.55
LST Haar	60.97	68.96	59.71	66.51	15.66	23.07	14.00	18.68
LST DCT	56.75	62.16	56.47	62.73	14.95	19.62	13.41	17.24
LST Adapt.	**24.41**	**37.26**	**31.16**	**37.14**	**7.51**	**17.94**	**7.40**	**14.62**

Estimated slownesses plotted in *Fig. 6.5, **Fig. 6.8, †Fig. 6.6, ††Fig 6.9, and ‡Fig 6.11.

from patch i as $\mathbf{y}_i = \mathbf{R}_i \mathbf{s}_\mathrm{s}$, and centered. The minimization problem is

$$\begin{aligned}\big\{\mathbf{C},\mathbf{D}\big\} =& \arg\min_{\mathbf{D}} \sum_i \arg\min_{T,\mathbf{c}_i^T=\pm 1} \|\mathbf{y}_i - \mathbf{D}\mathbf{c}_i\|_2^2 \\ =& \arg\min_{\mathbf{D}} \sum_i \arg\min_{T,\ c_i^t=\pm 1} \|\mathbf{y}_i - \sum_t \mathbf{d}_t c_i^t\|_2^2,\end{aligned} \tag{6.25}$$

where c_i^t is a non-zero coefficient and $\mathbf{d}_t$ is the corresponding dictionary atom. Expanding (6.25) and requiring $\|\mathbf{d}_t\|_2^2 = 1$,

$$\begin{aligned}&\min_{\mathbf{D}} \sum_i \min_{T,\ c_i^t=\pm 1} \left\{ \|\mathbf{y}_i\|_2^2 - 2\sum_t c_i^t \mathbf{d}_t^\mathrm{T}\mathbf{y}_i + B \right\} \\ &\quad = \|\mathbf{Y}\|_\mathcal{F}^2 + B - 2\max_{\mathbf{D}} \sum_i \max_{|K|=T} \sum_t \mathrm{abs}\big(\mathbf{d}_t^\mathrm{T}\mathbf{y}_i\big) \\ &\quad = \|\mathbf{Y}\|_\mathcal{F}^2 + B - 2\max_{\mathbf{D}} \sum_i \max_{|K|=T} \|\mathbf{D}_K^\mathrm{T}\mathbf{y}_i\|_1,\end{aligned} \tag{6.26}$$

where $\|\cdot\|_\mathcal{F}$ is the Frobenius norm, B is a constant, and K is the set of T dictionary indices having the largest absolute inner product $\mathrm{abs}(\mathbf{d}_t\mathbf{y}_i)$, by is by thresholding [3]

$$K(\mathbf{D},\mathbf{y}_i) = \arg\max_{|K|=T} \|\mathbf{D}_K^\mathrm{T}\mathbf{y}_i\|_1. \tag{6.27}$$

From (6.26), the dictionary learning objective is

$$\max_{\mathbf{D}} \sum_{i} \max_{|K|=T} \|\mathbf{D}_K^{\mathrm{T}} \mathbf{y}_i\|_1, \tag{6.28}$$

which finds $\mathbf{D}$ that maximizes the absolute norm of the T-largest responses from K. The ITKM solves (6.28) as a two-step algorithm. First, K is obtained from (6.27). Then, the dictionary atoms are updated per

$$\max_{\mathbf{D}} \sum_{i:l \in K(\mathbf{D},\mathbf{y}_i)} \mathrm{abs}\big(\mathbf{d}_l^{\mathrm{T}} \mathbf{y}_i\big). \tag{6.29}$$

(6.29) is solved using Lagrange multipliers, with the constraint $\|\mathbf{d}_l\|_2 = 1$. The Lagrangian function is

$$\Phi(\mathbf{d}_l, \lambda) = \sum_{i:l \in K(\mathbf{D},\mathbf{y}_i)} \mathrm{abs}\big(\mathbf{d}_l^{\mathrm{T}} \mathbf{y}_i\big) - \lambda\big(\mathbf{d}_l^{\mathrm{T}} \mathbf{d}_l - 1\big), \tag{6.30}$$

with λ the Lagrange multiplier. Differentiating (6.30) gives

$$\frac{d\Phi}{d\mathbf{d}_l} = \sum_{i:l \in K(\mathbf{D},\mathbf{y}_i)} \mathrm{sign}\big(\mathbf{d}_l^{\mathrm{T}} \mathbf{y}_i\big) \mathbf{y}_i - 2\lambda \mathbf{d}_l \tag{6.31}$$

The stationary point of (6.30), per (6.31), gives the update for $\mathbf{d}_l$

$$\mathbf{d}_l^{new} = \lambda_l \sum_{i:l \in K(\mathbf{D}^{old},\mathbf{y}_n)} \mathrm{sign}\big(\mathbf{d}_l^{old^{\mathrm{T}}} \mathbf{y}_i\big) \mathbf{y}_i, \tag{6.32}$$

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.

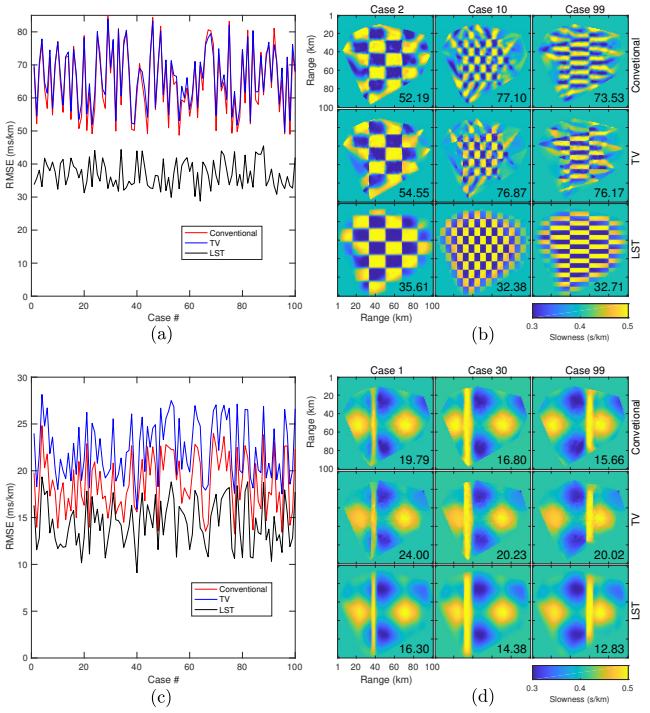
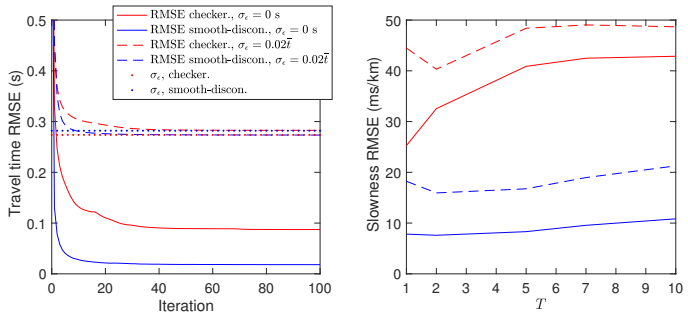


Figure 6.11: Conventional, TV, and LST tomography for 100 checkerboard and smooth-discontinuous slowness map configurations, each with 10 realizations of travel time error ($\sigma_\varepsilon = 0.02\bar{t}$). (a,c) RMSE error for 100 cases, and (b,d) mean slowness for 3 example cases from (a,c). RMSE (ms/km, (6.24)), is printed on 2D slownesses.



(a) (b)

Figure 6.12: (a) LST algorithm travel time RMSE convergence vs. iteration (Table 6.1) 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. 6.5(i), 6.8(g)) and smooth-discontinuous (Fig. 6.6(m), 6.9(i)) slowness maps.

6.9 Bibliography

[1] N. Rawlinson, S. Pozgay, and S. Fishwick, "Seismic tomography: a window into deep earth," *Phys. Earth Plan. Inter.*, vol. 178, no. 3, pp. 101–135, 2010.

[2] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter estimation and inverse problems*, 2nd ed. Elsevier, San Diego, 2013.

[3] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[4] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014.

[5] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, pp. 4311–4322, 2006.

[6] K. Schnass, "Local identification of overcomplete dictionaries," *J. Mach. Learn. Res.*, no. 16, pp. 1211–1242, 2015.

[7] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *ACM Proc. 26th Inter. Conf. Mach. Learn.*, pp. 689–696, 2009.

[9] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 4, pp. 791–804, 2012.

[10] A. Tarantola, *Inverse problem theory*. Elsevier Sci. Pub. Co., Inc., 1987.

[11] Y. Lin and L. Huang, "Quantifying subsurface geophysical properties changes using double-difference seismic-waveform inversion with a modified total-variation regularization scheme," *Geo. J. Int.*, vol. 203, no. 3, pp. 2125–2149, 2015.

[12] X. Zhang and J. Zhang, "Model regularization for seismic traveltime tomography with an edge-preserving smoothing operator," *J. Appl. Geo.*, vol. 138, pp. 143–153, 2017.

[13] L. Y. Chiao and B. Y. Kuo, "Multiscale seismic tomography," *Geo. J. Int.*, vol. 145, no. 2, pp. 517–527, 2001.

[14] R. Hawkins and M. Sambridge, "Geophysical imaging using trans-dimensional trees," *Geo. J. Int.*, vol. 203, no. 2, pp. 972–1000, 2015.

[15] I. Loris, G. Nolet, I. Daubechies, and F. A. Dahlen, "Tomographic inversion using ℓ_1–norm regularization of wavelet coefficients," *Geo. J. Int.*, vol. 170, no. 1, pp. 359–370, 2007.

[16] I. Loris, H. Douma, G. Nolet, I. Daubechies, and C. Regone, "Nonlinear regularization techniques for seismic tomography," *J. Comput. Phys.*, vol. 229, no. 3, pp. 890–905, 2010.

[17] J. Charléty, S. Voronin, G. Nolet, I. Loris, F. J. Simons, K. Sigloch, and I. C. Daubechies, "Global seismic tomography with sparsity constraints: comparison with smoothing and damping regularization," *J. Geophys. Research: Solid Earth*, vol. 118, no. 9, pp. 4887–4899, 2013.

[18] H. Fang and H. Zhang, "Wavelet-based double-difference seismic tomography with sparsity regularization," *Geo. J. Int.*, vol. 199, no. 2, pp. 944–955, 2014.

[19] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by v1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[20] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1–norm," *Geophys.*, vol. 44, no. 1, pp. 39–52, 1979.

[21] N. R. Chapman and I. Barrodale, "Deconvolution of marine seismic data using the l1 norm," *Geophys. J. Int.*, vol. 72, no. 1, pp. 93–100, 1983.

[22] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: the application of compressed sensing for rapid MR imaging," *Mag. Res. Medic.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[23] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1028–1041, 2011.

[24] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 260–271, 2014.

[25] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," *IEEE Sig. Proc. Let.*, vol. 23, no. 10, pp. 1469–1473, 2016.

[26] ——, "Multi-frequency sparse Bayesian learning for robust matched field processing," *J. Acoust. Soc. Am.*, vol. 141, no. 5, pp. 3411–3420, 2017.

[27] K. L. Gemba, W. S. Hodgkiss, and P. Gerstoft, "Adaptive and compressive matched field processing," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 92–103, 2017.

[28] M. Taroudakis and C. Smaragdakis, "De-noising procedures for inverting underwater acoustic signals in applications of acoustical oceanography," *EuroNoise 2015 Maastricht*, pp. 1393–1398, 2015.

[29] T. Wang and W. Xu, "Sparsity-based approach for ocean acoustic tomography using learned dictionaries," *OCEANS 2016 Shanghai IEEE*, pp. 1–6, 2016.

[30] M. Bianco and P. Gerstoft, "Compressive acoustic sound speed profile estimation," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. EL90–EL94, 2016.

[31] ——, "Dictionary learning of sound speed profiles," *J. Acoust. Soc. Am*, vol. 141, no. 3, pp. 1749–1758, 2017.

[32] ——, "Regularization of geophysical inversion using dictionary learning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[33] Y. Choo and W. Seong, "Compressive sound speed profile inversion using beamforming results," *Remote Sens.*, vol. 10, no. 5, p. 704, 2018.

[34] S. Beckouche and J. Ma, "Simultaneous dictionary learning and denoising for seismic data," *Geophys.*, vol. 79, no. 3, pp. 27–31, 2014.

[35] Y. Chen, "Fast dictionary learning for noise attenuation of multidimensional seismic data," *Geo. J. Int.*, vol. 209, no. 1, pp. 21–31, 2017.

[36] L. Zhu, E. Liu, and J. H. McClellan, "Sparse-promoting full waveform inversion based on online orthonormal dictionary learning," *Geophys.*, vol. 82, no. 2, pp. R87–R107, 2017.

[37] A. Li, Z. Peng, D. Hollis, L. Zhu, and J. H. McClellan, "High-resolution seismic event detection using local similarity for Large-N arrays," *Nature Scientific Reports*, vol. 8, no. 1, p. 1646, 2018.

[38] C. D. Rodgers, *Inverse methods for atmospheric sounding: theory and practice*. World Sci. Pub. Co., 2000.

[39] M. P. Barmin, M. H. Ritzwoller, and A. L. Levshin, "A fast and reliable method for surface wave tomography," *Pure and Appl. Geo.*, vol. 158, no. 8, pp. 1351–1375, 2001.

[40] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM J. Image Sci.*, vol. 1, no. 3, pp. 228–247, 2008.

[41] C.-F. Huang, P. Gerstoft, and W. S. Hodgkiss, "Effect of ocean sound speed uncertainty on matched-field geoacoustic inversion," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. EL162–EL168, 2008.

[42] C. C. Paige and M. A. Saunders, "LSQR: an algorithm for sparse linear equations and sparse least squares," *ACM Trans. Math. Software*, vol. 8, no. 1, pp. 43–71, 1982.

[43] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *27th. Annual Asilomar Conference on Signals, Systems and Computers, IEEE Proc.*, pp. 40–44, 1993.

[44] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Kluwer Academic, Norwell, MA, 1991.

[45] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Springer Sci. Bus. Media, 2009.

[46] I. Fedorov, B. D. Rao, and T. Q. Nguyen, "Multimodal sparse bayesian dictionary learning applied to multimodal data classification," *IEEE Conf. Acoust., Speech and Sig. Proc. (ICASSP)*, pp. 2237–2241, 2017.

[47] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging and Vision*, vol. 20, no. 1–2, pp. 89–97, 2004.

[48] M. Zhu, S. J. Wright, and T. F. Chan, "Duality-based algorithms for total-variation-regularized image restoration," *Comput. Opt. and App.*, vol. 47, no. 3, pp. 377–400, 2010.

[49] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 3rd ed. Pearson Ed., 2008.

[50] T. Bodin, M. Sambridge, and K. Gallagher, "A self-parametrizing partition model approach to tomographic inverse problems," *Inv. Prob.*, vol. 25, no. 5, 2009.

[51] C. M. Verlinden, J. Sarkar, W. S. Hodgkiss, W. A. Kuperman, and K. G. Sabra, "Passive acoustic source localization using sources of opportunity," *J. Acoust. Soc. Am. Expr. Let.*, vol. 138, no. 1, pp. EL54–EL59, 2015.

[52] P. Annibale, J. Filos, P. A. Naylor, and R. Rabenstein, "TDOA-based speed of sound estimation for air temperature and room geometry inference," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 2, pp. 234–246, 2013.

[53] R. Rabenstein and P. Annibale, "Acoustic source localization under variable speed of sound conditions," *Wirel. Comm. Mob. Comp.*, 2017.

Chapter 7

High-resolution seismic tomography of Long Beach, CA using machine learning: delineating aquifers

In the past decade, ambient noise tomography (ANT) methods have been used to image the interior of the Earth with great success [1–11]. Such methods are critical for improving our understanding of Earth's structure and characterizing seismic hazard [12, 13]. In ANT, seismic noise generated by ocean-atmospheric interactions or anthropogenic sources [14, 15] is cross-correlated between seismic sensors over periods of days to months to obtain travel times between sensors. The calculated travel times are used to estimate phase speed structure [2, 4]. The number of travel times in ANT can be very large, and the coverage of a region dense, as the number of travel times is $N(N-1)/2$ (with N the number of sensors, e.g. $N = 5204$ gives $\sim$13.5 million). Yet, the estimation of high-resolution phase speed structure with ANT remains an ill-posed inverse problem due to many factors, including irregular sensor distributions, phase ambiguities in the noise cross-correlations (especially for high-frequency surface waves), and non-isotropic noise distributions [6]. Hence, ANT is reliant on signal processing theory to obtain physically plausible

results. In this report, we demonstrate a machine learning-based tomography method, called locally sparse travel time tomography (LST) [16], on data obtained from the "large-N" Long Beach array (Fig. 7.1A–B), in Long Beach, CA, USA in 2011. It is shown that LST can improve ANT results over previous methods.

Recently, machine learning techniques have found many useful applications in seismology, including seismic waveform classification [17], event localization [10, 18, 19], earthquake prediction [20], and earthquake early warning [21]. In part, the success of these methods is derived from large amounts of training and ground-truth data. In ANT however, little training data exists. LST addresses this issue by using an unsupervised machine learning method, called dictionary learning, to constrain slowness features in the tomographic image. This procedure is derived from the *adaptive* dictionary learning paradigm from image processing [22–24], in which dictionaries are learned directly from corrupted measurements. In adaptive image denoising [22], the dictionary is trained on small rectangular groups of pixels, called patches, of a noisy image. In LST, slowness dictionaries are learned from patches of a least squares-regularized inversion, and are subsequently used to reconstruct a sparsity-constrained slowness image. The dictionary is initially unknown and is learned iteratively, assuming sufficiently dense ray sampling (see Supplemental Material). Relative to previous travel time tomography methods which enforce either smooth or discontinuous models, including conventional straight ray [25] and eikonal tomography [6], the sparse model in LST permits both smooth and discontinuous geophysical features. This approach is related to wavelet based methods [26], however in LST the atoms are adapted to the slowness features in the data using dictionary learning.

Using LST, we perform surface wave ANT with data from the Long Beach array (Fig. 7.1A–B). The Long Beach array was deployed from January to June 2011 as part of a petroleum industry survey. It was a very dense, "large-N" array with 5204 high-frequency vertical velocity sensors distributed over a 7×10 km area (Fig. 7.1A). We obtained Rayleigh surface wave travel times between all station pairs in the array by cross-correlating seismic noise recorded

during the period 5–25 March 2011. With $N = 5204$, this yielded $\sim$13.5 million travel times, which was reduced in preprocessing. We discretized the footprint of the Long Beach array into a 300×206 pixel (N-S$\times$E-W) phase-speed map (Fig. 7.1B), which corresponds to pixel sizes of 35×35 m. The phase speed for each pixel is estimated by LST with the Rayleigh wave travel times.

We use the 1 Hz Rayleigh surface wave band from the Long Beach data, which corresponds to near-surface geophysical features ($\sim$100-500 m depth). For each station pair, cross-correlations of all 1-h segments from the 3-week recording were normalized and stacked to obtain the causal and anti-causal travel times. The final travel times were obtained by averaging these causal and anti-causal components [6]. After quality control the number of useful cross-correlations for 1 Hz was $\sim$8.5 million. Quality control included SNR thresholding and removal of travel times with ranges less than one wavelength. The cross-correlations further suffered from phase ambiguities, which were reduced in preprocessing. This was done by clustering the rays [25] and filtering travel times that exceeded the median travel times of the clusters by one half-period (0.5 s for 1 Hz). Further, cross-correlations were rejected if travel times from different virtual sources disagreed by more than one-half period. This reduced the number of useful cross-correlations to $\sim$ 3 million. For further details and preprocessing steps, see Supplemental Material and [6].

The 1 Hz Rayleigh surface wave travel times are used by LST to estimate a 300×206 pixel phase speed image of the Long Beach region. We assume straight-ray surface wave propagation, which yields a simple linear measurement model (Eq. S1). Using the measuments, LST alternates between solving larger-scale, or global, phase speed features (see Supplemental Material) and smaller-scale or *local* phase speed features. The global problem (Eq. S4) is solved by least squares. Since the tomography matrix is sparse, we use the sparse least square program LSMR [27]. The local sparse problem (Eq. S5) is solved using orthogonal matching pursuit [22], and the dictionary is learned using the iterative thresholding and signed K-means algorithm [28]. The reference phase speed is constant, and estimated as the average phase speed of all ray paths.

The LST tuning parameters are $\{n, T, Q, \lambda_1, \lambda_2\}$. n is the number of pixels per patch. We use square, 10 by 10 pixel patches, giving $n = 100$ pixels per patch (yielding $300 \times 206 = 61,800$ patches). We assume sparsity $T = 2$ (see Eq. S5), meaning that each patch uses two atoms from the dictionary $\mathbf{D}$. Each atom in $\mathbf{D}$ is a vector with dimension $n = 100$ pixels (the patch size). We assume $\mathbf{D}$ has $Q = 200$ atoms, or twice the patch dimension. $\mathbf{D}$ is initialized with Gaussian random vectors of unit norm. λ_1, which is the ratio of travel time error variance to global slowness variance is set as $\lambda_1 = 13$ km^2 (see Eq. S4). λ_2, which is the ratio of patch slowness variance to global slowness variance, is set as $\lambda_2 = 0$ assuming a sparse slowness representation (see Eq. S5). When $\lambda_2 = 0$, the sparse slowness is simply the average of the patch slownesses (see Supplemental Material).

In the middle of the LST phase speed image (Fig. 7.1B), particularly to the West, a large fast anomaly is observed between 33.78° and 33.82° latitude. This fast anomaly corresponds well with the Upper Wilmington (UW) Quaternary formation, which includes the Silverado water bearing unit (Lower Pleistocene age, ∼300–580 ka) that supplies nearly 90% of the total ground water extracted in Long Beach [29,30]. Based on a 3D model [30], Fig. 7.2A) of the water-bearing structures in Long Beach obtained using borehole, seismic, and gravity surveys, the Silverado is significantly denser (2,290 kg/m^3) than the surrounding formations (2,050-2,100 kg/m^3) due to its coarse-grained facies. From empirical relations of density and seismic wave speeds, we expect the UW formation to increase V_s by ∼150% relative to the surrounding formations (see Supplemental Material). Since Rayleigh wave phase speed is dependent primarily on V_s, we conclude that this high velocity region of the map likely corresponds to the Silverado aquifer.

The proposed attribution of the fast anomaly to the Silverado aquifer is further supported by simulations the 1 Hz Rayleigh surface wave phase speed (Fig. 7.2(D)) from the Silverado depth range inferred in [30] and [29] (see Supplemental Material). In the region of the survey used from [30], where both Silverado depth and thickness were available, the simulation shows a gradual increase in phase speed from south to north. However, per [29], the Silverado is likely

absent about 1 km south of the NI fault in the region of the Long Beach array (Fig. 7.5). With this assumption for the simulation, the predicted trend and magnitude and phase speeds south of the NI fault compare well with the LST result (Fig. 7.1(B)). Relative to the eikonal tomography phase speed estimates of the Long Beach (Fig. 7.4(A)), the phase speed from LST shows better the broad fast region predicted by the simulation. It is clear that the Silverado is resolved particularly well by the LST in the west-central region. Moreover, well logs support an extension of the high-velocity anomaly toward the NE across the Newport-Inglewood (NI) fault zone, as observed in the LST phase speeds (see Fig. 7.2D), beyond the region of the gravity survey. The LST result appears to corroborate the older study [29], and contradicts some of the results of [30], which was based exclusively on a gravity survey in the region of the Long Beach array.

These results show that the LST method, by assuming that patches of seismic phase speed fields are repetitions of a set of few patterns, contained by the dictionary, can be used to further leverage existing seismic data to obtain high-resolution phase speed images from regions of interest. The dictionary (Fig. 7.3A), since it is learned directly from the phase-speed data via machine learning, is well adapted to the data itself, perhaps more so than predefined functions such as wavelets (Fig. 7.3B,C), and also provides a flexible model that is capable of modeling smooth and discontinuous slowness features. In the context of ambient noise tomography, LST leverages the dense sampling by learning the patterns directly from the cross-correlation data. Thus, we obtain high-resolution slowness maps that well complement other geophysical sensing modalities and existing studies, for better estimating at least near-surface Earth structure. In this application of LST to the Long Beach data, we believe it is likely that we have further-characterized key water-bearing aquifers in Long Beach.

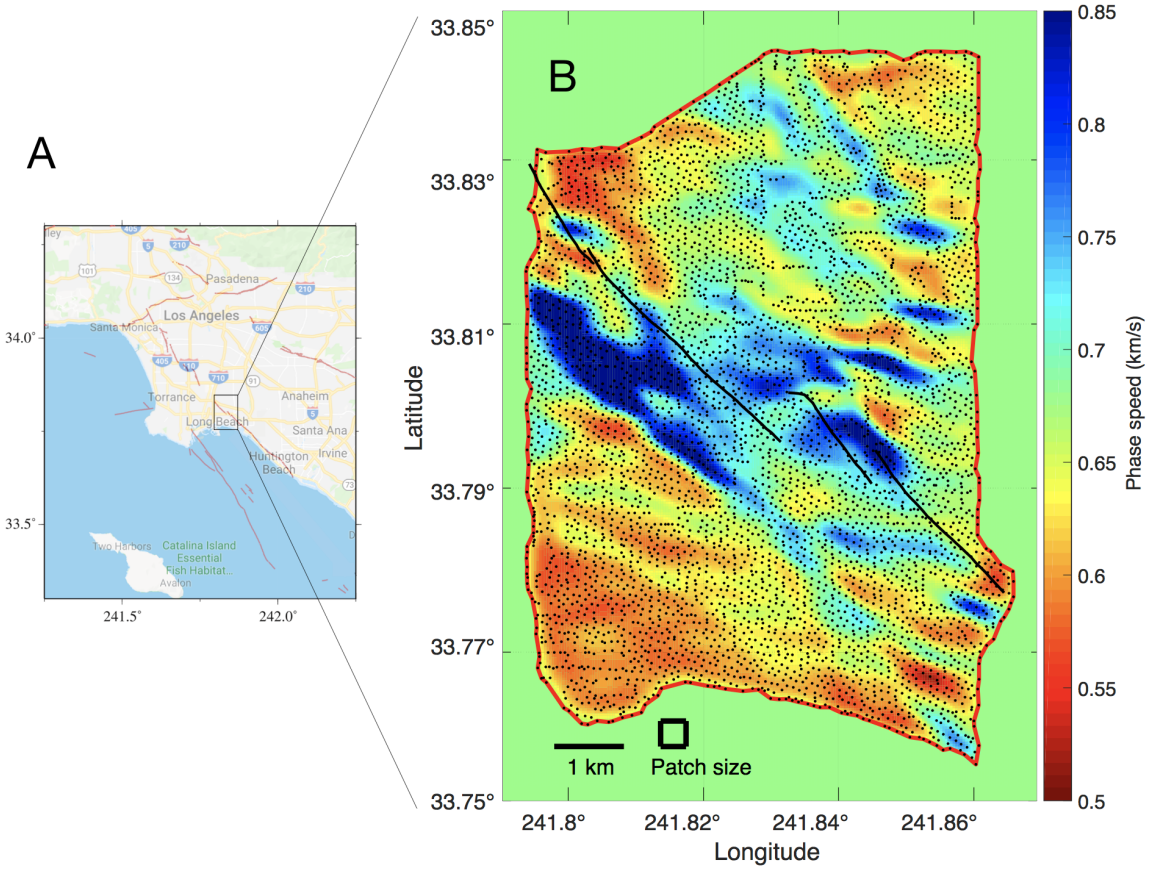


Figure 7.1: The Long Beach array located in (A), contained 5204 stations (dots in (B)) distributed over a ~ 70 km^2 area. (B) Locally sparse tomography (LST) phase speed maps of Long Beach, CA at 300×206 pixel resolution, using 3 million travel times. The Newport-Inglewood (NI) fault network (black lines) and the valid boundary for LST (red line) are shown.

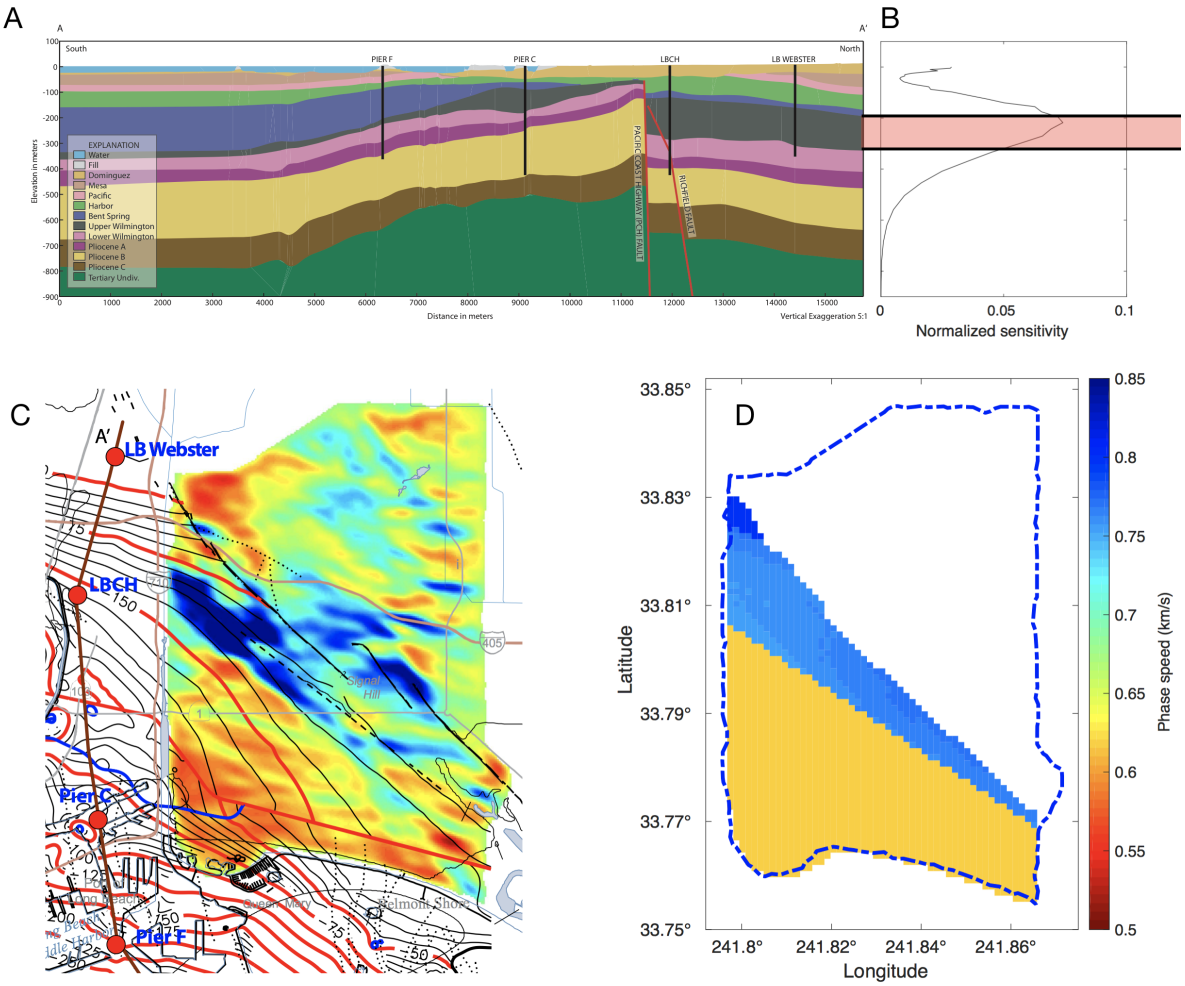


Figure 7.2: (A) Inferred stratigraphy along $\sim$ N-S profile in (C) [30], with the 4 deep wells used as constraints. The wells are located at Pier F, Pier C, Long Beach Cabrillo High School (LBCH), and Long Beach Webster (LB Webster) elementary school, 1 km west of the Long Beach array. (B) 1 Hz average V_s depth sensitivity kernel from [6] and overlap of northern end of the inferred Silverado (shown in (C)) with V_s sensitivity, indicated in red. (C) Inferred Silverado elevation [30] overlain with LST 1 Hz Rayleigh wave phase speed map (Fig. 7.1(B)), with wells (red dots), and the location of the stratigraphy profile (brown line). (D) Hypothesized 1 Hz Rayleigh surface wave phase speed calculated from geological properties and inferred Silverado elevation range (C), from [29, 30].

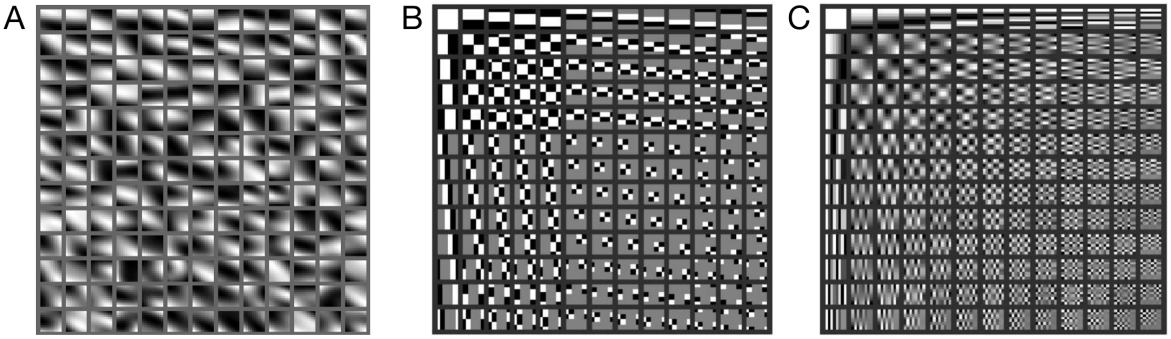


Figure 7.3: Comparison of (A) dictionary learned in LST inversion of Long Beach array data (Fig. 7.1(B), 7.2(C,D)), and generic dictionaries (B) Haar wavelet and (C) discrete cosine transform. All dictionaries shown with 169 atoms ($n = 100$). The atoms are sorted in order of decreasing variance from top to bottom (left to right). The learned dictionary atoms (A) with sharper, oriented gradients (higher variance) correspond to the sharper features in the LST phase speed map (e.g. the boundaries of the Silverado aquifer, Fig. 7.1(B)), whereas the smoother atoms (lower variance) are related to the smoother regions. Atom values stretched to full grayscale (0 to 1) for display.

7.1 Supplementary materials

7.1.1 LST theory and implementation

Our proposed locally-sparse travel time tomography (LST) approach obtains high resolution by assuming that small patches of discrete phase speed maps are repetitions of few elemental patterns from a dictionary of patterns. Such patterns, referred to as atoms (chemistry analogy) are learned in parallel with the inversion using dictionary learning, a form of unsupervised machine learning. Relative to conventional tomography methods, the sparsity of the dictionary representation permits smooth and discontinuous, high-resolution features where warranted by the data. In the following, we present an overview of the LST theory. For more details, please see [16].

In the LST, surface wave propagation is approximated as straight ray paths through an $N = W_1 \times W_2$ pixel phase speed map, and the travel time perturbations $\mathbf{t} \in \mathbb{R}^M$ from a known reference for M rays are modeled as

$$\mathbf{t} = \mathbf{A}\mathbf{s}_\mathrm{g} + \boldsymbol{\varepsilon}, \tag{S1}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the tomography matrix, $\mathbf{s}_\mathrm{g} \in \mathbb{R}^N$ is the perturbation global slowness (inverse of speed), and $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ is Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$. We call Eq. S1 the *global model*, as it captures the large-scale features that span the discrete map and generates $\mathbf{t}$.

We consider a second slowness model perturbation $\mathbf{s}_\mathrm{s} \in \mathbb{R}^N$, called the sparse slowness, in which $\sqrt{n} \times \sqrt{n}$ groups of pixels are represented as sparse linear combinations of atoms from a dictionary. The patches are selected from $\mathbf{s}_\mathrm{s} \in \mathbb{R}^N$ by the binary matrix $\mathbf{R} \in \{0,1\}^{n \times N}$, and modeled as

$$\widehat{\mathbf{x}}_i = \arg\min_{\mathbf{x}_i} \|\mathbf{R}_i \mathbf{s}_\mathrm{s} - \mathbf{D}\mathbf{x}_i\|_2^2 \text{ subject to } \|\mathbf{x}_i\|_0 = T, \tag{S2}$$

where $\mathbf{R}_i \mathbf{s}_s \in \mathbb{R}^n$ selects the i-th patch, $\mathbf{D} \in \mathbb{R}^{n \times Q}$ is the dictionary of Q atoms, $\widehat{\mathbf{x}}_i \in \mathbb{R}^Q$ coefficient estimates for the i-th patch, and T is the number of non-zero coefficients in $\widehat{\mathbf{x}}_i$. We consider all overlapping patches, and wrap the patches at the edges. Thus the number of patches is N. The

ℓ_0 pseudo-norm penalizes the number of non-zero coefficients [22]. We call Eq. S2 the local model, as it captures the smaller scale, localized features contained by patches. The dictionary $\mathbf{D}$ is assumed unknown and is learned from the data during the inversion.

The global (Eq. S1) and local (Eq. S2) models are combined into a Bayesian *maximum a posteriori* (MAP) objective

$$\left\{\widehat{\mathbf{s}}_{\mathrm{g}}, \widehat{\mathbf{s}}_{\mathrm{s}}, \widehat{\mathbf{X}}\right\} = \underset{\mathbf{s}_{\mathrm{g}}, \mathbf{s}_{\mathrm{s}}, \mathbf{X}}{\arg\min} \left\{ \frac{1}{\sigma_{\varepsilon}^2} \|\mathbf{t} - \mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2 + \frac{1}{\sigma_{\mathrm{g}}^2} \|\mathbf{s}_{\mathrm{g}} - \mathbf{s}_{\mathrm{s}}\|_2^2 + \frac{1}{\sigma_{p,i}^2} \sum_i \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i \mathbf{s}_{\mathrm{s}}\|_2^2 \right\} \quad \text{subject to } \|\mathbf{x}_i\|_0 = T \;\forall\, i, \tag{S3}$$

where $\widehat{\mathbf{s}}_{\mathrm{g}}$ is an estimate of the global slowness perturbation, σ_{g}^2 is the global slowness variance, $\widehat{\mathbf{s}}_{\mathrm{s}}$ is the estimate of the sparse slowness perturbation, $\sigma_{p,i}^2$ is the variance of the patch slowness, and $\widehat{\mathbf{X}} \in \mathbb{R}^{Q \times I}$ is the coefficient estimates.

We find the MAP estimates $\left\{\widehat{\mathbf{s}}_{\mathrm{g}}, \widehat{\mathbf{s}}_{\mathrm{s}}, \widehat{\mathbf{X}}\right\}$ using a block-coordinate minimization algorithm by decoupling the local and global models via substitution [22, 23]. The global objective is, from Eq. S3,

$$\widehat{\mathbf{s}}_{\mathrm{g}} = \underset{\mathbf{s}_{\mathrm{g}}}{\arg\min} \; \|\mathbf{t} - \mathbf{A}\mathbf{s}_{\mathrm{g}}\|_2^2 + \lambda_1 \|\mathbf{s}_{\mathrm{g}} - \mathbf{s}_{\mathrm{s}}\|_2^2, \tag{S4}$$

where $\lambda_1 = (\sigma_{\varepsilon}/\sigma_{\mathrm{g}})^2$ is a regularization parameter. The local objective is from Eq. S3, substituting $\mathbf{s}_{\mathrm{s}} = \widehat{\mathbf{s}}_{\mathrm{g}}$,

$$\widehat{\mathbf{x}}_i = \underset{\mathbf{x}_i}{\arg\min} \; \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}\|_2^2 \;\text{ subject to } \|\mathbf{x}_i\|_0 = T. \tag{S5}$$

Dictionary learning is added to the local problem (S5), by optimizing $\mathbf{D}$:

$$\widehat{\mathbf{D}} = \underset{\mathbf{D}}{\arg\min} \left\{ \min_{\mathbf{x}_i} \|\mathbf{D}\mathbf{x}_i - \mathbf{R}_i \widehat{\mathbf{s}}_{\mathrm{g}}\|_2^2 \text{ subject to } \|\mathbf{x}_i\|_0 = T \;\forall\, i \right\}. \tag{S6}$$

The dictionary learning problem (Eq. S6) is here solved using the iterative thresholding and signed k-means (ITKM) algorithm [28]. After $\widehat{\mathbf{D}}$ is obtained, the coefficients $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_I]$ are solved from Eq. S5 using orthogonal matching pursuit (OMP) with the same sparsity level T

as ITKM. Then with $\widehat{\mathbf{X}}$, $\widehat{\mathbf{D}}$, and global slowness $\widehat{\mathbf{s}}_\mathrm{g}$ from Eq. S4 we solve for $\mathbf{s}_\mathrm{s}$. Eq. S3 gives, assuming constant patch variance $\sigma_{p,i}^2 = \sigma_p^2$,

$$\widehat{\mathbf{s}}_\mathrm{s} = \underset{\mathbf{s}_s}{\arg\min}\ \lambda_2 \|\widehat{\mathbf{s}}_\mathrm{g} - \mathbf{s}_\mathrm{s}\|_2^2 + \sum_i \|\mathbf{D}\widehat{\mathbf{x}}_i - \mathbf{R}_i\mathbf{s}_\mathrm{s}\|_2^2, \tag{S7}$$

where $\lambda_2 = (\sigma_p/\sigma_\mathrm{g})^2$ is a regularization parameter. The solution to Eq. S7 is analytic

$$\widehat{\mathbf{s}}_\mathrm{s} = \frac{\lambda_2\widehat{\mathbf{s}}_\mathrm{g} + n\mathbf{s}_p}{\lambda_2 + n}, \tag{S8}$$

where n is the number of patches and $\mathbf{s}_p = \frac{1}{n}\sum_i \mathbf{R}_i^\mathrm{T}\mathbf{D}\widehat{\mathbf{x}}_i$. Eq. S8 gives $\mathbf{s}_\mathrm{s}$ as the weighted average of the patch slownesses $\{\mathbf{D}\widehat{\mathbf{x}}_i\ \forall\ i\}$ and $\widehat{\mathbf{s}}_\mathrm{g}$. When $\lambda_2 \ll n$, $\mathbf{s}_\mathrm{s} \approx \mathbf{s}_p$. When $\lambda_2 = n$, $\mathbf{s}_\mathrm{g}$ and $\mathbf{s}_p$ have equal weight. It is typical in image denoising to set $\lambda_2 = 0$ [24]. The expressions Eq. S4–S8 are solved iteratively until convergence. Before solving Eqs. S5, S6, the slowness patches $\{\mathbf{R}_i\widehat{\mathbf{s}}_\mathrm{g}\ \forall\ i\ \}$ are centered [24], i.e. the mean of the pixels in each patch is subtracted. The mean of patch i is $\bar{x}_i = \frac{1}{n}\mathbf{1}^\mathrm{T}\mathbf{R}_i\widehat{\mathbf{s}}_\mathrm{g}$. Hence, $\mathbf{R}_i\widehat{\mathbf{s}}_\mathrm{g} \approx \mathbf{D}\mathbf{x}_i + \mathbf{1}\bar{x}_i$.

We compare the results of the LST to eikonal tomography [6, 31], Fig. 7.4. Unlike conventional tomography, eikonal tomography avoids the inversion of very large tomography matrices in favor of solving a number of simpler subproblems and accounting partly for ray bending. These subproblems estimate a phase speed surface for each virtual source, using the eikonal equation on a smooth travel time surface, which are later averaged to obtain the overall map. In the case of LST, the global straight ray tomography solution is more similar to conventional than eikonal tomography. However, since relationships between neighboring pixels (e.g. smoothness) are not enforced, the tomography matrix $\mathbf{A}$ is sparse and is inverted more quickly than a non-sparse matrix. Where eikonal tomography enforces smoothness on travel time, LST does not enforce smoothness. It enforces sparsity of the patches.

We quantify the relative quality of the LST and eikonal phase speed maps (Fig. 7.4) by variance reduction, and by visual quality scores derived from natural images. With the

straight-ray assumption, the distribution of the travel time residuals from the reference phase speed are compared with the residuals from the LST phase speed, Fig. 7.7. The LST provides a 57% variance reduction, whereas eikonal gives a 25% reduction. Since there is no true phase speed map available, we use reference-less image quality metrics to help quantify the quality of the LST and eikonal phase speeds. While such metrics may not reflect the truth of estimated geophysical features, they can help quantify corruption of the geophysical features. We use the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [32], the Natural Image Quality Evaluator (NIQE) [33], and the Perception based Image Quality Evaluator (PIQE) [34]. With lower score better, the LST (eikonal) visual quality results were for BRISQUE 45.8 (43.7), for NIQE 5.9 (11.0), and for PIQE 72.5 (86.8). Overall, LST obtains a better score on 2 of 3 of the metrics. The BRISQUE metric has incorporated human opinions of image quality, whereas NIQE and PIQE do not. Hence BRISQUE may bill less suited to our application.

The LST provides improved results with less computational burden. Conventional tomography [35] complexity is dominated by square matrix inversion $O(N^3)$, though approximate solution methods are slightly less complex. For large tomography matrices $\mathbf{A}$, LST is dominated by matrix multiplication in the LSMR algorithm [27] $O(2MN)$. Since for LST the sparse matrix $\mathbf{A}$ is used directly, the memory required for a tomography problem with M travel times scales linearly with the map size N. For conventional, the memory required scales by N^2. Hence LST could be used for much larger maps than conventional tomography.

7.1.2 Geological interpretation

The LST sensitivity depth range ($\sim 100 - 500$ m) with 1 Hz Rayleigh surface waves is occupied by Pleistocene and Holocene deposits [30], which contain almost all the ground water resources in the area, Fig. 7.5(A–C). The Silverado formation (Lower Pleistocene age, $\sim 300 - 580$ ka, named in [29]), accounts for nearly 90% of the total ground water extraction in the region considered here [36]. The producing aquifers consist of sand and gravel, where in

particular the Silverado unit is characterized by coarse-grained sediments.

Ponti et al. (2007) [30] generated a 3D sequence stratigraphy model of Quaternary layers at the Dominguez Gap in a 16.5 x 16.1 km area that overlaps with our LST model region (see Figure 1). The study was aided by 5 reference boreholes drilled to more than 450 m depth (see Figure 1). In addition, more than 300 oil and water wells were compiled and used in the study. An important finding from this study was a fault, striking W-NW in agreement with the general trend of the area, named the Pacific Coast Highway (PCH) fault from its location (see Figure 1), with a progressive vertical throw (up to 200 m) causing displacement of all pre-Holocene formations down to the north (see Figure 2).

An area-wide gravity survey [30] was carried out to invert for stratigraphy along a N-S profile (see Figure 3), $\sim$ 1 km west of the LST result region (see Figure 1). The Silverado aquifer has significantly higher density (2290 kg/m^3), due to its coarse-grained facies, than the surrounding formations (2050–2100 kg/m^3). The Silverado gravity anomaly is in general agreement with the sequence-stratigraphic model (Figure S2), including the termination of the Silverado unit just south of LBCH. However, the Poland et al. (1956) survey [29] contradicts the Ponti et al. (2007) [30] inferred stratigraphy. In [29], it is concluded that the Silverado is missing about 1km south of the NI fault - near the termination of the LST phase speed anomaly.

Several established empirical relations positively correlate seismic wave speeds with density. Gardner et al. [37] relates density ρ to P-wave speed V_p as

$$\rho = 0.31 V_p^{0.25}, \tag{S9}$$

where V_p is in m/s. Eq. S9 gives a 40-50% increase in V_p for the Silverado formation, over the surrounding Pliocene, Pleistocene and Holocene layers. Brocher [38] related V_p, V_s by

$$V_s = 0.7858 - 1.2344 V_p + 0.7949 V_p^2 - 0.1238 V_p^3 + 0.0064 V_p^4, \tag{S10}$$

where V_s is in km/s. Eq. S10 gives a 150% increase in V_s using derived values of V_p [37] for the densities of the Silverado and surrounding layers. Since Rayleigh surface wave phase speed is dependent primarily on V_s [39], these results suggest that the Silverado aquifer should give rise to a significant phase-speed anomaly.

From the predicted V_s for the Silverado layer from the empirical relations Eq. S9, S10, we simulate the 1 Hz Rayleigh surface wave phase speed for the survey region to find the fast anomaly in the LST result (Fig. 7.1(B)). The method and results are summarized in Fig. 7.6. The inferred Silverado elevation and thickness (Fig. 7.6(A,B)) inferred in [30] are interpolated (Fig. 7.6(D,E)) to localize the depth ranges of the Silverado layer in the survey region. It is assumed per [29] that Silverado is missing south of the fast anomaly. The average V_s profile, estimated from Rayleigh wave dispersion measurements in [6], is assumed for the region. For each discrete Silverado depth range, the Silverado V_s is simulated by doubling the average V_s in the Silverado depth range (e.g. Fig. 7.6(C)). The 1 Hz Rayleigh surface wave phase speed for the survey region (Fig. 7.6(F)) is estimated using a numerical forward model [40].

7.2 Acknowledgments

The authors gratefully acknowledge Dan Hollis at NodalSeismic LLC, and Signal Hill Petroleum, Inc., for permitting them to use the Long Beach data. This project is supported by the Office of Naval Research, Grant No. N00014-18-1-2118.

Chapter 7, in part, is a reprint of the material soon to be submitted for publication. Bianco, Michael J.; Gerstoft, Peter; Olsen, Kim B.; Lin, Fan-Chi. The dissertation author was the primary investigator and author of this paper.

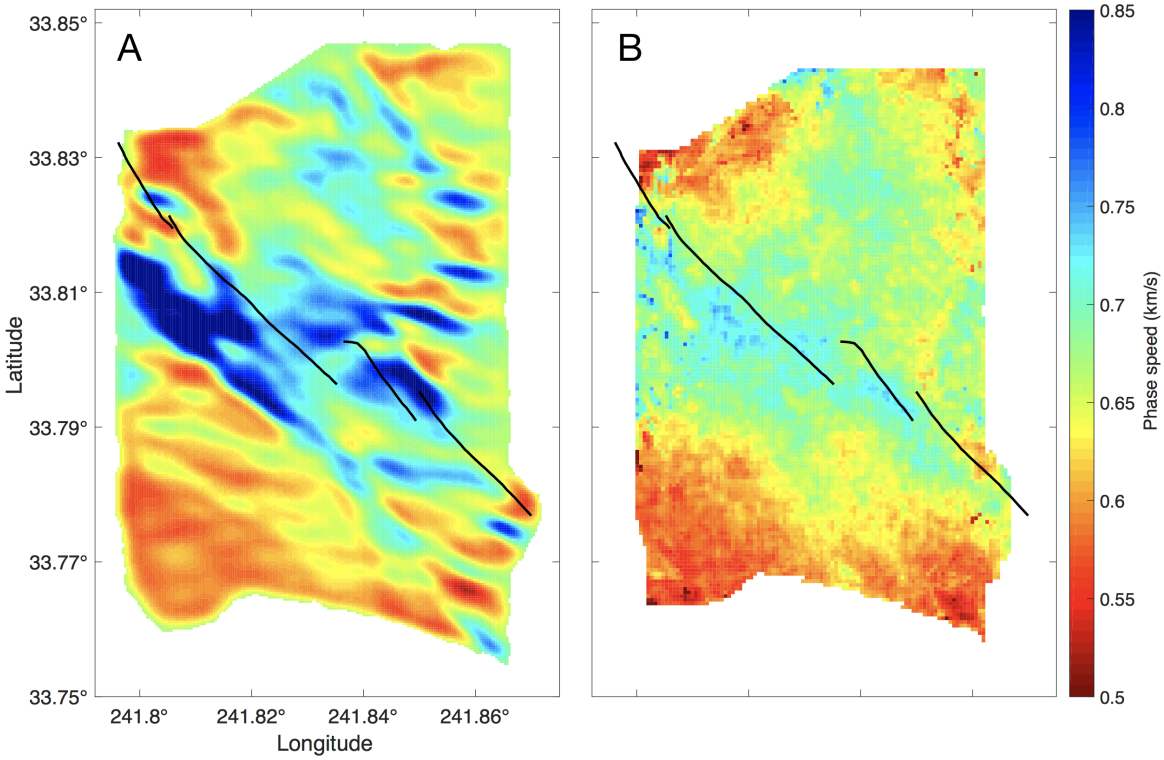


Figure 7.4: Comparison of 1 Hz Rayleigh surface wave phase speed maps from (A) eikonal tomography [6] and (B) LST with the NI fault network (black lines). The general trends are the same for eikonal and LST, though there is greater contrast and phase speed range observed in the LST map. There is greater contrast along the NI fault lines for LST. The largest disagreement the methods is in the western region of the map, where the LST is imaging the Silverado aquifer.

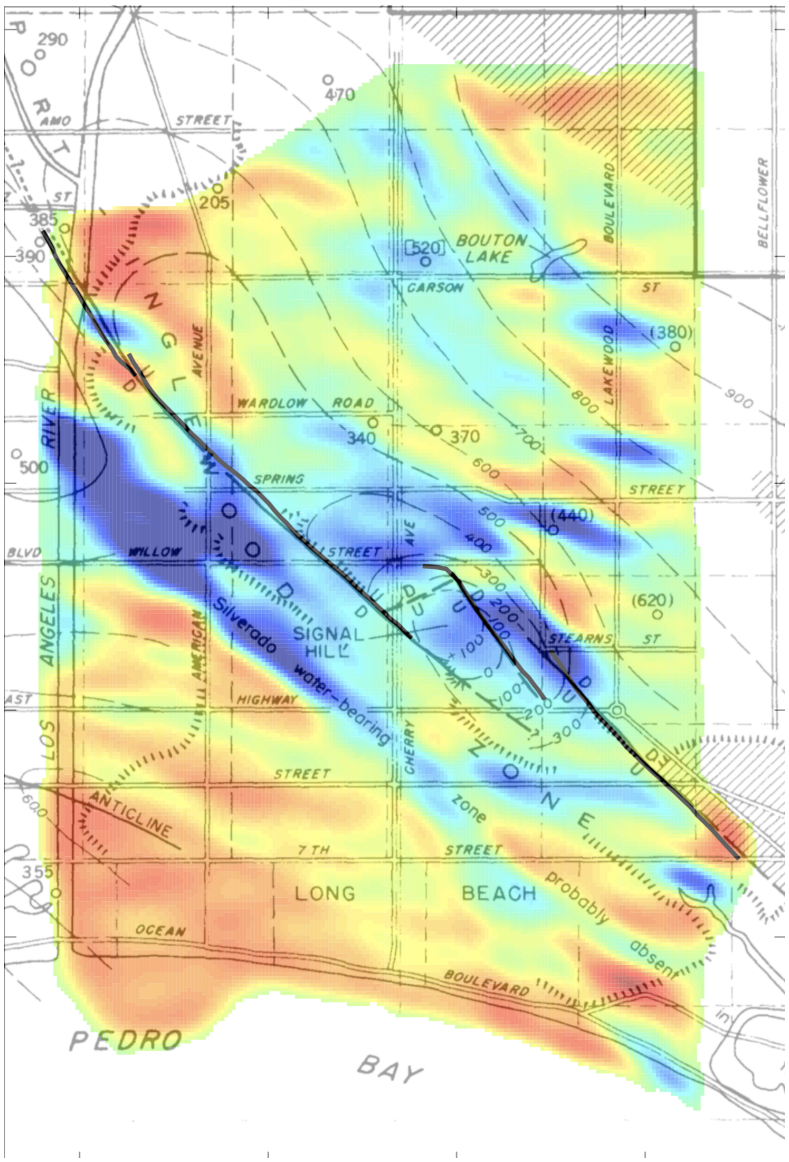


Figure 7.5: Silverado base depth (ft) inferred from borehole measurements in Poland et al. (1956) [29] overlain with LST phase speed (Fig. 7.1B, same colorscale). The results of [29] contradict [30], suggesting Silverado is absent south of the high phase speed anomaly in the west-central part of the LST map. The findings of [29] are corroborated by the LST result. The lower extent of the high-speed anomaly is used to generate a hypothesized phase speed map (Fig. 7.2D). We also note that the phase speed anomalies near and to the north of Signal Hill (indicated above) correlate well with the contours.

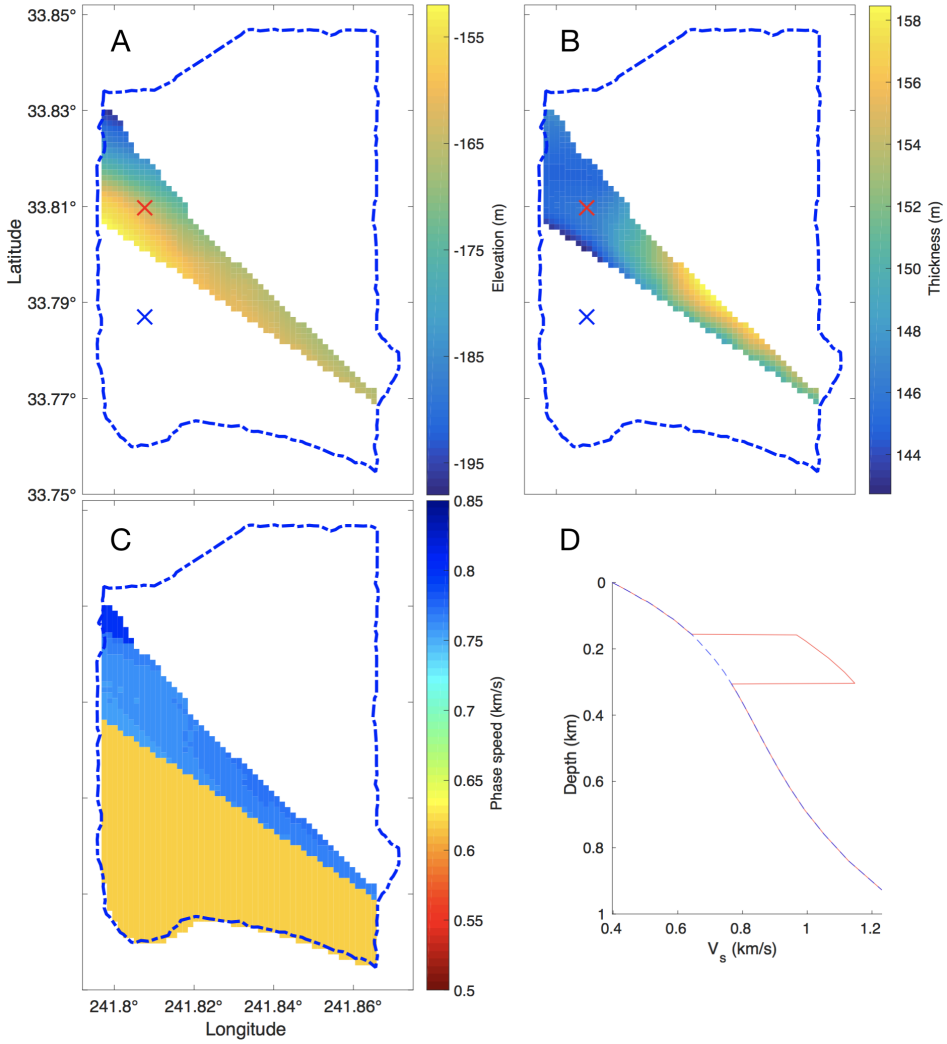


Figure 7.6: Interpolated Silverado (A) elevation and (B) thickness from the Ponti et al. (2007) survey [30], with Silverado missing south of LST high speed anomaly per Poland et al. (1956) survey [29] (see Fig. 7.5). (C) Phase speed estimated from (A,B), based on Silverado phase speed perturbation (150% of average V_s from [6] from Eq. S9, S10). (D) Two example V_s profiles: Silverado V_s predicted by Eq. S9, S10 (red line) superimposed on average V_s from [6] (blue dashed line), over depths corresponding to (A, B), from locations in (A,B) indicated by red and blue x's. Valid boundary for LST inversion shown as blue dashed line.

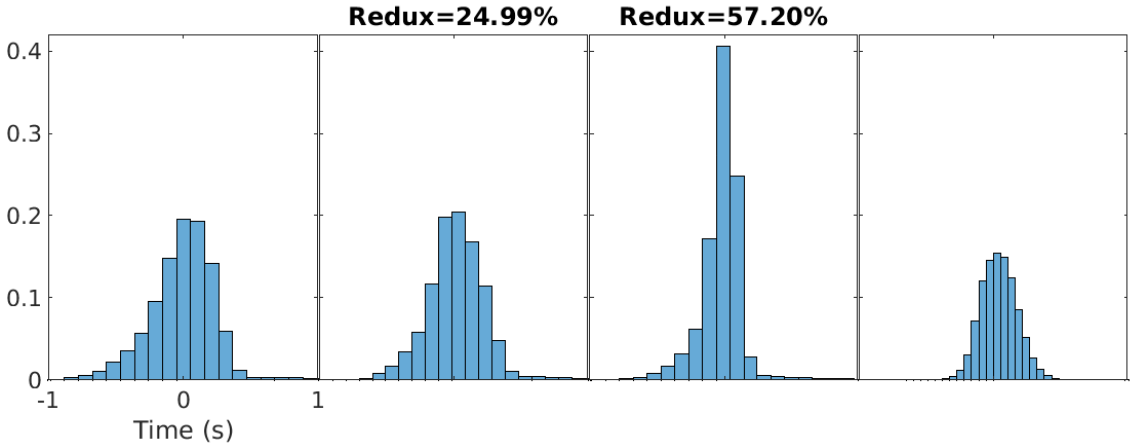


Figure 7.7: From left to right: residual histogram from straight-ray model, eikonal speed map (Fig. 7.4(B)), LST speed map (Fig. 7.4(A)), and residual difference between LST and eikonal tomography.

7.3 Bibliography

[1] O. I. Lobkis and R. L. Weaver, "On the emergence of the Green's function in the correlations of a diffuse field," *J. Acoust. Soc. Am.*, vol. 110, no. 6, pp. 3011–3017, 2001.

[2] N. M. Shapiro, M. Campillo, L. Stehly, and M. H. Ritzwoller, "High-resolution surface-wave tomography from ambient seismic noise," *Science*, vol. 307, no. 5715, pp. 1615–1618, 2005.

[3] K. G. Sabra, P. Gerstoft, P. Roux, and W. A. Kuperman, "Extracting time-domain green's function estimates from ambient seismic noise," *Geophys. Res. Letters*, vol. 32, no. 3, p. L14311, 2005.

[4] G. D. Bensen, M. H. Ritzwoller, M. P. Barmin, A. L. Levshin, F. C. Lin, M. Moschetti, N. M. Shapiro, and Y. Yang, "Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements," *Geo. J. Int.*, vol. 169, no. 3, pp. 1239–1260, 2007.

[5] F. Brenguier, M. Campillo, C. Hadziioannou, N. M. Shapiro, R. M. Nadeau, and E. Larose, "Postseismic relaxation along the San Andreas fault at Parkfield from continuous seismological observations," *Science*, vol. 321, no. 5895, pp. 1478–1481, 2008.

[6] F. C. Lin, D. Li, R. W. Clayton, and D. Hollis, "High-resolution 3D shallow crustal structure in long beach, california: Application of ambient noise tomography on a dense seismic array," *Geophysics*, vol. 78, no. 4, pp. Q45–Q56, 2013.

[7] B. Schmandt and R. W. Clayton, "Analysis of teleseismic p waves with a 5200?station array in Long Beach, California: Evidence for an abrupt boundary to Inner Borderland rifting," *J. Geo. Res. Solid Earth*, vol. 118, no. 10, pp. 5320–5338, 2013.

[8] H. H. Huang, F. C. Lin, B. Schmandt, J. Farrell, R. B. Smith, and V. C. Tsai, "The Yellowstone magmatic system from the mantle plume to the upper crust," *Science*, vol. 348, no. 6236, pp. 773–776, 2015.

[9] N. Nakata, J. P. Chang, J. F. Lawrence, and P. Boué, "Body wave extraction and tomography at Long Beach, California, with ambient?noise interferometry," *J. Geo. Res. Solid Earth*, vol. 120, no. 2, pp. 1159–1173, 2015.

[10] C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," *Sci. Advances*, vol. 1, no. 11, p. e1501057, 2015.

[11] A. Inbal, J. P. Ampuero, and R. W. Clayton, "Localized seismic deformation in the upper mantle revealed by dense seismic arrays," *Science*, vol. 354, no. 6308, pp. 88–92, 2016.

[12] N. Rawlinson, S. Pozgay, and S. Fishwick, "Seismic tomography: a window into deep earth," *Phys. Earth Plan. Inter.*, vol. 178, no. 3, pp. 101–135, 2010.

[13] D. C. Bowden, V. C. Tsai, and F. C. Lin, "Site amplification, attenuation, and scattering from noise correlation amplitudes across a dense array in Long Beach, CA," *Geo. Res. Let.*, vol. 42, no. 5, pp. 1360–1367, 2015.

[14] N. Riahi and P. Gerstoft, "The seismic traffic footprint: Tracking trains, aircraft, and cars seismically," *Geophys. Res. Letters*, vol. 42, no. 8, pp. 2674–2681, 2015.

[15] J. P. Chang, S. A. L. de Ridder, and B. L. Biondi, "High-frequency Rayleigh-wave tomography using traffic noise from Long Beach, California," *Geophys.*, vol. 81, no. 2, pp. B43–B53, 2016.

[16] M. Bianco and P. Gerstoft, "Travel time tomography with adaptive dictionaries," *IEEE Trans. Comput. Imag.*, 2018, DOI: 10.1109/TCI.2018.2862644.

[17] Z. E. Ross, M.-A. Meier, and E. Hauksson, "P wave arrival picking and first-motion polarity determination with deep learning," *Journal of Geophysical Research: Solid Earth*, vol. 123, pp. 5120–5129, 2018.

[18] A. Li, Z. Peng, D. Hollis, L. Zhu, and J. H. McClellan, "High-resolution seismic event detection using local similarity for Large-N arrays," *Nature Scientific Reports*, vol. 8, no. 1, p. 1646, 2018.

[19] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *Sci. Advances*, vol. 4, no. 2, p. e1700578, 2018.

[20] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, "Machine learning predicts laboratory earthquakes," *Geophys. Res. Letters*, vol. 44, no. 18, pp. 9276–9282, 2017.

[21] Q. Kong, R. M. Allen, L. Schreier, and Y. W. Kwon, "MyShake: A smartphone seismic network for earthquake early warning and beyond," *Sci. Advances*, vol. 2, no. 2, p. e1501055, 2016.

[22] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[23] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1028–1041, 2011.

[24] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014.

[25] M. P. Barmin, M. H. Ritzwoller, and A. L. Levshin, "A fast and reliable method for surface wave tomography," *Pure and Appl. Geo.*, vol. 158, no. 8, pp. 1351–1375, 2001.

[26] I. Loris, G. Nolet, I. Daubechies, and F. A. Dahlen, "Tomographic inversion using ℓ_1–norm regularization of wavelet coefficients," *Geo. J. Int.*, vol. 170, no. 1, pp. 359–370, 2007.

[27] D. C. L. Fong and M. Saunders, "Lsmr: An iterative algorithm for sparse least-squares problems," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2950–2971, 2011.

[28] K. Schnass, "Local identification of overcomplete dictionaries," *J. Mach. Learn. Res.*, no. 16, pp. 1211–1242, 2015.

[29] J. F. Poland, A. N. Piper *et al.*, "Ground water geology of the coastal zone, long beach, california," *U.S.G.S. Water Supply*, no. 1109, 1956.

[30] D. J. Ponti, K. D. Ehman, B. D. Edwards, J. C. Tinsley, T. Hildenbrand, R. T. H. J. W. Hillhouse, K. McDougall, C. L. Powell, E. Wan, and M. Land, "A 3-dimensional model of water-bearing sequences in the dominguez gap region, long beach, california," *US Geol. Surv. Open-File Rep.*, no. 2007-1013, 2007, https://pubs.usgs.gov/of/2007/1013/.

[31] F. C. Lin, M. Ritzwoller, and R. Snieder, "Eikonal tomography: surface wave tomography by phase front tracking across a regional broad-band seismic array," *Geo. J. Int.*, vol. 177, no. 3, pp. 1091–1110, 2009.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer." *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[34] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *IEEE Twenty-First Nat. Conf. Comm.* IEEE, 2015, pp. 1–6.

[35] C. D. Rodgers, *Inverse methods for atmospheric sounding: theory and practice*. World Sci. Pub. Co., 2000.

[36] S. D. Thomas, J. M. Liles, and T. A. Johnson, "Managing seawater intrusion in the Dominguez Gap area of Los Angeles County, California, USA," *First Int. Conf. Saltwater Intrusion and Coastal Aquifers, Essaouira, Morocco*, pp. 23–25, 2001.

[37] G. H. F. Gardner, L. W. Gardner, and C. L. Drake, "Formation velocity and density– the diagnostic basics for stratigraphic traps," *Geophys.*, vol. 39, no. 6, pp. 770–780, 1974.

[38] T. M. Brocher, "Empirical relations between elastic wavespeeds and density in the earth's crust," *Bulletin Seis. Soc. Am.*, vol. 95, no. 6, pp. 2081–2092, 2005.

[39] K. Aki and P. G. Richards, *Quantitative seismology*. University Science Books, Mill Valley, CA, USA, 2009.

[40] R. B. Herrmann, "Computer programs in seismology: an evolving tool for instruction and research," *Seism. Res. Letters*, vol. 84, no. 6, pp. 1081–1088, 2013.

Chapter 8

Conclusion

In this dissertation, methods for solving geophysical inverse problems were developed based on sparse modeling theory and dictionary learning, an unsupervised machine learning method. These methods, which employ more sophisticated model priors and latent representations than conventional methods, provided compelling results in the estimation of true ocean sound speed profile (SSP) models in geoacoustic inversion, improved compression of SSP structure over that obtained by empirical orthogonal functions (EOFs), and provide a more flexible travel time tomography framework based on machine learning. The locally sparse travel time tomography (LST) method leverages dense array sampling to learn dictionaries of geophysical features directly from the data under analysis. Thus LST does not require explicit training data.

A method for compressive inversion of ocean acoustic SSPs was developed and demonstrated. With medium SNR, *a priori* knowledge of the ocean sound speed statistics, and a dictionary of shape functions that sparsely represent the SSPs, fine-scale SSP structure was well estimated using compressive sensing (CS). Robust recovery of sparse SSP perturbations was shown using dictionaries containing either half-sinusoidal shape functions or EOFs.

The K-SVD dictionary learning algorithm was applied to ocean SSP data from the High Frequency '97 and South China Sea experiments. It was shown that the learned dictionaries (LDs)

generated describe ocean SSP variability with high resolution using fewer coefficients than EOFs. As few as one coefficient from a LD describes nearly all the variability in each of the observed ocean SSPs. This performance gain is achieved by the larger number of informative atoms in the LD relative to EOF dictionaries. Provided sufficient SSP training data is available, LDs can improve SSP inversion resolution with negligible computational expense. This could provide improvements to geoacoustic inversion, matched field processing, and underwater communication.

A travel time tomography method with a local sparse prior on patches of the slowness image, referred to as the LST algorithm, was derived. The LST can use predefined or learned dictionaries, though learned dictionaries gives improved performance. Relative to the conventional and total variation (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.

LST was tested in the context of surface wave tomography (2D), with both synthetic and real seismic data. For the dense sampling configuration and synthetic images used, LST well recovered true slowness maps with smooth and discontinuous features. LST was further used to perform ambient noise tomography (ANT) with data from the large and dense Long Beach array, installed in Long Beach, CA during 2011. The Long Beach inversion results show that the LST method, by assuming that patches of seismic phase speed fields are repetitions of few patterns, described by the dictionary atoms, can be used to further leverage existing seismic data to obtain high-resolution phase speed images.

LST exploits the dense sampling of ANT on large arrays by learning the patterns directly from the cross-correlation data. Thus, high-resolution slowness maps are obtained that well complement other geophysical sensing modalities and existing studies, for better estimating at least near-surface Earth structure. The LST phase speed map from the Long Beach data is likely to have further-characterized key water-bearing units in Long Beach. The LST is relevant to other tomography scenarios where slowness structure is irregularly but densely sampled, for instance in

ocean and terrestrial acoustics.

Future work could build-upon the approaches developed in this dissertation by further-accounting for the physics of the problems, including generative models and further physical assumptions about the geophysical parameters, and more sophisticated ray propagation modeling. Machine learning-based tomography approaches should leverage physical models to obtain synergy between the physical theory, and the enhanced, data-driven approaches from machine learning techniques. For example, LST in its current formulation [1] does not account for ray bending. Adding more physics to the propagation problem could improve tomographic resolution and characterize uncertainty in ocean acoustic and seismic tomography contexts.

Further, more sophisticated sparse modeling and machine learning methods and theory can be developed to account for the specific geophysical data issues. For instance, missing or incomplete measurements can be accounted-for using inpainting [2–4], a dictionary learning procedure which fixes missing pixels in images. Similar to natural images, geophysical structures in the ocean and solid earth contain local, or small scale, features that can be described by scaled, shifted, and rotated versions of a set of dictionary atoms. Further, these features are repeated or exist at different image resolutions and are correlated. This notion, originally developed in wavelet analysis (from image processing) [5], has given multi-resolution geophysical tomography methods, e.g. [6]. Recently in image processing, a convolutional sparse coding (CSC) [7] algorithm has been developed, which is a specific case of convolutional neural networks [8, 9] that use sparsity regularization on the weights. The CSC model accounts not only for shift variance, but also rotation and scaling of image features, as well as the for correlations between patches. The non-linearity of neural networks and the back-propagation method of their training [10], make neural networks "general function approximators" [11]. That is, given sufficient training data, a neural network can model nearly any function. The non-linear modeling capabilities of neural networks can be leveraged to generate improved data assimilation frameworks.

8.1 Bibliography

[1] M. Bianco and P. Gerstoft, "Travel time tomography with adaptive dictionaries," *IEEE Trans. Comput. Imag.*, 2018, DOI: 10.1109/TCI.2018.2862644.

[2] M. Elad, *Sparse and Redundant Representations*. Springer, New York, 2010.

[3] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014.

[4] V. Naumova and K. Schnass, "Dictionary learning from incomplete data for efficient image restoration," *IEEE 25th Euro. Signal Process. Conf. (EUSIPCO)*, pp. 1425–1429, 2017.

[5] S. Mallat, *A wavelet tour of signal processing*, 2nd ed. Elsevier, San Diego, CA, 1999.

[6] R. Hawkins and M. Sambridge, "Geophysical imaging using trans-dimensional trees," *Geo. J. Int.*, vol. 203, no. 2, pp. 972–1000, 2015.

[7] V. Papyan, Y. Romano, J. Sulam, and M. Elad, "Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks," *IEEE Sig. Process. Mag.*, vol. 35, no. 4, pp. 72–89, 2018.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Info. Process. Sys. (NIPS)*, pp. 1097–1105, 2012.

[10] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[11] K. Murphy, *Machine learning: a probabilistic perspective*, 1st ed. MIT Press, Cambridge, MA USA, 2012.