UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning to Combine: Model Fusion for Computational Imaging, Seismic Model Integration, and Mixture-of-Experts Learning

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

in

Electrical Engineering (Signal and Image Processing)

by

Zheng Zhou

Committee in charge:

Professor Peter Gerstoft, Chair
Professor Wenyuan Fan
Professor Florian Meyer
Professor Truong Quang Nguyen
Professor Kim B. Olsen
Professor Yuanyuan Shi

2024

The Dissertation of Zheng Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

TO MY FAMILY, MY FRIENDS AND COLLABORATORS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thank you to my advisor Prof. Peter Gerstoft for guiding me to complete this PhD degree with your insightful advice on the choice of research path. Sincere thanks to Prof. Kim B. Olsen for sharing his expertise on seismic tomography and for his indispensable assistance in editing the papers as a co-author. My sincere appreciation is also to all professors on my PhD committee including Prof. Truong Q. Nguyen, Prof. Florian Meyer, Prof. Yuanyuan Shi, and Prof. Wenyuan Fan.

Thanks to all my Noiselab labmates for making UCSD fun and memorable. Sincerely hope we could cross paths again.

Thanks to my family for supporting my education and career, and for the unconditional affection you have bestowed upon me.

Profound appreciation is extended to my cherished friends for their unwavering encouragement in my advancement through thorns.

This dissertation is a collection of papers that were published or have been prepared for publication. The text of Chapter 2, in full, is a reprint of the material as it appears in Zhou, Z., Bianco, M., Gerstoft, P. and Olsen, K.B. "High-resolution imaging of complex shallow fault zones along the July 2019 Ridgecrest ruptures", *Geophysical Research Letters*, 49(1), p.e2021GL095024, 2022. The dissertation author was the primary researcher and author of Chapter 2. The coauthors listed in this publication directed and supervised the research.

The text of Chapter 3, in full, is a reprint of the material as it appears in Zhou, Z., Gerstoft, P. and Olsen, K.B. "Fusion of Multi-Resolution Seismic Tomography Maps with Physics-Informed Probability Graphical Models", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 2024, and Zhou, Z., Gerstoft, P. and Olsen, K.B. "Graph-learning approach to combine multiresolution seismic velocity models", *Geophysical Journal International*, p.ggae212, 2024. The dissertation author was the primary researcher and author of Chapter 3. The coauthors listed in this publication directed and supervised the research.

The text of Chapter 4, in full, is a reprint of the material as it appears in Zhou, Z., Gerstoft, P. and Olsen, K.B. "3D multiresolution velocity model fusion with probability graphical models", *Bulletin of the Seismological Society of America*, 114(3), pp.1279-1292, 2024. The dissertation author was the primary researcher and author of Chapter 4. The coauthors listed in this publication directed and supervised the research.

The text of Chapter 5, in full, is a reprint of the material as it will appear in an under-review paper Zhou, Z., Zhang, Z., Li, J., and Gerstoft, P. "Integrating Mixture of Expert Models for Enhanced Predictive Analytics in Fiber Health Monitoring", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 2025. The dissertation author was the primary researcher and author of Chapter 5. The coauthors listed in this publication directed and supervised the research.

VITA

2016	B. S. in Applied Physics, Nanjing University of Science and Technology, China
2019	M. S. in Electrical Engineering (Signal and Image Processing), Rochester University
2024	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California San Diego

PUBLICATIONS

Zhou, Z., Bianco, M., Gerstoft, P., and Olsen, K.B. "High-resolution imaging of complex shallow fault zones along the July 2019 Ridgecrest ruptures", *Geophysical Research Letters*, 49(1), p.e2021GL095024, 2022.

Zhou, Z., Gerstoft, P. and Olsen, K.B. "Graph-learning approach to combine multiresolution seismic velocity models", *Geophysical Journal International*, p.ggae212, 2024.

Zhou, Z., Gerstoft, P. and Olsen, K.B. "3D multiresolution velocity model fusion with probability graphical models", *Bulletin of the Seismological Society of America*, 114(3), pp.1279-1292, 2024.

Zhou, Z., Gerstoft, P. and Olsen, K.B. "Fusion of Multi-Resolution Seismic Tomography Maps with Physics-Informed Probability Graphical Models", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 2024.

Zhou, Z.*, Zhang, Z.*, Li, J., and Gerstoft, P. "Integrating Mixture of Expert Models for Enhanced Predictive Analytics in Fiber Health Monitoring", *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 2025 (under review).

ABSTRACT OF THE DISSERTATION

Learning to Combine: Model Fusion for Computational Imaging, Seismic Model Integration, and Mixture-of-Experts Learning

by

Zheng Zhou

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California San Diego, 2024

Professor Peter Gerstoft, Chair

We conducted ambient noise tomography (ANT) using data from 342 seismographs distributed over a 50×50 km area encompassing the July 2019 Ridgecrest earthquakes (M7.1 and M6.4). By employing the locally sparse tomography (LST) method—an unsupervised machine learning approach—we effectively modeled small-scale geophysical structures using only data from the study region. The Rayleigh group speed derived from LST showed superior accuracy in predicting travel times compared to conventional regularized least-squares inversion techniques. From surface wave dispersion maps, we constructed a 3D shear velocity model, which revealed a heterogeneous low-velocity zone (LVZ) surrounding the causative faults, with

a significant 40% reduction in shear wave velocity primarily concentrated in the upper 2–3 km. Additionally, imaged LVZs associated with inactive portions of the Little Lake Fault System suggest the presence of enduring damage zones.

Building on this work, we introduced a method to fuse multiresolution seismic tomography models with physics-informed probabilistic graphical models (PIPGMs), which incorporate physical constraints like ray-path density. Using synthetic checkerboard models and fault zone structures from the Ridgecrest earthquake sequence, the PIPGM fusion method demonstrated marked improvements in travel-time residuals, image quality, and peak signal-to-noise ratios compared to conventional approaches. This novel fusion framework enables the integration of gridded velocity models of varying resolutions, offering a powerful computational tool for enhancing the quality and interpretability of seismic imaging data. We extended the Probability Graphical Model (PGM) to 3D velocity models combining tasks, including synthetic and real fault zone structures, and demonstrated significant improvements—reducing computed travel-time residuals by 44% for 3D models, respectively, over conventional methods. Unlike traditional techniques, the PGM's adaptive weighting preserves complex features from high-resolution data and effectively transfers these enhancements to the broader low-resolution background, making it a valuable tool for advancing seismic tomography and ground motion simulation.

We applied the newly proposed Mixture of Experts (MoE) to combine the features learned from multiple machine learning models and enhance the capabilities of medical Internet of Things (IoT) systems for blood pressure monitoring by integrating advanced technologies such as flexible sensors, wireless data communication, and intelligent algorithms. This integration facilitates efficient data collection, real-time analytical processing, and seamless online sharing of results. This model employs a dynamic MoE gating network, which intelligently directs incoming data to specialized expert models that are each optimized for distinct measurement scenarios. This model fusion through the MoE framework enables the system to adaptively respond to diverse environmental and individual physiological characteristics, thereby significantly enhancing the accuracy of blood pressure predictions by 15–18% over conventional methods.

Chapter 1

Introduction

We conducted ambient noise tomography (ANT) across a 50×50 km area encompassing the July 2019 Ridgecrest earthquakes (M7.1 and M6.4) using data from 342 seismographs. Employing the locally sparse tomography (LST) method—an unsupervised machine learning approach—we achieved high-resolution imaging of small-scale geophysical structures using only local data. LST outperformed conventional regularized least-squares inversion in predicting Rayleigh wave travel times, revealing a heterogeneous low-velocity zone (LVZ) with up to 40% reductions in shear wave velocity, concentrated in the upper 2–3 km. The imaged LVZs also include inactive portions of the Little Lake Fault System, indicating persistent fault damage.

To enhance seismic imaging further, we developed a physics-informed probabilistic graphical model (PIPGM) framework for fusing multiresolution tomography models. By incorporating physical constraints like ray-path density, PIPGM significantly improved image quality, travel-time residuals, and peak signal-to-noise ratios compared to traditional methods. Synthetic and real 3D fault zone models demonstrated a 44% reduction in residuals, preserving complex high-resolution features while enhancing broader low-resolution areas. This fusion framework provides a robust computational tool for advancing seismic tomography and improving ground motion simulations.

1.1 Ambient Noise Tomography for Ridgecrest

Ambient Noise Tomography (ANT) is a seismic imaging technique that utilizes ambient seismic noise to infer the subsurface structure of the Earth [1]. Unlike conventional seismic methods, which rely on active sources such as explosions or controlled vibrations, ANT exploits naturally occurring background noise, typically generated by ocean waves, atmospheric disturbances, and human activities. This makes it an efficient and non-invasive method for geophysical investigations [2].

The core principle of ANT lies in extracting the Green's function [3] $G(x,t)$ between two stations from the cross-correlation of ambient noise signals recorded at these locations. Mathematically, the cross-correlation [4] $C(\tau)$ between the signals $s_A(t)$ and $s_B(t)$ at stations A and B is given by:

$$C(\tau) = \int_{-\infty}^{\infty} s_A(t) s_B(t+\tau)\, dt, \tag{1.1}$$

where τ is the time lag.

Under ideal conditions, the cross-correlation converges to the Green's function $G(x,t)$, which describes the response of the medium to a point source. This relationship can be expressed as:

$$C(\tau) \propto G(x,t) + G(x,-t), \tag{1.2}$$

where $G(x,-t)$ represents the time-reversed Green's function. The symmetry arises due to the reciprocity of seismic wave propagation.

By analyzing the extracted Green's functions, group and phase velocities of surface waves can be measured. These velocities are frequency-dependent and provide insights into the shear wave velocity structure of the Earth. The phase velocity $c(\omega)$ at angular frequency ω is related to the dispersion relationship:

$$c(\omega) = \frac{\omega}{k(\omega)}, \tag{1.3}$$

where $k(\omega)$ is the wavenumber. These dispersion curves are inverted to obtain subsurface shear velocity profiles.

In practical applications, a dense network of seismic stations enables the construction of topographic maps, revealing lateral and vertical variations in the Earth's properties [5]. The resolution of these maps depends on the density of station coverage and the quality of ambient noise data. Solving ANT and the related inverse problem requires powerful techniques that utilize advanced mathematical models and computational algorithms to extract detailed information from data that would otherwise be inaccessible. These technologies continue to evolve, driven by advances in computational power and algorithmic innovation [6], broadening their application scope and enhancing their resolution and accuracy.

In Chapter 2, we investigated the complex shallow fault zones associated with the 2019 Ridgecrest earthquakes (M6.4 and M7.1) through high-resolution seismic imaging. Using data from 342 seismographs distributed across a 50×50 km area, the study applies ambient noise tomography (ANT) with a machine-learning-based locally sparse tomography (LST) approach. This innovative methodology improves upon traditional regularized least-squares inversion by effectively modeling small-scale geophysical structures using unsupervised machine learning.

The results highlight the heterogeneous and enduring nature of fault damage zones, with some LVZs correlating with inactive faults, suggesting the long-term persistence of these structures. This has profound implications for understanding fault mechanics and improving seismic hazard analysis. Our proposed method contributes a powerful computational approach for seismic imaging and a deeper understanding of fault zone structures, providing valuable tools for both academic research and practical seismic risk assessment.

1.2 PGMs for 2D Model Fusion

Probabilistic Graphical Models (PGMs) are a comprehensive way to represent and manipulate joint probability distributions. They are composed of vertices representing random

variables and edges representing probabilistic dependencies between these variables [7]. Among the two main types of graphical models—Bayesian Networks and Markov Random Fields—the latter is especially significant in the context of image processing due to its suitability for modeling undirected relationships.

Markov Random Fields (MRFs) are a subset of PGMs used extensively for spatial data analysis, particularly in image smoothing and restoration [8]. An MRF is defined over an undirected graph where each node corresponds to a random variable representing some attribute of the image, such as pixel intensity. The edges encode the conditional dependencies between these variables [9]. In the realm of image smoothing, the goal is to recover a noise-free image from a noisy observation by modeling the pixel values as an MRF [10]. It finds applications in various related fields such as computer vision [11], denoising problems [12], and spatial signal processing [13]. An MRF captures the local conditional independence properties of random variables, making it a powerful tool for modeling joint probability distributions over a set of variables.

Let $G = (V, E)$ be an undirected graph, where V represents the set of nodes (random variables) and E represents the edges that define dependencies between these variables. A random vector $\mathbf{X} = \{X_v : v \in V\}$ is said to form a Markov Random Field with respect to G if it satisfies the following Markov properties:

- **Pairwise Markov Property:** For any two non-adjacent nodes u, v, the variables X_u and X_v are conditionally independent given all other variables:

$$P(X_u, X_v \mid \mathbf{X} \setminus \{X_u, X_v\}) = P(X_u \mid \mathbf{X} \setminus \{X_u, X_v\}) P(X_v \mid \mathbf{X} \setminus \{X_u, X_v\}). \tag{1.4}$$

- **Local Markov Property:** Each node v is conditionally independent of all other nodes given its neighbors $N(v)$:

$$P(X_v \mid \mathbf{X} \setminus \{X_v\}) = P(X_v \mid \{X_u : u \in N(v)\}). \tag{1.5}$$

- **Global Markov Property:** For any two sets of nodes A and B, if all paths between A and B are blocked by a separating set C, then A and B are conditionally independent given C:

$$P(\mathbf{X}_A, \mathbf{X}_B \mid \mathbf{X}_C) = P(\mathbf{X}_A \mid \mathbf{X}_C) P(\mathbf{X}_B \mid \mathbf{X}_C). \tag{1.6}$$

The joint probability distribution of an MRF can be expressed in terms of potential functions $\psi_c(\mathbf{X}_c)$, defined over the cliques c of the graph.

We explore the integration of multi-resolution seismic velocity models to enhance the accuracy of seismic imaging and ground motion simulations in Chapter 3. Addressing the challenges of varying spatial resolutions due to differences in inversion methods, data coverage, and ray-path density, the study introduces a novel Physics-Informed Probabilistic Graphical Model (PIPGM). This framework combines high- and low-resolution models while preserving fine-scale details and ensuring smooth transitions between regions of varying resolution.

The proposed PIPGM framework incorporates physical constraints, such as ray-path density and gradient information, to adaptively enhance low-resolution models using high-resolution data. This approach reduces artifacts, sharp boundaries, and misalignments, improving overall image quality and interoperability. Evaluations using both synthetic checkerboard models and real fault-zone structures from the 2019 Ridgecrest earthquake sequence demonstrate the efficacy of PIPGM. Compared to conventional methods, PIPGM achieves a 38% reduction in travel-time residuals and superior performance in preserving fine-scale features. This work underscores the potential of combining computational techniques with physics-based models to improve seismic imaging, particularly for applications in fault zone analysis and earthquakes.

1.3 3D model Fusion with Generalized PGM

The fourth chapter focuses on the integration of multiresolution seismic velocity models, a critical challenge in geophysical research with applications in ground motion simulations and earthquake hazard analysis. Using Probability Graphical Models (PGMs), this study develops

a novel framework that addresses the limitations of traditional velocity model fusion methods like Gaussian smoothing and cosine tapering in 3D model fusion. The proposed approach leverages high-resolution (HR) subdomains to enhance low-resolution (LR) regions, preserving fine structural details while maintaining global model coherence.

In Chapter 4, we proposed a Bayesian Maximum A Posteriori (MAP) problem formulated to integrate HR and LR data using PGMs. The framework effectively captures spatial correlations and anisotropy, enabling accurate reconstruction of complex geological structures. The model employs a Markov Random Field (MRF)-based representation to iteratively adjust pixel values and cluster labels, ensuring smooth transitions across resolution boundaries. The integration process uses Gibbs sampling and Expectation-Maximization (EM) techniques [14] to optimize velocity distributions. The efficacy of the 3D PGM approach is demonstrated through synthetic checkerboard models and real-world data from the 2019 Ridgecrest earthquake sequence. Comparative analysis with benchmark methods (Gaussian filters, cosine tapering, and dictionary learning) reveals a 44% reduction in travel-time residuals for 3D models, underscoring the PGM's superior accuracy and resolution.

We also introduce a transformative computational approach to seismic tomography, providing a robust tool for integrating diverse geophysical datasets. It lays the groundwork for future advancements in seismic imaging, enabling more precise characterizations of Earth's subsurface and enhancing the predictive accuracy of seismic hazard models.

1.4 MoE-Based Fusion method from Multi-model Learning

Multi-model learning involves the integration of multiple learning algorithms to improve predictive performance, robustness, and generalization over what could be achieved with a single model. This approach is particularly useful in scenarios where different models capture various aspects of the data, leading to a more comprehensive understanding of the problem space. One prominent framework within multi-model learning is the Mixture of Experts (MoE), which

dynamically allocates the task of prediction among an ensemble of expert models [15], each specialized in different regions of the input space.

The MoE model is essentially a form of ensemble learning where each expert's prediction is weighted by a gating mechanism [16] that determines the relevance of each expert for a given input. The overall output of the system is a weighted combination of the experts' outputs. Mathematically, this can be represented as:

$$y = \sum_{i=1}^{K} g_i(\boldsymbol{x}) f_i(\boldsymbol{x}), \tag{1.7}$$

where y is the output, $\boldsymbol{x}$ is the input, $f_i(\boldsymbol{x})$ are the outputs of the K expert models, and $g_i(\boldsymbol{x})$ are the gating functions' outputs, which are typically implemented as softmax functions [17] to ensure that the gating outputs sum to one. Each expert can be a different model (e.g., linear models, neural networks), and the gating functions themselves can be learned alongside the experts, often using techniques such as Expectation-Maximization or backpropagation in a neural network context.

The strength of the MoE model lies in its ability to learn complex, non-linear decision boundaries by partitioning the input space into regions that are easier to model, effectively allowing experts to specialize. This specialization can lead to better performance on complex tasks that may be too challenging for a single model to handle effectively. MoEs have been successfully applied in various domains, including speech recognition, image classification, and natural language processing, where the diversity of data and task complexity make it a particularly suitable choice.

In Chapter 5, we developed an enhanced medical Internet of Things (IoT) system for blood pressure monitoring. This enhancement is achieved through the integration of fiber optic sensors, wireless data transmission, and intelligent algorithms. These technologies allow for efficient data collection, real-time analysis, and online result sharing, overcoming challenges faced by traditional blood pressure measurements which can be affected by individual physiological

differences such as skin conditions and body shapes.

The key innovation in the paper is the application of a Mixture of Experts (MoE) model within the IoT system, which improves blood pressure prediction accuracy by 15–18% compared to traditional methods [18]. The MoE model uses a gating network that dynamically adapts to varying measurement environments and individual characteristics by directing input data to specialized experts fine-tuned for specific scenarios. This methodology allows the system to handle complex scenarios more effectively by utilizing the strengths of various expert models in response to the dynamic nature of physiological and environmental changes. This adaptive capability makes the MoE model particularly effective in enhancing the predictive performance and adaptability of the health monitoring system.

Chapter 2

Imaging the Velocity Model around the Ridgecrest Area with Ambient Noise Tomography (ANT)

2.1 Introduction

The 2019 Ridgecrest earthquakes (M6.4 and M7.1) [19] in the Eastern California Shear Zone (ECSZ) highlight the need for accurate fault zone imaging to assess seismic hazards. These events occurred in zones with extensive distributed faulting, indicating underlying low-velocity zones (LVZs). Fault damage zones, characterized by depth-dependent dimensions and velocity reductions, influence seismic wave propagation, ground motion amplification, and rupture dynamics.

We perform ambient noise tomography (ANT) of the Ridgecrest region. Conventional 3D and machine learning-based 2D surface wave tomography are used to obtain a data-driven velocity model which helps illuminate the extent of the fault zone damage. Geological mapping studies [20] find extensive distributed faulting in a zone several kilometers wide around the 2019 Ridgecrest ruptures. While the depth extent is unknown, we assume that this documented off-fault deformation in Ridgecrest is the surface expression of underlying fault damage characterized by a low-velocity zone (LVZ).

Ambient noise is a reliable tool for seismic imaging of Earth's crust, due to the abundance

of shallow surface waves. Cross-correlations of noise obtained between station pairs provide estimates of group or phase velocities of the intervening media. The accuracy of the ANT depends on the spacing of the virtual-source receiver pairs. Deployments of dense seismic arrays have provided particularly high-resolution measurements for recent near surface tomography studies [21, 22, 23]. Rayleigh waves, with prominent significance on vertical-component records, are sufficient to obtain an abundance of ray paths. The ambient noise cross-correlation leads to depth-dependent velocity structure using surface wave dispersion inversion [24, 25].

The depth resolution of tomography depends on the frequency range used. We apply ambient noise correlation techniques to a rich dataset collected during 58 days in 2019 (07/13–09/08), consisting of a combination of a coarse regional deployment and dense 2D arrays across the M6.4 and M7.1 July 2019 Ridgecrest surface ruptures [26]. Specifically, we estimate the S-wave velocity structure to 1 km depth using 0.02–1 Hz Rayleigh waves. Group travel-times are extracted from narrow-band cross-correlation peaks. From these, 2D Rayleigh wave velocity models are estimated. Surface-wave dispersion analysis across a broad frequency range is then carried out at each grid point providing a local shear wave profile using a fully nonlinear direct-search algorithm [27]. Finally, we stack the resulting S-wave profiles to construct a 3D S-wave model from the surface to a depth of 5 km.

Typically, ANT is performed using regularized least squares and travel-time residuals or Eikonal equation inversion [28]. Recently, probabilistic neural network [29], random mesh projection [30] and dictionary learning [31, 32] have been used to generate the velocity models. [32] proposed improving tomography models by incorporating unsupervised machine learning. The method, locally sparse travel time tomography (LST), was demonstrated on a Nodal seismic array dataset from Long Beach, CA 2011 [23]. The LST approach separates the overall or *global* tomography map into patches of small-scale or *local* variation, and considers the local and global information separately [32]. LST, via sparse modeling, is capable of resolving sharp as well as smooth Earth features based on the data. Using the method for fault-zone imaging studies has the potential to improve resolution, for example at the boundaries of the LVZ, over conventional

tomography approaches. We resample the travel times using cross-validation resampling to compare two ANT approaches: LST and Gaussian smoothing-kernel regularized least squares (LSQ) tomography (Eq. S7). For both approaches, the parameters are chosen to minimize the travel time residual on withheld data (data not used to calculate model parameters).

2.2 Data Processing and Methodology

2.2.1 Data

We used the continuous seismic records (500 Hz sampling rate) from 07/13/2019 to 09/08/2019 around the Ridgecrest area by Catchings et al. (2020) on 342 seismic sensors, 65 nodes in a regional array with 8–15 km spacing and 277 nodes in 6 dense arrays with 60–170 m spacing (A1, A2, and B1–B4 in Fig. 2.1(a)). All seismic data are from the FDSN 3J network.

2.2.2 Noise Tomography with Locally-sparse Model

We first bandpass-filtered the data using a 0.02—1.00 Hz 2nd-order Butterworth filter and down-sampled the records to a 5 Hz sampling rate. An amplitude truncation technique was applied for removal of aftershocks [3], and spectral whitening [4] was then used to create data with a uniform amplitude spectrum. The data from every station pair were cross-correlated segment by segment, and the resulting correlograms were stacked over 6-hour intervals. Finally, a thresholding policy was adopted for quality control (see Section S1 for more details).

In conventional tomography, the least squares fit between observed and calculated travel times is minimized with regularization from the smoothness prior. Here, the slowness pixels are represented by the vector $\mathbf{s}' = \mathbf{s_0} + \mathbf{s_g} \in R^N$, where $\mathbf{s_0}$ is a reference slowness and $\mathbf{s_g}$ is a perturbation from the reference. The travel time observations are $\mathbf{t}' = \mathbf{t_0} + \mathbf{t_g} \in R^N$, where $\mathbf{t_0}$ and $\mathbf{t_g}$ are the reference travel times and perturbations. Since the reference slowness $\mathbf{s_0}$ is known, the reference travel time $\mathbf{t_0}$ is $\mathbf{t_0} = \mathbf{A}\mathbf{s_0}$, where $\mathbf{A} \in R^{M \times N}$ denotes the tomography matrix. We

estimate the perturbations by

$$\mathbf{t} = \mathbf{A}\mathbf{s}_\mathbf{g} + \varepsilon, \tag{2.1}$$

where $\varepsilon \in R^M$ is noise.

As a refinement of conventional tomography, LST introduces an additional slowness vector $\mathbf{s}_\mathbf{s} \in R^N$, which captures the local slowness features [32] using sparse modeling and dictionary learning. Here, $\mathbf{D} \in R^{n \times Q}$ is a dictionary of Q atoms, and $\mathbf{x}_i \in R^n$ are the sparse coefficients with n being the number of pixels in a patch. R is a binary matrix which selects a particular patch from $\mathbf{s}_\mathbf{g}$, and the Bayesian maximum a posteriori (MAP) estimations $\{\widehat{\mathbf{s}}_\mathbf{g}, \widehat{\mathbf{s}}_\mathbf{s}, \widehat{\mathbf{X}}\}$ are formulated as

$$\begin{aligned} \{\widehat{\mathbf{s}}_\mathbf{g}, \widehat{\mathbf{s}}_\mathbf{s}, \widehat{\mathbf{X}}, \mathbf{D}\} = \underset{\mathbf{s}_\mathbf{g}, \mathbf{s}_\mathbf{s}, \mathbf{X}, \mathbf{D}}{\arg\min} \{ & \left\| \mathbf{w}(\mathbf{t} - \mathbf{A}\mathbf{s}_\mathbf{g}) \right\|_2^{\,2} + \lambda_1 \left\| \mathbf{s}_\mathbf{s} - \mathbf{s}_\mathbf{g} \right\|_2^{\,2} + \lambda_2 \sum_i \left\| \mathbf{D}\mathbf{x}_i - \mathbf{R}_i \mathbf{s}_\mathbf{s} \right\|_2^{\,2} \} \\ & \text{subject to } \|\mathbf{x}_i\|_0 = T \ \forall i, \end{aligned} \tag{2.2}$$

with $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_I]$ and λ_1, λ_2 are hyperparameters. The dictionary $\mathbf{D}$ is learned during the optimization to represent local geophysical features with the sparse model. $\mathbf{w}$ is a weight matrix which normalizes the biased ray-azimuth distribution. Such normalization is necessary because the seismic stations within the deployment of coarse regional and dense 2D arrays (Fig. 2.1(b)) are highly unevenly-distributed, causing a significant bias of the the ray-path with certain azimuths. The normalization term $\mathbf{w}$ generates an unbiased azimuth distribution among the stations involved in our imaging (see Fig. 2.1(c)). We set $l = 10$ pixels (corresponding to $\sim$4 km) as our patch size, and $\lambda_1 = 1.58$ and $\lambda_2 = 0.40$ are the optimal hyperparameters picked from cross-validation tests (see Section 2.3 and the Supporting Information).

2.3 Cross-validation: Model Selection

Fixed training datasets can be subdivided or resampled to obtain statistics of the data and models designed to explain them [33]. Cross-validation is one such technique, which is often

used for selecting parameters for a model based on its performance on unseen samples [34]. In cross-validation, a given dataset is divided into multiple training and validation datasets, called folds. The model weights are estimated using the training data, and its performance is evaluated on the validation set for each fold. The purpose of this test is to evaluate the model's ability to predict new data that was not used for training, over a sequence of different subdivisions of the data. This approach provides insight on the model's generalization to unseen data samples, and can indicate if the model is over- or underfitting the training data. In this work, models are chosen based on minimum average error on the validation set over the individual folds. Furthermore, the velocity models from the individual folds are used to obtain model uncertainty.

To evaluate the quality of velocity models obtained by the LST algorithm, we compare results from the LST method along with the least-squared (LSQ) tomography method [35].

Both LST and LSQ methods are used with a 10-fold cross-validation — we evenly sampled the station-wise travel time pairs into 10 folds and the sampled data in each fold have the same azimuth distribution (see Fig. 2.1(c)) as the original data to avoid any direction-based bias. The LSQ model is solved assuming a constant reference velocity. The LST models are initialized using either a constant reference velocity, or the LSQ velocity model. The LST dictionary is initialized using Gaussian random atoms. Then, our LST and LSQ models are iteratively trained on 9-fold data and tested on the remaining fold. Our tests are performed on 0.7–0.9 Hz, 0.5–0.7 Hz, and 0.2–0.5 Hz data and we use the normalized travel time residual (NTTR) as the evaluation loss

$$loss_{\mathrm{NTTR}} = \left\|\mathbf{w}(\mathbf{t}-\hat{\mathbf{t}})\right\|_2^2 = \sqrt{\sum_i \mathbf{w}_i(\mathbf{t}_i-\hat{\mathbf{t}}_i)^2}, \tag{2.3}$$

where $\mathbf{w}$ is the direction normalization vector for each fold, $\mathbf{t}$ is the observed real travel time vector, and $\hat{\mathbf{t}} = \mathbf{A}\mathbf{s}_\mathbf{s}$ is the predicted travel time vector from tomography results. $\mathbf{w}_i$, $\mathbf{t}_i$ and $\widehat{\mathbf{t}}_i$ are the i-th elements in $\mathbf{w}$, $\mathbf{t}$ and $\hat{\mathbf{t}}$.

We summarize the averaged (mean) and the standard deviation of the Rayleigh group velocity (inverse slowness) maps for the 10-fold tests in different frequency bands in Fig. 2.2(a)–

(b), and the 10-fold mean and standard deviation of the NTTR losses in Fig. 2.2(c). Lower averaged loss indicates tomography maps which better predicted the travel time, and lower loss variance denotes higher confidence and more stable results. The LST method outperforms conventional LSQ on these terms. Initializing the LST learning algorithm with a good prior, which starts the learning algorithm closer to the optimal solution, can further improve the travel time residuals (see Fig. 2.2(a)). The lower frequency band corresponds to deeper velocity structure. We observe that the average velocity increases and the width of the LVZs shrinks as the sensitivity zone goes deeper. Generally, the LST provides lower velocities inside the LVZs, which results in a stronger contrast between the LVZs and the surrounding areas, and also sharpens the boundaries of the LVZs. Since each fold-test drops 10% of the ray paths randomly, the smaller standard deviation indicates that the LST model can provide results with higher confidence and stability (see Fig. 2.2(b)).

2.4 Synthetic experiments

2.4.1 2D Ambient Noise Tomography

Fig. 2.3(a) shows the Rayleigh group velocity map generated by the averaged 10-fold cross-validation tests on LST, representing a depth-average captured by the resolution kernel for 0.02–1 Hz, namely the upper $\sim$1 km. The image shows relatively disjoint and spatially complex low-velocity regions on the map, in particular surrounding the surface rupture of the M7.1 and M6.4 Ridgecrest earthquakes. The average velocities of the surrounding material and the low-velocity zones are 2.8 and 1.7 km/s, respectively, and the width of these zones vary from 0 to 5 km.

The phase gradients mapped from daily passes of the Sentinel satellites [36] are superimposed on the group velocity map in Fig. 2.3(a), indicating distributed faulting from the 2019 Ridgecrest events. There is a strong correlation between the LVZs and these phase gradients, suggesting that the distributed faulting occurred inside a kilometer-wide deformation zone.

Based on the Rayleigh wave group velocity resolution kernels, each frequency band senses different depths of the the shear wave velocity [37]. For the bands in Fig. 2.3(c)–(d) the peak depths are 300 m (0.7–0.9 Hz), 450 m (0.5–0.7 Hz), and 700 m (0.2–0.5 Hz). Notice that the average velocity of the LVZs decrease/width of the LVZs shrinks as the depth increases, in agreement with the expected localization at greater depth found by other fault zone studies such as [21, 22].

2.4.2 Surface Wave Dispersion Inversion

The Rayleigh wave group velocities obtained by cross-correlation of ambient noise at each station pair are used to estimate group velocity dispersion curves. We use the fundamental mode dispersion curves picked in the frequency range of 0.1–3.0 Hz, and the estimated group velocities are inverted for a 1D shear-wave profile using a stochastic damped least-squares inversion that minimizes the L2 norm misfit between synthesized and observed data (We use the SURF96 program [37]. Examples of dispersion curves and the resulting S-wave profile are shown in Fig. S3.). The technique [37] allows for the evaluation of partial derivatives of the Rayleigh wave group velocities with respect to the S-wave velocity and density for each layer. The model parameters are iteratively perturbed from the initial guess with a starting model taken from the Southern California Earthquake Center Community Velocity Model Version S-4.26 [38], which generally converges after a few iterations. The shear velocity profile is estimated for station pairs generated by every other station. The regions on either side of the surface rupture are modeled separately because of differences in the elevation of the stations (low elevations on the western side and high elevations on the eastern side).

Fig. 2.4(a) shows cross sections of shear wave velocities from the inverted surface wave dispersion curves at the dense arrays A1, A2, B1, B2, B3, and B4, and a composite 3D image of shear wave velocities obtained by the sparse arrays is shown in Fig. 2.4(b). We observe relatively wide LVZs in the upper 1 km range of Fig. 2.4(a) and Fig. 2.4(b) in our LST results. The surface traces from the M6.4 and M7.1 Ridgecrest events intersect the cross sections near

the center of the LVZs, and the shallow cross sections reveal distinct LVZ flower structures, as observed by [22]. However, the complexity of the fault zone, as characterized from the variation of particularly the width of the LVZ, is remarkable. For example, the fault zone at arrays A1 and B1 appear to delineate two or more separate low-velocity parts of the fault zone, which may represent a concentration of damage along different locations of the rupture for past events.

2.5 Conclusions

Our tomographic model of the Ridgecrest area obtained using ambient noise and the LST approach resolves up to 5 km-wide flower-shaped LVZs with a velocity contrast of 40% to the host rock. The imaged LVZs are intersected by the fault traces of the M7.1 and M6.4 Ridgecrest events, suggesting an origin as a fault damage zone. It is unlikely that sedimentary deposits significantly contributed to the LVZs, which in several areas are located on topographic relief (refer to Fig. 2.1(a)). In addition, the LVZs are relatively uncorrelated with depths to bedrock obtained from inversion of gravity data [39]. We have shown that our LST method, by being primarily data-driven, can yield improved seismic velocity images over conventional approaches in terms of smaller predicted travel time residuals. We further obtained data-driven estimates of the model uncertainty with a cross-validation resampling method. It was observed that the model obtained with LST had lower uncertainty than that from regularized LSQ.

2.6 Supplemental materials

2.6.1 Data Preprocessing

The cross-correlation is performed on the 6-hour interval-pairs from two stations, and the highest value of the absolute value of the cross-correlation curve is located to extract the travel time. The final travel time between the two stations is averaged over all 6-hour interval-pairs. Based on the Rayleigh wave travel time and the distance between the station-pairs, the averaged travel speed is calculated. We adopted the thresholding policy for the quality control. All the

station-pairs which provide an averaged-speed higher than 3.5 km/s or slower than 1 km/s are removed.

2.6.2 Locally Sparse Model

The global problem can be written as

$$\widehat{s_g} = \arg\min_{s_g} \frac{1}{\sigma_\varepsilon^2} \left\| W(t - As_g) \right\|_2^2 + \frac{1}{\sigma_g^2} \left\| s_g - s_s \right\|_2^2 \tag{2.4}$$

$$= \arg\min_{s_g} \left\| W(t - As_g) \right\|_2^2 + \lambda_1 \left\| s_g - s_s \right\|_2^2, \tag{2.5}$$

where $\lambda_1 = (\sigma_\varepsilon/\sigma_g)^2$ is a regularization parameter, and W is a direction normalization vector. The local problem is written from above, with each patch solved from the global estimate $\widehat{s_g}$ from (12) (decoupling the local and global problems), giving

$$\{\widehat{x_i}, \widehat{D}\} = \arg\min_{x_i, D} \left\| Dx_i - R_i \widehat{s_g} \right\|, \text{subject to } \|x_i\|_0 = T. \tag{2.6}$$

With the estimate of coefficients $\widehat{X} = [\widehat{x_1}, \ldots, \widehat{x_I}]$ and the global slowness $\widehat{s_g}$, we can solve for s_s

$$\widehat{s_s} = \arg\min_{s_s} \frac{1}{\sigma_g^2} \left\| \widehat{s_g} - s_s \right\|_2^2 + \frac{1}{\sigma_p^2} \sum_i \|D\widehat{x_i} - R_i s_s\|_2^2 \tag{2.7}$$

$$= \arg\min_{s_s} \lambda_2 \left\| \widehat{s_g} - s_s \right\|_2^2 + \sum_i \|D\widehat{x_i} - R_i s_s\|_2^2, \tag{2.8}$$

where $\lambda_2 = (\sigma_p/\sigma_g)^2$ is a regularization parameter.

Considering that λ_1 are λ_2 represent the ratios between σ_ε, σ_p, and σ_g, Fig. 2.5 shows the change of travel-time mean-squared errors with hyperparameters. The regularization parameter values for LST and conventional tomography were selected to minimize the normalized travel time residual (NTTR):

$$loss_{\text{NTTR}} = \|W(t - \hat{t})\|_2^2 = \sqrt{\sum_i w_i (t_i - \hat{t}_i)^2}, \tag{2.9}$$

where t is the observed real travel time vector, and $\hat{t} = As_s$ is the calculated travel time vector from our tomography results. w_i, t_i and $\widehat{t_i}$ are the i-th elements in W, t and $\hat{t}$.

2.6.3 Least-Squared Model

The LSQ method provides the maximum a posteriori (MAP) estimate of the slowness vector $\hat{s}$ by

$$\hat{\mathbf{s}} = (\mathbf{A}^T\mathbf{A} + \eta\boldsymbol{\Sigma}_L^{-1})^{-1}\mathbf{A}^T\mathbf{t}, \tag{2.10}$$

where η is the regularization parameter, and $\boldsymbol{\Sigma}_L(i,j) = exp(-\mathbf{D}_{i,j}/L)$ is a smoothing kernel. $\mathbf{D}_{i,j}$ is the distance between the i-th and j-th elements in the slowness image $\mathbf{s}$, and L is the parameter used to control the length of the smoothness scale. The smoothness kernel size L is set to 10 based on cross-validation and consistent with the LST patch size. It was found that values of L in the neighborhood of $L = 10$ pixels (roughly corresponding to $\sim$4 km) minimized the residual on validation data.

2.6.4 Parameter Selection

We carried out an additional test, where our LST model was iteratively trained on 9-fold data and tested on the remaining fold, and the NTTR loss was averaged over 10 iterations. We tested a group of hyperparameter combinations and summarize the corresponding NTTR loss in Fig. 2.5. Considering our tomography matrix has a size of 96×112, we choose $l = 10$ pixels as our patch size. $\lambda_1 = 1.58$ and $\lambda_2 = 0.40$ were selected as the optimal combination for the LST models, and $\eta = 0.40$ is selectd as the optimal value for the LSQ model from the parameter validation tests (shown in Fig. 2.5).

2.7 Acknowledgments

The text of Chapter 2, in full, is a reprint of the material as it appears in Zhou, Z., Bianco, M., Gerstoft, P. and Olsen, K.B. “High-resolution imaging of complex shallow fault zones along

the July 2019 Ridgecrest ruptures", *Geophysical Research Letters*, 49(1), p.e2021GL095024, 2022. The dissertation author was the primary researcher and author of Chapter 2. The coauthors listed in this publication directed and supervised the research.

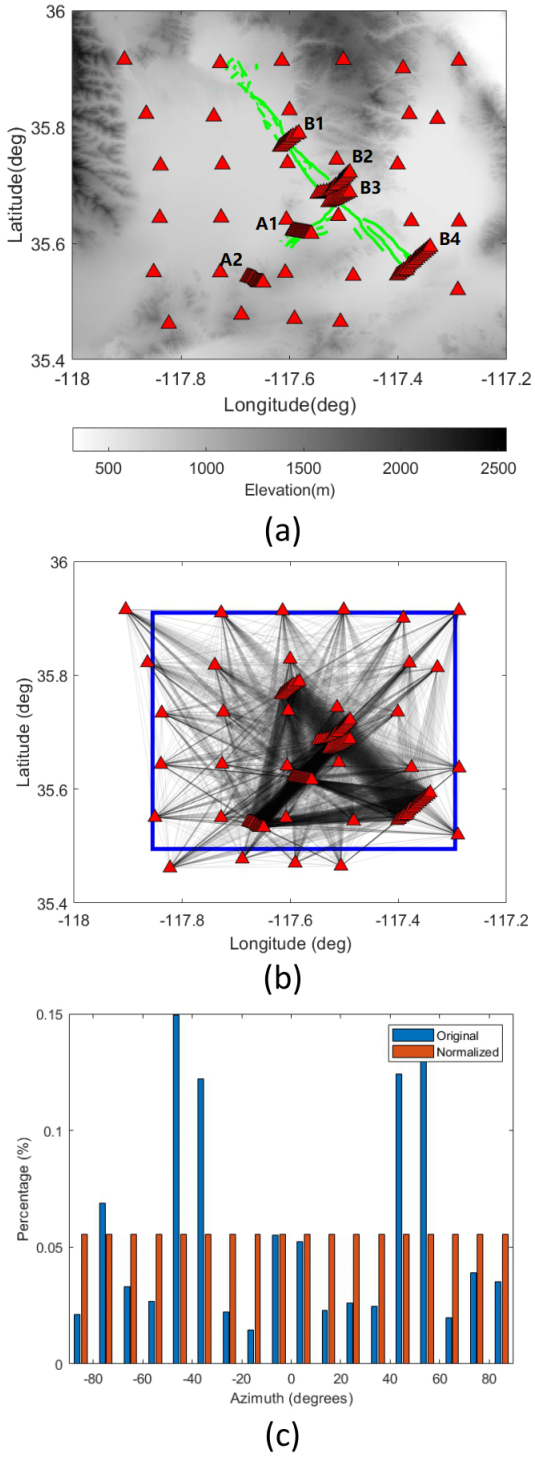

Figure 2.1. (a) Station locations (red triangles) and fault traces (green lines), (b) ray coverage in the model region (blue box), and (c) histogram of ray azimuths for the original and normalized distributions.

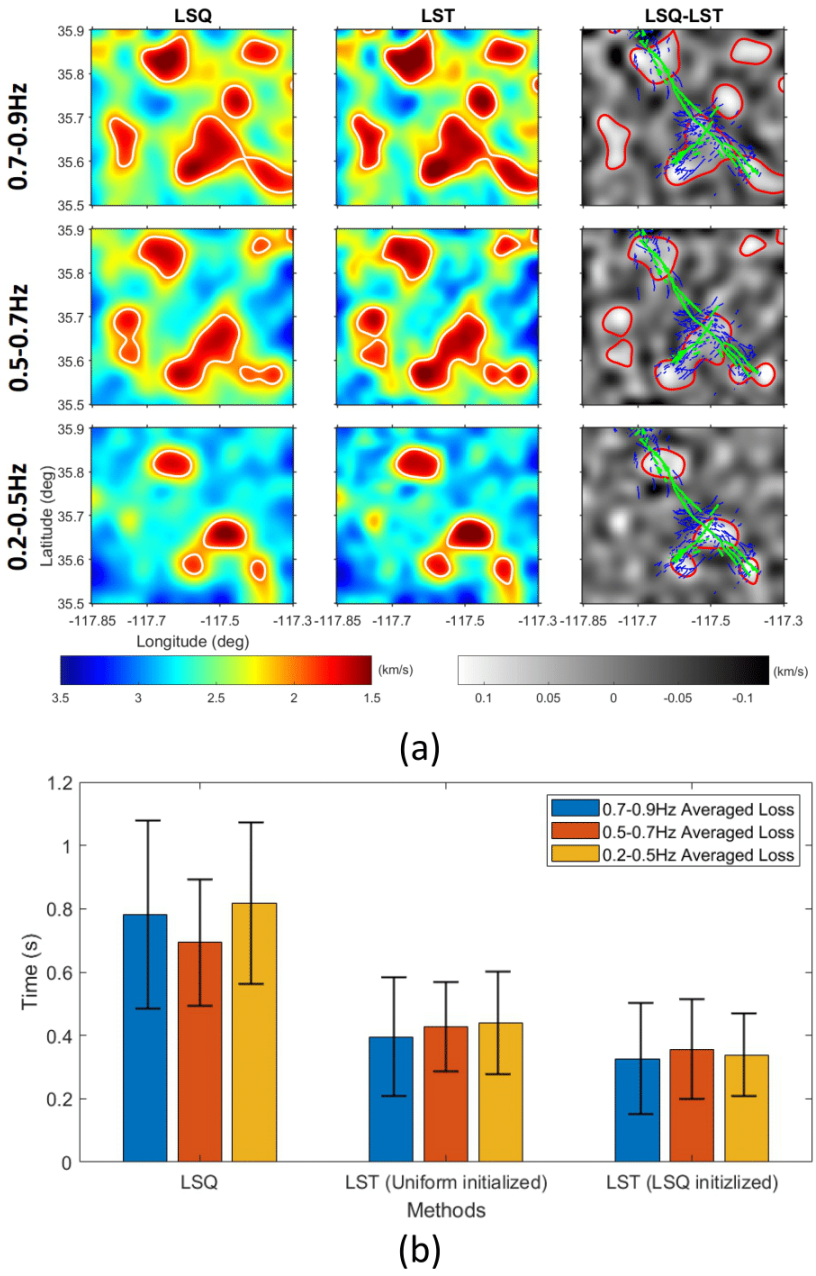

Figure 2.2. (a) 10 fold-averaged Rayleigh wave group velocity maps using least square (LSQ) and LST methods for 0.7–0.9, 0.5–0.7, and 0.2–0.5 Hz with the 2 km/s contour lines (white) superimposed. The LST is initialized by the least-square results. The difference maps (LSQ-LST) are superimposed with fault traces from field mapping (green lines), phase gradients of the Sentinel1 radar satellite (blue traces), and the 2 km/s Rayleigh wave velocity contour lines from the LST maps. (b) 10 fold-averaged mean (bins) and standard deviation (errorbars) of the normalized travel time residual (NTTR) losses derived by the LSQ and LST, which were initialized by either uniform or LSQ-estimated velocities.

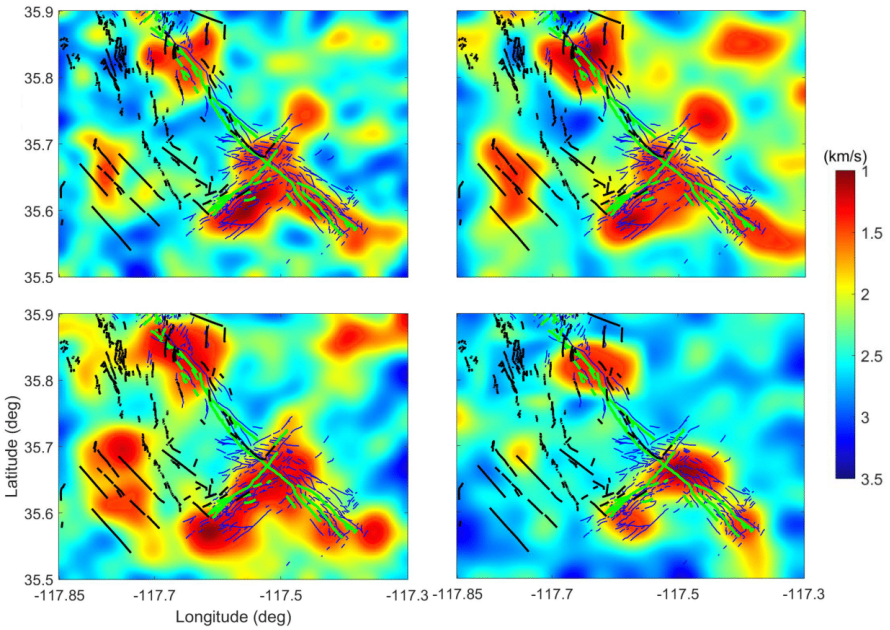

Figure 2.3. Rayleigh wave group velocity maps obtained using LST for bandwidths (a) 0.02–1, (b) 0.7–0.9, (c) 0.5–0.7, and (d) 0.2–0.5 Hz with superimposed fault traces that ruptured in the 2019 sequence (green lines), other Quaternary faults (black lines) and Sentinel1 phase gradients (blue traces). The velocity models are averaged over cross-validation folds.

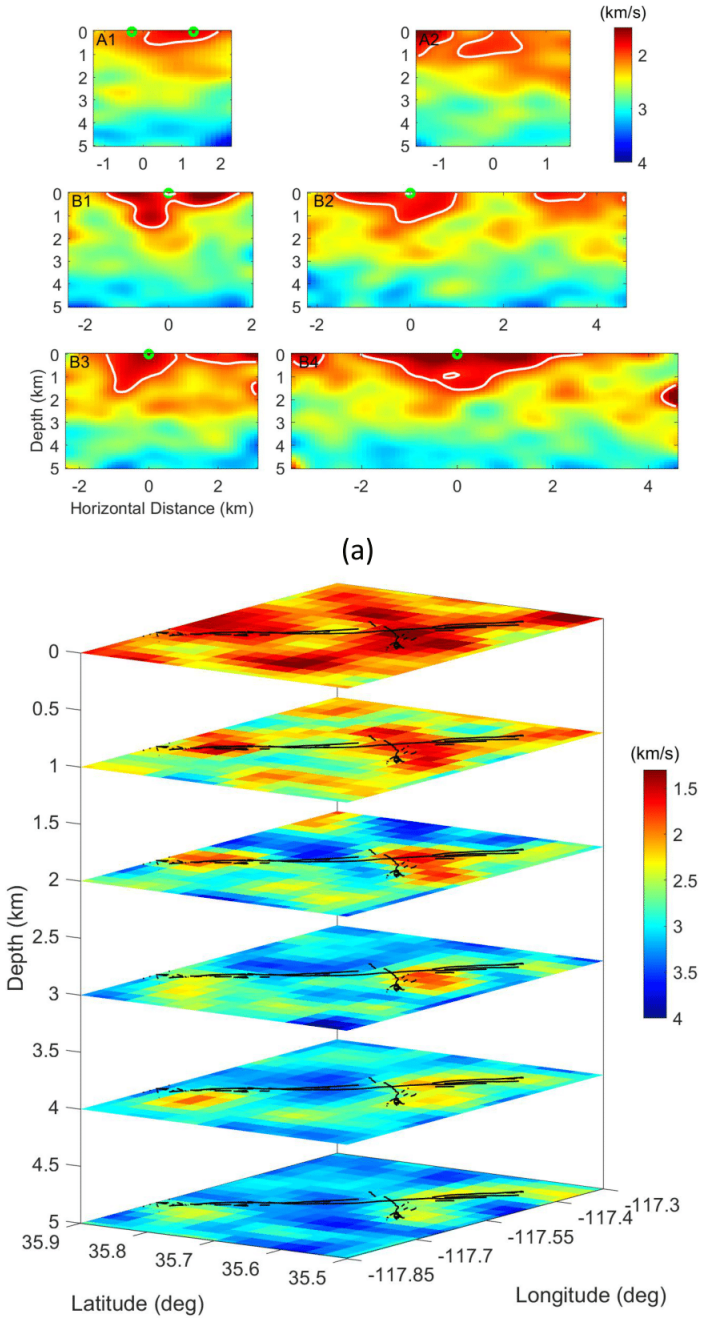

Figure 2.4. (a) Vertical cross-section of the shear wave velocities from the A1, A2, B1, B2, B3, and B4 station arrays. The intersection with the surface rupture of the M6.4 and M7.1 Ridgecrest events (green circles) and 2 km/s contours (white lines) are indicated. (b) Composite 3D image of shear wave velocities obtained from inversion of Rayleigh waves dispersion curves, delineating flower-shaped LVZs.

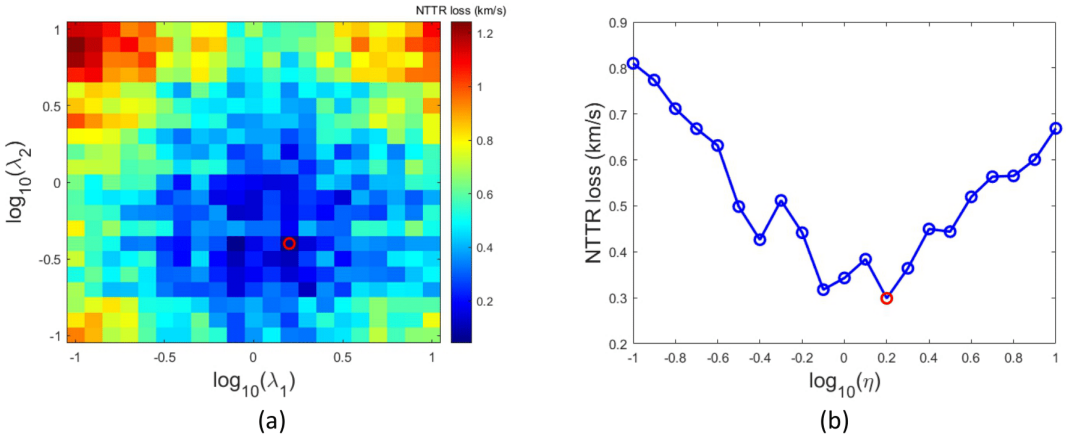

Figure 2.5. NTTR loss as a function of (a) hyperparameters λ_1 and λ_2 and (b) η. The red circle highlights the hyperparameters with the smallest NTTR loss.

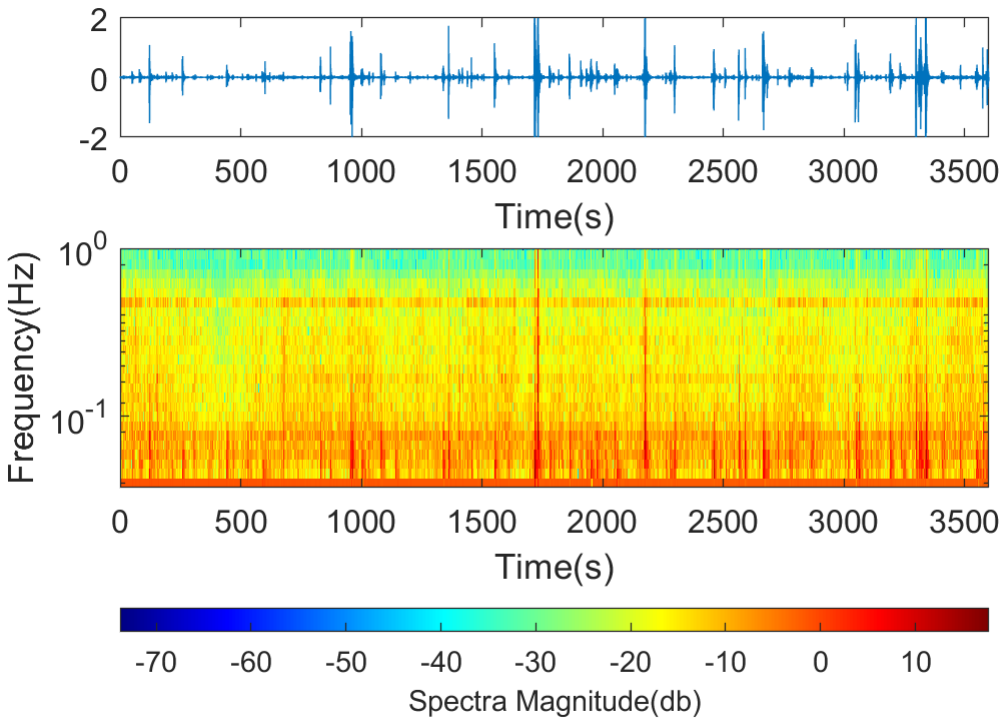

Figure 2.6. (a) An example of 1 hr of vertical component data recorded by one of the Nodes at the Ridgecrest area and (b) the corresponding spectrogram.

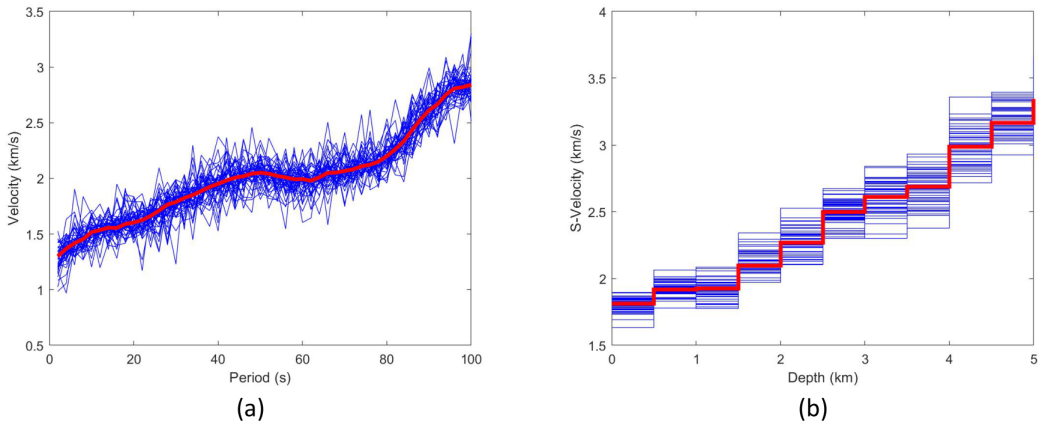

Figure 2.7. (a) Rayleigh wave dispersion curves and (b) inverted S-wave profiles for 50 station pairs picked from the B1 Nodal array (see Fig. 1 in our main manuscript). The red line is the median curve.

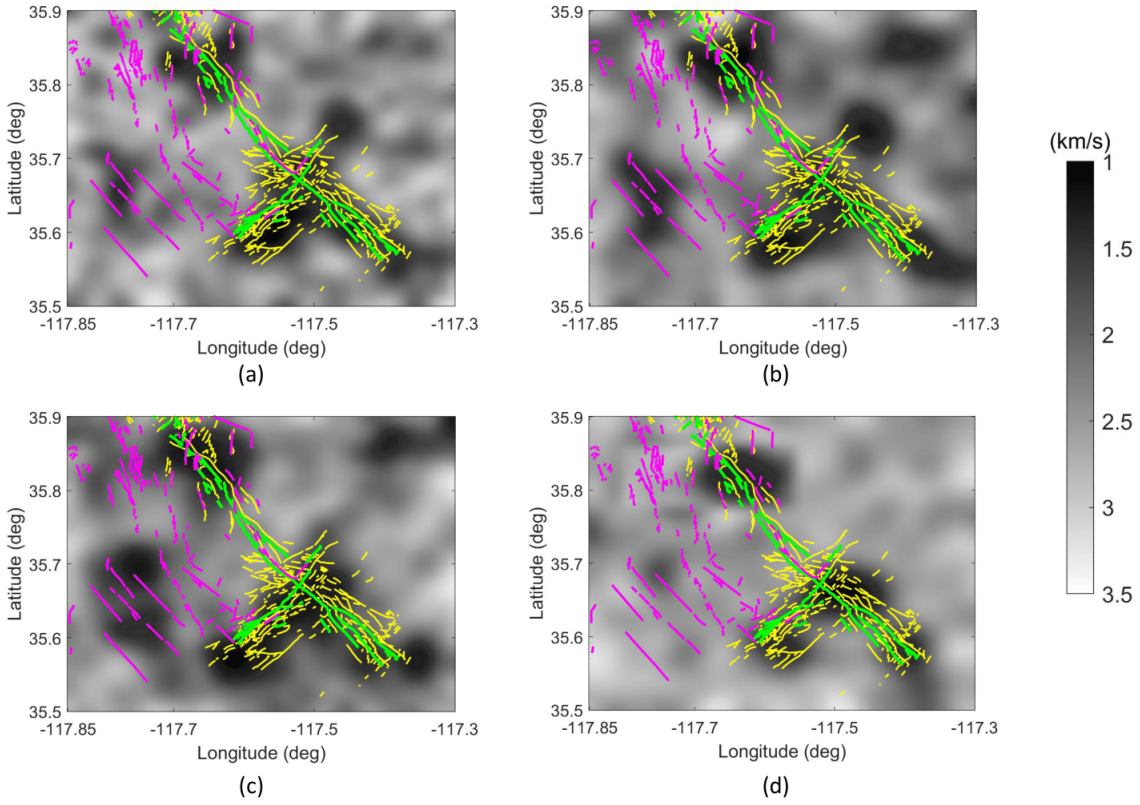

Figure 2.8. Rayleigh wave group velocity maps obtained using LST for bandwidths (a) 0.02–1, (b) 0.7–0.9, (c) 0.5–0.7, and (d) 0.2–0.5 Hz with superimposed primary fault traces (green lines), Quaternary faults (purple lines) and phase gradients (yellow lines). The velocity models are averaged over cross-validation folds.

Chapter 3

2D Multi-resolution Tomography Model Fusion with Physics-informed Probability Graphical Models

3.1 Introduction

The resolution of velocity models varies due to inversion approaches, ray coverage, etc. For example, large-scale Community Velocity Models (CVMs) are typically characterized by relatively low resolution, while high-resolution velocity models are limited to smaller areas with dense station coverage. Combining such velocity models with different resolutions is useful to improve CVMs, e.g., for ground motion estimation or dynamic rupture modeling, where a range of scales is needed [e.g., 40]. However, when models with different resolutions are directly superimposed, sharp boundaries caused by misaligned patterns may produce artifacts in the modeled seismic waves. The problem of fusing low- and high-resolution models is conventionally addressed by extracting patterns from various models and utilizing them as boundary conditions for wave propagation. Border merging has been performed by Gaussian smoothing [5] and by defining a weighting region [41]. However, this process depends on experts manually choosing models, and fine-tuning parameters is often time-consuming.

The exploration of multiscale problems has surged across diverse geophysical fields, encompassing full wave inversion, ground motion modeling, and surface-wave tomography.

A pivotal discovery is the scale-dependent nature of anisotropy, a finding that has substantial implications for how we perceive the structure of the Earth [42]. Analyses from both physical and signal processing viewpoints bring to light key challenges associated with the effect of spatial heterogeneities on seismic wave propagation. These studies have shown that seismic waves cannot distinguish between large-scale anisotropy and small-scale isotropic heterogeneities much smaller than a wavelength [43], enhanced by the potent spatial and directional dependence of tomographic resolution. Such dependency may provoke space- and direction-oriented smoothing [44], eliciting apparent anisotropy fluctuations that bear no intrinsic connection to Earth's anisotropy.

The construction of multiscale models has traditionally employed tactics such as multiscale element fitting [45], Gaussian kernel smoothing [46], and adaptive filtering methods, such as sparse learning [47, 48, 32]. Zhang et al.[49] proposed a data-driven DL method for transformations between seismic velocity models of different resolutions, which involves a linear decomposition of an input signal using a small set of basis signals, or atoms, learned from HR and LR pairwise data. Although these methods have proved efficient, they might not fully grasp the vast complexity of the earth's structure.

Probabilistic Graphic Methods (PGMs) can process images with complex structures, owing to their power to extract the underlying relations among images [50, 51]. In modeling pixel points and their interactions using a graph, a structured environment emerges, shaped by the spatial characteristics of the pixels' geometric connections in various applications, including denoising [52], segmentation [53], and seismic detection problems [54]. Beyond conventional graphs, PGMs have been extended to high-dimensional spaces, such as multilayer graphs [55] and hypergraphs [56]. Furthermore, graph neural networks (GNNs) and graph convolutional networks (GCNs) are important tools in image processing and computer vision [57, 58]. Bayesian methods provide a framework for modeling uncertainty, learning from data, and making predictions. They have found broad applications in a variety of seismic fields, such as seismic tomography [59], full-wave inversion [60], and ground motion prediction [61]. Among all the PGMs and

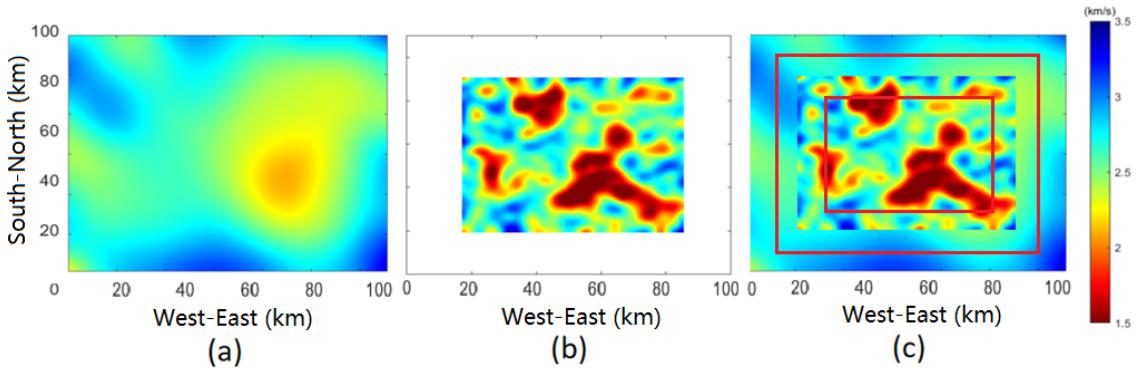

Figure 3.1. (a) Excerpt of S-wave velocities from the Southern California Earthquake Center (SCEC) Community Velocity Model (CVM) version S-4.26 (hereafter referred to as the low-resolution (LR) CVM) at 0.5 km depth around the Ridgecrest area. (b) High-resolution (HR) S-wave map from 1 Hz Rayleigh wave tomography from Zhou et al. [64]. (c) A direct superposition of the HR and LR models. These two models share some patterns in the low-velocity zones, but there are many mismatched detailed patterns where the two models overlap, which results in sharp and misaligned boundaries in those areas. PIPGM is applied to the mismatched boundary areas between the two red bounding boxes; note that the pixels in this area belong to the effective vertices set $\mathscr{V}$.

Bayesian methods, Markov Random Fields (MRFs) constitute a popular and effective approach for supervised structure learning tasks involving the mapping between complex geometric structures [62]. MRFs provide an image restoration procedure, first suggested by Geman et al. [8], which is based on Bayesian inference for a spatially stochastic model. In contrast to convolution-based methods, the MRF procedure yields an optimal and mathematically tractable result for image processing [9, 63].

Recently, many novel methodologies, within the category of physics-informed machine learning [65] have gained much attention. These methods combine conventional machine learning (primarily neural networks [66]) and weighting mechanisms. They adapt the behavior of the learning models to the non-linear features of the solution and introduce physical information as a guide, eventually improving the current limitations of learning models.

Combining the physics-informed mechanism and the MRF model, we propose a physics-informed probabilistic graphical model (PIPGM) that captures the relations between subdomains with multiple resolutions. The proposed method is inspired by the progress achieved in image

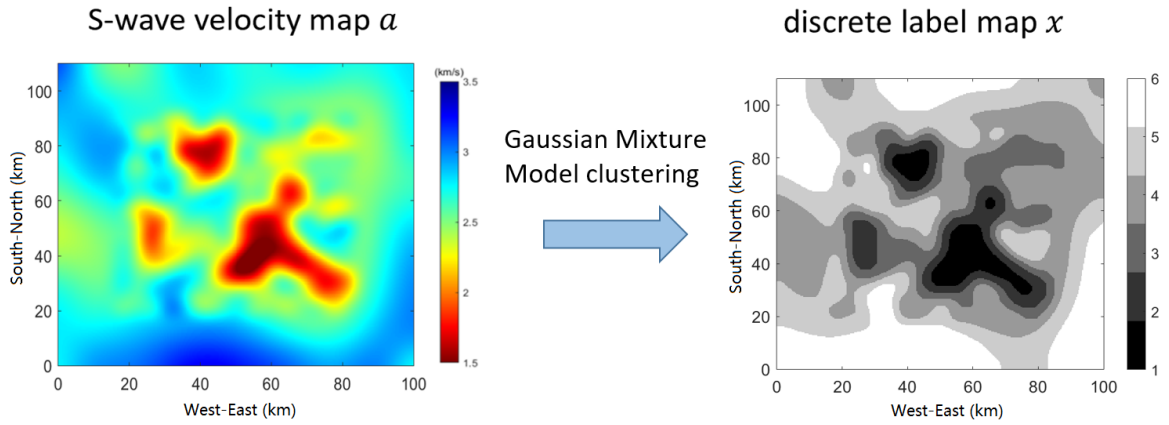

Figure 3.2. A 6-cluster Gaussian Mixture model clustering is applied on the continuous velocity map $\boldsymbol{A}$ (left), and this derives a 6-cluster discrete label map $\boldsymbol{X}$ (right). The pixels with similar velocity information have been assigned the same label.

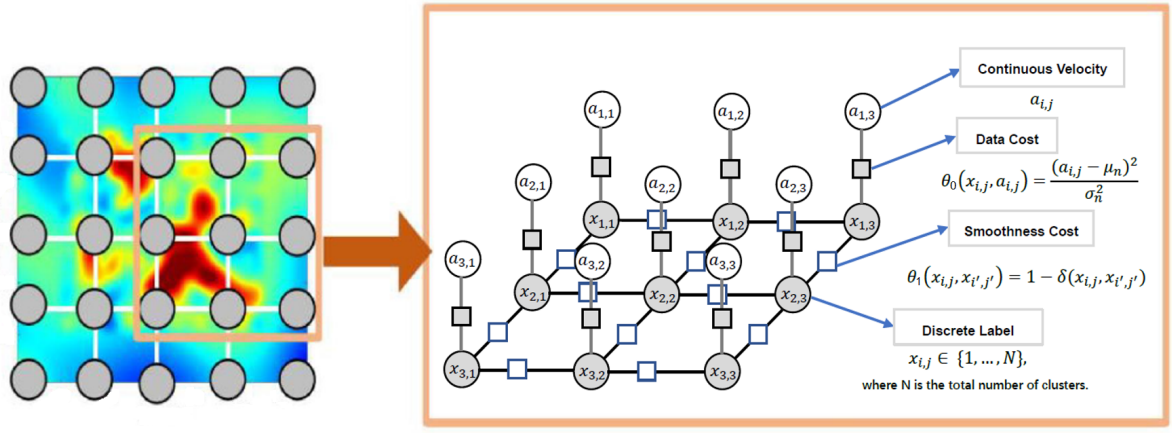

Figure 3.3. Each pixel has a continuous velocity value $a_{i,j}$ and a discrete label mask $x_{i,j}$. The object function designed for MAP estimation has two parts: (1) the data cost θ_0 (0th-order neighboring potential) that forces the pixels with the same label to follow the same Gaussian distribution and (2) the smoothness cost θ_1 (1st-order neighboring potential) that promotes smoothness among neighboring pixels [7].

super-resolution [67, 68] and image editing [44, 69], and is designed to fuse seismic velocity models with probabilistic graphical models (PGMs). We focus on models with well-defined, separate high-resolution (HR) and low-resolution (LR) areas, to enhance local HR structure and simultaneously preserve global smoothness in the fused model. By transferring the information from the HR subdomain, the details in the LR areas are enhanced by solving a maximum likelihood problem with prior knowledge from the HR areas. Evaluation tests on both a checkerboard and a fault zone model derived from the 2019 Ridgecrest, CA, earthquake are performed to demonstrate its efficacy. Quantitatively, model efficacy is evaluated by the deviation between observed and calculated travel times in fused models, relative to those from directly superimposed LR and HR models. Tests demonstrate that PIPGM is generally superior to widely used conventional methods (see Experiments section).

The contributions of this article are two-fold: (1) we introduce a PIPGM fusion method for combining velocity maps with various resolutions, and (2) we introduce physical information as a guide for the graph learning process.

3.2 Theory

3.2.1 Fundamental Model Setup

The objective is to estimate the true velocity field $\boldsymbol{A}$ from $\boldsymbol{A}_{\mathrm{LR}}$ (low-resolution) and $\boldsymbol{A}_{\mathrm{HR}}$ (high-resolution) models, focusing on optimal merging of their borders (see Fig. 3.1). A discrete class label map links the spatial velocity field together. It is initialized from a continuous velocity map $\boldsymbol{A}$ and a 6-cluster discrete label map $\boldsymbol{X}$. Each pixel, described by an (i, j) coordinate, contains a label $X_{i,j}$ and a velocity $A_{i,j}$, with velocities of the same label following the Gaussian distribution $\mathcal{N}(\mu_n, \sigma_n^2)$. Thus, in a graph, velocities $\boldsymbol{A}$ are on top of labels $\boldsymbol{X}$ (Fig. 3.3). d represents all possible labels of $X_{i,j}$ (namely, $d = \{1,2,3,4,5,6\}$ here), and $\boldsymbol{D}$ represents all possible combinations of labels $\boldsymbol{X}$ for the entire map. The velocity map $\boldsymbol{A}$ is connected by class labels $\boldsymbol{X}$.

3.2.2 Gaussian Mixture Model (GMM)

We use Gaussian mixture model (GMM) clustering to assign each continuous velocity pixel $A_{i,j}$ a discrete label $X_{i,j}$. We utilize these discrete labels to ensure the preservation of spatial coherence and the integrity of class boundaries, both essential for accurate data interpretation and model combining. GMM clustering is a widely used probabilistic clustering technique. In this model, data points are assumed to be generated from a mixture of a finite number of Gaussian distributions with unknown parameters [70]. Clustering aims at identifying these parameters, thus segmenting the data into clusters, each represented by a Gaussian distribution. The GMM clustering can be implemented using the Expectation-Maximization (EM) method. Here, the continuous velocity pixel is denoted by $A_{i,j}$ and the corresponding labels by $X_{i,j}$. The algorithm is described as follows:

1. **Initialization:** Define $N = 6$ clusters (the selection of cluster number will be discussed in Section 6.2) and initialize their means μ_n, variance σ_n, and mixing coefficients π_n are initialized either randomly or based on prior knowledge.

2. **E-Step:** Compute each data point's responsibility $p(x_i)$, the probability that belongs to the cluster n, using:
$$p(x_i) = \frac{\pi_n \mathcal{N}(A_{i,j}|\mu_n, \sigma_n)}{\sum_{n=1}^{N} \pi_n \mathcal{N}(A_{i,j}|\mu_n, \sigma_n)},$$

3. **M-Step:** Update parameters μ_n, σ_n, and π_n to maximize observed data's log-likelihood:
$$\mu_n^{new} = \frac{1}{N_n} \sum_{i=1}^{N_{\text{total}}} P(X_{i,j}) A_{i,j}, \tag{3.1}$$
$$\sigma_n^{new} = \frac{1}{N_n} \sum_{i=1}^{N_{\text{total}}} P(X_{i,j})(A_{i,j} - \mu_n)(A_{i,j} - \mu_n)^T, \tag{3.2}$$
$$\pi_n^{new} = \frac{N_n}{N_{\text{total}}}, \tag{3.3}$$

4. **Convergence:** Stop if parameters or log-likelihood change is below a threshold or a

maximum iteration count is reached. Assign $A_{i,j}$ to the cluster maximizing $P(X_{i,j})$.

3.2.3 Markov Random Fields

Bayesian estimation framework

Given the prior probabilities $P(\boldsymbol{X})$ of label $\boldsymbol{X}$ and the likelihood densities $P(\boldsymbol{A} \mid \boldsymbol{X})$ of the observed velocity $\boldsymbol{A}$, the posterior probability is computed using the Bayes rule:

$$P(\boldsymbol{X} \mid \boldsymbol{A}) = \frac{P(\boldsymbol{A} \mid \boldsymbol{X})P(\boldsymbol{X})}{P(\boldsymbol{A})} \propto P(\boldsymbol{A} \mid \boldsymbol{X})P(\boldsymbol{X}). \tag{3.4}$$

Here, $P(\boldsymbol{A})$, the probability density function (PDF) of $\boldsymbol{A}$, is a fixed probability distribution (for given $\boldsymbol{A}$) and does not affect the maximum a posteriori (MAP) estimation solution. The Bayesian labeling problem requires finding the MAP configuration. The MAP of labeling for observation $\boldsymbol{A}$ is given by:

$$\boldsymbol{X}^* = \arg\max_{\boldsymbol{X} \in D} P(\boldsymbol{X} \mid \boldsymbol{A}), \tag{3.5}$$

where D denotes a set of possible candidates of the discrete labels $\boldsymbol{X}$, and $\boldsymbol{A}$ represents the observation of the continuous velocities.

We need the prior probabilities and likelihood functions to derive the MAP solution. The likelihood function $P(\boldsymbol{A} \mid \boldsymbol{X})$ depends on the noise statistics and the underlying transformation from truth to observation.

Neighborhood System in MRF

The MRF, a model that analyzes spatial relations, uses a neighborhood system [11], defined as $\mathscr{N} = \{\mathscr{N}_{i,j} \mid \forall (i,j) \in \mathscr{V}\}$ (shown in Fig. 3.4), where $\mathscr{V}$ contains all pixel indices and $\mathscr{N}_{i,j}$ includes neighboring pixels. The neighboring system of the given pixel with index (i,j) can be decomposed as the union of 0th-, 1st-order, $\cdots$ neighboring systems as

$$\mathscr{N}_{i,j} = \mathscr{N}_{i,j}^0 \cup \mathscr{N}_{i,j}^1 \cup \cdots, \tag{3.6}$$

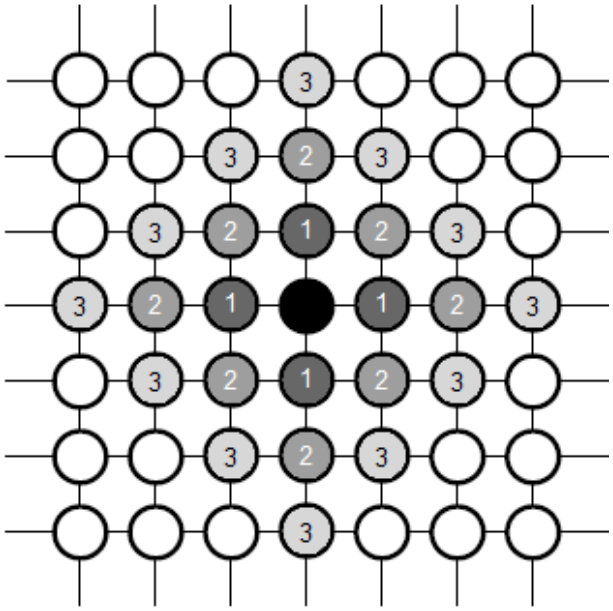

Figure 3.4. The neighborhood system $\mathcal{N}_{i,j}$ (marked in gray/black) of the given center node (i, j) (marked in black). The 1st-, 2nd-, and 3rd-order neighborhood systems of node (i, j) are marked with numbers '1', '2', '3', and they can be represented as $\mathcal{N}_{i,j}^{1}$, $\mathcal{N}_{i,j}^{2}$, and $\mathcal{N}_{i,j}^{3}$.

and we define the n-th-order neighboring system of (i, j) as

$$\mathcal{N}_{i,j}^{n} = \{(i', j') \mid |i - i'| + |j - j'| = n\}, \tag{3.7}$$

which means (i', j') and (i, j) have a Manhattan distance [71] of n. The pair $(\mathcal{V}, \mathcal{N})$ constitutes a graph, and the neighboring system N for the pixel (i, j) consists of the center node (0th-order neighboring system) $\mathcal{N}_{i,j}^{0} = \{(i, j)\}$ and a pair of neighboring pixels $\mathcal{N}_{i,j}^{1} = \{(i', j')\}$. To avoid artifacts at the boundary, we expand the image's boundaries by one pixel in all directions, where the values of the outermost pixels are directly replicated.

The MRF is on D with respect to $\mathcal{N}$ if (1) $P(X_{i,j}) > 0, \forall X_{i,j} \in d, \forall (i, j) \in \mathcal{V}$ (positivity) and (2) $P(X_{i,j} \mid X_{i',j'}, (i', j') \in \mathcal{V}) = P(X_{i,j} \mid X_{i',j'}, (i', j') \in \mathcal{N}_{i,j})$. Satisfying condition (1) above guarantees the model to be a random field. Condition (2) is the Markov local property. Considering that the label variable $\boldsymbol{X}$ depends on velocity $\boldsymbol{A}$ and is unobservable, it is assumed that the distribution of $P(\boldsymbol{X}|\boldsymbol{A})$ follows exponential distributions, using Bayes' rule and the conjugate distribution property [72].

MRF Prior and Posterior Energy

A model can be considered a valid MRF if and only if the probability distribution $P(\boldsymbol{X})$ of the configurations is an exponential distribution with normalization, defined as the following form

$$P(\boldsymbol{X}) = \frac{1}{Z_1} e^{-U_{\text{prior}}(\boldsymbol{X})}, \tag{3.8}$$

where Z_1 is a normalizing constant, and $U_{\text{prior}}(\boldsymbol{X})$ is the prior energy (Section 4.2 in [7]).

The prior energy $U_{\text{prior}}(\boldsymbol{X})$ can be expressed as the summation of neighboring potentials

$$U_{\text{prior}}(\boldsymbol{X}) = \sum_{n \in \mathcal{N}} \theta_n(\boldsymbol{X}) = \sum_{\{(i,j)\} \in \mathcal{N}_{i,j}^{0}} \theta_0\left(X_{i,j}\right) + \sum_{\{(i',j')\} \in \mathcal{N}_{i,j}^{1}} \theta_1\left(X_{i,j}, X_{i',j'}\right) + \cdots, \tag{3.9}$$

where $\mathcal{N}$ is the set of all the possible neighboring systems, and $\mathcal{N}_{i,j}^{0}$ and $\mathcal{N}_{i,j}^{1}$ denotes the 0th-

and 1st-order neighboring systems, with θ_0 and θ_1 representing the respective potentials. The 0th-order neighboring system consists of the ensemble of all feasible indices (i, j), identified as the central pixel. This study focuses exclusively on the neighboring potentials of 0th and 1st order while truncating higher order potentials as delineated in Equation (3.9).

Because we assume that the velocities $\boldsymbol{A}$ with given labels $\boldsymbol{X}$ follow Gaussian distributions, the likelihood function can be expressed in an exponential format

$$P(\boldsymbol{A} \mid \boldsymbol{X}) = \frac{1}{Z_2} e^{-U_{\text{like}}(\boldsymbol{A}|\boldsymbol{X})}, \tag{3.10}$$

where $U_{\text{like}}(\boldsymbol{A} \mid \boldsymbol{X})$ is called the likelihood energy. According to the Bayes rule Eq. (3.4), the posterior probability must be an exponential distribution

$$P(\boldsymbol{X} \mid \boldsymbol{A}) = \frac{1}{Z_3} e^{-U_{\text{post}}(\boldsymbol{X}|\boldsymbol{A})}, \tag{3.11}$$

where Z_2 and Z_3 are normalization constants. Taking the negative logarithm in Eqs. (3.10)–(3.11) gives the posterior energy

$$U_{\text{post}}(\boldsymbol{X} \mid \boldsymbol{A}) = U_{\text{prior}}(\boldsymbol{X}) + U_{\text{like}}(\boldsymbol{A} \mid \boldsymbol{X}) + C, \tag{3.12}$$

where C is a constant related to the normalization constants Z_1, Z_2, and Z_3. Hence, given a fixed $\boldsymbol{A}, X$ is also an MRF on d with respect to $\mathcal{N}$. The MAP solution is equivalently found by

$$\boldsymbol{X}^* = \arg\min_{\boldsymbol{X} \in D} U_{\text{post}}(\boldsymbol{X} \mid \boldsymbol{A}), \tag{3.13}$$

which minimizes the negative log-likelihood problem of Eq. (3.11).

In summary, the MRF modeling process consists of the following steps: Defining a neighborhood system $\mathcal{N}$, defining the prior potentials θ_0, deriving the likelihood energy $U_{\text{like}}(\boldsymbol{A}|\boldsymbol{X})$, and deriving the posterior energy $E(\boldsymbol{X})$ (which can be expressed as the summation

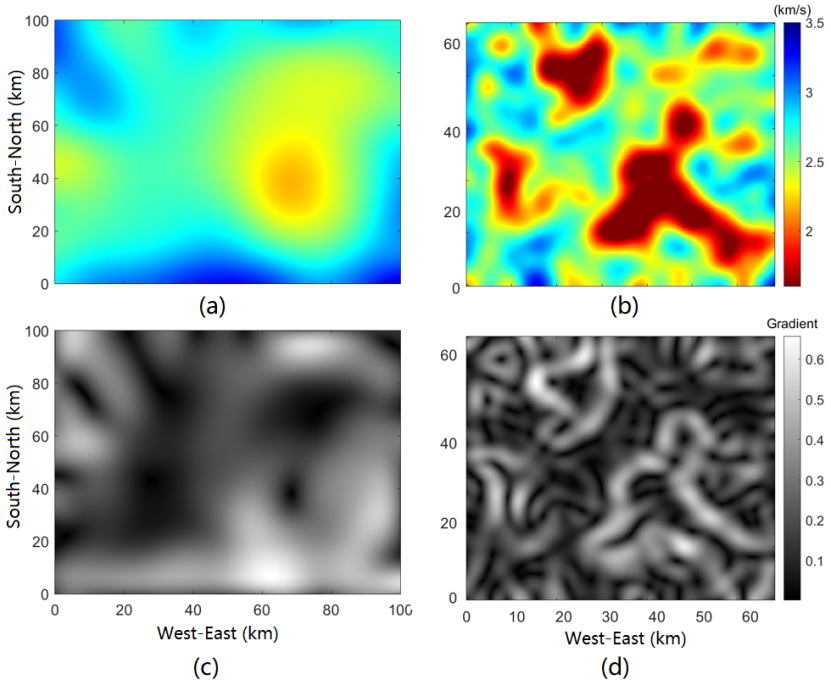

Figure 3.5. (a) LR CVM around the Ridgecrest area. (b) HR S-wave map from 1 Hz Rayleigh wave tomography. (c) and (d) corresponding gradient maps of (a) and (b). In the gradient maps, brighter (darker) areas indicate that velocities change sharply (slightly).

of neighboring potential functions). Eqs. (3.10)–(3.13) show that we can express the posterior probability $P(\boldsymbol{X}|\boldsymbol{A})$ into the prior energy $U_{\text{prior}}(\boldsymbol{X})$ (which can be measured by multiple potentials) and the likelihood function energy $U_{\text{like}}(\boldsymbol{A}|\boldsymbol{X})$, and this gives a good reason for using MRF priors, which means that we can measure conditional probabilities $P(\boldsymbol{X}|\boldsymbol{A})$ without knowing its specific expression.

3.3 PIPGM

3.3.1 Physics-Informed Mechanism

PIPGM follows a first-order MRF setting (Fig. 3.4) where each random variable has four neighbors on which it is conditionally dependent. The full conditional probability of the discrete random variable $X_{i,j} \in \{1,\ldots,6\}$ is the exponential of the sum of potentials (four 1st-order neighboring potentials θ_1 between cluster labels and one 0th-order center data potential θ_0

between cluster label and velocity) in conventional MRF settings. In image process problems, optimizing the entire map can be broken down into several suboptimization problems that iteratively optimize each pixel [73]. Inserting Eq. (3.9) into Eq. (3.12), we have

$$\begin{aligned} -\log p(X_{i,j} \mid A_{i,j}) &= U_{\text{post}}(X_{i,j} \mid A_{i,j}) \\ &\propto \theta_0(X_{i,j}, A_{i,j}) + \sum_{(i',j') \in \mathcal{N}_{i,j}} \theta_1(X_{i,j}, X_{i',j'}) + C, \end{aligned} \tag{3.14}$$

where C is a constant related to normalization factors Z_1, Z_2, and Z_3 (see Eqs. (3.8), (3.10), (3.11)), and

$$\theta_0\left(X_{i,j}, A_{i,j}\right) = \frac{(A_{i,j} - \mu_n)^2}{\sigma_n^2} \tag{3.15}$$

is the 0th-order neighboring potential [74] (also known as the data cost function) that relates $X_{i,j}$ with the observed velocity data $A_{i,j}$. μ_n and σ_n^2 are the mean and variance of all pixels with the same cluster label $n = X_{i,j}$. It promotes that continuous velocity values $\boldsymbol{A}$ sharing pixels with the same discrete label $\boldsymbol{X}$ follow the same Gaussian distribution.

$$\theta_1\left(X_{i,j}, X_{i',j'}\right) = 1 - \delta(X_{i,j}, X_{i',j'}) \tag{3.16}$$

is the 1st-order neighboring potential [74] (also known as the smoothness cost function) that relates $X_{i,j}$ to the 1st-order neighboring variable $X_{i',j'}$ (see Fig. 3.4). This function encourages the neighboring pixels to share the same discrete label $X_{i,j}$, promoting the model's local smoothness.

The effectiveness of conventional MRF approaches, or those based on potential functions, may be constrained when applied to complex geological structures. Assigning different neighboring pixels with various importance weights based on physical information (also known as the physics-informed mechanism) can effectively remove the anisotropic features of the model gradients. This approach can significantly improve inversion results, particularly in the context of geological structure.

In seismic tomography, the coverage of the ray path varies, with the coverage of the

densest ray path usually leading to more reliable estimations of velocities. Here, we introduce a confidence score $v_{R_{i,j}}$ based on the logarithmic format of the ray density for each pixel:

$$v_{R_{i,j}} = \alpha_R \log(D_{i,j} + 1) + \beta_R, \tag{3.17}$$

where $D_{i,j} \in [0, 315]$ is the number of ray paths through a given pixel. The logarithmic scaling parameters α_R and β_R are empirically chosen as 0.08 and 0.90, respectively, resulting in $v_{R_{i,j}} \in [0.9, 1.1]$.

We combine the gradient information from both the LR and HR maps as the prior estimation of the sharpness of the local patterns (as shown in Fig. 3.5). The gradient is calculated by the Prewitt operator approximating the 1st-order derivatives of 2D images [75]. The operator uses two 3×3 kernels convolved with the original image to calculate approximations of the derivatives - one for horizontal changes and one for vertical. The kernels for the Prewitt operator can be represented as

$$K^x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, \quad K^y = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}. \tag{3.18}$$

The convolution of these kernels ($\boldsymbol{K}^x$ and $\boldsymbol{K}^y$) with the original image yields two gradient images, one for the x-direction ($\boldsymbol{G}^x$), and one for the y-direction ($\boldsymbol{G}^y$).

Once $\boldsymbol{G}^x$ and $\boldsymbol{G}^y$ have been determined, we can find the gradient magnitude $G_{i,j}$ at the pixel with index (i, j) as

$$G_{i,j} = \sqrt{{G^x_{i,j}}^2 + {G^y_{i,j}}^2}. \tag{3.19}$$

The resulting image $\boldsymbol{G}$ is a gradient image showing the intensity of the edge. We applied the Perwitt operator on the LR and HR maps to generate the gradient images $\boldsymbol{G}^{\mathrm{LR}}$ and $\boldsymbol{G}^{\mathrm{HR}}$. We empirically choose the weighting parameter $\lambda = 0.2$ ($0 \leq \lambda \leq 1$) for balancing the LR and HR

gradients:

$$G'_{i,j} = (1-\lambda)G^{\mathrm{LR}}_{i,j} + \lambda G^{\mathrm{HR}}_{i,j}. \tag{3.20}$$

The overall range of the pixel values inside the gradient matrices $\boldsymbol{G}^{\mathrm{LR}}$ and $\boldsymbol{G}^{\mathrm{HR}}$ is $[0.05, 0.68]$. The scaling parameters α_G and β_G are empirically set as 0.36 and 0.85, respectively, and this guarantees $v_{G_{i,j}} \in [0.9, 1.1]$:

$$v_{G_{i,j}} = \alpha_G(1 - G'_{i,j}) + \beta_G. \tag{3.21}$$

Generally, the larger the gradient weight term $v_{G_{i,j}}$, the lower the importance of the local smoothness cost. The gradient weight term $v_{G_{i,j}}$ is joined with the previous ray-density weight $v_{R_{i,j}}$, defining a physics-informed weight $\omega_{i,j}$

$$\omega_{i,j} = v_{R_{i,j}} v_{G_{i,j}}. \tag{3.22}$$

With the parameters chosen above, the range of the physics-informed weight is $[0.81, 1.21]$ with a mean of around 1. The physics-informed weight $\omega_{i,j}$ can adaptively assign a larger weight to the trusted nodes based on existing physical information. We used ray-path density and gradient information, both exhibiting pronounced anisotropic characteristics (as shown in Fig. 3.5 and Fig. 3.8), thereby facilitating the model's capability to explore more complex formations within these regions. Assigning the physics-informed weight in Eq. (3.14), we obtain the posterior probability function in the PIPGM

$$-\log p(X_{i,j} \mid A_{i,j}) \propto \omega_{i,j}\theta_0(X_{i,j}, A_{i,j}) + \sum_{(i',j') \in \mathcal{N}_{i,j}} \omega_{i',j'}\theta_1(X_{i,j}, X_{i',j'}) + C, \tag{3.23}$$

where C is a constant related with to normalization factors Z_1, Z_2, and Z_3 (see Eqs. (3.8), (3.10),

Algorithm 1 MCMC Method for MRF

1. Input: $\boldsymbol{A}_{LR}$ and $\boldsymbol{A}_{HR}$
2. Initialize the velocity model $\boldsymbol{A}$ by superimposing $\boldsymbol{A}_{HR}$ over $\boldsymbol{A}_{LR}$
   Initialize $\boldsymbol{X}$, $\mu_{\mathbf{n}}$ and $\sigma_{\mathbf{n}}$ with Gaussian Mixture Model (GMM) clustering
3. **for** each EM iteration **do**
4. Construct PIPGM
5. **for** $t = 1$ to max iteration T **do**
6. (E-Step) Gibbs Sampling
7. **for** each pixel $(i, j) = (1, 1)$ to the maximum index (I, J) **do**
   (updating all the discrete labels $X_{i,j}$)
8. $X_{i,j}^{(t+1)} \sim P\left(X_{i,j} \mid X_{1,1}^{(t+1)}, \ldots, X_{i,j-1}^{(t+1)}, A_{i,j}^{(t)}, X_{i,j+1}^{(t)}, \ldots, X_{i,j}^{(t)}\right)$
9. **end for**
10. Update $P(X_{i,j} = n) = \frac{N_n}{N_{\text{total}}}$, where N_n is number of pixels with label n and N_{total} is number of pixels.
11. **for** each pixel $(i, j) = (1, 1)$ to the maximum index (I, J) **do**
   (updating all the continuous velocities $A_{i,j}$)
12. $A_{i,j} \sim \sum_{n=1}^{6} P(X_{i,j} = n) N(\mu_n, \sigma_n^2)$
13. **end for**
14. (M-Step) Update Gaussian parameters μ_n and σ_n^2 with the sample mean $\mu_n = \frac{1}{N_n} \sum_{X_{i,j}=n} A_{i,j}$ and sample variances $\sigma_n^2 = \frac{\sum_{X_{i,j}=n} (A_{i,j} - \mu_n)^2}{N_n}$
15. **end for**
16. **end for**
17. **return** $\boldsymbol{X}, \boldsymbol{A}$ (for each pixel)

(3.11)). The objective function of the MAP problem of $X_{i,j}$ becomes

$$
\begin{aligned}
X_{i,j}^{*} &= \arg\max_{\boldsymbol{X}_{i,j}} p(X_{i,j} \mid A_{i,j}) \\
&= \arg\min_{\boldsymbol{X}_{i,j}} \omega_{i,j} \theta_0(X_{i,j}, A_{i,j}) + \sum_{(i',j') \in \mathcal{N}_{i,j}} \omega_{i',j'} \theta_1(X_{i,j}, X_{i',j'}) + C.
\end{aligned}
\tag{3.24}
$$

3.3.2 Markov Chain Monte Carlo (MCMC) and Gibbs Sampling

MCMC is a statistical method used to sample probability distributions [76, 77]. Gibbs sampling is a specific MCMC algorithm that iteratively samples a multivariate probability distribution from the conditional distributions of each variable given the current values of the other variables [78]. Combining MCMC with Gibbs sampling enables estimating complex probability distributions without explicit knowledge of the distribution.

We employ the MCMC method with Gibbs sampling to solve Eq. (3.13). The algorithm

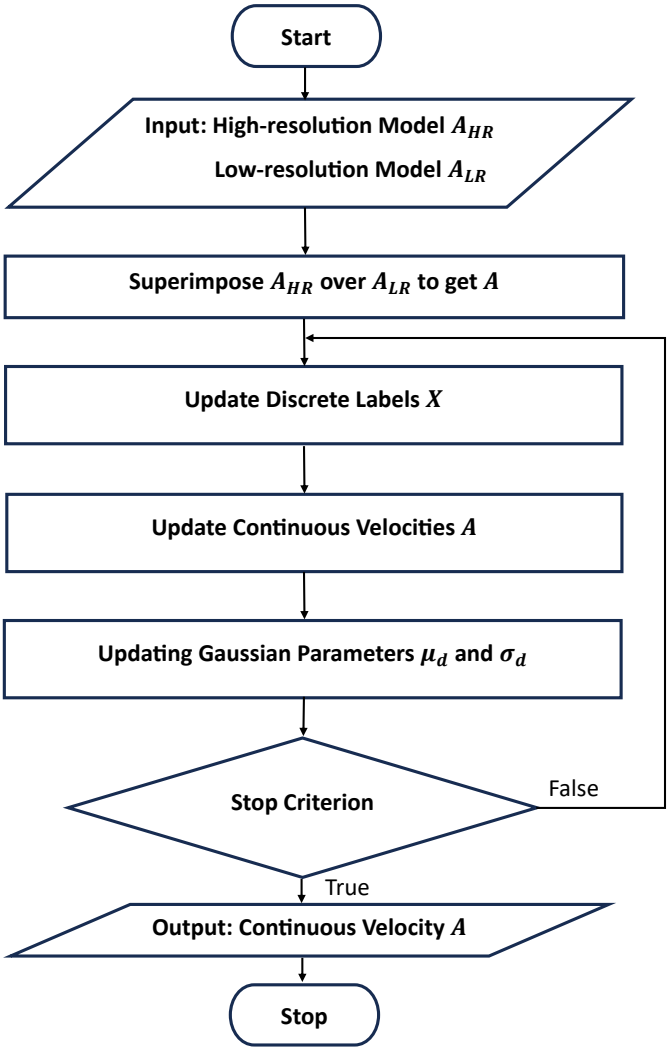

Figure 3.6. The pipeline of the iterative update policy for both pixel labels (discrete class labels) and pixel values (continuous velocity values).

assigns a discrete label to each pixel, creating a map-wide label distribution overview. This enables calculating the discrete label probability $P(X_{i,j} = n)$ by finding the frequency of each label n in Step 10, giving the map's label distribution. In Step 12, the continuous variable $A_{i,j}$ is sampled from a Gaussian mixture model (GMM), allowing velocities beyond a single label's distribution and considering neighboring clusters. In the MRF structure, the update is achieved by calculating the probability for each of the possible labels $n \in \{1, \cdots, 6\}$ at (i, j) using Eq. (3.23) and randomly selecting from this distribution, see Fig. 3.6.

The velocity map $\boldsymbol{A}$ is initialized with the superimposed HR and LR velocity maps, see Fig. 3.8 (a1, b1), and the label map $\boldsymbol{X}$ is initialized with a Gaussian mixture model clustering with the total cluster number $N = 6$ (will be discussed in Section 3.4.1), similar to Fig. 3.8 (a3, b3). All velocities with the label n follow the same Gaussian distribution $N(\mu_n, \sigma_n^2)$. The expectation–maximization (EM) algorithm [79], an iterative method to find MAP estimates of parameters, updates Gaussian parameters μ_n and σ_n^2. The termination criterion is either reaching 10,000 iterations or observing that the cumulative absolute difference across all pixels between consecutive iterations falls below 0.1 km/s, whichever is achieved first. The algorithm has been summarized in Algorithm 1.

3.4 Experiments

To evaluate the efficacy of the PGM fusion method, we employ both a checkerboard model and the fault zone structure imaged from the 2019 Ridgecrest, CA, earthquake sequence. The Ridgecrest fault zone image consists of a shallow (representing a depth of 0.5 km) high-resolution Rayleigh wave model [64], from which the S-wave velocity is approximated by dividing by 0.9.

Fig. 3.7 shows the smoothed results with different smoothing levels. The aggressive, simple smoothing policy in Fig. 3.7(c) (7×7 average filter) removes the sharp boundaries while the details are lost. Simple and mild smoothing in Fig. 3.7(d) (3×3 average filter) preserves the

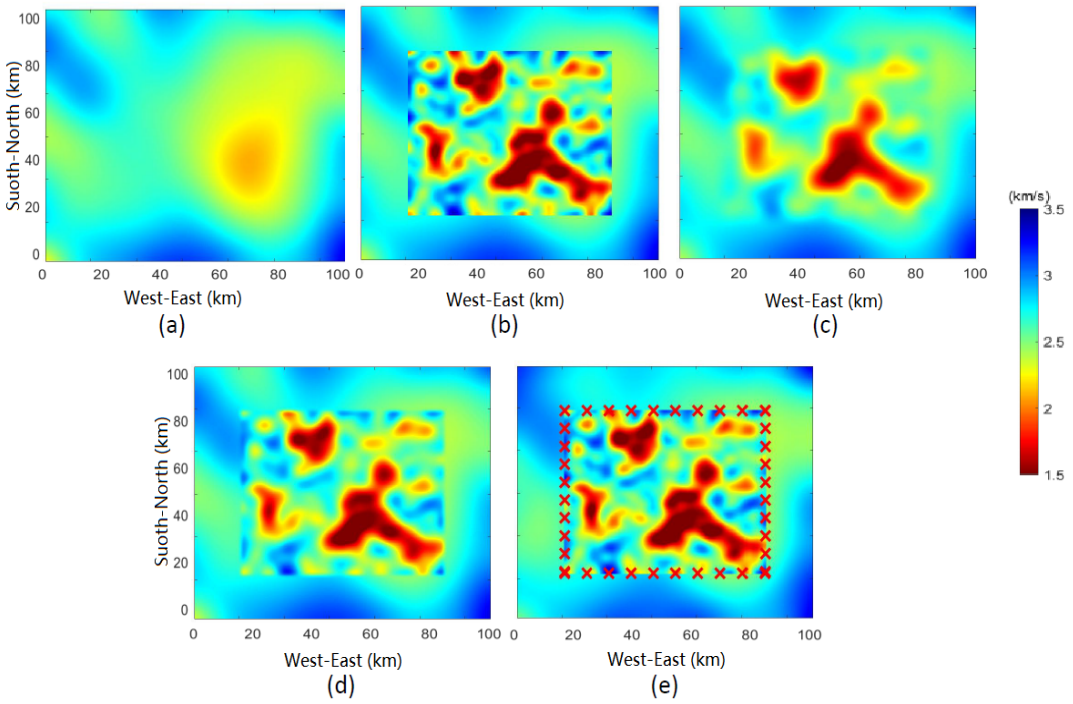

Figure 3.7. (a) LR model (Same as Fig 1a). (b) Same as Fig 1c. (c-d) Combined LR and HR models, smoothed by (c) 7×7 and (d) 3×3 average filters. (e) Synthetic stations (red 'X's) are deployed on the boundaries between HR and LR models for evaluation.

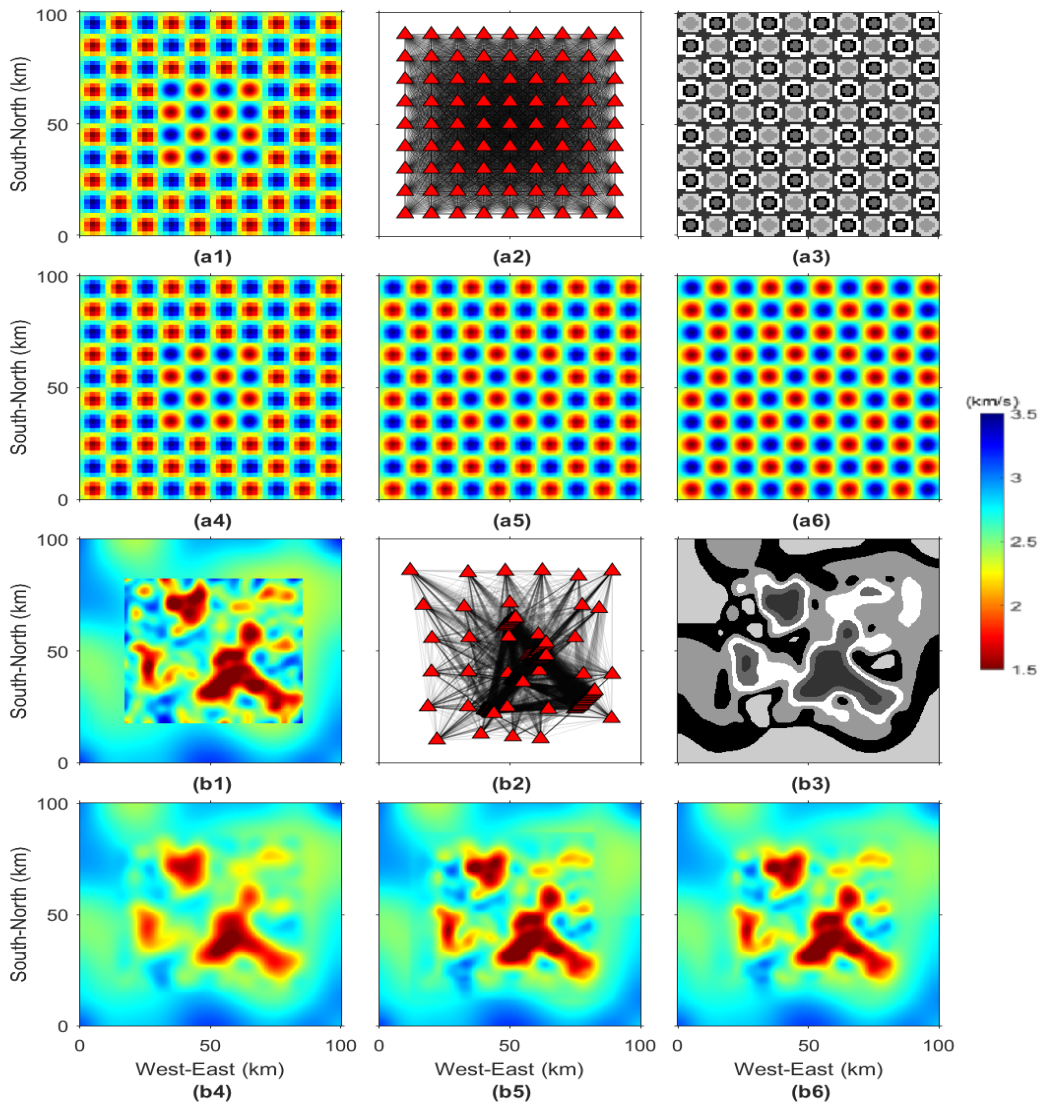

Figure 3.8. (top, 'a' panels) Checkerboard and (bottom, 'b' panels) Ridgecrest models. (1) Superimposed HR and LR models. (2) Station location and ray density. (3) 6-class label mask maps for HR models (pixels with the same label are learned together). (4) Smoothing results with a 5×5 Gaussian filter (GF). (5) and (6) Fusion results with DL and with PIPGM.

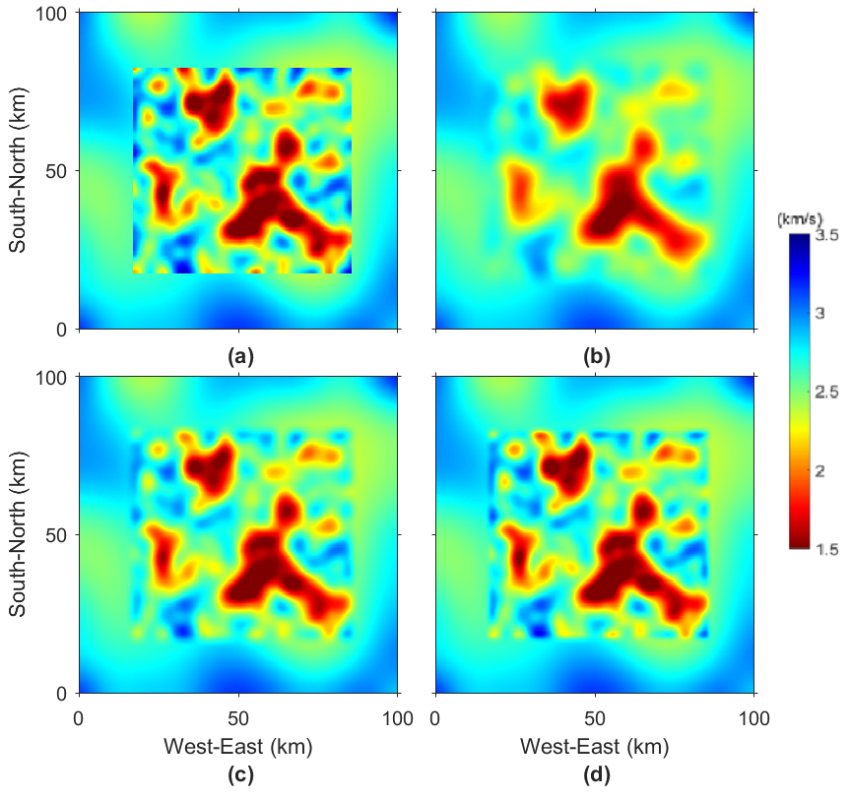

Figure 3.9. (a) Direct superposition of the HR 1 Hz Rayleigh wave tomographic velocity and CVM LR models for the Ridgecrest area. (b-d) Combined LR and HR models, smoothed by (b) 3rd-order, (c) 2nd-order, and (d) 1st-order MRFs. Generally, the higher the order is, the larger the range of neighboring pixels will be taken into consideration, and this leads to a model with smoother local patterns.

details together with the artificial boundaries. It is essential to achieve a trade-off between the two cases, this is the aim of PIPGM. To assess the balance between travel time among stations and their residuals, 36 evaluation points (marked as red 'X', with 10 situated along each edge) are positioned at the boundary dividing the LR and HR zones. These residuals gauge the extent of detailed information retained in the merged velocity model relative to the HR maps.

3.4.1 Graphical Structure Order Test

In an MRF, the variables are represented as nodes in the graph, and their dependencies with the neighbors are decided by the truncation order in Eq. (3.6). We name the MRF with the neighboring system truncated at *n*th order as 'n-th order MRF' for brevity. A 1st-order

neighboring system structure is a subset of nodes in which every two distinct nodes are directly adjacent. Usually, the neighboring system's order significantly affects the model's smoothing results. If an MRF model is based on a first-order neighboring system, it can account for interactions between immediate neighbors. In the context of image smoothing, it results in the enforcement of local smoothness, where each pixel is encouraged to be similar to its immediate neighbors. When a model involves a higher-order neighboring system (the Manhattan distance of the pixels in this set from the centering pixel may be greater than 1), it can capture more complex relations and dependencies among variables. Higher-order MRFs can enforce smoothness over larger regions, allowing the model to preserve the high-level structures (usually related to the low-frequency patterns) and avoid sharp boundaries or noisy details.

Since information from a wider range of neighbors is considered, higher-order MRFs can be computationally more intensive in one iteration and may reach the optimum with fewer learning iterations. On the other hand, first-order MRFs are simple and efficient, and they can sometimes preserve some noisy details. The choice between first-order and higher-order MRFs depends on the specific requirements of the task and the trade-off between model complexity and computational efficiency. In the velocity fusion problem, we demonstrate the fused models with 1st-, 2nd-, and 3rd-order MRF structures in Fig. 3.9. Considering that the Ridgecrest model measures the structure from a relatively limited region and that we prefer the rich detailed structures from the HR model, we choose the 1st-order MRF neighboring structure in the following comparison experiments.

3.4.2 Cluster Number Test

The number of GMM clusters influences the model complexity and interpretability of the results. More clusters result in a more complex model, which can capture intricate data structures better and lead to more detailed insights into the data. However, it increases the risk of overfitting.

The optimal number of clusters is crucial in GMM and other clustering techniques. Sev-

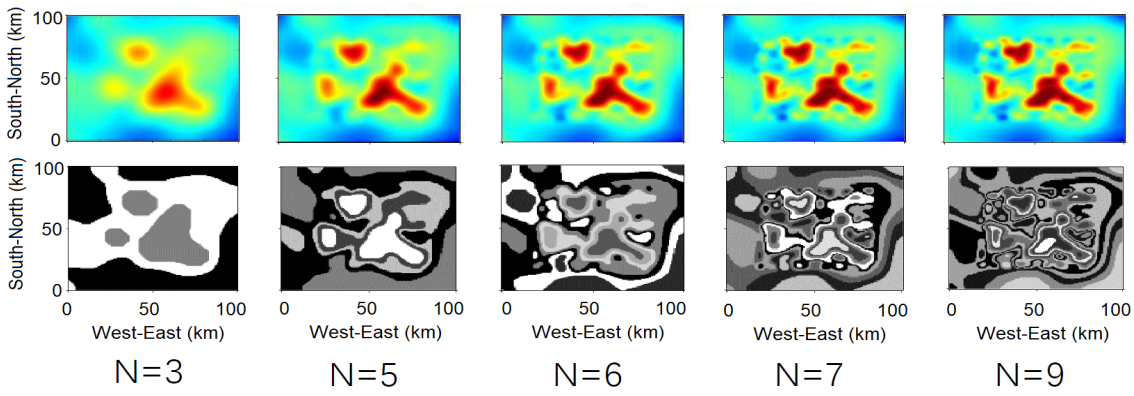

Figure 3.10. Fusion using cluster numbers n= 3, 5, 6, 7, and 9. (top row) fused velocity model and (bottom row) cluster distribution. Generally, the larger the number of clusters, the more detailed information is preserved in the HR models, and the larger the computation recourse it needs. A trade-off is required to balance the cost and the performance.

eral methods help determine an appropriate number of clusters, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the silhouette score. These methods balance the trade-off between the goodness-of-fit of the model and the complexity of the model.

For this experiment, we tested and compared the cluster number sequences $N \in \{3,4,5,6,7,9\}$, which is prevalent in practical uses of MRFs. Figure 3.10 shows that the larger the number of clusters, the more detailed information is preserved in the HR models, and the larger the computation recourse it needs. This implies that a trade-off between cost and performance needs to be achieved. Figure 3.11 demonstrates the number of clusters against RMSE error (left vertical axis) and computation time (right vertical axis) in the Ridgecrest test. The RMSE error experiences a significant decrease while the cluster number grows from 3 to 6, with limited decrease for larger cluster numbers, and there is a rapid growth in run time when the cluster number exceeds 6. For these reasons, we empirically choose the cluster number to be 6 as a trade-off between model performance and computation complexity.

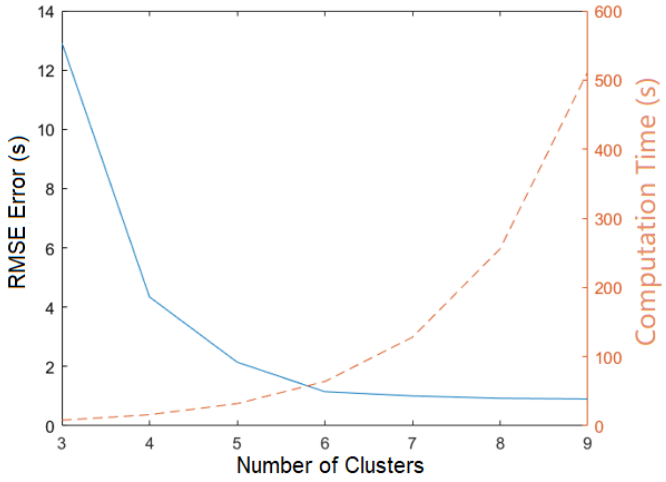

Figure 3.11. The number of clusters versus RMSE error (left vertical axis, corresponding to the blue solid line) and run time (right vertical axis, corresponding to the orange dashed line) on the Ridgecrest model. Generally, the larger the cluster number is, the smaller the RMSE error and the longer the running time is. We choose cluster number 6 to balance the model performance and the computation complexity.

3.4.3 Comparison with Conventional Methods

To demonstrate the advantage of the proposed PIPGM, we compare its performance with some commonly used conventional methods on both the synthetic checkerboard model and the real-data Ridgecrest model with inter-station S-wave travel time deviations before and after the merge. First, we briefly introduce several popular conventional fusion methods.

Gaussian Smoothing Filter (GF)

A GF is a data processing technique that reduces noise and smooths out signals or data distributions. This method involves applying a Gaussian kernel function to the data, a bell-shaped curve that weights the data points based on their distance from a central point. A smoothing parameter defines the Gaussian kernel called the kernel window size or standard deviation, which controls the amount of smoothing applied to the data. A larger window size leads to a wider, smoother curve, while a lower bandwidth results in a narrower, more detailed curve. GFs are commonly used in image and signal processing, and data analysis applications. It is a powerful

technique that can effectively remove noise and improve the clarity of data, but can also introduce bias and distortions in the data if the smoothing parameter is not chosen carefully.

DL Smoothing

DL Smoothing [80] is a data processing technique used to de-noise and smooth misaligned patterns in the signals or images. It involves representing the data as a sparse linear combination of a set of basis vectors or atoms, which is learned from the data itself through an iterative process. In this process, the algorithm attempts to identify an array of basis vectors capable of representing the data with minimal error. Once the dictionary is learned, it can be used to transform the data into a sparse representation, where most of the coefficients are zero. This sparse framework is subsequently employed in the denoising and smoothing of the data, achieved through the selective elimination or alteration of coefficients associated with noise or extraneous features. DL smoothing has been applied to a wide range of data processing applications, including image processing, audio processing, and signal processing. It is a powerful technique that effectively removes noise and preserves the underlying structure of the data, but it requires a large amount of training data and can be computationally expensive.

3.4.4 Results

We show the fusion of HR and LR components for directly superimposed checkerboard and Ridgecrest velocity models in Fig. 3.8 (a1, b1), which both have an HR region in the center, surrounded by LR velocities in the surrounding areas. The checkerboard model has 40×40 pixels in the 100×100 km LR area and 40×40 pixels in the 40×40 km HR area. The fused model has 100×100 pixels in the 100×100 km LR area. Similarly, the Ridgecrest model has 50×50 pixels in a 100×100 km LR area and 192×224 pixels in a 58×64 km HR area. The fused model has 330×350 pixels in a 100×100 km area. Fig. 3.8 (a2, b2) shows the checkerboard and Ridgecrest stations along with the ray-path density. For the checkerboard model, the stations are evenly distributed, whereas the stations for the Ridgecrest model are highly irregular, reflecting

Table 3.1. Evaluation results for Gaussian Filtering (GF), DL, classical PGM, and PIPGM for both checkerboard and the Ridgecrest model. Evaluation metrics are travel time root-mean-square error (RMSE), naturalness image quality evaluator (NIQE), peak signal-to-noise ratio (PSNR), and Fréchet inception distance (FID). ↓ indicates smaller is better, and ↑ opposite.

	RMSE/s↓	**NIQE↓**	**PSNR/dB↑**	**FID↓**
Checkerboard GF	1.65	7.68	14.58	45.75
” DL	1.18	5.44	15.70	33.85
” PGM	1.14	**5.40**	**16.14**	32.49
” PIPGM	**1.06**	5.41	**16.14**	**32.46**
Ridgecrest GF	3.52	12.41	21.80	61.39
” DL	2.61	7.29	22.36	54.25
” PGM	2.27	6.70	23.04	47.49
” PIPGM	**2.17**	**6.59**	**23.16**	**47.18**

the pattern used by Zhou et al. [64]. Fig. 3.8 (a3, b3) shows the label mask maps from the last iteration of the PIPGMs. Pixels with the same label indicate that these areas potentially share similar velocity patterns and are sampled from the same distribution. The smoothed fusion results with the 5×5 GF, DL [80], and PIPGM are shown in Fig. 3.8 (a4-a6) and (b4-b6). The results show that the learning methods (DL and PIPGM) preserve more detailed information than direct Gaussian smoothing. This is because the learning methods adaptively find the fusion parameters that optimize the accuracy of the representation, while Gaussian smoothing only combines the neighboring pixels with a predefined kernel.

We use multiple metrics to evaluate the model fusion efficacy: travel time Root-Mean-Squared-Error (RMSE, which measures information lost after model fusion [23]), Naturalness Image Quality Evaluator (NIQE, a common-used measurement for image quality [81]), Peak Signal-to-Noise Ratio (PSNR, measuring the sharpness of images [82]), and the Frechet inception distance (FID, capturing similarities between the original and fusion models [83]) in Table 3.1. In the checkerboard test, because the pattern is simple and the stations are evenly distributed, all the learning methods achieve similar performance. For the realistic Ridgecrest model, however, the PGMs outperform the DL model, as the latter is sensitive to the orientation of the patches while the graphical models are rotationally invariant. We observe further improvements after the

ray density information is informed in the PIPGM.

3.5 Conclusion

We developed an approach to fuse multiresolution seismic velocity maps with PGMs. PIPGM achieves a velocity model by balancing smoothing the generally undesired sharp boundaries between LR and HR components and preserving the detailed information from the HR models.

We tested the efficacy of the proposed fusion method using a checkerboard model and a realistic, complex fault zone model around the 2019 M7.1 Ridgecrest earthquakes. The tests for the checkerboard model, which is simple with evenly distributed stations, show that PIPGM outperforms all the baseline methods. All the learning-based techniques used to combine the maps (including the proposed PGM and PIPGM methods) significantly outperform the conventional methods, since the parameters are adaptively learned from the pairwise data. For Ridgecrest, the PIPGM methodology demonstrates a 38% reduction of travel time residuals compared to Gaussian kernel smoothing. This result is likely due to the poor performance of traditional techniques in handling the nonuniform-distributed data evenly, while PIPGM adaptively chooses the weights according to the intricate structure of the Ridgecrest model.

The proposed PIPGM outperforms traditional techniques in integrating maps of varying resolution, particularly when the map complexity escalates and the distribution of data points is uneven. Prospective future work involves addressing the challenge of irregular resolution within the HR domain, which is crucial for enhancing the fidelity and applicability of our models and potentially improves the understanding and application of various real-world models. Finally, we recommend testing PIPGM directly by comparison of synthetic and observed waveforms.

3.6 Acknowledgments

The text of Chapter 3, in full, is a reprint of the material as it appears in Zhou, Z., Gerstoft, P. and Olsen, K.B. "Fusion of Multi-Resolution Seismic Tomography Maps with Physics-Informed Probability Graphical Models", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 2024, and Zhou, Z., Gerstoft, P. and Olsen, K.B. "Graph-learning approach to combine multiresolution seismic velocity models", *Geophysical Journal International*, p.ggae212, 2024. The dissertation author was the primary researcher and author of Chapter 3. The coauthors listed in this publication directed and supervised the research.

Chapter 4

3D Multiresolution Velocity Model Fusion With Probability Graphical Models

4.1 Introduction

The problem of combining multiscale models appears across various geophysical fields, revealing the scale-dependent nature of anisotropy and introducing substantial implications for understanding Earth's structure [42]. The integration of velocity models with different resolutions is important for refining community models, especially in applications such as ground motion estimation or dynamic rupture modeling, where varying scales are imperative [e.g., 40, 84]. The fusion of high-resolution (HR) and low-resolution (LR) models poses challenges due to the potential emergence of sharp boundaries and misaligned patterns. Apart from being physically unrealistic, such patterns can result in artifacts in ground motion simulations.

To address the velocity model fusion problem, several notable techniques have been developed. The Gaussian kernel filter [85], widely used in image and signal processing, applies a Gaussian kernel to data for smoothing, with the degree of smoothing determined by the kernel's bandwidth or standard deviation. This technique is effective in enhancing the clarity of the data and requires precise parameter tuning to avoid data distortion. Cosine taper window interpolation [41, 86] employs a cosine taper to reduce signal amplitude at the sequence ends, thereby merging the two velocity models effectively while preserving their overall characteristics. However, both Gaussian filtering and cosine tapering have difficulties when considering the spatial and

directional dependency of the resolution, which might induce direction-oriented smoothing [44]. On the other hand, dictionary learning smoothing, explored in studies [80, 32, 87], offers an advanced approach for smoothing signals or images. This technique involves creating a sparse representation of data through dictionary learning, enabling effective noise removal while maintaining the data's underlying structure, albeit with high computational demands and extensive training data requirements.

Inspired by advancements in image super-resolution [67] and image editing [44, 69], we propose to fuse seismic velocity models employing Probability Graphical Models (PGMs). PGMs, capable of processing complex structures due to their ability to discern inherent relations among images [50, 51], represent a promising tool for seismic analysis, including the study of reflection and seismic attributes surrounding low-velocity zones. Expanding beyond standard graphs, PGMs have been extended to higher-dimensional spaces, such as multilayer graphs [55] and hypergraphs [56], and have been used in several seismic applications [59, 60, 61]. Within all the PGMs, Markov Random Field (MRF) is a prevalent and highly effective approach for tackling supervised structure learning tasks that encompass the intricate mapping of complex geometric structures, as articulated by Murphy et al. [62]. MRFs have been instrumental in the area of image restoration and editing, which was initially conceived by Geman et al. [8]. This approach is rooted in Bayesian inference principles, applied to a spatially stochastic model. In contrast to convolution-based methods, the MRF procedure has been empirically validated to yield optimal and mathematically tractable results in the context of image processing, as substantiated in [9].

To combine realistic tomography velocity models with unevenly distributed patterns, we propose a PGM that captures the relations between subdomains with different resolutions. Our focus is on models that distinctly segregate high-resolution (HR) and low-resolution (LR) areas. By learning information from the HR subdomain, we aim to enhance the details within the LR regions. This enhancement is achieved through a maximum likelihood formulation, which incorporates prior knowledge from the HR areas. Our proposed PGM fusion technique not only

accentuates the local HR structure, but also safeguards global smoothness in the resulting model, an important advance in addressing the complexities in 3D velocity model fusion. Due to the inherent characteristics of graphical methods, our approach embodies the property of invariance under rotations or angular transformations. Furthermore, it accommodates the extension of larger neighboring patch sizes, facilitating the adaptation to varying volumes of training data and allowing for the accommodation of varying data quantities during training. Our approach enables more inclusive, adaptive, and precision-enhanced modeling of Earth's subsurface structures, showing promise for the PGM in the fusion of 3D models spanning varied resolutions.

Tests are performed on both a checkerboard and a fault zone model derived from the 2019 Ridgecrest, CA, region to demonstrate its efficacy. Generally, a lower travel time deviation indicates a more accurate velocity model of the Earth's subsurface, which in turn leads to more precise ground motion simulations [88]. Our model is evaluated by the deviation between observed and calculated travel times and demonstrates that our PGM has great potential in combining velocity models with various resolutions (see Experiments Section).

4.2 Model Setup

For two partially observed velocity fields $\boldsymbol{A}_{\mathrm{LR}}$ and $\boldsymbol{A}_{\mathrm{HR}}$, the task is to estimate the true velocity field $\boldsymbol{A}$. Here, we let $\boldsymbol{A}_{HR}$ represent a high-resolution velocity field on pixels inside the low-resolution $\boldsymbol{A}_{\mathrm{LR}}$ velocity field. In this paper, we focus on optimally merging borders between $\boldsymbol{A}_{\mathrm{LR}}$ and $\boldsymbol{A}_{\mathrm{HR}}$, as illustrated in Fig 4.1 using an excerpt from the Statewide California Earthquake Center (SCEC) Community Velocity Model (CVM) version S-4.26 as well as a HR fault zone model from the Ridgecrest, CA, area. We illustrate our method using 6 labels, a choice that will later be shown to be optimal in the trade-off between model complexity and computational cost (see Section Experiments). In our graphical model, a discrete class label map helps tie the spatial velocity field together. The label maps represent different geological structures, defined by their association with certain velocity intervals. The label map is initialized from the continuous

velocity map $\boldsymbol{A}$ (Fig. 4.2, left), where we define a 6-cluster discrete label map $\boldsymbol{X}$ (Fig. 4.2, right) containing 6 velocity intervals (labeled 1–6), which is obtained from the continuous velocity maps $\boldsymbol{A}$.

In our 3D models, the pixels are described by (i, j, k) coordinates and contain both a label $X_{i,j,k}$ and a velocity $A_{i,j,k}$. The velocity $A_{i,j,k}$ with the label $X_{i,j,k} = n$ (n represents the labels, an example is shown in Fig. 4.2, right) follows a Gaussian distribution $\mathcal{N}(\mu_n, \sigma_n^2)$ with mean μ_n and variance σ_n^2. Velocities at different pixels but with the same labels follow the same distribution. Thus, in a graph, the velocities $\boldsymbol{A}$ are on top of the labels $\boldsymbol{X}$ (Fig. 4.3). d denotes the set of all possible labels of $X_{i,j,k}$ (here, $n = \{1, \ldots, 6\}$), and D represents the set of all possible combinations of labels $\boldsymbol{X}$ for the entire map. The whole map is tied together via the class labels $\boldsymbol{X}$ that depend on the neighboring class labels indicated by the graphical grid structure in Fig. 4.3. For each point (i, j, k) the neighboring class $\mathcal{N}_{i,j,k}$ is defined by its four immediate points. Note that we use the points in specific regions $\mathcal{V}$, i.e., at the border between the low- and high-resolution maps.

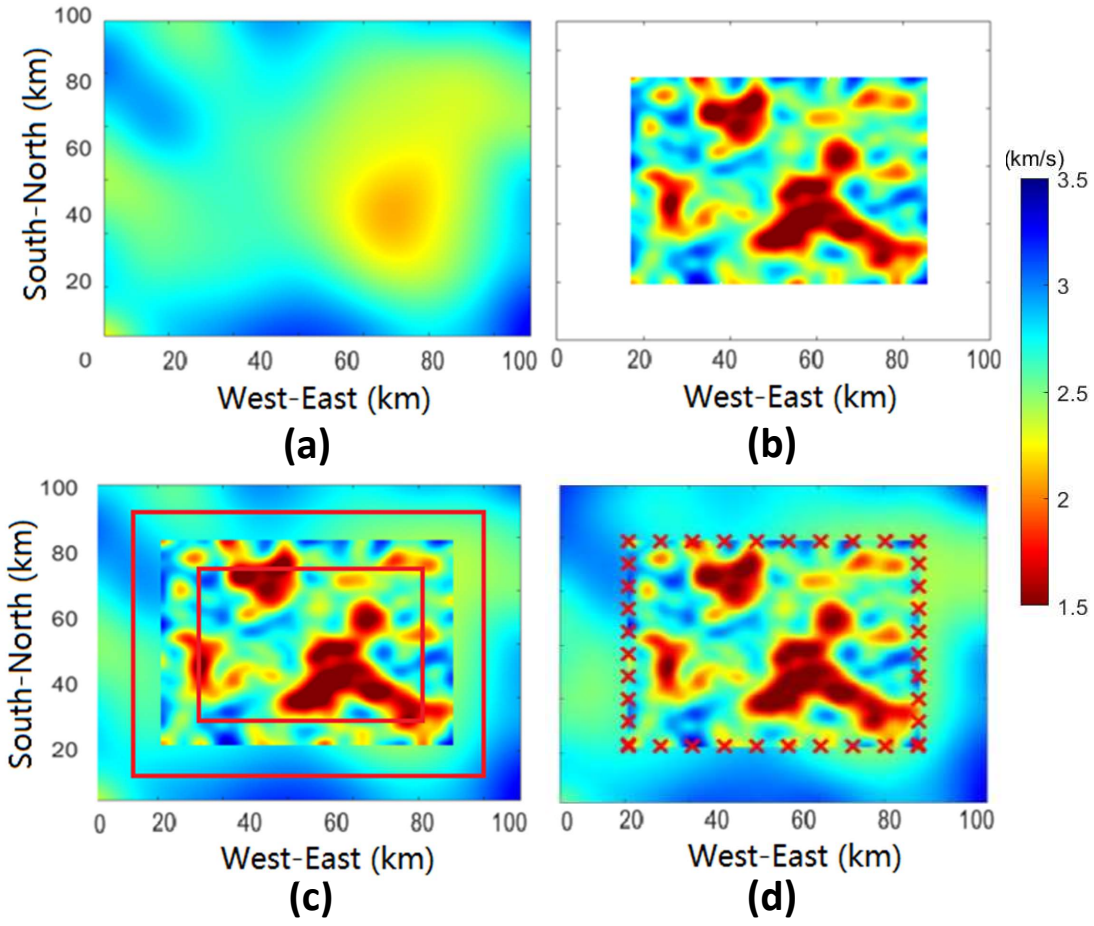

Figure 4.1. (a) Excerpt of S-wave velocities from the SCEC CVM version S-4.26 (hereafter referred to as the low-resolution (LR) CVM) at 0.5 km depth around the Ridgecrest area. (b) High-resolution (HR) S-wave map from 1 Hz Rayleigh wave tomography from [64]. (c) Direct superposition of the HR and LR models. (d) Evaluation points ('X') are located on the boundaries between the HR and LR models. These two models share some patterns in the low-velocity zones, but show many mismatched detailed patterns where the two models overlap, which results in sharp and misaligned boundaries in those areas. Our PGM is applied to the mismatched boundary areas between the two bounding boxes in (c). The pixels in this area belong to the effective vertices set.

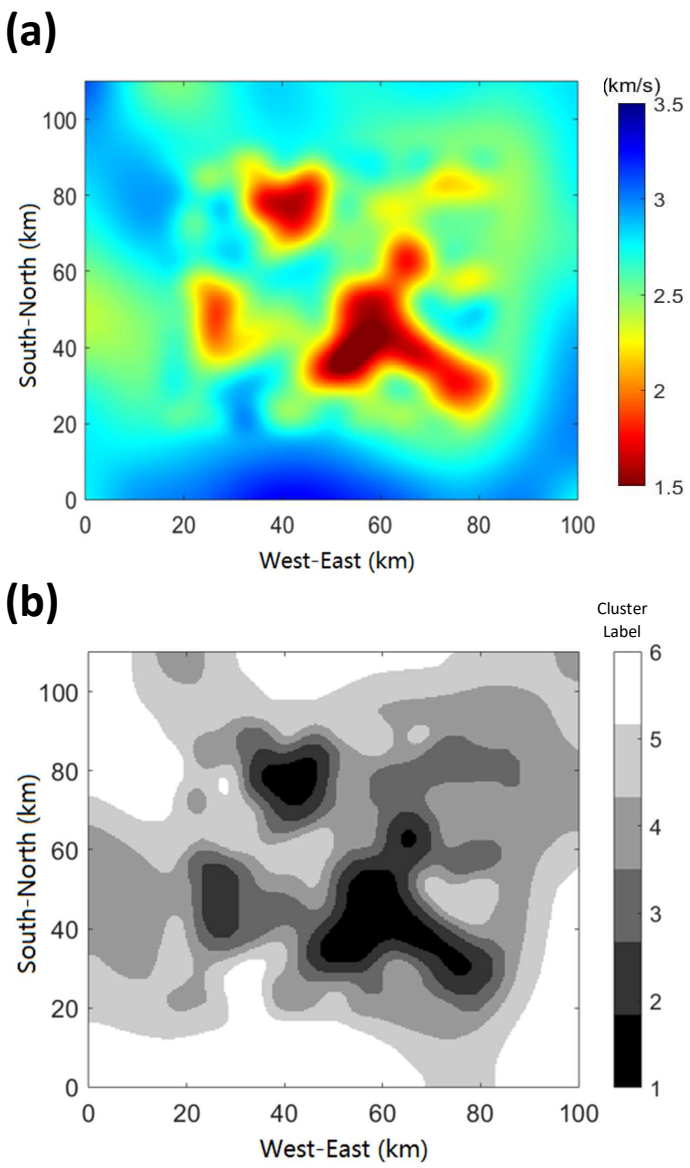

Figure 4.2. A 6-cluster Gaussian mixture model clustering is applied on the continuous velocity map $\boldsymbol{A}$ in (a). Each pixel is clustered to be a 6-cluster discrete label map $\boldsymbol{X}$ in (b). Pixels with similar velocity information have been assigned the same label.

4.3 Markov Random Field Models (MRFs)

4.3.1 Bayesian Estimation

Given the prior probabilities $P(\boldsymbol{X})$ of label $\boldsymbol{X}$ and the likelihood densities $P(\boldsymbol{A} \mid \boldsymbol{X})$ of the observed velocity $\boldsymbol{A}$, the posterior probability can be formulated through Bayes' theorem as:

$$P(\boldsymbol{X} \mid \boldsymbol{A}) = \frac{P(\boldsymbol{A} \mid \boldsymbol{X})P(\boldsymbol{X})}{P(\boldsymbol{A})} \propto P(\boldsymbol{A} \mid \boldsymbol{X})P(\boldsymbol{X}). \tag{4.1}$$

Here, the probability density function (PDF) $P(\boldsymbol{A})$ of $\boldsymbol{A}$ is a fixed probability distribution (for given $\boldsymbol{A}$) and does not affect the maximum a posteriori (MAP) estimation solution. The Bayesian labeling problem requires finding the MAP configuration. The MAP of labeling for observation $\boldsymbol{A}$ is given by:

$$\boldsymbol{X}^* = \arg\max_{\boldsymbol{X} \in D} P(\boldsymbol{X} \mid \boldsymbol{A}), \tag{4.2}$$

where D denotes a set of possible candidates of the discrete labels $\boldsymbol{X}$, and $\boldsymbol{A}$ represents the observation of the continuous velocities [89]. To derive the MAP solution, both the prior probability and the likelihood function are needed. The likelihood function $P(\boldsymbol{A} \mid \boldsymbol{X})$ captures the conditional relation between the observation (refers to the continuous velocity in our research) and the hidden states (the variable, which corresponds to the discrete labels, cannot be directly observed here).

4.4 Probability Graphical Model (PGM)

Our PGM follows a first-order MRF setting (Fig. 4.4) where each random variable has four neighbors on which it is conditionally dependent. The full conditional probability of the discrete random variable $X_{i,j,k} \in \{1, \ldots, n\}$ is the exponential of the sum of potentials (four 1st-order neighboring potentials θ_1 between cluster labels and one 0th-order center data potential θ_0 between cluster label and velocity) in conventional MRF settings. In image processing problems,

optimizing the entire map can be broken down into a group of suboptimization problems that optimize each pixel iteratively [73]. Inserting (3.9) into (3.12), we have

$$\begin{aligned} -\log p(X_{i,j,k} \mid A_{i,j,k}) &= U_{\text{post}}(X_{i,j,k} \mid A_{i,j,k}) \\ &\propto \theta_0(X_{i,j,k}, A_{i,j,k}) + \sum_{(i',j',k') \in \mathcal{N}_{i,j,k}} \theta_1(X_{i,j,k}, X_{i',j',k'}) + C, \end{aligned} \tag{4.3}$$

$$\theta_0\left(X_{i,j,k}, A_{i,j,k}\right) = \frac{(A_{i,j,k} - \mu_n)^2}{\sigma_n^2}, \tag{4.4}$$

where θ_0 is the 0th-order neighboring potential [74] (also known as the data cost function) that relates $X_{i,j,k}$ to the observed velocity data $A_{i,j,k}$. μ_n and σ_n^2 are the mean and variance, respectively, of all pixels with the same cluster label $n = X_{i,j,k}$. It promotes that continuous velocity values $\boldsymbol{A}$ sharing pixels with the same discrete label $\boldsymbol{X}$ follow the same Gaussian distribution.

$$\theta_1\left(X_{i,j,k}, X_{i',j',k'}\right) = 1 - \delta(X_{i,j,k}, X_{i',j',k'}) = \begin{cases} 0 & \text{for} \quad X_{i,j,k} = X_{i',j',k'} \\ 1 & \text{otherwise} \end{cases} \tag{4.5}$$

is the 1st-order neighboring potential [74] (or the smoothness cost function), where δ is the Dirac delta function, that relates $X_{i,j,k}$ to the 1st-order neighboring variable $X_{i',j',k'}$ (see Fig. 4.4). This function encourages the neighboring pixels to share the same discrete label $X_{i,j,k}$, promoting the model's local smoothness.

The performance of standard or potential function-based MRF schemes can be limited when dealing with complex geological structures ([90]). Assigning different neighboring pixels with various importance weights based on anisotropy patterns can effectively remove the non-uniform direction-dependent features of the model gradients, leading to improved inversion results, especially relevant for real, complex geological structures.

The objective function for the MAP problem of $X_{i,j,k}$ becomes

$$X^*_{i,j,k} = \arg\max_{\boldsymbol{X}_{i,j,k}} p(X_{i,j,k} \mid A_{i,j}) = \arg\min_{\boldsymbol{X}_{i,j,k}} \omega^0_{i,j,k}\theta_0(X_{i,j,k}, A_{i,j,k}) + \sum_{(i',j',k')\in\mathcal{N}_{i,j,k}} \omega^1_{i',j',k'}\theta_1(X_{i,j,k}, X_{i',j',k'}) + C, \tag{4.6}$$

where $\omega_{i,j,k}$ and $\omega_{i',j',k'}$ are the weights which balance the anisotropic characteristics. These weights are typically set to uniform default values, given by the number of pixels in the local neighborhood, here $\omega^0_{i,j,k} = 1/1$ and $\omega^1_{i',j',k'} = 1/6$. This implies an equal contribution from each neighboring pixel.

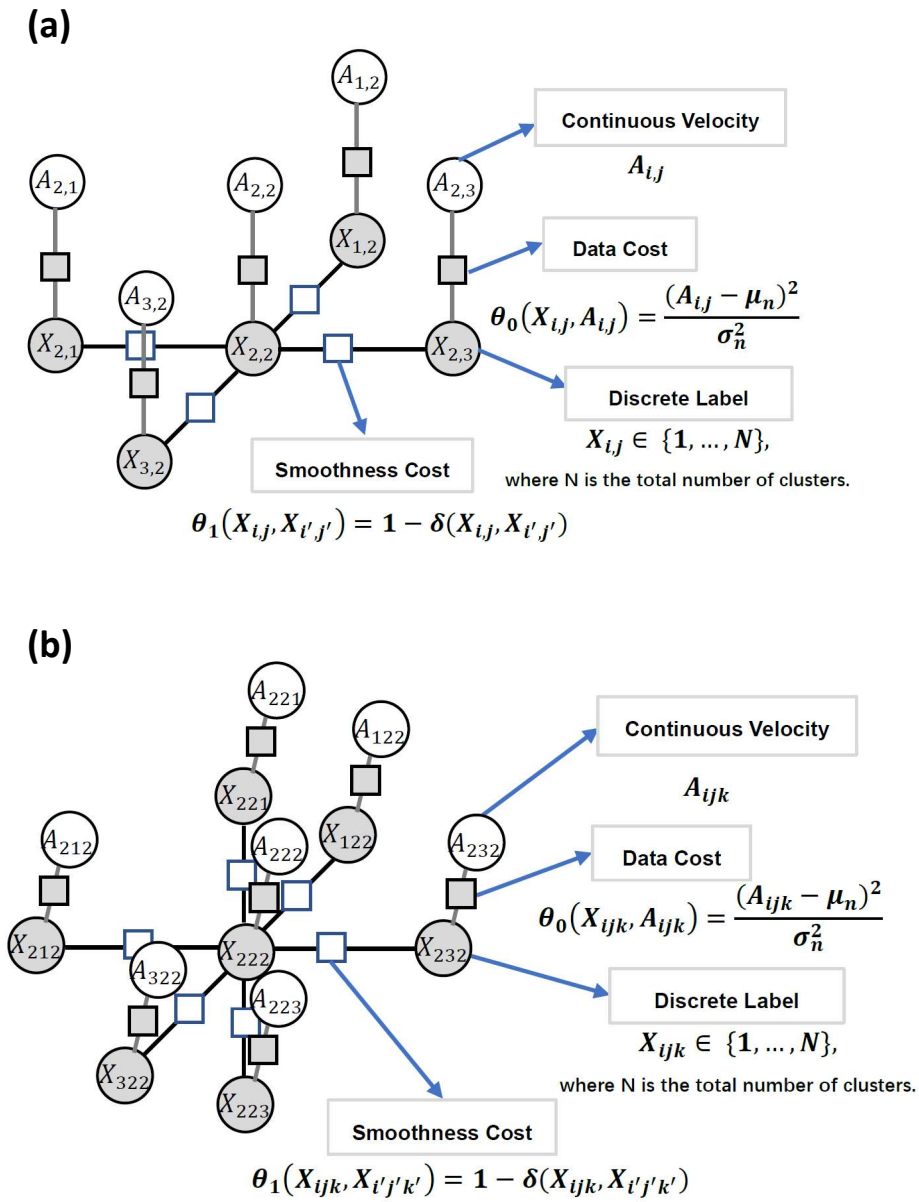

Figure 4.3. (a) In a 2D graphical model, each pixel has a continuous velocity $A_{i,j}$ and a discrete label mask $X_{i,j}$. The objective function has two parts: (1) the data cost θ_0 (0th-order neighboring potential) that forces a pixel to have a $A_{i,j}$ specified by the Gaussian distribution of its label $X_{i,j}$, and (2) the smoothness cost θ_1 (1st-order neighboring potential) that promotes smoothness among neighboring labels [7]. (b) For a 3D graphical model, the framework closely aligns with 2D (a). The 3D model encompasses 6 adjacent nodes (while in 2D, it has 4 nodes).

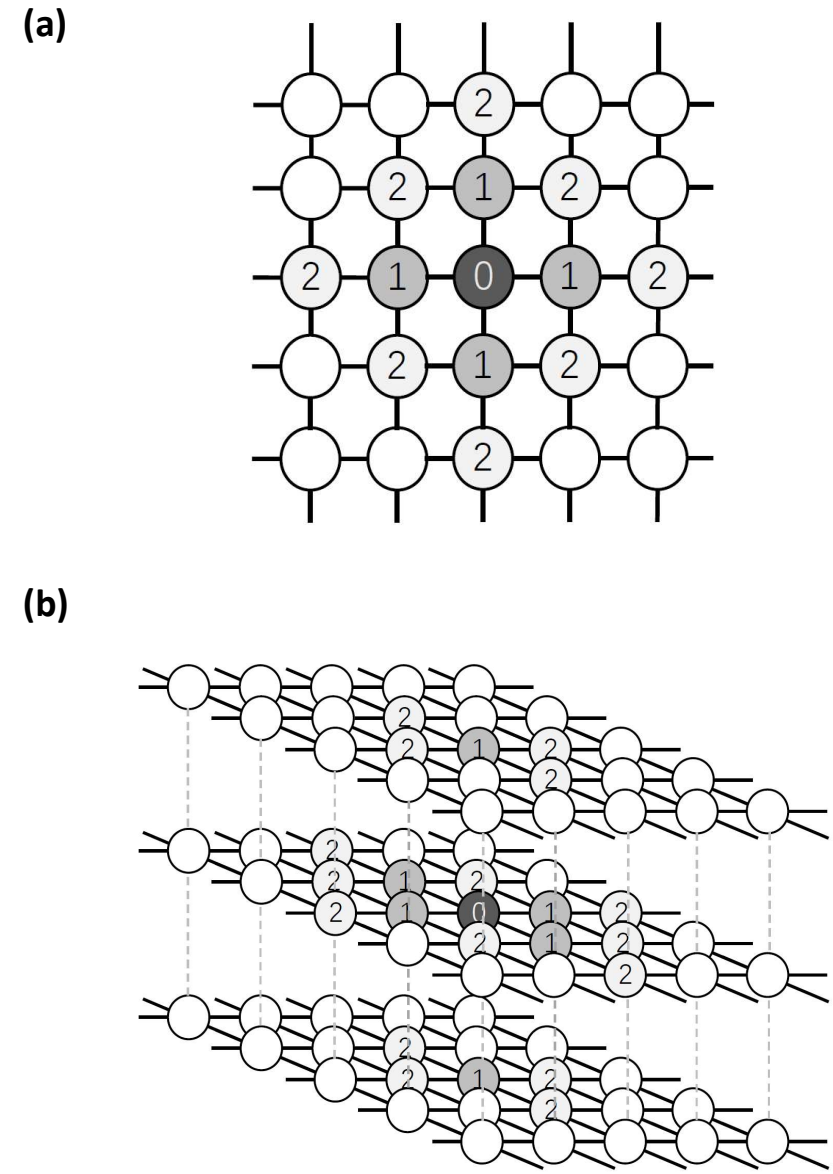

Figure 4.4. (a) Neighborhood system $\mathcal{N}_{i,j}$ (marked in gray/black) of the given center node (i, j) (marked in black) in a 2D graphical model. The 1st- and 2nd- neighborhood systems of node (i, j), marked with numbers 1 and 2, are represented as $\mathcal{N}_{i,j}^1$ and $\mathcal{N}_{i,j}^2$, respectively. Number 0 denotes the center pixel. (b) 3D graphical model similar to (a), where 0, 1, and 2 denote the center pixel and the 1st-, and 2nd- neighborhood systems, respectively.

4.4.1 MCMC and Gibbs Sampling

Markov Chain Monte Carlo (MCMC) is a statistical method used to sample probability distributions [76, 77]. Gibbs sampling is a specific MCMC algorithm that can be used to iteratively sample a multivariate probability distribution from the conditional distributions of each variable given the current values of the other variables [78]. Combining MCMC with Gibbs sampling enables estimating complex probability distributions without explicit knowledge of the distribution.

We employ the MCMC method with Gibbs sampling to solve Eq. (3.13). Gibbs sampling generates a new sample of $X_{i,j,k}$ directly from its distribution conditioned on the labels of its neighbors $X_{i',j',k'}$ and $A_{i,j,k}$. In the MRF structure, the update is achieved by calculating the probability for each of the possible labels (here, $n \in \{1, \cdots, 6\}$) at (i, j, k) using Eq. (4.6) and randomly selecting from this distribution (refer to Fig. 4.5).

The velocity map $\boldsymbol{A}$ is initialized with the superimposed HR and LR velocity maps, see Fig. 4.1 (a) and (b), and the label map $\boldsymbol{X}$ is initialized with a Gaussian mixture model clustering with the total cluster number N (here, 6), similar to Fig. 4.7 (a2, b2). All velocities with the label n follow the same Gaussian distribution $N(\mu_n, \sigma_n^2)$. We then apply the expectation–maximization (EM) algorithm [79], an iterative method to find the MAP estimates of the parameters, which updates the Gaussian parameters μ_n and σ_n^2. The termination criterion is either reaching a predefined maximum number of iterations (here 10,000) or observing that the cumulative absolute difference across all pixels between consecutive iterations falls below an error threshold, whichever is achieved first. The algorithm is summarized in Table 4.1.

Summarizing the algorithm from an intuitive perspective, our PGM adjusts each point in the grid-based method, not only on the point itself as in many conventional approaches, but also on the values of the surrounding points. The model processes each pixel, adjusting its values to align more closely with its neighbors, resulting in smoother and more consistent results. Our approach is analogous to a diffusion process, similar to introducing ink into clear water,

where the resulting patterns gradually spread throughout the entire system. In the context of image processing, the algorithm methodically traverses each pixel, recalibrating its coloration to achieve harmonious alignment with nearby pixels. This paradigm enhances overall smoothness and significantly reduces aberrations, thereby increasing the consistency of the entire model.

Table 4.1. Algorithm for 3D multiresolution velocity model fusion.

Algorithm 1 MCMC Method for MRF

1. Input: $\boldsymbol{A}_{LR}$ and $\boldsymbol{A}_{HR}$
2. Initialize velocity $\boldsymbol{A}$ by superimposing $\boldsymbol{A}_{HR}$ over $\boldsymbol{A}_{LR}$
   Initialize $\boldsymbol{X}$, $\mu_{\mathbf{n}}$ and $\sigma_{\mathbf{n}}$ with GMM clustering
3. **for** each EM iteration **do**
4. Construct PGM
5. **for** $t = 1$ to max iteration T **do**
6. (E-Step) Gibbs Sampling
7. **for** pixel $(i, j, k) = (1, 1, 1)$ to the max index (I, J, K) **do**
8. $X_{i,j,k}^{(t+1)} \sim P\left(X_{i,j,k} \,|\, X_{1,1,1}^{(t+1)}, \ldots, X_{i,j,k-1}^{(t+1)}, A_{i,j,k}^{(t)}, X_{i,j,k+1}^{(t)}, \ldots, X_{I,J,K}^{(t)}\right)$
9. **end for**
10. **for** pixel $(i, j, k) = (1, 1, 1)$ to the max index (I, J, K) **do**
11. $A_{i,j,k} \sim \sum_{n=1}^{6} P(X_{i,j,k} = n) N(\mu_n, \sigma_n^2)$
12. **end for**
13. (M-Step) Update Gaussian parameters μ_n and σ_n^2
    with sample means and variances of $\boldsymbol{A}^{(t+1)}$.
14. **end for**
15. **end for**
15. **return** $\boldsymbol{X}, \boldsymbol{A}$ (for each pixel)

Table 4.2. Model Coverage Range and Dimensions. Range indicates the physical coverage of the models, and Dimension denotes the number of pixels used for computation. 'LR', 'HR', and 'Fused' denote the low-resolution, high-resolution, and fused models, respectively.

	LR		**HR**		**Fused**	
	Range (km)	Dimension	Range (km)	Dimension	Range (km)	Dimension
2D Checkerboard	100×100	40×40	40×40	40×40	100×100	100×100
2D Ridgecrest	100×100	50×50	58×64	192×224	100×100	330×350
3D Ridgecrest	$100 \times 100 \times 5$	$50 \times 50 \times 11$	$54 \times 60 \times 5$	$108 \times 120 \times 21$	$100 \times 100 \times 5$	$200 \times 200 \times 21$

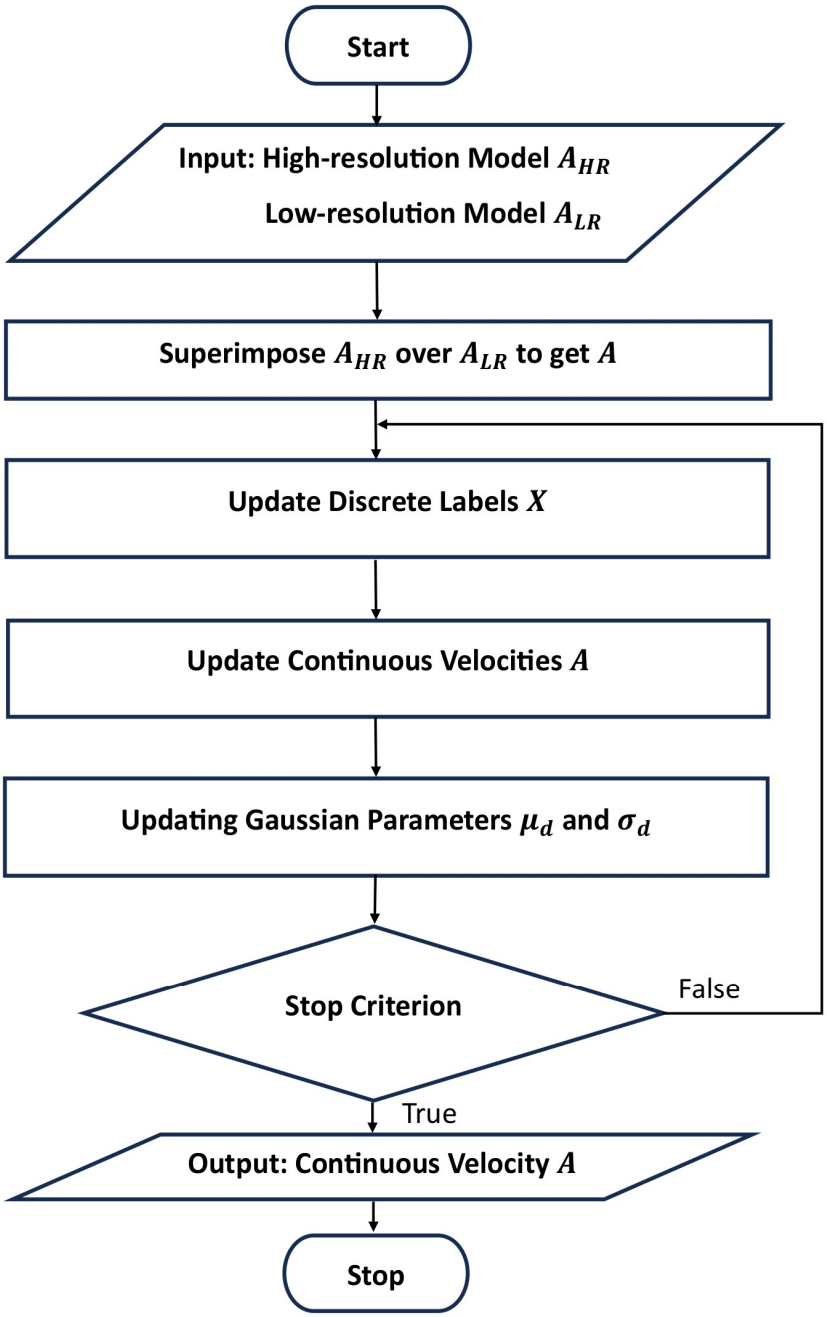

Figure 4.5. Pipeline of the iterative update policy for both pixel labels (discrete class labels) and pixel values (continuous velocity values).

4.5 Experiments

An aggressive smoothing policy removes sharp boundaries, while potentially important details are lost. On the other hand, gentle smoothing preserves details but leaves behind artificial boundaries between the LR and HR models. It is essential to achieve a trade-off between the two cases, this is the aim of PGM. To quantify this trade-off between presenting detailed information and minimizing artifacts we use the travel time between the stations and their residuals at 36 synthetic sensors (red Xs, 10 on each edge, see Fig. 1d on the border between the LR and HR areas). These residuals are then used to evaluate how much detailed information is preserved in the fused velocity model, compared to the HR maps.

4.5.1 Checkerboard Model

We used a 2D square checkerboard model with 100x100 pixels, each with 10 small squares along each edge, and each small square measuring 10x10 pixels in size. The pattern on the board alternates circular high- and low-velocity pixels in each small square.

4.5.2 Ridgecrest Fault Zone Model

To demonstrate the efficacy of the proposed PGM, we compare its performance with commonly used conventional methods (e.g., Gaussian filter and cosine taper window) on both the synthetic checkerboard model and the real-data Ridgecrest model. We have selected the high-resolution model of the Ridgecrest, CA, region obtained by ambient noise tomography to test the efficacy of our proposed PGM. The Ridgecrest fault zone image consists of a shallow (representing a depth of approximately 0.5 km) high-resolution Rayleigh wave model (Zhou et al., 2022), from which the S-wave velocity is roughly approximated by dividing by 0.9. This model reveals a 3D flower-shaped low-velocity zone surrounding the M7.1 and M6.4 earthquakes that ruptured in the 2019 Ridgecrest sequence. Te-Yang et al. [40] showed that including the fault zone model into the SCEC CVM-S V4.26M01 significantly improves the fit of simulations

to strong motion data from the M7.1 Ridgecrest earthquake, including at stations more than 200 km away in Los Angeles. The improvement in the fit to the data resulted from generating Love waves more accurately at the boundaries of the low-velocity zone around the faults, compared to the low-resolution model that lacked the fault zone model. Inspired by [40], who used the cosine-window taper fusion method by Ajala et al. [41], we compare the efficacy of our proposed PGM with other existing methods.

4.5.3 Comparison Methods

To evaluate the efficacy, a comparative evaluation was conducted, encompassing an array of smoothing techniques — Gaussian Smoothing Filter (GF) and Cosine Taper Smoothing (CT) — with machine learning-based approaches, Dictionary Learning (DL) and the newly proposed Probability Graphical Model (PGM). The GF and CT methods rely on predetermined parameters fixing the smoothing kernels, whereas DL and PGM feature adaptable parameters that optimize based on the data, suggesting a potential for enhanced adaptability.

The GF is a technique for noise reduction and data smoothing, utilizing a Gaussian kernel to weight data points by their distance from a central point [85]. The kernel size and standard deviation of the kernel control the level of smoothness, affecting the width and resolution of the curve. Specifically, the 1D GF kernel function is expressed as follows:

$$w_g(i) = \exp\left(-\frac{|i-\frac{L}{2}|^2}{2\sigma^2}\right), \tag{4.7}$$

where i and L denote the indices of the current position and the length of the kernel size, respectively, and σ represents the kernel's shape parameter, influencing the kernel's steepness.

CT smoothing reduces sharp patterns by applying a windowing function that reduces the amplitude in the boundary area, preserving the central portion [41, 86]. The cosine function shapes this taper, enhancing signal-to-noise ratio and retaining the characteristics of the original

models. A conventional 1D L-point CT window is defined by:

$$w_c(i) = \begin{cases} \frac{1}{2}\left\{1+\cos\left(\frac{2\pi}{r}[\frac{i}{L}-\frac{r}{2}]\right)\right\}, & 0 \le \frac{i}{L} < \frac{r}{2} \\ 1, & \frac{r}{2} \le \frac{i}{L} < 1-\frac{r}{2} \\ \frac{1}{2}\left\{1+\cos\left(\frac{2\pi}{r}[\frac{i}{L}-1+\frac{r}{2}]\right)\right\}, & 1-\frac{r}{2} \le \frac{i}{L} \le 1 \end{cases} \tag{4.8}$$

where i denotes the index within an L-point vector, and r signifies the fraction of the total window length occupied by the CT section, where r ranges from 0–1.

The combined velocity model $\boldsymbol{A}$ can be calculated as

$$\boldsymbol{A}(i) = w_c(i)\boldsymbol{A}_{\mathrm{HR}}(i) + (1-w_c(i))\boldsymbol{A}_{\mathrm{LR}}(i), \tag{4.9}$$

where $\boldsymbol{A}_{\mathrm{HR}}$ and $\boldsymbol{A}_{\mathrm{LR}}$ denote the high- and low-resolution velocity models. In the integration of methodologies employing fixed kernels, such as GF and CT, it is feasible to construct 2D and 3D kernels through the convolution of 1D kernels oriented in various directions. This process is accompanied by a normalization step, ensuring that the sum of all elements within the new kernel is equivalent to 1.

Dictionary learning eliminates Unaligned motifs and smooths the boundary areas by representing data as a sparse combination of learned basis vectors, or atoms, from the data itself. This iterative process, known as dictionary learning, aims to minimize the representation error [48, Eq. (12)]. The technique transforms data into a sparse format, which helps eliminate misaligned patterns while retaining essential features.

4.5.4 Optimal Parameter Selection

The number of clusters in Gaussian mixture model (GMM) clustering significantly impacts the results. Generally, the number of clusters can influence the complexity of the model and the interpretability of the results. More clusters result in a more complex model, which can

better capture intricate data structures and lead to more detailed insights into the data. However, it also increases the risk of overfitting.

Selecting the optimal number of clusters is crucial in GMM and other clustering techniques. Several methods can help determine an appropriate number of clusters, such as the Bayesian Information Criterion (BIC), and the silhouette score [91]. These methods balance the trade-off between the goodness of fit and the complexity of the model. In this experiment, we tested and compared the cluster number sequences $N = 3, 5, 6, 7, 9$, which are commonly used in practical applications of MRFs. Figure 4.6 (a) shows that the larger the number of clusters, the more detailed information is preserved in the HR models, and the larger the computation is required, implying a trade-off between computational cost and performance. Figure 4.6(b) demonstrates the number of clusters versus RMSE (root mean square error), which is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_t} (t_i - \hat{t}_i)}{N_t}}, \tag{4.10}$$

where N_t is the total number of ray paths, and t_i and $\hat{t}_i$ are the posterior and prior travel times (left vertical axis). The right vertical axis shows the computation time for the Ridgecrest test. RMSE decreases when the cluster number increases from 3–6, with a small decrease for larger clusters, and there is a rapid growth in run time for cluster numbers exceeding 6. For these reasons, we choose the cluster number as 6.

4.5.5 2D Performance

We demonstrate the fusion for the checkerboard and Ridgecrest velocity models described above, in both cases with directly superimposed HR and LR components, in Fig. 4.7 (a1, b1), both with an HR region in the center, surrounded by LR velocities. The dimensions of the models are summarized in Table 4.2.

Figure 4.7 (a2, b2) shows the checkerboard and Ridgecrest model station settings along with the ray-path density. For the checkerboard model, the stations are evenly distributed,

whereas the stations for the Ridgecrest model are highly irregular, reflecting the pattern used by Zhou et al. [64]. Fig. 4.7 (a3, b3) shows the label mask maps generated in the final iteration of the PGMs. In these maps, pixels sharing the same label suggest that the corresponding areas may exhibit comparable velocity patterns, implying that they are likely sampled from a similar distribution. In 2D experiments, we empirically selected the Gaussian filter kernel size as $(5,5)$ for south-north and west-east directions and the σ value as 1.5 [92, Chapter 6.9]. As for cosine-taper smoothing, we designed two 1D cosine-taper windows with parameters L and r (in Eq. 4.8) for analyzing south-north and west-east directions set to $(40,40)$ and $(0.5,0.5)$, respectively, in a checkerboard pattern, and $(50,50)$ and $(0.75,0.75)$, respectively, for the 2D Ridgecrest model. This design involved convolving two 1D windows to formulate a 2D Cosine-Taper window. The selection of the cosine-taper ratio parameter r was based on achieving the lowest RMSE among all tested combinations $\{(0.25,0.25),(0.5,0.5),(0.75,0.75)\}$. The setting of the dictionary learning adheres closely to the methodology described by Yang et al. [80], with the only modification being the adjustment of the kernel size to 7×7. The smoothed fusion results are summarized in Fig. 4.7 (a4-a6) and (b4-b6).

The outcomes suggest that the learning-based methods, e.g. DL and PGM, demonstrate the capacity to retain detailed information in comparison to the application of the direct Gaussian filter and cosine-taper smoothing. This enhanced performance is attributed to the adaptive nature of these learning methods in determining the optimal fusion parameters for accurate data representation. Conversely, Gaussian smoothing and cosine taper use a fixed kernel to blend neighboring pixels, which does not allow for such adaptive optimization and may lead to a less detailed final output.

We evaluate the efficacy of our model fusion with multiple metrics: travel time (RMSE, which measures the travel time deviations [23], calculated based on the observations from actual seismic stations in Fig. 4.7), naturalness image quality evaluator, NIQE, a commonly-used measurement for image quality [81], peak signal-to-noise ratio, PSNR, measuring the sharpness of images [82], and the Fréchet inception distance, FID, capturing similarities between the

original and fused models [83]) in Table 4.3. In the checkerboard test, due to the simplicity of the pattern and the uniform distribution of stations, all learning methods exhibit a comparable performance. In contrast, for the more complex and realistic Ridgecrest model, the PGMs outperform the DL model, as the latter is sensitive to the orientation of the patches while the graphical models are rotationally invariant.

Geological formations are often anisotropic, meaning that their properties vary depending on the direction in which they are measured, e.g., laterally continuous and vertically stratified. Standard Markov random field (MRF) schemes, which assume homogeneous properties (same properties in all directions), can lead to errors when applied to such formations. PGMs, on the other hand, consider the anisotropic nature of geological formations, generally leading to more accurate results. Seismic inversion is an ill-posed problem, meaning that it does not have a unique solution, and small changes in the input can lead to large changes in the output. In such cases, regularization is a technique used to stabilize the solution. Our PGM provides an edge-preserving regularization based on the information from neighboring pixels, effectively reconstructing subsurface models.

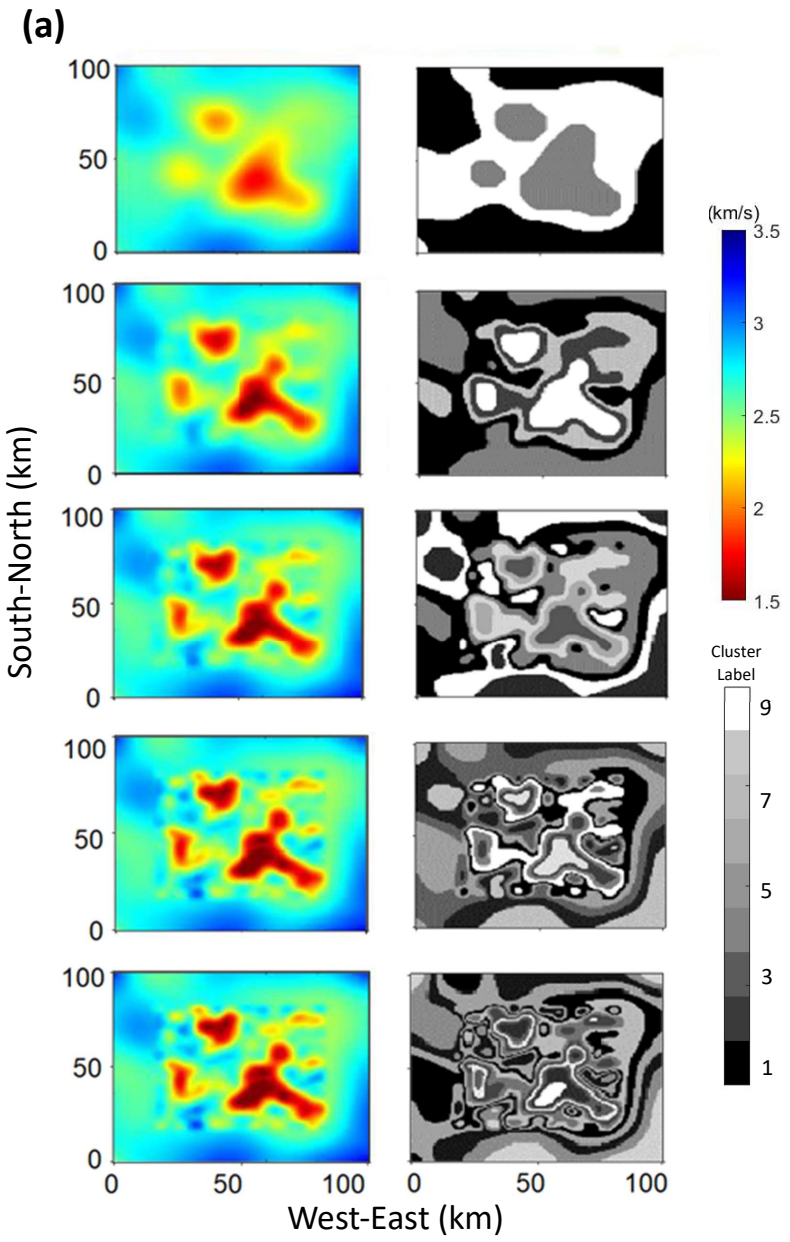

Figure 4.6. (a) Comparison of the fusion results using cluster numbers n= 3, 5, 6, 7, and 9. (left) fused velocity model and (right) cluster distribution. (b) The number of clusters versus RMSE error (left vertical axis, corresponding to the solid line) and run time (right vertical axis, corresponding to the dashed line) for the Ridgecrest model. Generally, the larger the cluster number, the smaller the RMSE error with a longer run time.

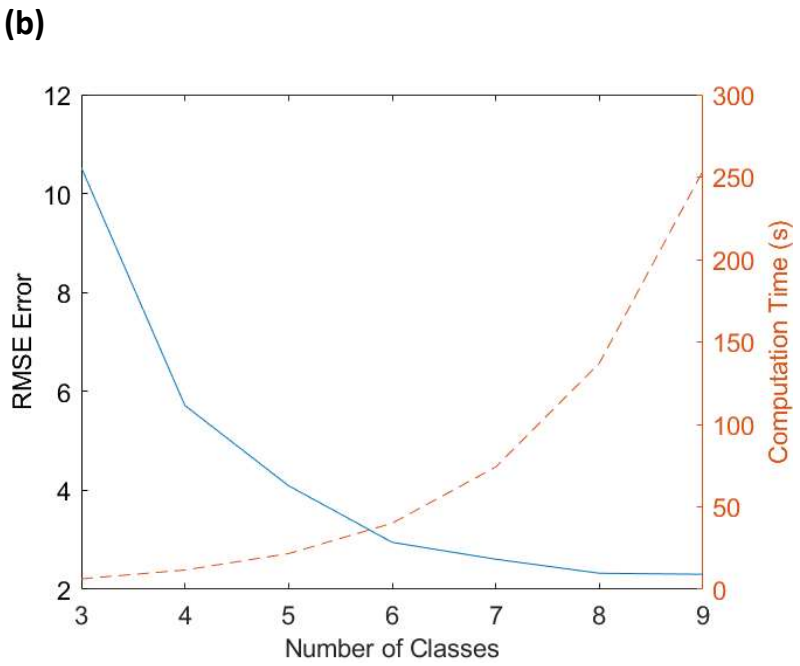

Figure 4.6. (a) Comparison of the fusion results using cluster numbers n= 3, 5, 6, 7, and 9. (left) fused velocity model and (right) cluster distribution. (b) The number of clusters versus RMSE error (left vertical axis, corresponding to the solid line) and run time (right vertical axis, corresponding to the dashed line) for the Ridgecrest model. Generally, the larger the cluster number, the smaller the RMSE error with a longer run time (Continued).

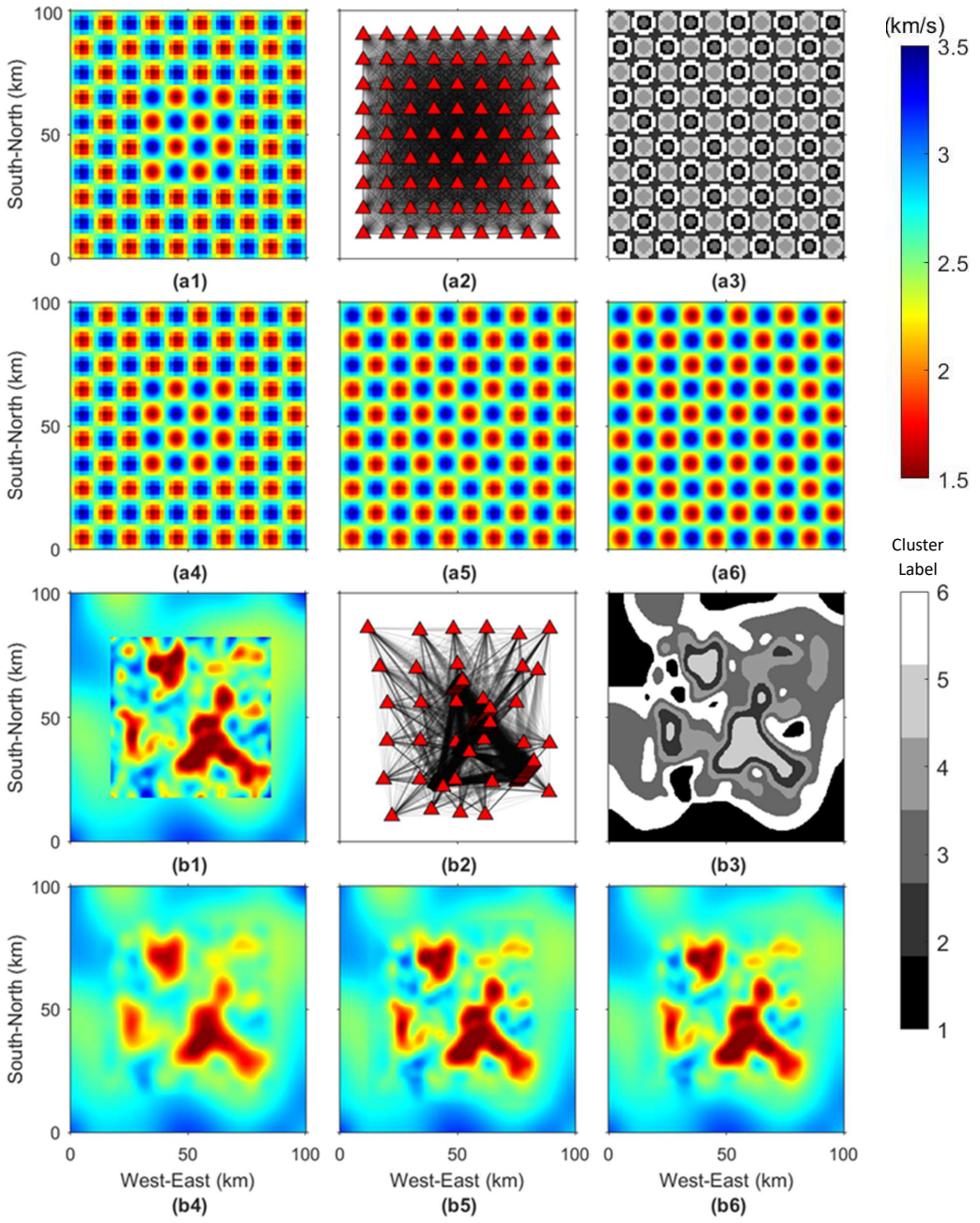

Figure 4.7. (top 2 rows) Checkerboard and (bottom 2 rows) Ridgecrest models. (1) Superimposed HR and LR models. (2) Station location and ray density. (3) 6-class label mask maps for HR models (pixels with the same label are learned together). (4) Smoothing results with a 5×5 Gaussian filter (GF). (5) and (6) Fusion results with dictionary learning (DL) and with our PGM, respectively.

4.5.6 3D Ridgecrest Model Fusion Comparison

To assess the proficiency of the PGM fusion approach, we have expanded our methodological framework to the integration of 3D models. Analogous to the 2D fusion experiments, the S-wave velocity model was extracted from the top 5 km around the 2019 Ridgecrest, CA, earthquake sequence, from the SCEC CVM-S4.26, serving as the LR model, while the 3D S-wave velocity model derived from surface wave dispersion inversion by [64] represents the HR model. Both LR and HR models were interpolated into horizontal slices with 100×100 pixels for each specified depth and resampled to a depth resolution of 250 meters.

In Fig. 4.8, the LR CVM-S4.26 model centered on the Ridgecrest domain is shown in panel (a), while panel (b) depicts the model obtained by directly incorporating the HR model (from surface wave dispersion inversion in [64]) into the LR matrix. We employ the cosine-taper smoothing technique and 3D dictionary learning (with kernel size $7 \times 7 \times 5$) as benchmark methodologies. Similar to the 2D experiments, we define three 1D cosine-taper windows (see Eq. (4.8)) with lengths L $(108, 120, 21)$ along south-north, west-east, and depth, respectively, and convolved the three 1D windows to generate a 3D cosine-taper window. The lowest RMSE was found for cosine ratios r of 0.75 along south-north (from test values 0.25, 0.5, and 0.75), 0.75 along west-east (from test values 0.25, 0.5, and 0.75), and 0.9 along depth (from test values 0.1, 0.3, 0.5, 0.7, and 0.9), respectively. The resultant fusion models via benchmark methods and PGM are shown in Fig. 4.8 panels (c)-(e), respectively. The cosine taper functions operate exclusively in the regions that overlap in the HR and LR data. For this reason, when there is a significant mismatch in the boundary areas, the cosine-taper smoothing may not fully correct misaligned patterns (Fig. 4.8(c)). However, the machine learning-based methods (dictionary learning and PGM) operate both at overlapping regions and areas with only LR information, which enhances their effectiveness in successfully aligning unmatched patterns from both sides. Note that the 3D PGM retains enhanced details from the HR models compared to the 3D dictionary learning procedure.

To quantify the performance in 3D, evaluation points were defined between the LR and HR models at each depth level (similar to 4.1d in the 2D case), where the travel times were computed before and after the fusion methodologies were applied on the HR and LR directly-superimposed model. Using the travel time preceding the fusion as a reference, we calculated the RMSEs corresponding to the post-fusion travel time deviation. The calculated RMSEs of travel time deviation for cosine taper smoothing, 3D dictionary learning fusion, and our 3D PGM for depths from the surface to 5 km are listed in Table 4.4 (with the evaluation points shown in Fig. 4.1). For the 2D fusion case, the RMSE from dictionary learning and PGM surpass those obtained through cosine taper smoothing. Our 3D PGM approach yields the best improvement, with an average 44% reduction in travel-time deviation relative to conventional cosine-taper methods. This indicates a minimal distortion of information from the HR model, as the 3D PGM preserves data well.

Six dense sensor arrays were deployed across the faults ruptured in the 2019 Ridgecrest earthquake sequence (see Fig. 4.9, left panel, A1, A2, B1 through B4). Owing to these densely distributed arrays, we computed surface wave dispersion inversion profiles for station pairs. Subsequently, we aggregated them to derive HR 2D vertical S-wave velocity models [64], as illustrated in Fig. 4.9 (top right). These derived models are compared with vertical cross-sections extracted from 3D models and combined with the LR background model (from the SCEC CVM-4.26) through various fusion methodologies. For instance, the B2 and B4 array panels (Fig. 4.9c, d) depict the 2D cross-sections extracted from the 3D DL fusion model and the 3D PGM. The performance of our 3D PGM approach is evident in its ability to define and preserve the accuracy of the boundary of the low-velocity zone more precisely. This improved accuracy can be attributed to the PGM's strategy of assigning differential weights to edges, which are oriented in various directions. In contrast, the efficacy of 3D DL is somewhat limited due to its inherent rotational invariance and the constraints imposed by a fixed patch dimension.

Table 4.3. 2D Evaluation Results. Evaluation metrics are root-mean-square error (RMSE) of the travel time deviation (with unit s), naturalness image quality evaluator (NIQE), peak signal-to-noise ratio (PSNR), and Fréchet inception distance (FID). ↓ indicates smaller is better, and ↑ opposite. GF: Gaussian filtering; CT: cosine tapering; DL: dictionary learning; PGM: probability graph model.

	RMSE/s↓	**NIQE↓**	**PSNR/dB↑**	**FID↓**
Checkerboard GF	1.65	7.68	14.58	45.75
” CT	1.32	5.87	15.27	36.49
” DL	1.18	5.44	15.70	33.85
” PGM	1.14	5.40	16.14	32.49
Ridgecrest GF	3.52	12.41	21.80	61.39
” CT	3.14	8.73	22.07	56.97
” DL	2.61	7.29	22.36	54.25
” PGM	2.27	6.70	23.04	47.49

Table 4.4. 3D Evaluation Results. Evaluation metric is the root-mean-square error (RMSE) of the travel time deviations (with unit s). CT: cosine taper; DL: dictionary learning; PGM: probability graph method.

Depth	**CT**	**DL**	**PGM**
0 km	1.67	1.57	0.86
1 km	1.52	1.39	0.73
2 km	1.73	1.38	1.04
3 km	1.58	1.53	0.96
4 km	1.57	1.62	1.13
5 km	1.79	1.43	1.04

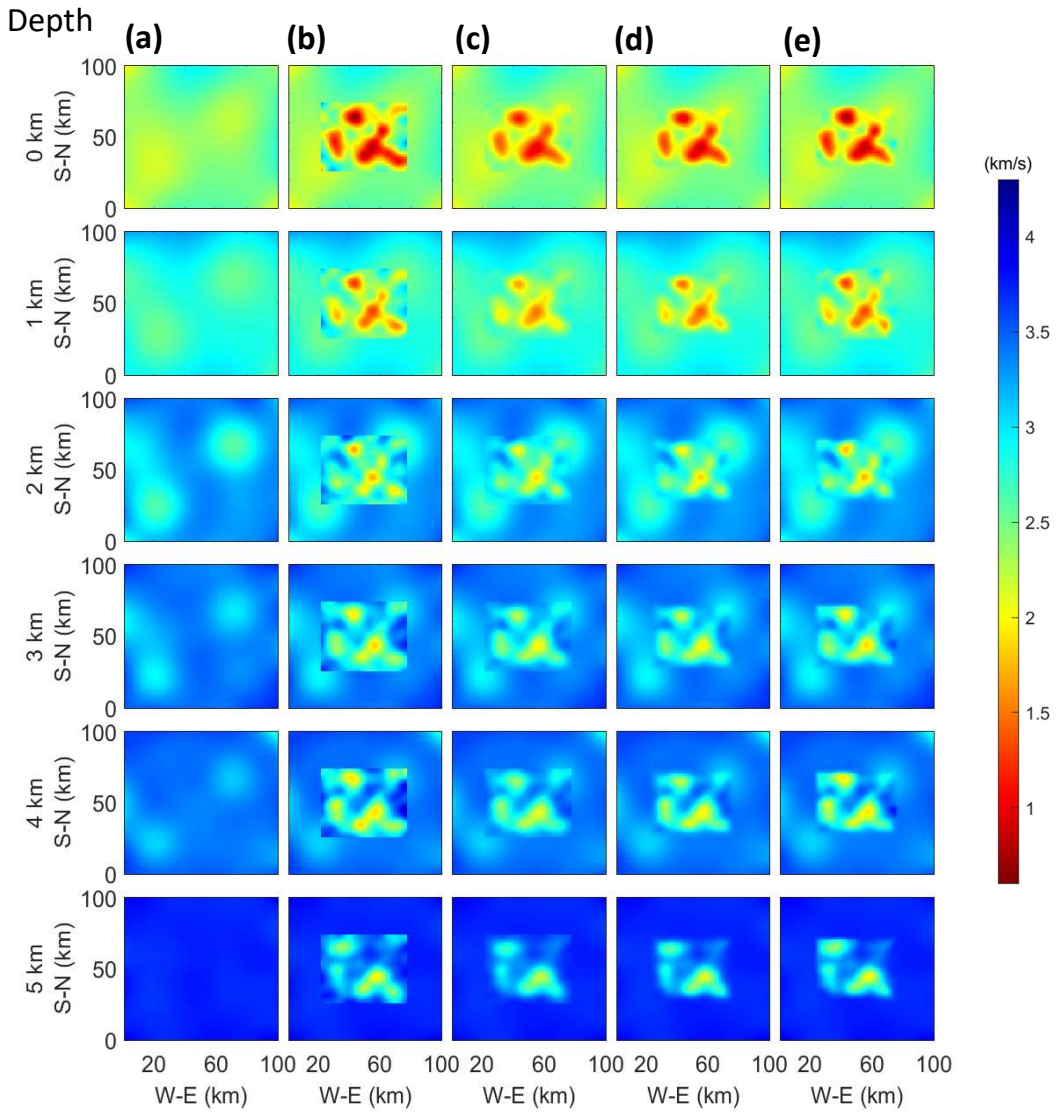

Figure 4.8. (a) The LR model extracted from CVM-S4.26 around the Ridgecrest area. (b) Direct superposition of the 3D HR surface wave dispersion inversion model and CVM LR models for the Ridgecrest area. (c-e) Combined LR and HR models, smoothed by (c) cosine-taper function, (d) dictionary learning, and (e) PGM.

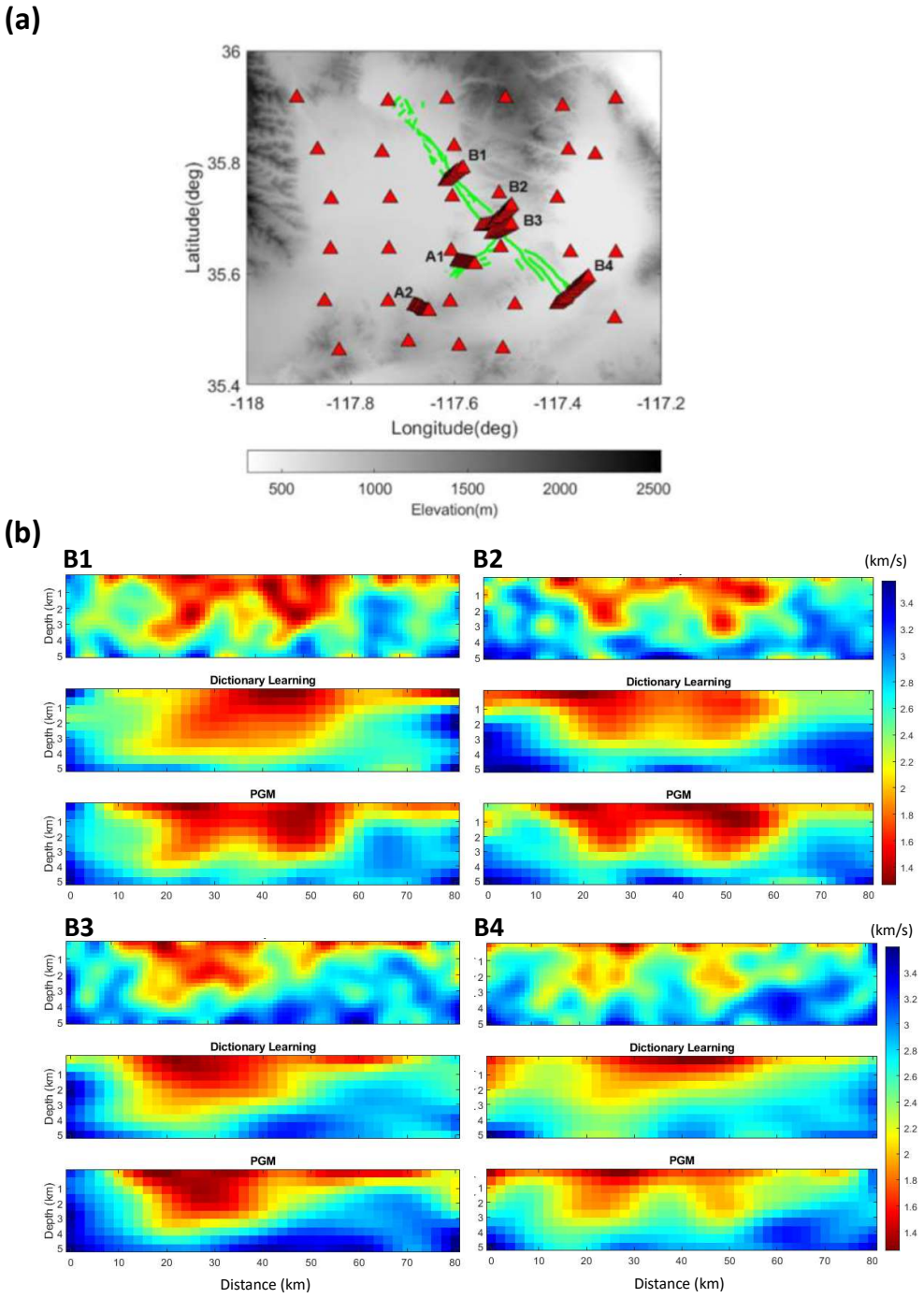

Figure 4.9. (a) Station locations (triangles) and main faults (lines) surrounding the Ridgecrest area. There are six dense sensor arrays across the main faults (A1-2 and B1-4). (b) Vertical cross-sections of the shear wave velocity along the B1-4 station arrays from (top) surface wave dispersion inversion, (center) the 3D fusion model from dictionary learning, and (bottom) the PGM.

4.6 Conclusions

We present a method for combining multiresolution seismic velocity maps usingPGMs. The performance of the PGM algorithm is evaluated through experiments, using both a checkerboard model and a complex fault zone model around the 2019 Ridgecrest earthquake sequence. The checkerboard model is characterized by inherent simplicity and uniform station distribution, the evaluation demonstrates that the PGM approach outperforms all tested established baseline techniques. The PGM outperformed traditional methods, potentially as a result of adaptive parameter learning. In the context of the Ridgecrest model, the PGM technique produces a 44% reduction in the computed travel-time residuals versus the conventional Gaussian and cosine-taper smoothing methods in 3D exploration models. This is due to the limitations of traditional methods in addressing anisotropic patterns, in contrast to the PGM which learns weights consistent with the complex structure of the Ridgecrest model. In summary, the PGM approach effectively minimizes the undesired sharp discontinuities often observed between LR and HR models, while preserving detailed information inherent in the HR models. A prospective area of investigation in future work involves addressing the challenge of irregular model resolution within the HR domain, which is crucial for enhancing the fidelity and applicability of our models and potentially improves the understanding and application in various real-world models. Finally, we recommend that the efficacy of the PGM be tested directly through a comparison of synthetic and observed waveforms.

4.7 Acknowledgments

The text of Chapter 4, in full, is a reprint of the material as it appears in Zhou, Z., Gerstoft, P. and Olsen, K.B. “3D multiresolution velocity model fusion with probability graphical models”, *Bulletin of the Seismological Society of America*, 114(3), pp.1279-1292, 2024. The dissertation author was the primary researcher and author of Chapter 4. The coauthors listed in this publication directed and supervised the research.

Chapter 5

Mixture of Expert Models for Enhanced Predictive Analytics in Fiber Health Monitoring

5.1 Introduction

Monitoring blood pressure, an indicator of cardiovascular health, is crucial for managing hypertension and enhancing healthcare efficiency. Traditional methods often rely on cumbersome inflatable cuffs and electronic monitors, which, while widespread, can be influenced by the cuff's presence [93]. Innovations in medical IoT, such as cuffless measurement techniques using wearable sensors like e-skin [94], electrocardiography (ECG) [95], and photoplethysmography (PPG) [96, 97], offer more seamless solutions. However, these technologies face challenges including susceptibility to electromagnetic interference, high power demands, and sensitivity to individual differences (e.g., skin conditions and body shapes), which can limit their effectiveness in complex scenarios [98].

Traditional monitoring techniques depend on single predictive models, which might not capture the complex behaviors and failure mechanisms. Mixture of Experts (MoE) models offers a solution by integrating multiple specialized expert systems. These models utilize a gating mechanism to dynamically select the most relevant expert based on input features, enhancing the understanding and prediction of different failure modes [99, 100]. We apply MoE models

to fiber health monitoring, potentially improving the accuracy and robustness of predictions compared to traditional methods, and offering a more comprehensive analysis of the health status of fiber-reinforced composites, thereby enhancing predictive performance and adaptability [101].

The in-fiber modal interferometer-based sensor was recently enhanced by a fiber optic micro-cavity sensor with increased sensitivity and compactness, enabling continuous blood pressure monitoring [102, 103]. We have developed a system that uses a flexible fiber optic sensor with a micro-open cavity Mach-Zehnder interferometer (MZI) [104] to capture the pulse waveform from the radial artery at the wrist accurately. This system is paired with a cloud-deployed machine learning model that processes denoised pulse features and wave signals, considering the user's physical characteristics to provide accurate blood pressure readings and real-time feedback. Combining with MoE learning models, our IoT system offers an accurate and adaptable solution for continuous blood pressure monitoring.

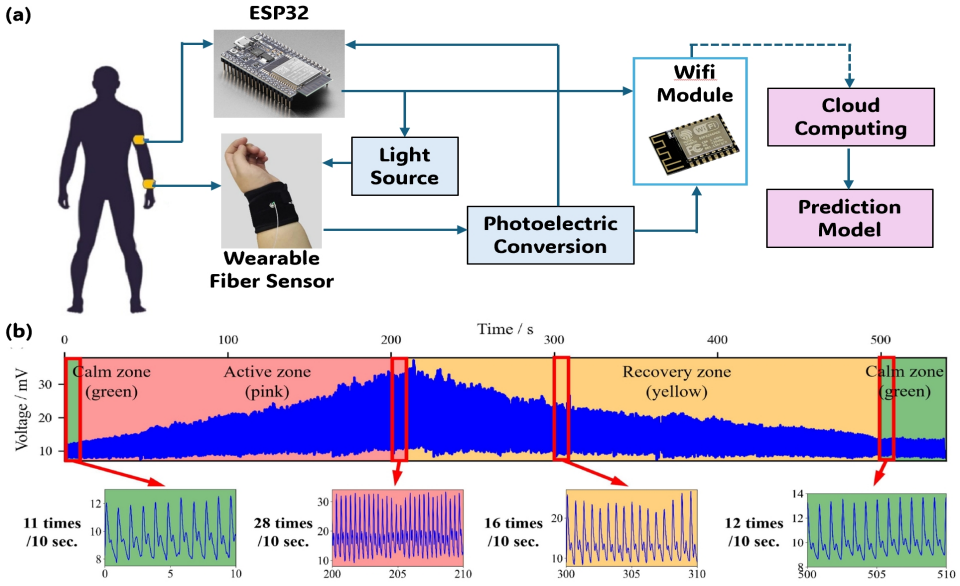

Figure 5.1. (a) Schematic of the IoT blood pressure monitoring system, and (b) dynamic blood pressure measurement data of one participant, demonstrating a calm–active–recovery–calm cycle.

5.2 Data Collection and Feature Extraction

Fig. 5.1(a) shows the architecture of our IoT-based blood pressure monitoring system. At the heart of the system lies the tw-MOC fiber optic sensor, which is optically linked to a control source. This control source, powered by an Espressif Systems ESP32 microcontroller [18], integrates a laser diode for signal emission, a photodetection module for signal reception, and a WiFi module for wireless communication. The WiFi module is essential for the seamless transmission of data to the cloud, where a sophisticated machine-learning model processes the incoming signals for blood pressure estimation. The processed results are transmitted back to a local display terminal, enabling continuous real-time monitoring and analysis of the patient's cardiovascular health.

To validate the effectiveness, we extended its application to assess dynamic blood pressure in active scenarios, including exercise. A total of 17 healthy volunteers were monitored through phases of rest, three minutes of stationary cycling, and recovery, with results segmented into calm,

active, recovery, and calm phases (Fig. 5.1(b)). The sensor maintained accurate pulse waveform readings without distortion throughout the test, showcasing its effectiveness for real-time blood pressure monitoring under varying usage scenarios. A dataset of approximately 60,000 real samples is used to train and evaluate our blood pressure prediction model.

The raw pulse data captured by the sensor undergoes conversion via a photo-predictor and is transmitted to the cloud through WiFi, where it is preprocessed and features are extracted before entering the blood pressure estimation model. Signal preprocessing includes using a wavelet transform (5-level decomposition with the Symlet-7 wavelet function [105]) to eliminate baseline drift (affected by temperature and the movement of the wearable device) and high-frequency noise above 15 Hz [106], effectively isolating the signal's main characteristics and normalizing it (Fig. 5.2(b)). The processed wave signal is then subjected to feature extraction to obtain some commonly used features for blood pressure prediction [107] (Fig. 5.2(c)). These three features are concatenated and input into the blood pressure estimation model (Fig. 5.3).

5.3 MoE Methodology

We proposed the methodology employed in integrating Mixture of Experts (MoE) models for enhanced predictive analytics in fiber health monitoring. We specifically focus on the dense [108] and sparse [109] MoE configurations to leverage their distinct capabilities in managing complex data characteristics associated with fiber composites.

5.3.1 Dense MoE Models

Given an input signal x, our model consists a set of $\boldsymbol{M}$ predictors $\{f_1, \ldots, f_M\}$ and a linear gating network h. Denote the parameters of the gating network as $\boldsymbol{\Theta} = [\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_M}] \in \mathbb{R}^{d \times M}$, the output of the gating network is $h(\boldsymbol{x}; \boldsymbol{\Theta})$, where d is the dimension of the embedded features of x [15]. Denote the output of the m-th predictor as $f_m(\boldsymbol{x}; \boldsymbol{W})$ with input x and parameter $\boldsymbol{W}$.

The route gate value for mth predictor is:

$$\pi_m(\boldsymbol{x}; \boldsymbol{\Theta}) = \frac{\exp(h_m(\boldsymbol{x}; \boldsymbol{\Theta}))}{\sum_{m'=1}^{M} \exp(h_{m'}(\boldsymbol{x}; \boldsymbol{\Theta}))}, \forall m \in [M], \tag{5.1}$$

and the output of MoE is given by:

$$F(\boldsymbol{x}; \boldsymbol{\Theta}, \boldsymbol{W}) = \sum_{m \in \mathscr{M}} \pi_m(\boldsymbol{x}; \boldsymbol{\Theta}) f_m(\boldsymbol{x}; \boldsymbol{W}). \tag{5.2}$$

This methodology ensures each aspect of the input data is thoroughly analyzed by every expert, making dense MoE models effective in scenarios where comprehensive data integration is crucial for accurate predictions [108].

5.3.2 Sparse MoE Models

Conversely, sparse MoE models activate only a subset of experts based on the input characteristics, significantly reducing computational overhead while still maintaining robust prediction capabilities [110]. The sparse MoE layer selects the top-k experts for activation,

defined as:

$$\pi(x;\Theta)_i = \text{softmax}(\text{TopK}(g(x;\Theta) + R_{\text{noise}}, k))_i, \tag{5.3}$$

where the TopK operation retains only the top-k gating values, setting others to $-\infty$, which after softmax, effectively zeroes them. The noise term R_{noise} is introduced to promote diversity and stability within the training process.

Integration of sparse MoE into predictive analytics for fiber health monitoring allows for efficient resource allocation, focusing computational efforts on the most relevant experts per input instance. This approach is particularly beneficial in real-time monitoring scenarios, where computational efficiency is as critical as predictive accuracy.

5.3.3 Model Implementation

To develop a method adapted to different predictors, we adopt a two-stage training strategy. First, we train predictors separately on each corpus. For the m-th predictor, the corresponding loss is

$$\begin{aligned} l_m &= -\sum_{c=1}^{C} \log \frac{\exp(f_{m,c}(\mathbf{x};\mathbf{W}))}{\sum_{c'=1}^{C} \exp(f_{m,c'}(\mathbf{x};\mathbf{W}))} y_{n,c}, \\ \mathscr{L}_m &= \frac{1}{N} \sum_{n=1}^{N} l_m, \end{aligned} \tag{5.4}$$

where C denotes the maximum number of variables that need to be predicted, N denotes the number of samples, and $y_{n,c}$ denotes the target value of n-th sample on c-th variable. In this stage, the parameters $\boldsymbol{\Theta}$ of the gating network are frozen.

Both dense and sparse MoE models are integrated into a unified framework as shown in Fig. 5.3 (b). This framework includes the dynamic selection parameter $\mathbf{W}$, which assesses the reliability of each expert model based on the complexity of the input data and the performance of each expert. This hybrid method optimizes performance across various expert models, enhancing the adaptability and robustness of the system for diverse data types in fiber health monitoring.

5.4 Experiments

The complexity of physiological factors and the multidimensional nature of pulse waveforms often challenge the efficacy of traditional linear methods in accurately predicting blood pressure – systolic blood pressure (SBP) and diastolic blood pressure (DBP) [111]. In response, we have developed a deep learning approach that models input features more effectively, accommodating individual variances and enhancing prediction accuracy. Our advanced hybrid model combines Convolutional Neural Networks (CNNs) [112], Long Short-Term Memory (LSTM) networks [113], and MoE Gating Networks. This integration allows for the processing of a diverse array of inputs, including normalized pulse signal, personal characteristics (PC) (e.g., age, height, body mass index (BMI), and sleep duration), and detailed pulse waveform features.

We aimed to evaluate how MoE models can adaptively assign weights, reflecting the confidence in each expert model. We selected 10,000 samples from participants aged 18-22, whose pulse signal patterns are simpler and clearer, and followed the conventional approach (Fig. 5.3 (b)), testing both dense and sparse MoE configurations. We found that the dense MoE model, which adaptively combines features from models with one to five CNN blocks, outperforms any single-branch CNN-block model. We selected a sparsity factor $k = 4$ using 5-fold cross-validation. The sparse MoE model assigned a weight of zero to the branch with five CNN blocks, aligning with the findings from conventional model selection, and demonstrating the best prediction errors and standard deviation.

We then fixed the weight for the CNN blocks 1 to 2 branch model, which was pre-trained on data from young individuals, and integrated a model [114] pre-trained on mid-aged samples. Additionally, we considered conventional PPG+ECG methods in our comparative experiments. All models were evaluated using 60,000 samples collected from participants aged 18 to 45. The results, presented in Table 2, indicate that models trained solely on young or mid-aged samples performed poorly as they lacked information across all age groups. The conventional methods demonstrated some robustness. Overall, our MoE-based method was able to adaptively assign

Table 5.1. Performance Comparison: Baseline vs. Proposed MoE Method [Mean Error (ME) ± Standard Deviation (STD)].

The number of CNN blocks	SBP(mmHg) ME ± STD	DBP(mmHg) ME ± STD
1	1.84 ± 2.01	-1.37 ± 1.69
2	0.28 ± 1.97	-0.06 ± 1.45
3	1.07 ± 2.26	-1.95 ± 1.74
4	-1.38 ± 2.51	-0.54 ± 1.85
5	-2.72 ± 2.83	-2.96 ± 2.33
Dense MoE 1–5	0.25 ± 1.81	0.05 ± 1.37
Sparse MoE 1–5	$\mathbf{0.23 \pm 1.74}$	$\mathbf{0.05 \pm 1.32}$

Table 5.2. Performance Comparison: Baseline vs. Proposed MoE Method [Mean Error (ME) ± Standard Deviation (STD)].

Prediction Models	SBP(mmHg) ME ± STD	DBP(mmHg) ME ± STD
PPG and ECG	1.75 ± 3.64	-1.93 ± 3.69
CNN Blocks 1-2	3.48 ± 4.16	-4.69 ± 3.52
Pre-trained CNN (PTCNN) [114]	-3.75 ± 5.49	4.14 ± 4.68
Dense MoE 1–5 + PTCNN	0.74 ± 2.81	0.95 ± 2.37
Sparse MoE 1–5 + PTCNN	0.87 ± 1.84	0.76 ± 2.13

weights to different models and achieve superior overall performance.

5.5 Conclusions

This paper presents a robust medical IoT system for human blood pressure monitoring, featuring a novel, highly flexible fiber optic sensor designed for precision and durability. Located in the cloud, our CNN-LSTM-MoE-based model processes multimodal pulse data to deliver accurate blood pressure readings, minimizing computational load on wearable devices. This system effectively handles simultaneous real-time monitoring of multiple subjects, demonstrating significant potential for advanced healthcare applications.

5.6 Acknowledgment

The text of Chapter 5, in full, is a reprint of the material as it will appear in an under-review paper Zhou, Z., Zhang, Z., Li, J., and Gerstoft, P. "Integrating Mixture of Expert Models for Enhanced Predictive Analytics in Fiber Health Monitoring", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* , 2025. The dissertation author was the primary researcher and author of Chapter 5. The coauthors listed in this publication directed and supervised the research.

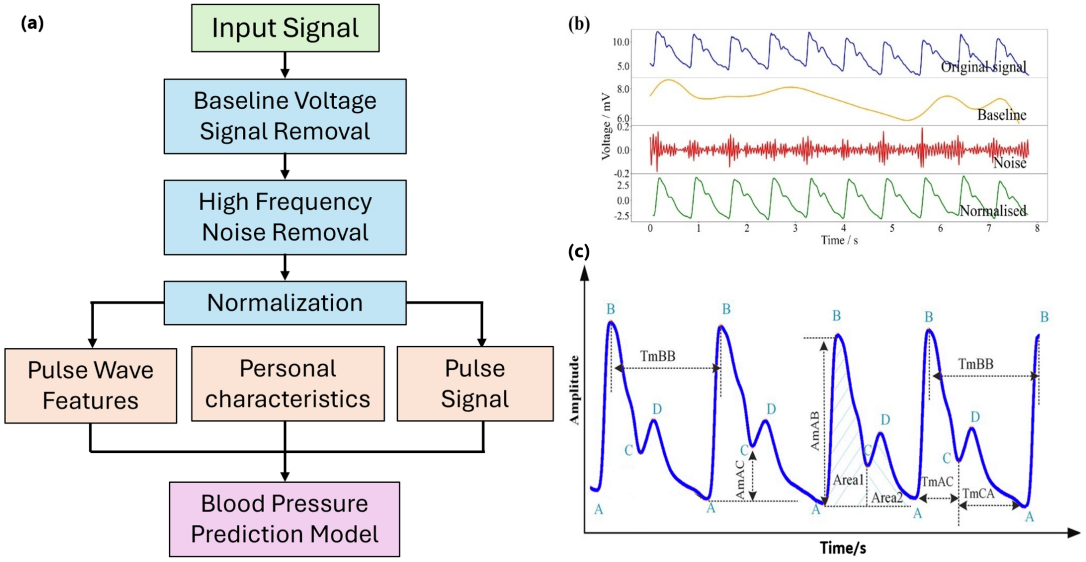

Figure 5.2. (a) A flowchart depicting the process of data processing and feature extraction. The normalized pulse wave signals are concatenated with pulse wave features and personal characteristic features. (b) Real data processing scenario using a segment of actual signals as an example; (c) Typical pulse wave features extracted from the normalized signals.

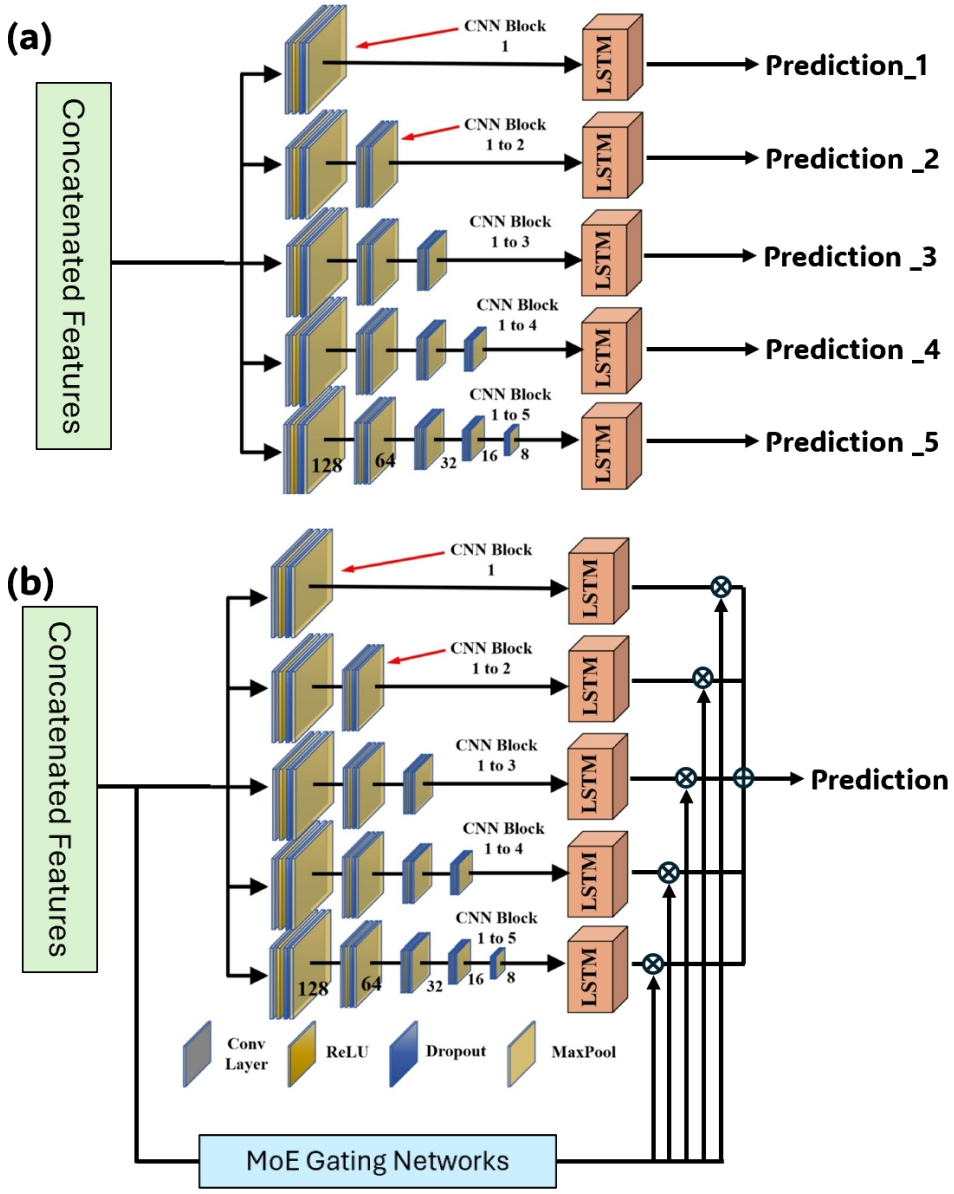

Figure 5.3. (a) The conventional approach to determining the optimal number of CNN blocks in the CNN-LSTM module involves using only wave signals as inputs to the hybrid mode. Each CNN-LSTM module operates independently and its performance is evaluated manually. (b) MoE gating networks are introduced to automatically assign a weight (confidence score) to each module, culminating in a single prediction. Here, $\otimes$ represents multiplication, and $\oplus$ indicates the summing operation.

Chapter 6

Conclusion

In this dissertation, our studies collectively represent significant advancements in the fields of seismic velocity modeling, fault zone imaging, and data fusion. Each contributes unique insights and methodologies aimed at improving our understanding and modeling of Earth's subsurface, with applications that extend from seismic hazard assessment to broader geophysical analyses. Together, they highlight the importance of integrating high-resolution data with large-scale models, addressing challenges such as irregular spatial resolutions, anisotropy, and the need for computationally efficient approaches.

The first study focuses on high-resolution imaging of fault zones, particularly the Ridgecrest fault system, revealing detailed low-velocity zones and their spatial characteristics. By integrating advanced imaging techniques, it not only improves the understanding of fault damage zones but also underscores the need to study their long-term evolution and implications for seismic hazards. The findings pave the way for future research to validate these models through ground motion simulations and extend the methodology to other fault systems worldwide.

The second and third papers emphasize the critical need for effective data fusion techniques to combine multiresolution seismic velocity models. The proposed Probability Graphical Models (PGMs) and graph-learning approaches offer powerful tools to address the challenges of integrating high-resolution (HR) and low-resolution (LR) data, reducing sharp discontinuities and preserving essential details. These methods have demonstrated substantial improvements

in travel-time residuals and have shown potential for broader applications, including real-time model updates and extension to diverse geophysical data types. Future work will focus on refining these models further, improving their computational efficiency, and validating them with observed seismic data.

In conclusion, these studies collectively advance the state of the art in geophysical modeling by addressing key challenges in resolution, data fusion, and computational efficiency. Their findings underscore the importance of adaptive methodologies that preserve critical details while ensuring smooth transitions in multiresolution data. By integrating high-resolution models with larger-scale frameworks, these approaches offer valuable insights into the structure and dynamics of Earth's subsurface. Continued research in this direction, including the validation and application of these models to diverse geological settings, promises to enhance our understanding of fault zones, seismic hazards, and broader geophysical phenomena.

We summarize the three papers within this dissertation and discuss our future work as below.

6.1 ANT with Locally Sparse Control Improves the Detailed Features

This study employs ambient noise tomography (ANT) enhanced with locally sparse tomography (LST) to provide detailed imaging of low-velocity zones (LVZs) surrounding the faults of the 2019 Ridgecrest earthquake sequence. The research reveals flower-shaped LVZs up to 5 km wide and a 40% velocity contrast with the surrounding host rock. These findings strongly suggest that the LVZs represent fault damage zones resulting from the Ridgecrest earthquakes.

Furthermore, the study identifies LVZs associated with faults that have not experienced recent activity, indicating the potential for long-lasting damage zones that may persist for thousands of years. This insight is supported by comparisons to other seismic zones such as the San Jacinto Fault and Calico Fault. The LST methodology offers improved resolution and

uncertainty quantification over conventional tomography, highlighting its potential for broader seismic imaging applications. These results advance the understanding of fault zone structures and their evolution, which is essential for seismic hazard assessment and improving models of fault dynamics.

6.2 PGM with Physical Informed Information for 2D and 3D models

Our study presents a novel Probabilistic Graphical Model (PGM) for fusing multiresolution seismic velocity maps. The approach balances the need to preserve detailed information in high-resolution (HR) regions while smoothing undesired sharp boundaries between HR and low-resolution (LR) areas. Tests using both synthetic and real-world data, including a Ridgecrest fault zone model, demonstrate that PIPGM outperforms conventional methods like Gaussian kernel smoothing and dictionary learning in reducing travel time residuals and maintaining model fidelity. By adapting weights based on physical parameters such as ray-path density and gradient information, the physics-informed PGM achieves a 38% reduction in residuals compared to Gaussian smoothing in 2D model fusion. The PGM method significantly reduces travel-time residuals compared to conventional methods like Gaussian smoothing and cosine-taper in 3D exploration models. Our proposed PGM achieves a 44% reduction in travel-time residuals. The results highlight the model's ability to adapt to geological complexities, making it a robust tool for seismic velocity model integration.

The methodology offers a significant advancement in integrating seismic velocity models of varying resolutions, paving the way for improved seismic hazard analysis and more accurate ground motion simulations. Future work includes addressing challenges related to irregular resolutions in HR domains and validating the approach through comparisons of synthetic and observed waveforms.

6.3 Future Work

Future work will be focused on advancing seismic velocity modeling, fault zone imaging, and data fusion techniques. One major area of focus is enhancing the resolution and adaptability of models. This includes addressing irregularities in high-resolution domains, developing methods to improve the spatial resolution of low-velocity zones, and refining the ability to model anisotropic features in complex geological settings. These advancements would not only improve the fidelity of seismic models but also ensure their applicability across diverse fault zones and geological structures globally.

Another significant direction involves validation and integration. Future studies aim to validate the developed models through comparisons with observed seismic waveforms and ground motion data, enabling a deeper understanding of the impact of fault damage zones on seismic wave propagation. Integrating these models with real-time data for dynamic updates could also provide critical insights for seismic hazard assessment and early warning systems. Expanding the methodology to broader geophysical contexts, such as gravity or electromagnetic surveys, could further enhance its utility.

Finally, computational efficiency and scalability present important challenges and opportunities. Improving the performance of advanced techniques like Probability Graphical Models (PGMs) and graph-learning approaches for large-scale 3D models is essential for their widespread adoption. Additionally, exploring new physics-informed constraints and innovative regularization methods could help balance detail preservation and smoothness in fused models. Together, these efforts will advance the field of geophysical modeling, contributing to more accurate and actionable insights into subsurface structures and seismic hazards.

Bibliography

[1] Heather Nicolson, Andrew Curtis, Brian Baptie, and Erica Galetti. Seismic interferometry and ambient noise tomography in the british isles. *Proceedings of the Geologists' Association*, 123(1):74–86, 2012.

[2] Hongfeng Yang, Yaohui Duan, Junhao Song, Xiaohuan Jiang, Xiaofeng Tian, Wei Yang, Weitao Wang, and Jun Yang. Fine structure of the chenghai fault zone, yunnan, china, constrained from teleseismic travel time and ambient noise tomography. *Journal of Geophysical Research: Solid Earth*, 125(7):e2020JB019565, 2020.

[3] Peter Gerstoft, Karim G Sabra, Philippe Roux, WA Kuperman, and Michael C Fehler. Green's functions extraction and surface-wave tomography from microseisms in southern california. *Geophysics*, 71(4):SI23–SI31, 2006.

[4] Philippe Roux, Karim G Sabra, Peter Gerstoft, WA Kuperman, and Michael C Fehler. P-waves from cross-correlation of seismic noise. *Geophysical Research Letters*, 32(19), 2005.

[5] Xinming Wu and Xavier Janson. Directional structure tensors in estimating seismic structural and stratigraphic orientations. *Geophysical Journal International*, 210(1):534–548, 2017.

[6] George Barbastathis, Aydogan Ozcan, and Guohai Situ. On the use of deep learning for computational imaging. *Optica*, 6(8):921–943, 2019.

[7] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[8] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians*, volume 1, page 2. Berkeley, CA, 1986.

[9] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov random fields for vision and image processing*. MIT press, 2011.

[10] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[11] Stan Z Li. Markov random field models in computer vision. In *Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, May 2–6 1994 Proceedings, Volume II 3*, pages 361–370. Springer, 1994.

[12] Adrian Barbu. Learning real-time mrf inference for image denoising. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1574–1581. IEEE, 2009.

[13] Tatsuya Yamazaki and Denis Gingras. Image classification using spectral and spatial information based on mrf models. *IEEE transactions on image processing*, 4(9):1333–1339, 1995.

[14] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[15] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293, 2014.

[16] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *Authorea Preprints*, 2024.

[17] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[18] Alexander Maier, Andrew Sharp, and Yuriy Vagapov. Comparative analysis and practical implementation of the esp32 microcontroller module for the internet of things. In *2017 Internet Technologies and Applications (ITA)*, pages 143–148. IEEE, 2017.

[19] Zachary E Ross, Benjamín Idini, Zhe Jia, Oliver L Stephenson, Minyan Zhong, Xin Wang, Zhongwen Zhan, Mark Simons, Eric J Fielding, Sang-Ho Yun, et al. Hierarchical interlocked orthogonal faulting in the 2019 ridgecrest earthquake sequence. *Science*, 366(6463):346–351, 2019.

[20] Susan E Hough, Zachary E Ross, and Timothy E Dawson. Introduction to the special section on the 2019 ridgecrest, california, earthquake sequence. *Bulletin of the Seismological Society of America*, 110(4):1395–1399, 2020.

[21] Dimitri Zigone, Yehuda Ben-Zion, Michel Campillo, and Philippe Roux. Seismic tomography of the southern california plate boundary region from noise-based rayleigh and love waves. *Pure and Applied Geophysics*, 172(5):1007–1032, 2015.

[22] Dimitri Zigone, Yehuda Ben-Zion, Maximilien Lehujeur, Michel Campillo, Gregor Hillers,

and Frank L Vernon. Imaging subsurface structures in the san jacinto fault zone with high-frequency noise recorded by dense linear arrays. *Geophysical Journal International*, 217(2):879–893, 2019.

[23] Michael J Bianco, Peter Gerstoft, Kim B Olsen, and Fan-Chi Lin. High-resolution seismic tomography of long beach, ca using machine learning. *Scientific reports*, 9(1):1–11, 2019.

[24] Fan-Chi Lin, Dunzhu Li, Robert W Clayton, and Dan Hollis. High-resolution 3d shallow crustal structure in long beach, california: Application of ambient noise tomography on a dense seismic array. *Geophysics*, 78(4):Q45–Q56, 2013.

[25] DC Bowden, Victor C Tsai, and Fan C Lin. Site amplification, attenuation, and scattering from noise correlation amplitudes across a dense array in long beach, ca. *Geophysical Research Letters*, 42(5):1360–1367, 2015.

[26] Rufus D Catchings, Mark R Goldman, Jamison H Steidl, Joanne H Chan, Amir A Allam, Coyn J Criley, Zhenning Ma, Daniel S Langermann, Garet J Huddleston, Andrian T McEvilly, et al. Nodal seismograph recordings of the 2019 ridgecrest earthquake sequence. *Seismological Society of America*, 91(6):3622–3633, 2020.

[27] J Julia, CJ Ammon, RB Herrmann, and Am M Correig. Joint inversion of receiver function and surface wave dispersion observations. *Geophysical Journal International*, 143(1):99–112, 2000.

[28] Fan-Chi Lin, Michael H Ritzwoller, and Roel Snieder. Eikonal tomography: surface wave tomography by phase front tracking across a regional broad-band seismic array. *Geophysical Journal International*, 177(3):1091–1110, 2009.

[29] Stephanie Earp and Andrew Curtis. Probabilistic neural network-based 2d travel-time tomography. *Neural Computing and Applications*, 32(22):17077–17095, 2020.

[30] Konik Kothari and Sidharth Gupta. Random mesh projectors for inverse problems. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[31] Hossein S Aghamiry and Ali Gholami. Interval-q estimation and compensation: An adaptive dictionary-learning approach. *Geophysics*, 83(4):V233–V242, 2018.

[32] Michael J Bianco and Peter Gerstoft. Travel time tomography with adaptive dictionaries. *IEEE Transactions on Computational Imaging*, 4(4):499–511, 2018.

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

[34] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine

learning. *arXiv preprint arXiv:1811.12808*, 2018.

[35] Clive D Rodgers. *Inverse methods for atmospheric sounding: theory and practice*, volume 2. World scientific, 2000.

[36] Chris Milliner and Andrea Donnellan. Using daily observations from planet labs satellite imagery to separate the surface deformation between the 4 july m w 6.4 foreshock and 5 july m w 7.1 mainshock during the 2019 ridgecrest earthquake sequence. *Seismological Research Letters*, 2020.

[37] Robert B Herrmann. Computer programs in seismology: An evolving tool for instruction and research. *Seismological Research Letters*, 84(6):1081–1088, 2013.

[38] Harold Magistrale, Keith McLaughlin, and Steven Day. A geology-based 3d velocity model of the los angeles basin sediments. *Bulletin of the Seismological Society of America*, 86(4):1161–1166, 1996.

[39] V. E. Langenheim. Setting of the ridgecrest and searles valley earthquakes from gravity and magnetic data. *Geological Society of America Annual Meeting*, 2019.

[40] T.-Y. Yeh and K.B. Olsen. Fault damage zone effects on ground motions during the 2019 mw 7.1 ridgecrest, ca, earthquake. *Bull. Seismol. Soc. Am.*, 113(4):1724–1738, 2023.

[41] R Ajala and P Persaud. Effect of merging multiscale models on seismic wavefield predictions near the southern san andreas fault. *Journal of Geophysical Research: Solid Earth*, 126(10):e2021JB021915, 2021.

[42] Paul Van Houtte, Anand Krishna Kanjarla, Albert Van Bael, Marc Seefeldt, and Laurent Delannay. Multiscale modelling of the plastic anisotropy and deformation texture of polycrystalline materials. *European Journal of Mechanics-A/Solids*, 25(4):634–648, 2006.

[43] Sergio Barbarossa and Stefania Sardellitti. Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing*, 68:2992–3007, 2020.

[44] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5213–5222, 2020.

[45] Andreas Fichtner, Jeannot Trampert, Paul Cupillard, Erdinc Saygin, Tuncay Taymaz, Yann Capdeville, and Antonio Villasenor. Multiscale full waveform inversion. *Geophysical Journal International*, 194(1):534–556, 2013.

[46] Hongjian Fang and Haijiang Zhang. Wavelet-based double-difference seismic tomography

with sparsity regularization. *Geophysical Journal International*, 199(2):944–955, 2014.

[47] Lingchen Zhu, Entao Liu, and James H McClellan. Seismic data denoising through multiscale and sparsity-promoting dictionary learning. *Geophysics*, 80(6):WD45–WD57, 2015.

[48] Michael Bianco and Peter Gerstoft. Dictionary learning of sound speed profiles. *The Journal of the Acoustical Society of America*, 141(3):1749–1758, 2017.

[49] Hao Zhang and Yehuda Ben-Zion. Enhancing regional seismic velocity models with higher-resolution local results using sparse dictionary learning. *Journal of Geophysical Research: Solid Earth*, 129(1):e2023JB027016, 2024.

[50] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[51] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

[52] Lina Liu, Jianwei Ma, and Gerlind Plonka. Sparse graph-regularized dictionary learning for suppressing random seismic noise. *Geophysics*, 83(3):V215–V231, 2018.

[53] Hao Zhang, Peimin Zhu, Zhiying Liao, and Zewei Li. Saltiscg: Interactive salt segmentation method based on cnn and graph cut. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[54] Flavio Cannavò, Andrea Cannata, Carmelo Cassisi, Giuseppe Di Grazia, Placido Montalto, Michele Prestifilippo, Eugenio Privitera, Mauro Coltelli, and Salvatore Gambino. A multivariate probabilistic graphical model for real-time volcano monitoring on mount etna. *Journal of Geophysical Research: Solid Earth*, 122(5):3480–3496, 2017.

[55] Pratyusha Das and Antonio Ortega. Graph-based skeleton data compression. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020.

[56] Songyang Zhang, Qinwen Deng, and Zhi Ding. Multilayer graph spectral analysis for hyperspectral images. *EURASIP Journal on Advances in Signal Processing*, 2022(1):1–25, 2022.

[57] Martijn PA van den Ende and J-P Ampuero. Automated seismic source characterization using deep graph neural networks. *Geophysical Research Letters*, 47(17):e2020GL088690,

2020.

[58] Gwantae Kim, Bonhwa Ku, Jae-Kwang Ahn, and Hanseok Ko. Graph convolution networks for seismic events classification using raw waveform data from multiple stations. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[59] Xuebin Zhao, Andrew Curtis, and Xin Zhang. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228(1):213–239, 2022.

[60] Xin Zhang, Angus Lomas, Muhong Zhou, York Zheng, and Andrew Curtis. 3-d bayesian variational full waveform inversion. *Geophysical Journal International*, 234(1):546–561, 2023.

[61] He-Qing Mu and Ka-Veng Yuen. Ground motion prediction equation development by heterogeneous bayesian learning. *Computer-Aided Civil and Infrastructure Engineering*, 31(10):761–776, 2016.

[62] Kevin Murphy. An introduction to graphical models. *Rap. tech*, 96:1–19, 2001.

[63] Zheng Zhou, Peter Gerstoft, and Kim Bak Olsen. 3d multiresolution velocity model fusion with probability graphical models. *Bulletin of the Seismological Society of America*, 114(3):1279–1292, 2024.

[64] Zheng Zhou, Michael Bianco, Peter Gerstoft, and Kim Olsen. High-resolution imaging of complex shallow fault zones along the july 2019 ridgecrest ruptures. *Geophysical Research Letters*, 49(1):e2021GL095024, 2022.

[65] Rongxi Gou, Yijie Zhang, Xueyu Zhu, and Jinghuai Gao. Bayesian physics-informed neural networks for the subsurface tomography based on the eikonal equation. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[66] Yunyue Elita Li, Daniel O'Malley, Greg Beroza, Andrew Curtis, and Paul Johnson. Machine learning developments and applications in solid-earth geosciences: Fad or future?, 2023.

[67] Gene Cheung, Enrico Magli, Yuichi Tanaka, and Michael K Ng. Graph spectral image processing. *Proceedings of the IEEE*, 106(5):907–930, 2018.

[68] Zheng Zhou, Peter Gerstoft, and Kim Olsen. Fusion of multi-resolution seismic tomography maps with physics-informed probability graphical models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2540–2544. IEEE, 2024.

[69] Zhongping Zhang, Yixuan Zhang, Zheng Zhou, and Jiebo Luo. Boundary-based image

forgery detection by fast shallow cnn. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2658–2663. IEEE, 2018.

[70] Mohammad Ali Shahrabi, Hosein Hashemi, and Mohammad Kazem Hafizi. Application of mixture of gaussian clustering on joint facies interpretation of seismic and magnetotelluric sections. *Pure and Applied Geophysics*, 173:623–636, 2016.

[71] Haiyan Liu, Cai Liu, Yang Liu, Ying Zhang, and Fengxia Gao. Application of coherence technique based on manhattan distance. *Global Geology*, 32(1):144–151, 2013.

[72] Edward I George, UE Makov, and Adrian FM Smith. Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, pages 147–156, 1993.

[73] Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov. Real-time computer vision with opencv. *Communications of the ACM*, 55(6):61–69, 2012.

[74] Stan Z Li. *Markov random field modeling in computer vision*. Springer Science & Business Media, 2012.

[75] Oscar A Zuniga and Robert M Haralick. Integrated directional derivative gradient operator. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(3):508–517, 1987.

[76] Dina E Melas and Simon P Wilson. Double markov random fields and bayesian image segmentation. *IEEE Transactions on Signal Processing*, 50(2):357–365, 2002.

[77] Malcolm Sambridge and Klaus Mosegaard. Monte carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3):3–1, 2002.

[78] Chain Monte Carlo. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, 581(540):3, 2004.

[79] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[80] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012.

[81] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[82] D Poobathy and R Manicka Chezian. Edge detection operators: Peak signal to noise ratio based comparison. *IJ Image, Graphics and Signal Processing*, 10:55–61, 2014.

[83] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020.

[84] Rasheed Ajala and Patricia Persaud. Ground-motion evaluation of hybrid seismic velocity models. *The Seismic Record*, 2(3):186–196, 2022.

[85] Sucharita Ghosh. *Kernel smoothing: Principles, methods and applications*. John Wiley & Sons, 2018.

[86] Rasheed Ajala, Patricia Persaud, and Alan Juarez. Earth model-space exploration in southern california: Influence of topography, geotechnical layer, and attenuation on wavefield accuracy. *Frontiers in Earth Science*, 10:964806, 2022.

[87] Hao Zhang and Yehuda Ben-Zion. Enhancing regional seismic velocity models with higher-resolution local results using sparse dictionary learning. *Journal of Geophysical Research: Solid Earth*, 129(1):e2023JB027016, 2024.

[88] Benjamin Edwards and Donat Fäh. A stochastic ground-motion model for switzerland. *Bulletin of the Seismological Society of America*, 103(1):78–98, 2013.

[89] Miroslav Dudik, Steven J Phillips, and Robert E Schapire. Performance guarantees for regularized maximum entropy density estimation. In *International Conference on Computational Learning Theory*, pages 472–486. Springer, 2004.

[90] Zheng Zhou, Peter Gerstoft, and Kim Olsen. Graph-learning approach to combine multiresolution seismic velocity models. *Authorea Preprints*, 2023.

[91] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[92] Ivanka Horová, Jan Kolacek, and Jiri Zelinka. *Kernel Smoothing in MATLAB: theory and practice of kernel smoothing*. World scientific, 2012.

[93] Daichi Shimbo, Nancy T Artinian, Jan N Basile, Lawrence R Krakoff, Karen L Margolis, Michael K Rakotz, Gregory Wozniak, American Heart Association, and the American Medical Association. Self-measured blood pressure monitoring at home: a joint policy statement from the american heart association and american medical association. *Circulation*, 142(4):e42–e63, 2020.

[94] Insang You, Bongsoo Kim, Jaeyoon Park, Kunsuk Koh, Sangbaie Shin, Sungjune Jung, and UJAM Jeong. Stretchable e-skin apexcardiogram sensor. *Advanced Materials (Deerfield Beach, Fla.)*, 28(30):6359–6364, 2016.

[95] Sen Yang, Yaping Zhang, Siu-Yeung Cho, Ricardo Correia, and Stephen P Morgan. Non-invasive cuff-less blood pressure estimation using a hybrid deep learning model. *Optical and Quantum Electronics*, 53:1–20, 2021.

[96] Liangqi Wang, Shuo Tian, and Rong Zhu. A new method of continuous blood pressure monitoring using multichannel sensing signals on the wrist. *Microsystems & Nanoengineering*, 9(1):117, 2023.

[97] Xiaoman Xing and Mingshan Sun. Optical blood pressure estimation with photoplethysmography and fft-based neural networks. *Biomedical optics express*, 7(8):3007–3020, 2016.

[98] Matt Y Cheung, Ashutosh Sabharwal, Gerard L Cot, Ashok Veeraraghavan, et al. Wearable blood pressure monitoring devices: Understanding heterogeneity in design and evaluation. *IEEE Transactions on Biomedical Engineering*, 2024.

[99] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[100] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *Journal of Machine Learning Research*, 21:188:1–188:37, 2020.

[101] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *Journal of Artificial Intelligence Research*, 82:991–1029, 2023.

[102] Liangye Li, Yanpeng Li, Liuyang Yang, Fang Fang, Zhijun Yan, and Qizhen Sun. Continuous and accurate blood pressure monitoring based on wearable optical fiber wristband. *IEEE Sensors Journal*, 21(3):3049–3057, 2020.

[103] Yi-Neng Pang, Bin Liu, Juan Liu, Sheng-Peng Wan, Tao Wu, Jinhui Yuan, Xiangjun Xin, Xing-Dao He, and Qiang Wu. Singlemode-multimode-singlemode optical fiber sensor for accurate blood pressure monitoring. *Journal of Lightwave Technology*, 40(13):4443–4450, 2022.

[104] Yang Ji, Yunchul Chung, D Sprinzak, Moty Heiblum, Diana Mahalu, and Hadas Shtrikman. An electronic mach–zehnder interferometer. *Nature*, 422(6930):415–418, 2003.

[105] Fen Miao, Nan Fu, Yuan-Ting Zhang, Xiao-Rong Ding, Xi Hong, Qingyun He, and Ye Li. A novel continuous blood pressure estimation approach based on data mining techniques. *IEEE journal of biomedical and health informatics*, 21(6):1730–1740, 2017.

[106] Md Sayed Tanveer and Md Kamrul Hasan. Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ann-lstm network. *Biomedical Signal Processing and Control*, 51:382–392, 2019.

[107] Jianjun Yan, Xianglei Cai, Guangyao Zhu, Rui Guo, Haixia Yan, and Yiqin Wang. A non-invasive blood pressure prediction method based on pulse wave feature fusion. *Biomedical Signal Processing and Control*, 74:103523, 2022.

[108] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

[109] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[110] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022.

[111] Ramakrishna Mukkamala, Jin-Oh Hahn, Omer T Inan, Lalit K Mestha, Chang-Sei Kim, Hakan Töreyin, and Survi Kyal. Toward ubiquitous blood pressure monitoring via pulse transit time: theory and practice. *IEEE transactions on biomedical engineering*, 62(8):1879–1901, 2015.

[112] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[113] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[114] Qing Wang, Weimin Lyu, Zhi Cheng, and Changyuan Yu. Noninvasive measurement of vital signs with the optical fiber sensor based on deep learning. *Journal of Lightwave Technology*, 41(13):4452–4462, 2023.