Unearthing InSights into Mars:
Unsupervised Source Separation with Limited Data

Ali Siahkoohi [1]   **Rudy Morel** [2]   **Maarten V. de Hoop** [1]   **Erwan Allys** [3]   **Grégory Sainton** [4]   **Taichi Kawamura** [4]

Abstract

Source separation involves the ill-posed problem of retrieving a set of source signals that have been observed through a mixing operator. Solving this problem requires prior knowledge, which is commonly incorporated by imposing regularity conditions on the source signals, or implicitly learned through supervised or unsupervised methods from existing data. While data-driven methods have shown great promise in source separation, they often require large amounts of data, which rarely exists in planetary space missions. To address this challenge, we propose an unsupervised source separation scheme for domains with limited data access that involves solving an optimization problem in the wavelet scattering covariance representation space—an interpretable, low-dimensional representation of stationary processes. We present a real-data example in which we remove transient, thermally-induced microtilts—known as glitches—from data recorded by a seismometer during NASA's InSight mission on Mars. Thanks to the wavelet scattering covariances' ability to capture non-Gaussian properties of stochastic processes, we are able to separate glitches using only a few glitch-free data snippets.

1. Introduction

Source separation is a problem of fundamental importance in the field of signal processing, with a wide range of applications in various domains such as telecommunications (Chevreuil & Loubaton, 2014; Gay & Benesty, 2012;
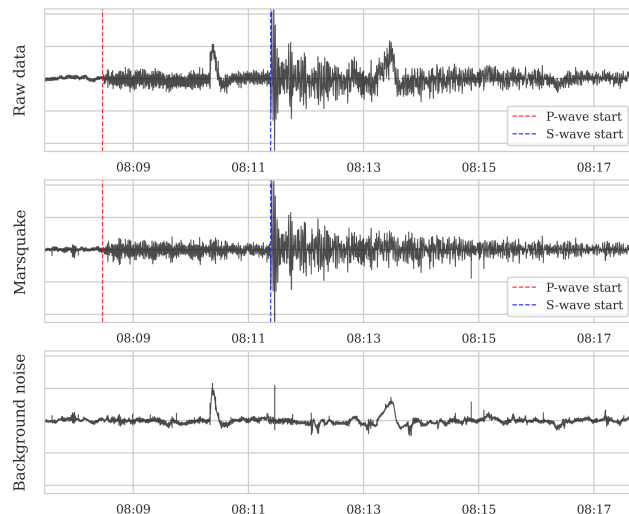

Figure 1. Unsupervised separation of background noise, including thermally induced microtilts (glitches), from a marsquake recorded by the InSight lander's seismometer on February 3, 2022 (InSight Marsquake Service, 2023). Approximately 30 hours of raw data from the U component, with no recorded marsquakes, were utilized for background noise separation without any explicit prior knowledge of marsquakes or glitches. The horizontal axis represents the UTC time zone.

Khosravy et al., 2020), speech processing (Pedersen et al., 2008; Chua et al., 2016; Grais et al., 2014), biomedical signal processing (Adali et al., 2015; Barriga et al., 2003; Hasan et al., 2018) and geophysical data processing (Ibrahim & Sacchi, 2014; Kumar et al., 2015; Scholz et al., 2020). Source separation arises when multiple source signals of interest are combined through a mixing operator. The goal is to estimate the original sources with minimal prior knowledge of the mixing process or the source signals themselves. This makes source separation a challenging problem, as the number of sources is usually unknown, and the sources are often non-Gaussian, nonstationary, and multiscale.

Classical signal-processing based source separation methods (Cardoso, 1989; Jutten & Herault, 1991; Bingham & Hyvärinen, 2000; Nandi & Zarzoso, 1996; Cardoso, 1998; Jutten et al., 2004) while being extensively studied and well understood, often make simplifying assumptions regard-

---

[1]Department of Computational Applied Mathematics & Operations Research, Rice University [2]Département d'informatique de l'ENS, ENS, CNRS, PSL University, Paris, France [3]Laboratoire de Physique de l'École normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France [4]Institut de Physique du Globe de Paris. Correspondence to: Ali Siahkoohi <alisk@rice.edu>.

ing the sources, e.g., sources being distributed according to Gaussian or Laplace distributions, which might negatively bias the outcome of source separation (Cardoso, 1998; Parra & Sajda, 2003). To partially address the shortcomings of classical approaches, deep learning methods have been proposed as an alternative approach for source separation, which exploit the information in existing datasets to learn prior information about the sources. In particular, supervised learning methods (Jang & Lee, 2003; Hershey et al., 2016; Ke et al., 2020; Kameoka et al., 2019; Wang & Chen, 2018) commonly rely on existence of labeled training data and perform source separation using an end-to-end training scheme. However, since they require access to ground truth source signals for training, supervised methods are limited to domains in which labeled training data is available.

On the other hand, unsupervised source separation methods (Févotte et al., 2009; Drude et al., 2019; Wisdom et al., 2020; Liu et al., 2022; Denton et al., 2022; Neri et al., 2021) do not rely on the existence of labeled training data and instead attempt to infer the sources based on the properties of the observed signals. These methods make minimal assumptions about the underlying sources, which make them a suitable choice for realistic source separation problems. Despite their success, unsupervised source separation methods often require tremendous amount of data during training (Wisdom et al., 2020), which is often infeasible in certain applications such as problem arising in planetary space missions, e.g., due to challenges associated with data acquisition. Moreover, generalization concerns preclude the use of data-driven methods trained on synthetic data in real-world applications due to the discrepancies between synthetic and real data.

To address these challenges, we propose an unsupervised source separation method applicable to domains with limited access to data. In order to achieve this, we embed inductive biases into our approach through the use of domain knowledge from time-series analysis and signal processing via the an extension of scattering networks (Bruna & Mallat, 2013). As a means of capturing non-Gaussian and multi-scale characteristics of the sources, we extract second-order information of scattering coefficients, known as the wavelet scattering covariance representation (Morel et al., 2022). We perform source separation by solving an optimization problem over the unknown sources that entails minimizing multiple carefully selected and normalized loss functions in the wavelet scattering covariance representations space. These loss function are designed to: (1) ensure data-fidelity, i.e., enforce the recovered sources to explain the observed (mixed) data; (2) incorporate prior knowledge in the form of limited (e.g., ≈ 50) training examples from one of the sources; and (3) impose a notion of statistical independence between the recovered sources. Our proposed method does not require any labeled training data, and can effectively separate sources even in scenarios where access to data is limited.

As a motivating example, we apply our approach to data recorded by a seismometer on Mars during NASA's Interior Exploration using Seismic Investigations, Geodesy and Heat Transport (InSight) mission (Giardini et al., 2020; Golombek et al., 2020; Knapmeyer-Endrun & Kawamura, 2020). The InSight lander's seismometer—known as the SEIS instrument—detected marsquakes (Horleston et al., 2022; Ceylan et al., 2022; Panning et al., 2023; InSight Marsquake Service, 2023) and transient atmospheric signals, such as wind and temperature changes, that provide information about the Martian atmosphere (Stott et al., 2022) and enable studying the interior structure and composition of the Red Planet (Beghein et al., 2022). The signal recorded by the InSight seismometer is heavily influenced by atmospheric activity and surface temperature (Lognonné et al., 2020; Lorenz et al., 2021), resulting in a distinct daily pattern. Among different types of noise, transient thermally induced microtilts, commonly referred to as glitches (Scholz et al., 2020; Barkaoui et al., 2021), are a significant component of the noise and one of the most frequent recorded events. These glitches, hinder the downstream analysis of the data if left uncorrected (Scholz et al., 2020). We show that our method is capable of removing glitches from the recorded data by only using a few snippets of glitch-free data.

In the following sections, after describing the related work, we introduce wavelet scattering covariance as a domain-knowledge rich representation for analyzing time-series and provide justification for their usage in the context of source separation. As a means to perform source separation in domains with limited data, we introduce our source separation approach that involves solving an optimization problem with loss functions defined in the wavelet scattering covariance space. We present two numerical experiments: (1) a synthetic setup in which we can quantify the accuracy of our method; and (2) examples involving seismic data recorded during the NASA InSight mission.

2. Related Work

Regaldo-Saint Blancard et al. (2021) introduced the notion of components separation through a gradient descent in signal space with indirect constraints with applications to to the separation of an astrophysical emission (polarized dust emission in microwave) and instrumental noise. In an extensive study, Delouis, J.-M. et al. (2022) attempts to separate the full sky observation of the dust emission with instrumental noise using similar techniques via wavelet scattering covariance representations. Authors take the nonstationarity of the signal into account by constraining statistics on several sky masks. Contrarily to a usual denoising approach, both of these works focus primarily on recovering the statis-

tics of the signal of interest. In a related approach, Jeffrey et al. (2022) use a scattering transform generative model to perform source separation in a Bayesian framework. While very efficient, this approach requires training samples from each component, which are often not available. Finally, Xu et al. (2022) similarly aim to remove glitches and they develop a supervised learning based on deglitched data obtained by existing glitch removal tools. As a result, the accuracy of their result is limited to the accuracy of the underlying data processing tool, which our method avoid by being unsupervised. As we show in our examples, we are able to detect and remove glitches that were undetected by the main deglitching software (Scholz et al., 2020) developed closely by the InSight team.

3. Wavelet Scattering Covariance

In order to enable unsupervised source separation with limited quantities of data, we propose to design a low-dimensional, domain-knowledge rich representation of data with which we perform source separation. This is partially motivated by recent success of self-supervised learning methods in natural language processing where high-performing representations of data—obtained through pretrained Transformers (Vaswani et al., 2017; Baevski et al., 2020; Gulati et al., 2020; Zhang et al., 2020)—are used in place of raw data to successfully perform various downstream tasks (Polyak et al., 2021; Gulati et al., 2020; Baevski et al., 2020; Zhang et al., 2020; Chung et al., 2021; Siahkoohi et al., 2022).

Due to our limited access to data, we cannot employ self-supervised learning with Transformers to acquire high-performing data representations. Instead, we propose to use wavelet scattering covariances (Morel et al., 2022) as means to transfer data to a suitable representation space for source separation. Rooted in scattering networks (Bruna & Mallat, 2013), wavelet scattering covariances provide interpretable representations of data and are able to characterize a wide range of non-Gaussian properties of multiscale stochastic processes (Morel et al., 2022)—a type of signals that we consider in this paper. The wavelet scattering covariance generally does not require any pretraining and its weights, i.e., wavelets in the scattering network, are often chosen beforehand (see Seydoux et al. (2020) for a data-driven wavelet choice) according to the time-frequency properties of data. In the next section, we introduce the construction of this representation space by first describing scattering networks.

3.1. Wavelet Transform and Scattering Networks

The main ingredient of the wavelet scattering covariance representation is a scattering network (Bruna & Mallat, 2013) that consists of a cascade of wavelet transforms fol-

lowed by a nonlinear activation function (akin to a typical convolutional neural network). In this network architecture, the wavelet transform, denoted by a linear operator $\mathbf{W}$, is a convolutional operator with predefined kernels, i.e., wavelet filters. These filters include a low-pass filter $\varphi_J(t)$ and J complex-valued band-pass filters $\psi_j(t) = 2^{-j}\psi(2^{-j}t)$, $1 \leq j \leq J$, which are obtained by the dilation of a mother wavelet $\psi(t)$ and have zero-mean and a fast decay away from $t = 0$. The wavelet transform is often followed by the modulus operator in scattering networks. The output of a two-layer scattering network S can be written as,

$$S(\mathbf{x}) := \begin{bmatrix} \mathbf{Wx} \\ \mathbf{W}|\mathbf{Wx}| \end{bmatrix}, \tag{1}$$

where $\mathbf{Wx} := \mathbf{x} \star \psi_j(t)$ denotes the wavelet transform that extracts variations of the input signal $\mathbf{x}(t)$ around time t at scale 2^j, and $|\cdot|$ is the modulus activation function (Bruna & Mallat, 2013). The second component $\mathbf{W}|\mathbf{Wx}|$ computes the variations at different time and scales of the wavelet coefficients $\mathbf{Wx}$. The scattering transform yields features that characterize time evolution of signal envelopes at different scales. Even though such representation has many successful applications, e.g., intermittency analysis (Bruna et al., 2015), clustering (Seydoux et al., 2020), event detection and segmentation (Rodríguez et al., 2021) (with learnable wavelets), it is not sufficient to build accurate models of multiscale processes as it does not capture crucial dependencies across different scales (Morel et al., 2022).

3.2. Capturing Non-Gaussian Characteristics of Stochastic Processes

The dependencies across different scales in scattering transform coefficients are crucial in characterizing and discriminating non-Gaussian signals (Morel et al., 2022). To capture them, we explore the outer product of the scattering coefficients matrix $S(\mathbf{x})S(\mathbf{x})^{\mathrm{H}}$:

$$\begin{bmatrix} \mathbf{Wx}\,(\mathbf{Wx})^{\mathrm{H}} & \mathbf{Wx}\,(\mathbf{W}|\mathbf{Wx}|)^{\mathrm{H}} \\ \mathbf{W}|\mathbf{Wx}|\,(\mathbf{Wx})^{\mathrm{H}} & \mathbf{W}|\mathbf{Wx}|\,(\mathbf{W}|\mathbf{Wx}|)^{\mathrm{H}} \end{bmatrix}. \tag{2}$$

In the above expression, $^{\mathrm{H}}$ denotes the conjugate transpose operation. The above matrix contains three types of coefficients:

- The correlation coefficients $\mathbf{Wx}\,(\mathbf{Wx})^{\mathrm{H}}$ across scales form a quasi-diagonal matrix, because separate scales do not correlate due to phase fluctuation, whether separate scales are dependent or not (Morel et al., 2022). We thus only keep its diagonal coefficients, which correspond to the wavelet power spectrum;
- The correlation coefficients $\mathbf{Wx}\,(\mathbf{W}|\mathbf{Wx}|)^{\mathrm{H}}$ capture signed interaction between wavelet coefficients. In particular, they detect sign-asymmetry and time-asymmetry in $\mathbf{x}$ (Morel et al., 2022). We also consider

a diagonal approximation to this matrix. For the same reason as $\mathbf{W}\mathbf{x}\,(\mathbf{W}\mathbf{x})^{\mathsf{H}}$, two separate scales on the last wavelet operator do not correlate. However, there may exist a correlation between $\mathbf{W}\mathbf{x}$ at a given scale and $|\mathbf{W}\mathbf{x}|$ at a distinct scale, and we retain these;

- Finally coefficients $\mathbf{W}|\mathbf{W}\mathbf{x}|\,(\mathbf{W}|\mathbf{W}\mathbf{x}|)^{\mathsf{H}}$ capture correlations between signal envelopes $|\mathbf{W}\mathbf{x}|$ at different scales. These correlations account for intermittency and time-asymmetry (Morel et al., 2022). Once again, we retain only those coefficients that demonstrate correlation between same-scale channels on the second wavelet operator.

We denote $\mathrm{diag}\left(S(\mathbf{x})S(\mathbf{x})^{\mathsf{H}}\right)$ as such diagonal approximation of the full sparse matrix $S(\mathbf{x})S(\mathbf{x})^{\mathsf{H}}$. The *wavelet scattering covariance* representation is obtained by computing the time average (average pool, denoted by Ave) of this diagonal approximation:

$$\Phi(\mathbf{x}) := \mathrm{Ave}\left(\begin{bmatrix} S(\mathbf{x}) \\ \mathrm{diag}\left(S(\mathbf{x})S(\mathbf{x})^{\mathsf{H}}\right) \end{bmatrix}\right). \qquad (3)$$

Non-Gaussian properties of $\mathbf{x}$ can be detected through non-zero coefficients of Φ. Indeed, let us separate real coefficients and potentially complex coefficients $\Phi(\mathbf{x}) = \left(\Phi_{\mathrm{real}}(\mathbf{x}), \Phi_{\mathrm{complex}}(\mathbf{x})\right)$, with $\Phi_{\mathrm{real}}(\mathbf{x})$ being the real coefficients $\mathrm{Ave}\left(|\mathbf{W}\mathbf{x}|, |\mathbf{W}\mathbf{x}|^2, |\mathbf{W}|\mathbf{W}\mathbf{x}||^2\right)$ and $\Phi_{\mathrm{complex}}(\mathbf{x})$ being the remaining potentially complex coefficients, that is the cross-layer correlations $\mathrm{Ave}\left(\mathbf{W}\mathbf{x}(\mathbf{W}|\mathbf{W}\mathbf{x}|)^{\mathsf{H}}\right)$ or the second layer correlations $\mathrm{Ave}\left(\mathbf{W}|\mathbf{W}\mathbf{x}|(\mathbf{W}|\mathbf{W}\mathbf{x}|)^{\mathsf{H}}\right)$ with different scale correlation on the first wavelet operator.

Proposition 3.1. *If* $\mathbf{x}$ *is Gaussian then* $\Phi_{complex}(\mathbf{x}) \approx 0$. *If* $\mathbf{x}$ *is time-symmetric, then* $\mathrm{Im}\,\Phi_{complex}(\mathbf{x}) \approx 0$.

More precisely, beyond detecting non-Gaussianity through non-zero coefficients up to estimation error, $\Phi(\mathbf{x})$ is able to quantify different non-Gaussian behaviors, which will be crucial for source separation. Appendix A.3 presents a dashboard that visualizes $\Phi(\mathbf{x})$ and can be used to interpret signal non-Gaussian properties such as sparsity, intermittency, and time-asymmetry.

The dimensionality of the wavelet scattering covariance representation depends on the number of scales J considered i.e. the number of wavelet filters of $\mathbf{W}$. In order for largest scale coefficients to be well estimated, one should choose $J \ll \log_2(d)$ where d is input data dimension. The maximum number of coefficients in Φ is smaller than $\log_2^3(d)$ for $d \geq 3$ (Morel et al., 2022). Contrary to higher dimensional representations or higher order statistics, scattering covariance $\Phi(\mathbf{x})$ are low-dimensional, low-order statistics that can be efficiently estimated on a single realization of a source and does not require tremendous amount of data for estimation to converge (Morel et al., 2022). In other

word, Φ is a low-variance representation. This point is key for our source separation algorithm to be applied on limited data. Wavelet scattering covariance Φ extracts average and correlation features from a two-layer convolutional neural network with predefined wavelet filters. It is analogous to the features extracted in Gatys et al. (2015) for generation, that considers however a pretrained convolutional neural network. In the following we will also make use of the scattering cross-covariance representation $\Phi(\mathbf{x}, \mathbf{y}) = \mathrm{Ave}\,\mathrm{diag}\left(S(\mathbf{x})S(\mathbf{y})^{\mathsf{H}}\right)$ that captures scale dependencies across two signals $\mathbf{x}$ and $\mathbf{y}$.

Proposition 3.2. *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are independent then*

$$\Phi(\mathbf{x}, \mathbf{y}) \approx 0$$

The above proposition shows that $\Phi(\mathbf{x}, \mathbf{y})$ detects independence up to estimation error, which will be useful when it comes to separating independent sources.

4. Unsupervised Source Separation

To enable high-fidelity source separation in domains in which access to training data—supervised or unsupervised—is limited, we cast source separation as an optimization problem in a suitable feature space. Owing to wavelet scattering covariance representation's ability to capture non-Gaussian properties of multiscale stochastic processes without any training, we perform source separation by solving an optimization problem over the unknown sources using loss functions over wavelet scattering covariance representations. Due to the inductive bias embedded in the design of this representation space, we gain access to interpretable features, which could further inform us regarding the quality of the source separation process.

4.1. Problem Setup

Consider a linear mixing of unknown sources $\mathbf{s}_i^*(t)$, $i = 1, \dots, N$ via a mixing operator $\mathbf{A}$,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}^*(t) + \boldsymbol{\nu}(t) = \mathbf{a}_1^\top \mathbf{s}_1^*(t) + \mathbf{n}(t), \qquad (4)$$

with

$$\mathbf{s}^*(t) = [\mathbf{s}_1^*(t), \dots, \mathbf{s}_N^*(t)]^\top, \ \mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \cdots \mathbf{a}_N^\top \end{bmatrix},$$
$$\mathbf{n}(t) = \boldsymbol{\nu}(t) + \sum_{i=2}^{N} \mathbf{a}_i^\top \mathbf{s}_i^*(t). \qquad (5)$$

In the above expressions, $\mathbf{x}(t)$ represents the observed data, and $\boldsymbol{\nu}(t)$ is the measurement noise. Here we capture the noise and the mixture of all the sources except for $\mathbf{s}_1^*(t)$ through the mixing operator in $\mathbf{n}(t)$ that does not longer depends on $\mathbf{s}_1^*(t)$. The matrices $\mathbf{x}(t)$ and $\mathbf{s}(t)$ have dimensions of $M \times T$ and $N \times T$, respectively, where T represents the

number of time samples. The mixing operator $\mathbf{A}$ has dimensions of $M \times N$. As a result, the product of $\mathbf{a}_1^\top$ and $\mathbf{s}_1(t)$ yields a matrix of dimensions $M \times T$, which corresponds to the contributions of source $\mathbf{s}_1(t)$ exclusively in $\mathbf{x}(t)$.

Objective. The aim is to obtain a point estimate $\mathbf{s}_1(t)$ given a single observation $\mathbf{x}(t)$ with the assumption that $\mathbf{a}_1$ is known and that we have access to a few realizations $\{\mathbf{n}_k(t)\}_{k=1}^K$ as a training dataset. For example, in the case of separating glitches from seismic data recorded during the NASA InSight mission, we will consider $\mathbf{n}_k(t)$ to be snippets of glitch-free data and $\mathbf{a}_1$ to encodes information regarding polarization. We will drop the time dependence of the quantities in equations (4) and (5) for convenience.

4.2. Principle of the Method

The inverse problem of estimating $\mathbf{s}_1$ from the given observed data $\mathbf{x}$, as presented in equation (4), is ill-posed since the solution is not unique. To constrain the solution space of the problem, we incorporate prior knowledge in the form of realizations $\{\mathbf{n}_k\}_{k=1}^K$. We achieve this through a loss function that emphasizes the wavelet scattering covariance representation of $\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1$ to be close to that of $\mathbf{n}_k$, $k = 1, \ldots, K$:

$$\mathcal{L}_{\text{prior}}(\mathbf{s}_1) := \frac{1}{K} \sum_{k=1}^K \left\| \Phi(\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1) - \Phi(\mathbf{n}_k) \right\|_2^2. \quad (6)$$

In the above expression, Φ is the wavelet scattering covariance mapping as described in equation (3). With the prior loss defined, we impose data-consistency via:

$$\mathcal{L}_{\text{data}}(\mathbf{s}_1) := \frac{1}{K} \sum_{k=1}^K \left\| \Phi(\mathbf{a}_1^\top \mathbf{s}_1 + \mathbf{n}_k) - \Phi(\mathbf{x}) \right\|_2^2. \quad (7)$$

The data consistency loss function $\mathcal{L}_{\text{data}}$ promotes estimations of $\mathbf{s}_1$ that for any training example from $\{\mathbf{n}_k\}_{k=1}^K$ the wavelet scattering covariance representation of $\mathbf{a}_1^\top \mathbf{s}_1 + \mathbf{n}_k$ is close to that of the observed data.

To promote the independence of sources, we penalize the scattering cross-covariance between $\mathbf{a}_1^\top \mathbf{s}_1$ and $\mathbf{n}_k$.

$$\mathcal{L}_{\text{cross}}(\mathbf{s}_1) := \frac{1}{K} \sum_{k=1}^K \left\| \Phi(\mathbf{a}_1^\top \mathbf{s}_1, \mathbf{n}_k) \right\|_2^2, \quad (8)$$

where $\Phi(\cdot, \cdot)$ is the scattering cross-covariance representation (see section 3.2).

4.3. Loss Normalization

The losses described previously do not contain any weighting term for the different coefficients of the scattering covariance representation. We introduce in this section a generic normalization scheme, based on the estimated variance of certain scattering covariance distributions. This normalization, which has been introduced in Delouis, J.-M. et al. (2022), allows to interpret the different loss terms in a standard form, and to include them additively in the total loss term without overall loss weights. Let us consider first the loss term given by equation (6), which compares the distance between $\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1$ and available training samples $\{\mathbf{n}_k\}_{k=1}^K$ in the wavelet scattering representation space. Specifying explicitly the sum on the M wavelet scattering covariance coefficients Φ_m, $m = 1, \ldots, M$, it yields

$$\mathcal{L}_{\text{prior}}(\mathbf{s}_1) = \frac{1}{MK} \sum_{m=1}^M \sum_{k=1}^K \left| \Phi_m(\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1) - \Phi_m(\mathbf{n}_k) \right|^2.$$

Let us consider the second sum in this expression. In the limit where $\Phi_m(\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1)$ is drawn from the same distribution as $\{\Phi_m(\mathbf{n}_k)\}_k^K$, the difference $\Phi_m(\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1) - \Phi_m(\mathbf{n}_k)$, seen as a random variable, should have zero mean, and the same variance as the distribution $\{\Phi_m(\mathbf{n}_k)\}_k^K$ up to a factor 2. Denoting $\sigma^2(\Phi_m(\mathbf{n}_k))$ as this variance, which can be estimated from $\{\Phi_m(\mathbf{n}_k)\}_k^K$, this gives a natural way of normalizing the loss:

$$\mathcal{L}_{\text{prior}}(\mathbf{s}_1) = \frac{1}{MK} \sum_{m=1}^M \sum_{k=1}^K \frac{\left| \Phi_m(\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1) - \Phi_m(\mathbf{n}_k) \right|^2}{\sigma^2(\Phi_m(\mathbf{n}_k))}$$

or in a compressed form

$$\mathcal{L}_{\text{prior}}(\mathbf{s}_1) = \frac{1}{K} \sum_{k=1}^K \frac{\left\| \Phi(\mathbf{x} - \mathbf{a}_1^\top \mathbf{s}_1) - \Phi(\mathbf{n}_k) \right\|_2^2}{\sigma^2(\Phi(\mathbf{n}_k))}, \quad (9)$$

which takes into account the expected standard deviation of each coefficient of the scattering covariance representation. This normalization allows for two things. First, it removes the normalization inherent to the multiscale structure of Φ. Indeed, coefficients involving low frequency wavelets tend to have a larger norm. Second, it allows to interpret the loss value, which is expected to be at best of order unity and to sum different loss terms of same magnitude.

We can introduce a similar normalization for the other loss terms. Loss term (7) should be normalized by the M-dimensional vector $\sigma^2(\Phi(\mathbf{a}_1^\top \mathbf{s}_1 + \mathbf{n}_k))$ that we approximate by $\sigma^2(\Phi(\mathbf{x} + \mathbf{n}_k))$, in order to have a normalization independent on $\mathbf{s}_1$, yielding

$$\mathcal{L}_{\text{data}}(\mathbf{s}_1) := \frac{1}{K} \sum_{k=1}^K \frac{\left\| \Phi(\mathbf{a}_1^\top \mathbf{s}_1 + \mathbf{n}_k) - \Phi(\mathbf{x}) \right\|_2^2}{\sigma^2(\Phi(\mathbf{x} + \mathbf{n}_k))}. \quad (10)$$

Finally, loss term (8) should be normalized by $\sigma^2(\Phi(\mathbf{a}_1^\top \mathbf{s}_1, \mathbf{n}_k))$ that we approximate by $\sigma^2(\Phi(\mathbf{x}, \mathbf{n}_k))$

$$\mathcal{L}_{\text{cross}}(\mathbf{s}_1) = \frac{1}{K} \sum_{k=1}^K \frac{\left\| \Phi(\mathbf{a}_1^\top \mathbf{s}_1, \mathbf{n}_k) \right\|_2^2}{\sigma^2(\Phi(\mathbf{x}, \mathbf{n}_k))}, \quad (11)$$

We can now sum the normalized loss terms defined in equations (9)–(11) to get the final optimization problem to perform source separation

$$\widetilde{\mathbf{s}}_1 := \arg\min_{\mathbf{s}_1} \left[\mathcal{L}_{\text{data}}(\mathbf{s}_1) + \mathcal{L}_{\text{prior}}(\mathbf{s}_1) + \mathcal{L}_{\text{cross}}(\mathbf{s}_1) \right]. \quad (12)$$

Due to the delicate normalization of the three terms, we expect that further weighting of the three losses using weighting hyperparameters is not necessary. We propose to initialize the optimization problem in equation (12) with $\mathbf{s}_1 := 0$. Such choice means that $\mathbf{n} = \mathbf{x} - a_1^\top \mathbf{s}_1$ is initialized to $\mathbf{x}$, which contains crucial information on the sources, as will be explained in the next section.

We have observed that as soon as we know the statistics $\Phi(\mathbf{n})$, our algorithm retrieves the unknown statistics of the source $\Phi(\mathbf{a}_1^\top \mathbf{s}_1^*)$. In other words the algorithm successfully separates the sources in the scattering covariance space, this constitutes a convergence result, that can be proved under simplifying assumptions (see theorem 4.1). Of course, in many cases as we will see in the next section, our algorithm retrieves point estimates of $\mathbf{s}_1(t)$ that is stronger.

Theorem 4.1. *Let* $\mathbf{x} = \mathbf{a}_1^\top \mathbf{s}_1^* + \mathbf{n}$ *with* $\mathbf{s}_1$ *and* $\mathbf{n}$ *two independent processes. Let us assume we have two processes* $\widetilde{\mathbf{s}}_1$ *and* $\widetilde{\mathbf{n}}$ *with* $\mathbf{x} = \mathbf{a}_1^\top \widetilde{\mathbf{s}}_1 + \widetilde{\mathbf{n}}$.

Under the following assumptions:

(i) $\mathbf{n}$ *has a maximum entropy distribution under moment constraints* $\mathbb{E}\{\Phi(\mathbf{n})\}$

(ii) $\widetilde{\mathbf{n}}$ *has a maximum entropy distribution under moment constraints* $\mathbb{E}\{\Phi(\widetilde{\mathbf{n}})\}$

(iii) $\mathbb{E}\{\Phi(\widetilde{\mathbf{n}})\} = \mathbb{E}\{\Phi(\mathbf{n})\}$

(iv) $\widetilde{\mathbf{s}}_1$ *and* $\widetilde{\mathbf{n}}$ *are independent*

(v) *The Fourier transform* $\widehat{p}_{\mathbf{n}}$ *of the distribution* $p_{\mathbf{n}}$ *of* $\mathbf{n}$ *is non-zero everywhere.*

one has $\mathbf{n} \overset{d}{=} \widetilde{\mathbf{n}}$ *and* $\mathbf{a}_1^\top \mathbf{s}_1^* \overset{d}{=} \mathbf{a}_1^\top \widetilde{\mathbf{s}}_1$ *where the equality is on the distribution of the processes.*

Essentially, it means that when the source $\mathbf{n}$ is statistically characterized by its scattering covariance descriptors the algorithm is able to retrieve statistically the other sources. The theorem is proved and its assumptions are discussed in appendix B. This emphasizes the choice of a representation Φ that can approximate efficiently the stochastic structure of multiscale processes (Morel et al., 2022).

5. Numerical Experiments

The main goal of this paper is to derive a unsupervised approach to source separation that is applicable in domain with limited access to training data, thanks to the wavelet scattering covariance representation. To provide a quantitative analysis to the performance of our approach, we first consider a stylized synthetic example that resembles challenges of real-world data. To illustrate how our method performs in the wild, we apply our method to data recorded on Mars during the InSight mission. We aim to separate transient thermally induced microtilts, i.e., glitches (Scholz et al., 2020; Barkaoui et al., 2021), from the recorded data by the InSight lander's seismometer. The code for partially reproducing the results can be found on GitHub. Our implementation is based on the original PyTorch code for wavelet scattering covariances (Morel et al., 2022).

5.1. Stylized Example

We consider the problem of separating glitch-like signals from increments of a multifractal random work process (Bacry et al., 2001). This process is a typical non-Gaussian noise exhibiting long-range dependencies and showing bursts of activity, e.g., see Figure 12 in the appendix for several realizations of this process. The second source signal is composed of several peaks with exponentially decaying amplitude, with possibly different decay parameters on the left than on the right. To obtain synthetic observed data, we sum increments of a multifractal random walk realization, which plays the role of $\mathbf{n}$ in equation (4), with a realization of the second source. The top three images in Figure 2 are the signal of interest, secondary added signal, and the observed data, respectively.

In order to retrieve the multifractal random walk realization, we solve the optimization problem in equation (12) using the L-BFGS optimization algorithm (Liu & Nocedal, 1989) using 500 iterations. We use a training dataset of 100 realizations of increments of a multifractal random walk, $\{\mathbf{n}_k\}_{k=1}^{100}$. The architecture we use for wavelet scattering covariance computation is two-layer scattering network with $J = 8$ different octaves with $Q = 1$ wavelet per octave. We use the same scattering network architecture throughout all the numerical experiments in the paper. Given an input signal dimension of $d = 2048$, this choice of parameters yields a 174-dimensional wavelet scattering covariance space. The bottom two images in Figure 2 summarizes the results. We are able to recover the ground-truth multifractal random walk realization up to small, mostly incoherent, and seemingly random error. To see the effect of number of training realizations on the signal recovery, we repeated the above examples and used varying number of training samples. Figure 4 shows that, as expected, the signal-to-noise ratio of the recovered sources increases the more training samples we have.

We also investigate the behaviour of our source separation algorithm in case there are no additional sources present in
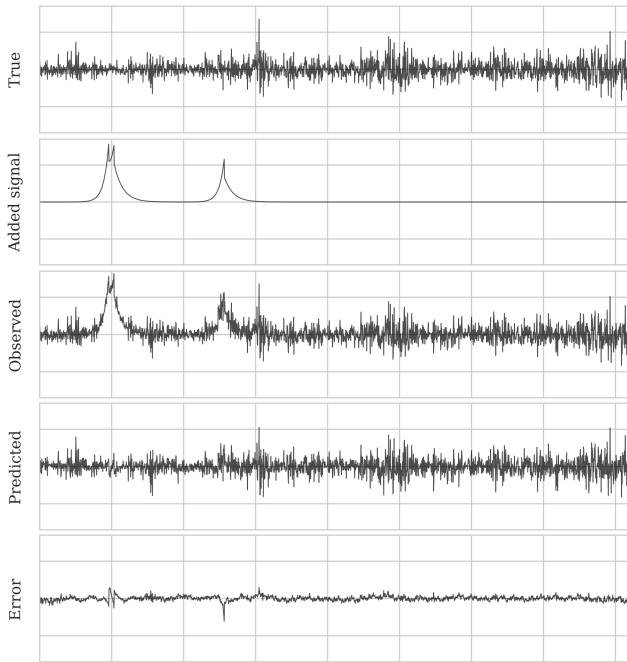
Figure 2. Unsupervised source separation applied to the multifractal random walk data. The vertical axis is the same for all the plots.
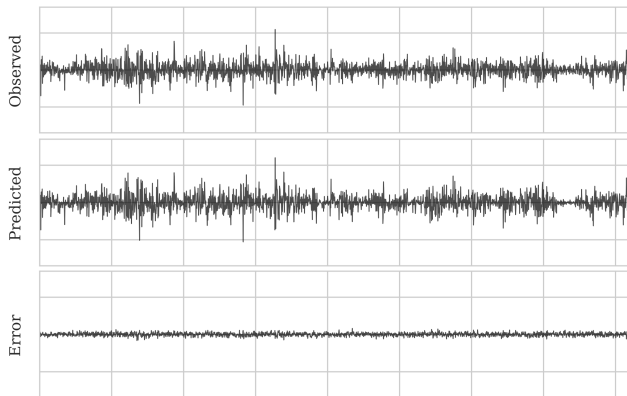

Figure 3. The behaviour when there are no sources to be removed, i.e., the observed data is a realization of the same stochastic process as the data snippets. The vertical axis is the same for all the plots.

the signal, i.e., the observed data is a realization of the same stochastic process as the data snippets $\{\mathbf{n}_k\}_{k=1}^{K}$. Ideally, the source separation algorithm should not unnecessarily remove important signals. We present the results of this experiment in Figure 3, which indeed confirms that only a negligible amount of energy has been removed from the observed data in this case. We argue that the undesired separated signal from the observed data by our method is mainly due to errors in estimating the scattering covariance statistics using a finite amount of data snippets.

To show our method can also separate sources that are not
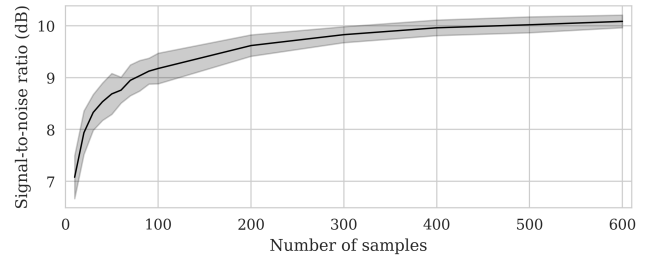
Figure 4. Signal-to-noise ratio of the predicted multifractal random walk data versus number of unsupervised samples. Shaded area indicates the 90% interval of this quantity for ten random source separation instances.
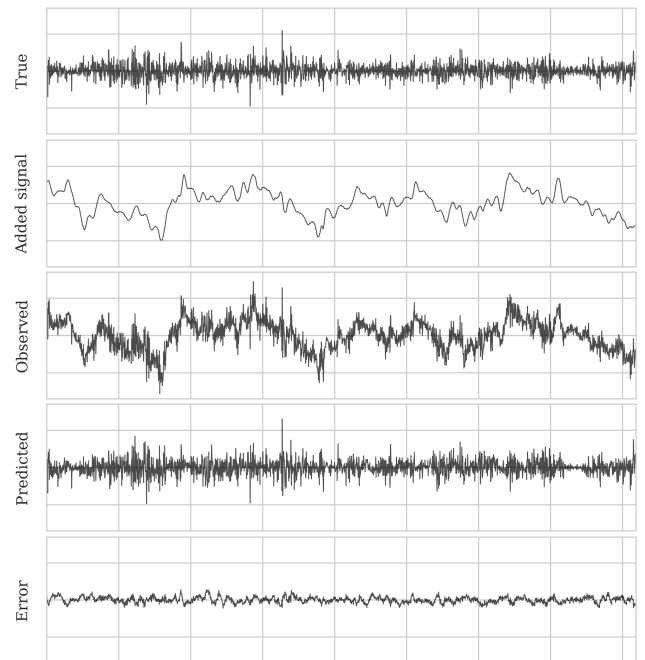

Figure 5. Unsupervised source separation applied to the multifractal random walk data with a turbulent additive signal. The vertical axis is the same for all the plots.

localized in time, we consider contaminating the multifractal random walk data with a turbulent signal (see second image from the top in Figure 5. Without any prior knowledge regarding this turbulent signal and by only using 100 realizations of increments of a multifractal random walk as training samples, we are able to recover the signal of interest with arguably low error: juxtapose the ground truth and predicted multifractal random walk realization in Figure 5. The algorithm correctly removes the low frequencies content of the turbulent jet, and makes a small, uncorrelated, random error at high frequencies. In this case the two signals having different power spectra helps disentangling them at high frequencies. In the above synthetic examples, the signal low frequencies are well separated and the algorithm in-

fers correctly the high frequencies. In the earlier example, the presence of time localized sources would facilitate the algorithm to "interpolate" the background noise knowing its scattering covariance representation. This case makes it more evident that the initialization $s_1 = 0$ informs the algorithm of the trajectory of the unknown source.

5.2. Application to Data from the InSight Mission

InSight lander's seismometer, SEIS, is exposed to heavy wind and temperature fluctuations. As a result, it is subject to background noise. Glitches are a widely occurring family of noise caused by a variety of causes (Scholz et al., 2020). These glitches often appear as one-sided pulses in seismic data and significantly affect the analysis of the data (Scholz et al., 2020). In this section we will explore the application of our proposed method in separating glitches and background noise from the recorded seismic data on Mars.

5.2.1. SEPARATING GLITCHES

We propose to consider glitches as the source of interest s_1 in the context of equation (4). To perform source separation using our technique, we need snippets of data that do not contain glitches. We select these windows of data using an existing catalog and glitches (Scholz et al., 2020) and by further eye examination to ensure no glitch contaminates our dataset. In total, we collect 50 windows of length 102.4 s during sol 187 (6 June 2019) for the U component. We show four of these windows of data in Figure 6. We perform optimization for glitch removal using the same underlying scattering network architecture as the previous example using 50 training samples and 1000 L-BFGS iterations. Figure 7 summarizes the results. The top-left image shows the raw data. Top-right image is the baseline (Scholz et al., 2020) (see Appendix C for description) prediction for the glitch signal. Finally, the bottom row (from left to right) shows our predicted deglitched data and the glitch signal separated by our approach. As confirmed by experts at the InSight team, indeed our approach has removed a glitch that the baseline has ignored (most likely due the spike right at the beginning of the glitch signal). More deglitching examples can be seen in Figures 13–16.

It is important to note that the separated glitch in our experiments may comprise some non-transient, non-seismic signals, potentially arising from atmosphere-surface interactions, as opposed to the the baseline glitch. Consequently, we anticipate the separation of these non-seismic signals in addition to the glitch when applying our approach. This results in "noisy" predicted glitches when compared to the baseline, which might be due to the the non-seismic signal. With this in mind, our approach extends the notion of glitch (as understood by the InSight team). This is one of the benefits of our unsupervised approach as the method—
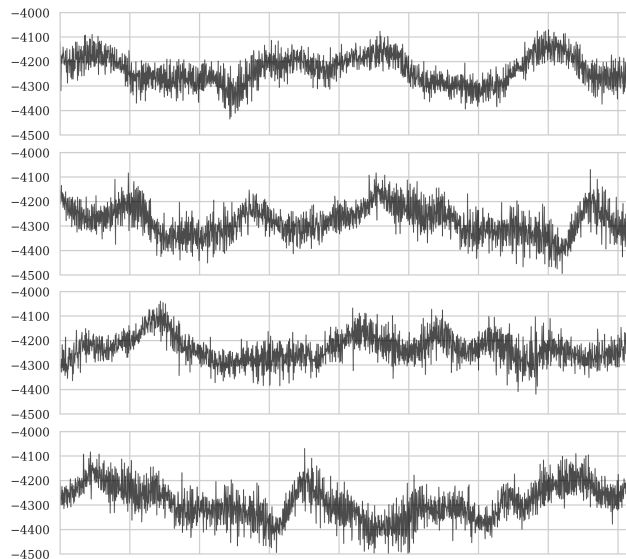

Figure 6. Glitch-free snippets of the seismic data from Mars (U component).

based on the statistics of the training data—identifies and removes events that do not seem to belong to the training data distribution..

Thanks to the interpretability of wavelet scattering covariance representations, stemming from our comprehension of scattering coefficients and covariances, we can perform a source separation quality control in domain where there is no access to ground truth source—as in our example. Figure 8 compares the power spectra of the reconstructed background noise (recorded data), a deglitched realization of the background noise and the mixed signal (observed data). It can be seen that the power spectrum of the background noise is correctly retrieved. In fact, the scattering covariance statistics, which extend the power spectrum, are correctly retrieved, which is due to the loss term in equation (6).

5.2.2. MARSQUAKE BACKGROUND NOISE SEPARATION

Marsquakes are of significant importance as they provide useful information regarding the Mars subsurface, enabling the study of Mars' interior (Knapmeyer-Endrun et al., 2021; Stähler et al., 2021; Khan et al., 2021). Recordings by the InSight lander's seismometer are susceptible to background noise and transient atmospheric signals, and here we apply our proposed unsupervised source separation approach to separate background noise from a marsquake (InSight Marsquake Service, 2023). To achieve this, we select about 30 hours of raw data (except for a detrending step)—from the U component with a 20Hz sampling rate—to fully characterize various aspects of the background noise through the wavelet scattering covariance representation. Next, we window the data and use the windows as training samples from
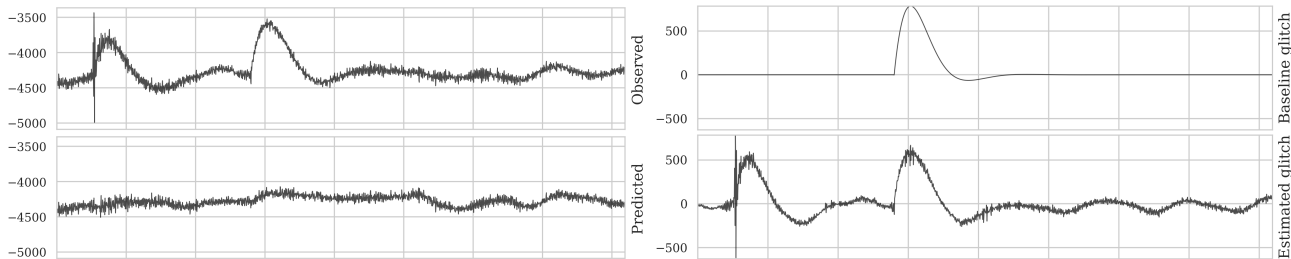
Figure 7. Unsupervised source separation for glitch removal. Juxtapose the predicted glitches on the right. Our approach is able to remove a glitch whereas the baseline approach fails to detect it.
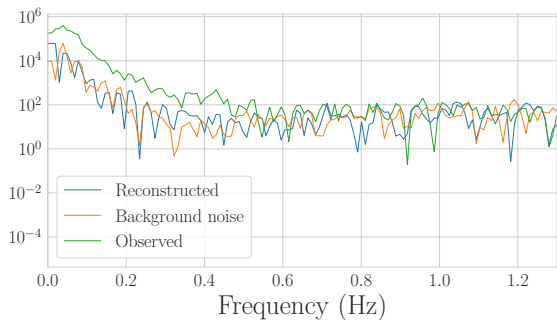

Figure 8. Power spectrum of the observed signal $\mathbf{x}$, the background noise $\mathbf{n}$ and the reconstructed background noise $\mathbf{x} - a_1^\top \widetilde{\mathbf{s}}_1$. We see that the reconstructed component statistically agrees with a Mars seismic background noise $\mathbf{n}$. The algorithm efficiently removed the low-pass component of the signal corresponding to a glitch.

background noise ($\mathbf{n}_k$ in the context of equation (4)) with the goal of retrieving the marsquake recorded at February 3, 2022 (InSight Marsquake Service, 2023).

We use the same network architecture as previous examples to setup the wavelet scattering covariance representation. We use a window size of $204.8\,\text{s}$ and solve the optimization problem in equation (12) with 200 L-BFGS iterations. The results are depicted in Figure 1. There are clearly two glitches that we have successfully separated, along with the background noise. This results is obtained merely by using 30 hours of raw data, allowing us to identify the marsquake as a separate source due to differences in wavelet scattering covariance representation.

6. Conclusions

For source separation to be effective, prior knowledge concerning unknown sources is necessary. Data-driven source separation methods extract this information from existing datasets during pretraining. In most cases, these methods require a large amount of data, which means that they are not suitable for planetary science missions. To address the challenge posed by limited data, we proposed an approach based on wavelet scattering covariances. We reaped the benefits of the inductive bias built into the scattering covariances, enabling us to obtain low-dimensional data representations that characterize a wide range of non-Gaussian properties of multiscale stochastic processes without pretraining. Using a wavelet scattering covariance space optimization problem, we were able to separate thermally induced microtilts (glitches) from data recorded by the InSight lander's seismometer with only a few glitch-free data samples. In addition, we applied the same strategy to separate marsquakes from background noise and glitches using only several hours of data with no recorded marsquake. Our approach did not require any knowledge regarding glitches or marsquakes, and proved to be more robust in separating glitches from recorded seismic data on Mars than existing techniques. An important characteristic of our approach is that it serves as an exploratory method for unsupervised learning, particularly beneficial for investigating complex and real-world datasets.

7. Acknowledgments

References

Adali, T., Levin-Schwartz, Y., and Calhoun, V. D. Multimodal data fusion using source separation: Application to medical imaging. *Proceedings of the IEEE*, 103(9): 1494–1506, 2015. doi: 10.1109/JPROC.2015.2461601.

Bacry, E., Delour, J., and Muzy, J. F. Multifractal random walk. *Phys. Rev. E*, 64:026103, Jul 2001. doi: 10.1103/PhysRevE.64.026103. URL https://link.aps.org/doi/10.1103/PhysRevE.64.026103.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

Barkaoui, S., Lognonné, P., Kawamura, T., Stutzmann, É., Seydoux, L., Maarten, V., Balestriero, R., Scholz, J.-R., Sainton, G., Plasman, M., et al. Anatomy of continuous mars seis and pressure data from unsupervised learning. *Bulletin of the Seismological Society of America*, 111(6): 2964–2981, 2021.

Barriga, E. S., Truitt, P. W., Pattichis, M. S., T'so, D., M.D., Y. H. K., M.D., R. H. K., and Soliz, P. Blind source separation in retinal videos. In Sonka, M. and Fitzpatrick, J. M. (eds.), *Medical Imaging 2003: Image Processing*, volume 5032, pp. 1591 – 1601. International Society for Optics and Photonics, SPIE, 2003. doi: 10.1117/12.481361. URL https://doi.org/10.1117/12.481361.

Beghein, C., Li, J., Weidner, E., Maguire, R., Wookey, J., Lekić, V., Lognonné, P., and Banerdt, W. Crustal anisotropy in the martian lowlands from surface waves. *Geophysical Research Letters*, 49(24):e2022GL101508, 2022. doi: https://doi.org/10.1029/2022GL101508. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL101508. e2022GL101508 2022GL101508.

Bingham, E. and Hyvärinen, A. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(01):1–8, 2000. doi: 10.1142/S0129065700000028. URL https://doi.org/10.1142/S0129065700000028. PMID: 10798706.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Bruna, J., Mallat, S., Bacry, E., Muzy, J.-F., et al. Intermittent process analysis with scattering moments. *Annals of Statistics*, 43(1):323–351, 2015.

Cardoso, J.-F. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing,*, pp. 2109–2112 vol.4, 1989. doi: 10.1109/ICASSP.1989.266878.

Cardoso, J.-F. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998. doi: 10.1109/5.720250.

Ceylan, S., Clinton, J. F., Giardini, D., Stähler, S. C., Horleston, A., Kawamura, T., Böse, M., Charalambous, C., Dahmen, N. L., van Driel, M., Durán, C., Euchner, F., Khan, A., Kim, D., Plasman, M., Scholz, J.-R., Zenhäusern, G., Beucler, E., Garcia, R. F., Kedar, S., Knapmeyer, M., Lognonné, P., Panning, M. P., Perrin, C., Pike, W. T., Stott, A. E., and Banerdt, W. B. The marsquake catalogue from insight, sols 0–1011. *Physics of the Earth and Planetary Interiors*, 333:106943, 2022. ISSN 0031-9201. doi: https://doi.org/10.1016/j.pepi.2022.106943. URL https://www.sciencedirect.com/science/article/pii/S0031920122001042.

Chevreuil, A. and Loubaton, P. Chapter 4 - blind signal separation for digital communication data. In Sidiropoulos, N. D., Gini, F., Chellappa, R., and Theodoridis, S. (eds.), *Academic Press Library in Signal Processing: Volume 2*, volume 2 of *Academic Press Library in Signal Processing*, pp. 135–186. Elsevier, 2014. doi: https://doi.org/10.1016/B978-0-12-396500-4.00004-1. URL https://www.sciencedirect.com/science/article/pii/B9780123965004000041.

Chua, J., Wang, G., and Kleijn, W. B. Convolutive blind source separation with low latency. In *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016. doi: 10.1109/IWAENC.2016.7602895.

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y. W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *arXiv preprint arXiv:2108.06209*, 2021.

Delouis, J.-M., Allys, E., Gauvrit, E., and Boulanger, F. Non-gaussian modelling and statistical denoising of planck dust polarisation full-sky maps using scattering transforms. *A&A*, 668:A122, 2022. doi: 10.1051/0004-6361/202244566. URL https://doi.org/10.1051/0004-6361/202244566.

Denton, T., Wisdom, S., and Hershey, J. R. Improving bird classification with unsupervised sound separation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 636–640, 2022. doi: 10.1109/ICASSP43922.2022.9747202.

Drude, L., Hasenklever, D., and Haeb-Umbach, R. Unsupervised training of a deep clustering model for multichannel

blind source separation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 695–699. IEEE, 2019.

Févotte, C., Bertin, N., and Durrieu, J.-L. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Computation*, 21(3):793–830, 03 2009. ISSN 0899-7667. doi: 10.1162/neco.2008.04-08-771. URL https://doi.org/10.1162/neco.2008.04-08-771.

Gatys, L., Ecker, A. S., and Bethge, M. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

Gay, S. L. and Benesty, J. *Acoustic signal processing for telecommunication*, volume 551. Springer Science & Business Media, 2012.

Giardini, D., Lognonné, P., Banerdt, W. B., Pike, W. T., Christensen, U., Ceylan, S., Clinton, J. F., van Driel, M., Stähler, S. C., Böse, M., et al. The seismicity of mars. *Nature Geoscience*, 13(3):205–212, 2020.

Golombek, M., Warner, N., Grant, J., Hauber, E., Ansan, V., Weitz, C., Williams, N., Charalambous, C., Wilson, S., DeMott, A., et al. Geology of the insight landing site on mars. *Nature communications*, 11(1):1–11, 2020.

Grais, E. M., Sen, M. U., and Erdogan, H. Deep neural networks for single channel source separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3734–3738, 2014. doi: 10.1109/ICASSP.2014.6854299.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Hasan, A. M., Melli, A., Wahid, K. A., and Babyn, P. Denoising low-dose ct images using multiframe blind source separation and block matching filter. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2(4):279–287, 2018. doi: 10.1109/TRPMS.2018.2810221.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35. IEEE Press, 2016. doi: 10.1109/ICASSP.2016.7471631. URL https://doi.org/10.1109/ICASSP.2016.7471631.

Horleston, A. C., Clinton, J. F., Ceylan, S., Giardini, D., Charalambous, C., Irving, J. C. E., Lognonné, P., Stähler, S. C., Zenhäusern, G., Dahmen, N. L., Duran, C., Kawamura, T., Khan, A., Kim, D., Plasman, M., Euchner, F., Beghein, C., Beucler, E., Huang, Q., Knapmeyer, M., Knapmeyer-Endrun, B., Lekić, V., Li, J., Perrin, C., Schimmel, M., Schmerr, N. C., Stott, A. E., Stutzmann, E., Teanby, N. A., Xu, Z., Panning, M., and Banerdt, W. B. The Far Side of Mars: Two Distant Marsquakes Detected by InSight. *The Seismic Record*, 2(2):88–99, 04 2022. ISSN 2694-4006. doi: 10.1785/0320220007. URL https://doi.org/10.1785/0320220007.

Ibrahim, A. and Sacchi, M. D. Simultaneous source separation using a robust radon transform. *GEOPHYSICS*, 79(1):V1–V11, 2014. doi: 10.1190/geo2013-0168.1. URL https://doi.org/10.1190/geo2013-0168.1.

InSight Marsquake Service. Mars seismic catalogue, insight mission; v13 2023-01-01, 2023. URL https://www.insight.ethz.ch/seismicity/catalog/v13.

Jang, G.-J. and Lee, T.-W. A maximum likelihood approach to single-channel source separation. *The Journal of Machine Learning Research*, 4:1365–1392, 2003.

Jeffrey, N., Boulanger, F., Wandelt, B. D., Regaldo-Saint Blancard, B., Allys, E., and Levrier, F. Single frequency cmb b-mode inference with realistic foregrounds from a single training image. *Monthly Notices of the Royal Astronomical Society: Letters*, 510(1):L1–L6, 2022.

Jutten, C. and Herault, J. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(91)90079-X. URL https://www.sciencedirect.com/science/article/pii/016516849190079X.

Jutten, C., Babaie-Zadeh, M., and Hosseini, S. Three easy ways for separating nonlinear mixtures? *Signal Processing*, 84(2):217–229, 2004. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2003.10.011. URL https://www.sciencedirect.com/science/article/pii/S0165168403002767. Special Section on Independent Component Analysis and Beyond.

Kameoka, H., Li, L., Inoue, S., and Makino, S. Supervised determined source separation with multichannel variational autoencoder. *Neural Computation*, 31(9):1891–1914, 2019. doi: 10.1162/neco_a_01217.

Ke, S., Hu, R., Wang, X., Wu, T., Li, G., and Wang, Z. Single channel multi-speaker speech separation based on quantized ratio mask and residual network. *Multimedia Tools Appl.*, 79(43–44):

32225–32241, nov 2020. ISSN 1380-7501. doi: 10.1007/s11042-020-09419-y. URL https://doi.org/10.1007/s11042-020-09419-y.

Khan, A., Ceylan, S., van Driel, M., Giardini, D., Lognonné, P., Samuel, H., Schmerr, N. C., Stähler, S. C., Duran, A. C., Huang, Q., Kim, D., Broquet, A., Charalambous, C., Clinton, J. F., Davis, P. M., Drilleau, M., Karakostas, F., Lekic, V., McLennan, S. M., Maguire, R. R., Michaut, C., Panning, M. P., Pike, W. T., Pinot, B., Plasman, M., Scholz, J.-R., Widmer-Schnidrig, R., Spohn, T., Smrekar, S. E., and Banerdt, W. B. Upper mantle structure of mars from insight seismic data. *Science*, 373(6553):434–438, 2021. doi: 10.1126/science.abf2966. URL https://www.science.org/doi/abs/10.1126/science.abf2966.

Khosravy, M., Gupta, N., Patel, N., Dey, N., Nitta, N., and Babaguchi, N. Probabilistic stone's blind source separation with application to channel estimation and multi-node identification in mimo iot green communication and multimedia systems. *Computer Communications*, 157:423–433, 2020. ISSN 0140-3664. doi: https://doi.org/10.1016/j.comcom.2020.04.042. URL https://www.sciencedirect.com/science/article/pii/S0140366420302516.

Knapmeyer-Endrun, B. and Kawamura, T. Nasa's insight mission on mars—first glimpses of the planet's interior from seismology. *Nature Communications*, 11(1):1–4, 2020.

Knapmeyer-Endrun, B., Panning, M. P., Bissig, F., Joshi, R., Khan, A., Kim, D., Lekić, V., Tauzin, B., Tharimena, S., Plasman, M., Compaire, N., Garcia, R. F., Margerin, L., Schimmel, M., Éléonore Stutzmann, Schmerr, N., Bozdağ, E., Plesa, A.-C., Wieczorek, M. A., Broquet, A., Antonangeli, D., McLennan, S. M., Samuel, H., Michaut, C., Pan, L., Smrekar, S. E., Johnson, C. L., Brinkman, N., Mittelholz, A., Rivoldini, A., Davis, P. M., Lognonné, P., Pinot, B., Scholz, J.-R., Stähler, S., Knapmeyer, M., van Driel, M., Giardini, D., and Banerdt, W. B. Thickness and structure of the martian crust from insight seismic data. *Science*, 373(6553):438–443, 2021. doi: 10.1126/science.abf8966. URL https://www.science.org/doi/abs/10.1126/science.abf8966.

Kumar, R., Wason, H., and Herrmann, F. J. Source separation for simultaneous towed-streamer marine acquisition — a compressed sensing approach. *GEOPHYSICS*, 80(6):WD73–WD88, 2015. doi: 10.1190/geo2015-0108.1. URL https://doi.org/10.1190/geo2015-0108.1.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, Aug 1989. ISSN 1436-4646.

doi: 10.1007/BF01589116. URL https://doi.org/10.1007/BF01589116.

Liu, S., Mallol-Ragolta, A., Parada-Cabaleiro, E., Qian, K., Jing, X., Kathan, A., Hu, B., and Schuller, B. W. Audio self-supervised learning: A survey. *Patterns*, 3 (12):100616, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100616.

Lognonné, P., Banerdt, W. B., Pike, W., Giardini, D., Christensen, U., Garcia, R. F., Kawamura, T., Kedar, S., Knapmeyer-Endrun, B., Margerin, L., et al. Constraints on the shallow elastic and anelastic structure of mars from insight seismic data. *Nature Geoscience*, 13(3):213–220, 2020.

Lorenz, R. D., Spiga, A., Lognonné, P., Plasman, M., Newman, C. E., and Charalambous, C. The whirlwinds of elysium: A catalog and meteorological characteristics of "dust devil" vortices observed by insight on mars. *Icarus*, 355:114119, 2021. ISSN 0019-1035. doi: https://doi.org/10.1016/j.icarus.2020.114119. URL https://www.sciencedirect.com/science/article/pii/S0019103520304632.

Morel, R., Rochette, G., Leonarduzzi, R., Bouchaud, J.-P., and Mallat, S. Scale dependencies and self-similarity through wavelet scattering covariance. *arXiv preprint arXiv:2204.10177*, 2022.

Nandi, A. and Zarzoso, V. Fourth-order cumulant based blind source separation. *IEEE Signal Processing Letters*, 3(12):312–314, 1996. doi: 10.1109/97.544786.

Neri, J., Badeau, R., and Depalle, P. Unsupervised blind source separation with variational auto-encoders. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 311–315, 2021. doi: 10.23919/EUSIPCO54536.2021.9616154.

Panning, M. P., Banerdt, W. B., Beghein, C., Carrasco, S., Ceylan, S., Clinton, J. F., Davis, P., Drilleau, M., Giardini, D., Khan, A., Kim, D., Knapmeyer-Endrun, B., Li, J., Lognonné, P., Stähler, S. C., and Zenhäusern, G. Locating the largest event observed on mars with multi-orbit surface waves. *Geophysical Research Letters*, 50(1):e2022GL101270, 2023. doi: https://doi.org/10.1029/2022GL101270. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL101270. e2022GL101270 2022GL101270.

Parra, L. and Sajda, P. Blind source separation via generalized eigenvalue decomposition. *J. Mach. Learn. Res.*, 4 (null):1261–1269, dec 2003. ISSN 1532-4435.

Pedersen, M. S., Larsen, J., Kjems, U., and Parra, L. C. *Convolutive Blind Source Separation Methods*, pp. 1065–1094. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-49127-9. doi: 10.1007/978-3-540-49127-9_52. URL https://doi.org/10.1007/978-3-540-49127-9_52.

Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., and Dupoux, E. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. *arXiv preprint arXiv:2104.00355*, 2021.

Regaldo-Saint Blancard, B., Allys, E., Boulanger, F., Levrier, F., and Jeffrey, N. A new approach for the statistical denoising of planck interstellar dust polarization data. *Astronomy & Astrophysics*, 649:L18, 2021.

Rodríguez, Á. B., Balestriero, R., De Angelis, S., Benítez, M. C., Zuccarello, L., Baraniuk, R., Ibáñez, J. M., and Maarten, V. Recurrent scattering network detects metastable behavior in polyphonic seismo-volcanic signals for volcano eruption forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–23, 2021.

Scholz, J.-R., Widmer-Schnidrig, R., Davis, P., Lognonné, P., Pinot, B., Garcia, R. F., Hurst, K., Pou, L., Nimmo, F., Barkaoui, S., de Raucourt, S., Knapmeyer-Endrun, B., Knapmeyer, M., Orhand-Mainsant, G., Compaire, N., Cuvier, A., Beucler, E., Bonnin, M., Joshi, R., Sainton, G., Stutzmann, E., Schimmel, M., Horleston, A., Böse, M., Ceylan, S., Clinton, J., van Driel, M., Kawamura, T., Khan, A., Stähler, S. C., Giardini, D., Charalambous, C., Stott, A. E., Pike, W. T., Christensen, U. R., and Banerdt, W. B. Detection, analysis, and removal of glitches from insight's seismic data from mars. *Earth and Space Science*, 7(11):e2020EA001317, 2020. doi: https://doi.org/10.1029/2020EA001317.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. d., Campillo, M., and Baraniuk, R. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, 11 (1):3972, Aug 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17841-x. URL https://doi.org/10.1038/s41467-020-17841-x.

Siahkoohi, A., Chinen, M., Denton, T., Kleijn, W. B., and Skoglund, J. Ultra-low-bitrate speech coding with pretrained transformers. In *Proc. Interspeech 2022*, pp. 4421–4425, 2022. doi: 10.21437/Interspeech.2022-10988.

Stott, A. E., Garcia, R. F., Chédozeau, A., Spiga, A., Murdoch, N., Pinot, B., Mimoun, D., Charalambous, C., Horleston, A., King, S. D., Kawamura, T., Dahmen, N., Barkaoui, S., Lognonné, P., and Banerdt, W. B. Machine learning and marsquakes: a tool to predict atmospheric-seismic noise for the NASA InSight mission. *Geophysical Journal International*, 233(2):978–998, 11 2022. ISSN 0956-540X. doi: 10.1093/gji/ggac464. URL https://doi.org/10.1093/gji/ggac464.

Stähler, S. C., Khan, A., Banerdt, W. B., Lognonné, P., Giardini, D., Ceylan, S., Drilleau, M., Duran, A. C., Garcia, R. F., Huang, Q., Kim, D., Lekic, V., Samuel, H., Schimmel, M., Schmerr, N., Sollberger, D., Éléonore Stutzmann, Xu, Z., Antonangeli, D., Charalambous, C., Davis, P. M., Irving, J. C. E., Kawamura, T., Knapmeyer, M., Maguire, R., Marusiak, A. G., Panning, M. P., Perrin, C., Plesa, A.-C., Rivoldini, A., Schmelzbach, C., Zenhäusern, G., Éric Beucler, Clinton, J., Dahmen, N., van Driel, M., Gudkova, T., Horleston, A., Pike, W. T., Plasman, M., and Smrekar, S. E. Seismic detection of the martian core. *Science*, 373(6553):443–448, 2021. doi: 10.1126/science.abi7730. URL https://www.science.org/doi/abs/10.1126/science.abi7730.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wang, D. and Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., and Hershey, J. Unsupervised sound separation using mixture invariant training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3846–3857. Curran Associates, Inc., 2020.

Xu, W., Zhu, Q., and Zhao, L. GlitchNet: A Glitch Detection and Removal System for SEIS Records Based on Deep Learning. *Seismological Research Letters*, 93(5):2804–2817, 07 2022. ISSN 0895-0695. doi: 10.1785/0220210361. URL https://doi.org/10.1785/0220210361.

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

A. Wavelet Scattering Covariance: Background Information

A.1. Wavelet Filters

A wavelet $\psi(t)$ has a fast decay away from $t = 0$, polynomial or exponential for example, and a zero-average $\int \psi(t) \, dt = 0$. We normalize $\int |\psi(t)| \, dt = 1$. The wavelet transform computes the variations of a signal x at each dyadic scale 2^j with

$$\mathbf{W}\mathbf{x}(t, j) = \mathbf{x} \star \psi_j(t) \text{ where } \psi_j(t) = 2^{-j}\psi(2^{-j}t).$$

We use a complex wavelet ψ having a Fourier transform $\widehat{\psi}(\omega) = \int \psi(t) \, e^{-i\omega t} \, dt$ which is real, and whose energy is mostly concentrated at frequencies $\omega \in [\pi, 2\pi]$. It results that $\widehat{\psi}_j(\omega) = \widehat{\psi}(2^j\omega)$ is non-negligible mostly in $\omega \in [2^{-j}\pi, 2^{-j+1}\pi]$.
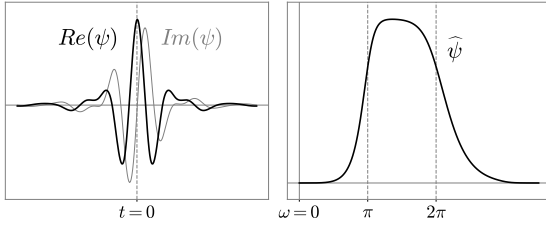
Figure 9. Left: complex Battle-Lemarié wavelet $\psi(t)$ as a function of t. Right: Fourier transform $\widehat{\psi}(\omega)$ as a function of ω.

We impose that the wavelet ψ satisfies the following energy conservation law called Littlewood-Paley equality

$$\forall \omega > 0 \ , \quad \sum_{j=-\infty}^{+\infty} |\widehat{\psi}(2^j\omega)|^2 = 1. \quad (13)$$

A Battle-Lemarié wavelet, see Figure 9, is an example of such wavelet. The wavelet transform is computed up to a largest scale 2^J which is smaller than the signal size d. The signal lower frequencies in $[-2^{-J}\pi, 2^{-J}\pi]$ are captured by a low-pass filter $\varphi_J(t)$ whose Fourier transform is

$$\widehat{\varphi}_J(\omega) = \Big(\sum_{j=J+1}^{+\infty} |\widehat{\psi}(2^j\omega)|^2 \Big)^{1/2}. \quad (14)$$

One can verify that it has a unit integral $\int \varphi_J(t) \, dt = 1$. To simplify notations, we write this low-pass filter as a last scale wavelet $\psi_{J+1} = \varphi_J$, and $\mathbf{W}\mathbf{x}(t, J+1) = \mathbf{x} \star \psi_{J+1}(t)$. By applying the Parseval formula, we derive from (13) that for all $\mathbf{x}$ with $\|\mathbf{x}\|^2 = \int |\mathbf{x}(t)|^2 \, dt < \infty$

$$\|\mathbf{W}\mathbf{x}\|^2 = \sum_{j=-\infty}^{J+1} \|\mathbf{x} \star \psi_j\|^2 = \|\mathbf{x}\|^2.$$

The wavelet transform $\mathbf{W}$ preserves the norm and is therefore invertible, with a stable inverse.

A.2. Scattering Network Architecture

A scattering network is a convolutional neural network with wavelet filters. In this paper we choose a simple two-layer architecture with modulus non-linearity:

$$S(\mathbf{x}) := \begin{bmatrix} \mathbf{W}\mathbf{x} \\ \mathbf{W}|\mathbf{W}\mathbf{x}| \end{bmatrix}.$$

The wavelet operator $\mathbf{W}$ is the same at the two layers, it uses $J = 8$ predefined Battle-Lemarié complex wavelets that are dilated from the same mother wavelet by powers of 2, yielding one wavelet per octave.

The first layer extracts $J + 1$ scale channels $\mathbf{x} \star \psi_j(t)$, corresponding to J band-pass and 1 low-pass wavelet filters. The second layer is $\mathbf{W}|\mathbf{W}\mathbf{x}|(t; j_1, j_2) = |\mathbf{x} \star \psi_{j_1}| \star \psi_{j_2}(t)$. It is non-negligible only if $j_1 < j_2$. Indeed, the Fourier transform of $|X \star \psi_{j_1}|$ is mostly concentrated in $[-2^{-j_1}\pi, 2^{-j_1}\pi]$. If $j_2 \leq j_1$ then it does not intersect the frequency interval $[2^{-j_2}\pi, 2^{-j_2+1}\pi]$ where the energy of $\widehat{\psi}_{j_2}$ is mostly concentrated, in which case $S\mathbf{x}(t; j_1, j_2) \approx 0$.

Instead of the modulus $|\cdot|$ we could use another non-linearity that preserves the complex phase, however it does not improve significantly the results in this paper.

A.3. Scattering Covariance Dashboard

The wavelet scattering covariance $\Phi(\mathbf{x})$ (3) contains four types of coefficients $\Phi(\mathbf{x}) = \big(\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \Phi_3(\mathbf{x}), \Phi_4(\mathbf{x})\big)$. The first family provides J order 1 moment estimators, corresponding to wavelet sparsity coefficients

$$\Phi_1(\mathbf{x})[j] = \text{Ave} \, |\mathbf{x} \star \psi_j(t)|. \quad (15)$$

The $J + 1$ second order wavelet spectrum associated to x are computed by

$$\Phi_2(\mathbf{x})[j] = \text{Ave} \, \big(|\mathbf{x} \star \psi_j(t)|^2\big). \quad (16)$$

There are $J(J+1)/2$ wavelet phase-modulus correlation coefficients for $a > 0$,

$$\Phi_3(\mathbf{x})[j; a] = \text{Ave} \, \big(\mathbf{x} \star \psi_j(t) \, |\mathbf{x} \star \psi_{j-a}(t)|\big). \quad (17)$$

Finally, in total the scattering covariance includes $J(J+1)(J+2)/6$ scattering modulus coefficients for $a \geq 0$ and $b < 0$,

$$\Phi_4(\mathbf{x})[j; a, b] = \text{Ave} \, \big(|\mathbf{x}\star\psi_j|\star\psi_{j-b}(t) \, |\mathbf{x}\star\psi_{j-a}|\star\psi_{j-b}^*(t)\big). \quad (18)$$

These coefficients extend the standard wavelet power spectrum $\Phi_2(\mathbf{x})$. After appropriate normalization and reduction that we describe below, scattering covariances can be visualized as a dashboard that displays non-Gaussian properties of $\mathbf{x}$, which is shown for example in Figures 10 and 11.
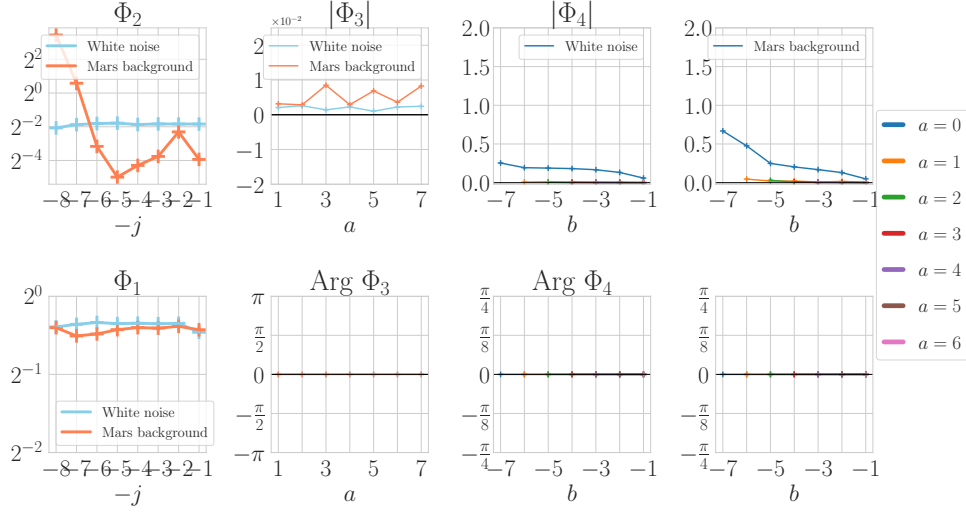
Figure 10. Scattering covariance visualization of the Mars background noise (no glitch) compared with a white noise. Estimation is performed on the same amount of data.
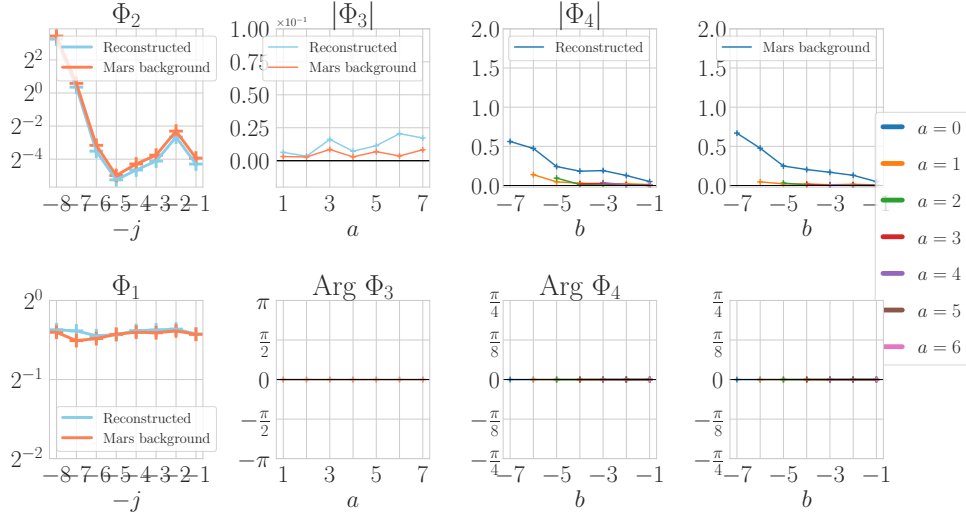

Figure 11. Scattering covariance visualization of the reconstructed Mars background noise compared with a true Mars background noise. This plots shows that beyond the wavelet power spectrum, other non-Gaussian properties of the background noise such as sparsity, long-range correlations match, up to a estimation error.

The power spectrum $\Phi_2(x)$ is plotted in a standard way, it is the energy of the scale channels of $\mathbf{x} \star \psi_j(t)$. This energy affects the other coefficients $\Phi_1(\mathbf{x}), \Phi_3(\mathbf{x}), \Phi_4(\mathbf{x})$. To deduct this influence, we normalize these coefficients by the power spectrum, $\Phi_1(\mathbf{x})[j]/\sqrt{\Phi_2(\mathbf{x})[j]}$, $\Phi_3(\mathbf{x})[j;a]/\sqrt{\Phi_2(\mathbf{x})[j]\Phi_2(\mathbf{x})[j-a]}$ and $\Phi_4(\mathbf{x})[j;a,b]/\sqrt{\Phi_2(\mathbf{x})[j]\Phi_2(\mathbf{x})[j-a]}$. Finally, we average $\Phi_3(x)$ and $\Phi_4(x)$ on j, in order to plot scaling invariant quantities, which reduces the number of coefficient to visualize (Morel et al., 2022).

B. Source Separation Guarantees

We prove theorem 4.1, discuss its assumptions for the deglitching example applied to data from Mars, and show how our implementation relates to these assumptions. For sake of simplicity we take $a_1 = 1$.

Proof. Part I. One can prove that there exists a unique process $\mathbf{n}$ that maximises entropy under moment constraint $\mathbb{E}\{\Phi(\mathbf{n})\}$, its distribution takes the form $p_\mathbf{n}(\cdot) = Z_{\boldsymbol{\theta}}^{-1}e^{-\boldsymbol{\theta}^\top \Phi(\cdot)}$ for certain Lagrange multipliers $\boldsymbol{\theta} \in \mathbb{R}^M$ where M is the dimension of Φ. Assumptions (i), (ii), (iii)

15

imply that $\mathbf{n}$ and $\widetilde{\mathbf{n}}$ are the same unique process, meaning $p_\mathbf{n} = p_{\widetilde{\mathbf{n}}}$.

Part II. Due to the independence of $\mathbf{s}_1^*, \mathbf{n}$ and $\widetilde{\mathbf{s}}_1, \widetilde{\mathbf{n}}$ (iv) we have $p_\mathbf{x} = p_{\mathbf{s}_1^*} \star p_\mathbf{n}$ and $p_\mathbf{x} = p_{\widetilde{\mathbf{s}}_1} \star p_{\widetilde{\mathbf{n}}}$. Since $p_{\widetilde{\mathbf{n}}} = p_\mathbf{n}$ we get $p_{\mathbf{s}_1^*} \star p_\mathbf{n} = p_{\widetilde{\mathbf{s}}_1} \star p_{\widetilde{\mathbf{n}}}$. This is a measure deconvolution problem. Taking the Fourier transform on measures yields

$$(\widehat{p}_{\mathbf{s}_1^*} - \widehat{p}_{\widetilde{\mathbf{s}}_1})\, \widehat{p}_\mathbf{n} = 0.$$

Under assumption (v) we get $p_{\widetilde{\mathbf{s}}_1} = p_{\mathbf{s}_1^*}$, which proves the theorem. $\qquad\square$

Assumption (i) is the main assumption. It implies that the processes $\mathbf{n}$ is fully determined by the values $\mathbb{E}\{\Phi(\mathbf{n})\}$, since there is a unique distribution satisfying (i). A maximum entropy process $\mathbf{n}$ under correlation constraints $\mathbb{E}\{\mathbf{n}\mathbf{n}^\top\}$ is a Gaussian process. A wavelet Scattering Covariance captures non-linear correlations, assumption (i) tells us that process $\mathbf{n}$ is a non-Gaussian noise fully characterized by $\mathbb{E}\{\Phi(\mathbf{n})\}$. Now, the Scattering Covariance $\mathbb{E}\{\Phi(\mathbf{n})\}$ was shown to characterize a wide range of non-Gaussian noises ((Morel et al., 2022)). In our case, the Mars seismic background noise $\mathbf{n}$ may not be fully characterized by its Scattering Covariance $\mathbb{E}\{\Phi(\mathbf{n})\}$, so that assumption (i) is only verified approximately, depending on the descriptive power of the representation $\mathbb{E}\{\Phi(\mathbf{n})\}$ for $\mathbf{n}$.

Assumption (ii) is approximately verified, requiring the entropy of $\mathbf{x}$ to be close to the entropy of $\mathbf{n}$, which is typically the case of time-localized signals such as glitch, of comparable amplitude than $\mathbf{n}$. The gradient descent algorithm implements (ii), reconstructed $\widetilde{\mathbf{n}}$ is initialized to $\mathbf{x}$ and is updated until $\Phi(\mathbf{x})$ matches the $\Phi(\mathbf{n}_k)$.

Assumption (iii) is imposed through the loss term $\mathcal{L}_{\text{prior}}$, up to estimation error of $\Phi(\mathbf{n})$ on a finite number of realizations.

Assumption (iv) relates to the loss term $\mathcal{L}_{\text{cross}}$ that imposes statistical independence up to the cross-Scattering Covariance.

Assumption (v) is a technical assumption satisfied for a Gaussian noise $\mathbf{n}$ for which the Fourier transform of $p_\mathbf{n}$ is a Gaussian. A non-Gaussian noise $\mathbf{n}$ satisfying (i) has a distribution of the form $p_\mathbf{n}(\cdot) = Z_{\boldsymbol{\theta}}^{-1} e^{-\boldsymbol{\theta}^\top \Phi(\cdot)}$. Apart from the coefficients $\text{Ave}(S(\mathbf{n}))$, the scattering covariance Φ is quadratic in $\mathbf{n}$, thus we may assume (v) is still satisfied.

C. Baseline Method

The glitch detection algorithm that we use as baseline is developed by Scholz et al. (2020) and consists of several processing steps applied to seismic data:

- Decimation: The data is downsampled to a uniform

rate of two samples per second to ensure consistent parameterization and improve computational efficiency;

- Deconvolution and band-pass filtering: Instrument response is removed from each component, transforming the data into acceleration. Additional band-pass filtering is also applied to highlight the significant features of acceleration;

- Time derivative calculation: The time derivative of the filtered acceleration data is computed, resulting in acceleration steps becoming impulse-like signals;

- Glitch detection: A constant threshold is applied to the time derivative to identify glitches. A window length is introduced to avoid false triggers on subsequent samples that are part of the same glitch event, serving as a safeguard against spurious detections.

After glitch detection, removal is based on obtaining a model (template) for the glitch signatures, followed by a separation techniques that assumes the observed data as a linear combination of the glitch and the glitch spike. To characterize each detected glitch, a glitch model is employed, consisting of three parameters: an amplitude scaling factor, an offset, and a linear trend parameter. The modeling process entails solving a nonlinear least squares data fitting problem to determine these parameters. Subsequently, the deglitched data is obtained by subtracting the fitted glitch (excluding the offset and linear trend) from the original data.

In comparison to our approach, the glitch modeling step in the mentioned method could be a significant limitation. Unlike their method, we do not make any assumptions about the functional form of the glitch or the unknown source. Instead, we focus on learning the wavelet scattering covariance statistics of the background noise. This allows us to overcome the potential limitations associated with explicitly modeling the glitches.

D. Multifractal Random Walk Realizations

Figure 12 shows realizations of the multifractal random walk process used in the stylized example.

E. Additional Glitch Separation Results

Here we provide more results regarding separating glitches from the seismic data recorded during the NASA InSight mission. Figures 13–16 provide glitch removal results for a more diverse set of glitches using the same setup as described in section 5.2.1.

We provide more comprehensive deglitching results by applying our approach to perform glitch separation on the U component for the nighttime (17:08–00:55 LMST) during
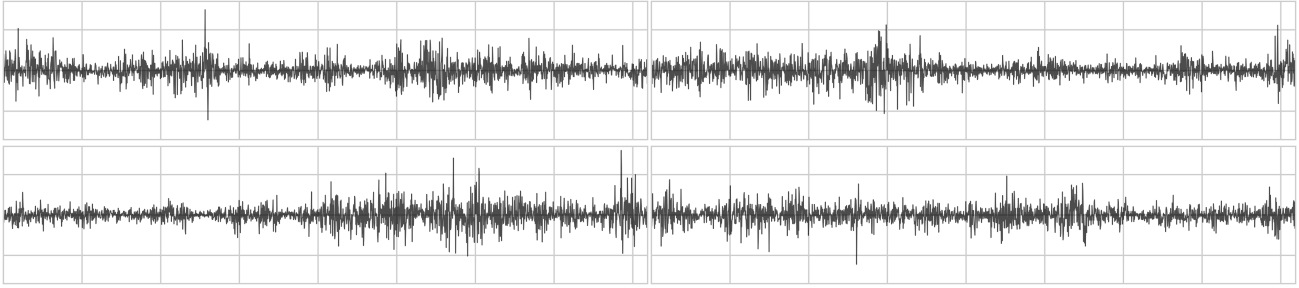
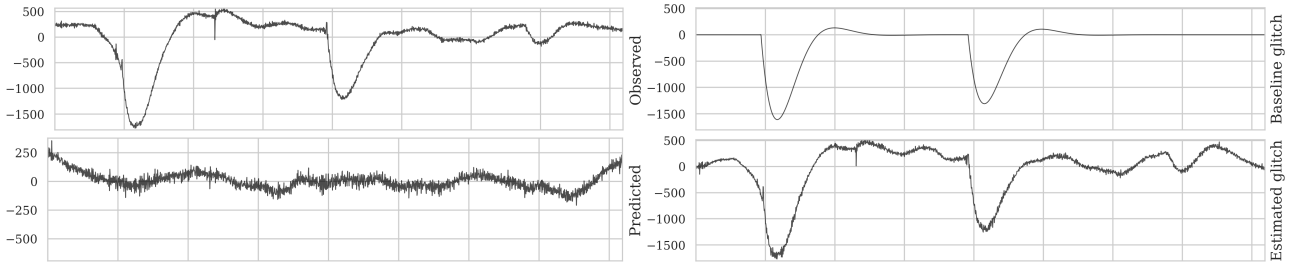*Figure 12.* Realizations of increments of the multifractal random walk process.

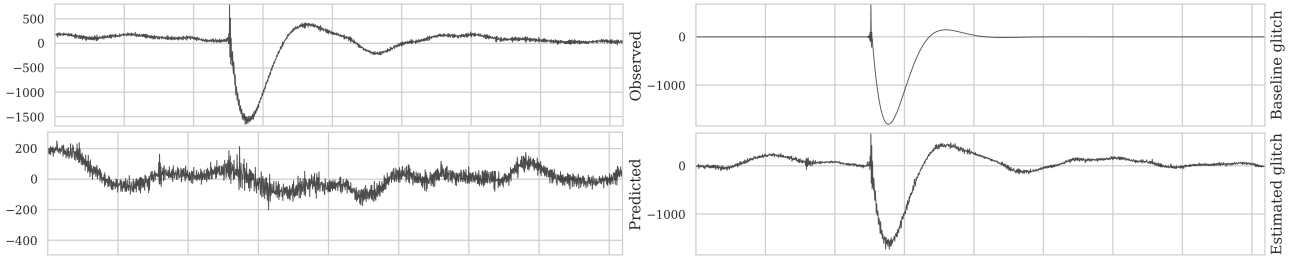*Figure 13.* Unsupervised source separation for glitch removal.

Figure 14. Unsupervised source separation for glitch removal.
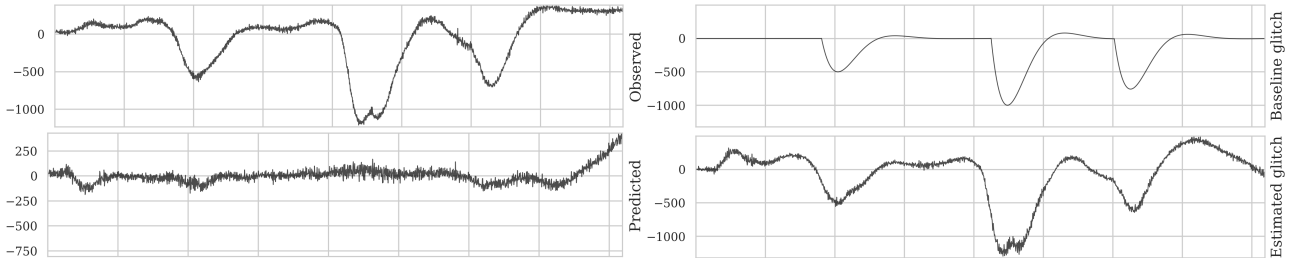

Figure 15. Unsupervised source separation for glitch removal.

sol 187 (June 6, 2019), as the glitches during the day are often obscured by daytime noise. We used a set of 50 snippets with window size of $204.8\,\text{s}$ and solved the source separation optimization problem using 200 L-BFGS iterations.

Our results indicate that the baseline method appears to overlook several anomalies in the U component that we believe to be glitches. In contrast, our method not only detects all the glitches identified by the baseline method, but it

also recognizes a significant number of additional glitches. Although it is true that our method appears to detect more glitches than the baseline, we must recognize that the baseline is the only dependable reference for identifying glitches and further verification by InSight experts is necessary to confirm the legitimacy of the identified events as glitches.
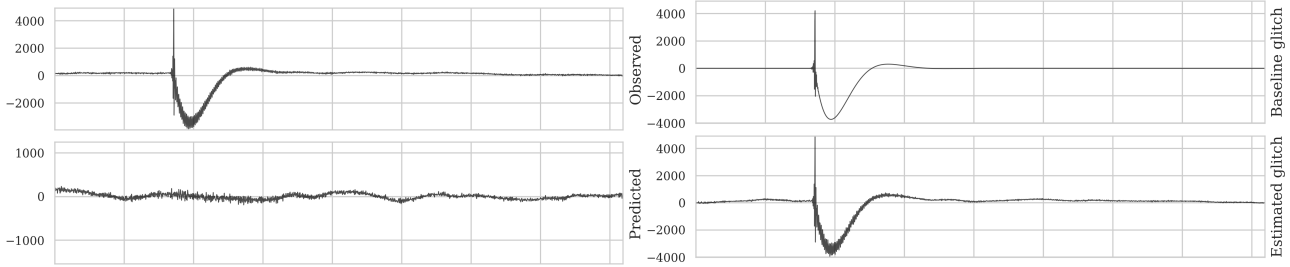
Figure 16. Unsupervised source separation for glitch removal.
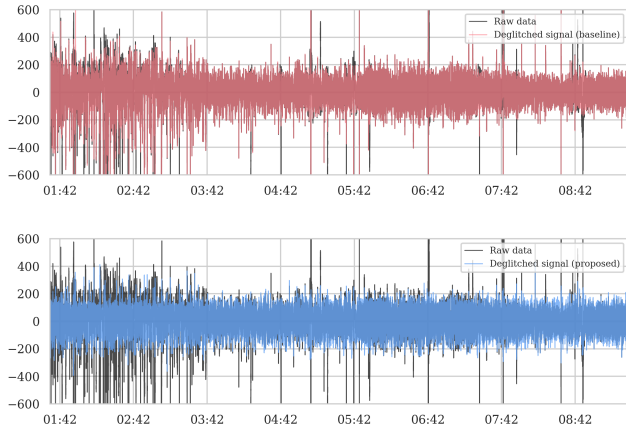

Figure 17. Unsupervised separation of glitches from seismic data recorded during sol 187 (June 6, 2019) from 17:08 to 00:55 Martian local time (the horizontal axis is in UTC time zone). The raw data is depicted in black, with the predicted deglitched data overlaid, represented by the baseline method in red and the proposed method in blue. The high-amplitude "spikes" observed in the raw waveform correspond to glitches. A successful deglitching outcome should exclude these spikes. Our deglitching results effectively separate a significant number of these high-amplitude events, whereas the baseline method fails to address a considerable portion of them.

F. Additional Marsquake Background Noise Separation Results

We present additional results on the separation of marsquake background noise and glitches, showcasing different marsquake characteristics. The first example pertains to a marsquake recorded on January 2, 2022 (InSight Marsquake Service, 2023). This particular marsquake exhibits a larger amplitude and a longer coda wave compared to the one presented in Figure 1. Although the background noise appears negligible and is not readily visible in the raw waveform, this provides an opportunity to demonstrate the effectiveness of our unsupervised source separation method when one source (the marsquake in this case) dominates in amplitude.
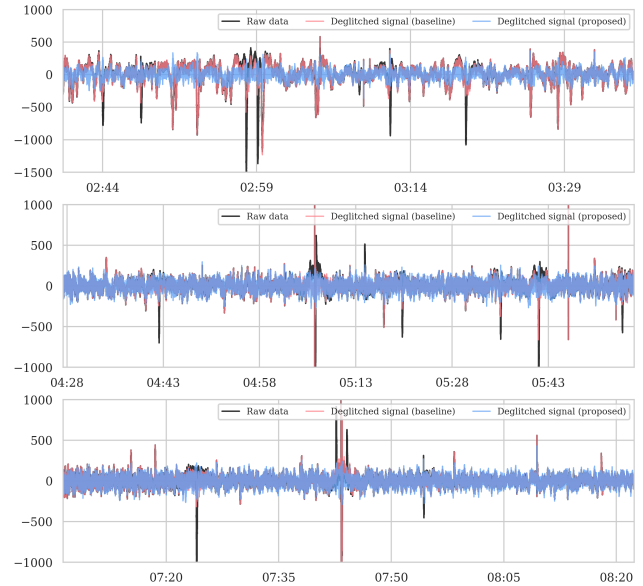

Figure 18. Three zoomed-in time intervals from Figure 17 to facilitate a detailed performance comparison between the baseline (red) and the proposed deglitching results. Both outcomes are overlaid on the raw waveform shown in black. The glitches manifest as high-amplitude one-sided pulses in the raw waveform, which we intend to separate. Within each of the aforementioned time intervals, it is evident that the baseline approach falls short in effectively separating several glitches. The horizontal axis represents the UTC time zone.

To achieve the separation of background noise, we selected approximately 36 hours of detrended raw data from the U component with a sampling rate of 20Hz. This ensured an accurate estimation of the wavelet scattering covariance statistics. The network architecture used is the same as in previous examples, and we employed a window size of $204.8\,s$. By solving the optimization problem outlined in equation (12) with 200 L-BFGS iterations, we obtained the results depicted in Figure 19. Notably, glitches occurring just before the P-wave arrival and towards the end of the marsquake were successfully separated. Moreover, the separated background noise exhibits a stationary characteristic, which is desirable as it indicates minimal leakage of the
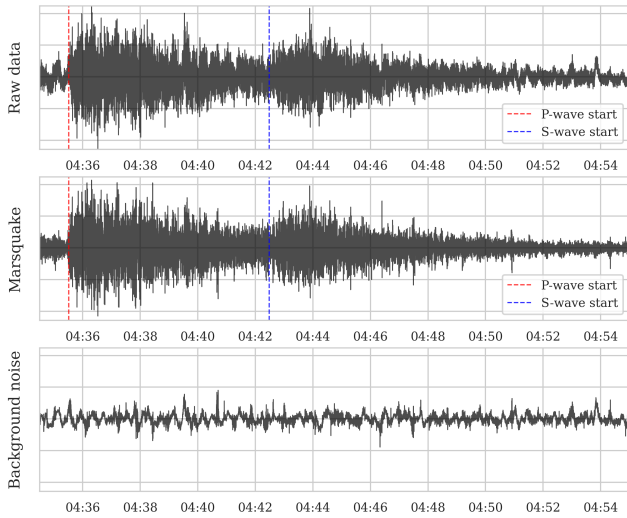
Figure 19. Unsupervised separation of background noise and glitches from a marsquake recorded by the InSight lander's seismometer on January 2, 2022 (InSight Marsquake Service, 2023). Approximately 36 hours of raw data from the U component were used without any additional prior knowledge of marsquakes or glitches. The horizontal axis is in UTC time zone.

Figure 20. Unsupervised separation of background noise and glitches from a marsquake recorded by the InSight lander's seismometer on July 26, 2019 (InSight Marsquake Service, 2023). Approximately 95 hours of raw data from the U component were used without any additional prior knowledge of marsquakes or glitches. The horizontal axis is in UTC time zone.

marsquake signal.

The final example involves a marsquake recorded on July 26, 2019 (InSight Marsquake Service, 2023). Separating the background noise in this case proves more challenging, as the P-wave arrival is barely discernible in the raw waveform shown in the top panel of Figure 20. Furthermore, the presence of background noise masks the detection of the S-wave, as well as the secondary PP- and SS-wave arrivals. To address these complexities and achieve accurate separation of the marsquake while minimizing signal leakage, we require 95 hours of detrended raw data from the U component. A window size of $409.6\,\text{s}$ is used, and the optimization problem in equation (12) is solved with 200 L-BFGS iterations. The results are depicted in Figure 20, where the separated marsquake is distinctly delineated. The accuracy of our approach is further confirmed by the independently picked arrival times by the InSight team (Scholz et al., 2020), shown as dotted lines in Figure 20. The alignment between their picked arrival times and our separated marsquake s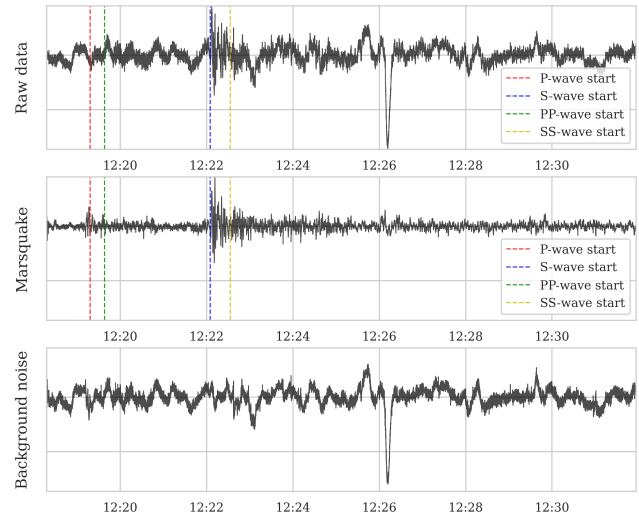erves as validation for the accuracy of our method.