Misfit functions based on differentiable dynamic time warping for waveform inversion

Fuqiang Chen*, KAUST, Daniel Peter, KAUST and Matteo Ravasi, KAUST

SUMMARY

Misfit functions based on differentiable dynamic time warping (DTW) have demonstrated an excellent performance in various sequence-related tasks. We introduce this concept in the context of waveform inversion and discuss a fast algorithm to calculate the first and second derivatives of such misfits. The DTW distance is originally calculated by solving a series of min sorting problems, thus it is not differentiable. The fundamental idea of differentiable DTW is replacing the min operator with its smooth relaxation. It is then straightforward to solve the derivatives following the differentiation rules, which results in both the differentiable misfit measurements and warping path. This path weights the traveltime mismatch more than its amplitude counterpart. We can construct a penalization term based on this warping path. The penalized misfit function is adaptable to measure traveltime and amplitude separately. Numerical examples show the flexibility and the performance of the proposed misfit function in mitigating local minima issues and retrieving the long-wavelength model features.

INTRODUCTION

Misfit functions play a crucial role in non-linear inverse problems, where estimated models are not linearly associated with predicted data. This non-linearity causes numerous local minima in the misfit function. Seismic waveform inversion is an example of non-linear inverse problems. In this case, an ideal scenario to reach a velocity model indicating plausible local minima is starting with low-frequency data, for example, below 3.0 Hz. However, those frequencies are generally not recorded in practice due to instrument response, source and receiver ghosts, and contamination by strong noise; this is especially true in the reflection part of seismic data.

A well-established strategy to alleviate such local minima issue is entitled wave-equation traveltime (WT) inversion (Luo and Schuster, 1991). The foundation of this strategy is that the traveltime inversion shows the chance to rid the implausible local minima. Various inversion methods (Luo and Schuster, 1991; Luo and Sava, 2011; Warner and Guasch, 2016; Zhu and Fomel, 2016) follow this principle where the misfit function focuses on the traveltime difference between observed and synthetic data. All these methods can be summarized as wave-equation inversion with tailored misfit function. Other misfit functions, e.g., based on a Wasserstein distance (Engquist et al., 2016; Métivier et al., 2016; Yang et al., 2018; Chen and Peter, 2018a,b) or dynamic time warping (DTW) distance (Ma and Hale, 2013; Yang et al., 2014), implicitly evaluate traveltime shifts by more complicated schemes. Numerical examples demonstrate their performance in waveform inversion. Based on our experience on applied optimal transport

with both entropy-regularized Kantorovich (Chen and Peter, 2018b) and 1D Monge (Chen and Peter, 2018a) formulations, we consider that adjoint sources of (some) current Wasserstein distances carry numerical artifacts. This bias might not be critical for the inversion of low frequencies but might affect the efficiency and robustness of inversions at higher frequencies.

This abstract focuses on DTW distance (Sakoe and Chiba, 1978). It is widely used for applications involving the misfit or similarity function evaluation between datasets even with different shapes (Cuturi and Blondel, 2017; Mensch and Blondel, 2018; Guen and Thome, 2019). In seismic data processing, it is used to estimate time-shifts between datasets (Hale, 2013) and demonstrated robustness compared to cross-correlation methods. In the geophysical community, Yang et al. (2014) and Ma and Hale (2013) used the original nondifferentiable DTW distance for waveform inversion. As we demonstrate later, the nondifferentiable DTW distance is not smooth and results in sharp changes in adjoint sources. We argue that these issues will affect the efficiency of updating seismic models by an iterative approach.

We expect misfit functions to be capable of exclusively focusing on the traveltime than the amplitude differences to retrieve the long-wavelength model. Even the DTW distance is a good indicator of traveltime difference but is still affected by amplitude differences. Guen and Thome (2019) propose to add a penalization term based on the differentiable warping path to the DTW distance to further highlight the “temporal” discrepancy. The “temporal” discrepancy has another name in the wave-equation inversion regime by traveltime difference. We introduce this penalized misfit to separately measure traveltime and amplitude differences. Luo and Hale (2013) have shown efforts to invert traveltime and amplitude separately. Additionally, (Blondel et al., 2021) develop a new misfit measurement entitled soft-DTW divergence and propose another prior distribution for warping path penalization. The soft-DTW divergence is still based on the differentiable DTW distance.

We use a simplified example and Chevron synthetic data to demonstrate how different hyper-parameter values will affect the performance of DTW distance. We expect the introduced misfits to alleviate the low-frequency dependency, invert the seismic data with no manual events identification, and work better than mean squared error in short-wavelength estimation.

THEORY

Given traces $f = \{f_1, \dots, f_{nt}\}$ and $g = \{g_1, \dots, g_{nt}\}$, the DTW distance can be defined as :

$$\varepsilon_0(f, g) := \min_{(i_k, j_k)} \sum_{k=1}^{np} \mathcal{D}_{i_k, j_k}, \quad (1)$$

where $\mathcal{D}_{i,j}$ measures the discrepancy between f_i and g_j . This ground measurement can be done by different methods, such

as the local cross-correlation and the squared Euclidean cost $\mathcal{D}_{i,j} = (f_i - g_j)^2$, the one we take in this abstract. The motivation of DTW is to find a np -node path parameterized by $(i_k, j_k)_{k=1}^{np}$ such that the accumulated cost along this path is minimized. In other words, an optimum of equation 1 warps the f to g , or vice versa. A greedy algorithm to find this warping path for practical problems is intractable. Sakoe and Chiba (1978) proposed to solve equation 1 by dynamic programming. The algorithm divides the overall problem in equation 1 into subproblems, where the solution to the original problem depends on the optimal solution to subproblems. The solution to the subproblem can be found by a sorting process with the min operator. The output of this min operator is however globally nondifferentiable. Locally, it is changing as a step function, which implies that the DTW distance is not even continuous. We still use the word derivative to stand for the change of DTW distance caused by a small change in synthetic data (zero almost everywhere except along the warping path). Even the approximated derivative information of $\varepsilon_0(f, g)$ is used to retrieve a model reducing the data misfit or image defocussing (Ma and Hale, 2013; Yang et al., 2014), this approximation at least will slow down the iteration performance if not ruin the final result.

Cuturi and Blondel (2017) developed a differentiable DTW distance, which replaces the min by its smooth alternative:

$$\min^\gamma(x_1, \dots, x_n) := -\gamma \log \sum_{i=1}^n e^{-x_i/\gamma} \quad (2a)$$

$$= x' - \gamma \log \sum_{i=1}^n e^{-\frac{x_i - x'}{\gamma}}, \quad (2b)$$

where $x' = \min(x_1, \dots, x_n)$. We use equation 2b to calculate the $\min^\gamma$ because it is numerically stable. An equivalent definition of DTW distance is

$$\varepsilon_\gamma(f, g) := \min^\gamma \{ \langle \mathcal{W}, \mathcal{D} \rangle, \mathcal{W} \in \mathbb{R}^{nt \times nt} \}, \quad (3)$$

where $\langle \mathcal{W}, \mathcal{D} \rangle = \sum_{i,j} \mathcal{W}_{i,j} \mathcal{D}_{i,j}$ denotes the Frobenius inner product. $\mathcal{W}_{i,j}$ can be interpreted as the probability of that f_i and g_j are connected. Equation 3 recovers the nondifferentiable DTW distance admitted by equation 1 if $\gamma = 0$. An arbitrary warping path is trivial and not helpful in misfit measurements. We apply restrictions on warping paths by, for example, only allowing them to arise following three directions: $\rightarrow$, $\downarrow$ and $\swarrow$. This restriction leads to a fast algorithm developed by Sakoe and Chiba (1978). It consists of a sweeping process following the pattern:

$$\mathcal{A}_{i,j} = \mathcal{D}_{i,j} + \min^\gamma(\mathcal{A}_{i-1,j-1}, \mathcal{A}_{i-1,j}, \mathcal{A}_{i,j-1}), \quad (4)$$

where $\mathcal{A}$ is the accumulated error array memorizing all paths satisfying the forward marching directions. Plunging the definition of $\min^\gamma$, we have

$$\mathcal{A}_{i,j} = \mathcal{D}_{i,j} - \gamma \log \sum_{(i',j') \in \mathbb{P}} e^{-\mathcal{A}_{i+i',j+j'}/\gamma}, \quad (5)$$

where $\mathbb{P} = \{(-1,0), (0,-1), (-1,-1)\}$. To get the DTW distance, we recursively evaluate equation 4 from $\mathcal{A}_{1,1}$ to $\mathcal{A}_{nt,nt}$. After the forward sweeping, we have $\varepsilon_\gamma(f, g) = \mathcal{A}_{nt,nt}$ (Mensch and Blondel, 2018). As pointed out by Blondel et al.

(2021) in their proposition 2, the minimum of differentiable DTW distance is not achieved at $f = g$ if $\mathcal{D}$ is the squared Euclidean cost. It could also give negative misfit measurements for γ larger than a certain bound. To address these issues, Blondel et al. (2021) proposed a differentiable DTW divergence as

$$\xi_\gamma(f, g) := \varepsilon_\gamma(f, g) - \frac{1}{2} \varepsilon_\gamma(f, f) - \frac{1}{2} \varepsilon_\gamma(g, g). \quad (6)$$

Both the differentiable DTW distance and divergence reflect the traveltime difference. This capacity can be further enhanced by introducing a penalty term (Guen and Thome, 2019). DTW distance with penalization is defined as:

$$\varepsilon_{\gamma,\lambda}(f, g; \mathcal{I}) := \varepsilon_\gamma + \lambda \langle \mathcal{W}^*, \mathcal{I} \rangle \quad (7a)$$

$$= \varepsilon_\gamma + \lambda \phi \quad (7b)$$

where $\mathcal{W}^*$ is the optimum for equation 3 and $\mathcal{I}$ represents the prior distribution of certain warping path, for example, $\mathcal{I}_{i,j} = (i-j)^2/nt^2$ (Guen and Thome, 2019) or $\mathcal{I} = \mathcal{D}$ (Blondel et al., 2021). The DTW divergence with penalization can be defined as:

$$\xi_{\gamma,\lambda}(f, g; \mathcal{I}) := \xi_\gamma + \lambda \langle \mathcal{W}^*, \mathcal{I} \rangle \quad (8a)$$

$$= \xi_\gamma + \lambda \phi. \quad (8b)$$

In the following, we explain the algorithm proposed by Cuturi and Blondel (2017) and Mensch and Blondel (2018) to efficiently calculate the derivative of $\varepsilon_{\gamma,\lambda}$ and $\xi_{\gamma,\lambda}$ w.r.t. f . Considering $\varepsilon_\gamma = \mathcal{A}_{nt,nt}$ and the dependency of $\mathcal{A}_{nt,nt}$ on all $\mathcal{A}_{i,j}$, Cuturi and Blondel (2017) discovered a recursive formula to calculate $\frac{\partial \varepsilon_\gamma}{\partial \mathcal{A}_{i,j}}$. Given the 3-point forward sweeping template in equation 4, we can have

$$\frac{\partial \varepsilon_\gamma}{\partial \mathcal{A}_{i,j}} = \sum_{(i',j') \in \mathbb{C}} \frac{\partial \varepsilon_\gamma}{\partial \mathcal{A}_{i+i',j+j'}} \frac{\partial \mathcal{A}_{i+i',j+j'}}{\partial \mathcal{A}_{i,j}} \quad (9)$$

where $\mathbb{C} = \{(1,0), (0,1), (1,1)\}$. Let $\mathcal{X}_{i,j} = \frac{\partial \varepsilon_\gamma}{\partial \mathcal{A}_{i,j}}$, then

$$\mathcal{X}_{i,j} = \sum_{(i',j') \in \mathbb{C}} \mathcal{X}_{i+i',j+j'} \frac{\partial \mathcal{A}_{i+i',j+j'}}{\partial \mathcal{A}_{i,j}} \quad (10)$$

Based on equation 5, we can have

$$\frac{\partial \mathcal{A}_{i+1,j}}{\partial \mathcal{A}_{i,j}} = \frac{e^{-\mathcal{A}_{i,j}/\gamma}}{\sum_{(i',j') \in \mathbb{P}} e^{-\mathcal{A}_{i+1+i',j+j'}/\gamma}} \quad (11)$$

Taking the log of both sides of equation 11 and plunging equation 5, we get

$$\gamma \log \frac{\partial \mathcal{A}_{i+1,j}}{\partial \mathcal{A}_{i,j}} = \mathcal{A}_{i+1,j} - \mathcal{A}_{i,j} - \mathcal{D}_{i+1,j} \quad (12)$$

Similarly, we can have

$$\begin{aligned} \gamma \log \frac{\partial \mathcal{A}_{i,j+1}}{\partial \mathcal{A}_{i,j}} &= \mathcal{A}_{i,j+1} - \mathcal{A}_{i,j} - \mathcal{D}_{i,j+1} \\ \gamma \log \frac{\partial \mathcal{A}_{i+1,j+1}}{\partial \mathcal{A}_{i,i}} &= \mathcal{A}_{i+1,i+1} - \mathcal{A}_{i,j} - \mathcal{D}_{i+1,j+1} \end{aligned} \quad (13)$$

SEG-2021-SOFT-DTW

Combining equations 10, 12, 13 and $\mathcal{X}_{nt,nt} = \frac{\partial \mathcal{A}_{nt,nt}}{\partial \mathcal{A}_{nt,nt}} = 1$, we can fill the matrix $\mathcal{X}$, which takes a backward sweeping process.

In summary, we obtain the DTW distance by one forward sweeping and the optimal warping path $\mathcal{W}^* = \mathcal{X}$ by one backward sweeping. Calculating the derivative of the penalization term ϕ w.r.t synthetic data involves the Hessian of DTW distance with $\mathcal{A}_{i,j}$ on all feasible paths. The direct calculation is however prohibitive. Fortunately, we only need Hessian-vector products. Pearlmutter (1994) proposed a fast algorithm to calculate such products without explicitly creating the Hessian matrix. We condense the main points associated with our application as below:

$$\frac{\partial \phi}{\partial \mathcal{A}_{k,l}} = \sum_{i=1}^{nt} \sum_{j=1}^{nt} \frac{\partial \mathcal{X}_{i,j}}{\partial \mathcal{A}_{k,l}} \mathcal{I}_{i,j}, \quad (14a)$$

$$= \underbrace{\sum_{i=1}^{nt} \sum_{j=1}^{nt} \frac{\partial^2 \varepsilon_\gamma}{\partial \mathcal{A}_{k,l} \partial \mathcal{A}_{i,j}} \mathcal{I}_{i,j}}_{\text{Hessian product}}, \quad (14b)$$

$$= \frac{\partial}{\partial \mathcal{A}_{k,l}} \underbrace{\sum_{i=1}^{nt} \sum_{j=1}^{nt} \frac{\partial \varepsilon_\gamma}{\partial \mathcal{A}_{i,j}} \mathcal{I}_{i,j}}_{\text{directional derivative}}. \quad (14c)$$

Compared to equation 14b, equation 14c converts the Hessian product into the first derivative of directional derivatives of the differentiable DTW distance along $\mathcal{I}$.

EXAMPLES

A single-parameter problem is commonly used to demonstrate how a misfit function can mitigate local minima. Here we extend the demonstration to a two-parameter problem to better highlight the different misfit landscapes. The governing equation becomes

$$f(t) = r(t; t_0, f_m) + r(t; t_1, f_m),$$

where $r(t; t_0, f_m)$ denotes the Ricker wavelet with the peak frequency f_m and the peak amplitude at time t_0 . Figure 1 displays the observed and two instances of synthetic data, with the peak frequency $f_m = 12.0\text{Hz}$. Figure 2 illustrates the misfit between observed and synthetic traces measured by differentiable DTW distance $\varepsilon_{\gamma,\lambda}$, where γ and λ take (100, 1, 0.01) and (0.9, 99), respectively. The case with $\gamma = 0.01$ represents the non-differentiable DTW distance. From Figure 2 we see: (i) the smoothness of DTW distance improves with larger γ ; (ii) with no penalization, a small γ gives unsatisfying measurements in global as shown in Figure 2g and a proper γ can improve as demonstrated in Figure 2d; (iii) the penalization term mitigates the local minima in all cases. Considering numerically $\gamma \rightarrow \gamma_0$, $\min^\gamma = \min$, a case in Figure 2g, we can conclude that the nondifferentiable DTW distance results in non-smooth misfit measurements, which will have an impact on efficiency in model updating based on gradient methods.

Next, we use Chevron blind test datasets to demonstrate that the differentiable DTW distance improves the misfit measure-

ments for waveform inversion compared to the nondifferentiable form. We start with the initial model released with seismic data. Figures 3, 4, and 5 show the adjoint source from data in 3.0Hz-3.5Hz, with different λ . Figures 6, 7 and 8 show their corresponding gradients. We conclude that a well-behaved misfit function such as nondifferentiable DTW distance does not ensure a plausible model estimation if we rely on its gradient information to determine the model update direction. In Figure 2, where only the misfit value is needed, we see a smaller γ (means stronger nonsmoothness) combined with a larger λ gives better measurements. However, from Figures 3 to 8, we clearly see that such a combination will result in a “noisy” gradient because of incoherent high-frequencies induced by sudden changes in the adjoint source. This observation has mathematical evidence. The DTW distance ε_γ and its gradient are $1/\gamma$ - and $2/\gamma$ -Lipschitz continuous, respectively. An α -Lipschitz continuous function means its change rate is bounded by α . Therefore, a smaller γ , for example, $\gamma = 0.01$ which makes $\min^\gamma$ equal $\min$ almost for every subproblem in this example, allows larger change rates. This explains why we see sharp changes in the adjoint sources as shown in Figures 3a, 3b, 4a, 4b, 5a, and 5b. Comparing Figure 5b with Figures 4b and Figure 3b, we see stronger penalization results in stronger discontinuities (sudden changes) because the adjoint source induced by the penalization term involves the second-order derivative, which admits a larger Lipschitz constant than $2/\gamma$. Besides, a smaller γ also results in a larger difference between adjoint sources solved by the definition and the algorithm. A larger γ can significantly reduce this discrepancy.

We find differentiable DTW distance with ($\gamma = 10$, $\lambda = 99$) gives the best results among these tests in terms of reducing the data mismatch. We also conducted tests with $\gamma = 0.001$ and $\lambda = 999$, however, those results were not providing further insights and therefore not presented.

CONCLUSION

We introduce a differentiable DTW distance for waveform inversion. It results in smooth misfit measurements and numerically accurate adjoint sources. It can also emphasize travel-time differences and reduce the impact of amplitude differences on misfit accumulation. Numerical tests demonstrate the improvements in efficiency from the introduced misfit function compared to the nondifferentiable form for model estimation. They also demonstrate the potential to invert the traveltimes misfits without prior event identification starting from practically available frequencies. We will consider an extension with DTW divergence in further studies.

ACKNOWLEDGMENTS

The authors would like to thank Chevron for making the synthetic data available. For computer time, this research used the resources of the Supercomputing Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

Figure 1: Observed trace (black) and two instances of synthetic traces.

Figure 2: Misfit measured by differentiable DTW distance $\varepsilon_{\gamma,\lambda}$ with different values of hyper-parameters γ and λ , where $\gamma = [100, 1, 0.01]$ from top to bottom and $\lambda = [0, 9, 99]$ from left to right. Every misfit map is normalized in $[0, 1]$ by doing $\varepsilon \leftarrow \varepsilon - \varepsilon_{\min}$ and $\varepsilon \leftarrow \varepsilon / \varepsilon_{\max}$. The minimum of misfit function is achieved around $(\delta t_0, \delta t_1) = (0, 0)$

Figure 3: The adjoint source of differentiable DTW distance with $\lambda = 0$ and $\gamma = (0.1, 1, 10, 100)$ for (a), (b), (c) and (d). The same order is used for Figures 4 and 5. We applied a window for observed and synthetic data. The window boundaries almost cover the non-zero values in the adjoint source. We do not show the muting scheme because of limited space.

Figure 4: The adjoint source of differentiable DTW distance with $\lambda = 9$ and $\gamma = (0.1, 1, 10, 100)$.

Figure 5: The adjoint source of differentiable DTW distance with $\lambda = 99$ and $\gamma = (0.1, 1, 10, 100)$.

Figure 6: The gradient of differentiable DTW distance with $\lambda = 0$ and $\gamma = (0.1, 1, 10, 100)$ for (a), (b), (c) and (d). The same order is used for Figures 7 and 8. We normalize every gradient in $[-1, 1]$ by doing $g \leftarrow g / \text{abs}(g)_{\max}$, also for those in Figures 7 and 8.

Figure 7: The gradient of differentiable DTW distance with $\lambda = 9$ and $\gamma = (0.1, 1, 10, 100)$.

Figure 8: The gradient of differentiable DTW distance with $\lambda = 99$ and $\gamma = (0.1, 1, 10, 100)$.

REFERENCES

- Blondel, M., A. Mensch, and J.-P. Vert, 2021, Differentiable divergences between time series: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, 3853–3861.
- Chen, F. and D. Peter, 2018a, Constructing misfit function for full waveform inversion based on sliced Wasserstein distance: 80th Annual International Conference and Exhibition, EAGE, Extended Abstracts, 1–5, doi: <https://doi.org/10.3997/2214-4609.201801030>.
- Chen, F. and D. Peter, 2018b, A misfit function based on entropy regularized optimal transport for full-waveform inversion: 88th Annual International Meeting, SEG, Expanded Abstracts, 1314–1318, doi: <https://doi.org/10.1190/segam2018-2995612.1>.
- Cuturi, M. and M. Blondel, 2017, Soft-dtw: A differentiable loss function for time-series: Presented at the ICML.
- Engquist, B., B. D. Froese, and Y. Yang, 2016, Optimal transport for seismic full waveform inversion: arXiv preprint arXiv:1602.01540.
- Guen, V., and N. Thome, 2019, Shape and time distortion loss for training deep time series forecasting models: Presented at the NeurIPS.
- Hale, D., 2013, Dynamic warping of seismic images: Geophysics, **78**, no. 2, S105–S115, doi: <https://doi.org/10.1190/geo2012-0327.1>.
- Luo, S., and D. Hale, 2013, Separating traveltimes and amplitudes in waveform inversion: 83rd Annual International Meeting, SEG, Expanded Abstracts, 969–974, doi: <https://doi.org/10.1190/segam2013-1284.1>.
- Luo, S., and P. Sava, 2011, A deconvolution-based objective function for wave-equation inversion: 71st Annual International Meeting, SEG, Expanded Abstracts, 2788–2792, doi: <https://doi.org/10.1190/1.3627773>.
- Luo, Y., and G. T. Schuster, 1991, Wave-equation traveltime inversion: Geophysics, **56**, 645–653, doi: <https://doi.org/10.1190/1.1443081>.
- Ma, Y., and D. Hale, 2013, Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion: Geophysics, **78**, no. 6, R223–R233, doi: <https://doi.org/10.1190/geo2013-0004.1>.
- Mensch, A., and M. Blondel, 2018, Differentiable dynamic programming for structured prediction and attention: Proceedings of the 35th International Conference on Machine Learning, 3462–3471.
- Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Journal International, **205**, 345–377, doi: <https://doi.org/10.1093/gji/ggw014>.
- Pearlmutter, B. A., 1994, Fast exact multiplication by the Hessian: Neural Computation, **6**, 147, doi: <https://doi.org/10.1162/neco.1994.6.1.147>.
- Sakoe, H., and S. Chiba, 1978, Dynamic programming algorithm optimization for spoken word recognition: IEEE Transactions on Acoustics, Speech, and Signal Processing, **26**, 43–49, doi: <https://doi.org/10.1109/TASSP.1978.1163055>.
- Warner, M., and L. Guasch, 2016, Adaptive waveform inversion: Theory: Geophysics, **81**, no. 6, R429–R445, doi: <https://doi.org/10.1190/geo2015-0387.1>.
- Yang, D., A. Malcolm, and M. Fehler, 2014, Using image warping for time-lapse image domain wavefield tomography: Geophysics, **79**, no. 3, WA141–WA151, doi: <https://doi.org/10.1190/geo2013-0424.1>.
- Yang, Y., B. Engquist, J. Sun, and B. F. Hamfeldt, 2018, Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion: Geophysics, **83**, no. 1, R43–R62, doi: <https://doi.org/10.1190/geo2016-0663.1>.
- Zhu, H., and S. Fomel, 2016, Building good starting models for full-waveform inversion using adaptive matching filtering misfit: Geophysics, **81**, no. 5, U61–U72, doi: <https://doi.org/10.1190/geo2015-0596.1>.