Optimal Transport and Geophysical Inversion

EMinar series,

19th May 2021.

Malcolm Sambridge[1], Andy Jackson[2], Andrew Valentine[3]

[1]Research School of Earth Sciences, Australian National University Canberra ACT, Australia;
[2]ETH Zürich, Institut für Geophysik, Sonneggstrasse 5, CH-8092 Zürich, Switzerland;
[3]Dept. of Earth Sciences, Durham University, UK.

Inversion problems and methods

Specific problems

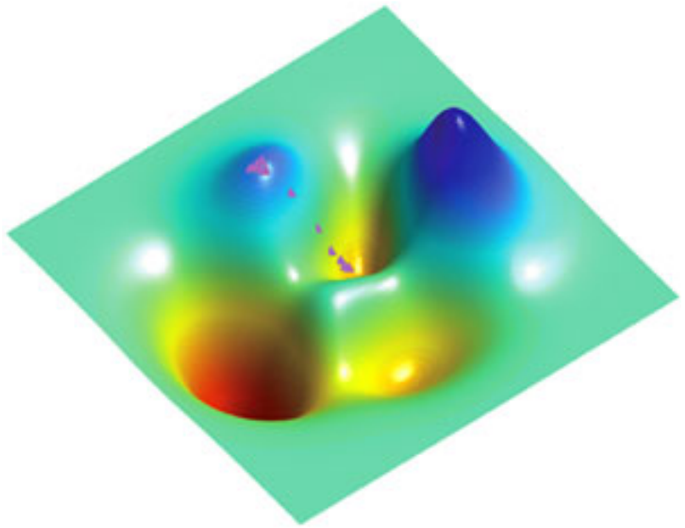
General methods

“No battle plan survives first contact with the enemy.”

Helmuth von Moltke the Elder (1800-1891)

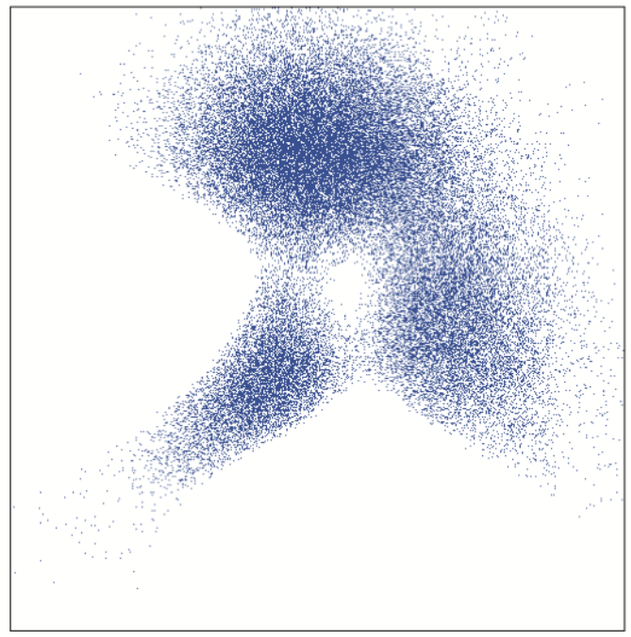
Two common approaches to inversion

Optimisation framework:
Minimising some misfit function…



Depends on a user defined data misfit function.

Bayesian Inference:
Probabilistic sampling

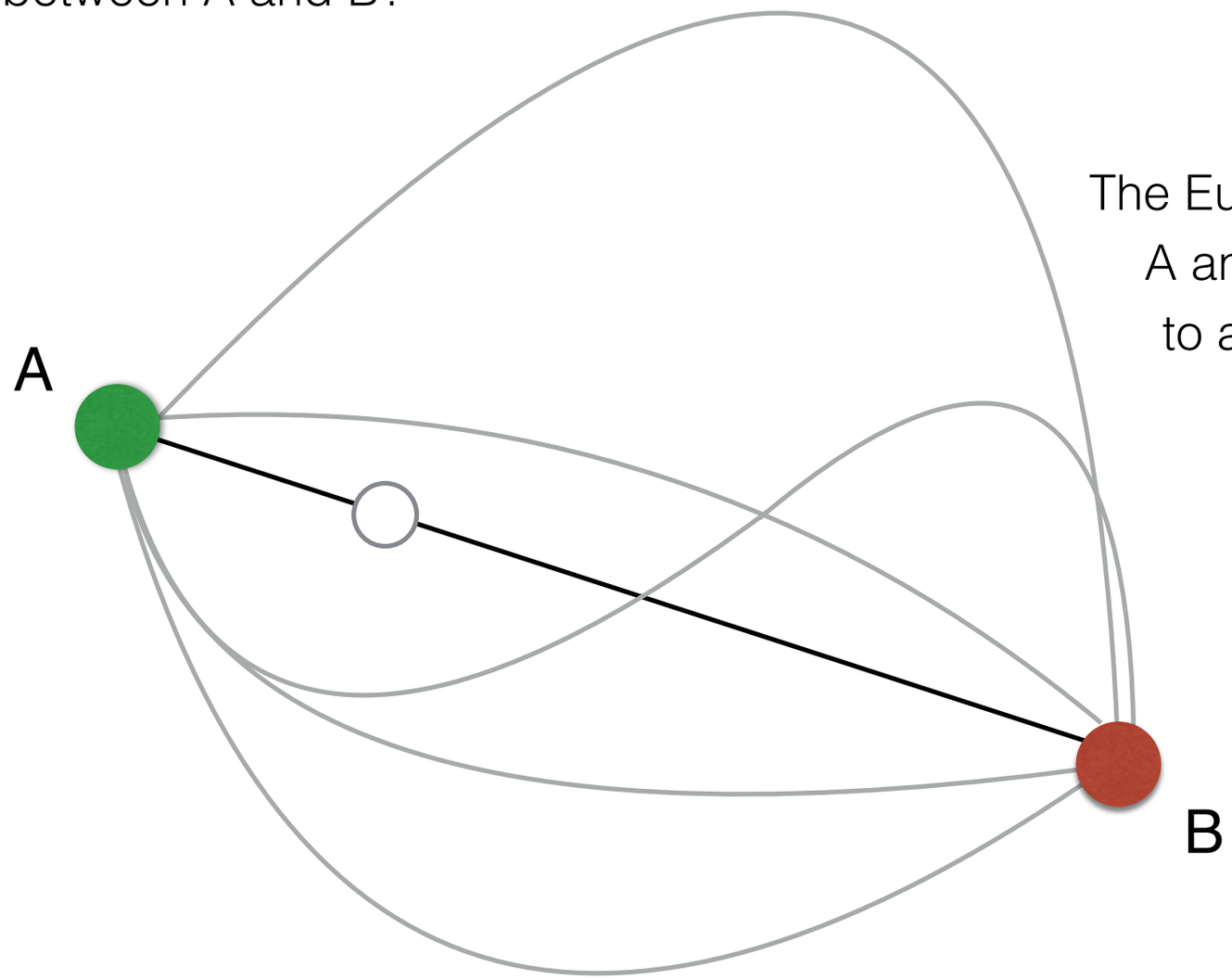


Requires a Likelihood function,
Involves probability density functions

Optimal Transport is directly relevant to both frameworks

The distance between A and B

What is the distance between A and B?

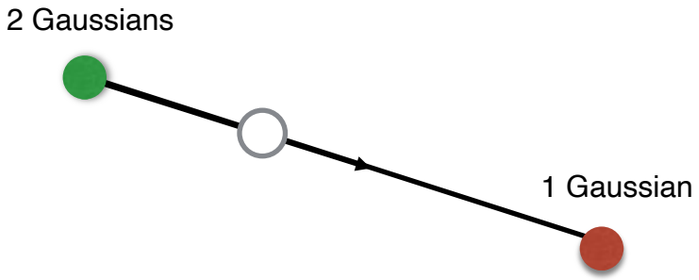


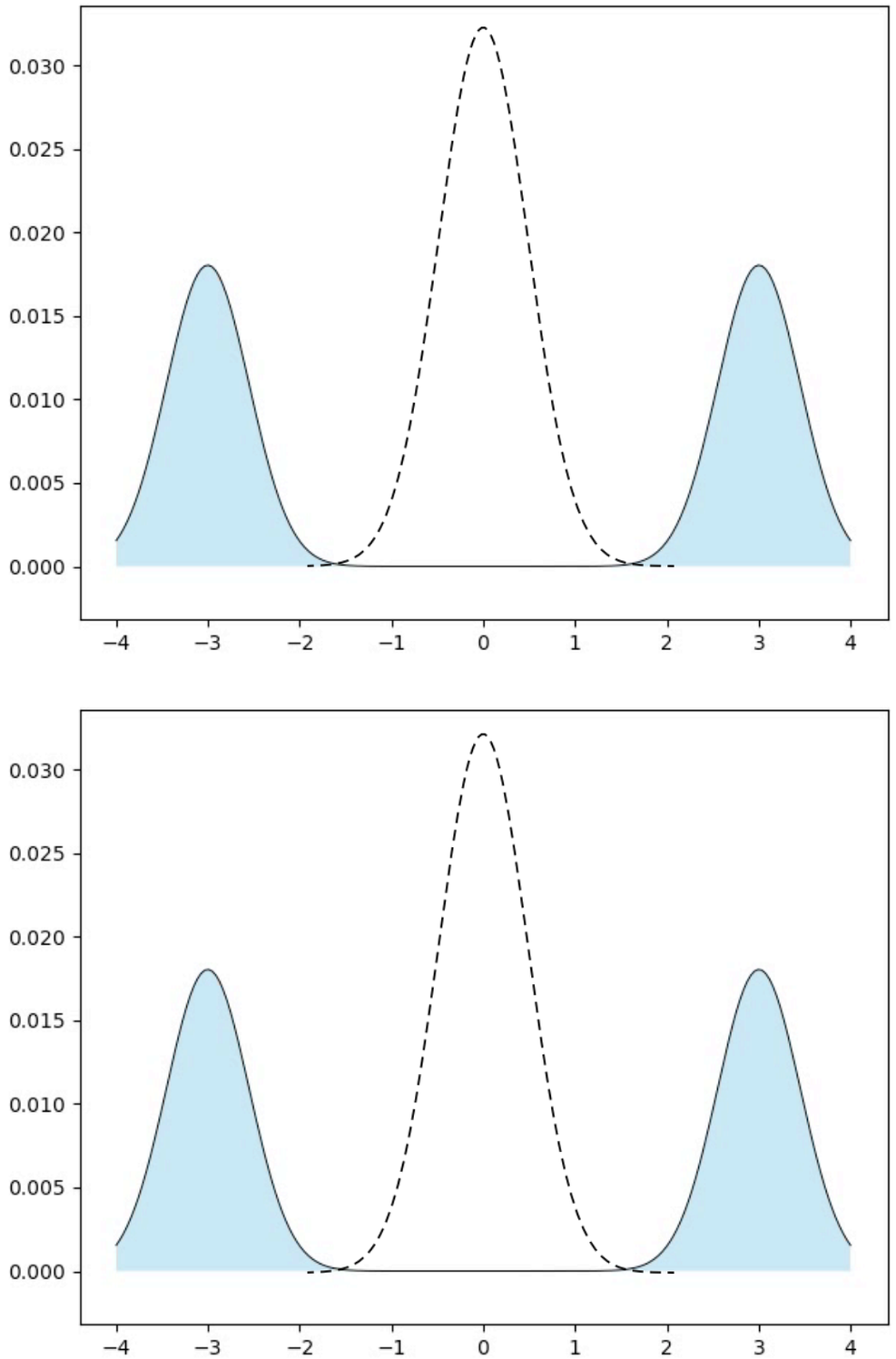
The Euclidean distance between A and B is itself the solution to an optimisation problem

A misfit function is a distance between two objects (observations and predictions, etc)
A misfit function corresponds to a transformation between predicted and observed data.

Transforming two Gaussians into one

Equal steps along the Euclidean path between endpoints

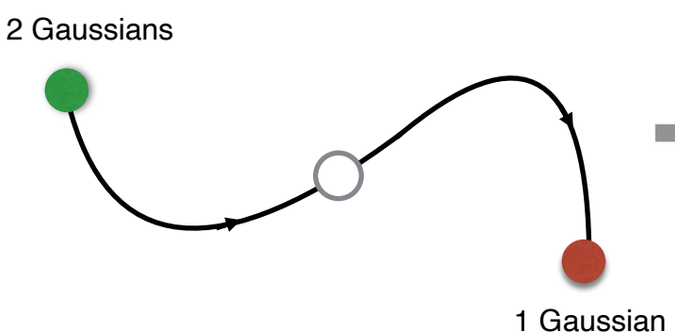




Animation of the linear (**least squares**) path between the start and end distributions

Only amplitude changes

Equal steps along the Optimal Transport path between endpoints



Animation of the **optimal transport** path between the start and end distributions

Amplitude and position changes

Optimal Transport of images

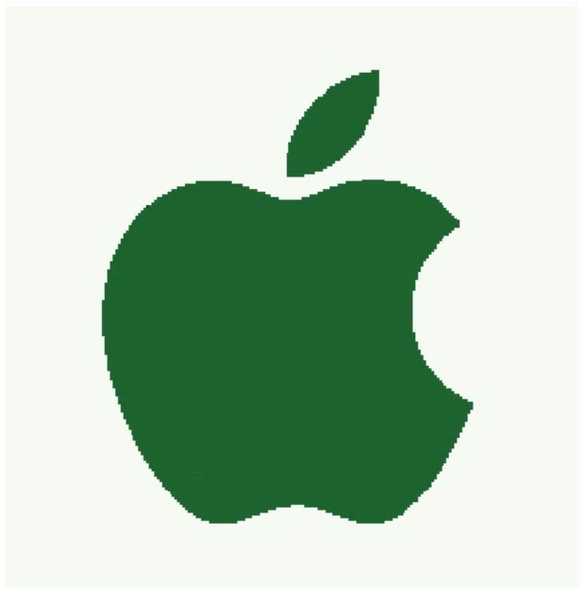
A mathematical topic that originated in the 19th century that has yielded two Fields medals and a Nobel prize. A vast literature exists, from mathematics to computer science.

Introduced into Geophysics by Engquist and Froese (2014).

Source

Linear

Source



Optimal transport

Target

$u_p(t)$

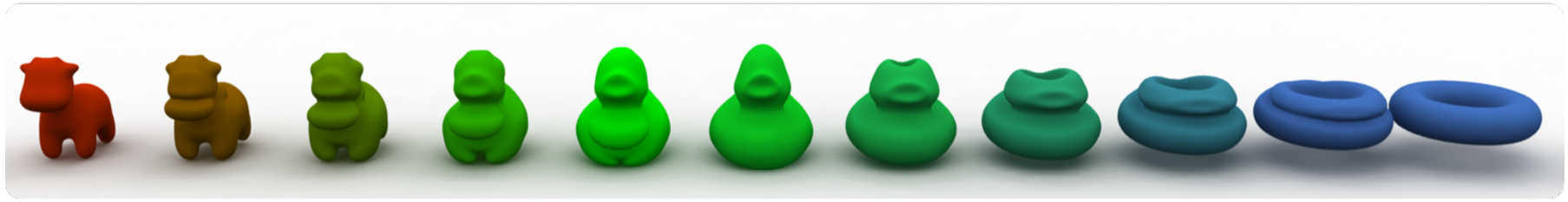
$u_o(t)$

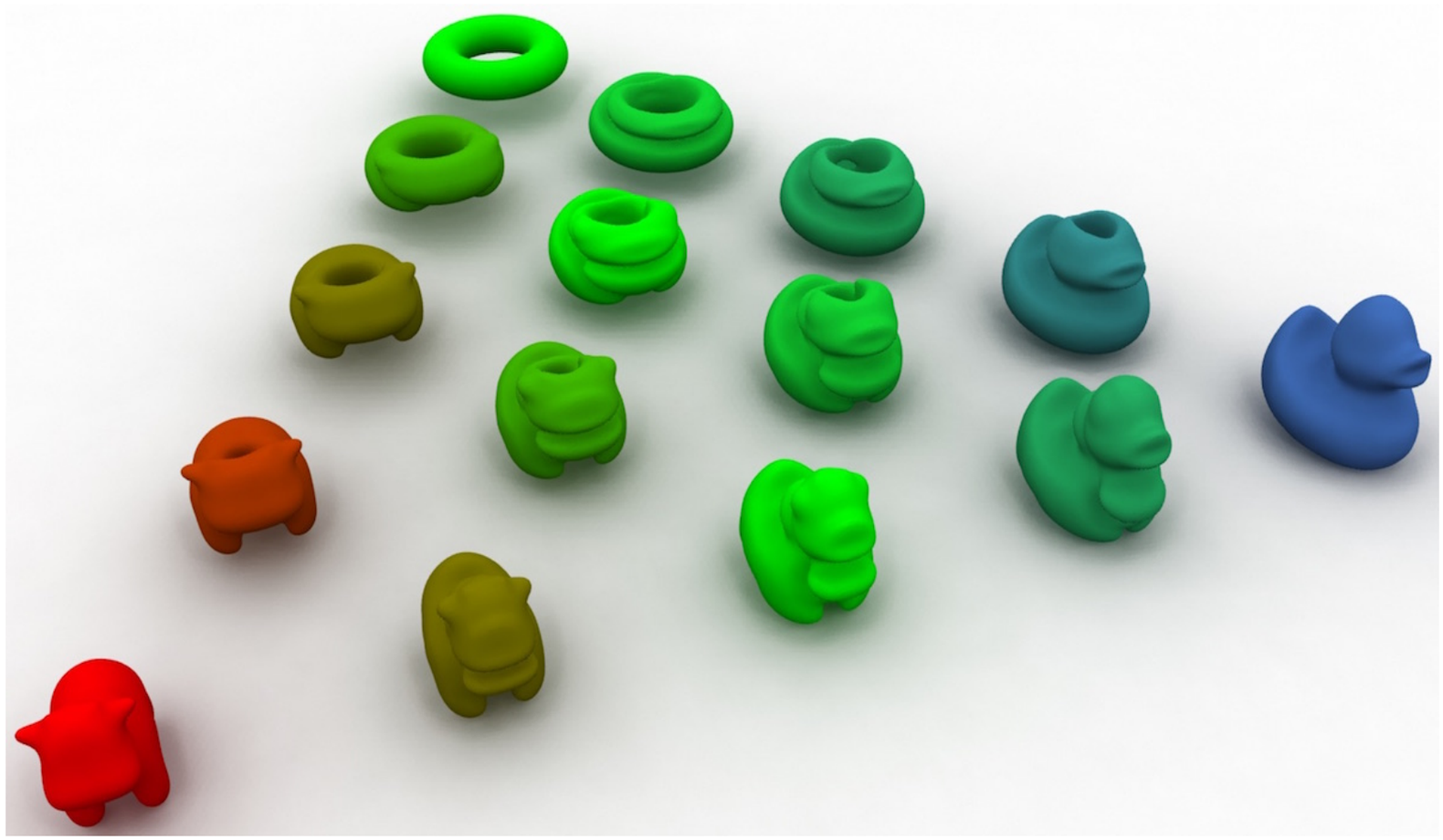
Makes use of Sinkhorn Convolutional Wasserstein algorithm of Solomon et al. (2015)

Rémi Flamary and Nicolas Courty, POT Python Optimal Transport library, 2017. https://pythonot.github.io/

Optimal transport of shapes in 3D

From **Cow** to **Duck** to **Torus**



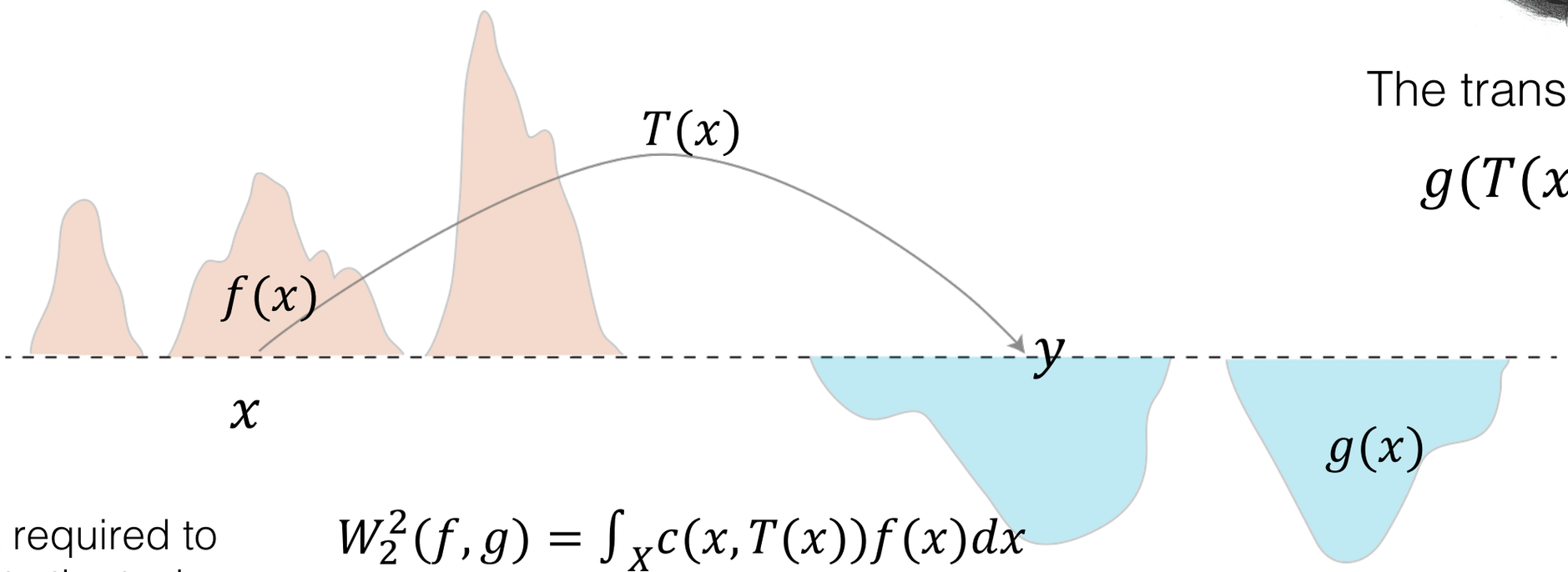


From Solomon et al. (2015)

Optimal Transport: Napoleon's problem

The modern subject of optimal transport traces its roots back to the tail of Napoleon and his mat

How to optimally transport the sand $f(x)$ to the holes $g(x)$?



The transport map

$$g(T(x)) = f(x)$$

The work required to complete the task

$$W_2^2(f,g) = \int_X c(x,T(x))f(x)dx$$

$c(x,y)$ is the distance between x and y

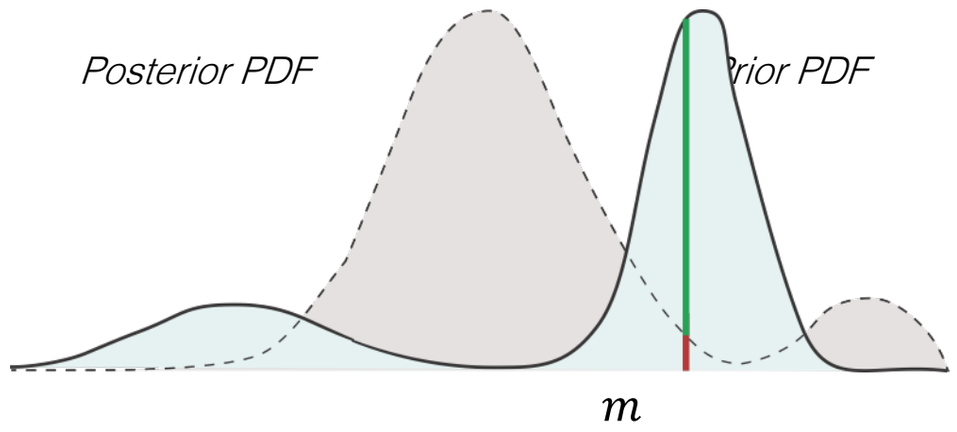
$$W_p = (\sum_i d_i^p m_i)^{1/p} \qquad p = 1 \text{ or } 2$$

$W_2^2 \propto$ (distance) ^{2} x mass $\propto$ Energy

The modern field derives from the work of Kantorovich (1942); Villani (2003, 2008).

Optimal Transport and Bayesian Inference

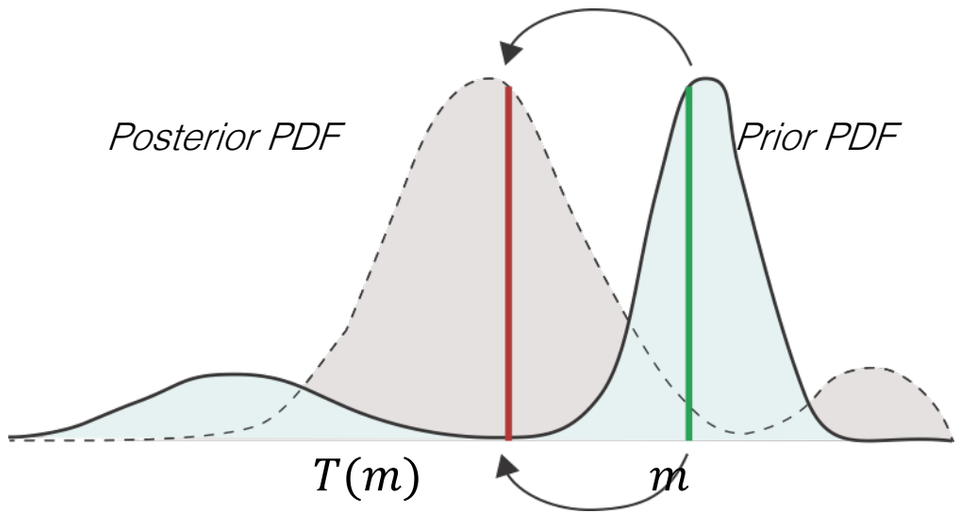
Bayes' rule $$p(m|d) = k \times L(d|m)p(m)$$



Optimal transport provides a `push forward' deterministic map between continuous PDFs.

$$p(T(m)|d) = p(m)$$

where $T(m)$ is the `Transport Map'



The transport map completely defines the posterior PDF and hence is an alternate way to describe the complete solution to the inverse problem.

El Moselhy, T. A., and Y. M. Marzouk (2012), *Bayesian inference with optimal maps*, Journal of Computational Physics, 231(23), 7815 – 7850.

Optimal Transport and Bayesian Inference

An optimal transport map (deterministic) or plan (probabilistic) can be viewed as a general solution to an inverse problem

A `push forward' deterministic map between prior and posterior PDFs.

$$p(m|d) = p(T(m))$$

No McMC required!

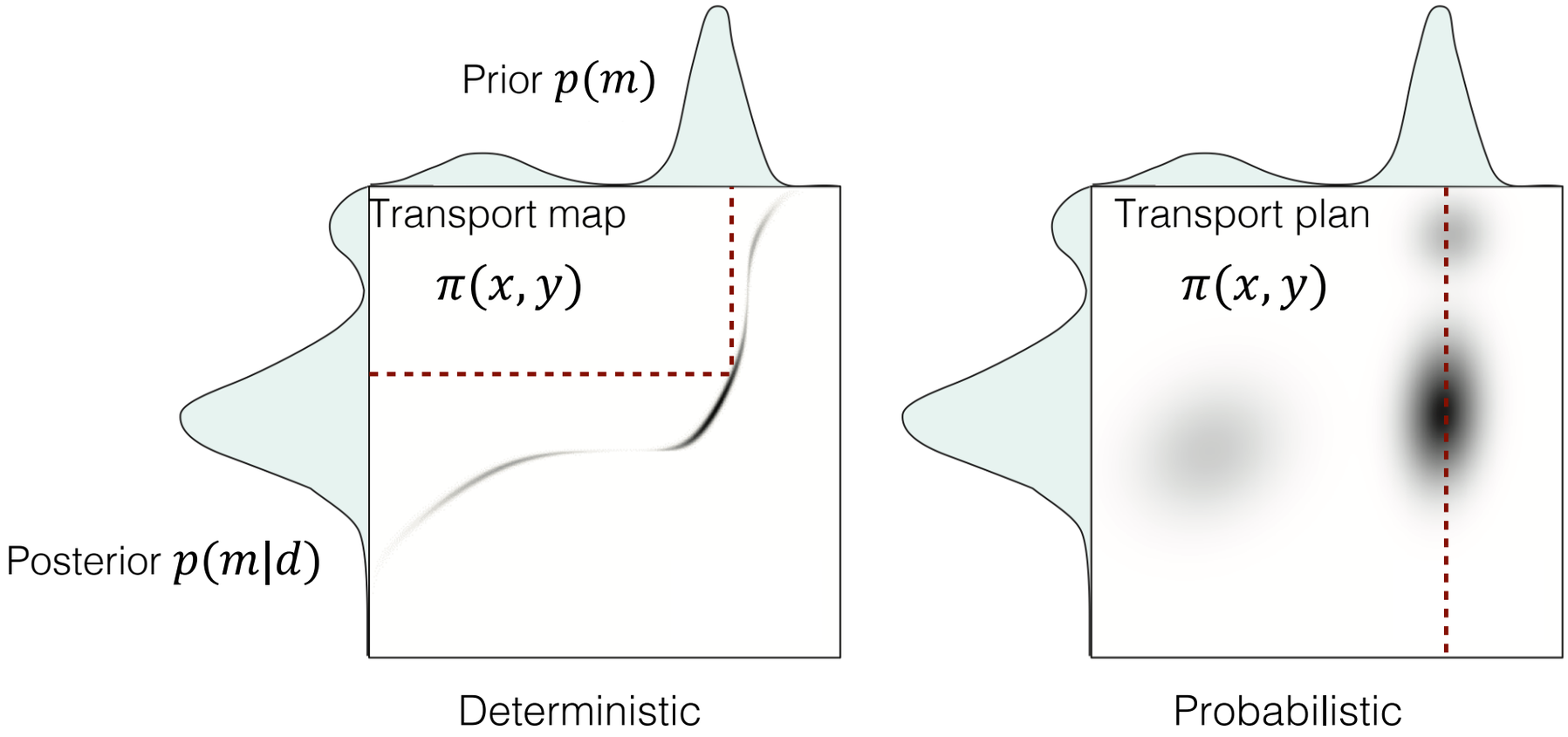
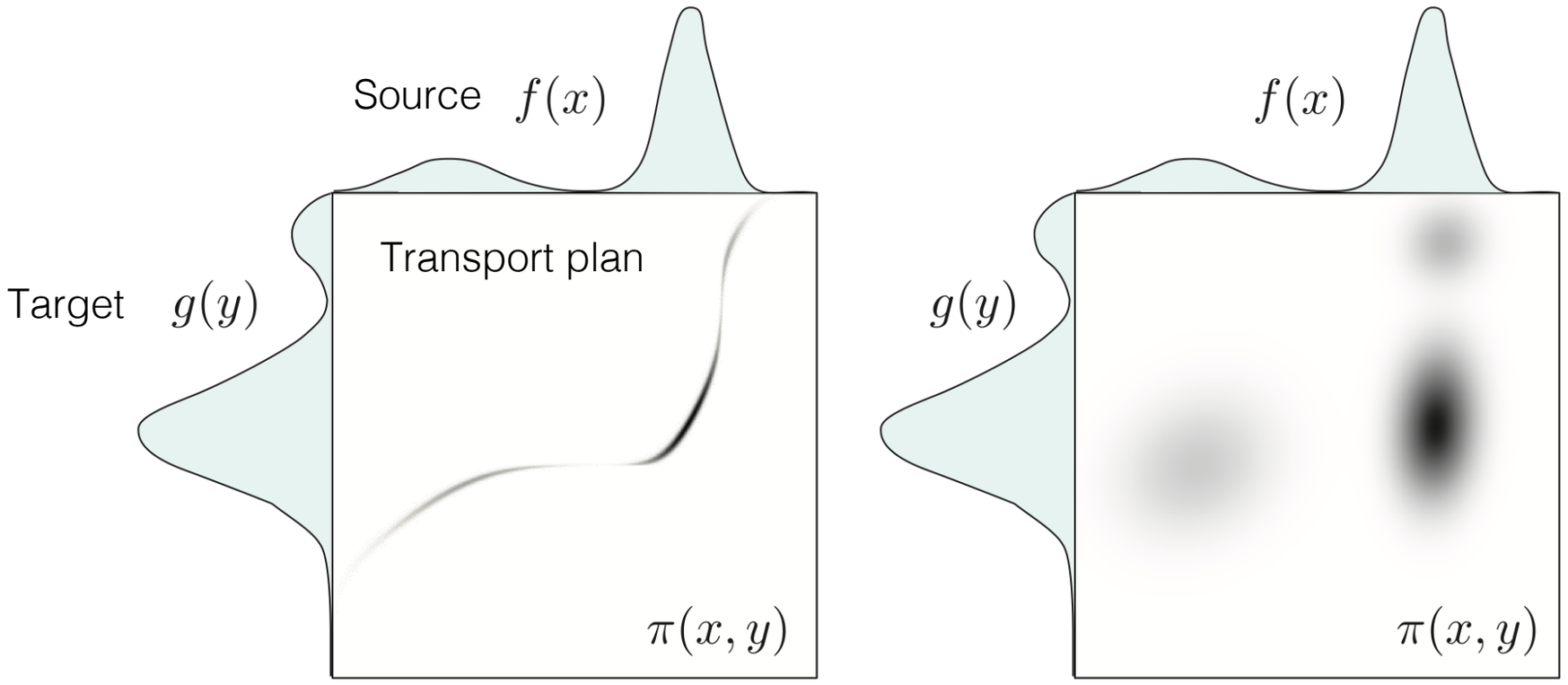


Figure adapted from Solomon (2015)

Optimal transport maps one PDF onto another

Linear programming formulation of Kantorovich (1942). Solve for transport plan $\pi_{i,j}$



$$\min_{\pi(x_i,y_j)} W_p^p = \sum_{i,j} c_{i,j}\pi_{i,j}, \sum_i \pi_{i,j} = f(x_i), \sum_j \pi_{i,j} = g(y_i)$$

$\pi(x, y)$ = Transport plan

$c(x, y)$ = distance between x and y

$$W_p^p = (\text{Distance})^{\text{p}} \times (\text{mass})$$

W_1 = distance x mass

W_2^2 = (distance) ^{2} x mass

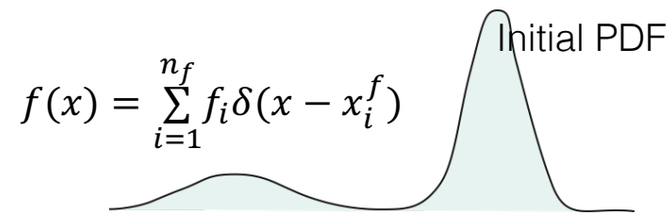
Figure adapted from Solomon et al. (2015)

Analytical OT solutions in 1D: Hooray!

Analytical solution for 1D continuous case in terms of inverse CDFs (Villani, 2003)

$$W_p(f,g) = \int_0^1 |F^{-1} - G^{-1}|^p dy$$

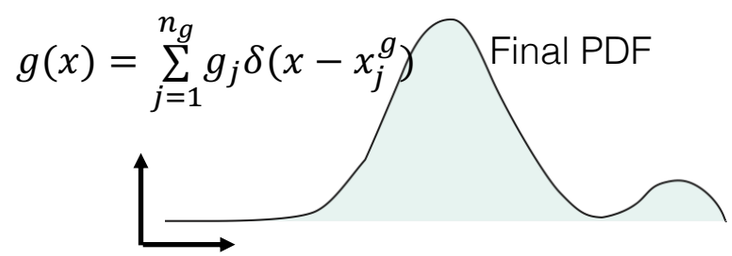
Point mass representation



Our 1D discrete solution

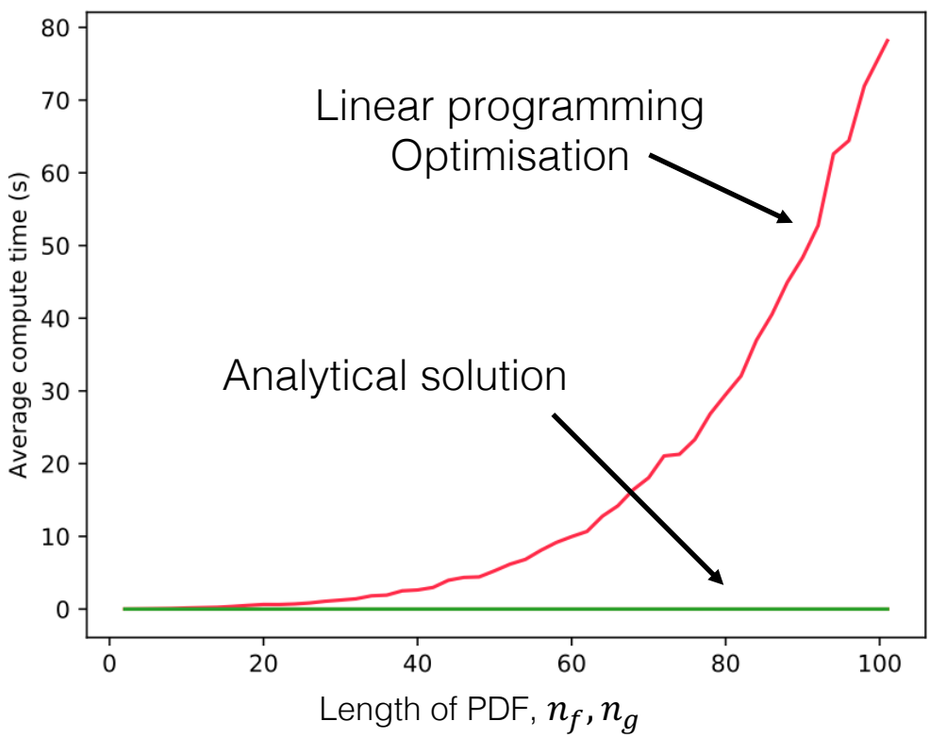
$$W_p(f,g) = [\Delta \mathbf{z}^T \Delta \mathbf{y}]^{1/p}$$

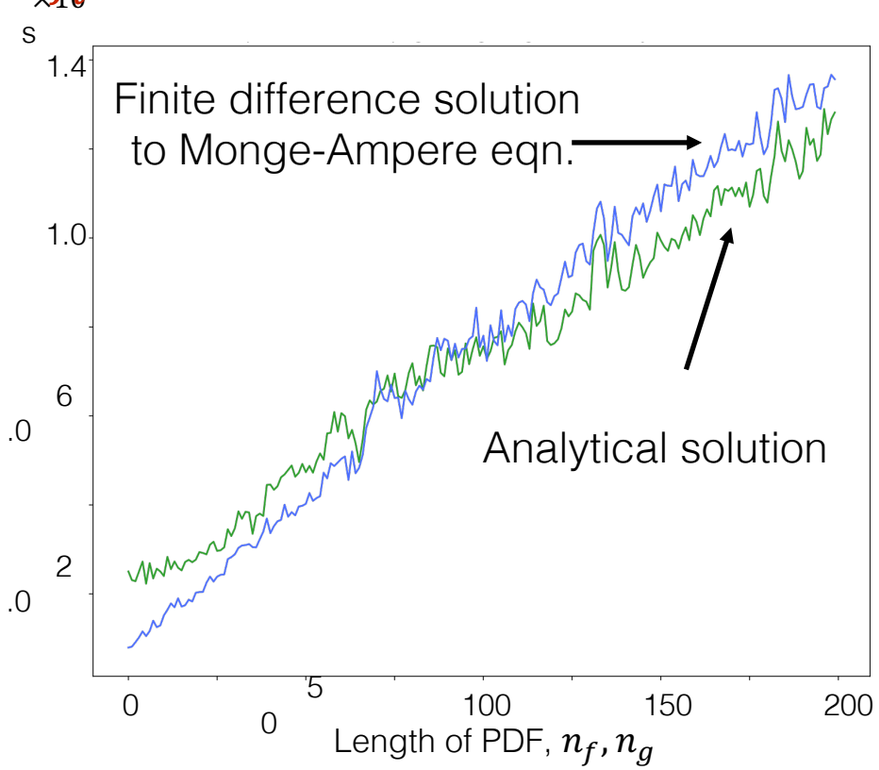
where $\mathbf{z}(x^f, x^g)$ depends only on point mass locations and $\mathbf{y}(f_i, g_j)$ depends only on point mass weights.



Exact for any p. Requires just one sort and a vector dot product. Exact derivatives available, $\partial W_p / \partial f_i$

Relative computational costs

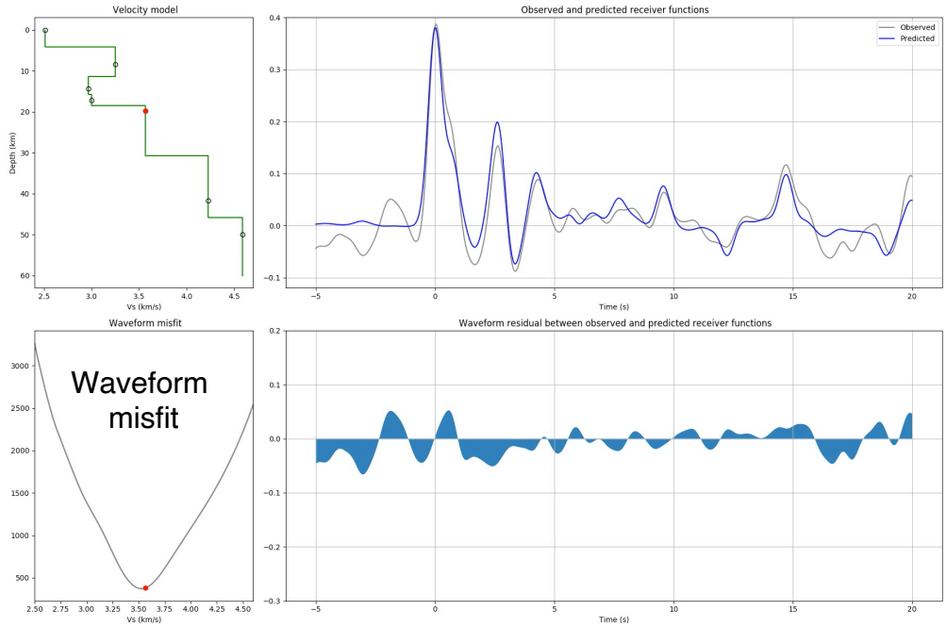




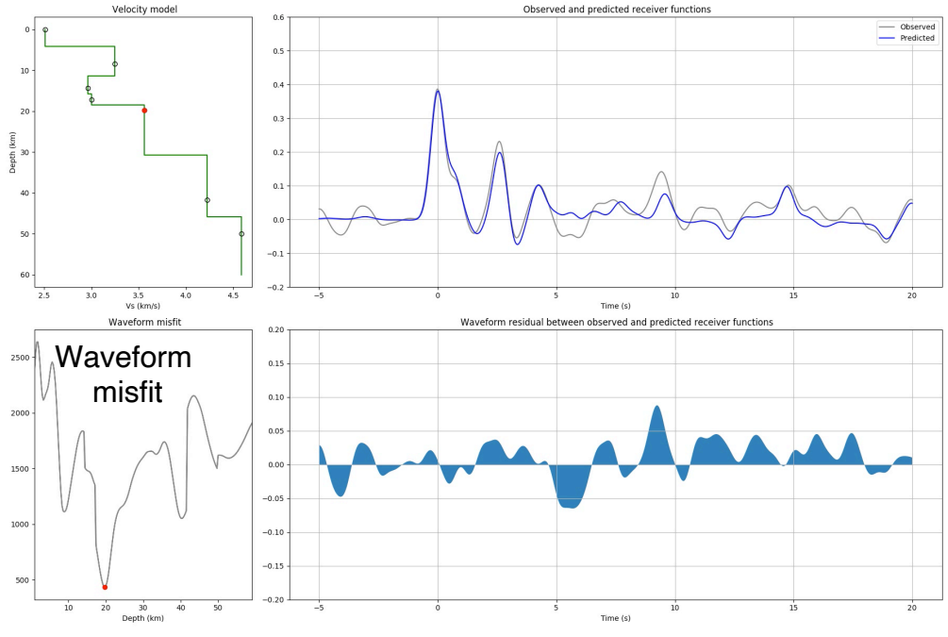
Averaged compute time over 10 random 1D problems

Optimisation challenges with squared waveform misfits

An animation of least squares data misfit vs model and waveform changes

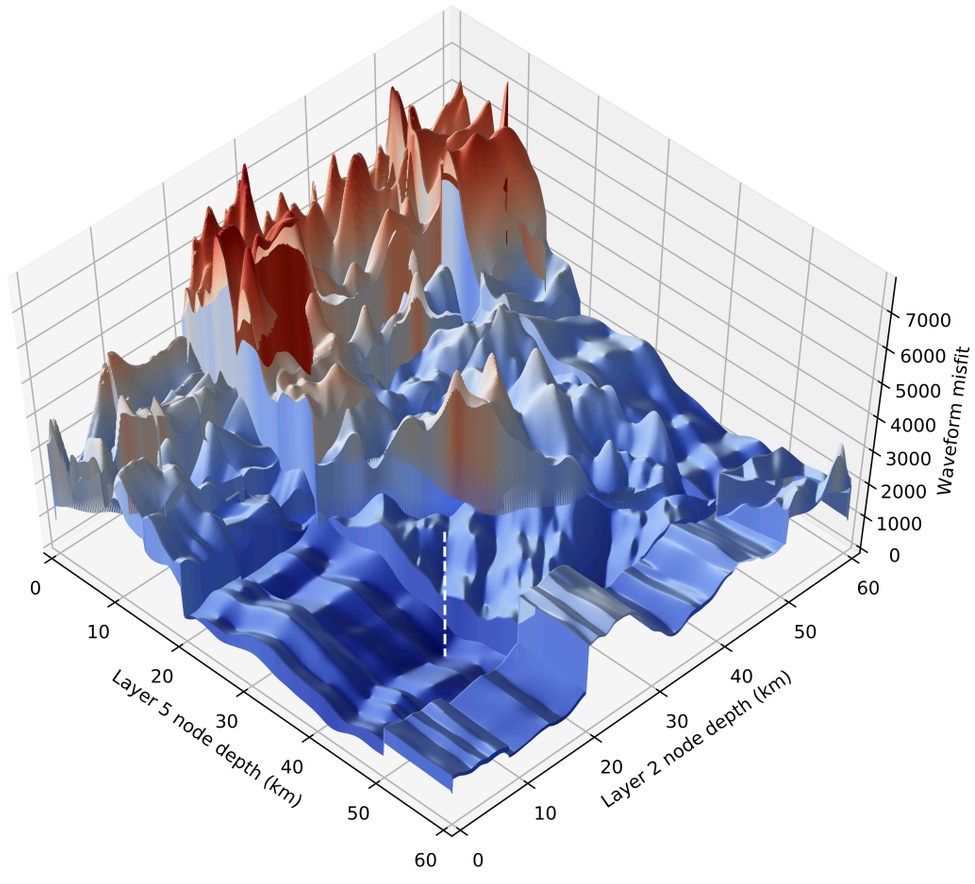


Linearisation will converge from any initial guess with velocity changes



Linearisation fails for interface changes unless we are very close to answer.

Inverse problems are either linear with simple misfit functions or nonlinear with potentially multi-modal misfits - **and sometimes both!**



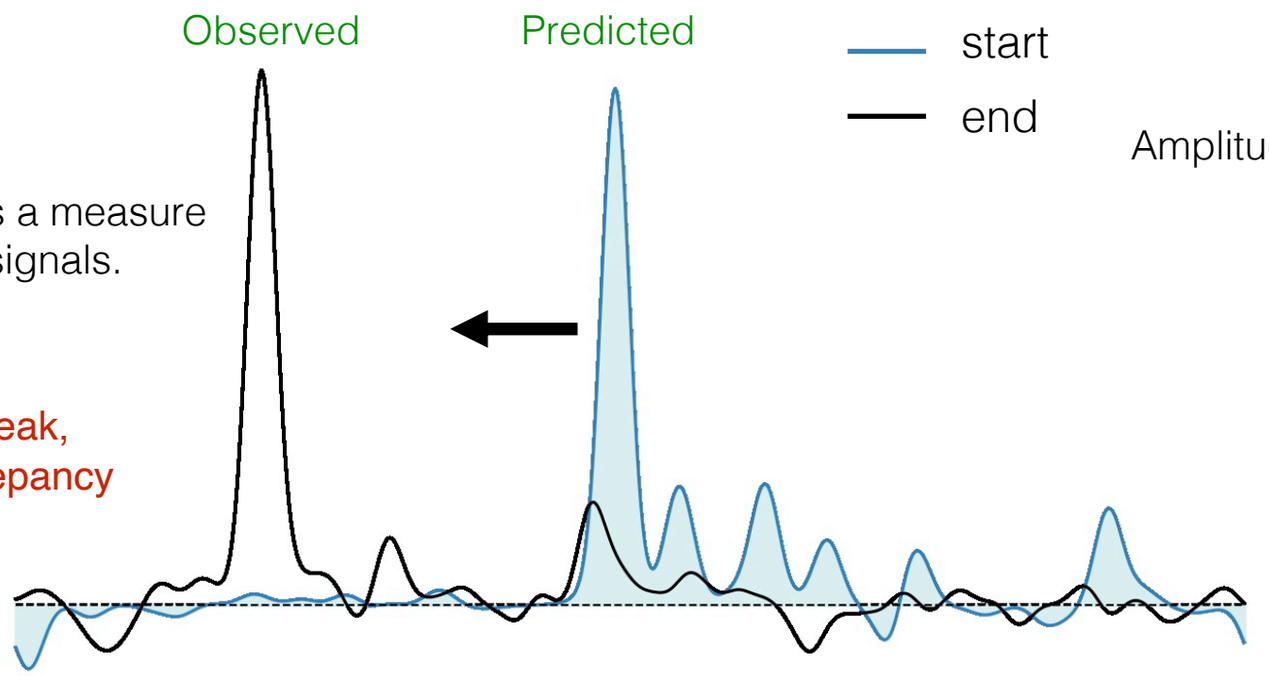
Most nonlinear inversion problems are easy if you start close enough to the minimum.

Waveform misfits: Least Squares and OT

Transformations between two seismic waveforms.

Uses the vertical separation as a measure of distance between two signals.

A least squares misfit does not `**see**' the other peak, just the local amplitude discrepancy

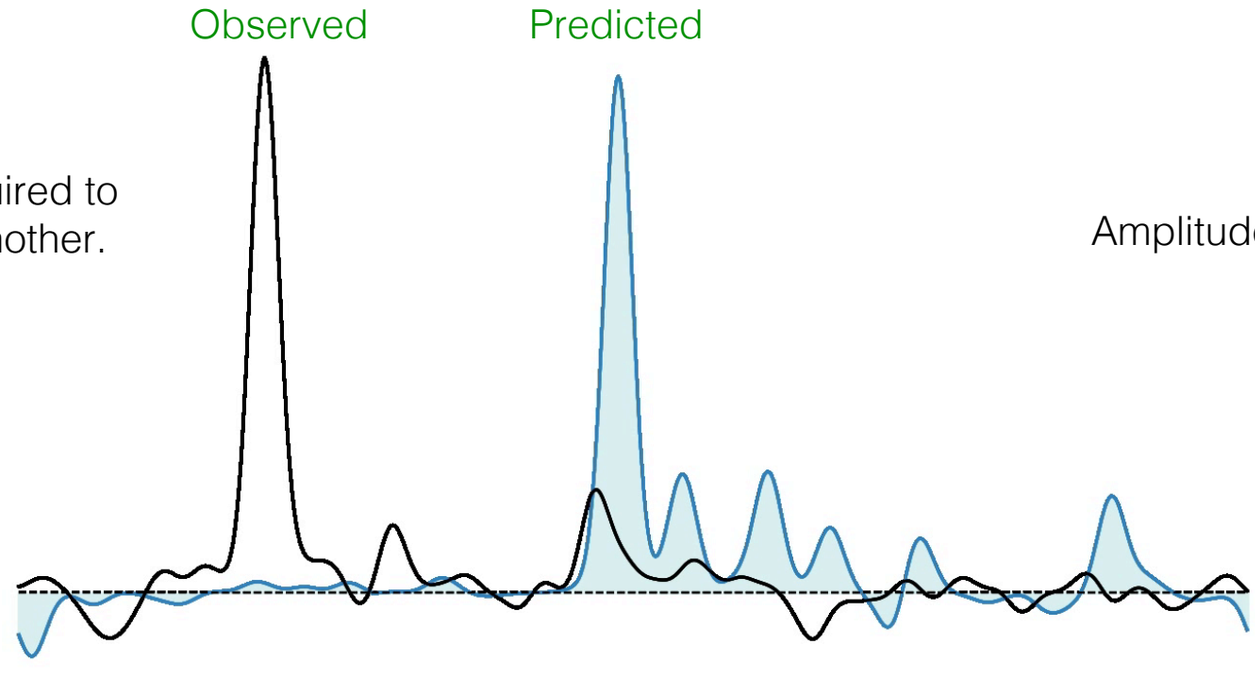


Amplitude, u changes at fixed time, t

$$\int (u_{obs}(t) - u_{pred}(t))^2 dt$$

L_2 has local sensitivity

Uses the minimum work required to transform one signal onto another.



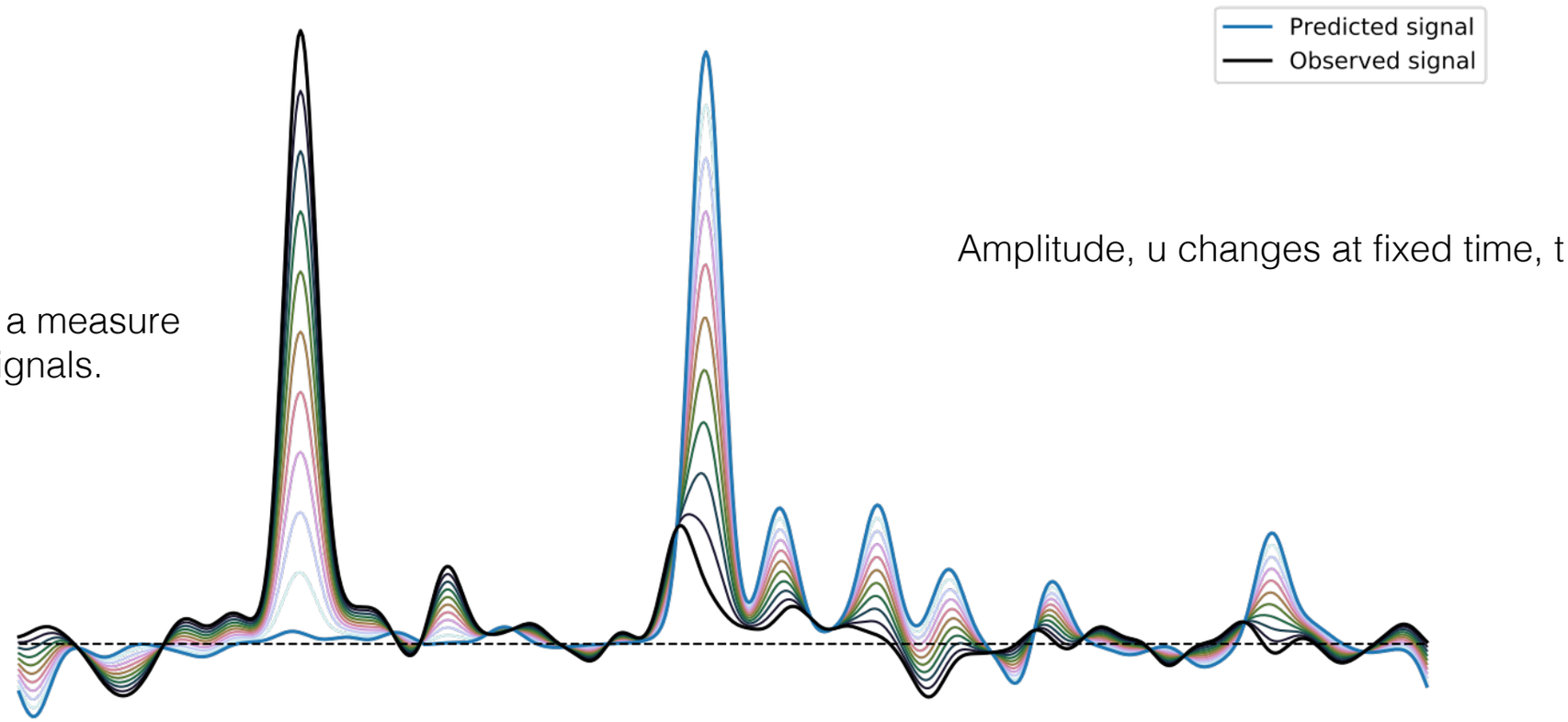
Amplitude, u and time, t change together.

W_2 has global sensitivity

Measuring the distance between complex objects

Least squares

Uses the vertical separation as a measure of distance between two signals.



Amplitude, u changes at fixed time, t

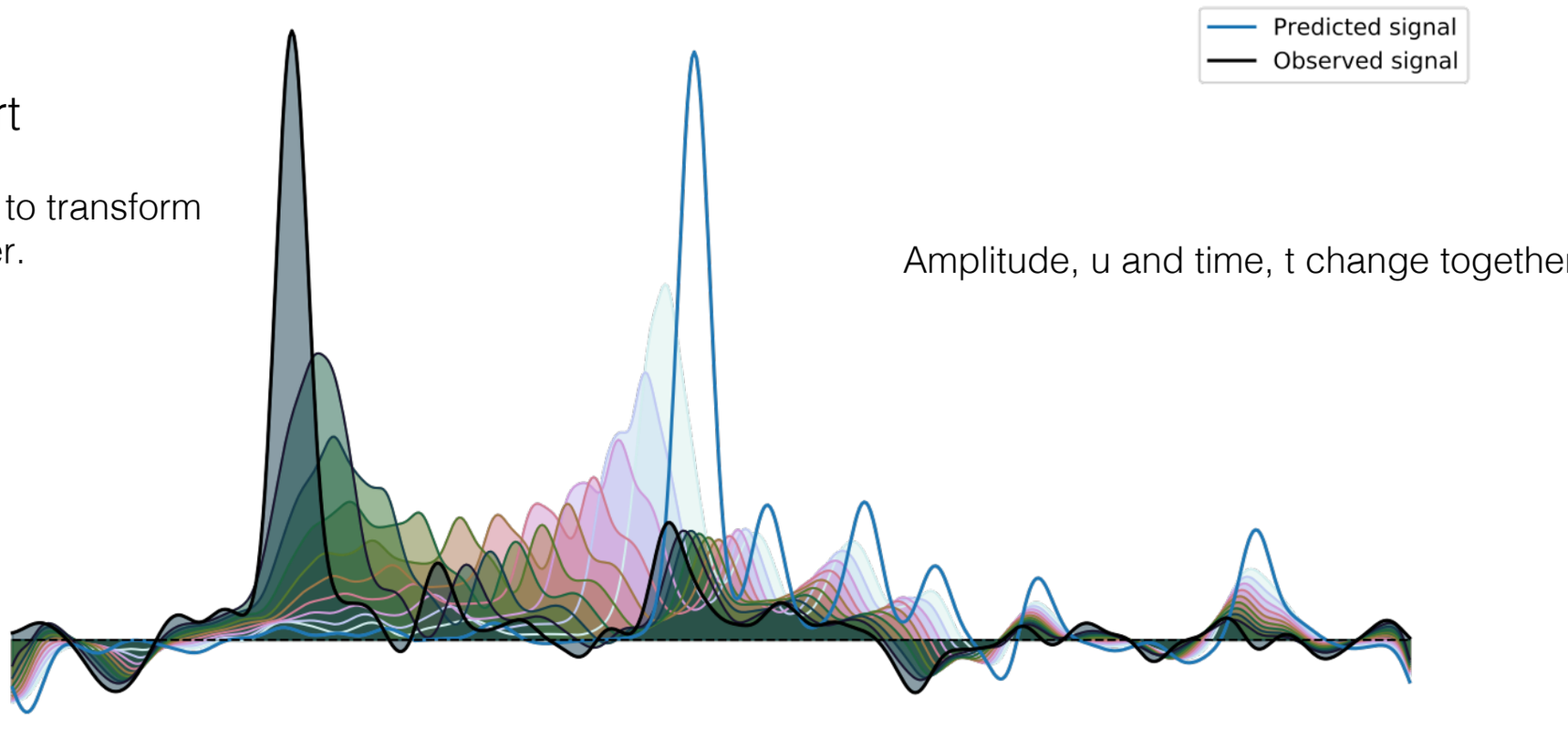
Optimal transport

Uses the minimum work required to transform one signal onto another.

W_1 = distance x mass

W_2^2 = (distance) ^{2} x mass

W_p^p = (distance) ^{p} x mass



Amplitude, u and time, t change together.

Optimal transport in seismic waveform inversion

A number of groups have applied variants of OT in geophysics, primarily to full waveform inversion (FWI) in exploration seismology.

Engquist and Froese (2014); Engquist et al. (2016); - Monge-Ampere PDE solver (p=2)
Yang and Engquist (2018); Yang et al. (2018);

Me´tivier et al. (2016 a,b,c,d); Me´tivier et al. (2018 a,b); - Dual formulation optimisation (p=1)
Me´tivier et al. (2019); Yong et al. (2018)

Hedjazian et al. (2019) - Seismic receiver functions; *Huang et al. (2019)* - Gravity inversion.

Books and lecture notes:

Villani (2003, 2008); Ambrosio (2003); Santambrogio (2015).

Approaches differ between studies:

- Solution method for Wasserstein distance, W_p, and also p value.
- Transform of seismic trace to a Probability Density Function (PDF).
- 1D OT Trace by trace or 2D reflection image.

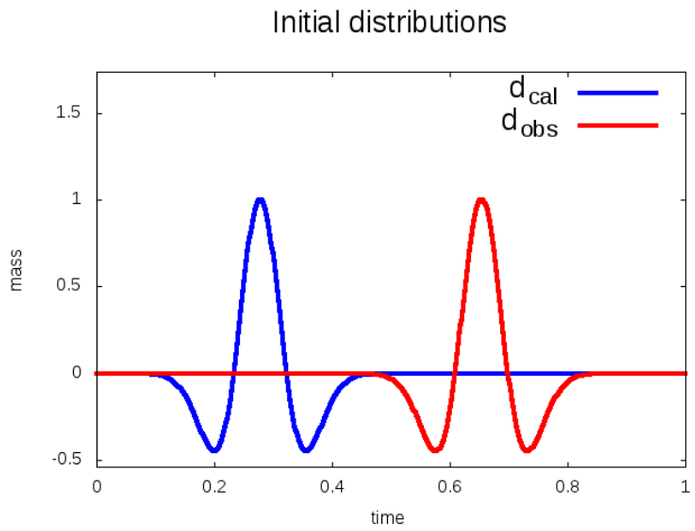
These are all **open** issues. It is an evolving field.

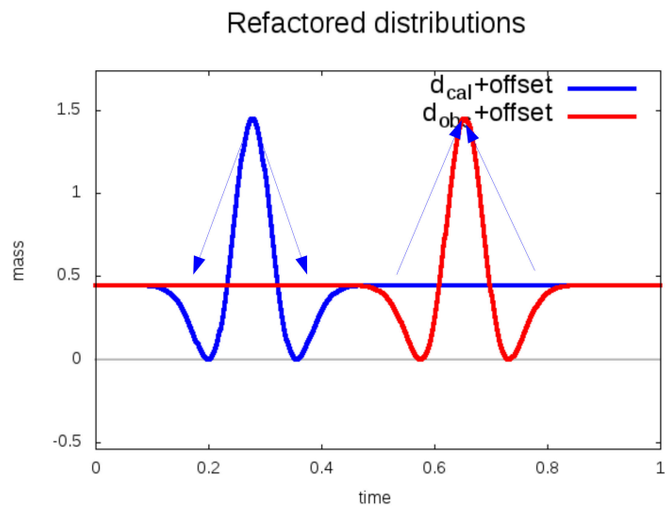
How to make a time series positive?

Several existing approaches in literature to transform a time series to be positive.

- Addition method: Add positive constant to f(x) and g(x)

$$\tilde{W}_p(f,g) := W(f+\alpha, g+\alpha)$$

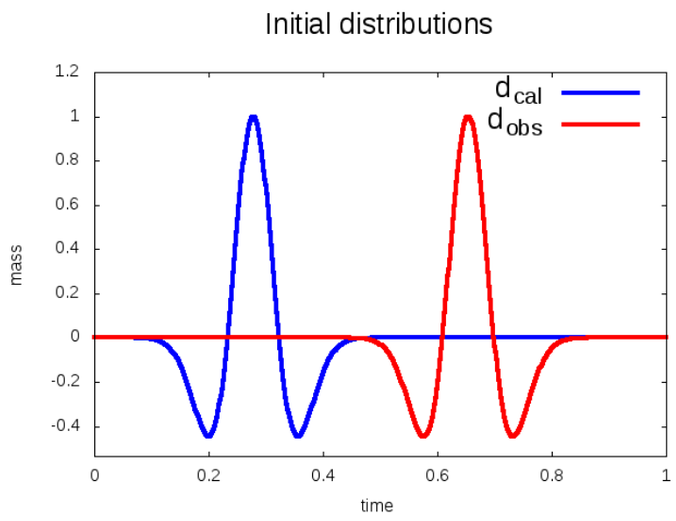


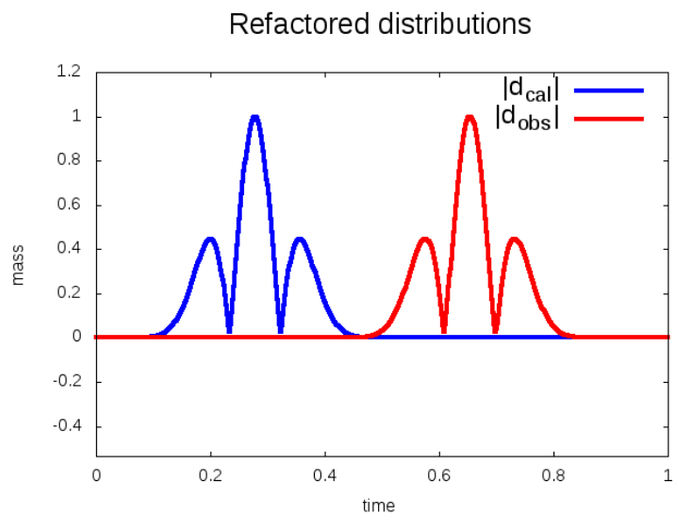


Issues: Loss of convexity w.r.t. time shifts; Transformation becomes local. => Reject.

- Take absolute values. Quite common solution.

$$\tilde{W}_p(f,g) := W(|f|, |g|)$$

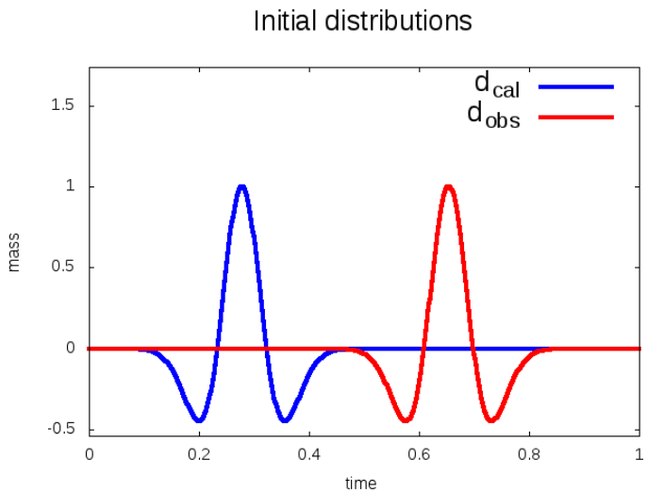


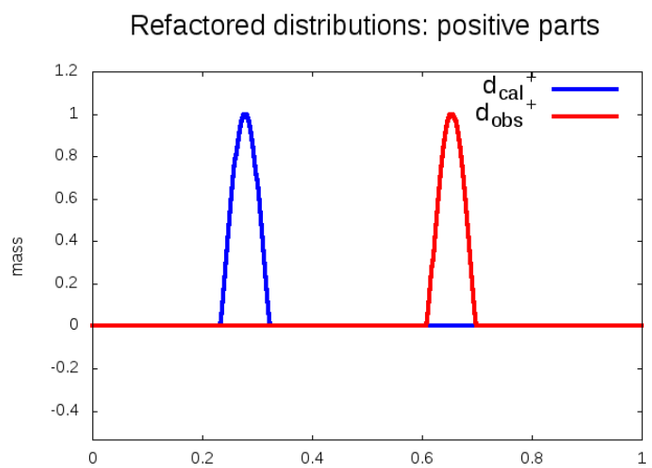


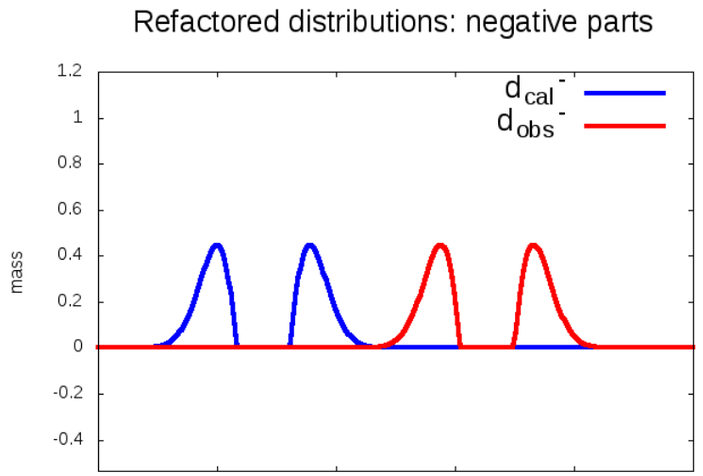
Issues: Loss of polarity information in signal. In FWI results in no sensitivity to impedance contrasts.

How to make a time series positive - 2

- Like with like: Separately transform +ve to +ve and -ve to -ve. $\tilde{W}_p(f,g) := W_p(f^+,g^+) + W_p(f^-,g^-)$

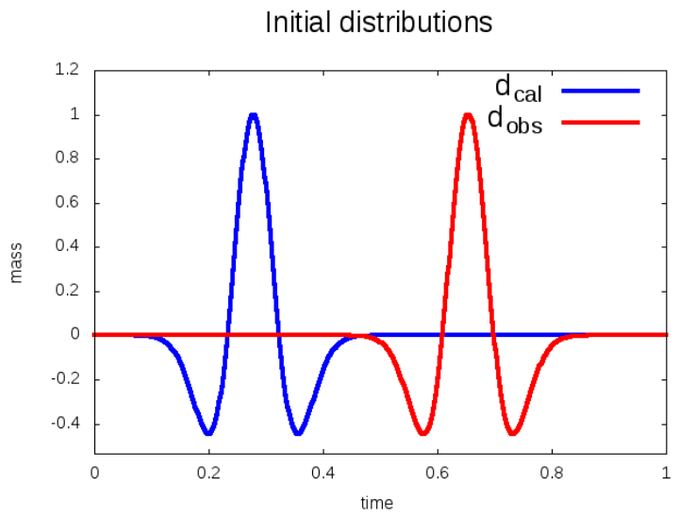


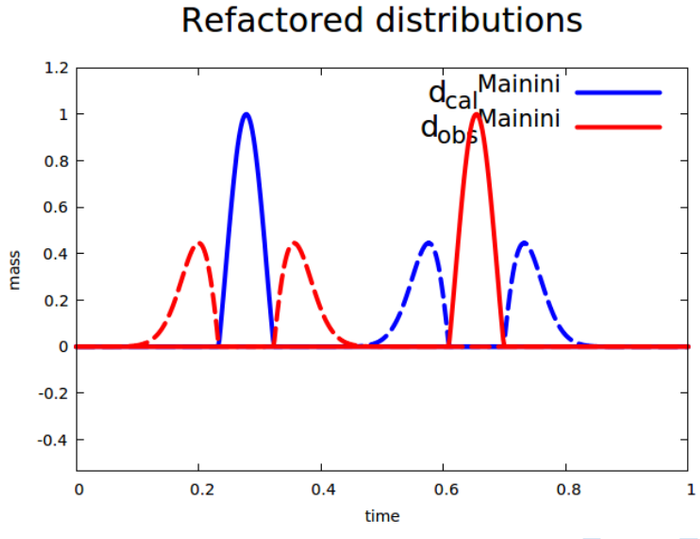




Issues: Loss of mass conservation; Artificial decorrelation between +ve and -ve parts.

- Global strategy: Mix +ve and -ve parts between f and g. $\tilde{W}_p(f,g) := W_p(f^+ + g^-, f^- + g^+)$





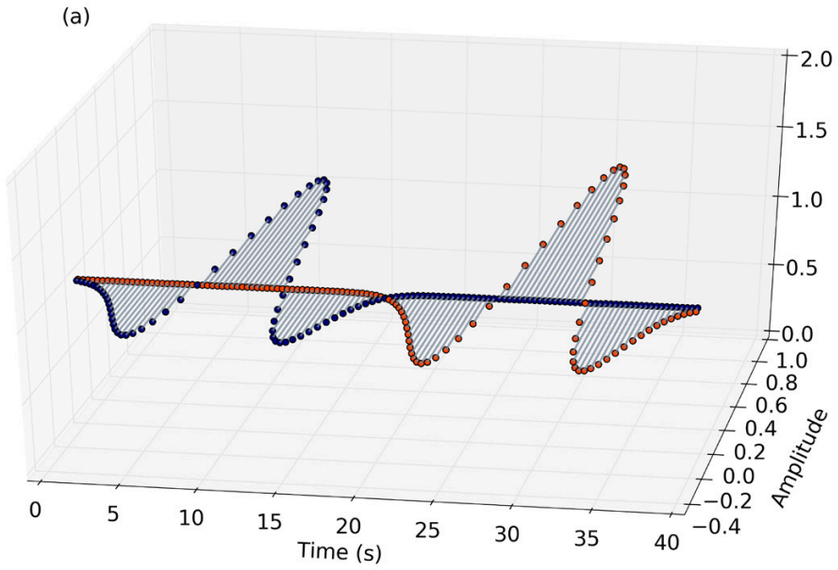
Issues: Ensures preservation of mass conservation;
If time signals are separated in time will map f+ to f- and g+ to g-.

From Engquist and Froese (2014)

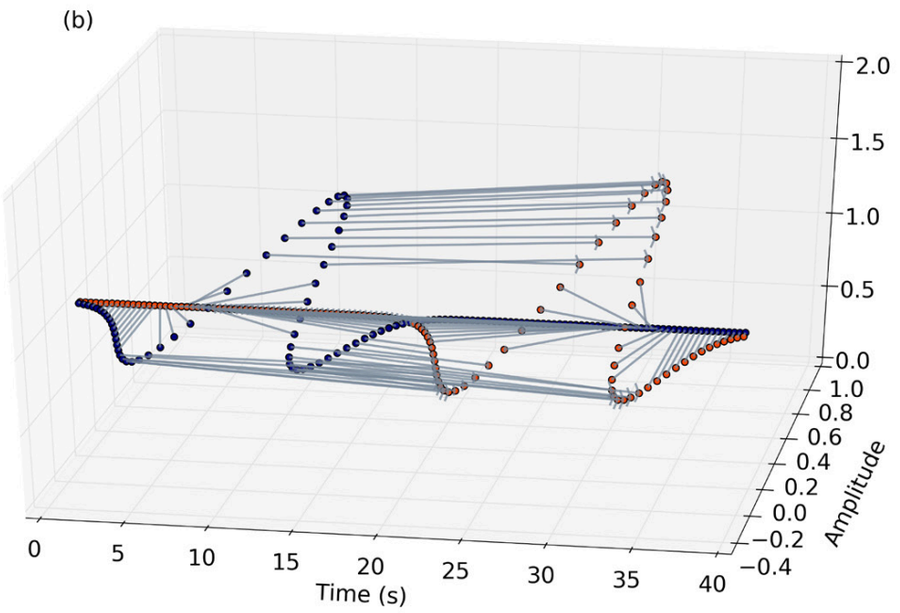
…and yet another

- A new idea by Metivier et al (2018) Graph Space OT is to replace the time series by a point cloud of 2D points distributed along f(t) and g(t). This creates a 2D point cloud where the OT problem is to map from one to the other.

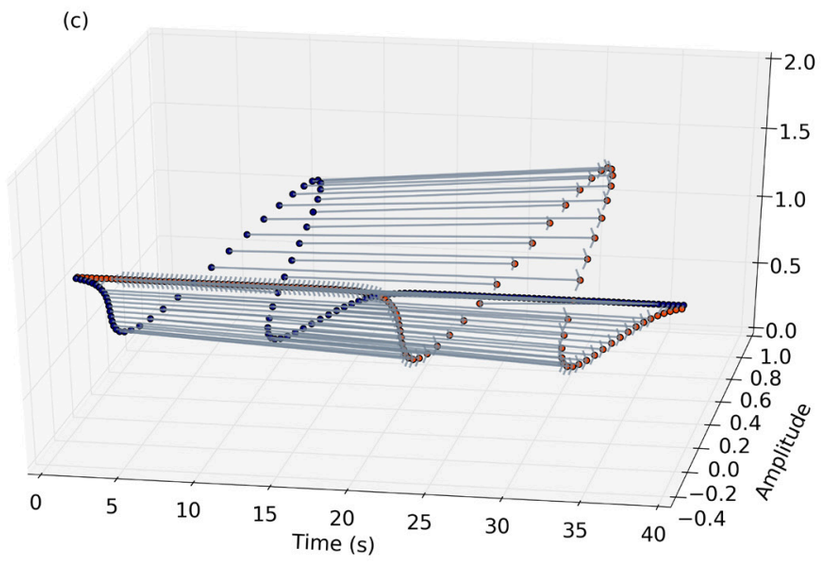
Initial



One GSOT solution



Another GSOT solution

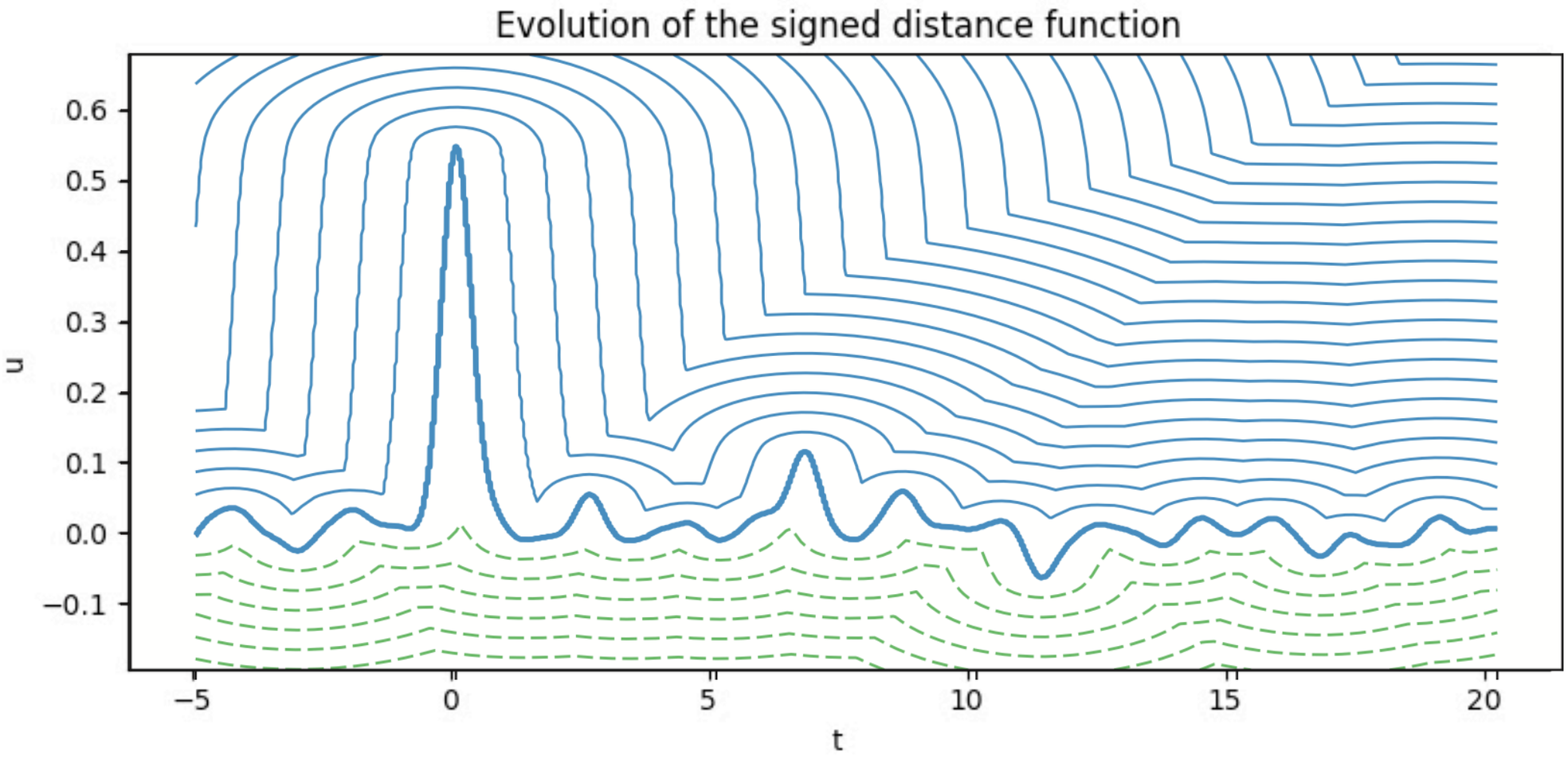


Different solutions are obtained depending on choices of distance scaling (which seems unavoidable).

Note how (b) seems to map some points of the blue signal ***to the zero line of the red signal*** while (c) does not. This would appear to be a failing of the approach as its more a characteristic of L2, than Wp.

Our way: Create a 2D `Fingerprint' from 1D waveform

Step 1: Rather than treating the +ve and -ve parts of a time series, $u(t)$, differently, we create a 2-D positive function, $d(u,t)$, which is the minimum distance from (u,t) to the waveform.

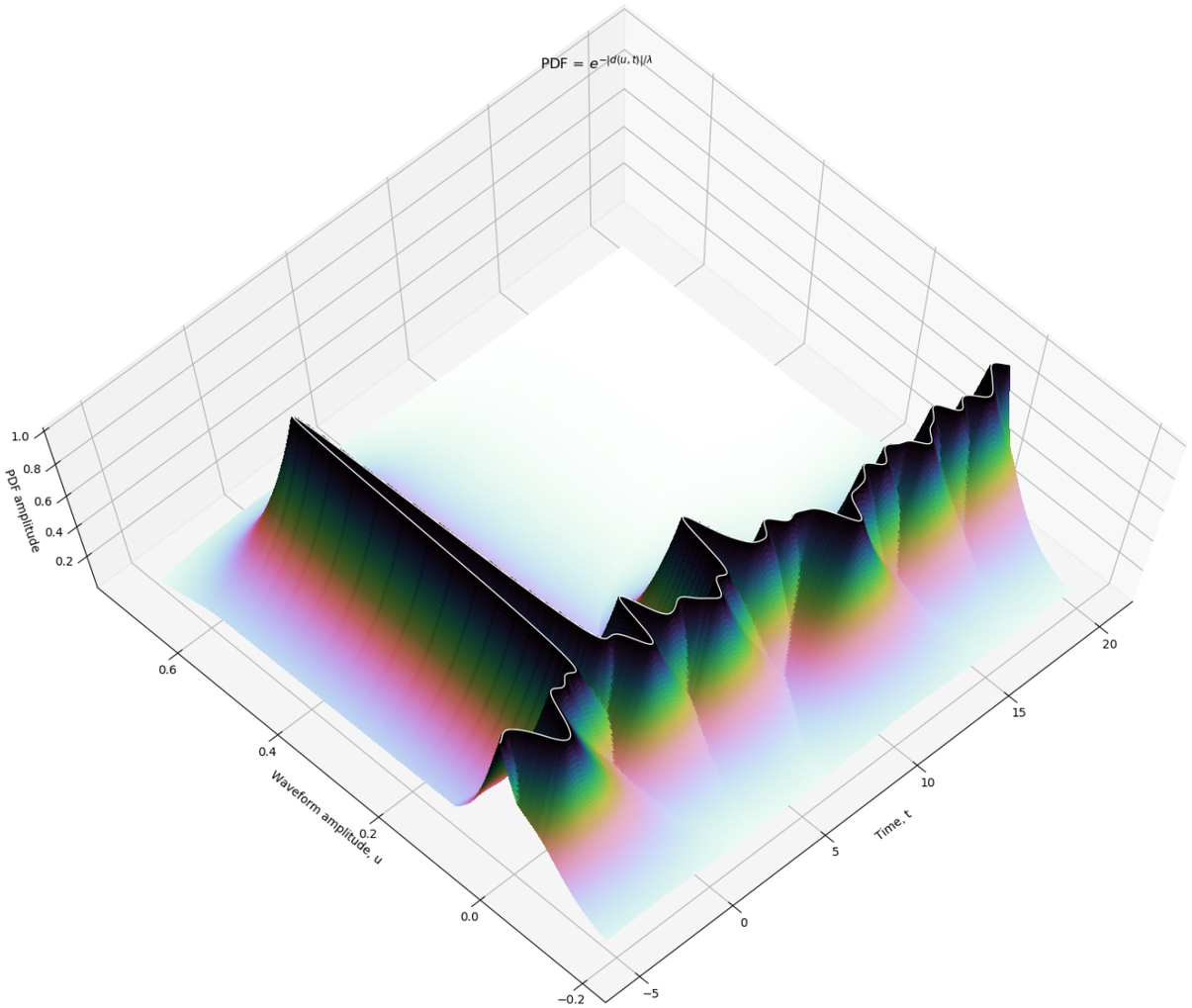


Seismologists will recognise this calculation because it is identical to that of propagating a seismic wavefront, initially in position $u(t)$, both upward and downward according to Huygens' p

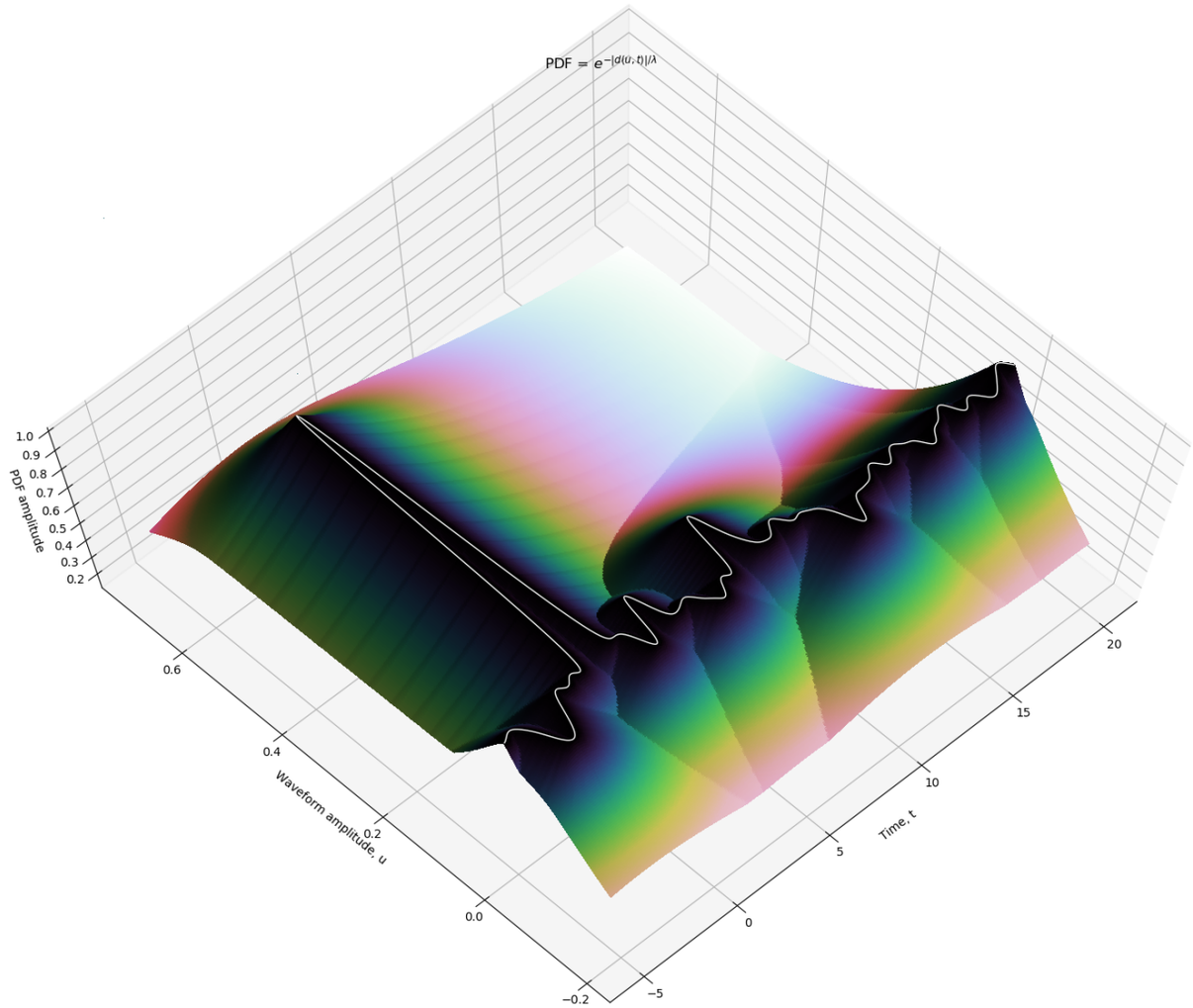
Efficiently done with the Fast Marching method of Sethian (1996,1999); Sethian and Popovici, (1999); Rawlinson and Sambridge, (2004).

Our way: Create a 2D `Fingerprint' from 1D waveform

Step 2: Take the exponential of the distance function, $\phi(u,t) = e^{-d(u,t)/\lambda}$



Small λ



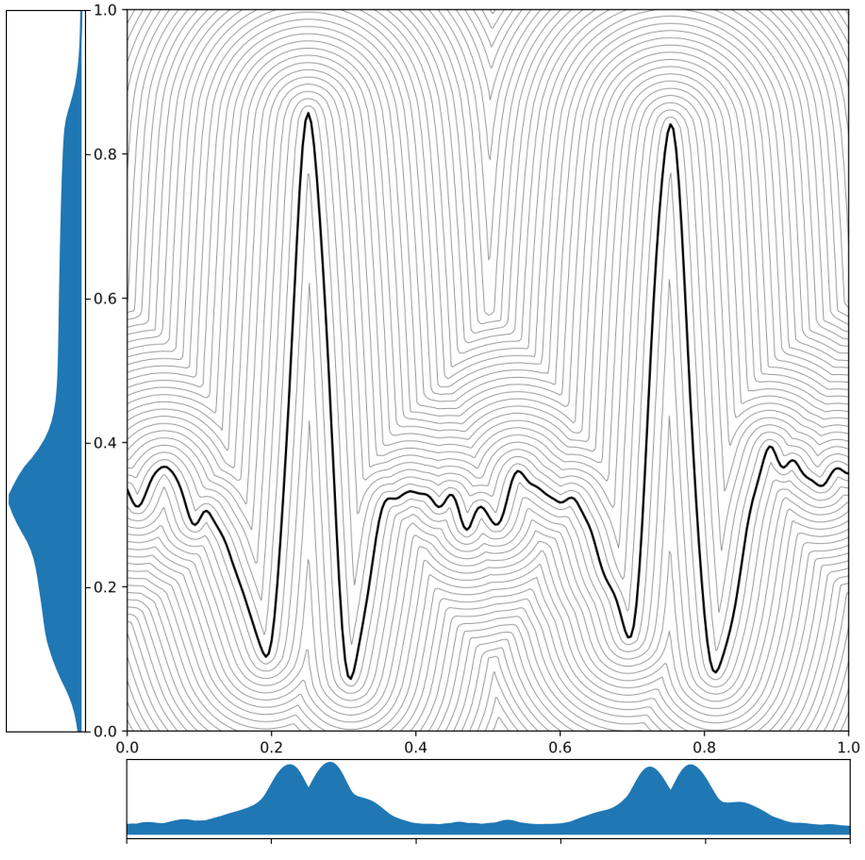
Large λ

Marginal Wasserstein in 2D

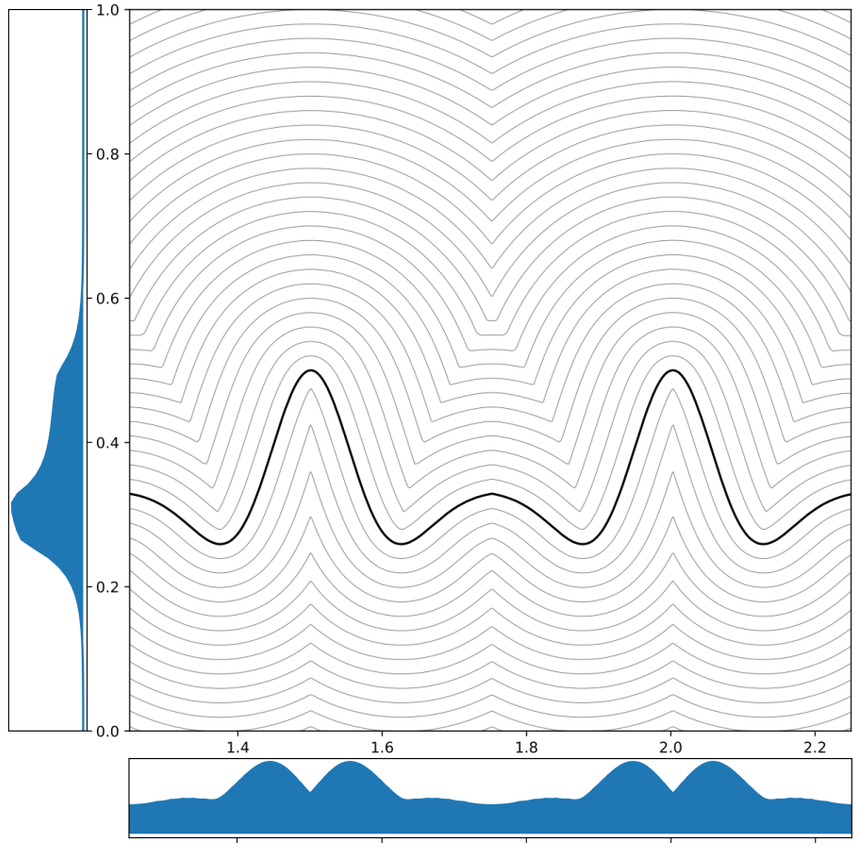
Our waveform misfit becomes the Wasserstein distance between observed and synthetic PDFs

We sum over each axis, and average Wasserstein distances between 1D Marginals.

Observed fingerprint



Synthetic fingerprint

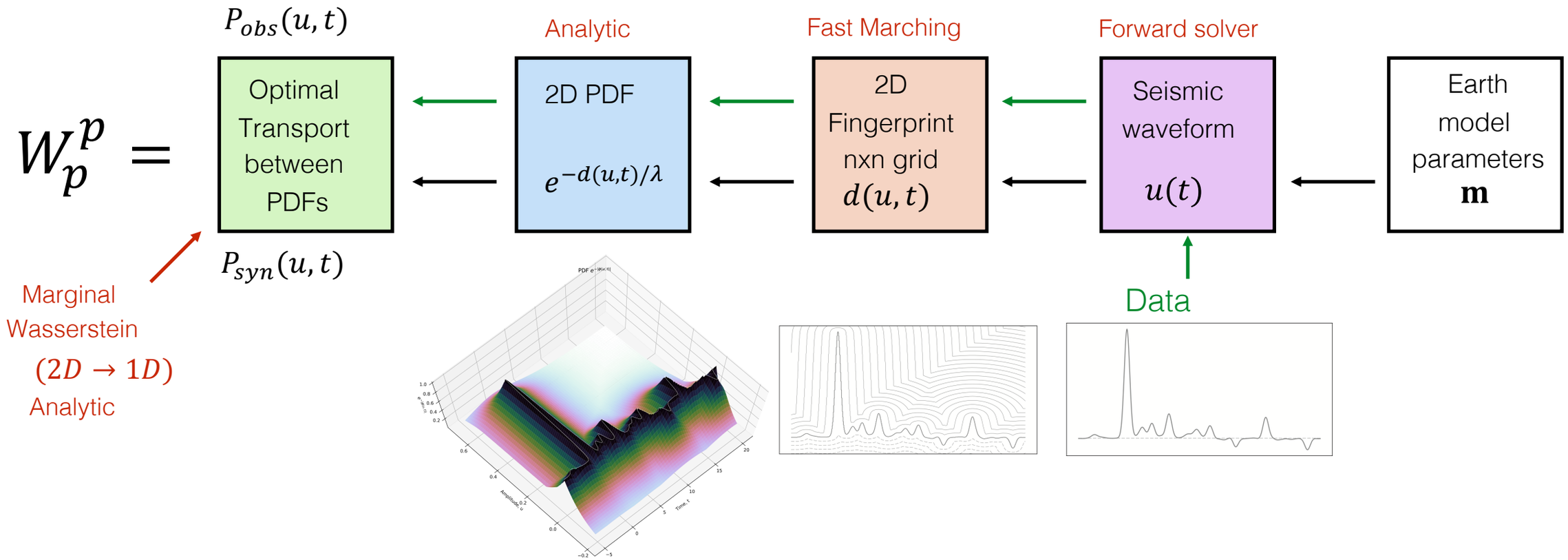


Advantages:

- Takes advantage of 1D analytical OT solutions.
- Handles different time windows about predicted and observed waveforms.
- Computational cost scales with n rather than n^2 for $n \times n$ grid.
  Faster than Sliced Wasserstein with similar results.
- Derivatives $\partial W_p^p / \partial u$ can be calculated.

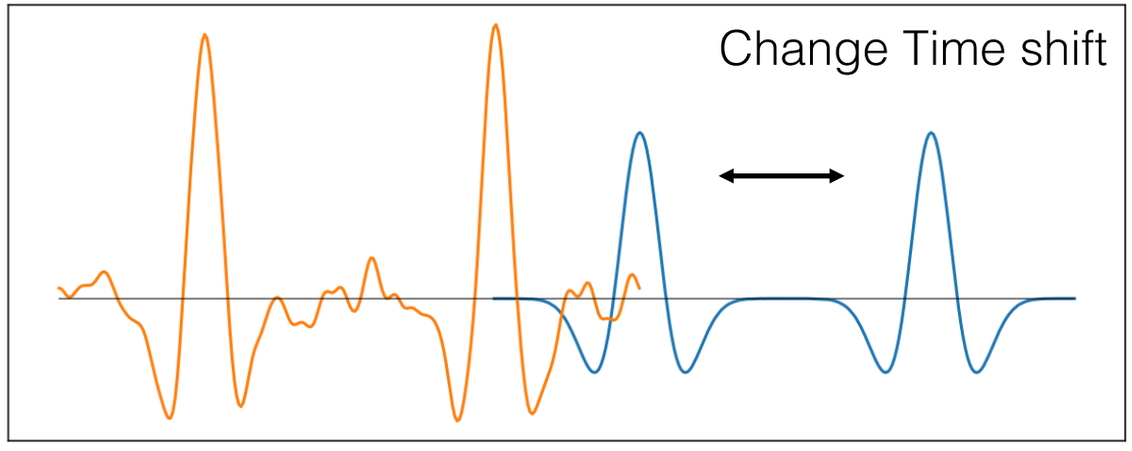
Computation of the Wasserstein distance between seismic fingerprints

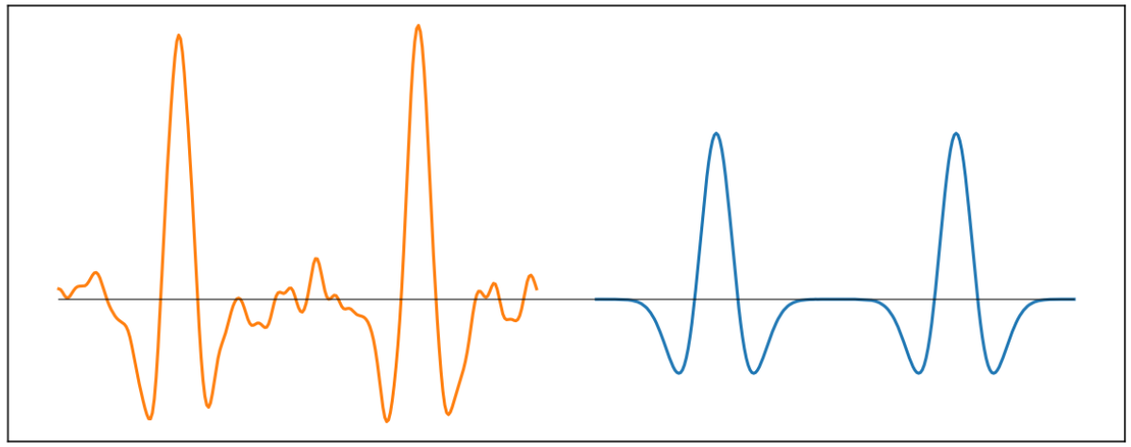
Breaking down the calculation into 4 steps:



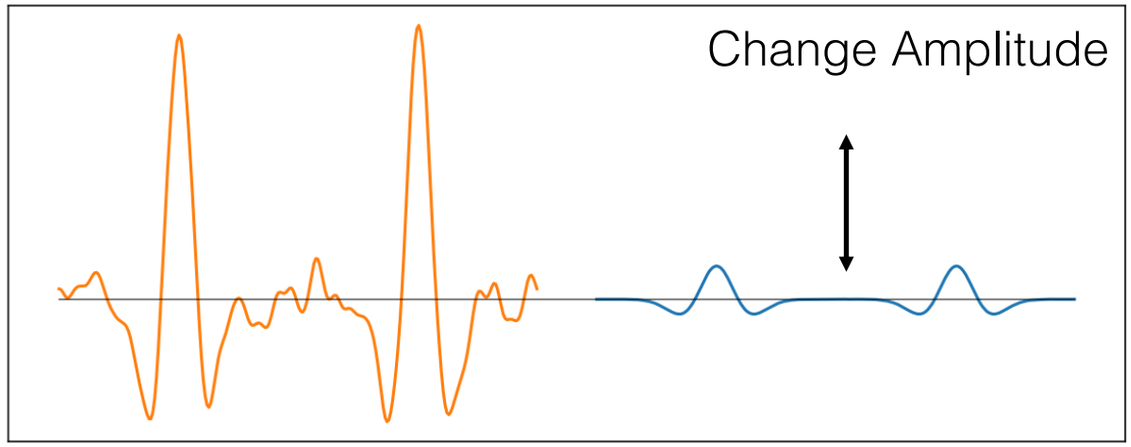
Example: Double Ricker wavelet fitting

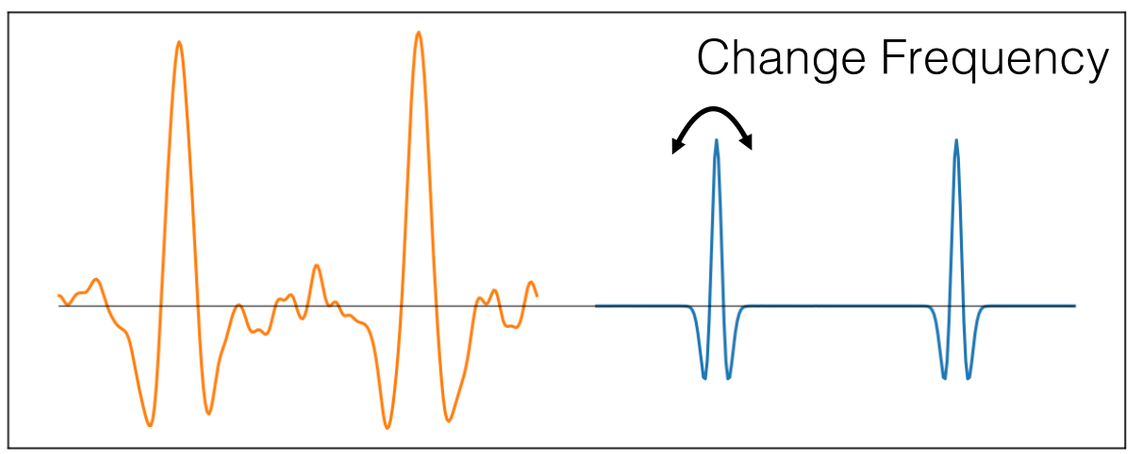
Fit the noisy waveform by adjusting three parameters





Observed Predicted



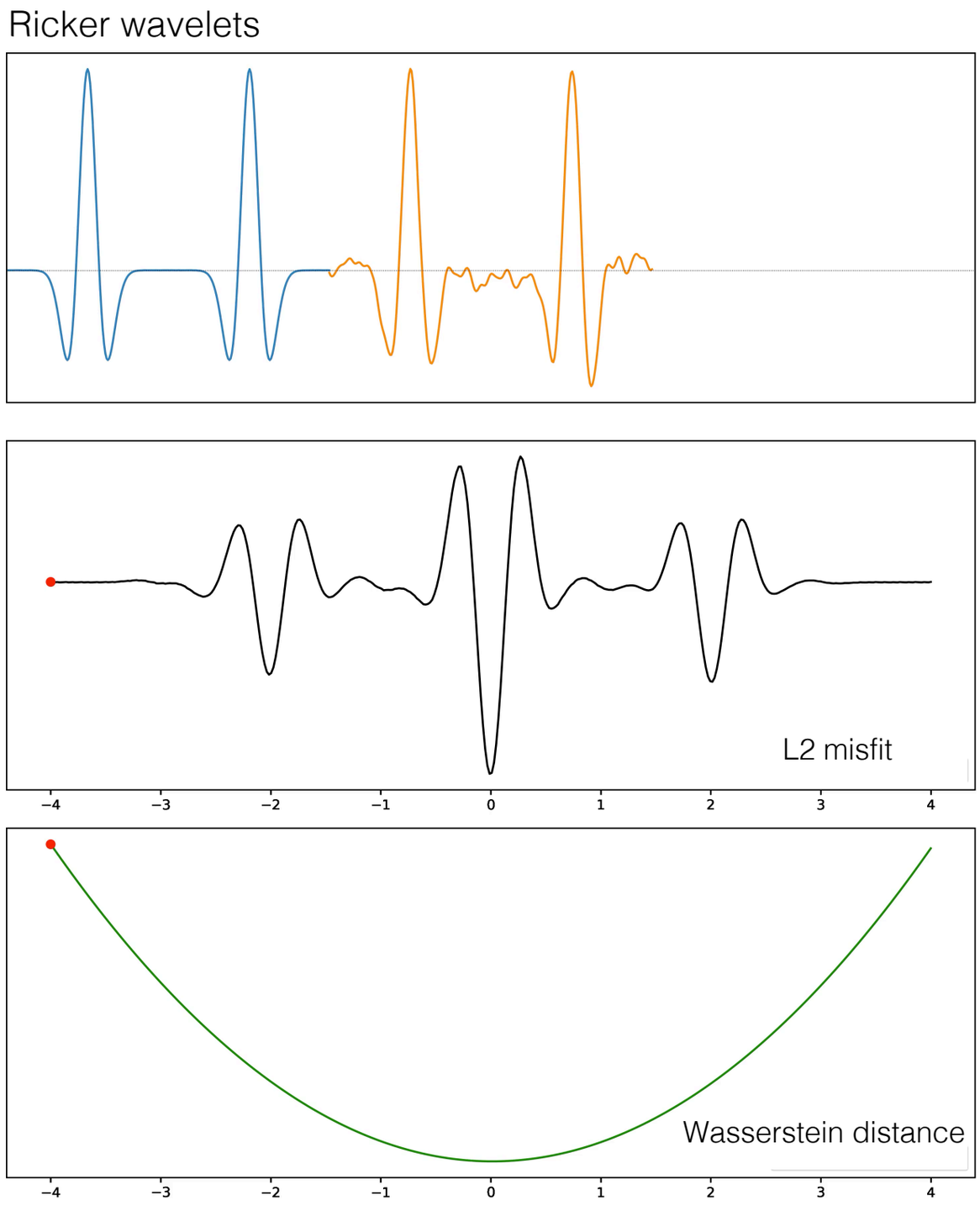


Noise is $N(\mu, \sigma^2)$

μ = 5% of maximum Ricker amplitude

σ = 50% of maximum Ricker period

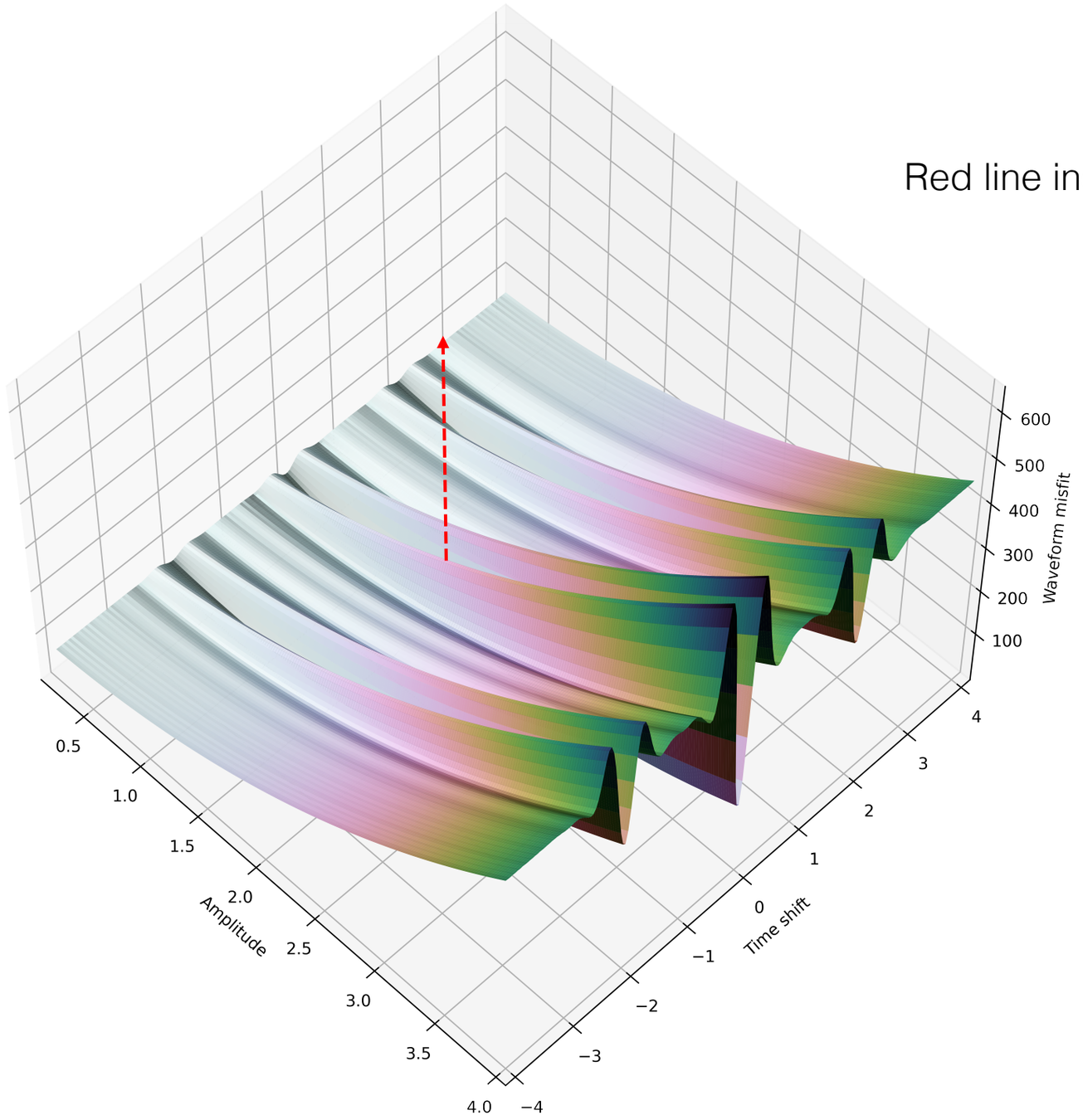
Least squares misfit and Wasserstein distance between a pair of double Ricker wavelets



After Engquist and Froese (2014)

L2 waveform misfit surface

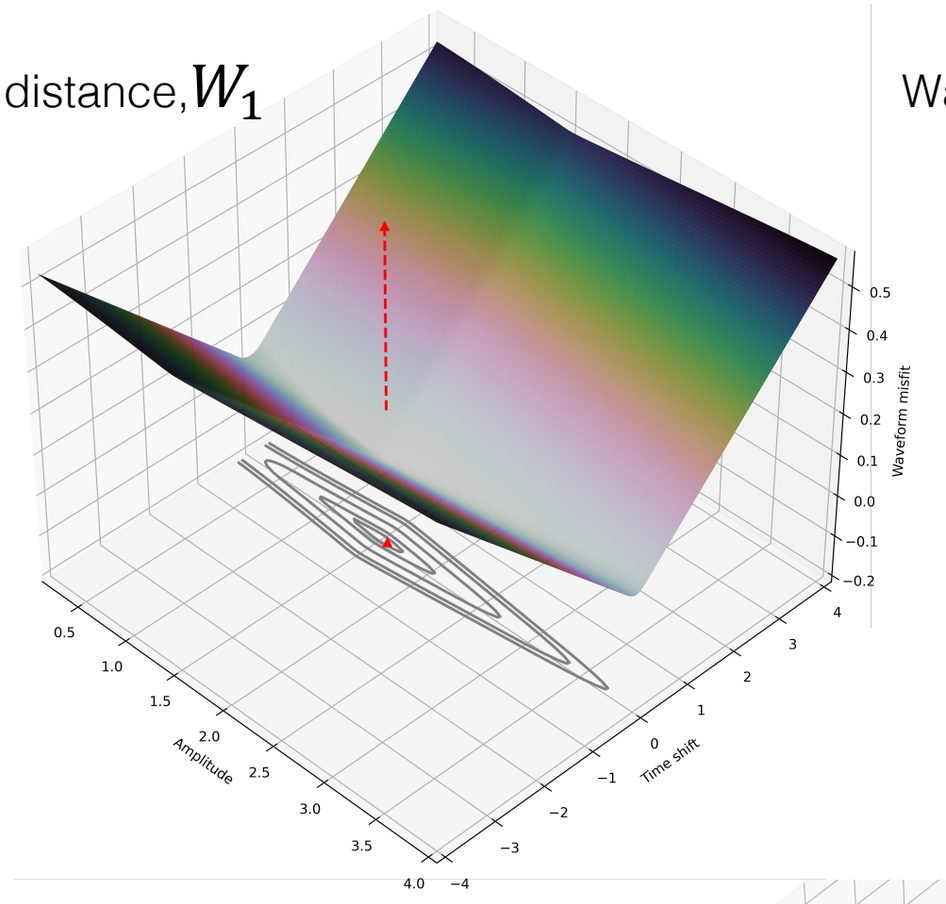
Red line indicates global minimum



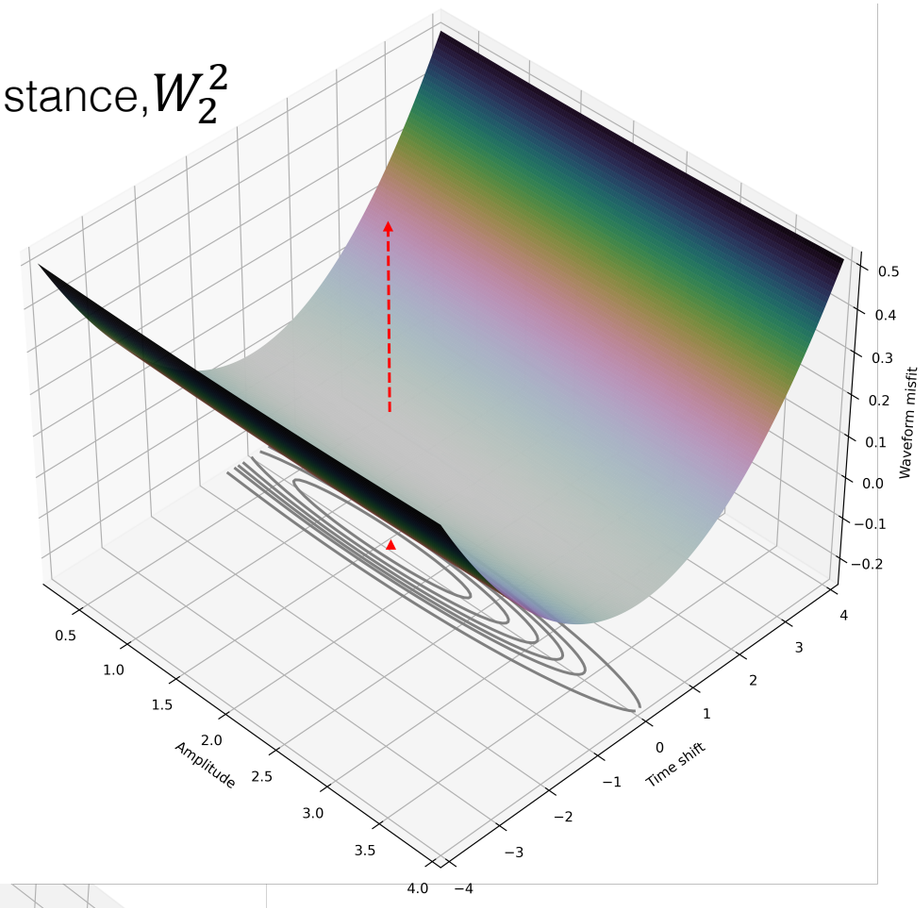
Least squares waveform misfit as a function of
Time shift and Amplitude parameters

Wasserstein and L2 waveform misfit surfaces

Wasserstein distance, W_1

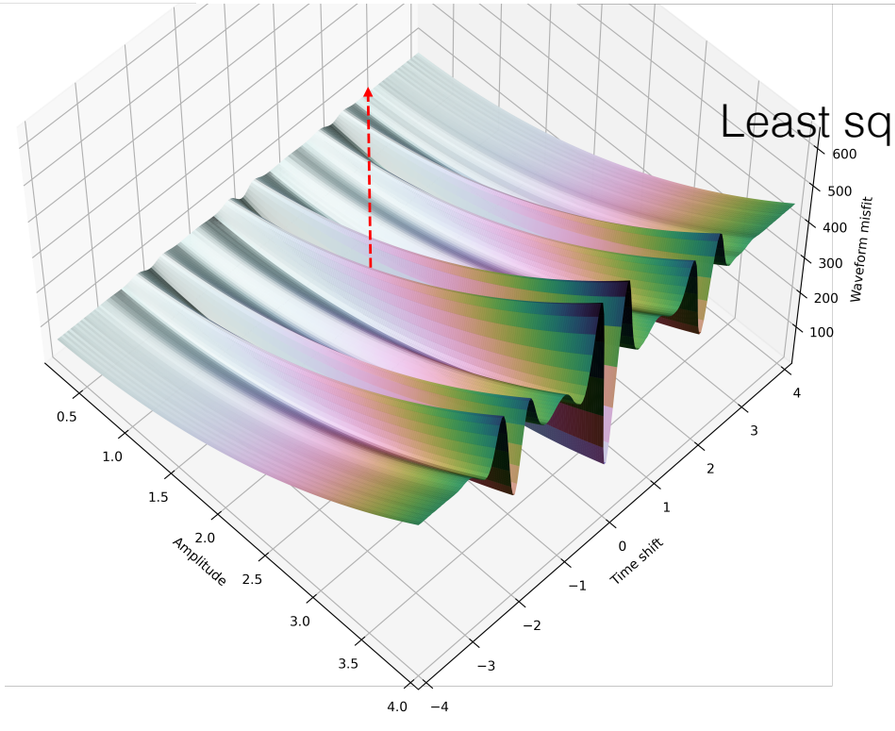


Wasserstein distance, W_2^2



Wasserstein based on
2D fingerprint PDF

Least squares misfit, L_2



Red line indicates global minimum

Calculating derivatives of Wasserstein distance

For the Wasserstein distance to be useful within an optimisation framework derivatives of the Wasserstein distance, $\frac{\partial W_p^p}{\partial m_j}, (j = 1, \ldots, M)$, with respect to underlying Earth model parameters are required. (***This is the discrete analogue of the Adjoint method in waveform inversion)***

The **chain rule** must be applied for each intermediate variable

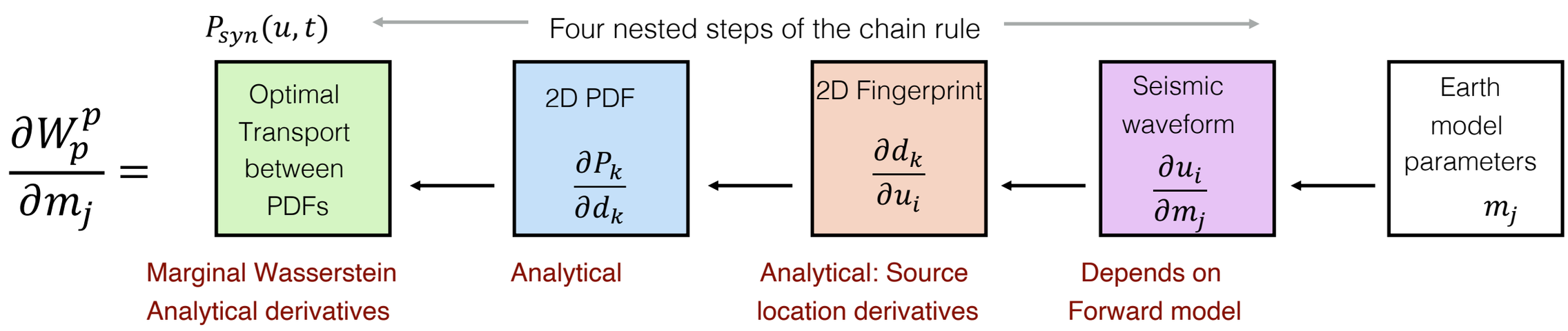
$$\frac{\partial W_p^p}{\partial m_j} = \sum_k \frac{\partial W_p^p}{\partial u_k} \frac{\partial u_k}{\partial m_j} (j = 1, \ldots, M).$$

One step of the chain rule

Other authors implementing OT to FWI have all done something similar with details depending on the choices made in applying Optimal Transport, e.g. Finite Difference solution of Monge-Ampere equations (Enquist and co workers, 2014-)

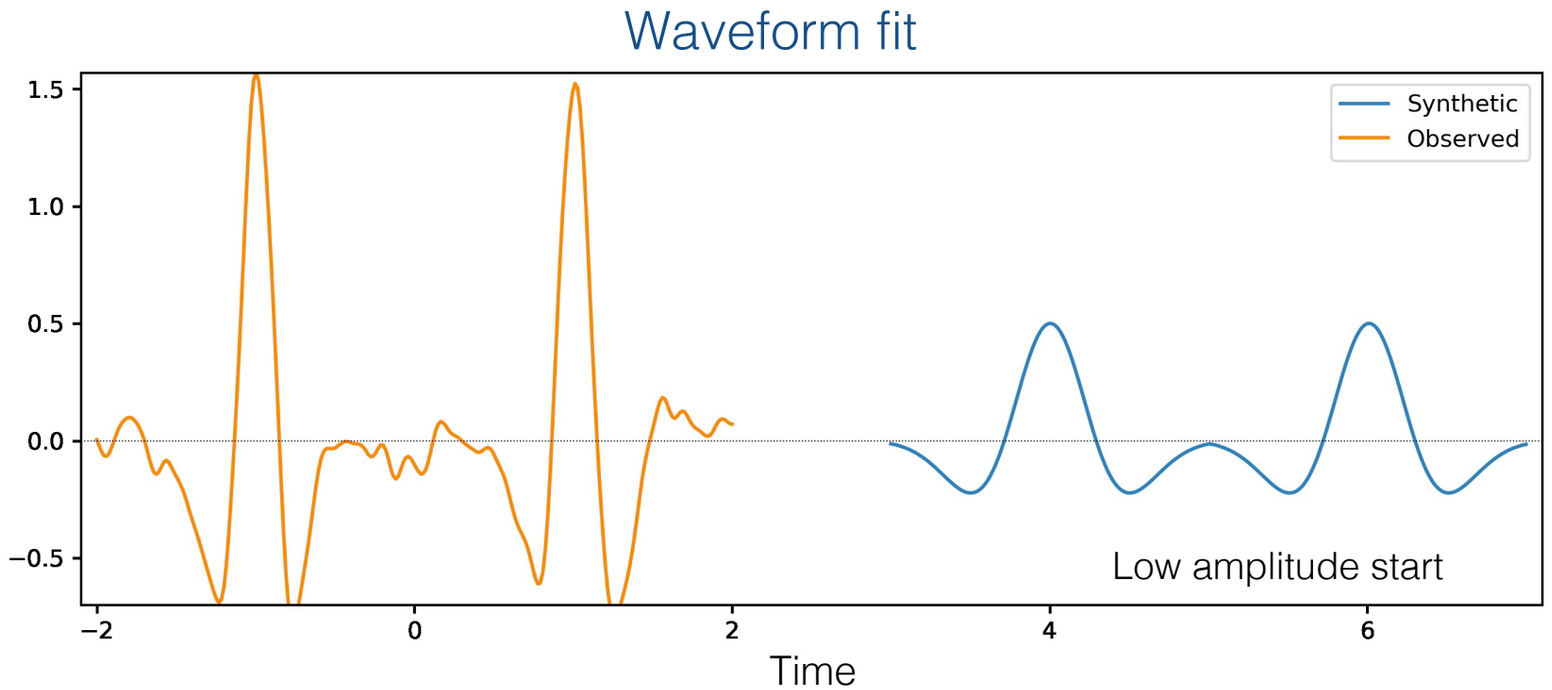
or constrained optimisation using the Monge-Kantorovich dual formulation (Metivier and co-workers 2016-).

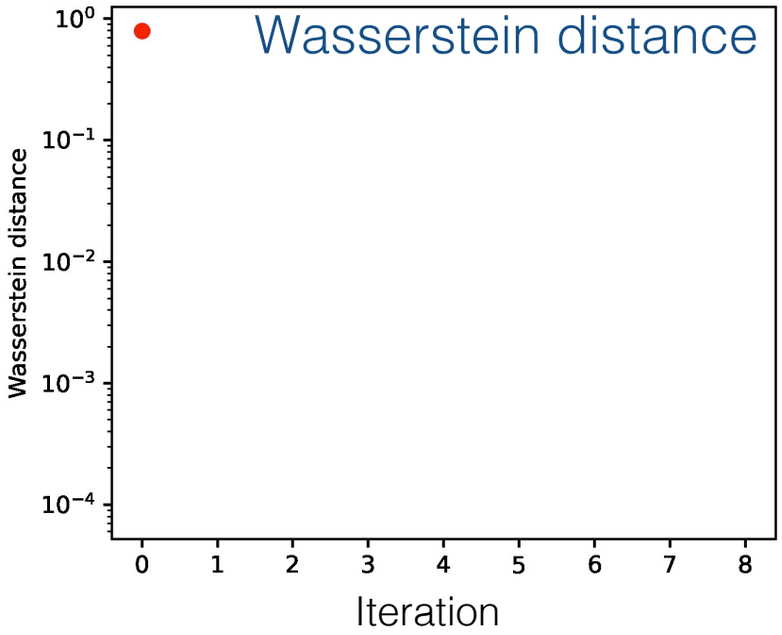
In our case



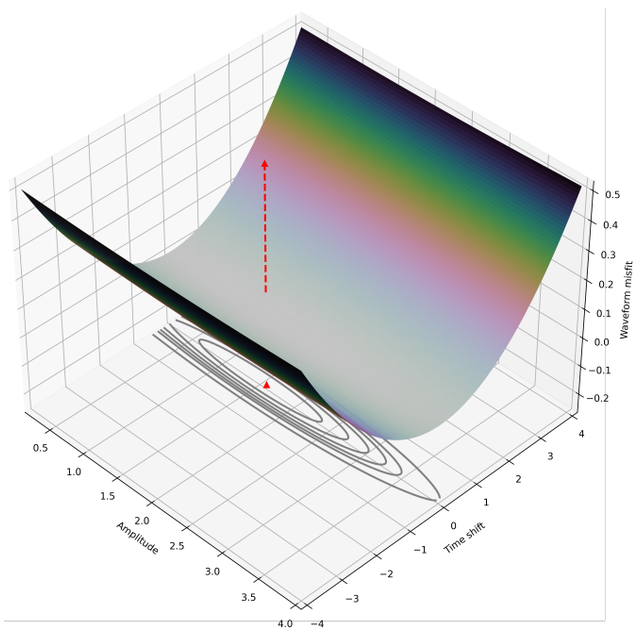
Effectiveness of optimisation requires accurate derivatives at every step. Analytical derivatives are **exact**!

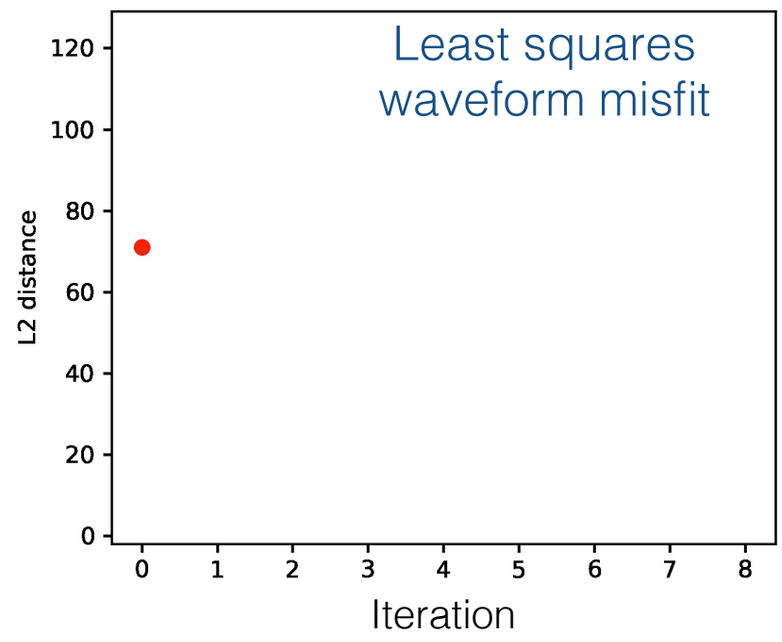
Minimizing the Wasserstein distance W_2^2





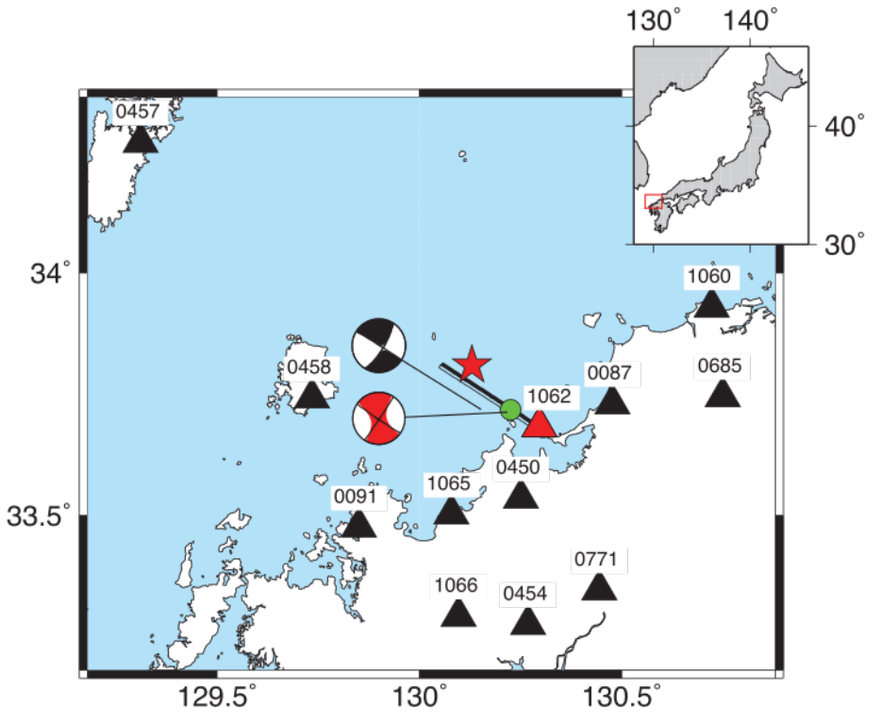
Wasserstein based on
2D PDF of fingerprint





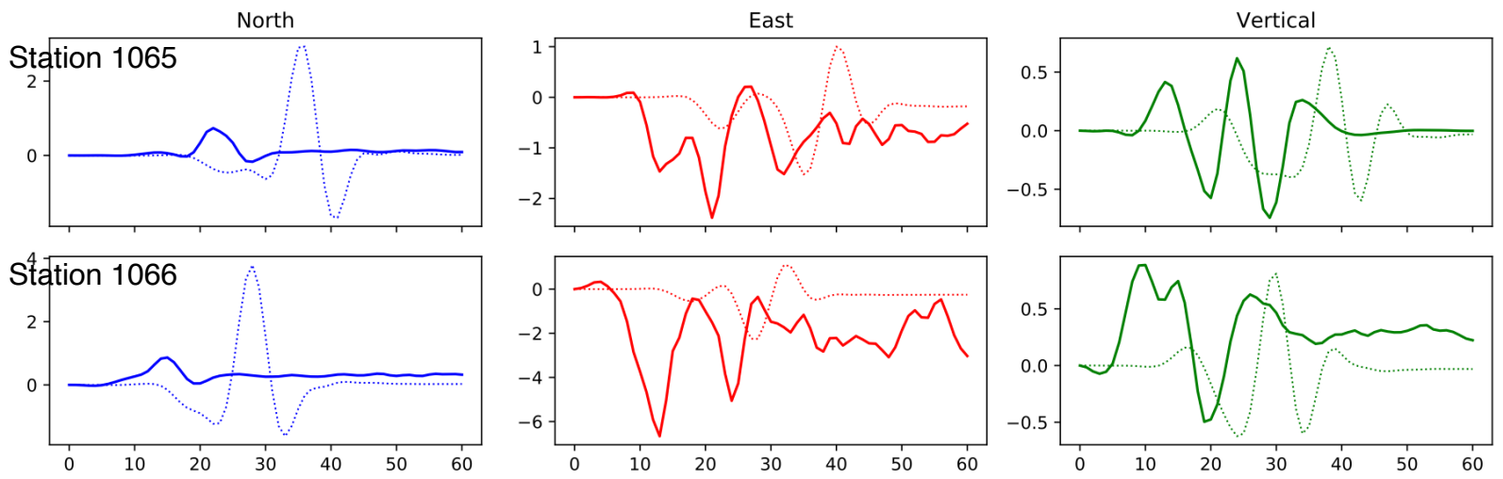
Source location and Centroid Moment Tensor inversion with Wasserstein Waveform misfit

An example of Wasserstein waveform misfit minimisation



GEONET High rate GPS station distribution detecting the $M_w 6.6$ 2005 Fukuoka earthquake, (From O'Toole, Valentine and Woodhouse, 2012.)

Displacement waveforms of true and starting guess



Synthetic noisy displacement seismograms (solid) and starting waveforms (dashed) (stations 1065, 1066)

Starting source is 56km away and CMT is equal to truth.

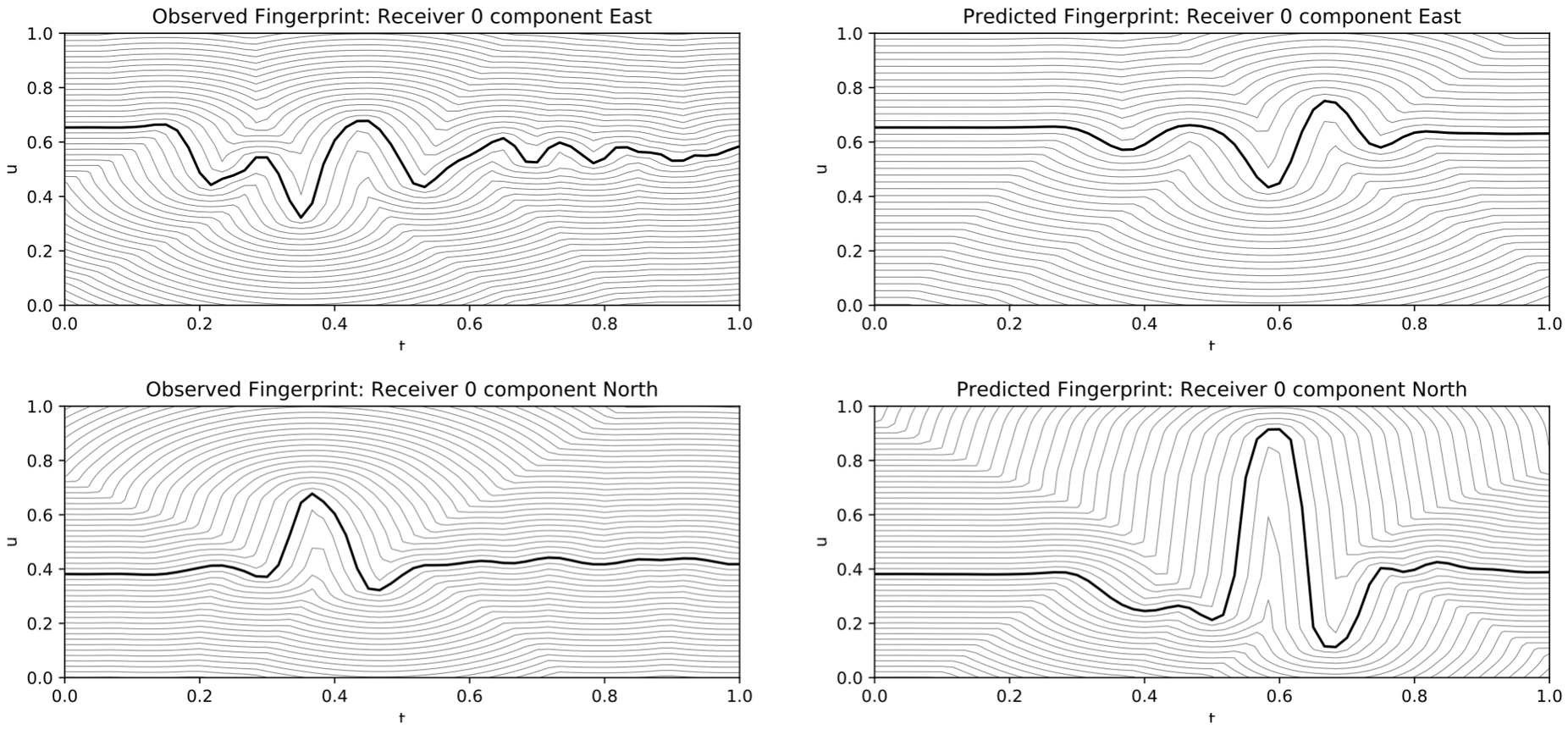
Synthetic example set up:

- 11 receivers each with 3 component (High rate GPS) displacement seismograms (with correlated noise)
- Alternating source location (x,y,z) and CMT solution (CMT solved using linear theory).
- Source location using gradient based minimisation of 1) Least squares waveform misfit and 2) Wasserstein distance, W_2, between observed and predicted 2D seismogram fingerprints.
- 1D Earth model, displacement seismograms and source derivatives calculated using the software package ***pyprop8*** (Valentine, 2021 in prep), based on the approach of O'Toole & Woodhouse (2011), O'Toole, Valentine & Woodhouse (2012).

Amplitude and time windows for Waveform fingerprints

Transform waveform amplitude, u, and time, t, to dimensionless (0,1) box

- Independent time windows with common **linear** scaling: $t' = \frac{t-t_0}{\Delta T_{obs}}$ $\quad t_0 \le t \le t_1$
- Independent amplitudes with common **nonlinear** scaling: $u' = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{(u-u_0)}{\Delta u} + \frac{(u-u_1)}{\Delta u}\right)$ $\quad -\infty < u < \infty$



$\Delta u = u_1 - u_0$

$\Delta T_{obs} = t_1 - t_0$

Observed and predicted Waveform fingerprints of horizontal components for station 1065.

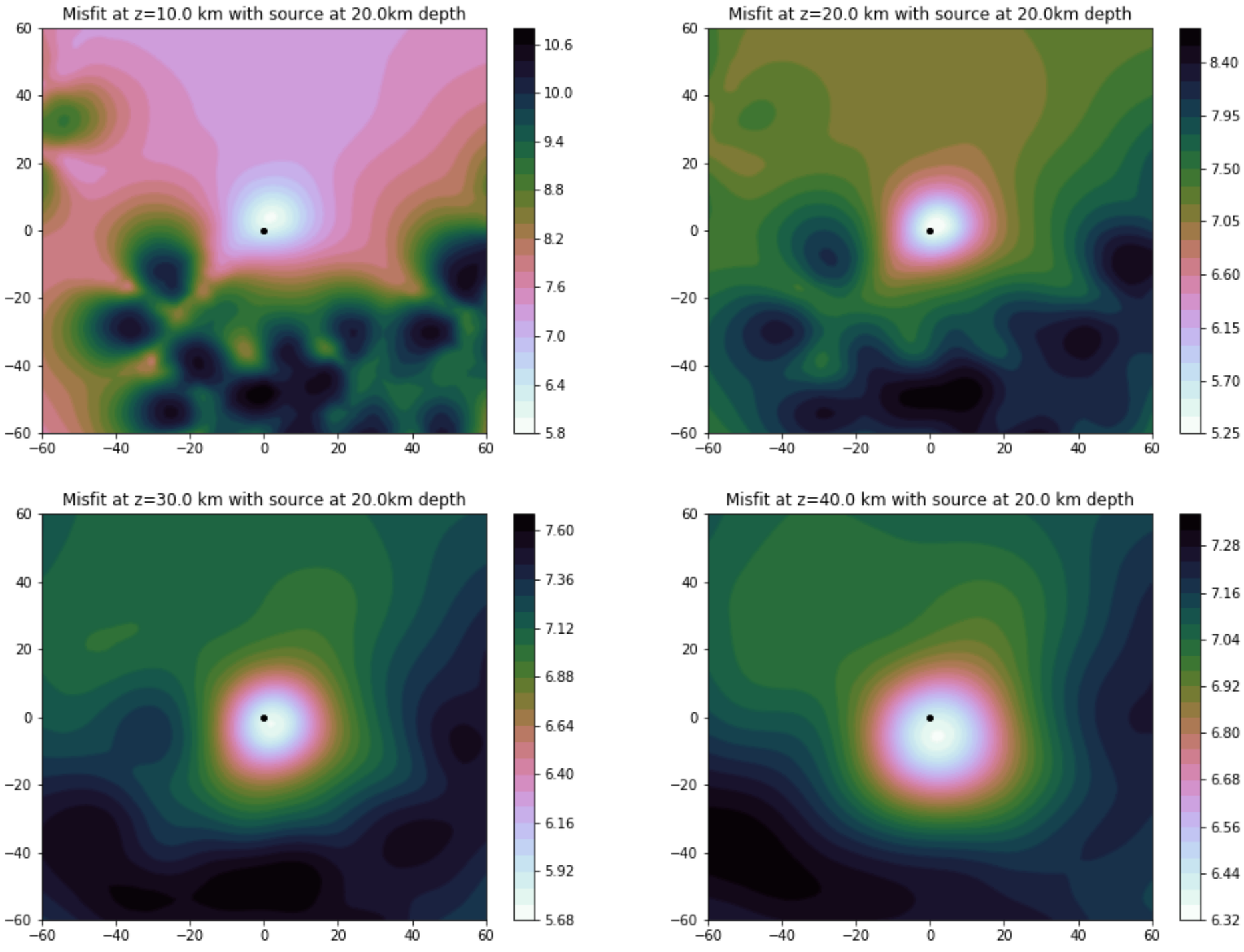
Amplitude range (u_0, u_1) defines main sensitivity interval for amplitude transform.

Time (t_0, t_1) and amplitude range parameters (u_0, u_1) defined by observed waveform.

Cross sections of least squares waveform misfit

- True source location

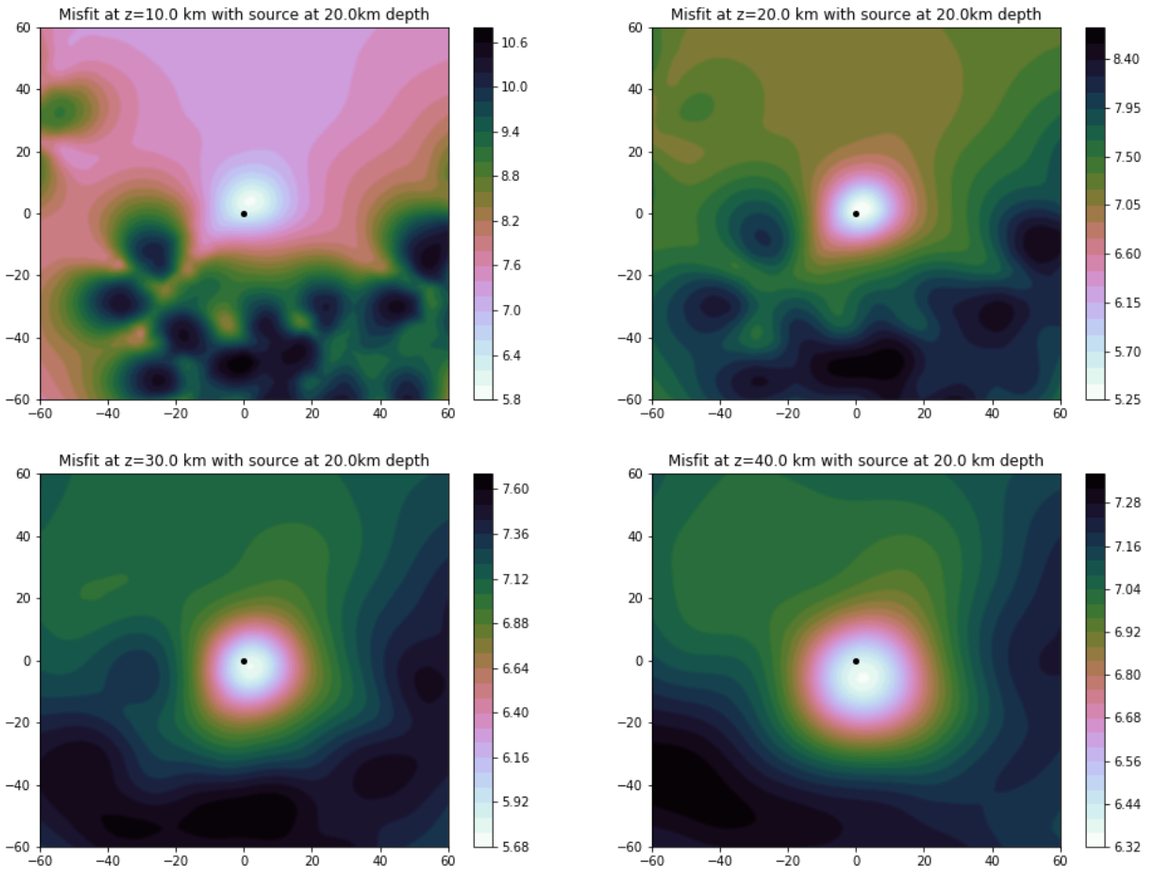
(Not at misfit minimum due to noise)



Cross sections through misfit function in source location.

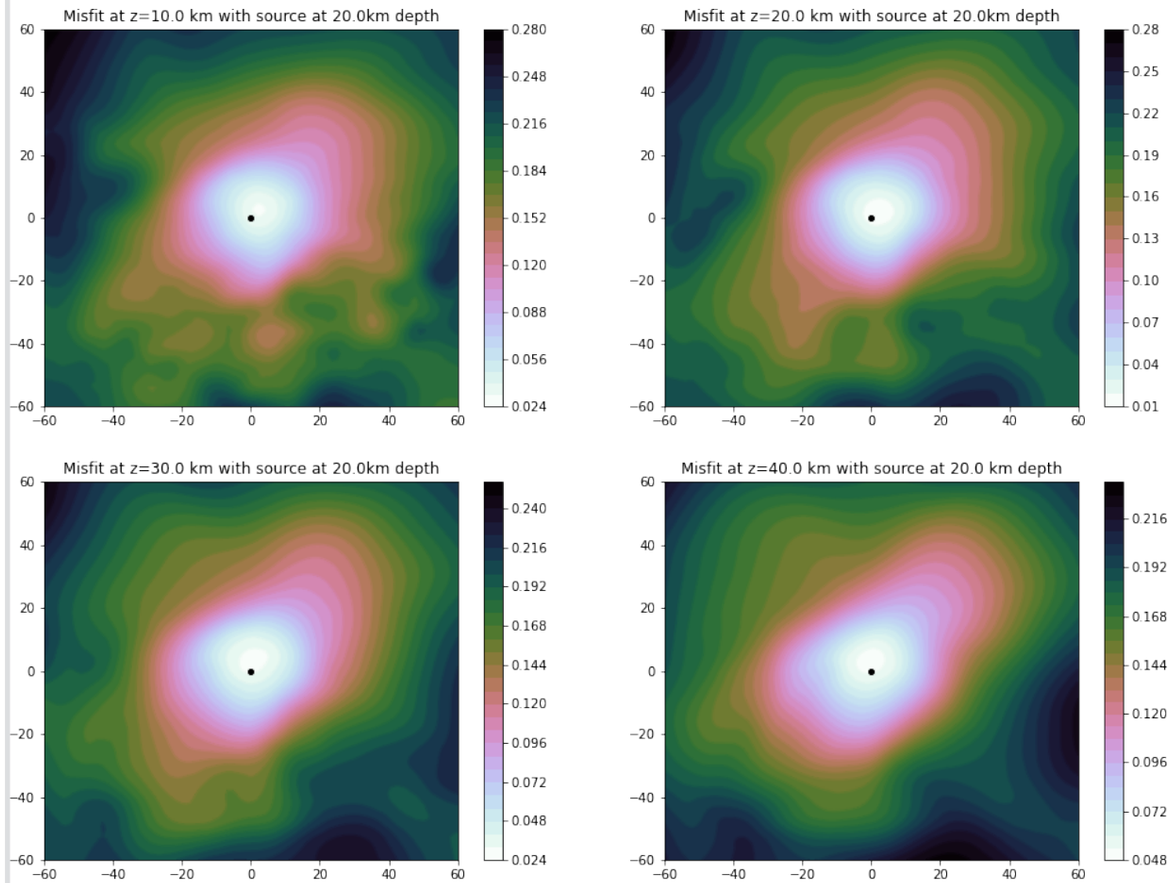
Cross sections of misfit functions with source location

Least squares waveform difference



$$\int_{t_o}^{t_1}(u_{obs}(t) - u(i,\mathbf{m}))^2 dt$$

Wasserstein distance, W_2^2



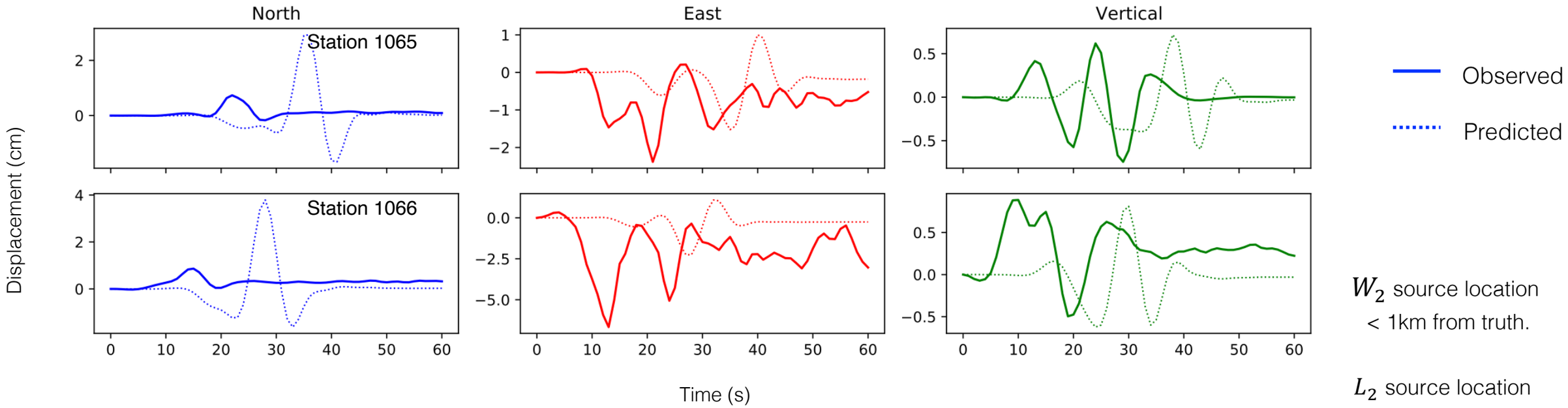
$$W_2^2(P_{obs}, P_{pred}) = \frac{1}{2}(W_t^2 + W_u^2)$$

- True source location
- $P(u,t) = \exp(-d(u,t)/\lambda)$ Fingerprint scale factor, $\lambda = 0.04$.
- Marginal Wasserstein algorithm used.

Gradient based minimisation of Wasserstein distance
Initial and final waveform fits

Example: Initial source 56km from true solution; source (x,y,z) solved for; simultaneous fitting of 33 seismogram pairs.

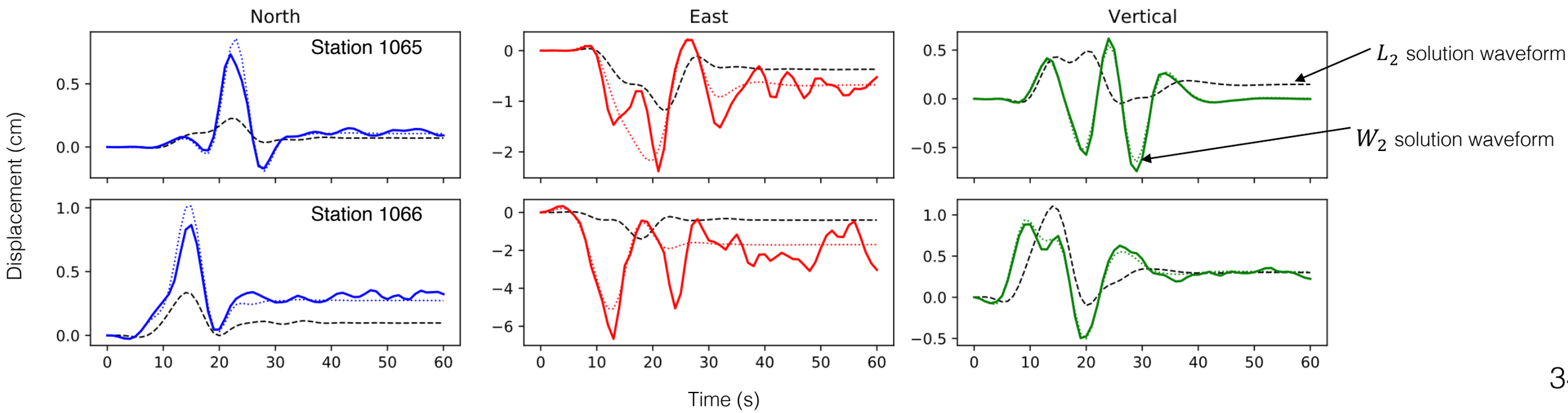
Initial waveform fit



W_2 source location
< 1km from truth.

L_2 source location
~32km from truth.

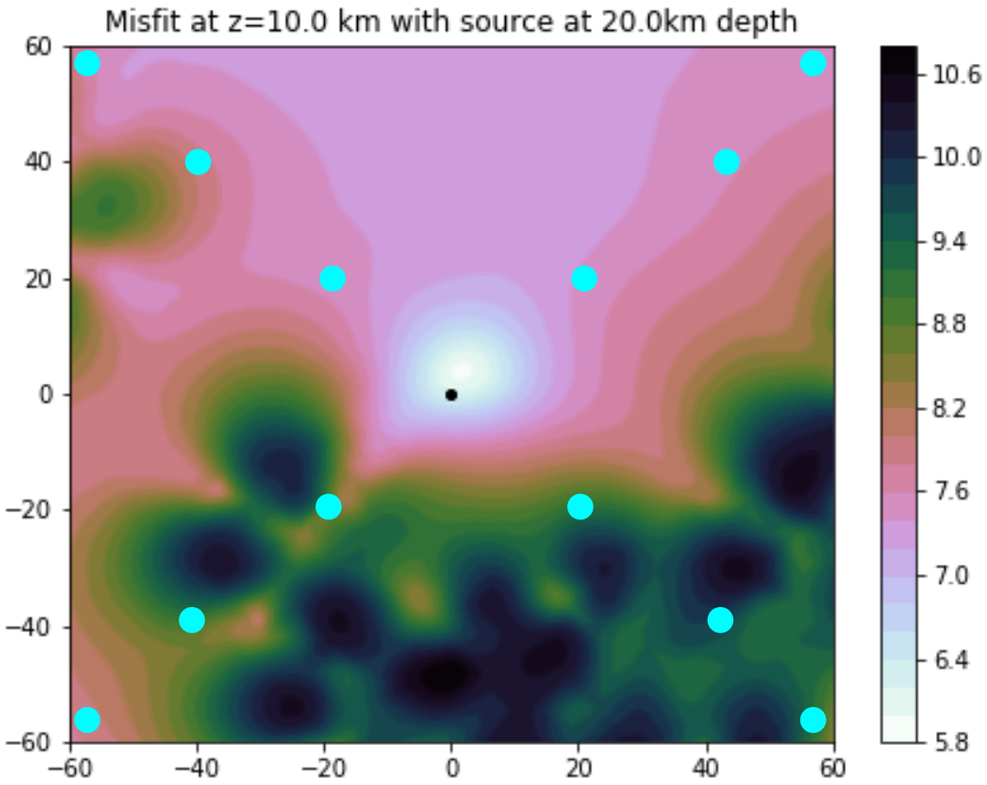
Final waveform fit



Optimisation performance:
Least squares vs Wasserstein distance

Results of 48 Repeat relocations from different starting guesses up to 85 km from true source location.

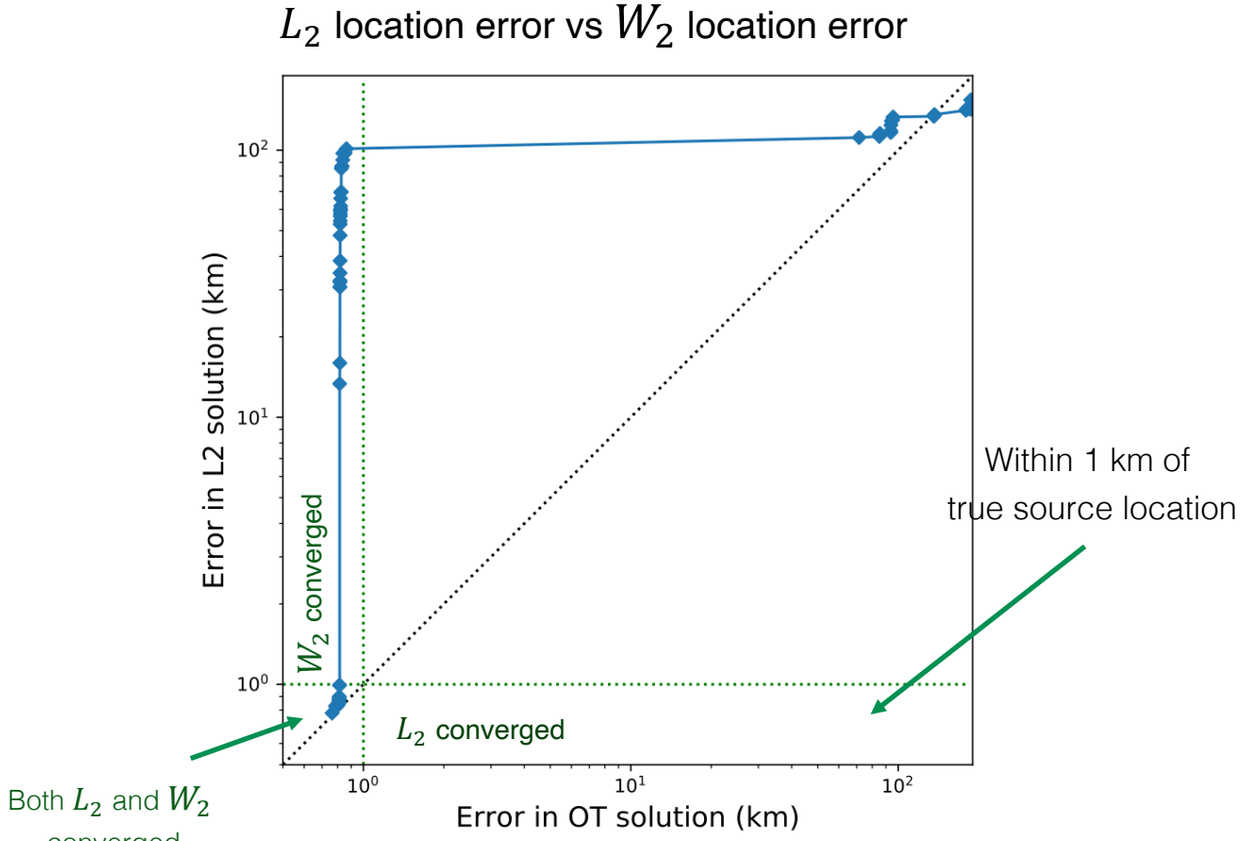
True source is at 20km depth. Trial locations 10-40km.



Optimisation Initial location

Initial depths, 10, 20, 30, 40 km

• True source location

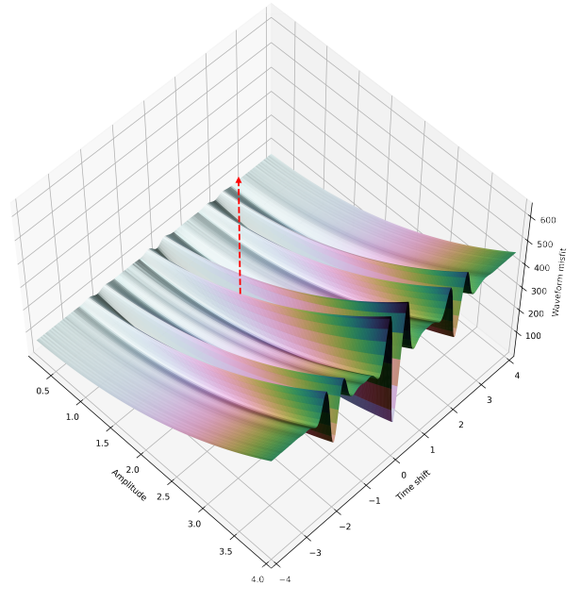


- 29% converged minimizing Least squares norm of waveform misfit
- 79% converged minimizing W_2 distance between waveforms

Conclusions and directions

- A new approach to optimal transport in waveform fitting
  Could be generalised to other time signals, surfaces…

- Exploiting analytical solutions and derivatives to facilitate optimisation framework.
  Quantitative evidence of improvement in convergence of derivative based optimisation

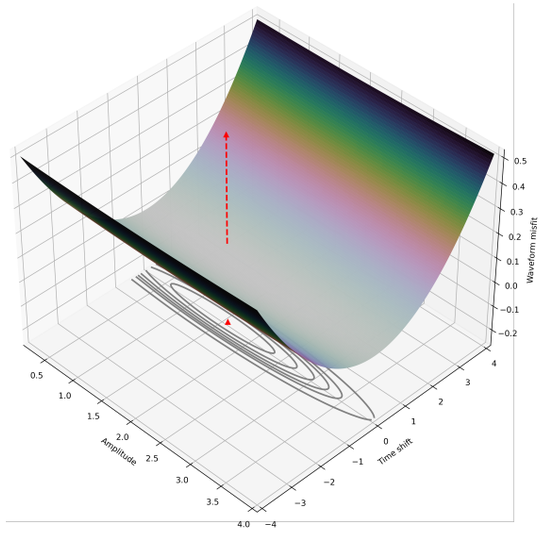


Least squares misfit

- Could be used as starting point for linearised uncertainty analysis or probabilistic sampling.

- Some new directions building on earlier work in seismic FWI.
  but many open issues…use of transport plans/maps

- Less attention on transport plans/maps, but may find new applications, e.g. Bayesian inference, and anywhere that makes use of transfer functions.



Wasserstein distance