Distance correlation for blind source separation: A study of machine learning techniques applied to synthetic and geodetic data

CALLANDER, ELIZABETH

Distance correlation for blind source separation: A study of machine learning techniques applied to synthetic and geodetic data

Elizabeth Callander

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
March 2025

Abstract

According to the WHO, one hundred twenty-five million people were affected by earthquakes between 1998 and 2017 [1]. Increasing our knowledge of the Earthquake cycle is an important task, and using machine learning techniques for the prediction of earthquakes is a promising research direction.

In recent years, the number of Global Navigation Satellite System (GNSS) receiver stations has significantly increased, providing daily data on their locations. Geodetic processes and errors associated with measuring the distance between satellites and receiver stations influence the apparent location of these receiver stations. The work in this thesis uses data that include key geodetic signals and underlying components representing non-geological activity, such as atmospheric components. The separation of these components is the core inspiration for my work, which I address using the blind source separation (BSS) technique to isolate seismic events from atmospheric and instrumental noise in geodetic time and spatial series (GNSS and SAR, respectively) for earthquake monitoring and post-seismic analysis.

In source separation techniques, it is common to assume that the underlying sources are independent. One challenge identified in this context is the difficulty in selecting an appropriate metric to quantify the dependence between sources while effectively optimising toward extrema to produce the most independent sources. To tackle this issue, I compare various independence metrics using the non-parametric test of Binary Phase Shift Keying over an additive white Gaussian noise (AWGN) channel, which serves as a well-established test in Communication Theory. Furthermore, I present an example that compares a binary signal to the average of other binary signals while gradually increasing the number of signals included in this average.

Then, I examine the suitability of these metrics as loss functions, particularly concerning their optimisation and the tailored algorithms required to compute challenging extrema. My research is comprehensive. I apply architectures and metrics to various benchmark datasets for widely adopted source-separation tasks; extend them to GNSS and SAR data to provide geological context and explore representation learning. This multifaceted approach validates my methods on both labelled (for supervised learning)

and unlabelled (for unsupervised learning) data, providing a robust foundation for my findings.

In this work, I introduce distance correlation as a metric for assessing signal independence and evaluate it on several distinct scenarios:

1. **Communication-Theory Benchmark**: Binary Phase Shift Keying (BPSK) signals transmitted over additive white Gaussian noise (AWGN) channels were used to compare distance correlation to a closed-form mutual information statistic.

2. **Synthetic and Hybrid Synthetic/Geodetic Mixtures**: The datasets for this task included combinations of three source signals formed by the linear mixing of sine, square and sawtooth waves; mixtures involving GNSS station pairs combined with synthetic seismic deformation signals or SAR data combined with additive signals. These datasets were used to evaluate various BSS methods, including comparisons with the popular FastICA algorithm.

3. **Geodetic Data**: Real GNSS time series collected around a known seismic event were used to investigate the separation of underlying geophysical sources.

4. **Representation Learning**: Modelling techniques aimed at extracting semantically meaningful features from datasets, including image-based classification across ten categories for CIFAR-10, and disentangled latent features from binary pedestrian mask sequences in the KITTI-Masks dataset.

For the first experiment, using the synthetic dataset, I extracted three waves from the input mixtures, such that the neural network was optimised to extract the most independent underlying sources. On average, the distance correlation method outperformed the established gold standard FastICA, a blind source separation technique based on non-Gaussianity.

I also applied this method to a dataset created by combining two signals from similar GNSS stations, considered to be one source, to a known synthetic signal representing an earthquake with post-seismic deformation at different epicentres. In this case, FastICA slightly outperformed distance correlation in separating the synthetic, seismic signal. When extracting a real seismic event from two actual GNSS stations, FastICA again outperformed distance correlation. It is important to note that the seismic signal in this scenario was compared against the decomposed trend of the GNSS stations and an element of afterslip, not a known ground truth. As such, this comparison should be regarded with caution.

In my final analysis, I applied distance correlation to more advanced representation learning tasks. For the CIFAR-10 dataset, I used a whitening technique for scattering and then brought positive pairs closer together using distance correlation. This approach achieved a Top 1 accuracy of 88.8%. However, it underperformed compared to the original W-MSE method, which achieved a Top 1 accuracy of 91.2%.

The previously mentioned whitening representation methods did not yield good results for the disentanglement task involving the KITTI-Masks dataset. However, when I updated the InfoNCE loss (Laplace, Unbounded) for double-centred inputs, as a proxy of distance correlation, I improved the state-of-the-art mean correlation coefficient (MCC) score by 0.6%.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I want to thank my supervisors, Dr. Ioannis Ivrissimtzis, Dr. Noura Al Moubayed, and Dr. Richard Walters, for their support and guidance during this project. I would also like to thank Geospatial Research Ltd., particularly Richard Jones, and the European Regional Development Fund for their backing.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Between 1980 and 2009, 314,634-412,599 deaths and 845,345-1,145,093 injuries have been reported to have been attributed to earthquakes, with 61 million people in total affected by earthquakes [7]. These numbers are likely underestimates. Due to an increase in population and urbanisation in earthquake-prone areas, the impact of earthquakes is expected to increase in the coming decades. In another example and from an economic perspective, the direct losses of the 1995 Great Hanshin earthquake, including losses to infrastructure and utilities, were estimated to be between $100 and $144 billion [8]. Therefore, learning more about earthquakes and their potential prediction is essential.

Earthquakes suddenly release strain energy within the crust of the Earth, radiating seismic waves from the epicentre. Suppose that a Global Navigation Satellite System (GNSS) station was within a given distance of the epicentre of an earthquake. In this case, the displacement time series recorded by said station may contain a step discontinuity. The Nevada Geodetic Laboratory (NGL) uses a threshold distance of $10^{0.5M-0.79}$ for a possible discontinuity in a GNSS time series associated with an earthquake, with distances in km and M being the magnitude of the earthquake on the Richter scale. Therefore, for a magnitude 4 earthquake, if a sensor is within 16km of the epicentre, a discontinuity in the GNSS time series may appear. However, it is not necessarily true that displacement will

occur at the time of a seismic event even when a sensor is within the threshold distance. The NGL provides caveats that the *depth, style, or directionality of displacement is currently not accounted for in the distance threshold* and that *the displacement may not have occurred at the specified time* [2].

Recently, there has been a dramatic increase in the number of GNSS stations deployed to record their location on the surface of the Earth over time, creating an opportunity for the application of big data methods. Given the applicability of big data and the potential of GNSS data to contain information relevant to the prediction of Earthquakes, GNSS displacement time series will be used as the primary geodetic dataset for the ensuing research.

NGL provides multi-purpose data products from more than 17,000 GNSS stations around the globe [2]. This increase in the volume and quality of the data also means improved spatial and temporal coverage and measurement accuracy. The displacement accuracy, now down to the millimetre, allows for better interpretations of underlying processes occurring at active fault zones, such as tectonic strain loading and slips at different points in the earthquake cycle.

The large volume and high resolution of GNSS data indicate that earthquake prediction could be explored through machine learning. In this thesis, machine learning techniques, predominantly using GNSS displacement time series or a synthetic dataset, will be investigated in terms of the viability of the independent component analysis of GNSS signals and its potential to help predict earthquakes.

Moreover, the selection of GNSS data was guided by the primary objective of this work, which was to study real-world applications, particularly the prediction of earthquakes as a part of a hazard mitigation strategy. This focus is in line with the original GNSS dataset proposed by the project stakeholders. Although more advanced datasets are available for BSS tasks, such as the Libri2Mix dataset, they are not directly applicable to the field of earthquake prediction.

There are three primary forms of geodetic data: SAR, InSAR, and GNSS data. SAR (Synthetic Aperture Radar) detects surface changes but does not provide explicit information on deformation. It has a high spatial resolution but lower temporal resolution compared to GNSS data. Similarly, InSAR (Interferometric Synthetic Aperture Radar)

maps deformation to the millimetre. However, it also has lower temporal resolution than GNSS data.

In contrast, GNSS data provides continuous, daily, and sub-daily motion vectors at discrete stations. This data allows researchers to calculate strain accumulation and plate-velocity models, providing a wealth of information about the earthquake cycle. Although GNSS data has high temporal resolution, it is limited by the density of the stations.

Given the advantages and disadvantages of each data type, I selected GNSS data for its potential to capture critical information in the lead-up to an earthquake that SAR data might miss due to its resolution constraints. For future work, it would be beneficial to combine InSAR and/or SAR data with GNSS data. Then machine learning methods similar to those applied to audio and video datasets could be explored.

The GNSS displacements are pseudoranges, consisting of components representing the distance from a satellite to a receiver and other elements associated with errors when calculating the actual range. The number of underlying processes, measured as underlying components of the overall GNSS displacement time series, varies from station to station. This work focuses on signal components with geological meaning, which often comprise a smaller proportion of the overall GNSS displacement time series. An example would be slow slip events, which are slip events that are undetectable by seismometers. Slow slip events have a longer duration compared to earthquakes of comparable seismic moment, causing a more gradual displacement within the GNSS time series over time. GNSS time series have aided in the discovery of several slow slip events in subduction zones, exemplifying the potential power and challenges of using GNSS time series in geological research.

1.1 Problem

This thesis investigates the applicability of machine learning analysis on GNSS data, focusing on seismic events. GNSS time series represent the position of a receiver station on the surface of the Earth, tracking how its three-dimensional coordinates evolve over time relative to a defined reference frame. As an example, Figure 1.1 presents the daily displacement over time for the J076 station in the east, north, and up directions, referenced

to the IGS14 frame. GNSS receivers measure their position to the nearest μm in a specific reference frame. Unless otherwise specified, the global reference frame *IGS14* [9], a 3-dimensional Cartesian geocentric coordinate system for International GNSS Service data products, is used.

The main focus of this research is to identify and separate components of the GNSS pseudoranges related to seismic processes, separating not only well-studied processes but also identifying new processes, ultimately aiming at improving our knowledge about active faults and the earthquake cycle. A successful example [10], well-documented in the literature, is the study of slow slip events, the understanding of which has increased tremendously with the use of GNSS data. It is worth noting nevertheless, that slow slip events still need to be successfully modelled in sufficient detail to remove their effect from the GNSS time series as a pre-processing step.

Defining the problem as that of blind source separation emerged from the modelling limitations identified at the early stages of the project. In particular, the work of Gualandi and Michel acted as the motivation for the research outlined within this thesis. [11] proposed variational Bayesian Independent Component Analysis, vbICA, to extract time series representative of different processes from GNSS position time series. Following on from this work, [5] used the proposed vbICA method on the Cascadia fault.

1.1.1 Problem definition

The definition of blind source separation varies, depending on the task and the available dataset. In the first instance, it will be defined in the context of the Cocktail Party Problem.

Imagine two people speaking concurrently, thought of as two individual sources, $s_1(t)$ and $s_2(t)$, and two microphones recording the mixed sounds, $x_1(t)$ and $x_2(t)$. The mixing system can be written as:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix}, \tag{1.1}$$

or in matrix-vector notation as:

$$X = AS. \tag{1.2}$$

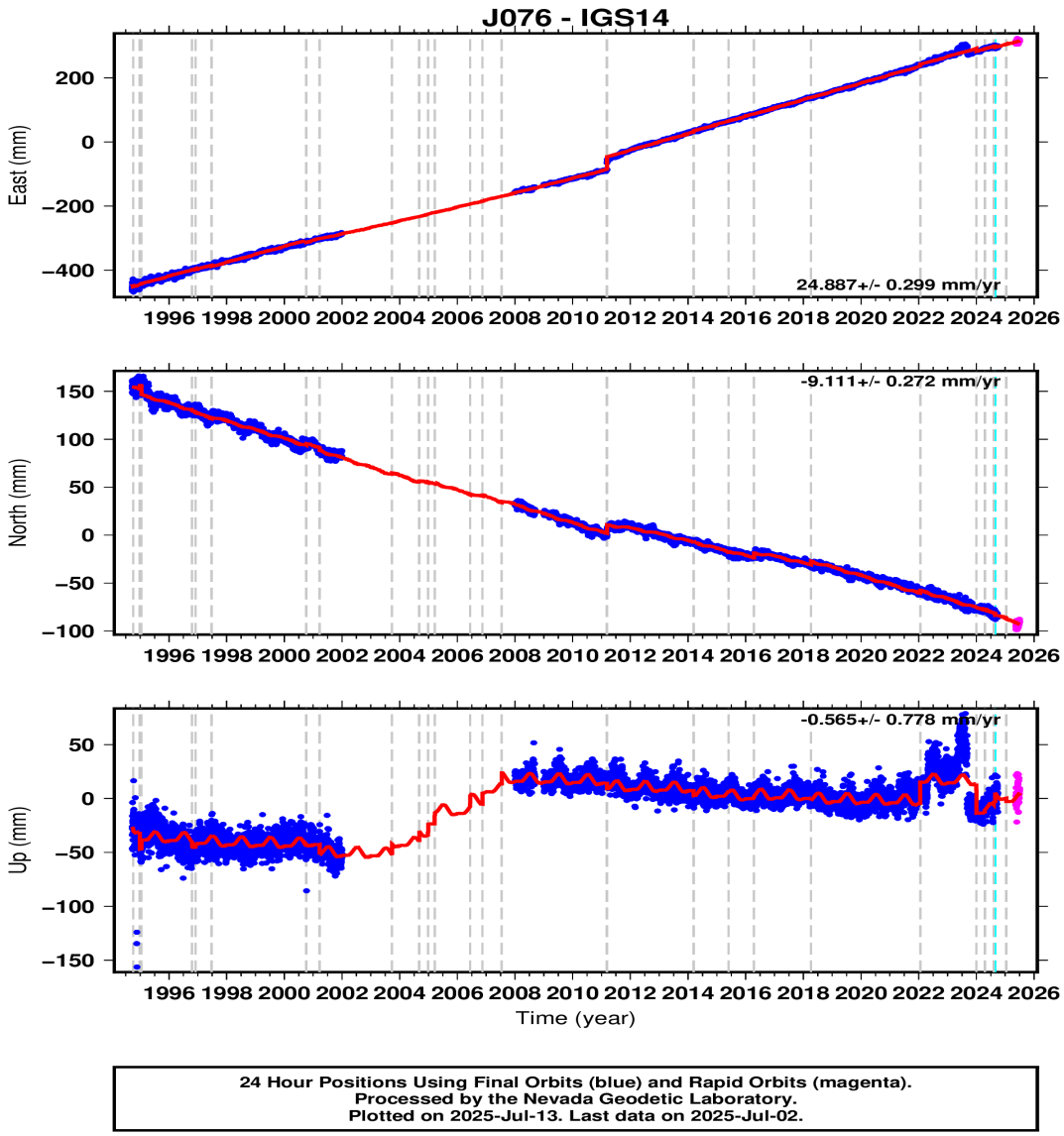


Figure 1.1: The daily displacement over time for the J076 station in the east, north, and up directions, referenced to the IGS14 frame. Blue points represent the daily data, while the red curve illustrates a model that estimates the best fit for this data. Grey dotted lines indicate the timing of nearby earthquakes, while the cyan line marks a possible step change, potentially resulting from equipment modification or software updates. The magenta points represent data with a 5-minute sampling rate collected over 24 hours.

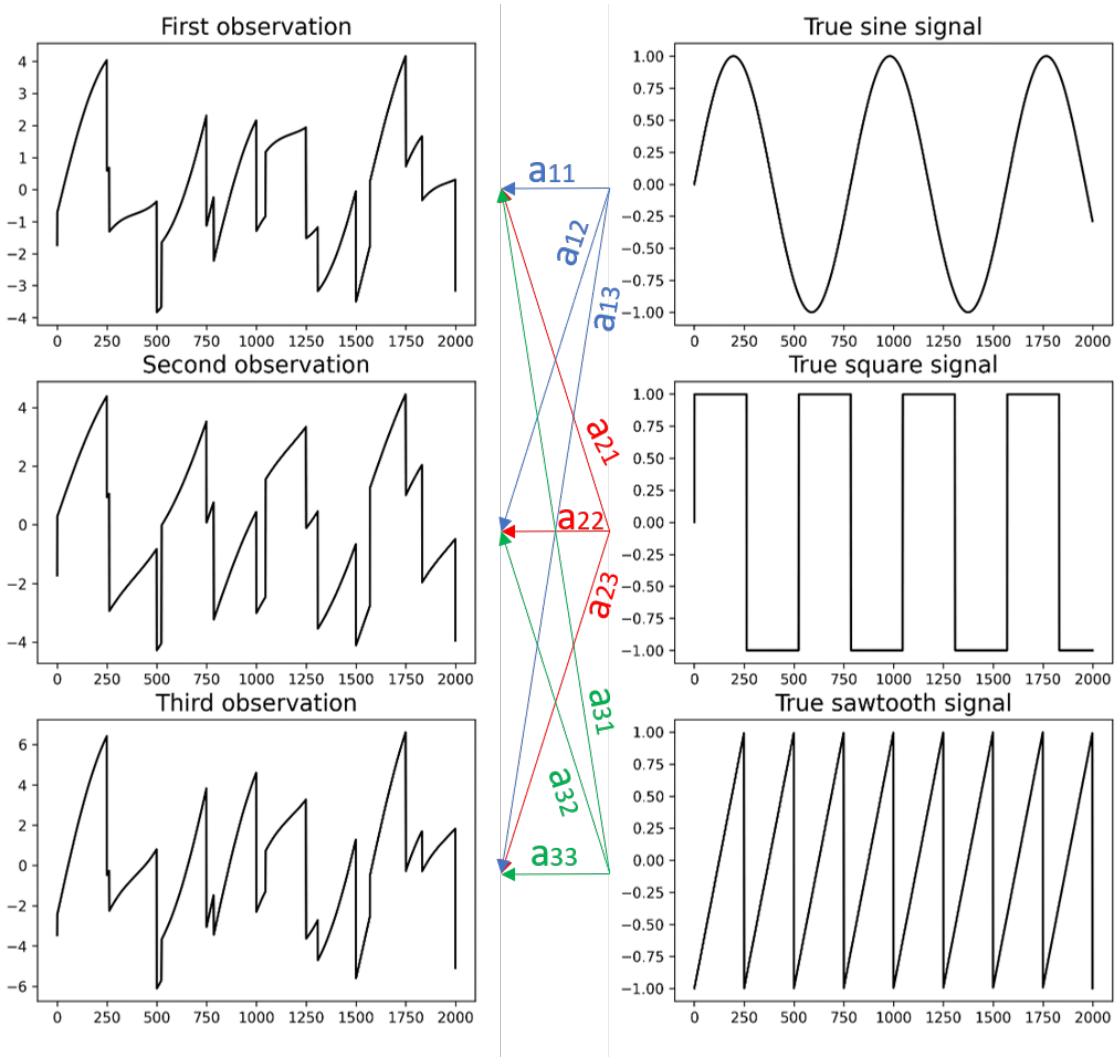


Figure 1.2: A graphical representation of blind source separation viewed as an inverse problem. On the left, we have the known observations, while on the right are the underlying sources. The mixing matrix operates on these underlying sources to produce the mixtures, which is why the arrows point from right to left.

Equation 1.2 is a general case where X is a column vector filled with the N recorded mixtures; A is the NxN mixing matrix, and S the column vector filled with the N sources or the Independent Components, in ICA terms. Figure 1.2 is an example with three mixed signals and three sources. Note that the arrows point from right to left and that the only known values are the three mixtures at the left-hand side.

The ICA formulation of this problem, with the time index omitted, is summarised by Equation 1.3.

$$x_j = a_{j1}s_1 + a_{j2}s_2 + ...a_{jN}s_N \qquad j = 1,2,\ldots,N \tag{1.3}$$

The time index was dropped because the ICA model assumes that each mixture, x_j, and source, s_k, are random variables rather than time series. Thus, the observed values x_j(t) are samples of the random variable x_j.

1.1.2 Motivation

In blind source separation, various source signals will be separated, with the aim of distinguishing processes not under investigation, 'noise', from valuable seismic information. What information is important is application-dependent and subjective. In this work, geodetic signals are fundamental. However, ionospheric noise, which varies periodically throughout the day, and seasonal tropospheric noise can have a greater amplitude than geodetic signals, often dominating them in the GNSS time series. Moreover, as stated in [11], 'we are mostly interested in understanding what is not already known, i.e. those signals for which we do not have any well-established pre-determined model'.

Problem definitions, other than blind source separation, may not provide the scope required to output information that will provide greater insights into seismic processes that are currently not well-modelled. Indeed, the previously mentioned slow slip events are an example of events that are challenging to model with more conventional methods. Slow slip events are examples of transient deformations, a non-periodic, non-secular accumulation of strain in the crust [11], and attempts to develop basic models for them, consisting of linear, cyclic, and offset components, have not been successful.

Understanding and modelling transient deformations can be valuable in assessing seismic hazards. However, as the research presented in this thesis progressed, it began to focus more on detailed investigations of machine learning techniques, independent of their specific applications.

1.2 GNSS displacement time series

GNSS is an umbrella term for global positioning satellite constellations. It encompasses systems such as the Global Positioning System (GPS), Galileo, GLONASS, and BeiDou. These systems are supplemented by ground-based and space-based augmentation systems [12].

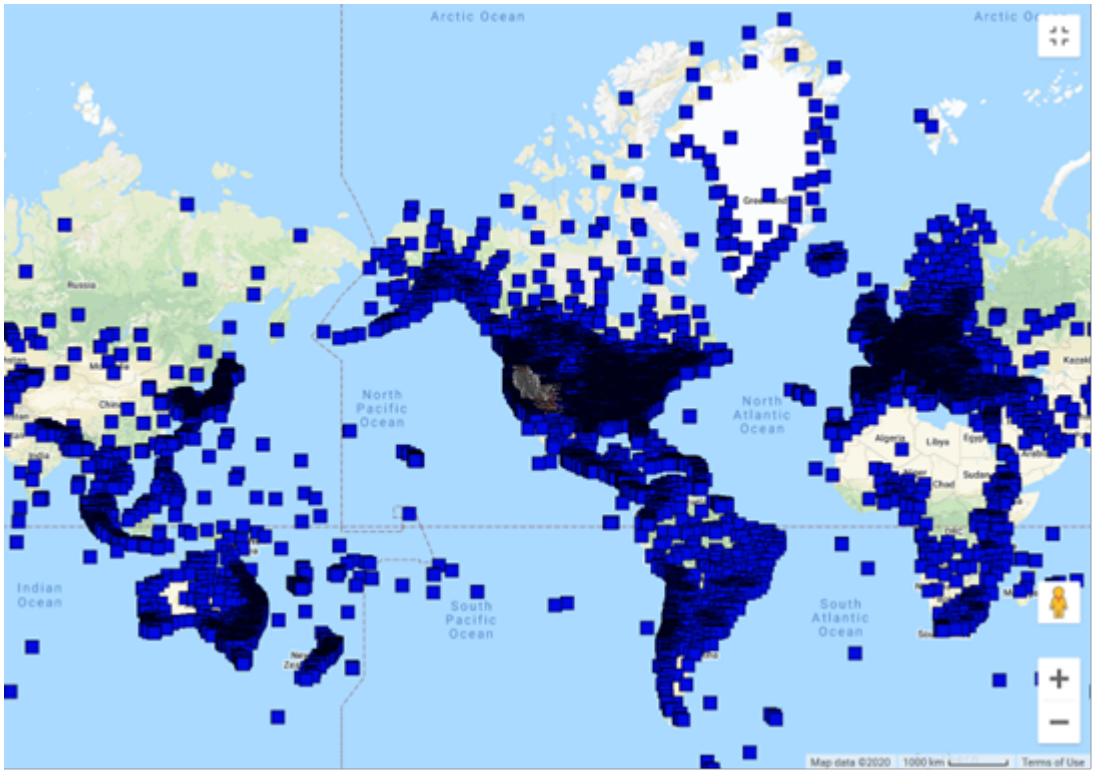


Figure 1.3: The global distribution of the GNSS receiver stations [2].

Using signals from the GNSS satellites, the NGL calculates and records the position coordinates of some 17,000 GNSS ground stations located around the globe [2]. Of those 17,000 stations, 10,000 stations have their positions taken once daily and the overall dataset is updated every week; 5,000 stations have their 5-minute position coordinates updated every day, and the final 2,000 have their 5-minute position coordinates updated every hour. Figure 1.3 depicts the global coverage of the GNSS receiver stations.

The positions of these receiver stations are determined by a process known as GNSS ranging. Given the time at which a signal is sent from a satellite, the position of the satellite when the signal is sent, and the time at which the GNSS station receives the signal, the distance from the receiver station to the satellite can be calculated. In ideal conditions, Equation 1.4 determines the range, ρ, of the signal, that is, the distance it travelled:

$$\rho = c \cdot \Delta t, \tag{1.4}$$

where Δt represents the signal propagation time and c is the speed of light, as the signal is a form of electromagnetic radiation.

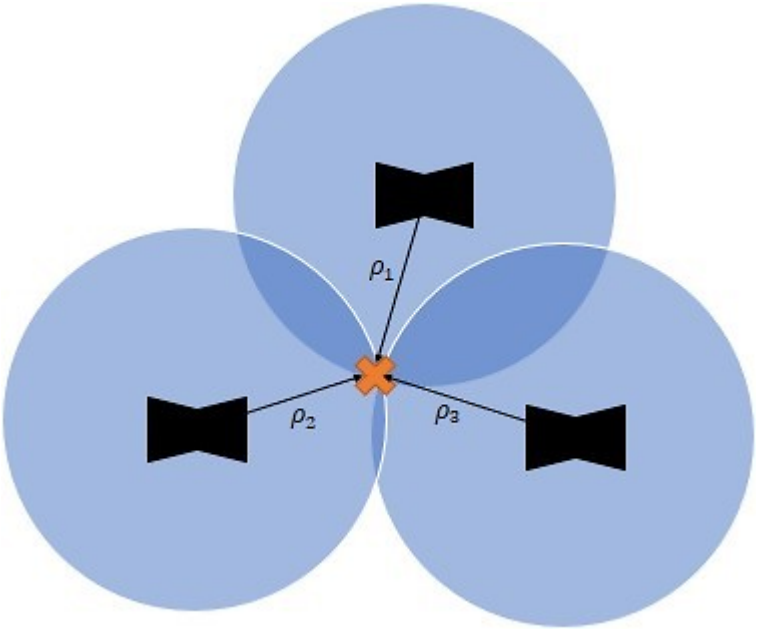


Figure 1.4: The 2D case, where three satellites, the black symbol, are located in the centres of circles with radii equal to their calculated range. The orange cross represents the intersection point and the location of the receiver station.

The range is the radial distance from the satellite to the receiver station, and thus to determine the location of the receiver station, the radial distances to 4 satellites are required, 3 for the spatial coordinates, and one for time. The first satellite locates the receiver somewhere on the surface of a sphere; two satellites locate the receiver somewhere on a circle created by the intersection of the two spheres; a third sphere provides two possible locations on the circle where the receiver may be. As stated before, 4 satellites are required to locate the receiver. The fourth satellite can be used to find the one point where all the spheres intersect (the location of the receiver). However, one of the two candidate points can usually be excluded because its location is nonsensical, that is, obviously far away from the Earth's surface.

Instead, the fourth satellite is more beneficial in addressing the receiver clock offset, as receivers usually contain quartz crystal clocks, which can drift around 0.1 nanoseconds to 1 second. For example, a drift of 0.1 nanoseconds compared to the satellite clock causes the range calculation to be out by around 0.03m, using the lower end of the drift ranges. To account for clock uncertainty, if the four spheres do not intersect, alternative ranges are taken from a series of times around the measurement, until there is an intersection point for all four of the range spheres. Figure 1.4 illustrates the simpler 2-dimensional case.

The scenario mentioned previously is an idealised case, and many processes affect the range calculation. Thus, the equation of the *pseudorange*, essentially an error decomposition, is used instead [13]:

$$p = \rho + d_\rho + c(\mathrm{d}t - \mathrm{d}T) + d_{ion} + d_{trop} + \varepsilon_{mp} + \varepsilon_p, \tag{1.5}$$

where p is the pseudorange; ρ is the true range; d_ρ is the satellite orbital errors; c is the speed of light, $\mathrm{d}t$ is the satellite clock offset; $\mathrm{d}T$ is the receiver clock offset; d_{ion} is the ionospheric delay; d_{trop} is the tropospheric delay, ε_{mp} is the multipath and ε_p is the receiver noise. In the case where the satellite and receiver clocks are offset by the same amount, the apparent signal propagation time would equal the actual propagation time as the offsets would cancel.

As an aside, the noise component of the GNSS displacement pseudorange is often assumed to be white noise due to model constraints, as stated in the literature [11] [14] [15]. However, this assumption is not strictly accurate. Fourier analysis has revealed that the noise is more likely to be Brownian or pink. The distinction between the two types of noise may arise from low-frequency linear trends in the data, which may be influenced by the detrending steps applied to the GNSS data. Further analysis is needed to explore this issue. In my work on BSS with GNSS data, I did not explicitly model the noise, but will nonetheless maintain the assumption that the noise is white.

Returning to the discussion of GNSS measurements, the position coordinates collected by the NGL are pseudoranges, not pure ranges, and are subject to several uncertainties. For instance, regarding ionospheric errors, the signal moves more slowly through the ionosphere when there is a higher density of electrons. This happens because free electrons are released when the UV radiation from the sun ionises them. Consequently, the impact of the ionosphere on the position-time series of the receiver station exhibits a regular daily cycle.

Unlike the refraction caused by the ionosphere, the electrically neutral troposphere does not cause refraction based on the frequency of the signal frequency. Tropospheric refraction arises because the refractive index of the troposphere decreases as altitude increases such that the signal bends towards the Earth [16]. The GNSS signal with the

shortest path through the troposphere will be the least delayed by it (i.e. the tropospheric effect will be minimised if a satellite is directly above the receiver). The refraction has two components. The more dominant is the dry component, which is closely correlated to the atmospheric pressure, and the less dominant is the wet component, which depends on the highly variable water vapour arrangement in the atmosphere.

The effects of both the ionosphere and the troposphere can be reduced by preprocessing the GNSS time series. In the case of the ionosphere, it is known that the delay depends on the frequency of the signal. Dual-frequency receivers can assess the frequency-dependent signal delays caused by the ionosphere by utilising two different signal frequencies, allowing for the correction of this error. Tropospheric modelling can be used in GNSS preprocessing to reduce the tropospheric influence on the GNSS data by up to 95%. Nevertheless, any residual of these preprocessing steps may still be on the order of magnitude of the geodetic signals of interest.

Finally, the 'noise' components present in the signal can differ depending on the receiver station's location. For example, a receiver station near the sea may have multipath due to the signal's reflection off the water, but a receiver station in a field may not.

Given the complexities of modelling GNSS signals, a useful and challenging research question would be '*How can all signals that are combined to form the GNSS position time series of a receiver station be separated from one other*?' or if this is not possible, '*How can signals representative of seismic processes be extracted from other processes within the mixed GNSS position time series*?'

A further complication in the latter, simpler form of the research question is that the user may not extract from the mixture processes which have not been modelled that are nevertheless part of the earthquake cycle if they somehow bias the process to extract signals that are already known. Consequently, the problem will be approached as a blind source separation problem.

Blind source separation of GNSS data

The blind source separation will now be redefined for the GNSS case. Take a system of N GNSS stations whose coordinates are updated daily. These coordinate readings would be analogous to a mixed microphone recording in the cocktail party formulation, with three

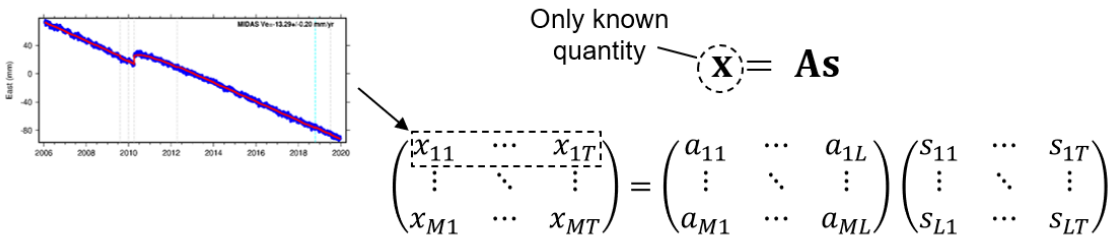


Figure 1.5: Summary of how the recorded GNSS time series are related to the mixing matrix A, and the unknown sources S. The GNSS time series is from NGL [2].

channels for the east, north, and up directions. The total number of elements in the time series would be $M = 3N$.

For $L < M$ sources, the pseudoranges can be written as a linear combination of the underlying sources. The problem can thus be written as:

$$X_{M\times T} = A_{M\times L} S_{L\times T} + N_{M\times T}, \tag{1.6}$$

where A is the mixing matrix; S is the source matrix; and N is a Gaussian noise matrix. Each row of S contains the time series of a given source, where each source is statistically independent of the other sources, corresponding to the M observed random variables, that is, the recorded coordinates, using a linear combination of L variables whose probability density functions describe the temporal evolution associated with each row of S [11]. Figure 1.5 summarises the GNSS blind source separation problem.

When applying BSS methods to GNSS time series, it is crucial to remember that both classical ICA algorithms (e.g. FastICA) and my newer distance-correlation-based BSS techniques assume the underlying sources are independent, identically distributed and stationary. In contrast, GNSS residuals typically violate stationarity as they contain a deterministic trend from tectonic plate motion; exhibit seasonal and periodic variations and can show amplitude changes over time from environmental factors. In my pipeline, I MIDAS-detrend the data, but residual non-stationarity will remain. However, I make the assumption that this non-stationary residual effect is negligible for the time window studied.

In terms of non-linearities, GNSS observables are typically modelled as a linear super-

position of error sources. After applying the standard corrections, remaining non-linear effects, such as multipath interference, are usually small relative to the dominant linear errors. Consequently, these non-linearities behave like high-frequency noise. Linear BSS methods can therefore separate the main components effectively.

The aim of blind source separation is to extract the underlying sources with limited supplemental information. In fact, in the definition of the above problem, the only known variables are the recorded mixed GNSS signals, X. Independent component analysis is a prevalent blind source separation technique, based on the assumption that the source signals that mix to form a recorded signals are statistically independent. Commonly, non-Gaussianity is used as a measure of independence due to the Central Limit Theorem. Heuristically, it can be said that under most circumstances, the sum of the probability distributions of two time series is more Gaussian than the individual distributions. That is, as more signals are mixed, the mixture becomes more Gaussian. Metrics used to describe Gaussianity in the context of ICA include the fourth-order moment, kurtosis and negentropy.

However, some fundamental limitations of ICA have not been resolved, and various alternatives to non-Gaussianity as a measure of signal independence have been proposed. A well-established alternative measure of signal independence is mutual information. In addition to the baseline ICA methods, mutual information will serve as a second comparator for the methods developed in this thesis.

A limitation of classical ICA is the assumption that each source has a unimodal, non-Gaussian distribution so that mixing drives the result closer to Gaussian. However, transient GNSS signals often exhibit multimodal distributions, such as bimodal peaks. Therefore, their mixtures remain distinctly non-Gaussian. In such cases, the core assumption of ICA breaks down and it cannot now reliably recover the original components [5]. This poses a real challenge for GNSS time series, where geophysical processes and underlying signals, such as the seasonal tropospheric component, may not be unimodal. Therefore, the use of another form of ICA, variational Bayesian ICA, may be more relevant for blind source separation of GNSS position-time series [11].

1.3 Aims and objectives

The broader aim of my work is to examine the suitability and reliability of using blind source separation and machine learning methods, both novel and well-established, on GNSS position-time series and well-established datasets. The significance of this research is considerable, as it may provide insight into the earthquake cycle and improve predictions for earthquakes.

My research objectives were:

- To compare distance correlation, non-Gaussianity and mutual information as measures of independence, especially regarding their suitability as loss functions for blind source separation.
- Assess the performance of these methods regarding the separation of a synthetic earthquake signal embedded in a known GNSS signal, and then on blind source separation of an actual seismic event (without a known ground truth).
- To determine the applicability of a distance correlation-based loss on other machine learning tasks: fine-tuning the source separation of the Libri2Mix [17] dataset; whitening for self-supervised representation learning; and the disentanglement task using the KITTI-Masks [18] [19] dataset.
- To investigate the relationship between natural signal whitening procedures and independence as determined by distance correlation.

1.4 Method

One major methodological challenge I encountered during this project was determining whether a function could effectively serve as an independence loss function. This involved both accurately describing independence and efficiently computing the extrema of the function to find its global minimum. Indeed, when tackling a blind source separation problem, it is not sufficient for the extrema of the independence metric to correspond to independent signals. One has also to be able to compute these extrema efficiently.

Addressing this challenge, in Chapter 3, I first apply non-parametric tests to compare the behaviours of independence metrics on some fundamental models of signals, which have been extensively studied in Communication Theory, before proceeding in later chapters to comparing the suitability of these independence metrics as loss functions. Perhaps most importantly, in Chapters 5-6, considerable effort is spent on improving the optimisation algorithms by adjusting them towards the specific characteristics of each independence metrics. Thus, the extrema of each independence metric are most often computed by multiple different optimisation methods. Some of these tailor-made optimisation algorithms can be seen as secondary research contributions to my thesis.

A second major methodological challenge encountered during the project was that in the context of applied problems such as blind source separation of GNSS data, signal independence becomes ill-defined. Traditional mathematical concepts of independence, such as the independence of probability distributions, might not be entirely applicable. As discussed in Section 1.2, I may not have a ground truth to measure against outputted sources. Moreover, the number and waveform of the sources themselves could be unknown.

Whitening is frequently applied to enhance BSS, but it is often used without careful consideration of the specific type being employed. Therefore, Chapter 4 provides a detailed discussion on whitening to shed light on this topic, with a particular focus on the difference between independence and decorrelation. Moreover, Chapters 5-7 use various types of data, including a standard synthetic data benchmark for testing blind source separation algorithms, as well as GNSS and SAR geological data. Additionally, image and audio data from multiple sources are included. The goal was to validate my approach by examining the relevance of my findings across as many different data types as possible.

The specific methods used within this thesis are broken down by research chapter below:

Ch.3 In my first research chapter, I compare independence metrics on the non-parametric test of Binary Phase Shift Keying (BPSK) over an Additive white Gaussian noise (AWGN) channel, exploiting the fact that the simplicity of the setting affords the efficient computation of the precise values of all metrics. The exercise is then repeated for different colours of noise, to identify whether white noise gives negentropy in

particular an advantage. Next, a similar problem involves comparing a binary signal to the average of several binary signals. This method exploits the central limit theorem to increase the Gaussian characteristics of the second signal, allowing for the investigation of the effects of heightened Gaussianity.

Ch.4 In my second research chapter, whitening is discussed as part of the blind source separation pipeline. I compare several whitening methods, and emphasise that decorrelation does not necessarily mean independence. Note that the choice of whitening method is an issue rarely discussed in applied blind source separation papers.

Ch.5 In my third research chapter, I compare the use of distance correlation, mutual information, and negentropy as loss functions. I test the metrics first on a standard synthetic benchmark for the blind source separation problem, and then on a mixture of synthetic and geodetic signals. The geodetic signals are GNSS displacement time series and SAR images. The synthetic signal and the mixing matrix were controlled and therefore known.

I implement a simple neural network for distance correlation and create a variant incorporating simulated annealing elements. The mutual information calculations are based on the Mutual Information Neural Estimation (MINE) algorithm [20]. The negentropy optimisation are based on the FastICA algorithm or its neural variant, PyFastICA. Two architectures are utilised for each loss function: one learns the outputted sources explicitly, and the other does so implicitly.

Ch.6 In my fourth research chapter, blind source separation using distance correlation is applied to a 2-mix-2-source GNSS problem, without any use of synthetic data. Moreover, I conduct a case study to extract 10 sources from 120 GNSS time series in Southern California.

Ch.7 In my final research chapter, I use distance correlation to extend the W-MSE method for the representation learning proposed in [4]. The model is tested on representation tasks using the CIFAR-10 and KITTI-Masks datasets. Its performance is compared to the InfoNCE contrastive method for KITTI-Masks and the whitening MSE

methods for CIFAR-10. Additionally, I explore fine-tuning a SepFormer model for the Libri2Mix task.

1.5 Contributions and Limitations

The main contribution of this thesis is the introduction of distance correlation as a part of the loss function to train machine learning architectures. While partial distance correlation has been used in order to compare the functional behaviour of neural network models in [21], it is believed that this work is the first to use distance correlation as a loss function.

The loss function itself is promising, having the ability to separate signals. However, its ability to train is not generally robust, leading to variable results. This is overcome by training networks multiple times and choosing the best results.

Another contribution emerged from comparing distance correlation to traditional correlation. When examining the highest variances described by the eigenvectors in PCA, an equivalent method for ICA emerged. Consequently, I developed a method to determine the optimal number of sources to extract through eigenvectors and distance variance, which effectively represent the directions of maximal dependence.

Moreover, I investigated the use of distance correlation for representation learning. I examined whitening MSE and InfoNCE losses for their respective representation learning tasks on CIFAR-10 and KITTI-Masks datasets. In the case of CIFAR-10, the W-MSE method with a distance correlation loss performed about 2.4% worse than its standard MSE counterparts. I also applied whitening MSE and its distance correlation extension to the KITTI-Masks dataset. However, these approaches did not surpass the performance of InfoNCE. This underperformance might be due to the sphering of the data not aligning with the natural geometry of the data. Ultimately, the InfoNCE loss, combined with a double-centred distance extension that I introduced, outperformed all other loss functions by at least 0.6%.

1.6 Organisation of the thesis

The following paragraph outlines the structure and content of this thesis. Chapter 2 reviews the literature relevant to this work. In Chapter 3, non-parametric tests are employed to compare different independence metrics. Chapter 4 examines various types of whitening techniques and explores their appropriate applications within the source separation pipeline. The core research is presented in Chapters 5 and 6, which investigate different blind source separation methods using both synthetic data and GNSS data, focusing on independence metrics as loss functions and on optimisation processes. In Chapter 7, I test distance correlation-based machine learning with deeper neural networks across other application domains, specifically using the KITTI-Masks and LibriMix datasets. Finally, Chapter 8 presents my conclusions and recommendations for future work. See Table 1.1 for an overview of the research questions and methods per chapter.

Returning to the detection of seismic activity, I harnessed distance correlation as an independence metric within a blind source separation framework to isolate tectonic ground deformation signals from non-tectonic noise in GNSS time series. I validated my approach on both synthetic benchmarks and real seismic GNSS data, aiming to recover a single seismic component from multiple underlying sources to input into earthquake-forecasting algorithms. Although this method did not yield a 'seismic' time series for such predictions, my experiments demonstrate the capacity of the method to separate complex, real-world GNSS mixtures and to extract a seismic step from a 2-mixture-2-source scenario. Furthermore, the self-supervised transfer learning strategy developed on KITTI-Masks, with natural transitions and where ground truths are available, offers a promising blueprint for capturing and modelling the subtle, periodic dynamics of seismic cycles using SAR data.

Chapter	Research Questions	Methods
3	Is distance correlation effective as an independence metric for blind source separation when benchmarked against empirical and MINE-based mutual information, and normalised negentropy across AWGN, coloured-noise, and noise-free averaging scenarios?	I benchmark distance correlation against empirical and MINE-based mutual information and normalised negentropy by computing all four metrics on BPSK signals over AWGN channels, under coloured-noise perturbations, and in noise-free binary-signal averages, using analytic formulas, a neural-network MINE estimator, and double-centred distance-matrix computations.
4	How do different whitening transformations and their placement within the BSS pipeline affect the ability of distance correlation to quantify source independence and enhance overall separation performance?	I evaluated five whitening transformations against established statistical benchmarks and analysed the placement of the whitening step in the BSS pipeline.
5	Can a properly centred, differentiable distance-correlation loss be optimised in neural BSS architectures to reliably converge and match or outperform traditional and MINE-based losses on synthetic mixtures, GNSS seismic extractions, and SAR images?	I incorporated distance correlation as a differentiable loss into both separation and reconstruction neural network architectures, and benchmarked their convergence and extraction quality on synthetic three-signal mixtures, GNSS seismic mixtures, and SAR images using SI-SDR and independence metrics.
6	Can blind source separation—using distance correlation ICA and variational Bayesian ICA reliably decompose GNSS displacement time series into independent seismic and non-seismic geophysical processes, matching or exceeding the performance and interpretability of standard methods like FastICA?	I constructed two-station synthetic mixtures and a 40-station Southern California network, then applied FastICA, distance-correlation ICA, PyFastICA, and variational Bayesian ICA to decompose seismic and non-seismic components, evaluating performance via source correlation, SI-SDR.
7	Can distance correlation serve as a unified, differentiable test of statistical independence that drives and evaluates neural source-separation and self-supervised representation learning, matching or surpassing existing benchmarks in convergence, efficiency, and accuracy?	I first benchmarked distance correlation's computational cost and SI-SDR separation accuracy on synthetic and LibriSpeech audio mixtures using three resampling schemes. Then I integrated whitened distance correlation and double-centred Laplace proxy of distance correlation as self-supervised losses in CIFAR-10 and KITTI-Masks pipelines, respectively, to evaluate their learned representation quality.

Table 1.1: Research questions and methods for each research chapter.

CHAPTER 2

Background

This chapter reviews the literature on blind source separation and independent component analysis (ICA) in Section 2.1. Section 2.2 provides an overview of classical ICA techniques, such as FastICA, and their application to geodetic data, specifically focusing on Variational Bayesian ICA. The research in this thesis was inspired by the use of Variational Bayesian ICA on GNSS data in [11], which showed more promise than classical ICA methods.

In Section 2.3, other independence metrics, including mutual information and distance correlation, are presented as potential loss functions for neural network adaptations of ICA and representation learning. Moreover, the Scale Invariant Source to Distortion Ratio (SI-SDR) is introduced as the evaluation metric in this work. While SI-SDR is used as an independence metric for training in other works, my research does not centre on supervised learning, which can use this metric. Section 2.5 details GNSS and audio source separation datasets, while also introducing KITTI-Masks and CIFAR-10 for representation learning. In Section 2.6, machine learning for source separation and representation learning, along with whitening in its various forms, is illustrated. The final subsection provides information on a range of machine learning tasks and techniques used on geodetic data, which is the focus of this work due to its task-oriented nature.

While this background chapter is comprehensive, relevant state-of-the-art research will also be discussed in each chapter.

2.1 Independent Component Analysis

Section 1.2 of the Introduction introduced the concept of blind source separation (BSS) for extracting 'noise' signals from seismic signals that have been linearly mixed to create GNSS displacement time series. Independent Component Analysis (ICA) is a well-established method within the BSS framework and operates under limited and mild assumptions regarding the mutual independence of the underlying sources.

ICA is often regarded as the gold standard in various fields, including feature extraction, brain imaging, and telecommunications, for the task of separating a signal into its constituent components [22]. The fundamental principle behind ICA is the central limit theorem, which states that when random variables are combined, their resulting distribution tends to approximate a Gaussian distribution in most circumstances. The goal of ICA is to ensure that the separated independent sources are as non-Gaussian as possible, thus minimising the chance that they are simply sums of actual independent sources.

Over the years, many variants of ICA have been developed, each tailored to different measures of non-Gaussianity, and employing various computational techniques for source extraction. In the literature, entropy serves as a common measure of Gaussianity, as entropy is maximal for Gaussian distributions with a specific variance. Numerous methods have been proposed to estimate a signal's entropy, each involving a trade-off between speed and accuracy.

The FastICA algorithm is one of the most popular implementations of ICA [23] [24] [25]. FastICA can identify the bases onto which the input mixtures are projected, resulting in a lower-dimensional space that minimises Gaussianity. It has been applied to epileptic seizure detection [26], facial recognition [27], and mechanical fault detection in wind turbines [28].

The statistical properties of the FastICA algorithm have been analysed in several studies [29] [23] [30] [31] [32] [33]. These studies compared the results of the FastICA algorithm with those of the Cramér-Rao bound [34] [35] [30]. Additionally, the FastICA

algorithm has been modified for random variable signals that include complex numbers, as discussed in [36], [37] and [38], with a study containing the Cramer-Rao bound presented in [39].

The Cramer-Rao bound serves as a lower limit on the variance of unbiased estimators for a parameter, indicating the highest precision achievable when estimating that parameter based on a given set of data. This bound establishes a limit on the efficiency of these estimators. The work in [34] found that the Cramer-Rao bound for source separation is equivalent to an expression of the signal-to-interference ratio (SIR) derived using Pham's method. Therefore, the highest precision for estimating independent sources is achieved with the best SIR.

The general stability of FastICA has been examined using test problems proposed in [40] measuring performance using the Amari error [41]. The first test involved applying multiple algorithms to a two-component ICA problem, which included eighteen different source distributions in [40]. In the second test, the algorithms were applied to mixtures produced by two, four, eight, and sixteen underlying sources, with the source distributions randomly selected from the same eighteen distributions mentioned previously.

The Amari score is a measure used to determine whether ICA has successfully converged on the actual unmixing matrix. It provides a quantitative comparison for source extraction, offering a clearer picture of when the extraction performance is poor compared to source comparison metrics. In [42], it was found that FastICA performed worse than other ICA methods, failing to find optimal solutions (or the most non-Gaussian directions) in several cases. Furthermore, in the case of deflation ICA, where independent components are extracted one at a time, training could become stuck in a local minimum that does not correspond to the separation of all sources when attempting to separate independent components with multimodal distributions [33]. Notably, the algorithm utilising kurtosis was the only version that did not exhibit this issue.

Since the emergence of FastICA, a variety of studies and extensions have been developed to enhance its functionality and application. In [43], approximations for negentropy were developed based on the maximum entropy principle. When only one non-quadratic function, G, is used, the negentropy approximation is defined as follows:

$$J(X) \propto [E(G(X)) - E(G(X_{Gaussian}))]^2, \tag{2.1}$$

where the random variable, X, and the Gaussian variable, $X_{Gaussian}$, have zero mean and unit variance.

The user-defined non-quadratic function that approximates negentropy was explored in [44]. The speed of the algorithm improved by substituting non-linear contrast functions, like the hyperbolic tangent function, with rational functions (fractions with polynomials in the numerator and the denominator). The authors of [44] introduced two such rational functions. The statistical properties in each case remain similar to previous methods.

Indeed, the impact of non-linearity on the statistical accuracy of algorithms is well understood. Many variants of FastICA use different non-linear functions, such as *tanh* and x^3. However, employing a piecewise-linear function can help assess important characteristics of the non-linear function. For instance, [45] found that simple piecewise-linear-output non-linearities can closely approximate linearity while being sufficiently non-linear to enable effective separation when using the FastICA algorithm.

The primary advantage of the FastICA algorithm is its speed. Although the kurtosis-based FastICA is widely used, its speed has not been rigorously assessed or compared to other methods. This raises questions about the factors contributing to its popularity, as noted by [46]. In [46], the authors consider speed in relation to the computational complexity needed to achieve a specific source extraction performance.

RobustICA performs an exact line search optimisation of the kurtosis contrast function. The step size toward the global maximum of the loss function was identified among the roots of a fourth-degree polynomial, with this rooting performed algebraically at low cost for every iteration [47]. The RobustICA method was found to have faster convergence and higher robustness for a range of initialisations compared to FastICA with kurtosis [46].

FastICA has become dominant as a BSS technique, primarily due to its speed, but that comes with drawbacks. In [48], it was found that FastICA failed to recognise the optimal projections to minimise Gaussianity, even when the pattern is evident to the naked eye. Therefore, [48] introduced an m-spacing approximation of entropy to overcome

the limitations of FastICA. m-spacing approximates entropy, $\widehat{H}$, by ordering the random variable, **X**, of length N in an *increasing order* and applying the following equation:

$$\widehat{H}(X_1, X_2, ..., X_N) = \frac{1}{N} \sum_{i=1}^{N-m} log\left(\frac{N}{m}(X_{i+m} - X_i)\right). \tag{2.2}$$

Note that the user chooses m, the index used to calculate the spacing.

The *m*-spacing entropy approximation approaches the actual entropy as sample sizes increase, although it is slow to calculate.

To reiterate, the primary advantage of FastICA is its speed, but this can come at the expense of accuracy. Nonetheless, it is important to note that FastICA will still be used as a baseline in the following chapters due to its prevalent use in the literature.

2.2 ICA on geodetic data

While classical ICA has been seen in the previous section to both be popular and well-established for the BSS problem, especially in its FastICA form, it does not always perform an optimal decomposition for non-unimodal signals, such as those of slow slip events in GNSS time series. To overcome this, a Variational Bayesian ICA (vbICA) method was introduced to recover sources from GNSS signals. Before [11], few applications of ICA on geodetic data had been presented in the literature.

In [49], the FastICA algorithm was applied to a continuous GPS network in the Neapolitan volcanic area. In [50] and [51], a modification of the JADE algorithm [52] was applied to the gravity recovery and climate experiment (GRACE) data. These papers applied PCA to the GRACE data and model-based total water storage (TWS) time series, comparing the results with those from ICA and the VARIMAX rotation. The VARIMAX rotation is an orthogonal rotation method that minimises the number of random variables that have a high correlation with each factor. In these studies, the actual source signals are often unknown. However, in simulated cases with known input signals and artificially generated noise, the components extracted through ICA are closer to the optimal decomposition, showing less signal mixing compared to PCA and VARIMAX. [50] noted that the results obtained through ICA provide a more comprehensive representation than merely decomposing signals into trend, seasonal, and residual components. Here, the

residuals include inter-annual and episodic features. PCA and ICA encompass all temporal information in their components. Therefore, ICA may be better suited for investigating the unknown underlying behaviours in hydrological observations.

In this thesis, following [11], I am interested in signals that may not have known models, such as transient signals [53]. Transient deformations are less frequent but last longer, occurring over hours to months, significantly longer than the sudden slip that happens during an earthquake. As stated previously, the probability density functions of transient signals can be non-unimodal and, therefore, classical ICA may not perform the optimal decomposition [54]. In the following section, the extension to ICA used by [11], vbICA, for use on non-unimodal distribution data will be described in more detail.

Variational Bayesian ICA

A generative model M can be described in terms of observed variables X, hidden variables H, and hidden parameters, θ, and the relationships between these quantities. However, only the observed variables can be directly measured. Hidden parameters can be characterised by model weights $W = \{H, \theta\}$. The goal of the generative model is to find the best weights to explain the observations and match a priori knowledge that is defined within the structure of the particular model.

In a Bayesian framework, given a model M and the observed data X, maximising the posterior probability distribution function over weights W given the data X provides the best choice for W:

$$p(W|X,M) = \frac{p(X|W,M)p(W|M)}{p(X|M)}. \tag{2.3}$$

The denominator, which acts as a normalising constant, is called the evidence and is denoted by $p(X|M)$. Its calculation is often intractable, requiring an integration over all weight space. In the following, assuming that the user has defined a specific model, I will consider that it is given, and thus, the evidence is a constant. Therefore, Equation 2.3 simplifies to Equation 2.4.

$$p(W|X) = \frac{p(X|W)p(W)}{p(X)} = \frac{p(X,W)}{p(X)} \tag{2.4}$$

This is when the Kullback-Leibler divergence should be introduced in the context of

the vbICA calculation. The Kullback-Leibler (KL) divergence measures the difference between the true posterior $p(W|X)$ and its approximation $q(W)$. The goal is to make $q(W)$ as close as possible to $p(W|X)$. The aforementioned KL divergence is given by:

$$\begin{aligned} D_{KL}(q(W)\|P(W|X)) &= \log(q(W)) - \log(W|X) \\ &= \log(q(W)) - \log(P(X,W)) + log(p(X)) \\ &= \log(p(X) + \log(\frac{q(W)}{P(X,W)}) \\ &= \log(p(X) + F(q(W)\|P(X,W))), \end{aligned} \tag{2.5}$$

where $P(X,W) = P(X|W)P(W)$ and $F(q(W)||P(X,W)))$ is the variational free energy between the approximate and actual posterior. As the evidence term is independent of the function approximating the posterior, to minimise the KL divergence in Equation 2.5, one minimises the variational free energy, $F(q(W)||P(X,W)))$, to obtain $P(W|X)$.

The assumption of independence allows the factorisation $\prod_{i=1}^{N}(q(w_i))$ and thus, the use of the expectation-maximisation algorithm to solve the negative variational free energy (NFE) maximisation problem. When the NFE reaches its maximum value, the Kullback–Leibler divergence between the approximating probability density function of the hidden variables in the model and the true posterior probability distribution function is at a minimum.

In every case study in [11], the starting values of the hyper-parameters were arbitrary with non-informative priors, to let the data to reveal their intrinsic structure as much as possible. The Bayesian framework allows for the automatic relevance determination (ARD) method [55] to be used to determine the best number of sources to output. Every column of the mixing matrix represents how large an effect the underlying sources have on the outputted mixtures, a characteristic exploited by ARD.

One value, the *precision*, is allocated to each column, as a hidden parameter within vbICA for the sources. The precision value allocated to a source indicates how relevant that source is to the explanation of the data. A large value corresponds to a posterior dominated by the prior density, setting elements of the associated signal to 0, i.e. minimising the effect of this source signal on the data explanation. Therefore, looking at the precision values associated with each of the sources, those with high values can be excluded as they

Limitation	FastICA	vbICA
Pre-whitening	Requires input whitening	Whitening not strictly necessary
Missing data	Cannot handle missing data	Variational Bayesian framework handles missing data
Computational cost	Low per iteration and fast convergence	Higher computational overhead
Initialisation sensitivity	Sensitive: can get stuck in local minima or saddle points if sources are initially correlated.	Sensitive: PCA initialisation can bias towards PCA-like solutions and random initialisations may converge suboptimally.
Number of independent components selection	Manual choice	Manual choice
Distribution flexibility	Limited to kurtosis	Models each source PDF as a Gaussian mixture model, which is more flexible.
Scalability	Scales linearly with dimension	Memory and CPU cost grow faster with network and data size/dimensions
Uncertainty quantification	Produces point estimates	Many hyperparameters to tune with associated uncertainties

Table 2.1: A comparative table summarising limitations of FastICA against vbICA

do not add to the description of the data, allowing for the most likely number of sources supported by the observation data to be determined [54].

The use of the vbICA method on synthetic continuous GNSS data, introduced in [11], has subsequently been adapted to study geodetic signals in several geographic regions, such as the Cascadia subduction zone in [5], the Baja California region in [3], Central California in [56], Central Italy in [57] and California in [58]. This list is not exhaustive.

To conclude this section on vbICA, I present a comparative summary of the popular FastICA versus vbICA, in Table 2.1, to contrast the two methods.

2.3 Independence metrics

For random variables X and Y, with corresponding distributions F_X and F_Y and a joint distribution of $F_{XY} = F_{Y|X}F_X$, independence testing identifies whether $F_{Y|X} = F_Y$. In terms

of hypothesis testing, this can be written as:

$$H_0 : F_{XY} = F_X F_Y \tag{2.6}$$

$$H_1 : F_{XY} \neq F_X F_Y, \tag{2.7}$$

with the null hypothesis, H_0, corresponding to X and Y being independent.

2.3.1 Mutual information

A natural measure of dependence is mutual information, defined as:

$$I(X;Y) = H(X) - H(X|Y), \tag{2.8}$$

where $I(X;Y)$ denotes the mutual information between the random variables X and Y, $H(X)$ is the marginal entropy of X, and $H(X|Y)$ the conditional entropy. When X is independent of Y, the conditional entropy becomes equal to the marginal entropy of X.

While mutual information can determine whether two random variables are independent, in practice, the use of the continuous mutual information is hampered by the intractability of its exact computation [20]. While there are many non-parametric methods to approximate mutual information [59], and independence more generally [60] [61] [2], they all have limitations in the form of underlying assumptions about the probability density functions, or they do not scale well with sample size or dimension.

This gap led to the development of the parametric method known as Mutual Information Neural Estimation (MINE) [20], which uses the Kullback-Leibler divergence definition of mutual information:

$$I(X;Z) = D_{KL}(P_{XZ}||P_X \otimes P_Z), \tag{2.9}$$

where P_{XZ} is the joint distribution and P_X and P_Z are the marginal distributions of the two respective random variables.

In [20], the following lower bound was derived from the Donsker-Varadhan represen-

tation of the KL divergence:

$$D_{KL}(P||Q) \geq^{sup}_{T \in F} \mathbb{E}_P[T] - \log(\mathbb{E}_Q[e^T]), \tag{2.10}$$

where, the supremum is over the class F of functions T, such that both expectations are finite, and P and Q denote distributions satisfying certain mild conditions.

The idea in MINE is to choose F as the family of functions $T_\theta : X \times Y \to \mathbb{R}$ parametrised by a deep neural network with parameters $\theta \in \Theta$. In this case, the actual mutual information $I(X;Z)$ is approximated by the tight lower bound $I_\theta(X,Z)$:

$$I_\theta(X,Z) =^{sup}_{\vartheta \in \theta} \mathbb{E}_{P_{XZ}}[T_\theta] - \log(\mathbb{E}_{P_X \otimes P_Z}[e^{T_\theta}]). \tag{2.11}$$

In the actual implementation, the expectation values are estimated using empirical samples, or by shuffling the joint distribution along the batch axis. As $I(X;Y) \geq I_\theta(X,Y)$, the parametrised mutual information $I_\theta(X,Y)$ is maximised by gradient ascent, to become as close to the actual mutual information $I(X;Y)$ as possible.

2.3.2 Distance correlation

Distance correlation (DistCorr) [62] [63] is a powerful multivariate independence test based on energy distance, with a fast implementation provided by [64].

The distance correlation between $\mathbf{X}$ and $\mathbf{Y}$ in an arbitrary dimension is defined as:

$$R_n^2(\mathbf{X},\mathbf{Y}) = \begin{cases} \dfrac{\nu_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\nu_n^2(\mathbf{X})\nu_n^2(\mathbf{Y})}} & \nu_n^2(\mathbf{X})\nu_n^2(\mathbf{Y}) > 0 \\ 0 & \nu_n^2(\mathbf{X})\nu_n^2(\mathbf{Y}) = 0, \end{cases} \tag{2.12}$$

where the distance covariance ν_n^2 is defined as:

$$\nu_n^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl} \tag{2.13}$$

and

$$\nu_n^2(\mathbf{X}) = \nu_n^2(\mathbf{X},\mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}^2, \tag{2.14}$$

with the following dependence statistics:

$$
\begin{aligned}
a_{kl} &= |X_k - X_l|_p & \overline{a}_{k\bullet} &= \frac{1}{n}\sum_{l=1}^{n} a_{kl} \\
\overline{a}_{\bullet l} &= \frac{1}{n}\sum_{k=1}^{n} a_{kl} & \overline{a}_{\bullet\bullet} &= \frac{1}{n^2}\sum_{k=1}^{n} a_{kl} \\
A_{kl} &= a_{kl} - \overline{a}_{k\bullet} - \overline{a}_{\bullet l} + \overline{a}_{\bullet\bullet} \ . & & \qquad (2.15)
\end{aligned}
$$

B is the equivalent statistic for the random vector Y. Note that while the original distance correlation was defined in [62] over continuous random variables, Equation 2.12 is the corresponding empirical distance correlation.

Distance correlation measures the strength of the relationship between random variables, be they linear or non-linear associations. The values of distance correlation range from 0 to 1, where 0 represents independent variables and 1 indicates that the linear subspaces of the random variables are equal.

Taking a step back to enhance intuition, I will describe distance correlation in geometric terms. Imagine you compute all pairwise distances among the samples within the two variables; centre the two distance matrices; reshape them into vectors and then measure the cosine of the angle between those two vectors. That cosine is the distance correlation. A value close to zero indicates no dependence, meaning the double-centred distance vectors are orthogonal to each other. Conversely, a value nearing one signifies that the distance patterns align, indicating the existence of dependence.

Additionially, partial distance correlation, introduced in [63], calculates the distance correlation between two variables, while excluding the effect of a third variable. It can be used to evaluate correlations between feature spaces of different dimensions. In [21], partial distance correlation was utilised to compare the functional behaviour of neural network models. This comparison was crucial in understanding what these models learn and identifying strategies to improve regularisation and/or efficiency. This methodology can be applied to various tasks, such as conditioning multiple deep learning models with respect to one another and learning disentangled representations. Although previous research has shown that comparing two different feature spaces is possible using Canonical Correlation Analysis (CCA), this approach has not been frequently used.

Finally, the Hilbert Schmidt Independence Criterion (HSIC) [65] [66] [67] is a kernel-based analogue of distance correlation, with a Gaussian median kernel by default used [68].

2.3.3 Other independence metrics

Other statistical definitions of independence include Canonical Correlation Analysis (CCA) [69] [70] [71] [72] and Rank Value (RV) [73] [74], which are multivariate versions of Pearson's correlation. CCA can model the correlation between datasets in two ways, either stating that one dataset is the dependent variable or not. RV is the multivariate generalisation of the squared Pearson correlation coefficient.

The Heller Heller Gorfine (HHG) [75] [76] independence test compares inter-sample distances and computes the Pearson statistic between the distance matrices. This test is accurate in certain settings but computationally complex and not easily interpretable. Kernel mean embedding random forest (KMERF) [77] uses a random forest-based similarity matrix to generate an independence statistic. KMERF is a highly accurate independence test and provides information on important features. However, it is slow, requiring training for each permutation, and is less viable for training a neural network in real-time. The Maximal Margin Correlation (MMC) [78] determines the highest pairwise correlation by analysing the correlations across all combinations of dimension pairs from $\mathbf{X}$ and $\mathbf{Y}$. This method can be effective but also computationally intensive.

The main advantage of CCA and RV methods is their speed. In comparison, HHG, MMC, and KMERF provide accurate independence measures but are slower.

2.4 Performance and metrics

Several metrics to evaluate source separation techniques were proposed in [79]. An estimation of the source as a function of time t can be decomposed as:

$$\widehat{S}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t). \tag{2.16}$$

In [79], it was proposed that this decomposition could be combined with orthogonal projections. Equations 2.17-2.19 represent three orthogonal projections. Equation 2.17 represents projection onto the subspace spanned by one specific source; Equation 2.18 onto the subspace of all other sources, and Equation 2.19 onto the subspace spanned by the noise component.

The computation of s_{target} is:

$$s_{target} = \frac{\langle \widehat{s}, s \rangle s}{\langle s, s \rangle}. \tag{2.17}$$

For e_{interf}, in the case that the sources are assumed to be mutually orthogonal, Equation 2.18 holds.

$$e_{interf} = \sum_{j \neq j'} \frac{\langle \widehat{s}_j, s_{j'} \rangle s_{j'}}{\langle s_{j'}, s_{j'} \rangle} \tag{2.18}$$

It is assumed that the noise signals, n_i, are mutually orthogonal and orthogonal to each source. Therefore:

$$P_{s,n}\widehat{s}_j \approx P_s \widehat{s}_j + \sum_{i=1}^{m} \frac{\langle \widehat{s}_j, n_i \rangle n_i}{\langle n_i, n_i \rangle}, \tag{2.19}$$

where m is the number of mixtures.

It is assumed that the actual source signals and any noise, if present, are known when calculating the above metric. Additionally, it is presumed that the noise remains consistent between the training and test datasets. Different types of noise can create problems if the training and testing conditions do not match.

A numerical performance criterion based on the computed energy ratio in decibels (dB, though I drop the unit when I report my results) is the Source-to-Distortion Ratio can be defined as:

$$SDR = 10log_{10} \frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}. \tag{2.20}$$

The vector norm is denoted as $||x||$, and e_{artif} refers to the error term associated with forbidden or artifact distortions.

Since 2016, there has been a significant uptake in the use of machine learning tech-

niques for source separation tasks, with the scale-invariant source-to-noise ratio (SI-SNR) gaining popularity. SI-SNR is defined in Equation 2.22.

$$e_{noise} = \widehat{s} - s_{target} \tag{2.21}$$

$$SI-SNR = 10log_{10}\frac{\langle s_{target}, s_{target}\rangle}{\langle e_{noise}, e_{noise}\rangle} \tag{2.22}$$

In 2018, Le Roux et al. [80] proposed a more robust metric in place of SDR, known as the scale-invariant SDR, such that $s-\widehat{s}$ is orthogonal to the target s. One way is to scale the target such that the residual is orthogonal to it by finding the orthogonal projection of the estimate $\widehat{s}$ on the line spanned by the target s.

SI-SDR can be defined as:

$$SI-SDR = \frac{|s|^2}{|s-\beta\widehat{s}|^2}, \tag{2.23}$$

for β subject to $s \perp (s-\beta\widehat{s})$.

The optimal scaling factor is obtained by $\alpha = \frac{\widehat{s}^T s}{||s||^2}$ and scaling the source by this factor outputs $e_{target} = \alpha s$. The estimate $\widehat{s}$ is decomposed as $\widehat{s} = e_{target} + e_{residual}$. Thus, SI-SDR can be written as follows:

$$\begin{aligned} SI-SDR &= 10log_{10}\frac{||e_{target}||^2}{||e_{residual}||^2} \\ &= 10log_{10}\frac{||\frac{\hat{s}^T s}{||s||^2}s||^2}{||\frac{\hat{s}^T s}{||s||^2}s-\hat{s}||^2}. \end{aligned} \tag{2.24}$$

2.5 Datasets

In blind source separation, datasets fall into one of two categories: *Application-orientated*, and *Diagnosis-orientated* datasets. Application-oriented datasets, such as professionally produced music recordings, involve real signals that simultaneously contain all issues

that can arise during source separation. Diagnosis-related datasets are artificially created to highlight one specific issue, such as noisy surroundings in speech separation.

Here, I introduce the benchmark datasets used within this thesis, such as GNSS displacement time series and datasets pertinent to machine learning tasks associated with source separation or representation learning.

2.5.1 GNSS data

As previously mentioned, the Nevada Geodetic Laboratory (NGL) provides routine updates of GNSS position coordinates across a global network. Approximately 10,000 stations offer daily-resolution position time-series, which are updated weekly. In addition, over 5,000 stations supply 5-minute-resolution data updated daily, while around 2,000 stations deliver 5-minute-resolution data on an hourly basis for near-real-time applications.

GNSS data is provided in plain text format and includes metadata such as station name and recording date, often redundantly encoded in multiple formats. The core content consists of both raw measurements and derived quantities. Two primary formats are commonly used: .tenv and .tenv3. My work has utilised the .tenv format, though the required data fields are common to both, making the distinction largely inconsequential for this application. In the .tenv format, the relevant fields are delta_e, delta_n, and delta_v, representing eastward, northward, and vertical displacements relative to a reference frame at each epoch. In contrast, the .tenv3 [81] format provides easting, northing, and vertical fractional coordinates, absolute values mapped via projection (e.g. longitude offset converted to metres along a parallel). The key difference lies in interpretation: delta_e reflects the time-series of eastward displacement (how far a station has moved from its baseline). Whereas, easting denotes the absolute projected coordinate at each epoch. While my analysis was based on .tenv, future workflows may benefit from adopting .tenv3 for its enhanced spatial context.

This data is made publicly available with related information, such as the position coordinates for each of the geodetic-quality GPS stations from hundreds of different organisations, including the International Global Navigation Satellite Systems Service (IGS) and UNAVCO, with multiple data intervals and reference frames that benefit different

users. To obtain as much useful GPS data as possible and assemble it in one point for easy access, the NGL uses more than 130 Internet Archives [82].

According to NGL, 'One person's noise may be someone else's signal.' For example, to determine a station's position, an estimate of the variability of the atmospheric refraction of the GPS signal is required, which is influenced by water vapour content. Data from this model is helpful for investigations where atmospheric refraction is deemed to be noise. Therefore, NGL has provided the tropospheric refraction parameters every 5 minutes since 1994 from over 18,600 stations.

NGL also updates station velocities in a global reference frame to determine the Earth's surface deformation rates. The velocities are estimated using the Median Interannual Difference Adjusted for Skewness (MIDAS, see usage in Section 4.3.2 and Chapter 6) method, a median-based velocity estimator for GPS stations. This method is designed to handle the increasing volume of data that can present issues which experts may identify but are often overlooked. MIDAS is robust against outliers, seasonality, and step functions caused by earthquakes or equipment changes, as well as variations in statistical data [82]. The velocity estimator is a variation of the Theil-Sen median trend estimator, a method for robust linear regression that chooses the median slope among all lines through pairs of two-dimensional sample points. The benefit of the Theil-Sen median trend estimator over least-squares trend estimates is that it is more robust to potential events within the data that would otherwise go undetected.

In order to mitigate seasonality and step discontinuities associated both with earthquakes and equipment changes, MIDAS selects pairs that are separated by a year. If there is missing data, this condition is relaxed in order for all of the data to be used. If a slope is from a pair that spans a step function, one-sided outliers can bias the median. To reduce this, MIDAS removes the outliers and recomputes the median.

Lastly, NGL provides data on the timing of known earthquakes. A ground station is identified as potentially containing a step function in its time series if it falls within a radius from the epicentre determined by the magnitude of the earthquake. Additionally, NGL maintains records of any equipment or code changes made that can produce steps in the GNSS time series.

2.5.2 LibriMix

LibriMix [17] is an open-source alternative to WSJ0-2mix [83] and its noisy extension, WHAM!. Based on LibriSpeech [84] (a corpus of approximately 1000 hours of 16kHz read English speech), LibriMix consists of two- or three-speaker mixtures without noise, in the clean case, or combined with ambient noise samples from WHAM! [85]. WSJ0-2mix has become the reference dataset for single-channel speech separation. Therefore, most deep learning-based speech separation models today are benchmarked on it. However, recent studies have shown important performance drops when models trained on WSJ0-2mix are evaluated on other similar datasets.

2.5.3 KITTI-Masks

KITTI-Masks [18] comprises pedestrian segmentation masks from the autonomous driving vision benchmark KITTI-MOTS [86]. The KITTI-Masks dataset consists of 2120 sequences of binary masks of pedestrians, with varying sequence lengths between 2 and 710.

I selected the KITTI-Masks dataset, along with CIFAR-10, to showcase representation learning. Representation learning can be viewed as a form of latent source separation. Rather than disentangling physical sensor signals, it isolates independent factors of variation in feature space. The KITTI-Masks dataset provides video sequences with pixel-level instance masks, yielding natural temporal transitions in latent representations, an ideal proxy for testing models that must capture evolving patterns, such as those found in seismic deformation time series.

In blind source-separation terms, the KITTI-MOTS dataset's instance segmentation masks serve as proxy ground-truth sources, since each pixel is assigned to a single object and thus to an independent component. Its video sequences naturally capture non-stationary mixing, mirroring the evolving mixtures found in seismic or other geophysical signals. With pixel-level masks, one can rigorously evaluate how well a model recovers latent components using metrics and quantify feature leakage between sources, as well as test robustness to occlusion. Moreover, its semi-automatic annotations enable self- or weakly-supervised methods to discover structured latent factors without the need for

dense manual labels, making KITTI-MOTS a powerful benchmark for bridging representation learning and real-world source-separation challenges.

2.5.4 CIFAR-10

CIFAR-10 [87] consists of labelled subsets of 80 million images in the tiny images dataset. The CIFAR-10 dataset consists of 60,000 32x32 colour images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.

The dataset is divided into five training batches and one test batch, each with 10,000 images. The test batch contains exactly 1,000 randomly selected images from each class. The training batches contain the remaining images in a random order. Some batches may contain more images from one class than another. Between them, the training batches contain exactly 5,000 images from each class.

CIFAR-10 is a widely recognised benchmark for representation learning (the separation of features in latent space) consisting of 10 classes, and it can be easily expanded to 100 classes through the CIFAR-100 dataset. It was selected as the initial task for representation learning because it serves as a baseline dataset for the W-MSE self-supervised representation learning task, which I modified to include distance correlation. Therefore, it was chosen to function as a baseline. However, it could also be used similarly for representation learning for labelled SAR images.

2.6 Machine learning

This section discusses relevant machine learning methods for source separation and representation learning, as well as an overview of various techniques applied to geodetic data.

2.6.1 Source separation

This thesis uses audio source separation, given that it is a well-developed field of study. It is important to note that in the audio source separation cases, the training sets include the underlying sources, such as individual instruments in a song. In contrast, geodetic

data is unlabelled. One of the most commonly used datasets for the speech separation task is Libri2Mix [17], which consists of a mixture from two speakers along with ambient noise from WHAM! [85], Several methods applied to this dataset will subsequently be discussed.

Deep learning techniques for single-channel speech separation fall either into the time-frequency (TF) domain, the most common technique until recently, or the end-to-end time domain. The TF features are produced by applying a short-time Fourier transform to data in the time domain. These features are then separated to obtain features from each source, with the source waveforms obtained by using the inverse of the aforementioned Fourier transform [88] [89] [90] [91] [92]. More recently, there have been significant advances in time-domain approaches, which input the mixture waveform directly into an encoder-decoder framework [93] [94] [95] [96].

Both of these methods require what is known as permutation invariant training (PIT), as source separation is not ordered. Therefore, one has to properly match the ground truth with estimated signals when evaluating the reconstruction errors. PIT [97] dynamically chooses the best label assignment. The assignment of labels in early epochs can be unstable, leading to slower convergence and reduced performance [98]. In [99], it was found that self-supervised pre-training could effectively stabilise the label assignment in PIT during the training of speech separation models, and significantly reduced label assignment switching during training, resulting in a faster convergence and improved performance.

This method was surpassed by TDANet [100]. The top-down attention in TDANet is extracted by the global attention (GA) module and the cascaded local attention (LA) layers. The GA module takes multi-scale acoustic features as its input to extract global attention signal, which then modulates features of different scales by direct top-down connections. The LA layers use features of adjacent layers as their input to extract the local attention signal, which is used to modulate the lateral input in a top-down manner.

In [101], they proposed Vocoder, a diffusion model pre-trained on single-speaker voices, and applied it to the output of a deterministic separation model, obtaining state-of-the-art separation results.

Finally, when evaluated on SI-SDRs, the best network as of writing is based on the

MossFormer model [102], known as the MossFormer2. It is an alternative approach to SepFormer [103], involving joint attention. MossFormer was found to emphasise longer-range, coarser-scale dependencies, with a deficiency in effectively modelling finer-scale recurrent patterns. In MossFormer2, a recurrent module based on a sequential memory network replaced the recurrent neural networks to capture recurrent patterns without recurrent connections, to produce a hybrid model that could model long-range coarse-scale and fine-scale recurrent patterns.

2.6.2 Representation learning

Whilst independent components are uncorrelated, decorrelated data is not necessarily independent. However, in many applications, whitening is used as a preprocessing step for data decorrelation, transforming the inputted data into an output with identity covariance matrix. Whitening, or synonymously sphering, takes an n-dimensional random variable $\mathbf{X}$, with mean μ and covariance matrix $\mathbf{\Sigma}$, and linearly transforms it into a new whitened random vector $\mathbf{Z}$, with unit diagonal covariance. Thus, the whitening transform matrix satisfies $\mathbf{W}^T\mathbf{W} = \mathbf{\Sigma}^{-1}$. Notice that this condition does not uniquely determine the whitening matrix $\mathbf{W}$. Due to the rotational freedom in whitening, there are infinitely many matrices that conform to this constraint.

In the study conducted by Kessy et al. [6], five types of whitening methods (ZCA-Mahalanobis, ZCA-cor, PCA, PCA-cor, and Cholesky) were examined to determine if any statistical function is optimised. All methods, except for Cholesky whitening, showed some form of optimisation. The findings from this research will be discussed in greater detail in Chapter 4. In that chapter, I will explore these five whitening methods as part of a source separation task to assess whether any form of whitening encourages independence, specifically in terms of distance correlation.

Whitening has the potential to aid the search for independence. In the case of FastICA, whitening reduces signal redundancy, preventing the sources from tending towards a single solution by ensuring they remain uncorrelated. Whitening has been integrated into machine learning tasks, such as self-supervised representation learning, which will appear in Chapter 7. In most instances, self-supervised representation learning (SSL) methods are based on contrastive losses for a discrimination task, whereby augmented

versions of an image instance (labelled as positive) are compared to instances extracted from other images (labelled as negative), with balanced sets of labels.

Contrastive methods have seen much uptake in recent work given the availability of large unlabelled datasets. Several theories about contrastive learning and its informative representations have been proposed, with clashes between theoretical and empirical observations [104].

In [105], it was found that the contrastive loss converges asymptotically to a combination of two components: one that encourages the representations of positive pairs to be similar, and another that acts as a uniformity term, which maximises the entropy of the learned latent distribution. This approach has been shown to yield results for downstream tasks that are comparable to, or even better than, previous methods.

The triplet loss is an effective method for determining metric spaces based on human perceptual similarity [106] [107] [108] [109] [110]. It works by encouraging the positive sample to be close to the anchor while ensuring that the negative sample is at least a certain margin away from the positive sample. Both triplet and contrastive losses are sensitive to the quantity and quality of negative samples [111]. To address these challenges, whitening mean squared error (W-MSE) was proposed as an alternative method in [4]. This whitening process, along with normalisation, imposes a uniformity constraint on the distribution of features on the hypersphere, while the MSE loss applied to positive pairs enhances the mutual information through alignment. The semantics of positive samples are shared while being different from those of the negative samples. These properties encapsulate the critical aspects of contrastive loss as outlined in [105].

A commonly used contrastive function is InfoNCE, a probabilistic contrastive loss which encourages the latent space to capture information that is maximally useful to predict future samples [112]. InfoNCE is based on the concept of noise contrastive estimation (NCE) introduced in [113], which effectively estimates complex statistical models without the need for normalisation. The InfoNCE loss was defined in [112] for a set $X = \{x_1, \ldots, x_N\}$ of N random variables. This set contains one positive sample drawn from the distribution $p(x_{t+k}|y_t)$ and $N-1$ negative samples from $p(x_{t+k})$. The InfoNCE loss is formulated as follows:

$$L_N = -E_X \left[\log \frac{f_k(x_{t+k}, y_t)}{\sum_{x_j \in X} f_k(x_j, y_t)} \right]. \tag{2.25}$$

Here, y represents the context random variable (denoted as variable c in [112]), and k indicates the number of steps used to define future observations. When optimising Equation 2.25, the function f estimates the density ratio:

$$f_k(x_{t+k}, y_t) \propto \frac{p(x_{t+k}|y_t)}{p_{t+k}}. \tag{2.26}$$

.

It is important to note that Equation 2.25 is related to mutual information as follows:

$$I(x_{t+k}, y_t) \geq \log(N) - L_N. \tag{2.27}$$

Thus, minimising the InfoNCE loss maximises to the lower bound of mutual information, as described in [112].

In [114], the work of [115] was used to show that for InfoNCE distribution matching implies parameter matching. In [116], learned latent representations were associated with ground-truth generative factors to determine under what conditions data generation can be inverted to recover the true latent factors.

In the context of the InfoNCE case, the theoretical motivation for contrastive learning involves examining the mutual information between different views, or positive pairs, as discussed in various studies [112] [111] [117] [104] [118]. The InfoMax principle can be employed to maximise this mutual information using Jensen-Shannon Divergence. However, in [119], the InfoMax principle was applied to contrastive learning to evaluate a loss that was tightly bound to mutual information. It was discovered that having a tighter bound on mutual information could adversely affect the separations produced, making it less favourable for contrastive learning.

2.6.3 On geodetic data

The application of machine learning to geodetic data is an emerging field that is gaining traction. The ideal outcome of this thesis would have been the prediction of earthquakes.

While efforts have been made towards this goal, it has not been realised in this work.

Other research in this field has used GNSS-based and geodetic data. In the case of GNSS data, there are strain accumulation models utilising standard [120] and high-rate [121] [122] [123] GNSS velocity fields for real-time tracking. Moreover, ionospheric anomalies [124] [125], in the form of Total Electron Content (TEC) detections, may serve as precursors for seismic events. Some techniques use a combination of SAR and GNSS data, such as in [126], to analyse coseismic and interseismic activities. Recently, there has been a growing trend in employing synthetic datasets to train deep learning models, particularly convolutional neural networks and transformers, to estimate earthquake magnitudes. However, these methods face challenges when it comes to generalising across different tectonic systems.

In the case of InSAR data, machine learning has been used to: detect, locate, and classify the presence of co-seismic-like surface deformation within an interferogram using a CNN called SarNet [127]; classify interferometric fringes in wrapped interferograms with no atmospheric corrections using AlexNet [128]; obtain a relative landslide hazard map [129] using Artificial Neural Networks, Generalised Boosting Models, and Maximum Entropy, on InSAR datasets; and, finally, obtain a full coverage map of the groundwater-induced land subsidence using boosted regression trees and extreme gradient boosting algorithms for InSAR analysis [130].

In the case of GNSS data, there has been an uptake in the use of machine learning techniques to deal with the large quantity of data that GNSS provides. Before discussing machine learning for GNSS data analysis in the context of geodetic data, it is worth noting that there are other contexts for using machine learning on GNSS data. For instance, Google DeepMind is enhancing Google Maps by integrating deep learning with Street View. It uses the GNSS location of the Street View car and address information from the imagery to refine their existing knowledge [131].

In the systematic review [132], which focuses on machine learning techniques applied to GNSS data, it was found that the primary methods used include neural networks, decision trees, and random forests. However, this list is not exhaustive, as multiple other techniques have also been explored.

In [124], an approach to the prediction of earthquakes and determination of their mag-

nitude using neural networks and ionospheric disturbances was proposed. Vertical Total Electron Content from the National Oceanic and Atmosphere Administration was used as the training data, achieving an accuracy of 85.71% in validation assessment to predict Tres Picos Mw = 8.2 earthquake from 1:30 UTC to 04:00 UTC, approximately 3 hours before the seismic event. This exemplifies how different 'noise' terms within the GNSS signal are subjective. However, this work has not gained much traction.

In [133], researchers examined the data characteristics that enhance the performance of machine learning binary classifiers for predicting imminent slip events. Their goal was to improve the understanding of deformation associated with the seismic cycle in subduction zones. The machine learning classifiers used GNSS-like surface deformation data derived from seismotectonic scale modelling, as discussed in [134].

The findings indicated that the timing of when an event was identified as imminent significantly influenced the performance of the classifiers. Additionally, factors such as the density of monitoring stations and the duration of data collection also contributed to performance. The study concluded that accurate earthquake predictions were not achievable with the algorithms used, even in a simplified scenario with an optimally designed monitoring network. However, the predicted alarm periods aligned reasonably well with actual earthquakes, particularly when multiple seismic cycles were recorded and a longer time-frame was considered imminent for an event.

Another interesting paper is [135], which looks at slip events known as slow earthquakes, which slip little and have a high frequency. The slow slip history between 2007 and 2017 for the Cascadia fault was used to assess predictability of such events, as multiple slow earthquakes have occurred in different parts of the region over a relatively short period of time. The system dynamics were characterised using embedding and extreme value theory, with a non-linear chaotic system found instead of a stochastic system. It was found that the prediction power of this setting was on the order of days, with long-term predictions impossible, similar to weather forecasting. It was thought that regular earthquakes might similarly be predictable but with a limited prediction horizon. Their analysis also implies that it should be possible to forecast the onset of large slow slip events ahead of time, based on an explicit deterministic representation of the system dynamics or some machine learning algorithm that would implicitly capture it.

In a more controlled setting, frictional motion for a laboratory fault as it passes through the stability transition from stable sliding to unstable motion was studied in [136]. It was found that the seismic cycle of a lab 'earthquake' exhibits characteristics similar to those of natural slow earthquakes. The 'labquakes' were best modelled by a random attractor based on sliding rate- and a state-dependent friction whose dynamics are stochastically perturbed. Small variations of the shear and the normal stress applied to the fault greatly affected the macro dynamics and recurrence time of these 'labquakes', and the non-linearity of the friction also reduced the predictability of otherwise periodic macroscopic dynamics. Regarding tectonic faults, they found that small stress field fluctuations can lead to variations in earthquake repeat time of a few percent.

CHAPTER 3

Distance correlation as an independence metric

3.1 Introduction

In this chapter, I will utilise non-parametric tests to evaluate the effectiveness of distance correlation as a metric for assessing independence. Rather than using it as a loss function for an optimisation algorithm, I will first compute distance correlation values on various signals. I will then compare its behaviour with two other commonly used metrics for measuring signal independence: mutual information and negentropy.

Section 3.2 employs a fundamental model from Communication Theory [137] [138] [139], which models the transmission of a digital signal over a noisy analogue channel. My approach is essentially similar, though more simplistic, to the one employed in [140]. Specifically, it models the transmission as binary phase shift keying (BPSK) modulation over an additive white Gaussian noise (AWGN) channel. Using this non-parametric communication model, I calculated distance correlation, mutual information, and negentropy to quantify how effectively they differentiate between the input and output signals with various levels of AWGN noise.

In this particular case, the simplicity of the BSPK with AWGN setting means that the computation of the mutual information can be done directly through a closed form for-

mula, rather than relying on MINE. This also offers an opportunity to assess the accuracy of MINE and compare its various configurations. See Section 3.2.3.

Section 3.3 compares the behaviour of the distance correlation against negentropy for various noise colours. The motivation for this broadening of the investigation was two-fold. First, the observation that Gaussian noise may intrinsically favour the negentropy independence metric, which itself is based on the notion of non-Gaussianity. Secondly, certain findings in the literature suggest that noise in time series data is better modelled by a combination of white and other colours of noise [141].

Finally, for completeness of the investigation, Section 3.4 departs from the additive noise models of the previous two sections, which create the noisy input channel by adding a continuous noise function to a discrete signal, and instead generates an input signal by adding together a number of discrete signals. As the number of signals increases, the distribution will approach a Gaussian. In a manner similar to Section 3.3, I will compare distance correlation with negentropy.

In all cases, the simplicity of the setting means that the values of the various independence metrics can be computed quickly and accurately.

3.2 Binary Phase Shift Keying in an AWGN channel

In the BPSK with AWGN example, the first signal X, called the *input channel*, is a random variable with two equiprobable values of -1 and 1. The second signal Y, called the *transmitted signal*, is the input channel with added Gaussian noise $N \sim \mathcal{N}(0, \sigma)$, which has a mean of zero and a specified variance. That is $Y = X + N$.

3.2.1 Computation of metrics

While the closed-form formula for distance correlation is presented in Section 2.3.2, I will also introduce the empirical mutual information formula for the BPSK over an AWGN channel and provide a brief recap of negentropy. Note that the computation of the distance correlation and negentropy are consistent for all of the tests in this chapter.

Mutual Information

The BPSK modulation problem allows for a non-parametric comparison of distance correlation with a known empirical mutual information. The mutual information for binary phase shift keying over a Gaussian channel has been introduced to compare the estimated independence metric with an empirical result. The input channel can take values of 1 and -1, with equal probabilities of occurrence. The transmission adds Gaussian noise with zero mean and variance, σ^2. The empirical mutual information is then calculated using Monte Carlo integration.

The mutual information is calculated as:

$$\begin{aligned} I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\ &= E_{P(X,Y)}\left[log\left(\frac{P(X,Y)}{P(X)P(Y)}\right)\right] \\ &\approx \frac{1}{K}\sum_{k=1}^{K} log\left(\frac{P(x_k,y_k)}{P(x_k)P(y_k)}\right). \end{aligned} \tag{3.1}$$

In Equation 3.1, there are K samples of the joint distribution used to calculate the empirical mutual information corresponding to the sample, using Monte Carlo sampling techniques. x_k and y_k are samples from the joint distribution. The above is equivalent to:

$$\frac{1}{K}\sum_{k=1}^{K} log\left(\frac{P(x_k,y_k)}{P(x_k)P(y_k)}\right) = \frac{1}{K}\sum_{k=1}^{K} log\left(\frac{P(y_k|x_k)}{P(y_k)}\right). \tag{3.2}$$

As the component of X can be equiprobably -1 or 1, the denominator can be expanded for the conditional cases of a sampled y_k being produced by an X corresponding to one of these two values, as seen in Equation 3.3.

$$I(X;Y) \approx \frac{1}{K}\sum_{k=1}^{K} \log\left(\frac{P(y_k|x_k)}{0.5\cdot P(y_k|X=-1)+0.5\cdot P(y_k|X=1)}\right). \tag{3.3}$$

In Equation 3.3, the numerator is the distribution of a Gaussian centred at x_k, with a variance of σ^2. The denominator contains two Gaussians centred at -1 or 1, with a variance equal to that of the numerator. Thus, the mutual information can be calculated, with known input signals X and Y, using the Gaussian probability density function.

Negentropy

Negentropy, $J(X)$, can be seen as the opposite of entropy, measuring the difference between the entropy, $H(X)$, of a system, X, and its maximum entropy, $H(X_{Gaussian})$, where entropy is a measure of a state's randomness or uncertainty. The differential entropy of a random vector X with a probability density function f(X) is defined as:

$$H(X) = -\int f(X) log(f(X)) dX \ . \tag{3.4}$$

Equation 3.5 provides the mathematical definition for negentropy:

$$J(X) = H(X_{Gaussian}) - H(X), \tag{3.5}$$

where $X_{Gaussian}$ is a Gaussian random variable with the same variance as the system, X.

As central limit theorem states that under most conditions, the sum of random variables tends towards a Gaussian, maximising the negentropy of signals can separate underlying signals from their sums.

While negentropy would naturally be an optimal measure of non-Gaussianity, estimating negentropy using Equations 3.4 and 3.5 would require a probability density function.

In this chapter, I calculate the entropy as the negative of the expectation of the logarithm of the Softmax function (the Softmax acting as a probability density function), defined as $H(X) = -\mathbb{E}(log(p(X)))$. From the entropy, negentropy is computed using Equation 3.5.

Since negentropy is not a pairwise metric, it is only applied to the output signal as the input signal remains constant throughout the experiment.

3.2.2 Results

The mutual information (directly computed using Equation 3.3), the negentropy and the distance correlation for the BPSK examples with various variances can be seen in Figure 3.1. The signal in the experiment consisted of 1,000 samples.

In this example, a single binary input signal was produced and kept constant through-

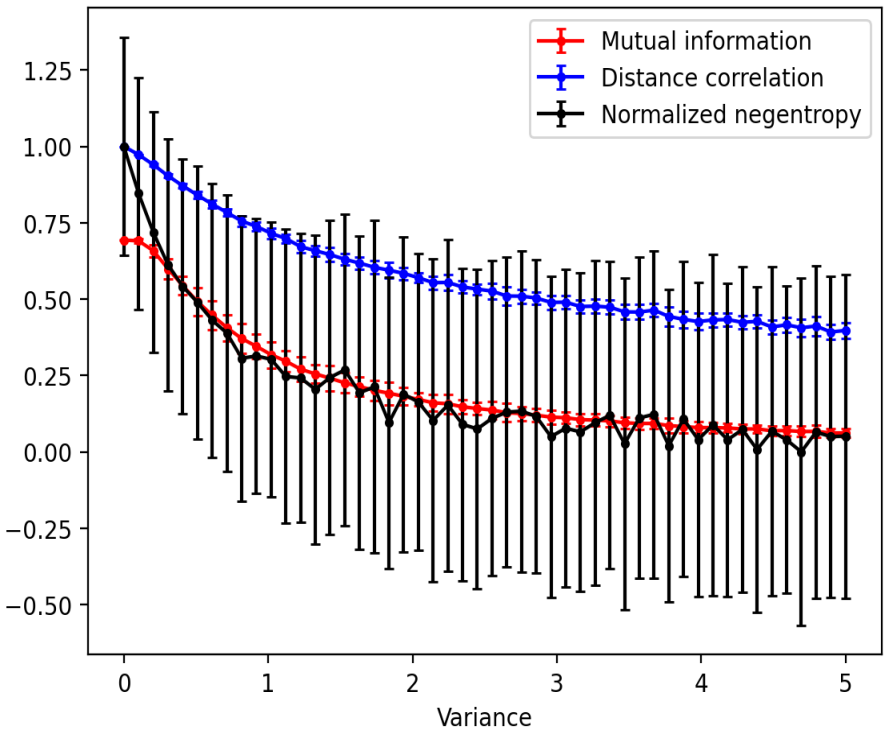


Figure 3.1: Independence metrics in relation to changing AWGN variances for the BPSK problem. The red line is the empirical mutual information (Equation 3.3), and the blue and black lines are the distance correlations and the *normalised* negentropy. For each Monte Carlo estimate of mutual information, I computed the standard error from both the sampling variability and the variability across input signals. I then treated these two sources of uncertainty as independent, summing their variances, taking the square root to get a combined standard error, and used that to generate the error bars.

out the experiment. For each of 50 variance values equally spaced between 1e-16 and 5, 10 noisy signals were produced. These 10 repeats of the computation allow for a standard deviation to be calculated, seen as error bars on the plots.

The Pearson correlation between the distance correlation and the empirical mutual information is 0.984, identifying a high correlation between the two sets of values. Both the metrics decrease as the variance increases, plateauing with increased variance.

While there is some uncertainty associated with the calculation of distance correlation, the standard deviation remains within acceptable limits, reaffirming its reliability as a metric. Additionally, there is minimal overlap among the different points associated with other variances shown in Figure 3.1. Although the distance correlation metric is higher than the ground truth mutual information and the normalised negentropy, I will demonstrate that this is an advantage of distance correlation in future.

In the closed-form curve of mutual information, there is a plateau at very low variances, the left-hand side of Figure 3.1. This plateau corresponds to the curve approaching the theoretical maximum, where the input and output signals are equal. In this case, Equation 3.3 would reduce to $log(1/0.5) = log(2) = 0.693$. The plateaus for small and large variances are explained by the effect of the denominator of Equation 3.3, which, on average, is slow to increase with slight changes in the variance for variances close to zero and then plateaus for higher variances. The plateau close to zero represents the case where the empirical mutual information is calculated between a signal and a signal very close to itself, producing high mutual information values.

Negentropy, as defined in Equation 3.5, tends to approach zero as noise increasingly dominates the signal at higher variances. It is important to note that the error bars for negentropy are significantly larger. This is mainly due to the calculation of the differential entropy using an estimated Gaussian term derived from sampling a distribution. Therefore, it is preferable to standardise the signal and ignore the entropy of the Gaussian, as this remains constant when variances are equal, rather than using negentropy with an estimation of Gaussian entropy. For easier comparison, I normalised the negentropy values; its actual range is between -0.00907 and 0.0659.

Distance correlation decreases as the variance of the AWGN added to the transmitted signal increases, producing a transmitted signal tending toward white noise. As the transmitted signal tends toward white noise, the inputted and transmitted signals become more independent. The distance correlation does not approach zero. Upon further investigation, this is also true of Pearson's correlation coefficient, suggesting a linear relationship between the inputted and transmitted signals. As a result, the distance correlation should not tend toward zero since there is at least a linear relationship between the two signals. This demonstrates that distance correlation is a more effective independence metric for machine learning based on this test.

The two methods differ as distance correlation examines pairs of random variables, whereas negentropy focuses on an individual random variable. Therefore, to eliminate redundancy in solutions, negentropy requires an extra step, such as whitening, to remove the correlation between variables.

When analysing BPSK over an AWGN channel, I consider it an independence test

between the X and Y signals. As I increased the noise level in Y, it began to dominate the signal, leading X and Y to tend to independence. At high variances, when the signals are nearer to independence, the distance correlation changes more with variance than empirical mutual information. Normalised negentropy exhibits a near linear relationship with mutual information, indicating that it is also effective in describing source dependence. However, more subtle features can be captured using the distance correlation metric.

Additionally, at lower variances, the normalised negentropy and distance correlation do not plateau like the empirical mutual information. This characteristic may allow for better discrimination among sources close to being independent.

Therefore, both negentropy and distance correlation can identify slight differences between sources near independence. Conversely, at high levels of dependency, distance correlation may serve as the more effective metric, enabling the development of fine-grained representations.

3.2.3 Closed form vs MINE mutual information computations

Figures 3.3 and 3.4 illustrate the empirical 'true' BPSK mutual information (depicted in red and calculated using Equation 3.3) alongside the MINE estimates (shown in black) averaged over ten runs, as a function of AWGN noise variance. These figures use two different MINE architectures to produce values for the joint and marginal distributions of the variables. These values are then used to calculate an estimate of MI using Equation 2.11. The estimates are optimised through gradient ascent to approximate the actual mutual information as closely as possible. In Figure 3.3, where the information-bottleneck style network is used (see Figure 3.2a), the MINE mutual information estimates almost overlay the empirical line of best fit across low- and high-noise regimes, showing tight tracking. In Figure 3.3, where the alternating linear layer and Leaky ReLU network is used (see Figure 3.2b), the estimates still follow the overall shape of the true MI but begin to undershoot and display greater scatter at very low-noise (high-MI) settings, reflecting slightly more bias and variance from the deeper architecture.

To describe in more depth architectures and their corresponding outputs, I will present Figure 3.3 which corresponds to the architecture which initially applies two linear layers, one to each random variable. The outputs of the linear layers are summed, and a ReLU

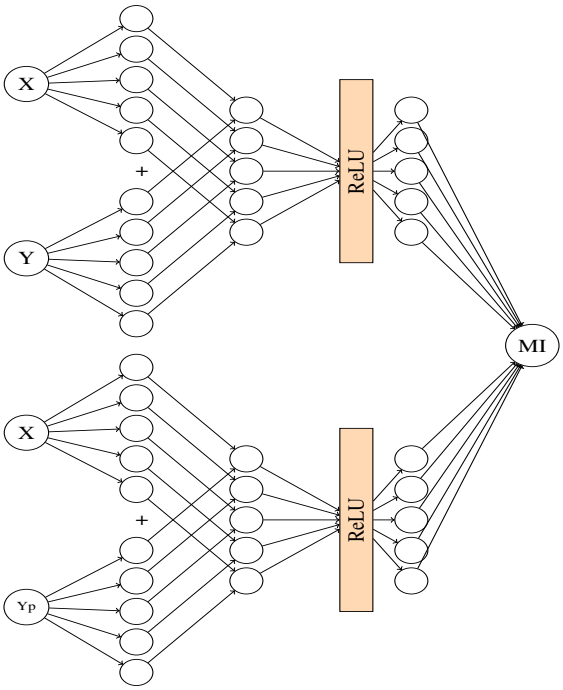
activation is applied. Finally, one more linear layer is applied. This MINE implementation was inspired by the information bottleneck outlined in [20], using a ReLU activation instead of ELU. The deep information bottleneck referenced above was based on the work in [142], which states that the deep variational information bottleneck objective is to '*outperform those that are trained with other forms of regularisation, in terms of generalisation performance and robustness to adversarial attack*'. In the case of 3.4, the network contains six repeats of a linear layer followed by a Leaky ReLU activation, with a final linear layer applied to conclude the architecture. In this instance, I used a deeper neural network architecture to identify whether more parameters could better maximise the approximate mutual information to the actual mutual information. The two architectures are exemplified by the diagrams in Figure 3.2.

The lines of best fit in Figures 3.3 and 3.4 illustrate the average of the final 50 mutual information approximations (black) alongside the empirical mutual information (red). In Figure 3.3, the two lines appear to be approximately the same. In Figure 3.4, the approximation slightly underestimates the empirical mutual information, although it is within one standard deviation of the mean value. As MINE maximises the approximate mutual information to converge on the actual mutual information, the empirical mutual information is a ceiling value that can be underestimated, epitomised by the Information Bottleneck method outperforming the deeper neural network, reinforcing the robustness results from [142]. These figures are used to exemplify the potential importance of the choice of MINE architecture and that mutual information can be used as a metric in a tractable manner.

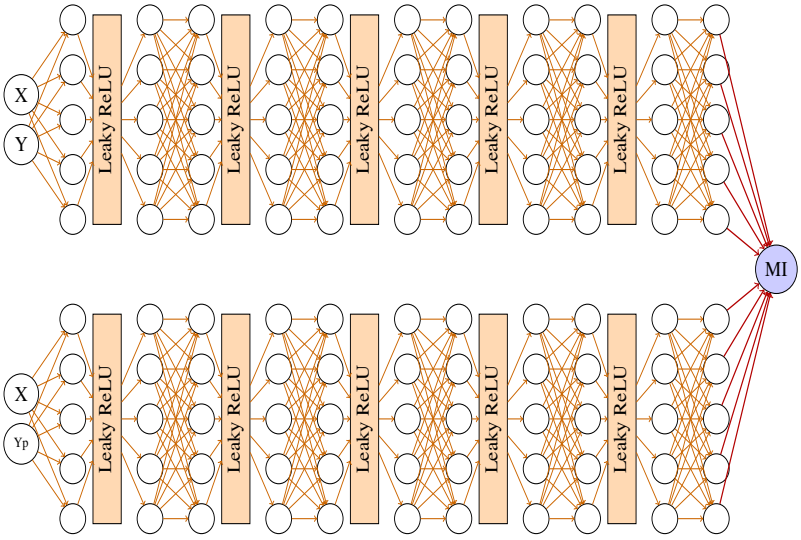
However, as MINE is parametric, it can cause issues when being integrated into source separation problems, as will be discussed later in the thesis.

3.3 Noise colour

In the previous definition of the problem, white Gaussian noise was added to an equally probable sampled binary signal to simulate random processes related to certain natural phenomena, as commonly used in information theory examples. Adding white noise may favour negentropy over distance correlation, as the former is a measure of Gaussianity



(a) Information bottleneck architecture



(b) Deep architecture

Figure 3.2: The upper MINE architecture utilises the information bottleneck, while the lower features an architecture with a deeper network. The upper and lower networks represent the calculations of marginal and joint distributions. When these calculations are combined using Equation 2.11 the final step, they approximate the lower bound of mutual information through gradient ascent. Y_p represents a random permutation of Y, permuted in the time dimension.

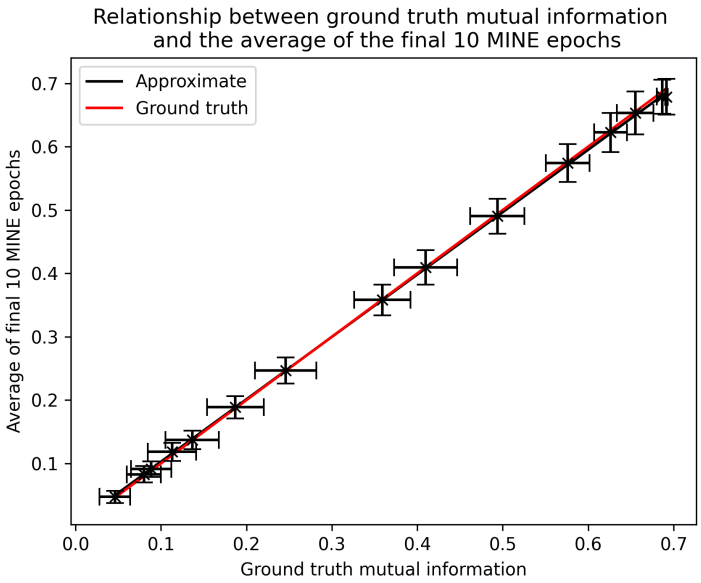


Figure 3.3: Comparison of the average MINE mutual information at the final epoch, calculated over 10 repeats, with the empirical mutual information for BPSK over an AWGN channel, using the first or, more explicitly, the information bottleneck architecture.

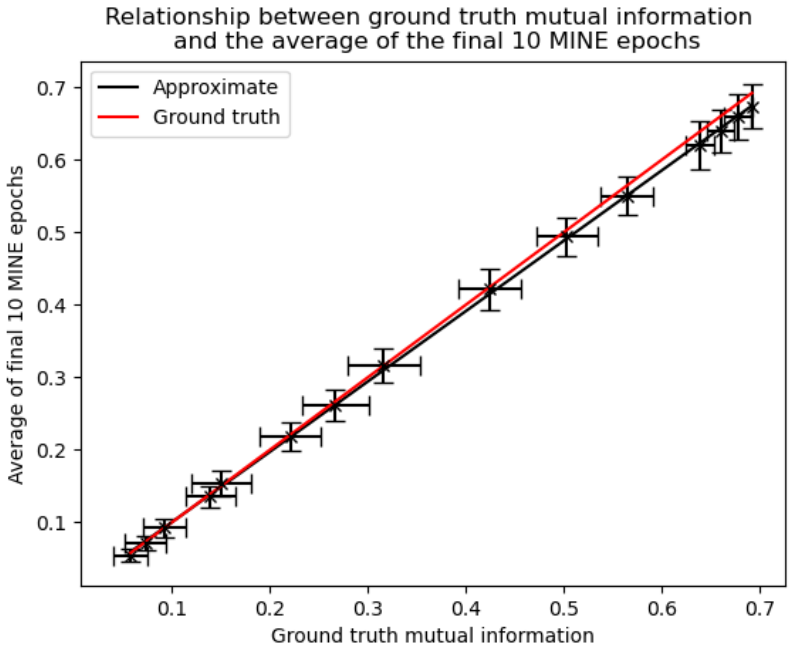


Figure 3.4: Comparison of the average final MINE epoch for ten repeats with the empirical mutual information for BPSK over an AWGN channel, specifically for the second or alternating linear layer and Leaky ReLU architecture.

and, therefore, benefits from the noise being the colour white. Moreover, some research has indicated that time series noise should be represented by a combination of white noise and power-law processes [141] rather than only white noise.

In this section, several other colours of noise were added to the binary signal X. In each case, the noise was standardised and then scaled by variance for comparison. The output signal ($Y = X + N$) is standardised to align with the central limit theorem, which requires matching variances to compare methods. Colour noise, otherwise known as power law noise, indicates a power law spectral density. Spectral density per unit bandwidth is proportional to $\frac{1}{f}^{\beta}$, where β is an integer representing a colour of noise and f is frequency. β is 0 for white noise, 1 for pink noise, 2 for Brownian noise and -1 for blue noise.

One method used to produce data exhibiting a power law spectrum, $S(f) \sim \frac{1}{f}^{\beta}$, is provided in Equation 3.6 [143]:

$$N(t) \sim \sum_{2\pi f} \sqrt{S(2\pi f)} cos(2\pi f t - \phi(2\pi f)) \ , \tag{3.6}$$

where ϕ is a random phase. The amplitude, N(t), is deterministic for each frequency and only the phase is randomised. This amplitude is the noise to be added to the signal.

The reader can find a wider variety of types of added noise in appendix A. In this section, I will compare two examples: Brownian noise, as random walks are applied in various fields, including fluid dynamics, and velvet noise, characterised by a sparse sequence of positive and negative impulses defined by the density of impulses. The higher densities approximate white noise, enabling the study of density with known behaviour at high values.

Brownian noise has an exponent, β, of 2. In this case, the *signalz.brownian_noise* function, from the signalz Python module, which generates Brownian noise (also known as red noise or random walk noise) was used. This module generates Brownian noise by integrating white Gaussian noise with a given standard deviation. In Figure 3.5, I used a leaky 10% integration to keep the noise within a reasonable range. The leaky integrator is the result of integrating a specific differential equation, which leaks some of the input over time, constraining the possible values of the random walk.

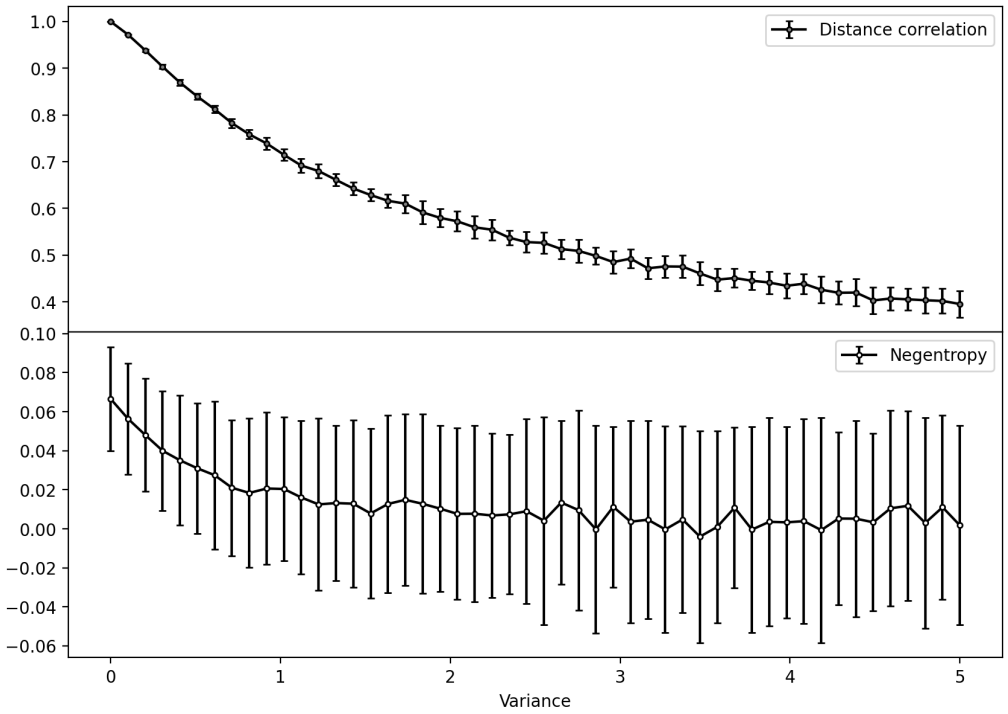


Figure 3.5: The relationship between a BPSK signal and the BPSK signal with added Brownian noise while varying the noise component's variance, as measured by distance correlation and negentropy. A leaky integration of 10% was employed to maintain the noise within a reasonable range.

Velvet noise is a sparse noise sequence [144]. I implemented velvet noise using the scipy.sparse.random module to generate both positive and negative impulses with a specific density. As I decreased the sparsity (or increased the density), the velvet noise, as anticipated, began to resemble white noise.

For the velvet case, I conducted two experiments. First, I generated velvet noise signals with varying densities, ranging from 0.05 to 1, across 50 equidistant points. I observed two competing effects based on the distance correlation and negentropy results: the signal shows a tendency toward white noise at higher densities, and the standard deviation of the signals changes with the density. Therefore, I will present the negentropy and distance correlation results, both with and without noise standardisation, alongside the variations in standard deviation with density.

The three results will help one identify the effects of density. At high-density values, the signals tend to resemble white noise. First, I will examine this behaviour with and without the inherent changes in the standard deviations of the velvet noise due to density,

allowing the effect of changing standard deviation to be observed while also isolating it when analysing how the probability distribution of the noise changes with varying densities. The variability in standard deviation impacts the test's objective to examine how the independence metrics change as the probability distribution of the noise approaches a Gaussian distribution (at high densities). It is important to note that the noisy signal, represented as Y=X+N, is standardised in all cases, though N is not.

Secondly, I created a velvet noise signal with a density of 0.1. Then I standardised this signal and scaled it by a factor between 0 and 5, which adjusted the standard deviation range while maintaining the same signal probability distribution. I applied the same methods to the blue, pink, and violet noise case tests (See Appendix A). This test, which varies the standard deviation, allows the noise component to dominate the signal Y, making it more independent of X as the noise becomes dominant.

Figure 3.6 illustrates how varying the density affects the noise without standardising it. Meanwhile, Figure 3.7 shows how the standard deviation of the velvet noise changes with density. The noise standardisation results can be seen in Figure 3.8. There is a notable connection between the change in the distance correlation and the variance of the noise, both with density, as seen in Figures 3.6 and 3.7, respectively. In Figure 3.6, the noise has not been standardised. These results emphasise the importance of the standardisation step, ensuring that no competing elements exist in the experiments. In Figure 3.8, one observes that both metrics decrease and eventually plateau when examining how negentropy and distance correlation change with increasing density. However, the distance correlation decreases more slowly. As intimated by Equation 3.5, the negentropy approaches zero because, with higher density, the terms tend to cancel each other out. On the other hand, distance correlation reveals a relationship between the random variables. This suggests that it may serve as a more effective metric for source separation. Unlike negentropy, which is calculated only on the BPSK signal with added velvet noise, distance correlation is computed using the original BPSK signal and the noisy signal. It is important to note that this analysis does not consider the effects of whitening in the source separation pipeline, which will be discussed further in Chapter 4.

For the results of the second experiment, where the velvet noise had a density of 0.1 and was standardised and scaled by a factor of between 0 and 5, I direct the reader to

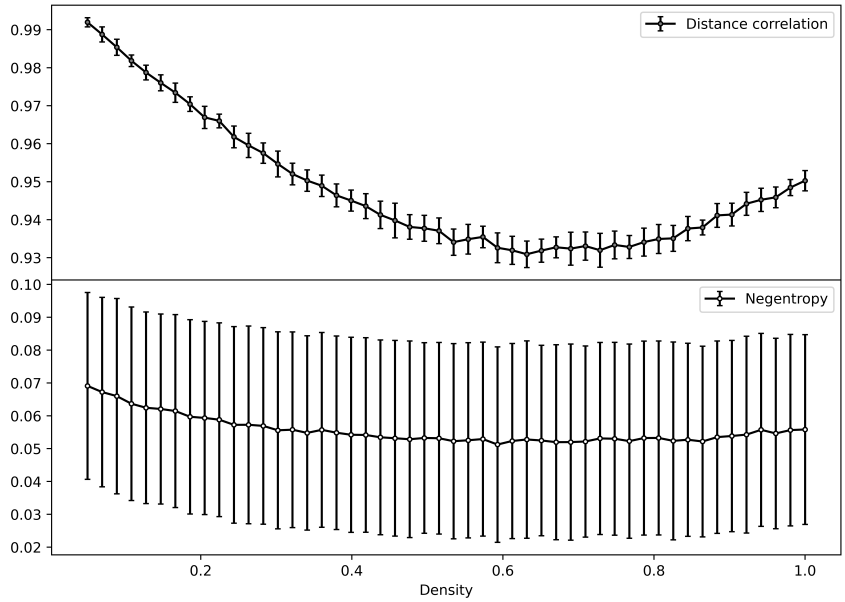


Figure 3.6: The relationship between a BPSK signal and the BPSK signal with added velvet noise while varying the noise component's variance, as measured by distance correlation and negentropy. The velvet noise has **not** been standardised in this case.

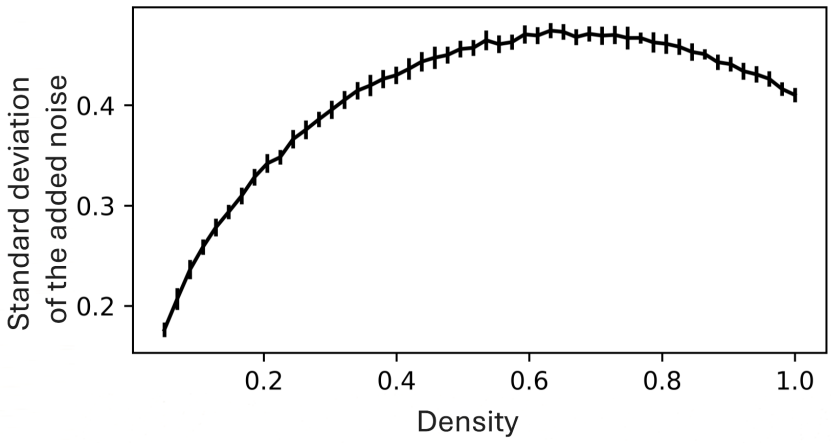


Figure 3.7: The standard deviation for a given density related to the velvet noise in Figure 3.6.

Figure 3.9.

As I am focused on the cases where sources are independent, in the case of additive noise, the equivalent would be to add noise with a high variance such that it dominates the outputted signal. For each of the colours of noise, the distance correlation has a greater gradient for higher variances when compared to negentropy, which tends to plateau. For a more in-depth summary of the impact of different colours of noise, please see Table 3.1.

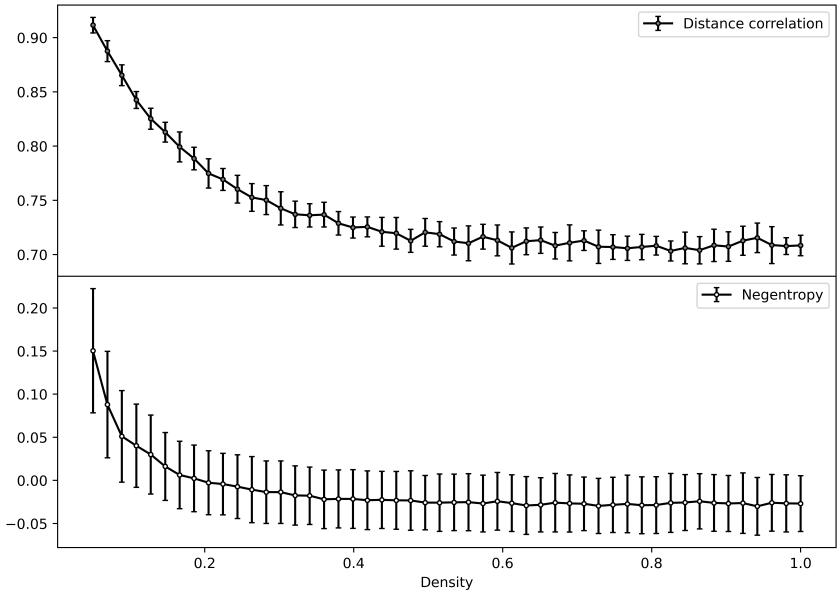


Figure 3.8: The relationship between a BPSK signal and the BPSK signal with added velvet noise while varying the noise component's variance, as measured by distance correlation and negentropy. The velvet noise has been standardised in this case.

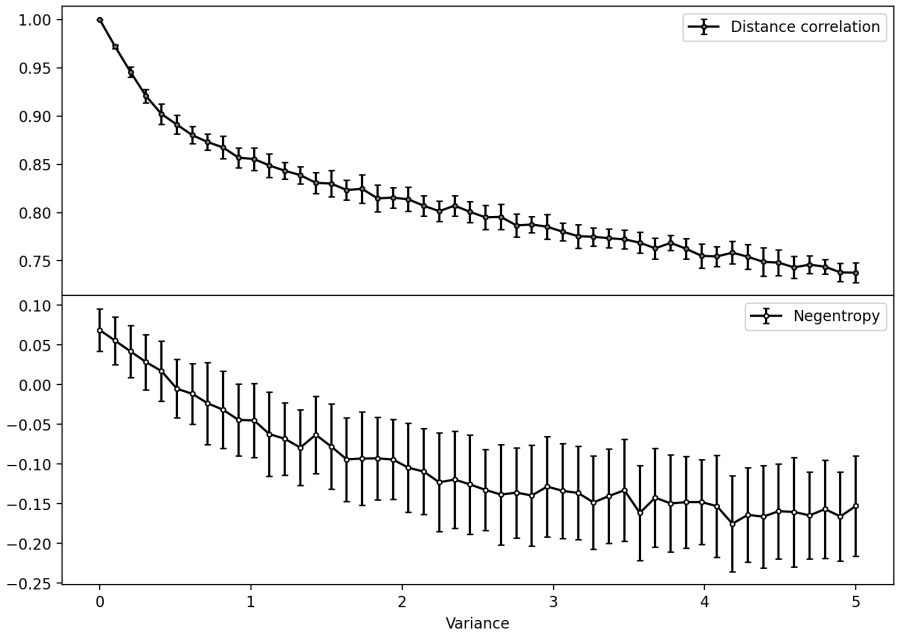


Figure 3.9: Distance correlation and negentropy for a binary signal transmitted through a velvet noise signal, where the noise has been standardised and scaled by a user-defined variance within a specific range. This approach allows the observation of how these metrics change when the density of the velvet noise signal remains constant, at 0.1, while the standard deviation varies.

The lack of plateau opens up a fascinating area for further research, as the larger gradient for distance correlation at high variances could lead to a more fine-grained comparison

Noise type	PSD $\propto 1/f^k$	Temporal Memory	Impact on Distance correlation	Impact on negentropy
White	$k = 0$	None	Smooth, low-bias decline as variance increases	Smooth decline to a plateau with increased variance; large associated uncertainty
Pink	$k = 1$	Moderate long-range memory	Similar decline to the DistCorr white noise case	Similar to white noise case negentropy
Brownian (Red)	$k = 2$	Strong long-range memory	Similar decline to the DistCorr white noise case	Similar to white noise case negentropy but with more scatter and uncertainty
Blue	$k = -1$	Short, anti-correlated	Similar decline to the DistCorr white noise case	Similar to white noise case negentropy
Violet	$k = -2$	Very short, anti-correlated	Similar decline to the DistCorr white noise case	Similar to white noise case negentropy
Velvet (white)	$\approx$ flat	Irregular impulses	A pronounced drop in DistCorr that levels off at a non-zero value as density increases	Density increases produce greater Gaussianity, causing a sharp decline in negentropy that then plateaus near zero

Table 3.1: Impact of coloured noise on distance correlation and negentropy. Note that some researchers consider Brownian noise representative of the noise in geodetic time series. However, care must be taken to distinguish genuine low-frequency geophysical signals, such as the linear drift from tectonic plate motion, from noise, to avoid misclassifying these trends as noise, particularly Brownian instead of pink noise.

between random variables close to independence. Such insights could significantly enhance the applicability of distance correlation in gradient descent optimisation when used within loss functions for BSS.

Note that in this comparison, I am not focused on the metric values themselves, as the negentropy for example could be normalised in order to increase its range for optimisation.

3.4 No additive noise

In a separate experiment, I generated 40 input channels labelled as $Signal_1$ to $Signal_{40}$, each containing equiprobable values of -1 or 1. I calculated the distance correlation and negentropy between $Signal_1$ and the average of the signals from indexes 1 to 40, varying the number of signals included in this average. I computed averages with up to 10,000 signals. However, the average of the first 40 signals accounted for most of the change in distance correlation and negentropy before they plateaued. This thesis presents findings on the average of up to 40 BPSK signals, which constitutes most of change as additional BPSK signals are added to the average signal.

Additionally, the second signal tends to approximate a Gaussian distribution. The outcomes are based on 25 repetitions, each involving a different set of 40 binary signals, each with a length of 1,000. It is important to note that the initial and the average signals were standardised.

Due to the nature of the mean signal and the law of large numbers, an increase in the number of signals will lead to the sum at each time point approaching zero, which reduces the variance of the signal. This creates competing effects: on one hand, the distribution of the mean of a large number of signals tends to resemble a Gaussian, and on the other hand, the variance moves towards zero. This effect, analogous to the previous section, emphasises the importance of standardising the second signal.

In Figure 3.10, it is evident that the distance correlation has a value of 1.0, as $Signal_1$ is inherently dependent on itself. In both scenarios, the metrics decrease rapidly as the number of signals in the average increases, eventually plateauing. The initial decline is more pronounced in the case of negentropy and gentler for a higher number of signals in the distance correlation case.

Tying this to the Central Limit Theorem, as one sums k independent sources, their normalised sum tends to approach a Gaussian distribution, even if the underlying sources are non-Gaussian. In BSS, ICA algorithms can use non-Gaussianity to measure independence, in the form of kurtosis or negentropy in the case of FastICA. The reduction in the non-Gaussianity, in terms of negentropy, of the mixtures due to the Central Limit Theorem scales as $1/\sqrt{k}$. This limits the accuracy of source separation and explains the plateau observed in Figure 3.10. However, distance correlation captures pairwise statistical de-

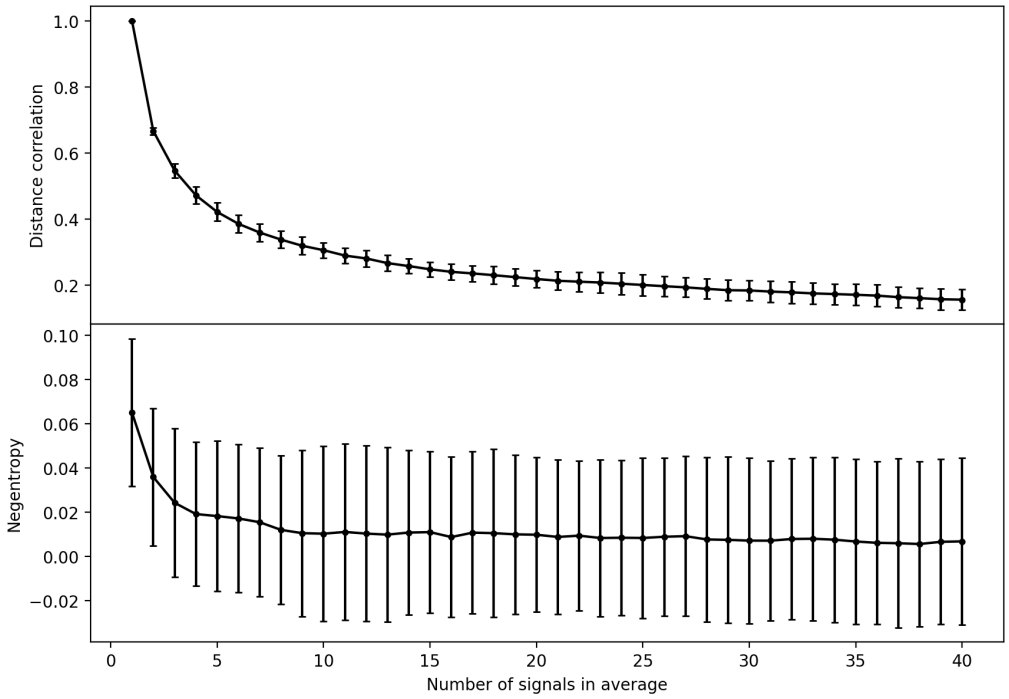


Figure 3.10: Distance correlation or negentropy between an initial signal and the average of all signals up to the number specified by the x-axis.

pendencies, both linear and non-linear, unlike negentropy, which primarily measures the non-Gaussianity of an individual signal. This includes small residual dependencies when the number of signals in the average is high. Therefore, it remains effective even when the Central Limit Theorem has driven a marginal distribution toward normality, supporting its superior performance.

In the case of distance correlation, its value decreases (though its uncertainty increases) until it reaches a plateau as the sources become more independent. This indicates a point where the binary input source becomes independent of the mean of many BPSK signals, resulting in similar but non-zero distance correlation values. Additionally, the absolute correlation coefficient between the negentropy of the average signals and the distance correlation between one of the signals and the average signal is over 0.8, demonstrating a strong linear relationship between these metrics in this example.

3.5 Conclusion

In this chapter, I employed non-parametric tests to thoroughly examine the suitability of distance correlation as an independence metric.

I first transmitted a binary signal through a Gaussian noise channel characterised by a mean of zero and a known variance. I examined the independence of the input and output signals using distance correlation and negentropy. In the case of negentropy, I calculated it solely based on the output signal. These values were then compared to empirical mutual information, which was determined using Monte Carlo integration.

Both the distance correlation and the empirical mutual information decreased and eventually plateaued as the variance increased. Similarly, the normalised negentropy also decreased and then plateaued with increasing variance. In theory, the minimum negentropy value should be zero, as the $Signal_1$ component becomes dominated by noise. The divergence is an artifact of the calculation technique.

As noise becomes more prevalent, the distance correlation decreases, indicating increased independence. However, it does not tend toward zero, nor does the Pearson correlation coefficient. The Pearson correlation coefficient indicates a linear relationship between the signals, confirming that a linear relationship exists and that the distance correlation should not reach a minimum value. The fact that the distance correlation does not tend to zero implies that it could serve as a better loss function for source separation in a machine-learning context. It provides a more nuanced description of the relationships between the data, encompassing both linear and non-linear dynamics, which allows for more effective minimisation of these relationships.

I compared negentropy and distance correlation in scenarios where different types of noise, including Brownian (pink, blue, violet) and velvet noise, were added to a binary signal. The results showed that the distance correlation exhibited less uncertainty and plateaued more gradually. Distance correlation may be more beneficial for gradient descent updates due to increased variation in values. In the case of velvet noise, when the added noise was normalised, the uncertainty associated with negentropy was higher as a proportion of the signal (i.e. the actual uncertainty value for negentropy may be lower, but the metrics are not on the same scale). Additionally, negentropy plateaued earlier than distance correlation but showed a similar reduction before stabilising.

Finally, I examined the independence of a binary signal with a standardised average of other binary signals. As the number of signals included in this average increased, the result became more similar to white noise. In this case, I found that the distance correlation exhibited a greater variation in values as the number of signals increased compared to negentropy and had less associated uncertainty.

This chapter concludes that distance correlation offers a more comprehensive representation of data by considering both linear and non-linear relationships. In contrast, negentropy is minimised for BPSK over an AWGN channel, approaching zero when the random variables have a linear relationship. Furthermore, distance correlation demonstrates a steeper gradient when dealing with higher variances in examples of coloured noise. In a proxy scenario for independence, the steeper gradient indicates that a distance correlation metric can more effectively learn independent sources using gradient descent optimisations.

CHAPTER 4

Whitening

After a brief introduction of the most commonly used whitening methods in Section 4.1, in Section 4.2, I study, empirically and theoretically, the effect of whitening on the distance correlation between the whitened signals. In Section 4.3, I give some recommendations on the use of whitening within the BSS pipeline, and provide a conclusion for this chapter in Section 4.4.

4.1 Whitening methods

Several BSS algorithms employ decorrelation to facilitate the estimation of sources by reducing the number of free parameters. One method commonly used to reduce the number of parameters is to linearly transform the data into a form that has an identity covariance matrix, effectively eliminating linear relationships present in the original data. This *whitening* transformation has the form of a matrix that, when multiplied by the data, in this case, an n-dimensional vector of random variables $\boldsymbol{x}$, produces a result that is decorrelated.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ with mean $E(\boldsymbol{x}) = (\mu_1, \mu_2, \ldots, \mu_n)^T$, and let the covariance matrix of $\boldsymbol{x}$ be $\boldsymbol{\Sigma}$. To apply a whitening transformation to $\boldsymbol{x}$, $\boldsymbol{\Sigma}$ must be invertible, and thus,

the covariance must be positive definite. Then, by applying a whitening matrix $\boldsymbol{W}$, one can transform $\boldsymbol{x}$ into a new random vector:

$$\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^T = \boldsymbol{W}\boldsymbol{x}, \tag{4.1}$$

with the covariance matrix of $\boldsymbol{z}$ being the identity matrix. Simple computations, see in [6], show that $\boldsymbol{W}$ is a whitening matrix if and only if:

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}. \tag{4.2}$$

Equation 4.2 contains a rotational degree of freedom. Therefore, $\boldsymbol{W}$ is not uniquely determined. This can be seen from the *polar decomposition* of $\boldsymbol{W}$,

$$\boldsymbol{W} = \mathbf{Q_1}\boldsymbol{\Sigma}^{-\frac{1}{2}}, \tag{4.3}$$

where $\boldsymbol{Q}_1$ is an orthogonal matrix. Indeed, inputting Equation 4.3 into Equation 4.2, one can check that $\boldsymbol{W}$ satisfies the whitening condition irrespective of the chosen orthogonal rotation matrix, because $\boldsymbol{Q}_1^T\boldsymbol{Q}_1 = \mathbf{I}$. Interpreting 4.3 geometrically, the whitening transformation first rescales the input vector by $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ and then rotates the output by $\boldsymbol{Q}_1$.

A second decomposition of $\boldsymbol{W}$, used in [6] to characterise the various whitening methods they study, is:

$$\boldsymbol{W} \;=\; \boldsymbol{Q_2}\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}} \;=\; \boldsymbol{Q_2}\boldsymbol{G}\boldsymbol{\Theta}^{-\frac{1}{2}}\boldsymbol{G}^T\boldsymbol{V}^{-\frac{1}{2}}, \tag{4.4}$$

where $\boldsymbol{Q_2}$ is again an orthogonal rotation matrix; $\boldsymbol{V}$ and $\boldsymbol{P}$ are defined through the decomposition $\boldsymbol{\Sigma} = \boldsymbol{V}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{V}^{\frac{1}{2}}$ of the covariance matrix into the correlation matrix $\boldsymbol{P}$ and the diagonal variance matrix $\boldsymbol{V}$; and $\boldsymbol{G}$ and $\boldsymbol{\Theta}$ are defined through the eigendecomposition of the correlation matrix $\boldsymbol{P} = \boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{G}^T$. The eigendecomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ will also be used.

Before I introduce in more detail the five natural whitening procedures studied in this Chapter, chosen for comparison to [6], I will introduce the cross-covariance and cross-correlation matrices between the whitened vector, $\boldsymbol{z}$ and the original random vector, $\boldsymbol{x}$.

The cross-covariance matrix between $\boldsymbol{z}$ and $\boldsymbol{x}$ is:

$$\boldsymbol{\phi} = (\phi_{ij}) = cov(\boldsymbol{z},\boldsymbol{x}) = cov(\boldsymbol{W}\boldsymbol{x},\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{\Sigma} = \boldsymbol{Q}_1\boldsymbol{\Sigma}^{\frac{1}{2}}. \tag{4.5}$$

Similarly, the cross-correlation matrix is:

$$\boldsymbol{\psi} = (\boldsymbol{\psi}_{ij}) = cor(\boldsymbol{z},\boldsymbol{x}) = \boldsymbol{\phi}\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{Q}_2\boldsymbol{P}^{\frac{1}{2}}. \tag{4.6}$$

Note that the derivation in Equation 4.6 has been shortened for brevity, and I have omitted the discussion regarding the connections between the rotations in the two decompositions of $\boldsymbol{W}$ in Equations 4.3 and 4.4 ($\boldsymbol{Q}_1 = (\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}})\boldsymbol{Q}_2$, see [6] for more details).

The rotational degrees of freedom of $\mathbf{W}$, represented either by $\mathbf{Q}_1$ or by $\mathbf{Q}_2$, can be used to optimise measures of cross-covariance and cross-correlation between the original and whitened vector. Following [6]. I will verify the previously established relationships between four such statistical measures, which I will introduce in the following paragraph, and five whitening transforms. Then, I will extend these results away from whitening for decorrelation and move towards whitening for independence.

4.1.1 Common whitening methods

In [6], five natural whitening methods were compared in terms of optimisation of various measures. Here is a summary of these five popular and commonly used whitening methods, along with the whitening matrix $\boldsymbol{W}$ from Equation 4.1 in brackets:

1. ZCA [145] [146] [6] ($\boldsymbol{W}_{ZCA} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$): The ZCA whitening transformation is the unique transformation that minimises the squared distance between the original and whitened vectors. Simple but non-trivial computations in [6] showed that this squared distance is minimised when the trace of the cross-covariance matrix $tr(\boldsymbol{\phi})$ of Equation 4.5 is maximised. This happens when the orthogonal rotation matrix $\boldsymbol{Q}_1$ in the decomposition of Equation 4.3 equals the identity matrix. Therefore, the whitening transformation that minimises the squared distance between the original and whitened vectors is the ZCA transformation.

2. PCA [147] [25] [6] ($\boldsymbol{W}_{PCA} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}^T$, where $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ is the eigendecomposition

of $\mathbf{\Sigma}$): The PCA whitening transformation involves rotating the vectors using the matrix $\boldsymbol{U}^T$ composed of the eigenvectors of the covariance matrix, much like PCA. In this setup, the directions associated with the highest eigenvalues represent the directions that account for the most variation. Although this rotation results in orthogonal components, these components typically do not have unit variances. To address this issue, the components are scaled by the square root of the eigenvalues, $\mathbf{\Lambda}^{-\frac{1}{2}}$, producing whitened data.

In [6], the optimality properties of the PCA whitening are studied through the row sums of the squared cross-covariances: $\boldsymbol{\phi}_i = \sum_{j=1}^{d} \text{cov}(z_i, x_j)^2$ which correspond to the diagonal elements of the matrix $\boldsymbol{\phi}\boldsymbol{\phi}^T$. PCA whitening maximises the $max(diag(\boldsymbol{\phi_m}\boldsymbol{\phi_m^T}))$, and even though it is not unique in that respect, following [6] we use it as an evaluation metric.

3. ZCA-Cor [6] ($\boldsymbol{W}_{ZCA-Cor} = \boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}$): The ZCA-Cor whitening transform minimises the squared distance between the original standardised and whitened vectors. Similar to ZCA whitening, [6] shows that only one component of this squared distance varies during the whitening process: the trace of the cross-correlation: $tr(\boldsymbol{\psi})$. Computations similar to the ZCA whitening show that the squared distance is minimised when $tr(\boldsymbol{\psi})$ is maximised, and that happens when $\boldsymbol{Q}_2$ in Equation 4.4 equals the identity matrix, resulting in the ZCA-Cor whitening transform.

4. PCA-Cor [6] ($\boldsymbol{W}_{ZCA-Cor} = \mathbf{\Theta}^{-\frac{1}{2}}\boldsymbol{G}^T\boldsymbol{V}^{-\frac{1}{2}}$, where $\boldsymbol{P} = \boldsymbol{G}\mathbf{\Theta}\boldsymbol{G}^T$ is the eigendecomposition of the correlation matrix): Like PCA whitening, PCA-Cor additionally ensures scale invariance by optimising for cross-correlations rather than total correlations. Analogously, this transform maximises $max(diag(\boldsymbol{\psi_x}\boldsymbol{\psi_x^T}))$.

5. Cholesky [6] [148] ($\boldsymbol{W}_{Chol} = \boldsymbol{L}^T$): This unique procedure gives matrices $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ that are both lower triangular and have positive diagonal elements. It is important to note that this is the only whitening method that does not optimise a metric based on cross-correlation or cross-covariance.

Note that in the above, the input and whitened vectors are all centred.

As mentioned earlier, whitening transformations assist with source separation by eliminating linear relationships within the data. However, as we just saw, the conditions for

whitening do not yield unique linear transforms $\boldsymbol{W}$. In the machine learning literature, whitening transforms are often applied without a clear statistical justification. Instead, it seems that factors such as ease of implementation and consistency with previous studies are the driving forces behind the choice of whitening transform. Here, I aim at providing a deeper rationale for my choice of whitening transforms, as well as their positioning within the source separation pipeline.

4.2 Whitening and distance correlation

I experimented with a small set of synthetic data, applying the five whitening methods described in Section 4.1, computing the same metrics on the original and the whitened signals as in [6], as well as distance correlations. My aim is to obtain insights into how the distance correlations between signals are affected by the various whitening methods and, where possible, to support my empirical observations with theoretical justifications.

4.2.1 Synthetic data and experimental design

I utilised the synthetic data from the Scikit-Learn FastICA source separation example. This choice ensures compatibility with existing literature, as this synthetic example is the sole validation test used in [149], which introduced MINE for source separation. The data comprises three sources: a sine wave ($\boldsymbol{s_1}$), a sawtooth wave ($\boldsymbol{s_2}$), and a square wave ($\boldsymbol{s_3}$). I combined these sources to create three mixtures ($\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}$), via the mixing matrix:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0.5 & 2 & 1 \\ 1.5 & 1 & 2 \end{bmatrix}. \tag{4.7}$$

Three sources will be extracted from these three mixtures using blind source separation in Chapter 5. In this chapter, I investigate the effect of whitening on the mixtures and the underlying sources, which I assume to be known. In future, this problem will be referred to as the synthetic problem.

I applied the five commonly used whitening transformations of Section 4.1 to the three synthetic mixtures. Following this, I calculated $tr(\boldsymbol{\phi_x})$, maximised by ZCA whitening,

$tr(\boldsymbol{\psi_x})$, maximised by ZCA-Cor whitening, and $max(diag(\boldsymbol{\phi_x}\boldsymbol{\phi_x^T}))$ and $max(diag(\boldsymbol{\psi_x}\boldsymbol{\psi_x^T}))$, maximised by the PCA and PCA-Cor whitening transformations, respectively.

I extended my analysis to include the distance correlation between pairs of initial and whitened mixtures, and the pairwise correlations and distance correlation between the whitened mixtures and the underlying sources as a small test to identify if any form of whitening pushes inputs towards independence.

In my comparisons, I allowed permutations between inputs and outputs. First I select the highest distance correlation between all pairs of whitened vectors and sources, next I select the highest distance correlation from the remaining pairs, and finally the last result comes of the only remaining pair. This step produces a permutation-invariant comparison. It is important to note that while permutations were allowed, only the PCA whitening versions exhibited their highest values off-diagonal, and those values were anti-diagonal. Consequently, I presented the results keeping a fixed order where $\boldsymbol{s_1}$ is always the sine wave, $\boldsymbol{s_2}$ the square wave, and $\boldsymbol{s_3}$ the sawtooth wave, which enhances interpretability.

4.2.2 Experimental results

The experimental results are presented in Tables 4.1, Table 4.2. Table 4.1 records values between whitened mixtures and original mixtures or sources, while Table 4.2 records values between whitened sources and original sources. The layout of both tables follows the format in [6], meaning that the maximum variable of the experiment, with regard to choice of whitening technique, should be traced row-wise.

The first six rows of Table 4.1 show that Cholesky whitening achieves a maximum value of 1 for one pair of input and whitened vectors for both correlation and distance correlation. This is not a coincidence as in Cholesky whitening the final whitened variable is always a scaled version of the final original variable. Table 4.1 illustrates this effect in the values of cor($\boldsymbol{x_3}, \boldsymbol{z_3}$) and DistCorr($\boldsymbol{x_3}, \boldsymbol{z_3}$).

The first six rows of Table 4.2 show the same behaviour on whitened original sources rather than mixtures. It is important to note that before any processing or computation the sources had been standardised. I emphasise this point as standardisation removes the distinction between PCA and PCA-cor as well as between ZCA and ZCA-cor, and the correlation and covariance matrices also become equivalent. This can been seen in Table 4.2,

	ZCA	PCA	Cholesky	ZCA-cor	PCA-cor
cor($\boldsymbol{z_1}, \boldsymbol{x_1}$)	0.564	0.992	0.150	0.679	0.993
cor($\boldsymbol{z_2}, \boldsymbol{x_2}$)	0.839	0.410	0.637	0.822	0.381
cor($\boldsymbol{z_3}, \boldsymbol{x_3}$)	0.840	0.034	1.000	0.762	0.065
DistCorr($\boldsymbol{z_1}, \boldsymbol{x_1}$)	0.532	0.988	0.191	0.633	0.989
DistCorr($\boldsymbol{z_2}, \boldsymbol{x_2}$)	0.863	0.590	0.769	0.848	0.559
DistCorr($\boldsymbol{z_3}, \boldsymbol{x_3}$)	0.799	0.137	1.000	0.721	0.144
cor($\boldsymbol{z_1}, \boldsymbol{s_1}$)	0.927	0.755	0.776	0.950	0.729
cor($\boldsymbol{z_2}, \boldsymbol{s_2}$)	0.933	0.787	0.920	0.914	0.763
cor($\boldsymbol{z_3}, \boldsymbol{s_3}$)	0.945	0.595	0.714	0.975	0.565
DistCorr($\boldsymbol{z_1}, \boldsymbol{s_1}$)	0.922	0.743	0.765	0.944	0.716
DistCorr($\boldsymbol{z_2}, \boldsymbol{s_2}$)	0.951	0.797	0.938	0.934	0.769
DistCorr($\boldsymbol{z_3}, \boldsymbol{s_3}$)	0.936	0.548	0.677	0.970	0.516
$tr(\boldsymbol{\phi_x})$	**5.115**	2.556	4.369	5.069	2.408
$tr(\boldsymbol{\psi_x})$	2.243	1.369	1.787	**2.263**	1.309
$max(diag(\boldsymbol{\phi_x}\boldsymbol{\phi_x^T}))$	7.115	**13.818**	12.930	5.606	13.797
$max(diag(\boldsymbol{\psi_x}\boldsymbol{\psi_x^T}))$	1.296	2.745	2.519	1.000	**2.749**

Table 4.1: Rows 1-3 display the pairwise correlations between the whitened and original mixtures, and Rows 4-6 the corresponding distance correlations. Rows 7-9 show the pairwise correlations between the whitened mixtures and the underlying sources, and Rows 10-12 the corresponding distance correlations. In rows 13 and 14, I provide the traces of the cross-covariance and cross-correlation matrices between the original and whitened mixtures. As described in [6], these traces are maximised by the ZCA and ZCA-Cor whitening transforms, respectively. In rows 15 and 16, I give the maximum values of the diagonal of the row sum of the squared cross-covariances and cross-correlations. These diagonals are maximised by the PCA and PCA-Cor whitening transformations, respectively.

	ZCA	PCA	Cholesky	ZCA-cor	PCA-cor
cor($\boldsymbol{W}(\boldsymbol{s_1}),\boldsymbol{s_1}$)	0.999	0.590	0.997	0.999	0.590
cor($\boldsymbol{W}(\boldsymbol{s_2}),\boldsymbol{s_2}$)	0.999	0.039	0.998	0.999	0.039
cor($\boldsymbol{W}(\boldsymbol{s_3}),\boldsymbol{s_3}$)	0.999	0.381	1.000	0.999	0.381
DistCorr($\boldsymbol{W}(\boldsymbol{s_1}),\boldsymbol{s_1}$)	0.999	0.555	0.996	0.999	0.555
DistCorr($\boldsymbol{W}(\boldsymbol{s_2}),\boldsymbol{s_2}$)	0.999	0.072	0.998	0.999	0.072
DistCorr($\boldsymbol{W}(\boldsymbol{s_3}),\boldsymbol{s_3}$)	0.999	0.342	1.000	0.999	0.342
$tr(\boldsymbol{\phi_S})$	**2.999**	0.170	2.996	**2.999**	0.170
$tr(\boldsymbol{\psi_S})$	**2.997**	0.170	2.994	**2.997**	0.170
$max(diag(\boldsymbol{\phi_S}\boldsymbol{\phi_S^T}))$	1.001	**1.118**	1.006	1.001	**1.118**
$max(diag(\boldsymbol{\psi_S}\boldsymbol{\psi_S^T}))$	1.000	**1.117**	1.005	1.000	**1.117**

Table 4.2: Rows 1-3 of this table display the pairwise correlations between the corresponding elements of the whitened and original sources. Rows 4-6 illustrate the distance correlations between the whitened and the original sources. The whitened sources $\boldsymbol{s_i}$ are denoted $\boldsymbol{W}(\boldsymbol{s_i})$ rather than $\boldsymbol{z_i}$ to avoid conflict with the notation of Table 4.1, where $\boldsymbol{z_i}$ denotes whitened mixture. In rows 7 and 8, I provide the traces of the cross-covariance and cross-correlation matrices between the original and whitened source random vectors. As discussed in [6], these traces are maximised by the ZCA and ZCA-Cor whitening transformations, respectively. In rows 9 and 10, I present the maximum values of the diagonal of the row sum of the squared cross-covariances and cross-correlations between the original and whitened sources. These are maximised by the PCA and PCA-Cor whitening transformations, respectively.

where a small discrepancy between the traces of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ arise from computational approximations and rounding errors in the variances of the whitened sources, which do not equal one.

The last four rows of Tables 4.1 and 4.2 show the maximality of the values of $tr(\boldsymbol{\phi})$ and $tr(\boldsymbol{\psi})$ for ZCA and ZCA-cor whitening, respectively, and similarly, the maximality of the values of $diag(\boldsymbol{\phi}\boldsymbol{\phi}^T)$ and $diag(\boldsymbol{\psi}\boldsymbol{\psi}^T)$ for PCA and PCA-cor, respectively. That is, as expected, the highest values of these statistics were achieved by the whitening transformation that theoretically maximises them. Indeed, upon examining the last four rows of the two Tables, the reader will notice that the maximum results highlighted in bold correspond to the transform that theoretically maximises them. Cholesky whitening does not optimise any of the four metrics, as it was not defined to do so.

In Table 4.2, the maximum values of $tr(\boldsymbol{\phi})$ and $tr(\boldsymbol{\psi})$ nearly reach their theoretical maximum of three, that is, the dimension of the covariance and correlation matrix. Indeed, the covariance and correlation matrices will be approximately the identity matrix

when working with standardised data and inputs that are not highly correlated. Consequently, the ZCA, ZCA-Cor and the Cholesky whitening transforms will all be close to the identity matrix. Table 4.2 verifies that these three whitening methods nearly maximise the shared information between standardised inputs and whitened outputs, leading to high correlations, distance correlations, and high values of $tr(\boldsymbol{\phi})$ and $tr(\boldsymbol{\psi})$.

The top six rows of Table 4.2 show that the correlations and distance correlations between sources and whitened sources are markedly lower for the PCA methods. This is because its input is an individual underlying source; therefore, compressing them into as few whitened vectors as possible, as the PCA transform attempts to do, leads to a loss of information about the relationship between the inputs and their corresponding whitened outputs.

Finally, from the Rows 4-6 of Table 4.1, one can see that ZCA-cor achieves the highest average distance correlation across the three pairs of input and whitened mixtures. However, it does not yield the maximum for any individual pair. Additionally, from Rows 10-12, one can see that ZCA-cor has the highest average and the best two out of three pairwise distance correlations between the whitened mixtures and the underlying sources. Thus, the ZCA and ZCA-cor whitening transforms appear to be effective for source separation, removing linear relationships from the whitened vectors while maximising either covariance or correlation, thus preserving the content of the original input for a distance correlation or negentropy loss optimisation, which will aid in removing remaining non-linear relationships. It is also noted that whitening can be harmful, as evidenced by the PCA transforms, which in Rows 10-12 of Table 4.1 show an average distance correlation between the whitened mixtures and sources of 0.712 for PCA, and 0.686 for PCA-cor. Both of these values are lower than the average distance correlation between the unwhitened mixtures and sources, which I have separately computed to be 0.733.

As a summary, to select the appropriate whitening transform, consider the following criteria. If you aim to preserve the spatial or topological structure, ZCA is preferable. Conversely, if you require the maximal integration with the original features or a specific ordering of components, consider using PCA or PCA-Cor.

Techniques like FastICA typically focus on removing second-order structure before assessing higher-order, non-linear dependencies. The choice of whitening affects both

the convergence speed and the solution quality. For ICA techniques that emphasise strong non-Gaussian signals, such as FastICA, PCA whitening is a suitable option. In contrast, for downstream metrics like distance correlation and mutual information, both of which focus on capturing pairwise dependencies, whitening transformations that minimise geometric distortion, such as ZCA-Cor or ZCA, are ideally suited for end-to-end machine learning. These transformations enhance the influence of pairwise metrics on the learning process. Similarly, representation learning, including techniques like scattering in self-supervised representation learning, benefits from minimal distortion of the representation, which is best achieved through ZCA-based whitening.

In what follows, I often chose the Cholesky method for whitening because it produced the cleanest gradient updates in my PyTorch neural network architecture during my early research. When it became easier to integrate various forms of whitening, I adjusted the whitening method to ZCA. However, unless otherwise specified, Cholesky whitening was used.

4.2.3 Whitening and independence

From the Rows 7-12 of Table 4.1, it can be observed that the whitening method maximising the correlation of between a whitened mixture and a source, also maximises the distance correlation. While the choice of whitening transform can lead to smaller or greater distance correlations between the whitened mixtures and the original sources. Here, I discuss why it is challenging to select a whitening transformation that maximises the dependence between the whitened mixtures and the sources by maximising the corresponding distance correlations.

In [6], the ZCA whitening method was shown to decorrelate the input while making the whitened *mean centred* random vector $\boldsymbol{z}$ as similar as possible to the original *mean centred* random vector $\boldsymbol{x}$. Their approach was based on the work of [150], which focused on minimising the total squared distance between the original and whitened variables. Therefore, the objective there was to minimise $E(\mathbf{z}-\mathbf{x})^T E(\mathbf{z}-\mathbf{x})$, for mean-centred vectors.

Generally, maximising independence among random variables is not the direct primary purpose of whitening transforms. These transformations are designed to remove

linear correlations and produce uncorrelated random variables. Thus, it is not surprising that they cannot consistently eliminate various types of non-linear dependencies. Nevertheless, under specific conditions, different whitening techniques can enhance independence, depending on the nature of the data.

As discussed in detail in Section 4.1, each whitening method corresponds to a distinct orthogonal rotation of the generic whitening transformation applied to the original data, and thus, the alignment between the whitening transform and any non-linear dependencies will vary, depending on this orthogonal rotation. Consequently, no single whitening technique can consistently minimise dependencies, whether linear or non-linear, because different whitening methods yield different alignments, which may affect the non-linear dependencies in the data in different ways. Instead, the effectiveness of whitening transformations in this task depends on the distributions and dependencies of the random variables within the vector, and the alignment of those dependencies with the whitening transform applied.

Taking PCA whitening as an example, if non-linear dependencies in the data are strongly aligned with the direction of maximum variance, PCA whitening may amplify them. However, the same transform can reduce non-linear dependencies when their directions do not align with the principal axis. In that case, by aligning the data along the principal components, PCA whitening may suppress non-linear dependencies in directions of lower variance.

No single linear whitening method universally maximises statistical independence because whitening only removes second-order correlations, not the higher-order dependencies ICA exploits. FastICA depends on PCA whitening to emphasise directions of high variance, often containing non-Gaussian structures, before hunting for the optimal orthogonal rotation. By contrast, distance correlation losses reward minimal residual cross-structure and align with ZCA and ZCA-Cor (see Table 4.2, rows 4–8), which preserve spatial geometry and produce high pairwise distance correlations. The ideal whitening hinges on the spectral content of the signal, the colour of the noise, and the degree of non-Gaussianity. As PCA whitening rescales the axes by $1/\sqrt{\lambda_i}$, the low variance noise components can be amplified and swamp the underlying non-Gaussian characteristics. Unmodelled Gaussian noise in the covariance can push the mixture closer to true

Gaussianity, thwarting ICA's contrast functions unless the noise structure is accounted for [151].

To transcend these limits, I propose applying a whitening transform to the underlying components at each epoch of a neural BSS pipeline. Since all whitening transforms differ only by a rotation, a properly designed algorithm (e.g. with a Restart strategy) can steer any initial whitened mixture toward the same global independence optimum, though convergence speed will vary with the data and whitening choice. In practice, ZCA-based methods remain attractive when preserving the geometric structure of the sources is critical.

In summary, whitening focuses on decorrelation rather than maximising independence, the latter being an objective that no whitening method explicitly addresses. The rotational degree of freedom among whitening transforms means that each transform can produce outputs that suppress, magnify, or distort non-linear dependencies in various ways.

In future chapters, unless otherwise specified, I will use Cholesky whitening. This choice is made for consistency with the work by [4], which noted that Cholesky decomposition is fully differentiable and easy to implement in frameworks like PyTorch, establishing Cholesky whitening as a standard practice in machine learning contexts. The only exception is Appendix B, where ZCA whitening was employed, for the needs of the research in that chapter.

4.3 Whitening within the BSS pipeline

In this section, I discuss two issues related to the use of whitening within blind source separation pipelines that utilise distance correlation. In Section 4.3.1 I argue that whitening should be applied before the double-centring step of the distance correlation computation, and in Section 4.3.2 that it should not be used as a pre-processing step within a BSS pipeline that uses distance correlation computation as loss function.

4.3.1 Whitening and double-centring

Here, I briefly discuss why whitening should be applied before the double-centring step of the distance correlation calculation. The main argument is that if the variable x to be whitened is double-centred, its mean is zero and the covariance matrix $\mathbf{\Sigma}$ has determinant zero. Since $\mathbf{\Sigma}$ is singular, it is not invertible. Therefore, no matrix $\mathbf{W}$ can satisfy $\mathbf{W}^T\mathbf{W} = \mathbf{\Sigma}^{-1}$, so the whitening condition fails.

Nevertheless, the pseudo-inverse of singular matrices can be approximated using singular value decomposition by replacing the diagonal matrix of singular values with its reciprocal, and if the matrix is invertible, then the pseudo-inverse and the inverse become the same.

For example, in the official code referenced in [4], the authors employed Cholesky whitening and added a small perturbation to the covariance matrix to prevent it from becoming singular during training. That is, their modification to the covariance definition is $\mathbf{\Sigma} = (1 - eps) \times \mathbf{\Sigma} + eps \times \boldsymbol{I}$. In my later research, I discovered that even minor changes in the covariances could adversely affect training and lead to the learning of poor representations when the perturbation method was used. As a result, I chose not to employ pseudo-inverse methods, which would have allowed whitening to be applied before the double-centring step.

4.3.2 Whitening within the BSS pipeline

Whitening can be applied at multiple points in a BSS pipeline. Here, I will explain why readers should avoid using whitening as a pre-processing step, as it may lead to suboptimal results when processing a larger number of mixtures from fewer sources. In the case of GNSS data, typically, there are many more mixtures than sources. I will argue that using a whitening transformation directly on the inputted mixtures is not advisable.

Indeed, whitening the GNSS signals, using PCA whitening in this instance, results in the same number of outputs as there were inputs. However, by the definition of whitening, all these outputs have unit variance, which means that the variance explained by each component will be equal. As none of the whitened mixtures account for more variance than the others, whitening limits the user's ability to reduce the number of extracted

components. Increasing their number, aiming at capturing more variance and reaching a satisfactory user-defined threshold of variance explained, may lead to overfitting.

The effect of whitening the input mixtures is demonstrated in Figures 4.1 and 4.2, which show the cumulative distance covariance of the original input mixtures, and their whitened versions, respectively. The mixtures comprise 120 variables in total, from forty stations in the South-Western United States of America across three directions. In both cases, I also show the corresponding graphs for the Median Interannual Difference Adjusted for Skewness (MIDAS) detrended data, which is a standard pre-processing technique for GNSS data. The percentage of distance covariance explained is calculated similarly to the percentage of covariance explained in PCA and is described in more detail in Appendix C.

In Figure 4.1, the first five components of the non-whitened data describe over 50% and around 80% of the distance variance, with and without MIDAS detrending, respectively. The primary component of distance covariance in the non-MIDAS-detrended case corresponds to the dominant trend of the time series, explaining over 60% of the distance variance. This trend is likely removed through MIDAS detrending, accounting for the difference in distance variance explained with and without this processing step. Meanwhile, in the case of whitening, the distance covariance described by each component is equal. Thus, the user will need a larger number of sources to reach a specified percentage of the distance covariance explained.

However, the importance of the cumulative percentage of variance or the distance variance explained is slightly exaggerated by the previous statements. By incorporating the inverse of the whitening transform into the mixing or unmixing layer, one can find the true mixing or unmixing matrix in the original space. Consequently, the percentage of variance or distance variance explained by the pre-whitened data can also be represented in the whitened data through a linear transformation.

That being said, I do not recommend whitening the raw mixtures upfront. This approach requires the network to first undo an arbitrary rotation and equal-variance scaling before isolating the signal. As a result, it diminishes the natural variance hierarchy that highlights the signal subspace, inflating low-variance noise directions to unit scale and increasing the effective search space. Moreover, the gradients from noise modes compete

with true signal directions. This competition slows down convergence and may lead to suboptimal local minima. Even if one uses the pre-whitened variance to determine the number of components, the initial directions typically account for a small amount of variance or distance covariance explained. Consequently, the model has to work harder to learn the mixing and unmixing matrix.

Therefore, it is not advisable to use whitening as a preprocessing step for source separation, but once the sources have been extracted before calculating the loss function.

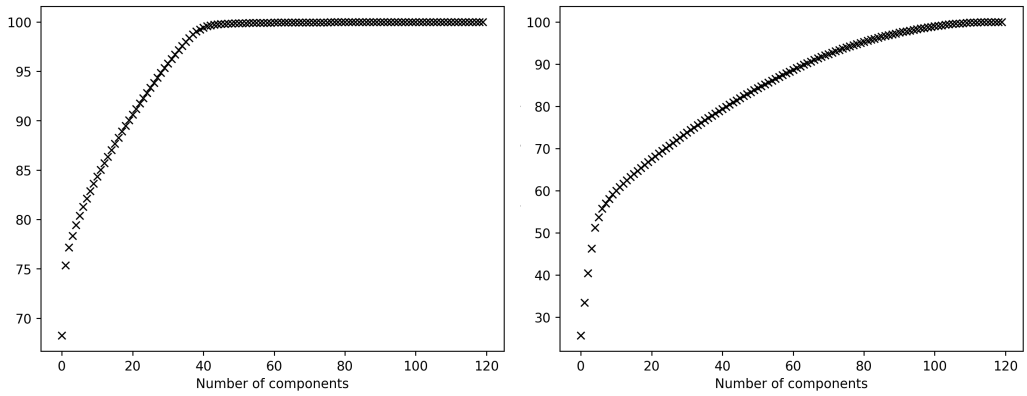


Figure 4.1: **Non-whitened data (as a pre-processing step applied to the GNSS mixtures):** Cumulative percentage of the distance covariance described by number of components. The GNSS data was provided by the Gualandi et al. as part of a case study of Post-large earthquake seismic activities in the region [3]. The left-hand figure is not MIDAS detrended, whilst the right-hand column is. The cumulative distance-covariance curves rise more quickly than for the cases of the PCA whitened GNSS mixtures (Figure 4.2). See its caption for more detail.

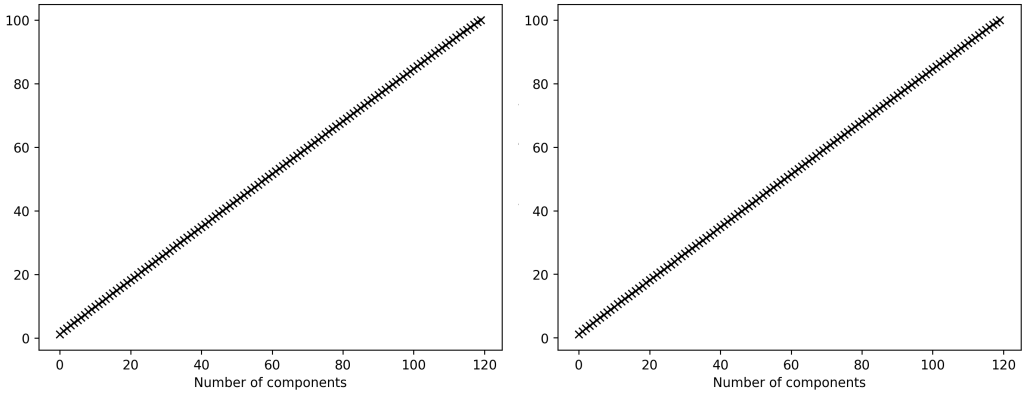


Figure 4.2: **Whitened data (as a pre-processing step applied to the GNSS mixtures):** Cumulative percentage of the distance covariance described by number of components. The GNSS data was provided by the Gualandi et al as part of a case study of Post-large earthquake seismic activities [3]. The left-hand figure is not MIDAS detrended, whilst the right-hand is. After PCA (or any other form of) whitening, each component has unit variance, so residual dependence (distance-covariance) is distributed evenly across dimensions. Consequently, the cumulative distance-covariance curves rise more slowly than for the cases of the raw GNSS mixtures (Figure 4.1), requiring more components to capture the same fraction of total dependence. In a high-mixture, low-source regime, (ZCA-, PCA- or Cholesky-based) whitening as a preprocessing step flattens the few dominant source variances, so initial features like principal or independent components no longer highlight independent directions. By rescaling every axis to unit variance, whitening removes the variance-based cues that would otherwise rank the strongest source axes. As a result, all components appear equally strong, which impairs the prioritisation of the leading sources for extraction, through maximising their independence. To preserve efficiency, enforce unit-covariance constraints within the source-estimation loop, i.e. re-whitening during iterative updates of a neural network, rather than solely as an initial preprocessing step.

4.4 Conclusion

In this chapter, I examined various forms of whitening in relation to independence, specifically through the lens of distance correlation. The investigation revealed that no particular form of whitening could be identified as optimal for enhancing independence, as independence encompasses both linear and non-linear relationships. Nonetheless, in most cases, ZCA-based whitening demonstrated strong performance, often yielding the best pairwise dependence between the original and the whitened variables. Additionally, the analysis indicated that the most effective time to apply whitening is after the source separation step. While ZCA-based whitening was found to be superior in this chapter, I used Cholesky

whitening in the subsequent chapters of this thesis to maintain consistency with the work by [4]. Their research noted that Cholesky decomposition is fully differentiable and easy to implement in frameworks like PyTorch, which has made Cholesky whitening a standard practice in machine learning contexts. The Cholesky transformation was also found to outperform PCA-based whitening in the BSS research in this Chapter. As an exception, in Appendix B, ZCA whitening was used, in line with the research discussed in this chapter.

After whitening, each component shows unit variance, eliminating second-order structure. Standard ICA pipelines, which use whitening followed by rotation, are effective when the number of mixtures and sources is comparable, with PCA whitening being a strong choice. However, in cases with a high mixture-to-source ratio, such as GNSS data, whitening as a preprocessing step is not advisable, as it slows the rise of the cumulative distance-covariance curve, requiring more components to represent dependence and increasing overfitting risk in neural networks. If fewer components are selected, applying whitening as a preprocessing step would require the network to work harder to account for the variance or distance variance explained that is missing from the first components, when compared to their unwhitened counterparts.

For BSS networks, it is beneficial to integrate whitening constraints within the iterative source-estimation loop by re-whitening during each training epoch. This maintains unit covariance throughout optimisation, improving convergence and source extraction. ZCA whitening aligns well with preserving cross-structure, aiding the search for independence through gradient updates. However, all whitening methods relate through a rotation and potentially converge to the same global minimum.

CHAPTER 5

Distance correlation as a loss function

5.1 Introduction

In this chapter, I will compare the use of *distance correlation*, *mutual information* and *negentropy* as loss functions used to describe source independence, highlighting the importance and relevance of the comparison in data science and signal processing.

Bakirov et al. [152] introduced distance correlation to measure linear and non-linear relationships between random variables. Moreover, the authors proposed that distance correlation could be an effective loss function for optimising blind source separation algorithms due to its distinctive properties. Note that in this chapter, I will consider vectors as random variables.

Mutual information is a natural metric for evaluating the relationship between two random variables from an information-theoretic standpoint. Recently, its application in blind source separation has gained popularity [149], largely due to the development of computationally efficient neural networks for estimating mutual information, such as MINE [20].

Negentropy represents a fundamental conceptual difference compared to distance correlation and mutual information. Unlike these metrics, which are calculated using pairs of random variables, negentropy is derived from a single random variable. It quantifies

the Gaussianity of a signal by measuring the Kullback-Leibler distance between the signal's probability distribution and that of a Gaussian distribution with the same mean and standard deviation. In most cases, minimizing the Gaussianity of the estimated sources enhances their independence in BSS, a concept supported by the Central Limit Theorem.

In Chapter 3, I compared the three independence metrics discussed using non-parametric tests to evaluate their effectiveness in source separation and optimisation. In this chapter, I will focus on the potential of these three metrics as loss functions for neural networks, with the aim of addressing two research goals.

Firstly, I will examine how the independence metrics relate to the best extraction of estimated sources by calculating the SI-SDR values between the estimated sources (produced solely through the gradient descent algorithm) and the ground truth. In addition, I will compute Pearson correlation coefficients to compare the final loss for each metric with their SI-SDR values, assessing how effectively each metric extracts the sources.

Secondly, I will examine the sources related to the best minima and assess how effective gradient descent optimisation is at finding these solutions. This analysis will explore the effectiveness of source separation through the gradient descent of the different loss functions, which can have complex loss landscapes. These complexities may hinder the effective learning of optimal minima with various random initialisations. To address this issue, I will employ a Restart algorithm to determine which loss functions extract better sources at good local minima when compared to the ground truth, independently of their random initialisations.

The two objectives distinguish between the best possible source separation that can be achieved with a particular metric and the complexity of training using the gradient descent algorithm for each metric. In other words, this highlights how likely it is that optimal outputs will be obtained through standard training methods.

The main contribution of the chapter is a systematic comparison between distance correlation, mutual information, and negentropy as measures of signal independence. In this approach:

- I conducted experiments using two optimisation algorithms, one to compute the extrema of the independence measures and the other to compute the independence measures through standard gradient descent. My goal was to thoroughly evaluate

the effectiveness of these independence measures in source separation, distinctly from the practical aspects of calculating their extrema.

- I used/introduced the use of BSS on several data types: the synthetic data problem and real GNSS and SAR data containing embedded known signals.

5.2 Implementation

Recently, neural networks, specifically deep neural networks, have been increasingly used for blind source separation, achieving state-of-the-art results. Sometimes, training is supervised ([103] [153]) using labelled data from a specific application domain. In other cases, unsupervised neural networks ([154] [11] [5]) are used instead of more traditional iterative methods ([23] [155]) to solve the optimisation problem.

As an extension to the gradient descent algorithm, a new algorithm was implemented to explore more of the loss landscape, helping the training process escape local minima. For every epoch in which the independence metric was minimised, I compared the independence loss for a neural network model with learned parameters (updated by gradient descent) to that of a model with randomly initialised parameters. If the randomly initialised model exhibited a lower loss function, its parameters were adopted and the learning was restarted. This strategy enables a thorough exploration of the weight space, making it easier to find lower loss values by using randomly initialised weights. These values are more likely to converge to a global minimum, which can improve the source separation. Additionally, this approach allows for comparisons of optimal source separation outputs across different metrics. The Restart algorithm, Algorithm 1, outlines this method.

I have included a visualisation that illustrates the difference between the original network and the Restart method. See Figure 5.1. On the left, you can see the original method, which uses gradient-descent-updated weights. On the right are two networks from the Restart method: one with the same gradient-descent-updated weights as before and another that is randomly initialised. In the Restart Algorithm case, every epoch, one of the model's weights is updated using gradient descent, whilst the other is reinitialised each epoch. If the second network has a lower loss (as is the case in the visualisation), its

Algorithm 1 Pseudocode representing the **Restart** technique.

Randomly initialise θ
for $epoch = 1, 2, \ldots$ **do**
  Compute S and independence loss L with weights θ
  Randomly initialise θ_1
  Compute S_1 and independence loss L_1 with weights θ_1
  if $L_1 < L$ **then**
    $\theta \leftarrow \theta_1$
    $L \leftarrow L_1$
    $S \leftarrow S_1$
  else
    Pass
  end if
  optimise L wrt. θ
  $\theta_{epoch+1} \leftarrow \theta_{epoch} - \eta \Delta \theta_{epoch}$
end for

parameters replace that of the gradient descent network, and the process continues until training ends.

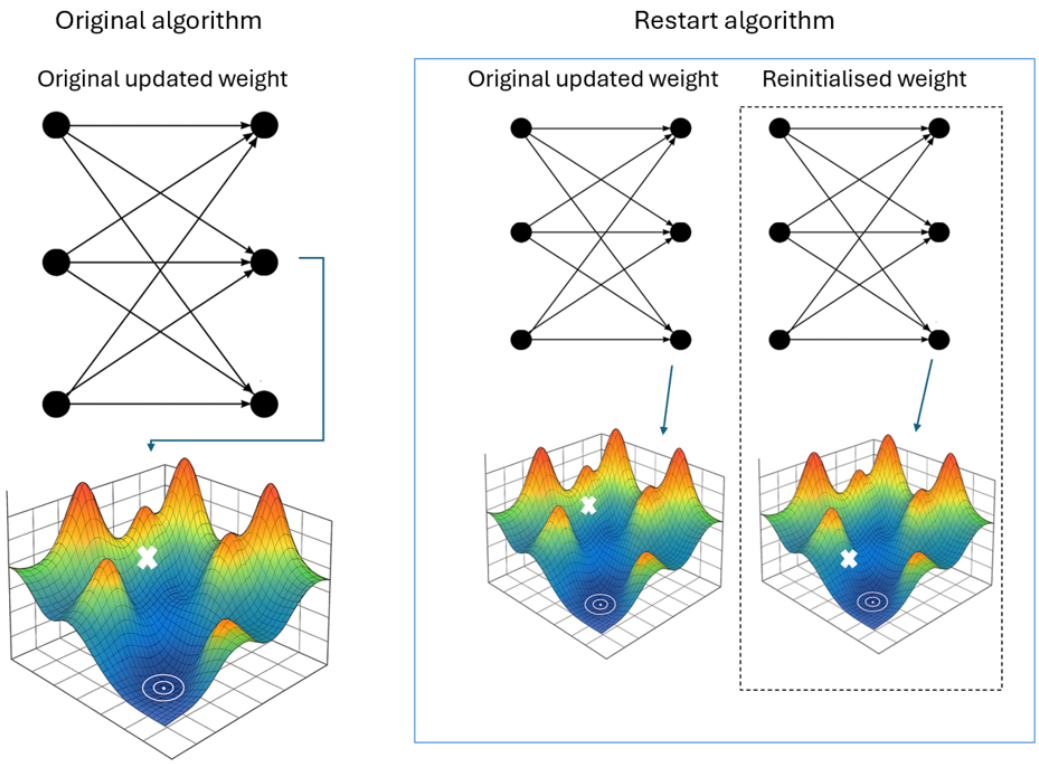


Figure 5.1: A visualisation of the Restart method in comparison to the original neural network.

If one considers the white cross on the loss landscape to represent the loss associated with separation, choosing the second option for the weights brings the results closer to the global minimum. This approach also helps to avoid the risk of the training process ending up in a poor local minimum or getting stuck on a saddle point.

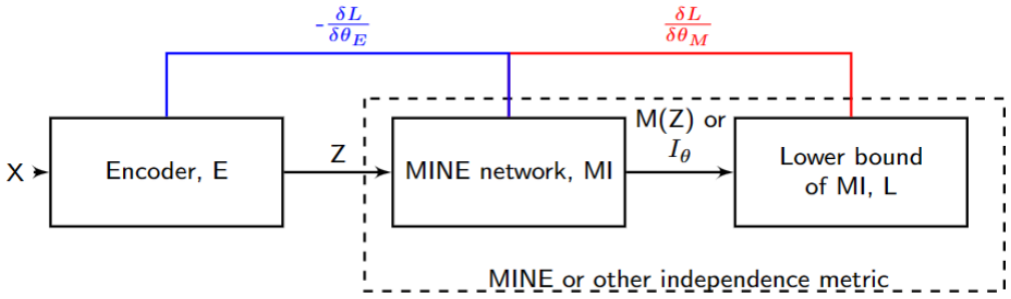
I estimated the underlying sources using two methods for each metric (though not for the FastICA method). The first mixture-input method employs the more common formulation, in which mixtures serve as the input to the neural network. The mixtures are passed through a linear layer, generating underlying sources (with an equal number of mixtures and sources). The goal is to maximise the independence of these outputted sources by optimising negentropy, mutual information through MINE, or distance correlation. This approach is called the **Separation** method. See Figure 5.2a. It is important to note that when using loss functions other than MINE, the MINE component of the architecture is omitted.

The second method is the source-input case, where the underlying sources are learnable parameters within the network. This allows the mixing matrix to be rectangular (i.e. not square), such that the number of sources and mixtures can differ. I apply a linear layer to the sources to produce the known mixtures. Then, I employ a reconstruction loss to ensure that the sources can accurately reconstruct the known mixtures and an independence loss to maximise the independence of the sources. This approach is the **Reconstruction** method. See Figure 5.2b.

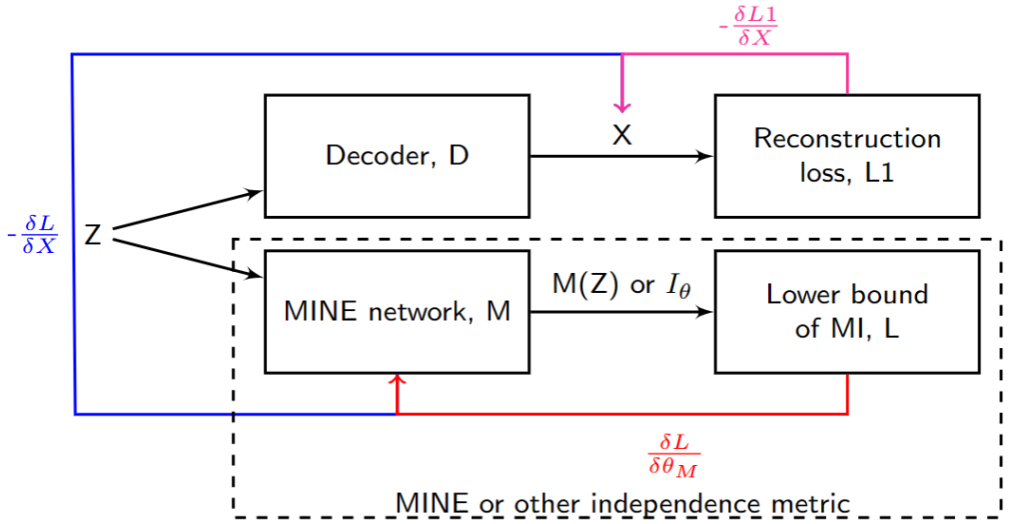
In the **Reconstruction** method, I found that I needed to prioritise the reconstruction loss to achieve outputs similar to the known mixtures. If the weighting factor places too much emphasis on the source independence, it can lead to the development of white noise sources, which do not adequately combine to form the desired mixtures.

In the context of network architecture, factors such as the number of layers, their sizes, and the types of non-linearities and normalisation can significantly impact convergence speed when using distance correlation, MINE-mutual-information or negentropy-based losses. These elements also influence the final independence scores, including both distance correlation and SI-SDR values, as well as the sensitivity of the network to initialisation and overfitting.

Shallow networks, such as those consisting of only a linear layer or shallow multilayer



(a) In MINE, a neural network learns statistically independent outputs through the alternate optimisation of an encoder E and a MINE network M, which are parametrised by θ_E and θ_M, respectively. When using distance correlation or a negentropy-based loss, these metrics can be expressed in closed form. As a result, the sources are optimised for independence using one of these non-parametric metrics at each epoch unless MINE is being investigated. The figure presented illustrates the **Separation** architecture.



(b) For MINE, the neural network learns statistically independent learnable parameters Z, the sources, by alternating the optimisation of the decoder D and the MINE network M parametrised by hidden parameters, which include the decoder parameters and learnable sources. In the MINE case, mutual information is estimated and minimised between sources. The non-parametric distance correlation and the negentropy-based losses are calculated in closed form, allowing them to be computed for each epoch without the need for an additional neural network to approximate mutual information like MINE. The figure presented illustrates the **Reconstruction** architecture.

Figure 5.2: The neural network architectures for the **Separation** and **Reconstruction** methods.

perceptrons, may be limited to capturing only linear or mildly non-linear mixing. In contrast, deeper neural networks can better approximate complex source transformations, although they run the risk of encountering vanishing gradient issues or the sources losing meaning, or becoming more suited to representation learning. Wider layers enhance the capacity of the network to disentangle subtle higher-order dependencies but may lead to overfitting due to capturing spurious noise.

Regarding non-linearities, if activation functions are employed, those with zero-slope regions (like ReLU) can freeze parts of the network. In contrast, smooth saturating functions (such as tanh) may obscure fine dependencies. Moreover, incorporating batch normalisation or a custom whitening layer during training can enforce unit covariance, potentially stabilising the distance correlation objective. Additionally, residual connections help preserve identity mappings, allowing the network to focus on decorrelating features rather than reconstructing the entire transformation, which often enhances convergence, especially in deeper networks.

Considering the data types, GNSS/SAR data can be classified as either time series or spatial information. Thus, RNNs or LSTM networks are suitable for GNSS time series data, while CNNs are more appropriate for SAR spatial data. Convolutional filters can effectively capture local dependencies, influencing the speed and accuracy of independence maximisation. However, for source separation in a simple synthetic scenario, this problem is theoretically solvable with a single linear layer. To simplify my analysis of the loss function, I aimed to keep the neural network as straightforward as possible. Regarding the GNSS data, its additive error equation indicates that it could also be modelled as a linear layer. In both cases, the inputs were treated as random variables rather than time series. I found no benefit to temporal modelling (RNN/LSTM) in the synthetic case, so I treated inputs as IID features for simplicity.

In the following sections, I will outline how to implement source separation algorithms using mutual information, distance correlation, and negentropy.

FastICA

In my work, I use two forms of ICA. The first is the FastICA implementation provided by Scikit-learn [156] [25]. The form of whitening applied to the outputted sources in this

framework is a unit variance version of PCA whitening.

Additionally, I will introduce PyFastICA, a neural network version of FastICA developed in PyTorch. This implementation uses a gradient-based iteration scheme rather than a fixed-point iteration scheme. In this case, I opted for Cholesky whitening because it aligns well with other machine-learning tasks involving source separation [149] and representation learning [4]. More importantly, after conducting 100 repetitions, Cholesky whitening consistently yielded higher average SI-SDR values than the other whitening techniques. It is important to note that whitening techniques are related through rotations, which means that different whitening methods can lead to the same separations as the reverse of any rotation can then be incorporated into the unmixing matrix. In this case, the non-quadratic function referenced in Equation 2.1 for PyFastICA is the logarithm of the softmax function, which I found to be more stable during training than other functions. Moreover, as the data is whitened, each feature has a variance of 1 and, therefore, the Gaussian entropy is constant. Thus, the negentropy loss, Equation 2.1, is simplified to be only the entropy of the whitened data for stability in training.

5.2.1 MINE

The Separation architecture with the MINE extension was introduced in the foundational paper by Hlynsson and Wiskott [149]. In this work, I utilise the MINE architecture described in [149], which computes the joint and marginal distributions of input vectors using a deep neural network consisting of six linear layers followed by Leaky ReLU activations, culminating in a final linear layer. The output sample size is set to 64 for every linear layer. The outputs of this architecture are used to calculate the parametrised mutual information, as specified in Equation 2.11. I aim to maximise the estimated mutual information to approximate the actual mutual information closely. This result is then employed as the mutual information, serving as the independence metric for the subsequent minimisation step.

As the MINE network is parametric, the outputted sources are kept constant by freezing the linear layer weights for seven out of every eight epochs when the MINE section of the architecture is iterating. For the final epoch, the weights are unfrozen, and the MINE architecture's weights are frozen, to minimise the parametrised mutual information.

MINE requires many epochs of maximisation within the network to effectively approximate the lower bound of the actual mutual information before minimisation takes place. I discovered that if too few steps are used during the maximisation phase, the transition between maximisation and minimisation tends to converge to zero without crossing it. Consequently, the lower bound of mutual information may not be accurately estimated. As a result, source separation may be trained on a value that does not truly represent the source dependence whilst appearing to have mutual information that is close to the minimum. The parametrised mutual information maximisation converges more reliably on its true value after many epochs. However, incorporating MINE with additional epochs of maximisation into a source separation algorithm will significantly increase the training time.

5.2.2 Distance correlation

In this section, I will examine the foundation of the whitening Mean Squared Error (W-MSE) method used for contrastive representation learning. W-MSE involves scattering data by whitening it and minimising the distance between positive pairs. Although the W-MSE method is not the main focus of this chapter, its introduction aligns well with the introduction of distance correlation as a loss function. In the next chapter, I will utilise the W-MSE method and its distance correlation extension.

Representation learning is an advanced technique utilised in the field of machine learning that involves the development and application of algorithms designed to identify and learn meaningful patterns from a given dataset. This process can occur in two primary ways: supervised learning, where the model is trained on labelled data, or unsupervised learning, where it seeks out patterns without any predefined labels.

The aim of representation learning is to create representations that are not only informative but also interpretable. These representations can help uncover hidden features within the data, making it easier to understand complex datasets.

One of the most effective methods in representation learning is the use of contrastively trained models. These models focus on measuring the similarity or dissimilarity between data elements, operating under the premise that similar data points share semantics. As a result, the distance between these similar points is minimised. Conversely, for dissim-

ilar pairs, the model ensures that their distances remain maximised (or minimised when compared to a threshold margin distance).

In W-MSE, a whitening transform projects the latent space representation onto a spherical distribution. The L_2 normalisation then brings the whitened representation onto the unit hypersphere. The MSE loss function is then used to bring positive pairs together, as depicted in Figure 5.3 for the distance correlation extension. The original W-MSE would be represented by this figure if you replace the pairwise distances **A** and **B** with **X** and **Y**.

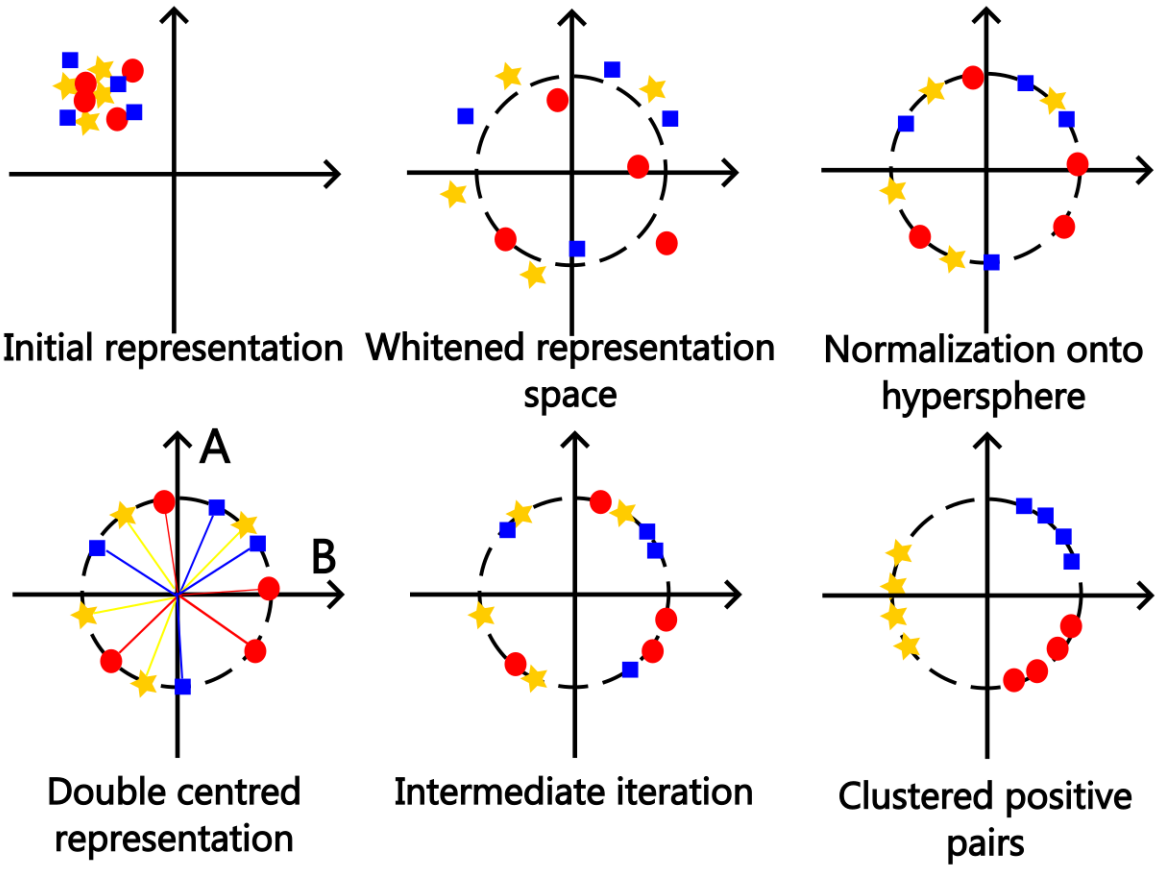


Figure 5.3: Depiction of self-supervised learning through whitening and minimising the angles between positive pairs. An extension of the work in [4] for a distance correlation-based loss.

As stated in the seminal paper [4], the relation between the mean squared error of the

normalised whitened representation and the cosine similarity is:

$$\begin{aligned} MSE(\mathbf{X},\mathbf{Y}) &= \left\| \frac{\mathbf{X}}{\|\mathbf{X}\|} - \frac{\mathbf{Y}}{\|\mathbf{Y}\|} \right\|_{\mathbf{2}}^{\mathbf{2}} \\ &= \frac{\mathbf{X^T X}}{\|\mathbf{X}\|_2^2} + \frac{\mathbf{Y^T Y}}{\|\mathbf{Y}\|_2^2} - \frac{\mathbf{X^T Y}}{\|\mathbf{X}\|_2 \|\mathbf{Y}\|_2} - \frac{\mathbf{Y^T X}}{\|\mathbf{X}\|_2 \|\mathbf{Y}\|_2} \\ &= 2 - 2\frac{\langle \mathbf{X},\mathbf{Y} \rangle}{\|\mathbf{X}\|_2 \|\mathbf{Y}\|_2}. \end{aligned} \tag{5.1}$$

By expanding the empirical distance correlation, seen in Equation 2.12, and substituting in Equations 2.13 and 2.14, I derive Equation 5.2, with the index n, in $\mathbf{X}_n$, representing the sample size of the random variable $\mathbf{X}$ (which is not necessarily a mixture in this instance).

$$R_n^2(\mathbf{X},\mathbf{Y}) = \frac{\frac{1}{n^2}\sum_{k,l=1}^{n} \mathbf{A}_{kl}\mathbf{B}_{kl}}{\sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} \mathbf{A}_{kl}\mathbf{A}_{kl} * \frac{1}{n^2}\sum_{k,l=1}^{n} \mathbf{B}_{kl}\mathbf{B}_{kl}}} \tag{5.2}$$

In Equation 5.2, the n components of the numerator and denominator cancel out. Applying vectorisation to Equation 5.2 yields Equation 5.3:

$$R_n^2(\mathbf{X},\mathbf{Y}) = \frac{1}{\|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2} \sum_{i=1}^{n^2} \boldsymbol{A}_i \boldsymbol{B}_i = \frac{\langle \boldsymbol{A},\boldsymbol{B} \rangle}{\|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2}. \tag{5.3}$$

Thus, the relation between the mean squared error and the square of the sample distance correlation, equivalent to Equation 5.1, is:

$$\begin{aligned} MSE(\mathbf{A},\mathbf{B}) &= \left\| \frac{\mathbf{A}}{\|\mathbf{A}\|} - \frac{\mathbf{B}}{\|\mathbf{B}\|} \right\|_2^2 \\ &= \frac{\mathbf{A}^T\mathbf{A}}{\|\mathbf{A}\|_2^2} + \frac{\mathbf{B}^T\mathbf{B}}{\|\mathbf{B}\|_2^2} - \frac{\mathbf{A}^T\mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} - \frac{\mathbf{B}^T\mathbf{A}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} \\ &= 2 - 2\frac{\langle \mathbf{A},\mathbf{B} \rangle}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = 2 - 2R_n^2(\mathbf{X},\mathbf{Y}). \end{aligned} \tag{5.4}$$

In the notation for the Separation Architecture, the inputs are the k mixture signals $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_k)$ and the output is the k separated sources $(\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots \boldsymbol{S}_k)$, all of length n. Therefore, I calculate the pairwise distance-correlation-based loss between the sources as:

$$\sum_{1\leq i<j\leq k}(1-R_n^2(\boldsymbol{S}_i,\boldsymbol{S}_j)). \tag{5.5}$$

To increase pairwise independence between sources, this loss function is maximised, so that the distance correlation, R_n, is minimised.

For comparison, I will present the W-MSE method from [4], which serves as the basis for the aforementioned method. This method utilises random variables as its input. Here, I will use the notation $\boldsymbol{X}$ and $\boldsymbol{Y}$ to refer to the random variables for consistency with the literature. Taking two random variables with a mean of zero and which are normalised, I can write the Pearson correlation coefficient, which measures the linear relationship between variables, as:

$$r(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{n}\sum_i(\boldsymbol{X}_i\cdot\boldsymbol{Y}_i). \tag{5.6}$$

If the random variables are normalised, the MSE between random variables is:

$$MSE(\mathbf{X},\mathbf{Y}) = \frac{1}{n}\sum_i^n(\boldsymbol{X}_i-\boldsymbol{Y}_i)^2 = 2\left(1-\frac{1}{n}\sum_i^n(\boldsymbol{X}_i\cdot\boldsymbol{Y}_i)\right). \tag{5.7}$$

Therefore, it can be seen that the mean squared error is related to Pearson's correlation r for normalised variables,

$$MSE(\mathbf{X},\mathbf{Y}) = 2(1-r(\mathbf{X},\mathbf{Y})). \tag{5.8}$$

Equations 5.8 and 5.4 share a similar structure.

Search space strategy

Comment: In the case of k=3, the components of the loss function of Equation 5.5 are $R_n^2(\boldsymbol{S}_1,\boldsymbol{S}_2)$, $R_n^2(\boldsymbol{S}_1,\boldsymbol{S}_3)$, and $R_n^2(\boldsymbol{S}_2,\boldsymbol{S}_3)$. Minimising the average can reveal interesting patterns in various local minima. For instance, in one minimum of the synthetic example, I noticed a pattern where the training effectively minimised two of the three distance correlations but at the cost of the third. As a result, two reconstructed signals became similar, impacting the average SI-SDR. For more details, see Appendix B.

5.3 Data

In this chapter, I will address the synthetic problem by extracting three known sources from three mixtures. Refer to Section 4.2.1 for more details.

The standard test bed for developing general blind source separation techniques is the audio domain. Audio BSS involves extracting underlying signals from mixtures of human voices and/or other sounds captured by one or more microphones, typically in a supervised manner. However, BSS in the audio domain will not be addressed in this chapter but in Chapter 7.

Blind source separation of GNSS and SAR signals presents a more challenging problem, allowing for more robust comparisons between various independence metrics. This issue is more complex for geodetic data than audio datasets, where the underlying sources are known and included in the training data. In geodetic datasets, the number and types of sources contributing to each mixed GNSS/SAR signal are poorly understood and can vary by location. For example, slow slip events may occur in the Cascadia subduction zone but not in southern California. In this chapter, I will use a simplified two-mixture-two-source problem to compare the independence metrics more easily.

5.3.1 GNSS data

For the temporal ICA case, I created two mixtures by adding a synthetic signal that simulates an earthquake, followed by post-seismic deformation, to actual GNSS signals from NGL [80]. The goal was to establish source separation tasks with varying difficulty levels controlled by the scaling factor associated with different epicentres located along a line between the two GNSS stations. Here, I aimed to extract two sources from two mixtures.

For the initial set of tasks, the vertical component of the GNSS time series data from two nearby receivers, ORWA and P445, located 1.19 km apart, is used. Due to their proximity, the time series from these receivers exhibit a high correlation, with a Pearson coefficient of 0.903 (and coefficients between the two stations and the added signal of -0.249 and -0.172). The synthetic signal, *Synth*, consists of a vector containing 400 zeros followed by the values generated by the sigmoid function of 600 equidistant points ranging from 0 to 8. This vector represents a step function that transitions into a sigmoid

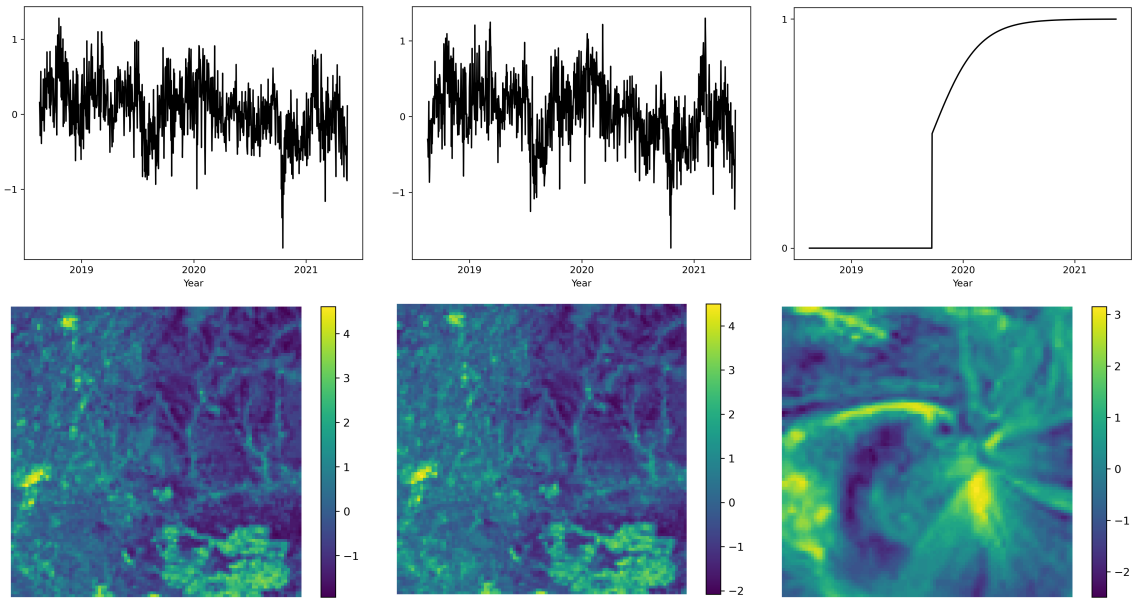


Figure 5.4: **From left to right. Top:** ORWA, P445, and the synthetic signal. The time series have been centred and scaled, producing a unitless y-axis. **Bottom:** the 100×100 SAR1 and SAR2 images, and the additive sun image.

curve, simulating an earthquake event followed by post-seismic deformation. These two signals are illustrated in Figure 5.4 (Top).

The synthetic signal is generated at a point between ORWA and P445. The synthetic signal is combined with the GNSS signals, with weights based on the ratio, r, of the distance from ORWA to the total distance between the stations. The mathematical equations are as follows:

$$\begin{aligned} mix1 &= ORWA + r \times Synth \\ mix2 &= P445 + (1-r) \times Synth\ . \end{aligned} \tag{5.9}$$

In order to take the geodetic signals as one source, the two GNSS displacement time series, ORWA and P445, must be similar. A suitable blind source separation algorithm should ideally produce one source closely resembling these GNSS series and a second source reflecting the added synthetic signal. I created five pairs of mixtures, corresponding to ratios of $r = \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}$. It is noteworthy that $r = \frac{1}{2}$ results in two mixtures that are very similar to each other, making the two-source separation task quite challenging. In contrast, $r = \frac{1}{4}$ and $\frac{3}{4}$, as well as $r = \frac{1}{8}$ and $\frac{7}{8}$, present progressively more manageable

separation tasks.

I followed the same procedure for the second experiment but used the ORWA and GRSV stations, which are 26km apart. The signals had a lower Pearson coefficient of 0.742 (and coefficients between the two signals and the synthetic signal of -0.192 and -0.238). Due to the spatial proximity between ORWA and P445, their atmospheric noise is similar, but that is not necessarily the case for the ORWA/GRSV case.

The proposed hybrid dataset approach benefits from a well-defined ground truth for the synthetic source. This enables more rigorous and precise comparisons between metrics than using solely GNSS data.

5.3.2 SAR data

To evaluate the effectiveness of different metrics on spatial data, I created two source separation problems: one straightforward and one more complex problem. I based the problems on two 100×100 pixel crops from synthetic aperture radar (SAR) maps with a high degree of similarity and a non-geodetic signal dissimilar to both. The two SAR maps (sar1 and sar2) were collected from sweeps across Carlisle toward Sunderland on June 29, 2023, and July 11, 2023 [157]. The second signal, *sun*, is a 100×100 pixel section of an image of the Sun. I chose the Sun image because it contains spatial information yet is not derived from a geological context, ensuring that it does not share information with the SAR maps. The Pearson correlation coefficient between sar1 and sar2 was 0.940, while the correlations between the SAR maps and the Sun image were -0.0287 and -0.0298, respectively, for sar1 and sar2.

For the first (Defined) problem, I added a multiple of 0.5 and 2 of the *sun* signal to one of the SAR signals, sar1 and sar2, creating two distinct mixtures. That is:

$$\begin{aligned} mix1 &= sar1 + 0.5 \times sun \\ mix2 &= sar2 + 2 \times sun. \end{aligned} \tag{5.10}$$

For the second (Undefined) problem, I added a 0.5 scaled *sun* signal to both sar1 and sar2, resulting in two similar mixtures that made extracting the added sun more challeng-

ing. That is:

$$\begin{aligned} mix1 &= sar1 + 0.5sun \\ mix2 &= sar2 + 0.5sun. \end{aligned} \tag{5.11}$$

5.4 Results

To compare the performance of the methods, I adopted the Scale Invariant Signal-to-Distortion Ratio (SI-SDR) [42], a metric widely used to assess blind source separation problems, especially audio source separation. I conducted each experiment 10 times and provided the means and standard deviations of the outputs as the results. The standard deviation serves as a measure of reproducibility. The target and output sources are both mean-centred to prevent any bias term from influencing the SI-SDR values, as this would not accurately reflect the content and interdependence of the signals.

5.4.1 Synthetic problem

In the Synthetic problem, I used two distinct architectures: the Separation architecture and the Reconstruction architecture. The Separation architecture takes the known mixtures as its input, processes them through a linear layer, and learns the unmixing matrix. On the other hand, the Reconstruction architecture incorporates learnable sources as parameters of the neural network. I initialised the learnable sources for the Reconstruction architecture as a tensor of random numbers from a normal distribution with mean 0 and variance 1. These sources are passed through a linear layer to generate mixtures that closely approximate the known mixtures.

The Synthetic problem involves three mixtures and three sources, allowing both architectures to be employed. I calculate various independence measures (MINE, PyFastICA and distance correlation) for both architectures to maximise the independence of the sources. The Reconstruction method requires an additional MSE reconstruction element in the loss function to ensure that the outputted mixtures are similar to the known mixtures. The losses were optimised by either gradient descent or the Restart Algorithm

(Algorithm 1).

Tables 5.1 and 5.2 are the mean SI-SDR values between the known and the outputted sources for the Separation and Reconstruction architectures, respectively, for the PyFastICA, MINE and distance correlation losses (and for the MSE loss in the Reconstruction architecture case). Since ICA permits permutations, the resulting sources are rearranged to maximise the average SI-SDR values across the three outputted sources. This means that even if the sine wave is intended to be the first source but is instead output as the third source, the average SI-SDR is not negatively affected due to the permutation-invariant nature of the evaluation. The FastICA method is a standard fixed-point iterative scheme that corresponds to the separation method and utilises the method's default unit-variance whitening.

I will provide the hyperparameters used for the results shown in Tables 5.1 and 5.2. For the PyFastICA results in the Table 5.1, the learning rate was set to 0.005, and Cholesky whitening was applied. In the case of the MINE results, the architecture had hidden layers with a size of 64, with a learning rate of 0.005 and Cholesky whitening. For the distance correlation, the learning rate was adjusted to 0.0003, and ZCA whitening was utilised. In all instances, a Nadam optimiser was used.

In the Separation architecture (Table 5.1), the FastICA method demonstrates the highest robustness, achieving the best average performance across all inputs. Although both the PyFastICA method and MINE also show relatively high average performance, they do not surpass FastICA. It should be noted that while the distance correlation exhibits a lower SI-SDR, it has a higher variance, which I intend to leverage to identify the optimal solutions for achieving good distance correlation minima.

In the case of the Reconstruction method, the results presented in Table 5.2 for the neural network optimisation were not whitened, as whitening led to suboptimal outcomes. I randomly initialised the learnable source parameters for the MSE, PyFastICA, and distance correlation methods. In contrast, the MINE results used the known mixtures for the initialisation, which yielded a better source output. The learning rates for the MSE, MINE, PyFastICA, and distance correlation methods are 0.003, 0.003, 0.003, and 0.0003, respectively.

When comparing the Separation and Reconstruction architectures, the average SI-

Input	FastICA	PyFastICA	MINE	DistCorr
Sine	21.9±0.0	19.2±7.5	13.8±5.7	3.2±8.1
Square	30.9±0.0	23.9±9.5	27.8±4.7	11.0±24.7
Sawtooth	21.6±0.0	20.4±5.4	14.7±4.4	5.9±6.6

Table 5.1: SI-SDR values for the Separation architecture in Fig. 5.2a. Provided is the average over ten repetitions.

Input	MSE	FastICA	PyFastICA	MINE	DistCorr
Sine	9.9±6.4	21.9±0.0	0.5±1.2	7.0±0.6	6.1±1.8
Square	8.6±6.0	30.9±0.0	7.9±4.3	19.2±1.0	26.3±2.4
Sawtooth	9.0±4.0	21.6±0.0	2.1±1.3	8.7±0.4	5.5±2.3

Table 5.2: SI-SDR values of the Reconstruction architecture in Fig. 5.2b. Provided is the average over ten repetitions.

SDR values are higher for the Separation method. The lower values observed in the Reconstruction method stem from its more complex loss function, which must balance the trade-off between reconstructing mixtures and achieving independent sources. For instance, while white noise signals are independent of each other, they cannot combine to create the known mixtures. MSE was used as a baseline to demonstrate the improvement or failure in training when employing an independence metric versus not using one.

I will revisit the Reconstruction architecture later, but for now, I will focus on the Separation architecture. I will examine the high standard deviations in the metrics to determine if the best outputs from each method are competitive with those of FastICA. Among the metrics studied, the distance correlation loss has the lowest average performance compared to PyFastICA and MINE, as indicated in Table 5.1. However, it does exhibit a high standard deviation, which I hope to leverage through more intensive training. I aim to determine whether, when it performs well, distance correlation outperforms the other independence metrics. For a more thorough examination of the sources produced using a distance correlation loss and the Separation architecture, both with and without whitening, please refer to Appendix B.

Here, I performed 100 repeats to produce sufficient final SI-SDR results for a metric comparison. The best output from every 10 trials, the top 10 results out of 100 and the Restart method (Algorithm 1) results are shown in Table 5.3. In Table 5.3, PyFastICA

		Best of 10		Best 10 of 100		Restart method	
Input	FastICA	PyFastICA	DistCorr	PyFastICA	DistCorr	PyFastICA	DistCorr
Sine	**21.9±0.0**	20.9±0.9	19.4±0.1	21.2±0.9	19.3±0.2	19.6±1.9	19.3±0.1
Square	30.9±0.0	34.0±3.5	**53.5±2.7**	33.9±2.5	**53.8±2.7**	27.3±6.6	**57.4±3.0**
Sawtooth	21.6±0.0	**26.1±1.3**	19.0±0.1	**26.1±1.2**	18.9±0.3	**23.7±4.7**	19.1±0.2

Table 5.3: The average SI-SDR across ten outputs for the PyFastICA and DistCorr methods is analysed using different strategies to achieve the best results. The strategies discussed include the Restart method, selecting the best output from every ten generated, and choosing the best ten outputs from a set of one hundred. All of these methods utilise the loss function rather than the SI-SDR as a guiding metric. It is important to note that I dropped the MINE method because it became increasingly unstable and less competitive when employing the Restart method. This instability arose from its moderate Pearson correlation of 0.596, measured between the average mutual information between all pairs of sources and the average SI-SDRs for those sources compared to their ground truth over 100 repeats. Consequently, using a lower average mutual information does not necessarily reduce the average SI-SDR values of the outputted sources. The SI-SDR values for the sine, square, and sawtooth waves were 9.8 ± 4.5, 15.1 ± 4.5, and 11.1 ± 6.4, respectively, for the Restart method.

employs Cholesky whitening with a learning rate of 0.005, while DistCorr utilises ZCA whitening and has a learning rate of 0.0003.

The results in Table 5.3 show that I can utilise the high variance observed in the PyFastICA and DistCorr separation methods to achieve effective source separation when optimisation during training converges to the global minimum for these metrics. The best overall results and the best results every ten iterations from 100 repeats for the PyFastICA and DistCorr methods yield similar SI-SDR values to those obtained using the gold-standard FastICA method. This suggests that the choice of method is primarily a matter of training to attain the best optimisation outcomes. On average, the distance correlation method slightly outperforms the two FastICA methods. Furthermore, the Pearson correlation between the final loss and the SI-SDR of the final source outputs for PyFastICA is 0.577. This correlation coefficient suggests that using negentropy to guide the Restart method may lead to a degradation in SI-SDR results, as seen in Table 5.3.

The MINE algorithm performs poorly when the Restart approach is applied. There is a correlation of only 0.596 between the average mutual information approximation between sources and the SI-SDR values for these sources compared to their ground truths. This limitation arises because the Restart method does not fully consider the parametrised

nature of MINE.

MINE estimates the lower bound of mutual information, and performing a gradient ascent step is essential for fine-tuning the parameters to closely approximate mutual information at this lower bound. However, convergence to this lower bound is not guaranteed, especially when only a few epochs are used during the gradient ascent process. As a result, when the Restart method picks the lowest approximated mutual information loss values, it is possible that the training of the MINE network has not fully converged to a value that accurately reflects the mutual information between the extracted sources for the learned and randomly initialised architectures. A poor MINE convergence to a lower parametrised mutual information may indicate that a less accurate approximation is selected, suggesting that more epochs of MINE training are needed to obtain valid results. Nevertheless, due to the increased complexity introduced by the parametrised MINE loss, it will only be used sporadically for BSS in this thesis and is why the MINE method has been dropped in Table 5.3.

The distance correlation improves significantly with the Restart algorithm, resulting in the highest average SI-SDR of 31.9 ± 1.0, along with the best SI-SDR value for the square wave signal. In contrast, the FastICA algorithm achieves an average SI-SDR of 24.8 ± 0.0 across all tested signals. It is important to note that while the average performance of the FastICA algorithm is strong for this example, SI-SDR is not a linear metric and that the research in [152] demonstrated that FastICA may not perform as well as other ICA algorithms when evaluated using the Amari index on various test cases. Despite concerns regarding FastICA's ability to extract the most non-Gaussian sources, the example provided in Scikit Learn illustrates a specific scenario in which FastICA performs exceptionally well.

The results in Table 5.3 indicate that no single method is superior, as all yield similar competitive outcomes. Among the methods, FastICA, distance correlation, and PyFastICA perform best on the sine, square, and sawtooth waves, respectively. A two-sample t-test (n=10) revealed that for the sine wave, all comparators had p-values below 0.05 when compared to FastICA. Similarly, for the highest value observed with the square wave, specifically the Restart method with DistCorr loss, all other methods also had p-values below 0.05. In contrast, for the sawtooth signal, while most p-values were below

0.05, the results from the other PyFastICA methods were not statistically different. Additionally, the average for the Restart distance correlation method is statistically different from the averages of the other methods.

The main differences in SI-SDR values stem from the additive noise present in the reference signal. This noise influences how well the sine, square, and sawtooth components approximate the input signal. Once the non-noise component of the signal is accurately approximated, the difference between that SI-SDR and the theoretical maximum (which is infinity) depends on the quality of the additive noise approximation. Since the noise is not explicitly extracted during BSS but significantly impacts the SI-SDR, its influence may often be underestimated, even though all methods remain competitive.

In Table 5.4, I compare the average distance correlation between signal pairs for the original method, the top one every ten results, the top 10%, and the Restart technique. I haven't specifically prioritised distance correlation for any pairs through weighting in the examples provided. The standard deviation of the distance correlations and the distance correlations decreased when I compared the original to other methods. This reduction occurs because I only select the best outputs as defined by the independence metric for this analysis.

The distance correlations obtained from the Best of 10 and Best 10 of 100 methods are identical. Additionally, the Restart method produce similar results. The minor differences in SI-SDR values for the different pairs illustrate the trade-offs in reducing the distance correlation between pairs by modifying the unmixing matrix, affecting all output sources. It is essential to recognise that merely achieving maximum independence in source separation does not guarantee equal pairwise distance correlations. See Table 5.4 to verify the different distance correlation values for each pair of underlying sources. If these sources were independent, one would expect a distance correlation loss value of -2.

Regarding the outputted sources being similar for each repeat of the training, the network weights can be initialised using the eigenvalues associated with the PCA of the input mixtures to improve the robustness of distance correlation without using the Restart algorithm. While independent components have zero covariance, zero covariance alone does not imply independence. This weight initialisation will likely bring the system to a good local minimum. The results are presented in Table 5.5. These values slightly underper-

Input	Sine-Square	Sine-Sawtooth	Square-Sawtooth
Original	-1.963±0.026	-1.981±0.002	-1.969±0.015
Best of 10	-1.997±0.000	-1.984±0.000	-1.999±0.000
Top 10%	-1.997±0.000	-1.984±0.000	-1.999±0.000
Restart	-1.997±0.000	-1.984±0.000	-1.999±0.000

Table 5.4: Means and standard deviations for the ten outputs of the distance correlation broken down by pairs. The Hungarian method is used to find the best output order

Input	PCA eigenvalues	Simulated annealing	Separate optimisation
Sine	19.5±0.1	12.5±8.2	15.3±6.5
Square	29.1±0.2	22.3±4.6	24.4±17.9
Sawtooth	17.5±0.0	11.5±7.1	14.2±8.2

Table 5.5: The means and standard deviations were calculated over ten outputs using the distance correlation method. PCA eigenvalue initialisations, simulated annealing, a probabilistic technique for approximating the global optimum of a function, and the optimisation of distance correlation pairs individually were investigated. This approach was used to determine whether these methods improved the robustness and mean SI-SDR values provided by the distance correlation method.

form compared to the FastICA method for this problem and the global optimum for the distance correlation metric.

Simulated annealing is a technique that identifies global solutions without exploring all possible options. It improves upon the results for distance correlation shown in Table 5.1, although the PCA method remains more robust and yields higher mean values. Additionally, optimising the loss associated with each pair of sources with its own optimiser demonstrated a slight improvement compared to the results presented in Table 5.1.

While these methods help reduce the standard deviation in the outputted SI-SDR values, they do not consistently achieve the global minimum. Therefore, I will proceed with the Restart method to identify the most extreme distance correlation values and best outputs. This will allow me to compare the best results from this approach with those of the baseline FastICA method.

Now, I will revisit the Reconstruction case to enhance its Greedy version for distance correlation. In the Reconstruction case, the loss comparison in Algorithm 1 should concentrate exclusively on the reconstruction component of the loss or give it significant priority. This focus is crucial to ensure the sources combine linearly to produce the mix-

tures. If too much emphasis is placed on independence, sources such as white noise, which cannot combine linearly to create the mixtures, may be mistakenly identified and incorrectly treated as suitable outputs.

The findings from this analysis are presented in Table 5.6. The high SI-SDRs indicate that applying the Restart method primarily to the reconstruction component of the loss can effectively extract suitable sources. Moreover, these SI-SDRs are associated with low distance correlation, through the nature of the Restart Algorithm. However, it is important to emphasise that the reconstruction aspect of the loss is necessary to constrain the possible solutions.

Method	Whitening	Sine	Square	Sawtooth
Reconstruction	Yes	6.9±3.7	14.1±13.5	6.5±4.9
Reconstruction	No	13.2±1.1	31.0±0.5	14.6±1.6

Table 5.6: Mean and standard deviations of the SI-SDRs for the reconstruction method for the 3-mix-3-source example. The Restart method is applied only to the reconstruction element of the loss in this example.

When comparing the results from Table 5.6 for the SI-SDRs obtained using the Reconstruction method, both with and without whitening, to the results for the sine, square, and sawtooth waves presented in Table 5.3, it is evident that the Reconstruction method does not perform as well as the Separation method (in all cases the p-value is less than 0.05). This observation holds even after applying the Restart algorithm to the reconstruction component of the loss function in the Reconstruction method.

Nonetheless, the Reconstruction method shows promise. With some fine-tuning, it could become competitive. Additionally, this method allows for varying numbers of input sources to create output mixtures, suggesting an exciting direction for future research.

5.4.2 GNSS ICA

ICA was applied to separate two sources from two different mixtures in two distinct cases. In the first example, I utilised the close ORWA and P445 stations, which resulted in similar atmospheric components. In contrast, the second example involved the ORWA and GRSV stations, which are farther apart and may exhibit slightly different atmospheric components.

For the geodetic data, I report only the results from the FastICA method, as they were generally more precise and accurate than those produced by the PyFastICA method in previous tests. In all optimisation methods, the learning rate was set to 0.0003. Cholesky whitening was applied for the geodetic data in all instances, except for FastICA, which utilises unit-variance whitening, and the GNSS case of distance correlation, which employed ZCA whitening.

Tables 5.7 and 5.8 depict the average SI-SDR with 5 different epicentres for the ORWA/P445 and ORWA/GRSV examples. In both scenarios, I have structured the problem to become ill-posed for BSS when the epicentre is at the midpoint between the GNSS stations. The problem is ill-posed because the number of sources to extract effectively exceeds the number of known mixtures, which essentially reduces to just one when the scaling factor for the synthetic source becomes equal for each station. Unlike audio source separation, which relies on known training signals, true BSS assumes that the sources are independent without knowing the signals themselves. The further from the midpoint, the closer the output sources are to their ground truths, as indicated by the SI-SDR values increasing away from the midpoint. Additionally, the distances between stations and the atmospheric effects have a minimal impact on SI-SDR values, as the equivalent SI-SDR values for each station pair are similar.

The distance correlation metric is an effective tool for distinguishing between GNSS and synthetic signals, making it valuable for source separation compared to FastICA. While both methods yield similar results when whitening is applied, the Restart method is exclusively used with the distance correlation metric, as it can negatively affect the MINE methods. This issue was highlighted by the negative SI-SDR values observed at each epicentre distance for the synthetic signal (although I have not reported these values here as they are worse than the whitened results). Consequently, using the Restart method and selecting the appropriate whitening technique is essential for achieving optimal source extraction. However, these steps are more challenging to implement and can be more demanding for MINE-based losses, often requiring more trial and error.

Neural estimators like MINE suffer because of the high variance of their exponential variational bound and non-convex loss training. Even after apparent convergence, stochastic mini-batches and small gradients near the optimum cause estimates to oscillate.

Over-parametrised mutual information approximations can exploit this noise, producing spurious fluctuations. Using MINE within a BSS pipeline, freezing the MINE architecture for separation and then unfreezing it, amplifies these instabilities. The frozen phase relies on static, biased MI signals; unfreezing then injects high-variance gradients that can derail the unmixing matrix.

Downstream one may see lower SI-SDR, longer training, and blurred hyperparameter effects (causing challenges in hyperparameter tuning). In many source-separation scenarios, the slight gain from a tighter MI bound is outweighed by training instability. Moreover, in representation learning, it has also been found that a tighter bound on mutual information does not necessarily lead to better learned representations [158].

ORWA/P445	Signal	FastICA	MINE	DistCorr
1/8	GNSS	11.5±0.0	10.5±2.1	11.8±0.0
	Synthetic	6.6±0.0	5.8±1.1	6.6±0.0
1/4	GNSS	9.8±0.0	9.6±0.3	9.8±0.0
	Synthetic	3.3±0.0	2.8±0.9	3.2±0.0
1/2	GNSS	4.0±0.0	-2.4±11.6	3.2±0.0
	Synthetic	-22.4±0.0	-17.8±6.3	-19.0±0.0
3/4	GNSS	9.0±0.0	8.2±1.2	8.4±0.0
	Synthetic	2.4±0.0	2.0±0.6	2.0±0.0
7/8	GNSS	10.5±0.0	9.0±5.4	11.9±0.0
	Synthetic	6.0±0.0	5.0±2.4	5.9±0.0

Table 5.7: The synthetic signal was placed at various distance ratios between the ORWA station and station P445. Average SI-SDR values over ten repetitions.

GRSV/ORWA	Signal	FastICA	MINE	DistCorr
1/8	GNSS	10.1±0.0	10.8±1.0	11.2±0.0
	Synthetic	5.6±0.0	5.2±0.7	5.4±0.0
1/4	GNSS	8.5±0.0	6.4±2.5	7.3±0.1
	Synthetic	2.1±0.0	1.2±1.3	1.6±0.0
1/2	GNSS	3.6±0.0	-1.7±9.8	3.0±0.0
	Synthetic	-20.2±0.0	-17.9±5.9	-17.3±0.0
3/4	GNSS	9.0±0.0	8.3±0.8	8.9±0.0
	Synthetic	2.7±0.0	2.4±0.4	2.7±0.0
7/8	GNSS	10.9±0.0	11.0±0.6	11.2±0.0
	Synthetic	6.0±0.0	5.6±0.4	5.9±0.0

Table 5.8: The synthetic signal was placed at various distance ratios between the GRSV and ORWA stations. Average SI-SDR values over ten repetitions.

The SI-SDR values for the GNSS signals are consistent across all the analysed meth-

ods. The similarity observed between FastICA and MINE can be explained by the following observation: Mutual information, defined as $I(X;Y) = H(Y) - H(Y|X)$, when minimised using the MINE approach, maximises the associated conditional entropy $H(Y|X)$. This process tends to create a conditional distribution that is as close to Gaussian as possible, aligning with the definition of negentropy. Maximising mutual information may not always achieve a Gaussian joint distribution. As a result, there can be differences in the outcomes between MINE and FastICA. Furthermore, while the MINE technique is less robust, its results are still generally within one standard deviation of those produced by the other methods.

5.4.3 SAR ICA

In this section, I will extract two underlying sources from two mixtures, consisting of a linear combination of 100x100 pixel multi-temporal SAR images and a 100x100 pixel section of a picture of a star.

In Table 5.9, we can see that when extracting two sources from two mixtures, both in the Defined and Undefined cases, FastICA, MINE, and DistCorr produce similar SI-SDR values for the SAR and SUN signals. Overall, the FastICA method yields the best results. It is worth noting that the SI-SDR values for SAR and SUN are higher in the Defined case than in the Undefined case when the problem is less ill-posed, and the input mixtures are more distinct. In the temporal example, an ill-posed scenario occurs when the two input mixtures are similar, much like the midpoint epicentre between the two GNSS stations.

In the Undefined case, where in effect two sources are separated from one mixture, the FastICA method produces the best output for SAR. Distance correlation and MINE show comparable results, differing by no more than three standard deviations from the FastICA results.

In the Defined case for MINE, the results showed one significant loss divergence. Removing this output resulted in a mean of 10.1 ± 0.6 for the SAR signal and 12.3 ± 0.8 for the Sun image. Introducing a threshold to halt training in MINE may help mitigate loss divergence. Similarly, in the Undefined case, there was one divergence in the distance correlation case. After removing the data associated with the divergence, the results improved to 4.3 ± 0.3 for the SAR signal and -18.2 ± 0.3 for the synthetic signal. However,

in the latter case, the resulting SI-SDR remains lower than that achieved with the FastICA method. The negentropy-based loss aims to extract the strongest non-Gaussian component in this apparent one-mixture-two-source undefined case, corresponding to the SAR signal. However, complete separation of both sources is not guaranteed. Furthermore, distance correlation does not inherently emphasise non-Gaussianity but rather the linear and non-linear relationships between random variables. This may limit its capacity to identify the most independent source in cases where prioritizing one more distinguishable source is necessary.

	Signal	FastICA	MINE	DistCorr
Defined	SAR	10.8±0.0	8.1±6.1	10.8±0.0
	SUN	13.1±0.0	12.0±1.3	13.0±0.1
Undefined	SAR	5.7±0.0	4.4±0.5	3.6±0.8
	SUN	-40.2±0.1	-17.0±3.5	-16.7±2.1

Table 5.9: Average SI-SDR values over 10 repeats for the spatial ICA case.

5.5 Conclusion

In this chapter, I proposed several metrics and architectures to maximise the independence of BSS using both synthetic and geodetic data, aiming to achieve spatially or temporally independent components.

The baseline FastICA method outperformed the other techniques on the synthetic dataset. The Restart algorithm allowed for optimal performance of the distance correlation metric, which, in turn, surpassed the baseline results on one out of three metrics and outperformed it when assessed on the average SI-SDR across all sources. Regarding the Reconstruction method, a Restart approach, based on minimising the MSE loss, may yield SI-SDR values comparable to those achieved with the Separation method. However, the Reconstruction architecture, with the hyperparameters outlined in this chapter, does not perform as well as the Separation method.

For the hybrid GNSS/synthetic signal, the distance correlation metric extracted sources for each epicentre performed similarly to the FastICA baseline for all points except the midpoint, with some minor issues for ill-posed problems. These findings suggest that the method can differentiate seismic signals from non-seismic ones effectively in most cases.

Similarly, the FastICA, MINE, and DistCorr loss functions yielded similar results for the SAR and SUN outputs in the Defined case for spatial ICA with two mixtures and two sources. As in the GNSS formulation, distance correlation slightly underperforms when the SAR problem is ill-posed.

This proof-of-concept study indicates that minimizing distance correlation for source separation is as good as the baseline methods included in this chapter and may be superior for some geodetic signals. It can accommodate the non-unimodal probability distribution functions commonly found in transient signals, which traditional ICA methods often struggle to handle. This limitation is a key motivation behind the proposal of the vbICA method by [11].

Further development and fine-tuning of the methods described in this chapter would be necessary to handle larger GNSS datasets effectively. My goal, which I will elaborate on in Chapter 6, is to enhance the technique's capability to process pure GNSS data and accurately separate underlying deformation signals from noise within a GNSS time series. This will enable direct comparisons between the neural network-based methods proposed in this chapter and more traditional yet state-of-the-art optimisation methods, such as the vbICA method utilised in [5].

CHAPTER 6

Blind source separation for GNSS data

6.1 Introduction

This study focuses on the blind source separation of GNSS data, which is a less explored area. Separating GNSS data is particularly challenging because establishing a reliable ground truth solution is not straightforward [159].

In this chapter, I propose the use of distance correlation (DistCorr) as a loss function for the source separation of GNSS time series through unsupervised learning. DistCorr's application in the BSS of GNSS data remains unestablished, making my proposal a unique contribution to the field.

Evaluating a specific loss function for a BSS task presents challenges, even when a generally reliable metric for separation quality, such as the SI-SDR, is available. In this case, SI-SDRs typically work well but are not always robust to noise or offsets. However, SI-SDR is a widely used metric for audio source separation due to its interpretability and robustness; therefore, I use it in my work.

Unlike standard error metrics commonly used in classification problems like HTER, SI-SDR is not a linear measure. Therefore, the reader should assess the mean averages with caution. Additionally, I want to emphasise that when I report the best SI-SDR values

for a metric, I refer to the SI-SDRs associated with the best loss function values rather than the highest SI-SDR values. To address the potentially poor optimisation of an independence metric, the Restart algorithm, as introduced in Algorithm 1 in Chapter 5, is utilised again in this chapter. For a detailed analysis of source separation through optimisation of distance correlation for the synthetic problem, I encourage the reader to refer to Appendix B.

The main contribution of this chapter are summarised as follows:

- Testing BSS methods on GNSS data to distinguish between seismic and non-seismic time series, and to separate various geodetic signals in the Southern California region.

As noted in Section 5.1, the primary goal is to assess the effectiveness of distance correlation for blind source separation. This includes examining its optimal values at the extremes and its efficiency when trained using gradient descent, particularly here in the context of GNSS data.

6.2 Experimental set-up and test data

6.2.1 Two mixture problem

In this chapter, I discuss source extraction using the Separation architecture. Figure 6.1 illustrates the simple neural network employed to evaluate the effectiveness of the loss functions. Notably, the linear layer in the encoder restricts the source separation to situations where the number of mixtures equals the number of sources.

I utilised a linear layer to extract the sources, which were subsequently ZCA-whitened and normalised. I computed the average loss for each pair of sources, whether that loss was distance correlation or a negentropy-based approach. This loss was then optimised using gradient descent for PyFastICA and the Restart method, as described in Algorithm 1, for DistCorr.

I selected two sets of Midas detrended GNSS easterly time series from Japan, covering the period from 2010 to 2012. This timeframe includes a significant step function in the GNSS time series that corresponds to the earthquake on March 11, 2011.

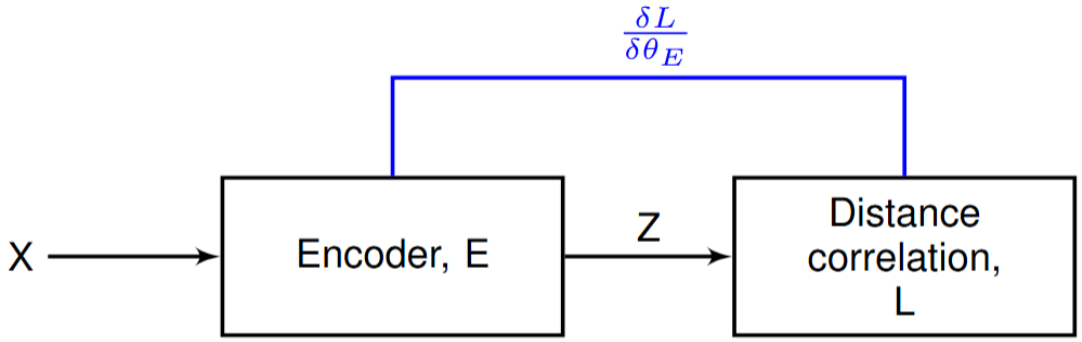


Figure 6.1: A neural network learns statistically independent outputs through a distance correlation loss and optimising an encoder E parametrised by θ_E.

I used stations J076 and G119 to analyse nearby stations. They are 1.34 km apart near Mount Koyama. For sites that are further apart, I selected stations J076 and G025, which are 18.32 km apart. To address any missing data, I applied linear 1D interpolation.

This problem is presented as a two-mixture-two-source case with the aim of extracting both seismic and non-seismic signals despite the absence of ground truth data.

The GNSS signal is decomposed into three additive components: *trend*, *seasonality*, and *residual* to identify a proxy for the underlying seismic signal. This decomposition is performed using the seasonal decomposition function from the statsmodels library. In the additive model, the time series can be expressed as $X(t) = T(t) + S(t) + e(t)$, where $T(t)$ represents the trend, $S(t)$ denotes the seasonal component, and $e(t)$ represents the residuals.

A convolution filter is first applied to extract the trend from the data. In this case, the series is processed using a simple centred moving average filter that has a length equal to the seasonal period, applying uniform weighting across the data. The edge values are dropped rather than using padding. However, the fixed-width moving average filter assumes constant seasonality and can under- or over-smooth if the cycle drifts. After the trend has been removed from the time series, the average of the detrended series is calculated for each user-defined period. Once the seasonal component is removed, the remaining data corresponds to the residuals.

To create the ground truth signal for comparison, I assumed it was a 2-mix, 2-source problem, with one seismic signal that contained a step function and another non-seismic signal that was common to both stations within the pair. I also assumed that the time

series was stationary over the time frame. This is important as BSS algorithms rely on the stability of the statistical properties of the sources over time. However, the trend estimated by seasonal decomposition may violate stationarity when a step is present.

To establish a shared seismic signal, I performed a seasonal decomposition on each GNSS series in the pairs (J076/G119 or J076/G025) using an additive model with a period of 2, which corresponds to a biannual period. I then analysed the trend from each time series, which included the step function, to define the seismic signal. However, when I averaged these two signals, after centring and standardising them, to determine the ground truth, the SI-SDR was lower than in previous experiments. The results improved when I incorporated an element from the residual data (spanning 3 days before and after the earthquake) that appeared to originate from seismic activity. Consequently, for each case, I took the average of the two sums of the trend and the residual around the earthquake as the target signal.

I have introduced several biases due to my choices. For example, I selected a biannual period because there is a low-frequency oscillation with a biannual cycle. Additionally, the residuals I chose, which correspond to the days around the earthquake, appear to be associated with inhomogeneous afterslip. While other time intervals may contain seismic information, they were less noticeable and therefore missed due to my biased selection.

Using a more geological lense, the trend serves as the main component for the ground-truth seismic signal, along with an additional element derived from the residual observed around March 11, 2011, as depicted in Figure 6.2. The selected section of the residual appears to correspond with models of afterslip. Afterslip refers to the slipping that occurs during aftershocks following a significant earthquake, characterised by a gradual trough followed by a peak after the main shock event. For further clarification, I recommend the paper by [160], which includes figures that illustrate the concept of afterslip. It is important to note that my methodology is subjective. My methodology drew inspiration from the work of [5], which applied vbICA to deformation data from the GRACE satellites to identify seasonal components in Southern California.

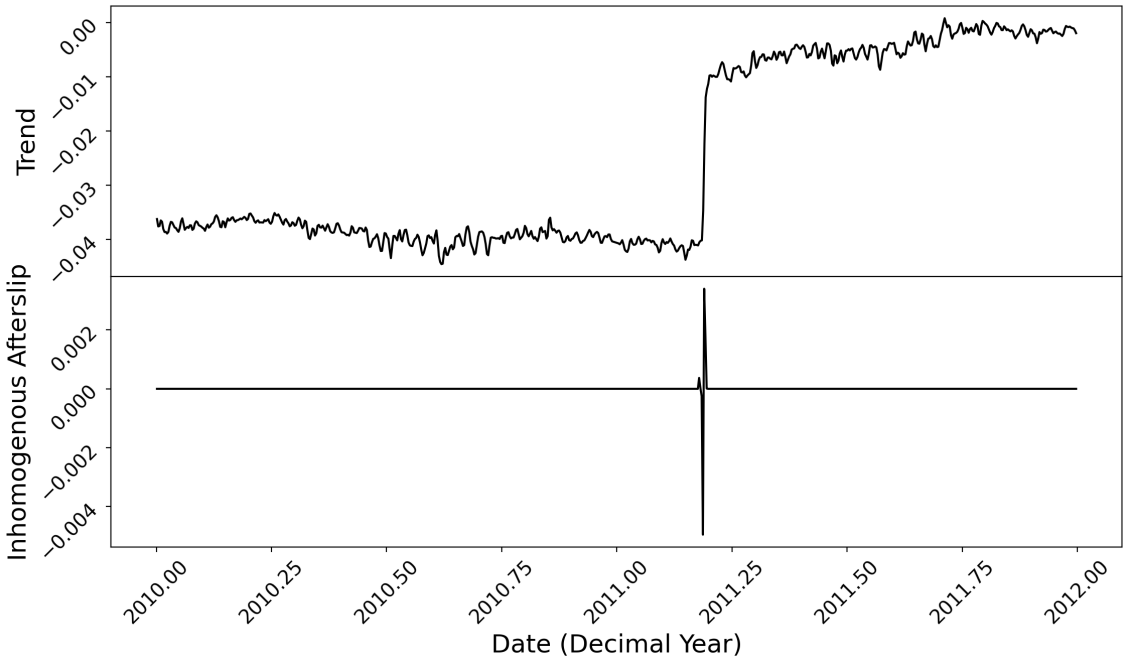


Figure 6.2: Underlying elements taken to be the seismic signal. The decomposition of J076 by (using statsmodels seasonal_decomposition) following an altered version of [5]. The trend and manually selected element of the residual representing inhomogenous afterslip are summed to form the 'ground truth'. The plots are displayed on different scales for better readability.

Results

As shown in Figure 6.2, the trend is the main component of the ground truth for the seismic signal, with an added component from the residual around the 11th of March 2011. The residual component was subjectively identified when I examined the seasonality and residual signals around the time of the mainshock, indicated by the step function in the trend. Around the time of the mainshock, I located a shape in the residual that may correspond to afterslip.

Table 6.1 presents the SI-SDR values comparing the estimated underlying seismic source to its ground truth, estimated using a seasonal decomposition of the GNSS data. Among the evaluated methods, PyFastICA is the least robust, as it underperformed compared to FastICA and DistCorr in terms of mean values, except for the G119 Restart case, where it outperformed the distance correlation method. The distance correlation method showed strong performance, producing means comparable to the baseline FastICA, al-

though it fell slightly short, particularly in the J076/G119 case. This underperformance in the closer case may indicate that distance correlation methods struggle compared to negentropy-based approaches when dealing with ill-posed problems, as discussed in the previous chapter. However, the slightly higher SI-SDR values seen in the J076/G025 case may result from a slight offset between the GNSS signals, leading to more distinct mixtures. Figure 6.3 shows this offset.

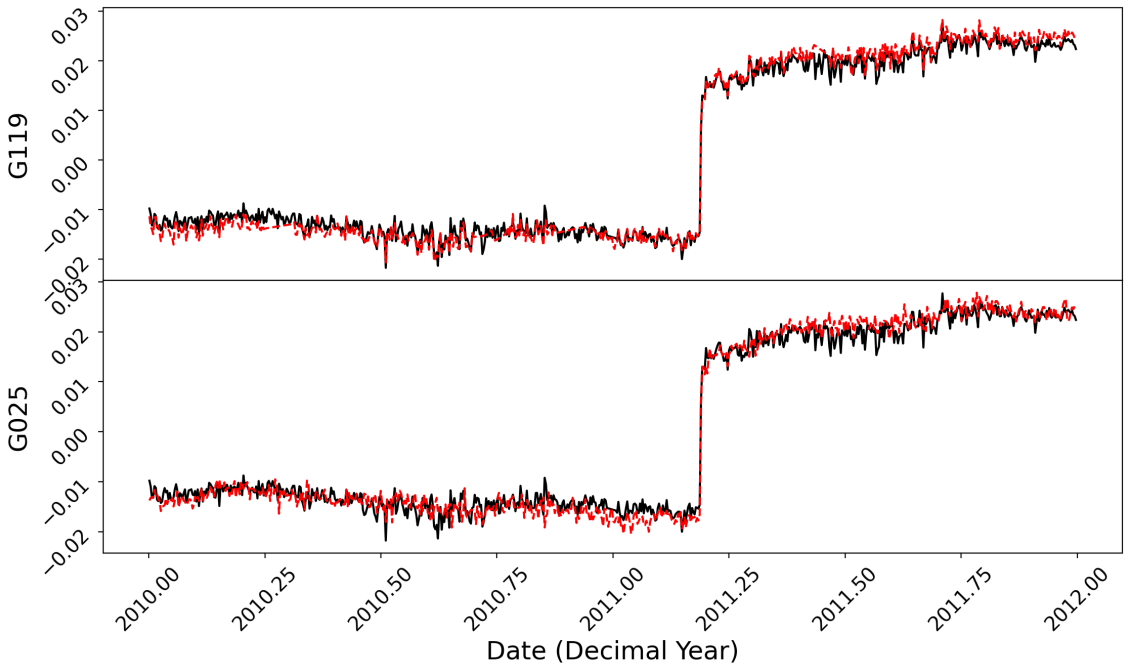


Figure 6.3: MIDAS detrended easterly GNSS time series for the J076 (black), G119 (red upper) and G025 (red lower) sites. The time series have been centred for comparison purposes.

It is important to note that the actual SI-SDR values may vary slightly due to the assumptions made while producing the ground-truth seismic signal for the SI-SDR computations.

Referring to Table 6.1, for the case documented as G119, the Restart method performed worse than the Original method when using distance correlation. I found that reducing the distance correlation past a certain low threshold, to what seems to be the best local minimum, resulted in a slight decrease in SI-SDR, which explains this underperformance.

The corresponding waveforms of seismic and non-seismic signals for the J076/G119

Method	Input	FastICA	PyFastICA	DistCorr
Original	G119	28.2±0.0	19.3±9.1	26.0±0.0
Original	G025	30.0±0.0	17.4±8.1	29.9±0.0
Best of 10	G119	28.2±0.0	20.8±3.8	25.7±0.2
Best of 10	G025	30.0±0.0	22.8±5.6	29.9±0.0
Top 10%	G119	28.2±0.0	22.2±3.7	25.6±0.1
Top 10%	G025	30.0±0.0	24.8±5.3	29.8±0.0
Restart	G119	28.2±0.0	27.3±1.8	25.8±0.2
Restart	G025	30.0±0.0	23.7±1.4	29.9±0.0

Table 6.1: Means and standard deviations of SI-SDRs between the average ground truth of the J076 and second site, produced by the by trend plus afterslip. The J076/G119 and J076/G025 are the closer and further sites, respectively.

Method	Input	PyFastICA	DistCorr
Pearson	G119	0.686	0.972
Pearson	G025	0.775	0.931

Table 6.2: The Pearson correlation between the final epoch loss and the average SI-SDR of the outputted and ground truth signals, calculated over 100 repetitions.

and J076/G025 examples can be seen in Figures 6.4 and 6.5, respectively, for the distance correlation, FastICA and PyFastICA methods.

As shown in Table 6.2, the Pearson correlation between the distance correlation losses and the SI-SDR values between the outputted and ground truth sources exceeded 0.9 in the cases of G119 and G025. This strong correlation suggests that the Restart algorithm would likely provide higher final SI-SDR values in most instances when compared to the Original method, though the Original method performed well in these test cases. Additionally, distance correlation is shown to be an effective and explainable approach for source separation.

In Table 6.1, the results from PyFastICA for the G119 case outperform those from distance correlation in the Restart method. However, the lower Pearson correlation shown in Table 6.2 indicates that this discrepancy is likely a result of variability in the runs themselves, rather than the Restart method consistently identifying good local minima.

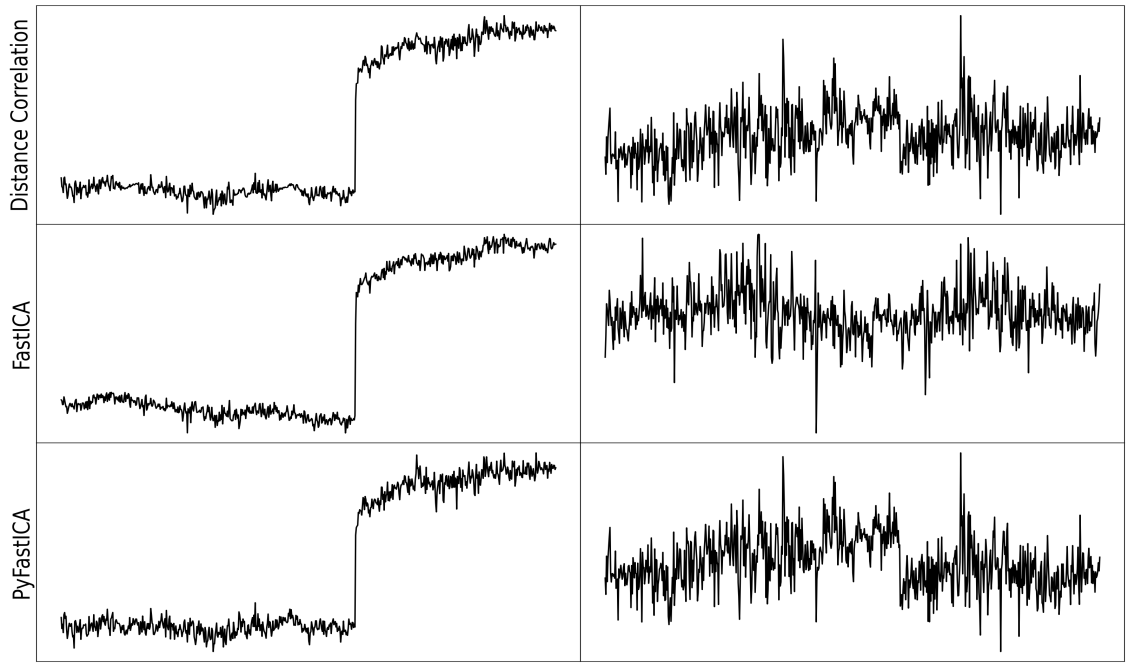


Figure 6.4: Seismic (left) and non-seismic (right) time series extracted using FastICA, PyFastICA, and distance correlation methods with a linear layer for the latter two methods for the J076/G119 station case.

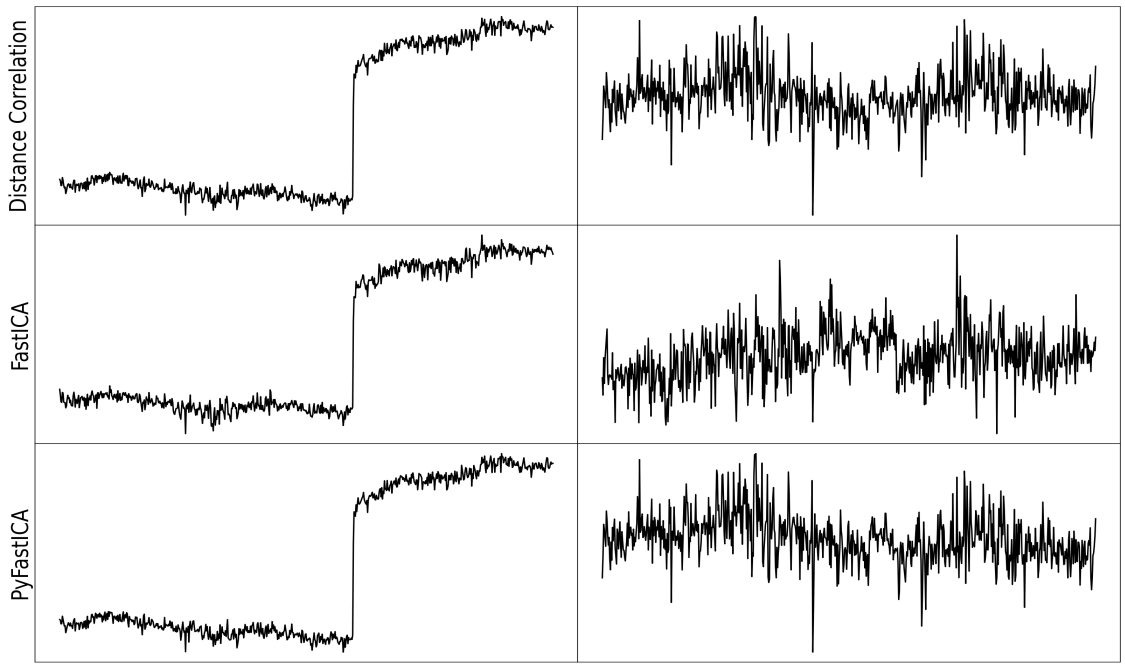


Figure 6.5: Seismic (left) and non-seismic (right) time series extracted using FastICA, PyFastICA, and distance correlation methods with a linear layer for the latter two methods for the J076/G025 station case.

6.3 BSS of GNSS data in SoCal

Since GNSS problems lack ground truths, the best Reconstruction architecture identified for the synthetic dataset in Chapter 5, using a distance correlation-based loss, will be compared with the sources produced by the Variational Bayesian ICA method. I used 40 IGS14 GNSS stations with three components, representing the east, north, and up directions, with latitudes between 32 and 36.5 and longitudes between -112 and -118. The longitudes and latitudes align with the case study provided with the code for vbICA, and contain a subset of the 125 stations used in [3].

In my work, I have selected the number of sources to extract based on a method I introduced in Appendix C, which is based on the percentage of variance explained by each principal components in PCA, but adapted for Independent Components. This method identifies directions that describe the highest distance variance instead of variance. The extraction of the sources using an ICA version of PCA is also described in this Appendix. However, a limitation of this approach is that it results in a higher-dimensional distance space, necessitating a decision on how to reduce dimensions. This is particularly important when ensuring, for example, that the linearity of unmixing is maintained, as the vectors representing the distances are not unique.

Consequently, I concluded that while this method is interesting, using distance correlation as a loss function is superior, as a neural network can be constrained by prior knowledge. The source separation by minimising distance correlation is unique within the limitations of ICA for the synthetic problem (see Appendix B).

Nevertheless, the PCA-ICA method has advantages, as it offers insight into the number of sources a user might want to consider. When a user analyses the cumulative percentage of distance variance explained by various sources, they will receive an indication of how many Independent Components are necessary to achieve a specified level of independence.

In [3], the daily displacement-time series used were generated by the Jet Propulsion Laboratory. These time series were cleaned and MIDAS detrended, with corrections for long-term linear trend and event offsets, both seismic and otherwise, using PCA around the offset to determine the step. The number of ICs, 12, was selected using the automatic relevance determination (ARD) approach proposed for a Bayesian framework in [56].

As vbICA is based on finding the best weights for given data, different pre-processing steps on the inputted GNSS data can provide vastly different underlying independent components. Therefore, I will use the MIDAS detrended GNSS signals for the same region as the case study in the IGS14 reference frame, including event steps and outliers. The outliers present in the GNSS dataset are extreme data points that do not accurately represent the dataset's probability distribution function. These outliers enable the testing of the robustness of various methods for extreme data points. I have not removed the step functions to assess how effectively seismic signals can be distinguished from other underlying sources within the GNSS displacement-time series. Additionally, I aim to determine whether these signals can be classified solely as seismic signals. If successful, step functions may enable clearer labelling of signals as originating from seismic activity, potentially enhancing the scalability of GNSS data in both semi-supervised and supervised machine learning contexts.

I used GNSS data from the University of Nevada Geodetic Laboratory to maintain consistency with my previous GNSS work that employed distance correlation, as seen in this chapter and Chapter 5. As mentioned previously, the pre-processing in this case has been limited to MIDAS detrending. Figure 4.1 showed the cumulative distance covariance explained by the number of independent components. In PCA, the percentage of variance explained helps users decide how many sources to extract, relying on eigenvalues that indicate the amount of variance associated with each eigenvector. Similarly, in ICA, distance covariance can be used. Instead of the covariance matrix, the squared distance covariance eigenvalues help identify both linear and non-linear relationships. The eigenvalues represent the amount of distance variance explained by the eigenvectors, indicating how many sources to retain. I have selected 10 as the number of ICs to extract because it accounts for approximately 60% of the explained distance variance. Increasing the number of components beyond this point only marginally affects the explained percentage. It is worth noting that in [3], 12 sources were extracted.

As seen in the previous sections, architectures and whitening, as well as the inputted data, can alter the outputted independent components of the GNSS signal. The Reconstruction method, both with and without whitening, offers flexibility in selecting the number of mixtures and sources. Thus, the Reconstruction method has been chosen to address

the GNSS problem.

One of the seismic components, IC9 from [3] or the Brawley swarm deformation, is not extracted in this case study by the vbICA or distance correlation method. An earthquake swarm is a sequence of seismic events that occur in a small region, such as near Brawley in California, which do not follow a mainshock-aftershock sequence. The swarm signal component is prominent in both stations, P506 and P499. However, only P499 is included in this case study. As a result, the swarm signal is absent from the source separation process. This observation reveals a limitation in the data used for the case study and emphasises the importance of the input data for analysing seismic activity. Having a higher density of stations used within the Brawley region is likely to produce more ICs representing the seismic processes in this region. It is important to note that the three seismic sources extracted from the JPL case study data for the L=12 scenario are also extracted in the L=10 case. This suggests that the reduction in the number of sources is unlikely to significantly and negatively affect the results presented in this chapter.

To understand Figure 6.6 and subsequent source figures, it is important to note that they consist of two components, with the temporal component located on the top and the spatial extent displayed below. On the top, you will find a normalised time series plotted over the years, with blue dashed vertical lines indicating the dates of known earthquakes. The bottom part of the figure shows the latitude and longitude of known stations, along with the direction and magnitude of each source.

Because I scale each source to unit variance, the y-axis is unitless, and the mixing coefficients carry all amplitude information. Moreover, the overall sign is arbitrary. The map in the bottom shows the mixing weights reshaped into a three-component vector (w_E,w_N,w_U) at each of the 40 stations' latitudes / longitudes. Horizontal movements (east/north) appear as arrows whose orientation and length encode (w_E,w_N); the Up component (w_U) is shown via a colour bar. Together, these vectors depict how strongly and in which direction each GNSS station contributes to that source. These figures were created using GeoPandas. Note that for the uncertainties associated with the time series, both the residual-based standard error estimator, I used for the distance correlation methods, and the variational Bayesian ICA approximation treat the inferred mixing/unmixing parameters as fixed, attributing all variability to residual noise or the mean-field posterior respec-

tively. As a result, they both systematically underestimate the true uncertainty of source estimates. To address this, one should run the algorithm using various initial parameters to discover alternative local minima and evaluate the robustness of the results. This approach acknowledges that each extraction relies on hidden, unenumerated optimisation choices. The outputs from the distance correlation method included in this document reflect the best gradient-descent optimisation achieved through multiple random initialisations. However, increased runs may enhance our understanding of the uncertainties and minima involved, and potentially lead to a better optimised solution.

Figure 6.6 illustrates three sources that may have a geodetic origin, generated using the vbICA method applied to JPL data. Figures 6.7 and 6.8 show four sources with potential geodetic origins identified through the use of the vbICA method on UNR data. Figures 6.11 to 6.15 present the results from the use of the Reconstruction method, applied to the UNR data. I employed a distance correlation loss and the sources were initialised randomly. Meanwhile, Figures 6.16 to 6.19 display the results of the Reconstruction method using distance correlation, but with principal components used for source initialisation. I set up each method to produce ten sources.

Moreover, combining different initialisation schemes, PCA, random or others, into an ensemble yields multiple local minima and diverse source extractions. This diversity is valuable in applications like hazard prediction or weather forecasting, since it generates a spread of plausible scenarios, but it does not guarantee the single best separation. Ensemble aggregation reduces variance and often improves stability and generalisation, even if it sacrifices the most independent or interpretable components.

Bootstrapping time-series data by sampling with replacement creates several training sets; you train a separate model on each, then average or vote to produce a final prediction. This approach mitigates overfitting and lets you derive confidence intervals, but it will not smooth out a non-convex loss landscape. One can tune voting weights, using, for example, distance correlation, to favour the most reliable models. However, those weights hinge on how well each segment's statistic reflects the full signal: a brief burst of correlated noise can exhibit artificially high distance correlation compared to the overall series, become overweighted, and thus skew the ensemble's results.

Neither ensemble modelling nor bootstrapping changes the underlying gradient-based

optimisation: every base learner still relies on gradient descent.

In this case of the Reconstruction method, I applied only the Reconstruction loss for the first 50,000 epochs out of a total of 500,000 epochs and then introduced the distance correlation element. Given the large number of computations required for calculating pairs of distance correlations, I computed the average distance correlation for a random sample of seven pairs of sources every epoch.

From Figure 6.8, it is evident that Source 6 exhibits a step function corresponding to the magnitude 5.4 earthquake that occurred on August 26, 2012. This source could illustrate the relative fault movement that triggered the earthquake. The map of the epicentre, shown in Figure 6.9, provides additional context. However, while Source 6 is included in the IGS14 GNSS dataset, it is absent from the cleaned JPL data. This absence suggests that Source 6 represents an offset removed during the JPL data cleaning process, as discussed in Section 6.3. Similarly, Source 2 also displays a significant step function at the time of the 2012 earthquake.

In the UNR case, the lack of outlier removal has posed a challenge for the vbICA model, as its parameters are tailored to specific data sets. However, this situation has also highlighted the sensitivity of vbICA, as demonstrated by Source 1 in Figure 6.10. The outlier in the time series suggests that the data should be cleaned. The stations with the highest spatial extent are linked to those with GNSS time series that contain the outlier, which involved two main stations. The findings related to the outlier highlight the increased necessity for data cleaning in the vbICA method compared to the Reconstruction method.

For the distance correlation case, I chose to ignore the noise element of the signal, instead of learning white noise as in the case of vbICA, and assumed it to be explained by an error in the reconstruction of each mixture. For the random initialisation case, Figures 6.11 and 6.12 show two outputs from the distance correlation run, potentially representing afterslip and viscoelastic post-seismic deformation, respectively.

In my work, I used two source initialisation methods. First, I applied the first n principal components from PCA to capture the most variance, since they are uncorrelated. Second, I initialised data randomly to explore a larger search space and potentially avoid suboptimal training.

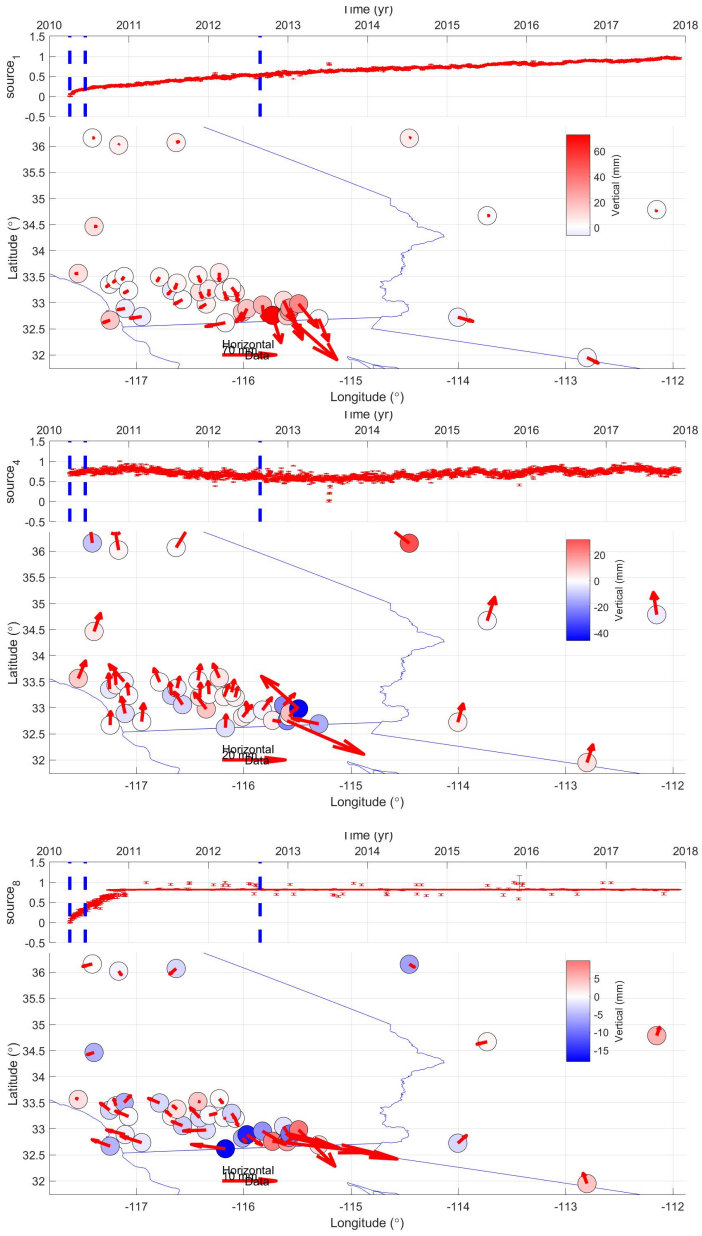


Figure 6.6: Of the 10 outputted sources, both the temporal and spatial components, from the vbICA method using the data provided by Gualandi from JPL, 3 sources of potential seismic origin are shown that correspond to those from [3]. The sparsity of the data in the test case likely has led to one of the sources not being picked up from [3].

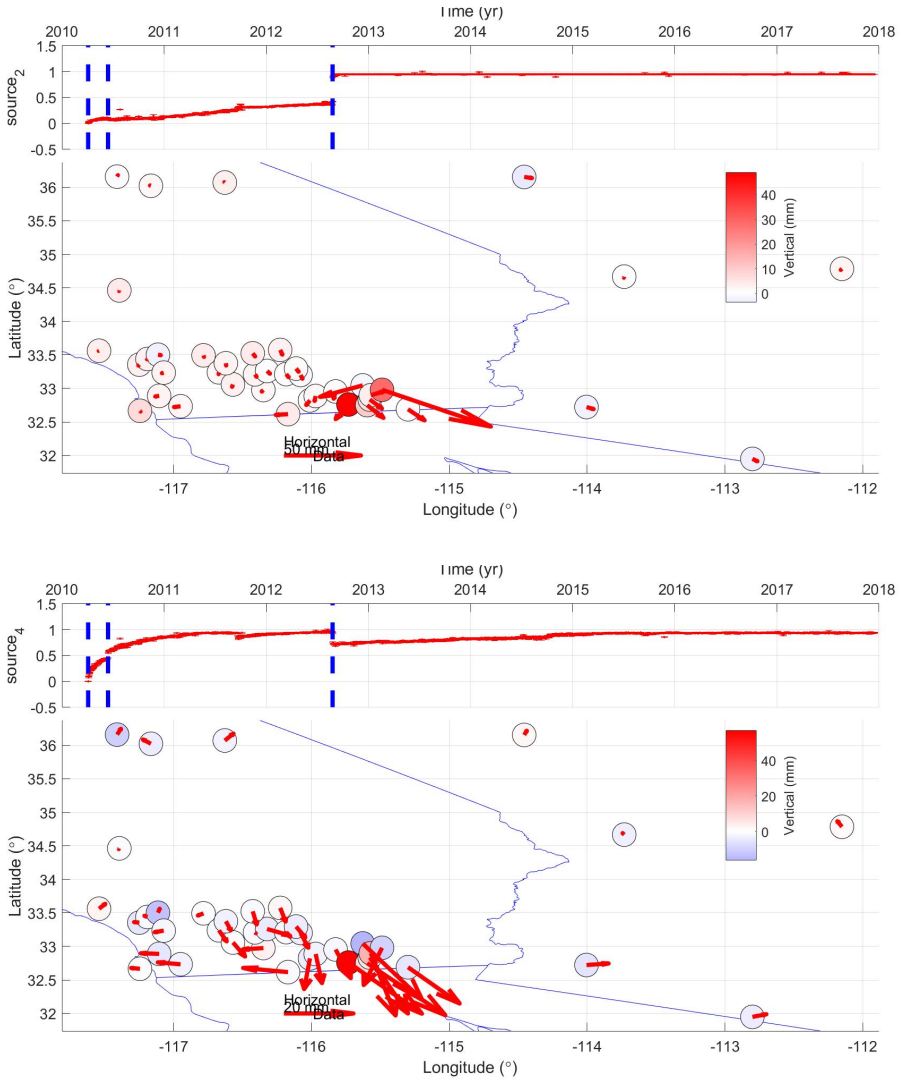


Figure 6.7: Of the 10 outputted sources, both the temporal and spatial components, from the vbICA method using IGS14 GNSS data provided by the University of Reno Nevada, the first 2 of 4 sources of potential seismic origin are shown. The only processing step for the GNSS data was being MIDAS detrended.

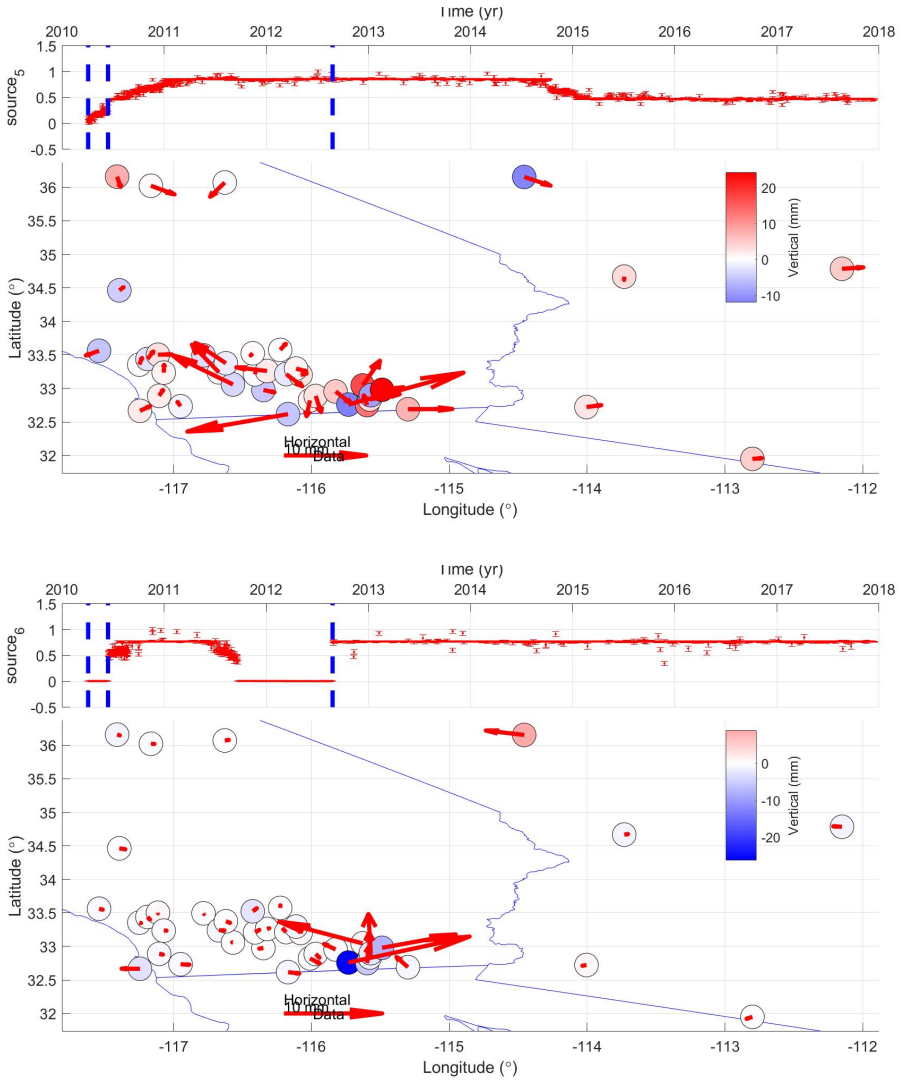


Figure 6.8: Of the 10 outputted sources, both the temporal and spatial components, from the vbICA method using IGS14 GNSS data provided by the University of Reno Nevada, the second 2 of 4 sources of potential seismic origin are shown. The only processing step for the GNSS data was being MIDAS detrended.

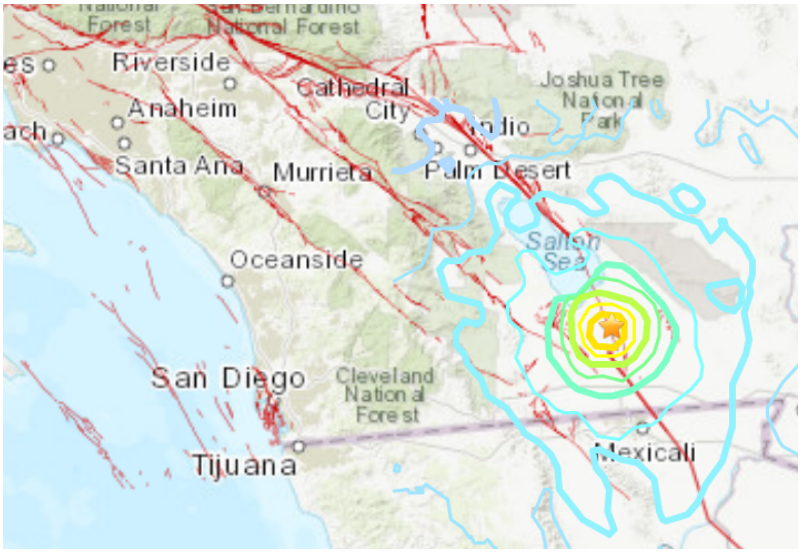


Figure 6.9: Map of the area surrounding the 2012 Brawley earthquake's epicenter (star). The contours indicate the Modified Mercalli Intensity and the red lines indicate US Faults and Tectonic Plates. Data was provided by the Caltech/USGS Southern California Seismic Network (SCSN), doi:10.7914/SN/CI, operated by the Caltech Seismological Laboratory and USGS, which is archived at the Southern California Earthquake Data Center (SCEDC), doi:10.7909/C3WD3xH1

Combining these initialisation strategies or others through an ensemble method with various base and weak learners can yield different minima and varied source extractions. Ensemble modelling with different initial parameters is helpful for hazard predictions and weather forecasts, as it provides diverse scenario probabilities, though not guaranteed robustness. While ensemble methods can improve robustness by aggregating predictions and reducing variance, their aim is not always to find the single best separation but rather a more stable and generalised solution, which may not be the most independent or meaningful sources.

Additionally, bootstrapping can create different batches in time series data, enhancing training and reducing overfitting while offering confidence intervals. However, it may not address the complex loss landscape effectively. Bootstrapping generates multiple training datasets by randomly sampling the original dataset with replacement; the neural network is then trained independently for each of the bootstrapped datasets, and finally, the results are averaged or voted on for a final, more stable prediction. However, it may be possible to determine the best weighting parameters for voting. The bootstrap method with the lowest distance correlation could be sufficient. However, the dependence on these parameters

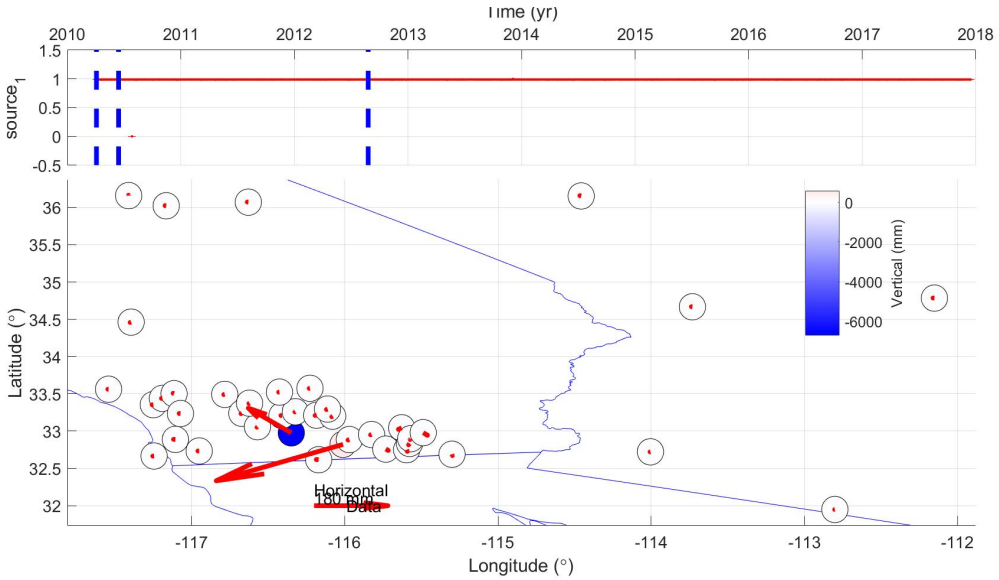


Figure 6.10: In the case of vbICA, its first source, when the data is not cleaned, is sensitive to outliers.

may vary based on the content of that segment. Therefore, one must make assumptions regarding whether it is representative of the overall joint distribution of each pair within the signal.

In Figure 6.7, the components more clearly identifiable as products of geodetic processes in the vbICA case exhibit a step function related to the 2012 earthquake. However, this does not apply to the distance correlation case, as demonstrated in Figures 6.13, 6.14, and 6.15. Keep in mind that such offsets are absent in the JPL data due to the offset removal process.

Both Figures 6.14 and 6.15 contain steps and produce a spatial representation of the 2012 Earthquake. However, they do not have an obvious sole geological meaning, with the latter potentially even having a tropospheric element.

Repeating the Reconstruction method but initializing with the top 10 PCs produced sources that could be representative of viscoelastic post-seismic deformation, post-seismic deformation, afterslip, and Brawley swarm deformation in Figures 6.16, 6.17, 6.18 and 6.19, respectively.

In the Reconstruction case with PC initialisation, the step functions are found in sources that may represent seismic signals. However, similar to the case of vbICA, these step functions are not present in just one of the extracted sources.

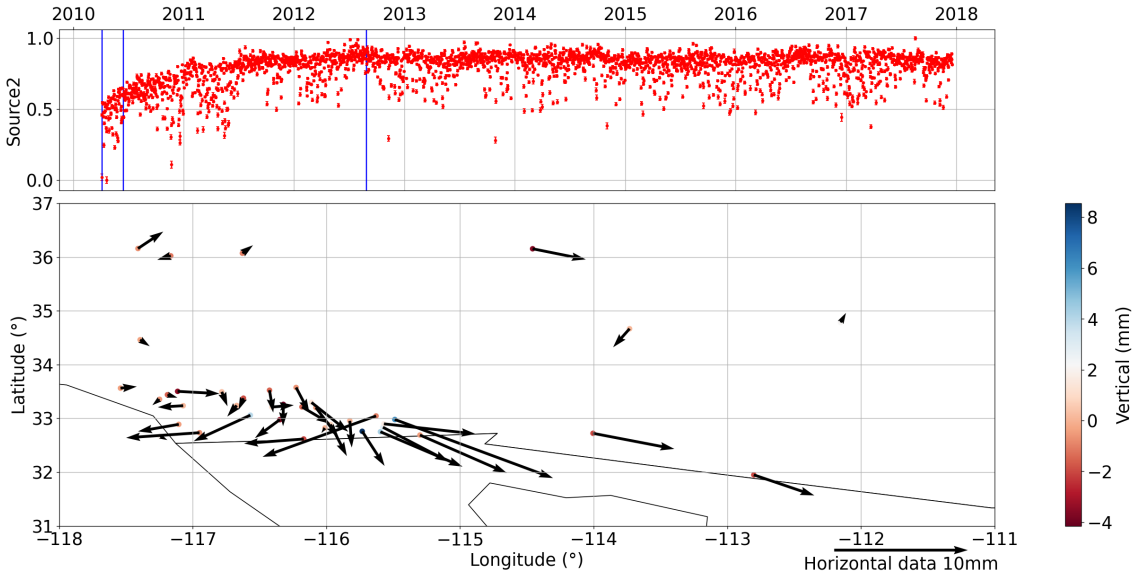


Figure 6.11: The second source of ten extracted from 40 GNSS stations in the SoCal region, with random source initialisation. This source appears to be representative of afterslip.

The task of distinguishing a meaningful seismic or geodetic signal from a noise signal is more successful in Section 5.4.2 than attempting to isolate all underlying signals with the anticipation that one would represent a seismic signal. The seismic step function has been observed in various sources, making extracting a single earthquake source that includes the mainshock for straightforward dataset labelling impractical. As a result, automating the labelling of these sources is unfeasible using the method outlined in this chapter.

6.4 Conclusion

In this chapter, I evaluated the effectiveness of distance correlation for the BSS of GNSS time series, using real data without synthetic known signals, to obtain temporally independent components.

When non-parametric PCA-ICA was introduced, it did not perform as well as the Restart method because it did not strictly enforce linear separation, as a linear layer encoder does. Therefore, its results can be found in Appendix C. However, PCA-ICA provided a technique that could help users determine how many sources to extract.

I compared FastICA, PyFastICA, and distance correlation methods for extracting seis-

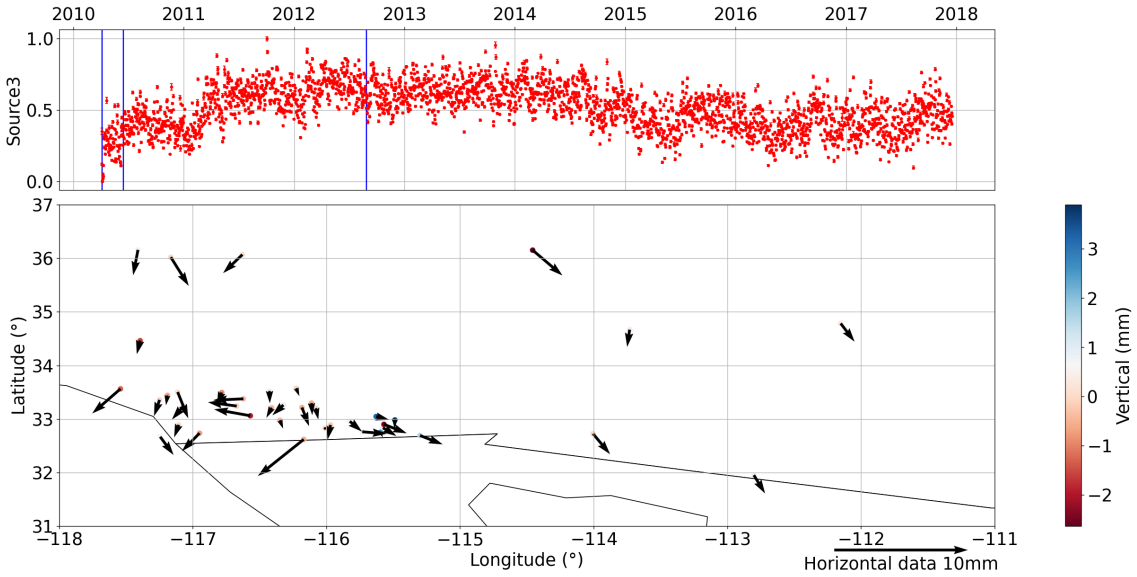


Figure 6.12: The second source of ten extracted from 40 GNSS stations in the SoCal region, with random source initialisation. This source potentially could represent viscoelastic post-seismic deformation.

mic and non-seismic signals from two pairs of GNSS signals, one set positioned closer together and the other set farther apart. The distance correlation method was able to extract a decent seismic signal. Still, it slightly underperformed compared to the ground truth provided by FastICA, particularly in the case of the closer J076/G119 pair. While distance correlation demonstrates competitive potential, further research is needed to assess its performance against FastICA on GNSS data thoroughly. Since the ground truth is synthesised, readers should interpret these findings cautiously. However, distance correlation showed poorer performance in the closer station example, consistent with the results from the previous chapter.

Additionally, I utilised the Reconstruction method with a distance correlation loss and the vbICA method on the UNR data to extract 10 sources from 120 GNSS time series. In both scenarios, the step function associated with the 2012 earthquake appeared in multiple independent components. Thus, while it is possible that I extracted geodetic signals, the use of the step function to potentially identify an earthquake source was not successful.

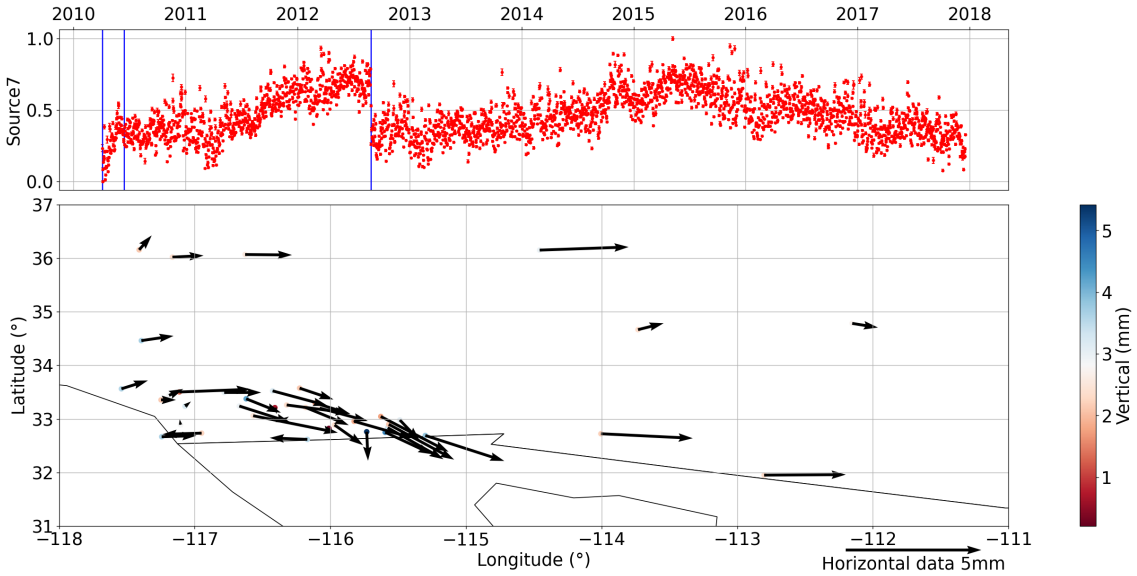


Figure 6.13: The seventh source of ten extracted from 40 GNSS stations in the SoCal region, with random source initialisation. This source could partly be representative of the Brawley storm deformation.

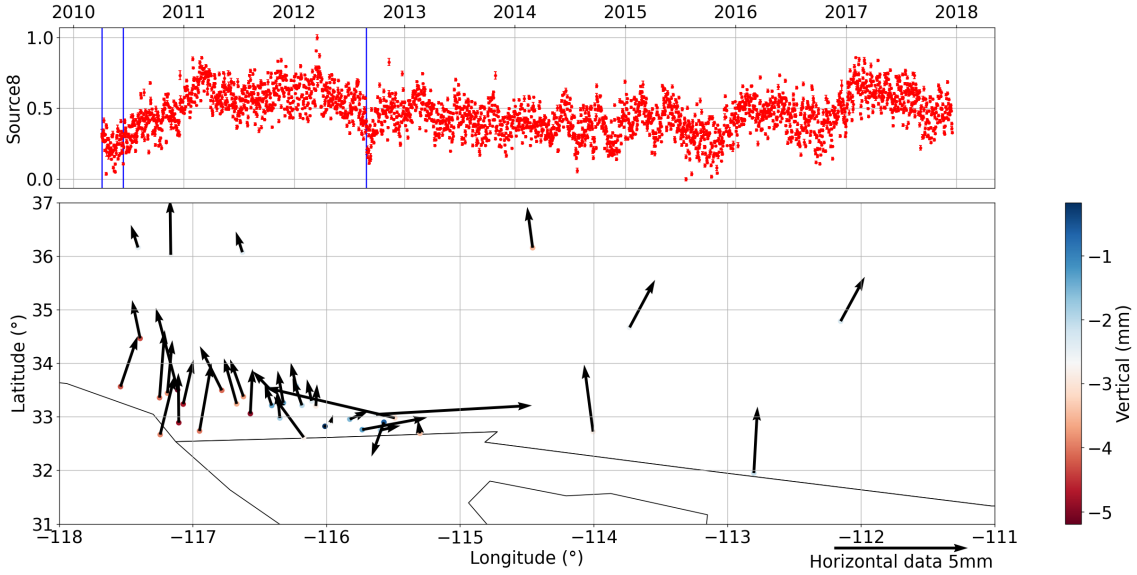


Figure 6.14: The eighth source of ten extracted from 40 GNSS stations in the SoCal region, with random source initialisation. This source has elements similar to post-seismic deformation.

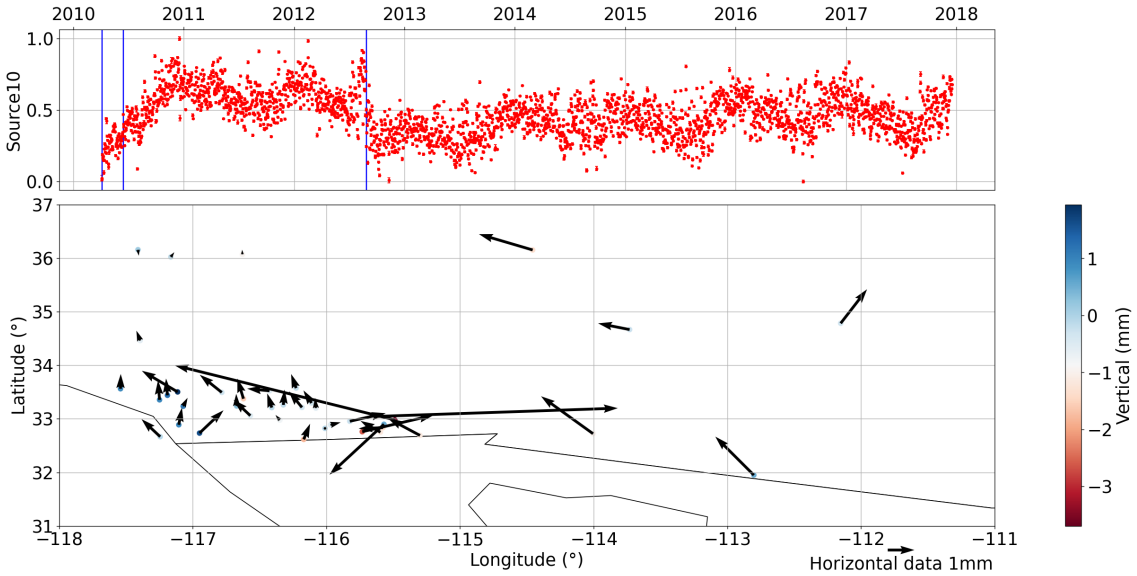


Figure 6.15: The tenth source of ten extracted from 40 GNSS stations in the SoCal region, with random source initialisation. This source has elements similar to post-seismic deformation, along with an element with an annual periodicity.

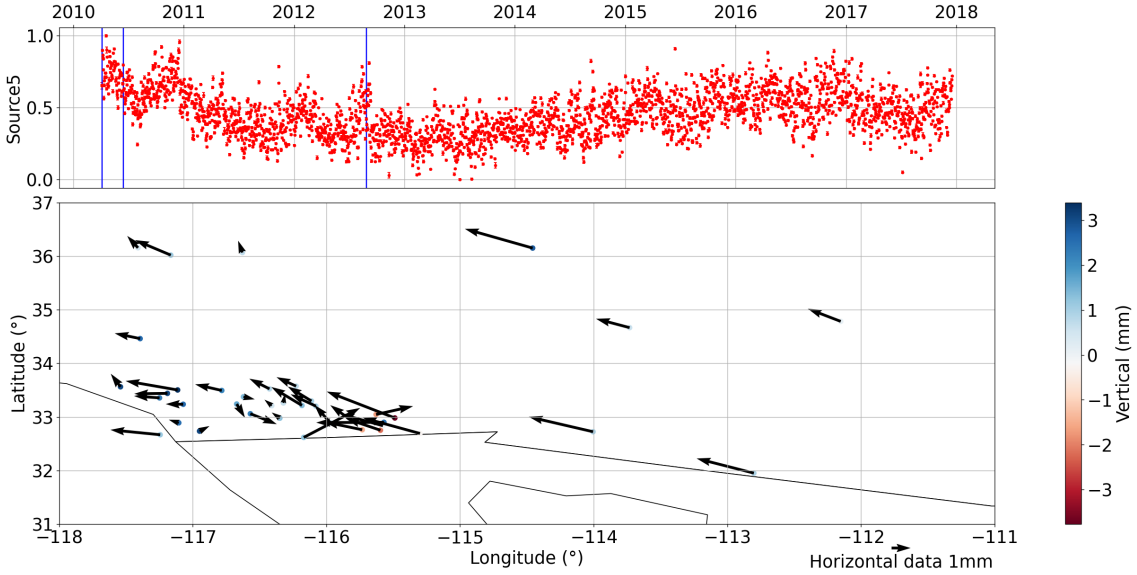


Figure 6.16: The fifth source of ten extracted from 40 GNSS stations in the SoCal region, with PC source initialisation. This source potentially could represent viscoelastic post-seismic deformation.

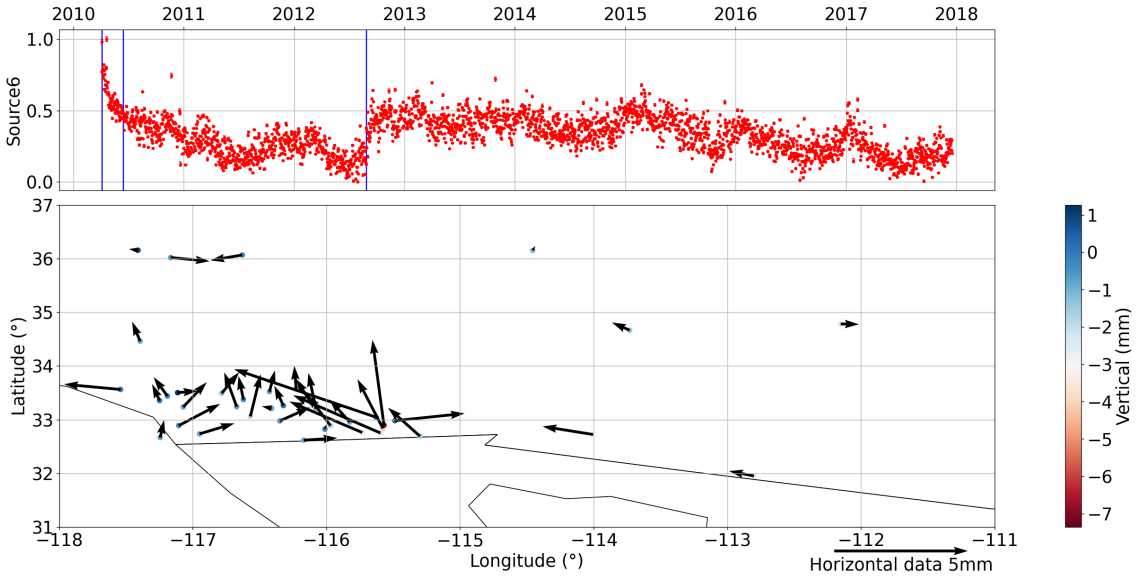


Figure 6.17: The sixth source of ten extracted from 40 GNSS stations in the SoCal region, with PC source initialisation. This source could partly represent post-seismic deformation.

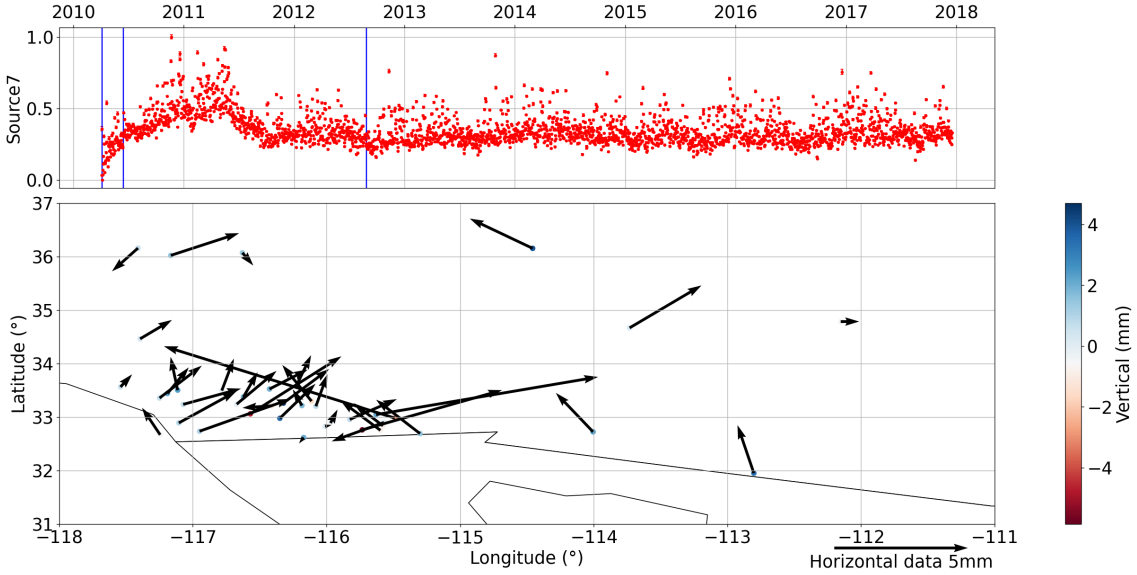


Figure 6.18: The seventh source of ten extracted from 40 GNSS stations in the SoCal region, with PC source initialisation. This source could represent afterslip.

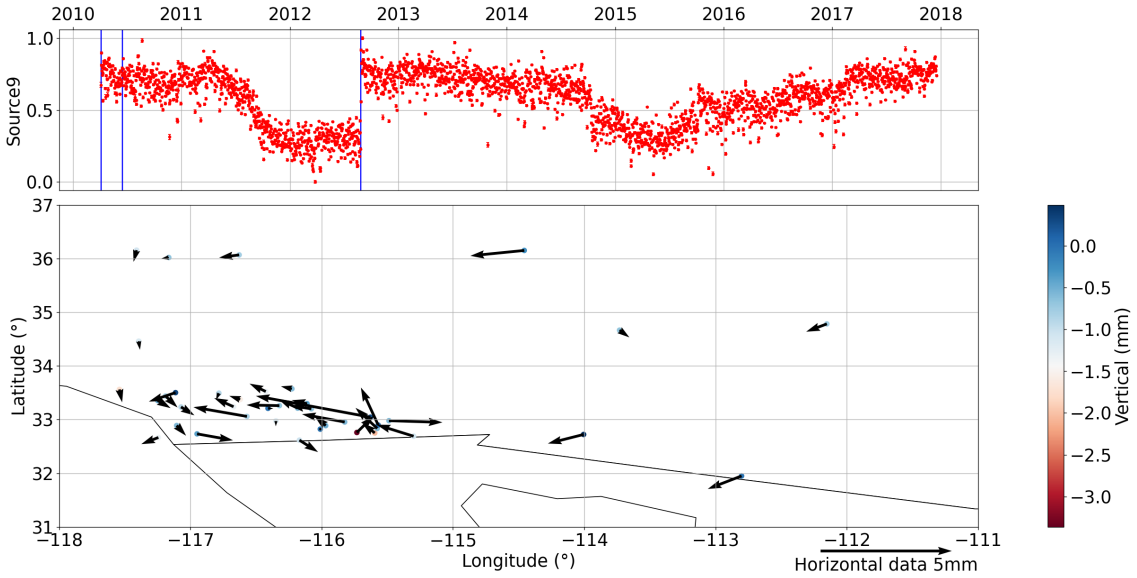


Figure 6.19: The ninth source of ten extracted from 40 GNSS stations in the SoCal region, with PC source initialisation. This source appears to partly represent the Brawley swarm deformation.

CHAPTER 7

Distance correlation for machine learning applications

7.1 Introduction

This chapter focuses on using distance correlation as both an evaluation metric and a loss function for the well-established auditory BSS and disentanglement machine learning tasks. ICA has been used extensively for many such tasks, including disentanglement [18] [116], image classification [161] [162] [163], and time series analysis [164] [165]. Therefore, I will use ICA and disentanglement problems to test distance correlation as a loss function.

Distance correlation, as proposed in [152], provides a non-parametric test for the independence of two or more variables based on relationships established through inter-point distances, as defined in Section 2.3.2. Distance correlation is formulated in a closed form, which has been examined for non-parametric tests in Chapter 3 and for BSS tasks in Chapters 5 and 6, but not for well-established machine learning tasks. This chapter aims to remedy this.

For the BSS task, I chose the Libri2Mix audio source separation dataset [17], utilising the state-of-the-art pre-trained SpeechBrain-SepFormer method [166]. Moreover, a small synthetic case, where I separated three linearly combined LibriSpeech signals, was

investigated.

Moving away from blind source separation, I will introduce disentanglement. Disentangled representation learning focuses on creating a representation of the given data that captures distinct factors of variation, ensuring each factor has its own individual component within the representation. One of the tasks I explored is the unsupervised non-linear disentanglement of underlying factors in videos, using the KITTI-Masks dataset [18] as an example. This dataset includes pedestrian segmentation masks derived from the KITTI-MOTS videos, which serve as a benchmark for evaluating vision systems in autonomous driving. Positive pairs of frames from the videos are defined based on a user-specified time interval, denoted as $\overline{\Delta t}$. For the disentanglement process involving KITTI-Masks, the most advanced loss function utilised is the contrastive InfoNCE loss. This loss function aims to increase the mutual information between positive pairs while increasing the distance between negative pairs.

The other dataset examined for disentanglement is CIFAR-10 [87]. This dataset was used to introduce whitening for representation learning [4], a technique aimed at scattering the data and bringing positive samples closer together. In this chapter, I have expanded this whitening technique to incorporate distance correlation. While whitening-based contrastive representation learning has been explored in [4], the InfoNCE loss remains the gold standard for the KITTI-Masks (see Section 2.5.3) disentanglement task.

Linking this work back to geodetic data, blind-source separation of the time-domain Libri2Mix dataset parallels decomposing GNSS displacement series, with spatial measurements potentially playing the role of frequency-domain inputs. Though, Libri2Mix dataset is labelled, unlike geodetic data. I introduced CIFAR-10 primarily as a baseline for disentanglement using W-MSE and related losses. Nevertheless, its labelled spatial patterns also serve as a proof of concept for classifying geospatial structures. The KITTI-Mask dataset, chosen for its natural scene transitions, mirrors the abrupt shifts observed in geodetic data, such as SAR, due to seismic events. Broadly speaking, the representation learning of CIFAR-10 and KITTI-Masks can be seen as an extension of source separation, operating in a learned latent space.

7.1.1 Contributions

This chapter explores the application of distance correlation loss to well-established machine learning tasks related to audio speech separation and representation learning. My specific contributions are outlined below:

- I applied the distance correlation loss to the KITTI-Masks non-linear disentanglement task. This included an adaptation of W-MSE with distance correlation, denoted as $W_{DistCorr}$, as well as an implementation of InfoNCE using double-centred distances. I discovered that the InfoNCE with double-centred distance outperformed the state-of-the-art methods and improved the MCC scores for sparse results. This is demonstrated by the Uniform and Laplace results.
- I compared different combinations of whitening loss functions for use on the CIFAR-10 dataset.
- I investigated the use of distance correlation to fine-tune auditory source separation on the Libri2Mix dataset, utilising the pre-trained SpeechBrain SepFormer model.

7.2 Distance correlation test for independence

In Section 5.2.2, I introduced self-supervised contrastive learning. It utilises a whitening process, which was based on the work in [4], eliminating the requirement for negative samples. The key aspects of contrastive learning, as introduced in Section 5.2.2, involve ensuring the closeness of features from positive pairs and maintaining a uniform distribution of the normalised features on the hypersphere, as highlighted in [105]. The whitening process helps to enforce these properties. In my investigation, I will explore several loss functions aimed at decreasing the distance between positive pairs. These include distance correlation and multiple equivalent mean squared error losses, which can be found in Equations 5.4 (reiterated in Equation 7.1) and 7.2.

$$MSE(\mathbf{A},\mathbf{B}) = 2 - 2R_n^2(\mathbf{X},\mathbf{Y}) \tag{7.1}$$

$$MSE(\mathbf{X},\mathbf{Y}) = 2(1 - r(\mathbf{X},\mathbf{Y})) \tag{7.2}$$

Implementation of loss functions

Distance correlation is a robust metric, capturing both linear and non-linear relationships in the data, making it valuable for assessing independence between random variables. The complexity of computing distance correlation increases with the square of the dimension. Therefore, for a sample size n, the computational complexity is n^2, which makes it quite expensive in terms of computation. To address this issue, I explore methods to segment the data in order to reduce this complexity.

The optimisation for the Libri2Mix synthetic problem involves two main steps. The first step is whitening, which helps to scatter the data on a hypersphere. After that, the objective is to maximise $2(1 - R_n^2)$, which minimises distance correlation.

Given the complexity of the distance calculation, which increases with the length of the squared signal, three resampling methods were examined:

- Average Method: In this approach, distance correlation is calculated for every segment of 1,000 samples from the signal. If the total number of samples is not divisible by 1,000, the distance correlation for the remaining data points is also included. The distance correlation values are then averaged.

- Alternating Method: In this approach, the 1,000 sample segments are randomly ordered every epoch to compare the distance correlation between the samples. This loss term is combined with the loss calculated using the Average method to form the overall loss function. The weight assigned to the intersample distance correlation is 0.05, while the loss from the previous method carries a weight of 0.95. Since the intersample loss is noisy, relying on it alone can cause issues during training.

- Resample Method: This method involves sampling at regular intervals by selecting one data point for every six samples in the overall signal. The starting point for this sampling is chosen randomly at every epoch. The length of the resampled signal is defined such that their lengths remain constant each epoch.

These methods aim to enhance efficiency while maintaining the accuracy of the optimisation process.

In representation learning for the CIFAR-10 dataset, I compare the traditional whitening mean squared error (W-MSE) loss with various alternative whitening loss functions. In these alternatives, the mean squared error (MSE) term is defined in different ways, including its dot product definition (Dot); a definition based on the cosine function (Cos) and a definition based on the Pearson correlation coefficient as seen in Equation 7.2. Additionally, I applied the whitened distance correlation method ($W_{DistCorr}$) to the CIFAR-10 task. This method utilises whitening to spread the data on a hypersphere, while distance correlation is employed to reduce the distance between positive pairs.

The W-MSE and $W_{DistCorr}$ were both applied to the KITTI-Masks dataset. The gold-standard InfoNCE loss serves as the baseline for this dataset. Additionally, I modified the InfoNCE loss to incorporate double-centred distances as input, using it as a proxy for distance correlation. The intuition for using this came from Equations 7.1 and 7.2.

7.3 Experimental set-up

7.3.1 Source Separation

For a synthetic problem, I linearly mixed three ground-truth sources from the Libri2Mix \ LibriSpeech dataset to create a straightforward source separation case. This synthetic audio source separation problem utilised the mixing matrix from the previous synthetic example, as defined in Equation 4.7.

Additionally, to add complexity, I selected the clean Libri2Mix supervised speech separation task, as it presents a more challenging scenario compared to the synthetic problem. The Libri2Mix dataset, described in Section 2.5.2, was selected as it represents a natural progression from the blind source separation task to a well-established example in machine learning.

For the simple synthetic problem, the architecture is a single linear layer, which suffices for linear unmixing. I employed three different segmentation techniques in order to deal with the complexity of the distance correlation. See Section 7.2 for details.

In each case, the Restart Algorithm was used for the Separation Architecture with

ZCA whitening. I investigated two signal lengths: 30,000 points and 25,000 points, to examine the effect of silence on the signal. In the 25,000-point scenario, I removed 5,000 points of leading silence. Due to the complexity mentioned earlier, various segmentation methods were used to partition the overall audio signal. However, in some cases, this segmentation can result in distance correlation being calculated between segments of data that contain only silence, combined with some noise from the recording. The distance correlation calculated for silence would be nearly 0, with any variations attributed to the recording noise. As a result, the perceived dependence of the signal may appear artificially high in these segments, which would increase the average distance correlation and influence the outcomes of the distance correlation minimisation process.

The second Libri2Mix task involves a more complex challenge of separating two overlapping speech signals. This work is based on the pre-trained SepFormer architecture [166]. It includes a comparison between fine-tuning using SI-SNR and distance correlation. Both SI-SNR and distance correlation are utilised for fine-tuning as well as for comparing the outputted signals. This approach ensures that neither loss function is favoured by using it as both an optimisation function and an evaluation metric.

7.3.2 Representation Learning

The CIFAR-10 dataset [87] consists of 60,000 colour images, each measuring 32x32 pixels, distributed across ten classes. The dataset's representations are learned through self-supervised learning techniques, such as the W-MSE loss introduced in [4], with a ResNet-18 architecture, with four positive samples extracted from each image. This loss function organises the representations onto a sphere through whitening and aligns positive pairs using an MSE, or equivalent, loss.

The other representation learning task utilises the KITTI-Masks dataset [18]. This dataset contains 2120 sequences of binary masks of pedestrians, with varying sequence lengths. In this experiment, I used a randomly initialised custom ConvNet encoder with ReLU activation, followed by a linear projection, as in the implementation of [116], rather than ResNet-18 architecture mentioned in their paper. Once disentanglement training is complete, following [116], I compute the Mean Correlation Coefficient (MCC) to evaluate performance. To calculate the MCC, you first compute the correlation between each latent

Method	Length	Signal 1	Signal 2	Signal 3
Average	25,000	47.1±13.4	32.3±10.4	43.4±4.6
Alternating	25,000	45.8±13.0	34.0±12.3	42.3±5.9
Resample	25,000	41.1±0.9	29.6±0.3	38.5±1.2
Average	30,000	43.4±14.7	39.0±14.7	39.1±4.5
Alternating	30,000	43.6±14.0	38.9±14.6	39.7±4.7
Resample	30,000	39.1±1.2	30.2±0.5	38.1±0.9

Table 7.1: Mean and standard deviation SI-SDR values for 10 repeats of the audio source separation problem, using a variety of resampling methods and signal lengths.

dimension and every ground-truth factor across all samples. For each factor, you then select the highest correlation among the latent dimensions and average these values to produce the final MCC score. As noted in my notation, the distance between time frames within a KITTI-Masks pair is Δt, the maximum of this distance being a hyperparameter. In related works, the maximum distance is not used when referring to results; rather, the mean, $\overline{\Delta t}$, is used. A maximum value of 1s or 5s relates to a $\overline{\Delta t}$ of 0.05s and 0.15s, respectively. Further details can be found in Section 2.5.3.

7.4 Results

This section presents results for speech separation from the LibriSpeech synthetic example, the Libri2Mix separation task, and representation learning outcomes for the CIFAR-10 and KITTI-Masks datasets.

7.4.1 Source Separation

The synthetic auditory source separation task involves isolating three ground-truth sources from the LibriSpeech dataset, which I combined to form three mixtures. In Table 7.1, I compared various sampling methods in relation to the three-mixture and three-source separation problem. The shorter lengths of the samples contained reduced amounts of silence at the beginning of the auditory signal. The SI-SDR values for the different sampling methods and the three signals are presented in Table 7.1, with examples of the outputted 30,000-point signals shown in Figure 7.1.

The Resample method exhibits the lowest standard deviation among the evaluated

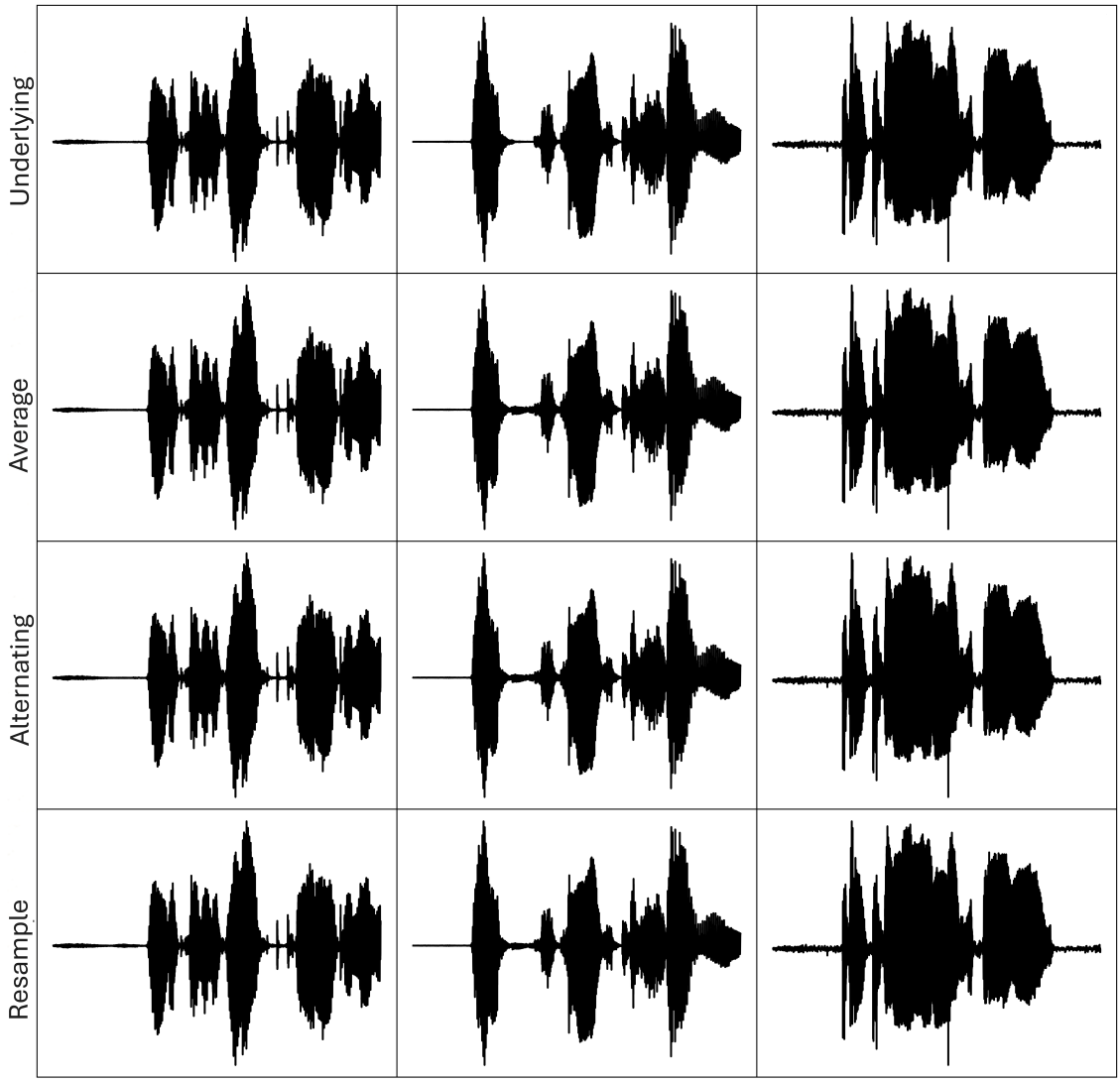


Figure 7.1: Examples of the three outputted 30,000-point audio signals from the synthetic LibriSpeech problem. The first row represents the ground truth, followed by the outputs associated with the three resampling methods defined in Section 7.2. These results correspond to the SI-SDR values presented in Table 7.1.

techniques. By resampling at every *6th* point within the signal, the resulting resampled signals are likely to retain more shared information compared to randomly selected sections of data. This leads to reduced noise during the training phase. Additionally, the average SI-SDR values were nearly identical for the 25,000- and 30,000-point cases for the Resample method. This indicates that silence has a low impact, which is consistent with how the Resample technique reduces the sparsity of the sample. As a result, silence has less influence, and the resampled signal provides a more accurate representation of the overall signal.

I investigated how different segmentation lengths affect the distance correlation using the synthetic problem. The results show that the distance correlation seems to plateau at a segment length of around 1,000 and with a slight increase for the segment of length 10,000, as depicted in Figure 7.2. The variation in distance correlation calculation is not only influenced by the signal within the segment but also by aspects of the inputted signal. This understanding is crucial in interpreting the results and their implications.

Figure 7.3 illustrates the run time in relation to segment length. The complexity of distance correlation is n^2. When averaging over segments, the complexity reduces to n^2/m, where m represents the number of segments.

For the speech separation task using the Libri2Mix dataset, I employed an SI-SNR pretrained SepFormer model, which was subsequently fine-tuned on Libri2Mix dataset with either SI-SNR or distance correlation as the optimisation metric.

As shown in Table 7.2, fine-tuning with either SI-SNR or distance correlation improved separation performance as indicated by the metric used for fine-tuning, while negatively impacting the other metric. This demonstrates that while distance correlation and SI-SNR can complement each other during training, they do not yield equivalent results.

Following standard practice, below I also report SI-SNRi, that is, the difference between the SI-SNR of the extracted sources and targets, and the SI-SNR of the mixtures and targets. In our experiments, the SI-SNRi (and SDRi in brackets) results are as follows: 17.0dB (17.5dB) for no fine-tuning, 17.6dB (18.1dB) for SI-SNR fine-tuning, and 13.4dB (13.7dB) for distance correlation fine-tuning. For context, SI-SNRi values of 12.2dB, 16.5dB, 16.6dB, and 22.0dB correspond to the Conv-TasNet [93], SepFormer [103], WaveSplit [167], and SepReformer-M systems [168], respectively. The results indicate

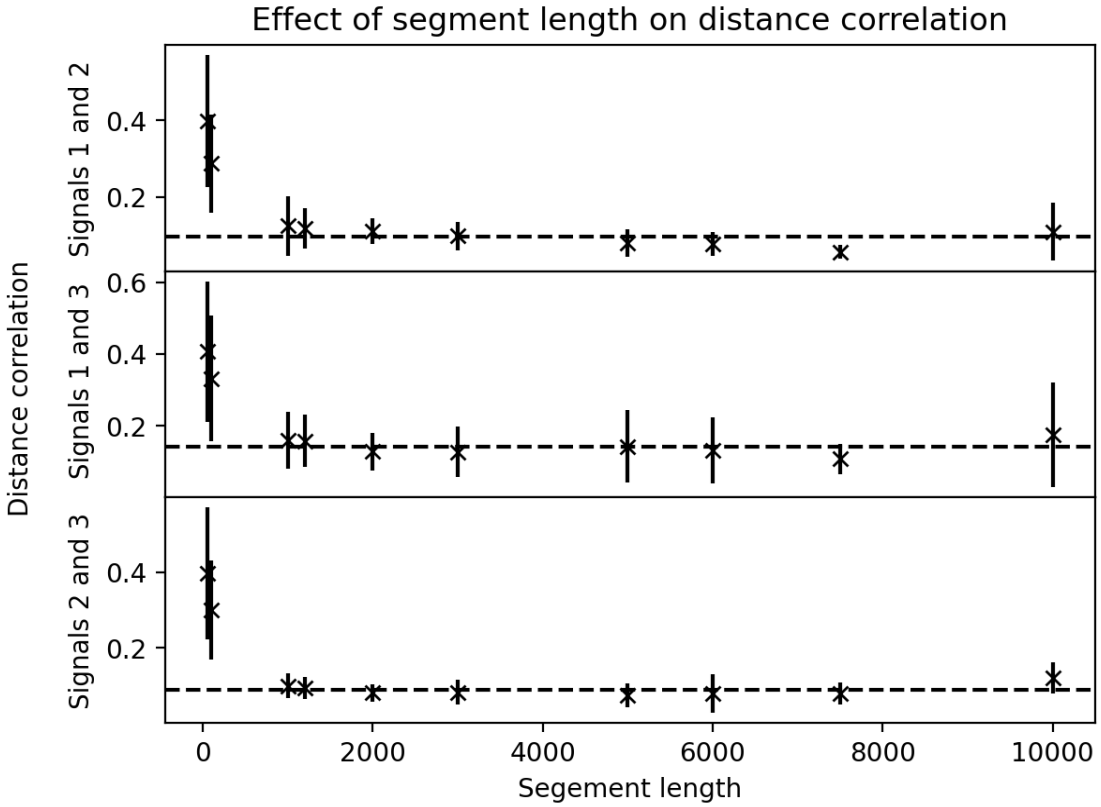


Figure 7.2: The average pairwise distance correlation among the underlying three LibriSpeech signals is analysed across all segments of varying lengths, with the segment length displayed on the x-axis. The dotted line indicates the average distance correlations for segment lengths of 1,000 and greater.

Finetuning \ Metric	SDR	SI-SNR	DistCorr
None	17.7	17.0	0.4332
SI-SNR	18.3 (+0.6)	17.6 (+0.6)	0.4308 (-0.0014)
DistCorr	13.8 (-3.9)	13.4 (-3.6)	0.4603 (+0.0281)

Table 7.2: Two different loss functions (SI-SNR) and (DistCorr) for the fine-tuning of the pre-trained SepFormer model used on the Libri2Mix dataset. SDR, SI-SNR and DistCorr are used as the evaluation metric, with bracketed numbers giving the change from the case without finetuning.

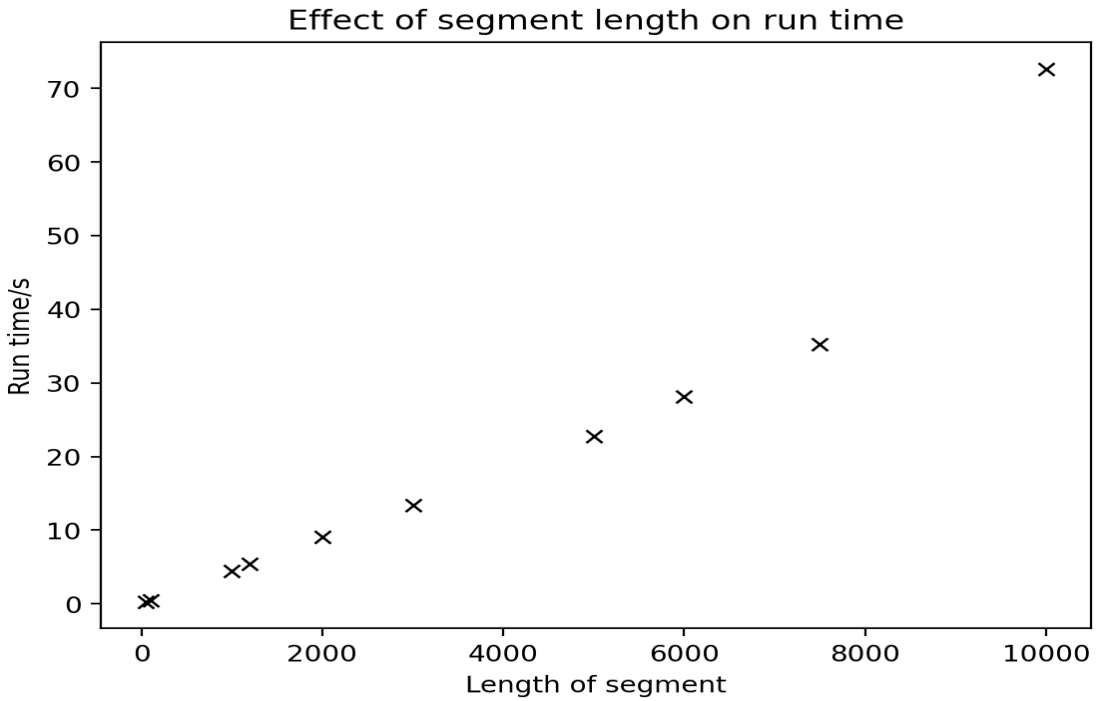


Figure 7.3: Segment length vs time required for calculation.

that the proposed distance correlation method achieves a reasonable trade-off of fidelity for independence. In fact, the fidelity measured by SI-SNRi still surpasses the least effective of the comparator methods, though potentially in large part due to the pretraining.

Although SI-SNRi normalises performance by subtracting the mixture baseline, making it ideal for comparing models across varying input conditions, my experiments use a constant test set with consistent mixture SI-SNR. In this scenario, absolute SI-SNR directly reflects each model's output fidelity without extra baseline correction, simplifying evaluation and focusing squarely on final signal quality.

7.4.2 Representation Learning

This section presents the results for the CIFAR-10 dataset, which illustrates the whitening mean squared error (W-MSE) contrastive learning method. Additionally, the section includes the KITTI-Masks pedestrian segmentation task to compare the whitening contrastive methods with the more traditional InfoNCE contrastive method, described in Section 2.6.2, which when minimised maximises to the lower bound of the mutual information between random variables.

From Table 7.5, one can see equivalent whitening MSE methods. They aim to achieve better uniformity, whilst reducing the distance between positive pairs using mean squared error (MSE), or equivalent metric. The W_{MSE} methods yielded similar accuracy results of approximately 91.2%.

The distance correlation method, $W_{DistCorr}$, I proposed to compare with W_{MSE} performed 2.4% worse than the best MSE results. As shown in Tables 7.4, 7.6 and 7.7, the $W_{DistCorr}$ method is less resilient to extreme learning rates, specifically when I chose learning rates of $3e^{-5}$ and $3e^{-2}$. Under these conditions, the chosen hyperparameters result in the network learning a poor representation that is not positive definite, either due to divergence or convergence to a suboptimal local minimum. The Cholesky decomposition is a differentiable method for decomposing a Hermitian positive definite matrix into the product of a lower triangular matrix and its conjugate transpose. Consequently, by definition, the Cholesky decomposition fails during training under these circumstances.

Two approaches were investigated to determine whether the results of Cholesky whitening could be improved. Firstly, the covariance matrix LL^T was regularised to ensure that it is positive definite. In this regularised form, LL^T is expressed as $(1-eps)LL^T + eps\ I$, where *eps* is a user-defined variable, is the lower triangular matrix, and I is the identity matrix. This regularisation allows the Cholesky whitening process to work on a weighted average of the learned matrix and the identity matrix, with the latter being inherently positive definite. Consequently, there exists a value for *eps* (specifically $eps = 1$) at which the Cholesky whitening will always succeed without failure. I conducted several trials. However, they did not produce good representations of the data.

Secondly, I substituted the Cholesky whitening step with ZCA whitening. ZCA whitening eliminates the requirement for the input to be positive definite, which allows training to proceed even for representations that may be inadequate for continued training using Cholesky whitening.

When using a small diagonal element for numerical stability in the Cholesky decomposition or replacing that decomposition with ZCA whitening, I found that the resulting representations far underperformed Cholesky whitened representation, without significant alterations to the network. The same can be said for altering the nature of whitening. These methods seem only to produce poor representations instead of enhancing the out-

	Top1	Top5	5-nn
W_{MSE} (MSE: lr=$3e^{-4}$)	0.9122	0.9967	0.8886
W_{MSE} (Dot: lr=$3e^{-3}$)	0.9122	0.9974	0.8907
W_{MSE} (Cosine: lr=$3e^{-3}$)	0.9119	0.9981	0.8927
W_{MSE} (Pearson: lr=$3e^{-3}$)	0.9127	0.9976	0.8966
$W_{DistCorr}$ (lr=$3e^{-3}$)	0.8882	0.9950	0.8639

Figure 7.4: Comparison between Top 1 and Top 5 accuracies and the 5 Nearest Neighbours classifiers for the W-MSE methods, described in Section 7.3.1, and related distance correlation method. The task in this case was CIFAR-10 and 250 epochs were used in each case.

Method \ LR	$3e^{-5}$	$3e^{-4}$	$3e^{-3}$	$3e^{-2}$
W_{MSE}	0.8193	0.9122	0.8316	0.6716
W_{MSE}	0.7965	0.8944	0.9122	0.8242
W_{MSE}	0.8058	0.8949	0.9119	0.8289
W_{MSE}	0.8031	0.8964	0.9127	0.8310
$W_{DistCorr}$	-	0.8687	0.8882	-

Figure 7.5: Top 1 accuracy of a linear classifier for the CIFAR-10 dataset regarding the impact of learning rate on various whitening methods. Note '-' represents a null result due to a poor representation that was not positive definite being learned.

puts or producing good representations.

From Figures 7.4, 7.6 and 7.7 it can be observed that the $W_{DistCorr}$ method is more prone to learning poor representations due to getting stuck in local minima or diverging, compared to the W_{MSE} methods. However, Table 7.8 shows that, across different learning rate ranges, the Top 1, Top 5, and 5-nn metrics yield similar values to those of the W_{MSE} methods. It is important to note that the W_{MSE} was scaled by a factor of 64 compared to the equivalent losses (evident in Table 7.5), causing its optimal learning rate to be an order of magnitude different to the other losses. For a learning rate of $3e^{-2}$, this method results in a positive definite representation that is inferior compared to other learning rates used, with Top 1, Top 5, and 5-nn values being 0.6716, 0.9749, and 0.6543, respectively. The MSE methods demonstrate a broader range of outcomes than the distance correlation method. Thus, while MSE methods are generally easier to train, they can yield worse representations compared to those produced by the distance correlation method.

The whitening representation learning methods are now applied to a new dataset, the

Method \ LR	$3e^{-5}$	$3e^{-4}$	$3e^{-3}$	$3e^{-2}$
W_{MSE}	0.9910	0.9967	0.9940	0.9749
W_{MSE}	0.9901	0.9961	0.9974	0.9920
W_{MSE}	0.9901	0.9958	0.9981	0.9925
W_{MSE}	0.9908	0.9956	0.9976	0.9933
$W_{DistCorr}$	-	0.9936	0.9950	-

Figure 7.6: Top 5 accuracy of a linear classifier for the CIFAR-10 dataset regarding the impact of learning rate on various whitening methods. Note '-' represents a null result due to a poor representation that was not positive definite being learned.

Method \ LR	$3e^{-5}$	$3e^{-4}$	$3e^{-3}$	$3e^{-2}$
W_{MSE}	0.7759	0.8886	0.8069	0.6543
W_{MSE}	0.7578	0.8735	0.8907	0.7950
W_{MSE}	0.7582	0.8714	0.8927	0.7964
W_{MSE}	0.7601	0.8768	0.8966	0.7945
$W_{DistCorr}$	-	0.8404	0.8639	-

Figure 7.7: 5-Nearest Neighbours classifier accuracy for the CIFAR-10 dataset regarding the impact of learning rate on various whitening methods. Note '-' represents a null result due to a poor representation that was not positive definite being learned.

Accuracy \ LR	1e^{-4}	3e^{-4}	1e^{-3}	3e^{-3}
Top 1	0.8281	0.8687	0.8770	0.8882
Top 5	0.9922	0.9936	0.9924	0.9950
5-nn	0.7859	0.8404	0.8393	0.8639

Figure 7.8: Top 1 and 5 of a linear classifier and 5-Nearest Neighbours accuracies for the CIFAR-10 dataset regarding the impact of a smaller range of learning rates on the $W_{DistCorr}$ method.

KITTI-Masks dataset, and compared to the more traditional contrastive learning InfoNCE loss. Moreover, I created an updated version of the InfoNCE loss that uses the double-centred pairwise distances of the random variables as the input rather than the random variables themselves, as a proxy for distance correlation. InfoNCE (DC) builds on the work in [116], which demonstrated that a carefully selected contrastive loss on convex bodies corresponds to the cross-entropy between the ground-truth conditional distribution and the inferred latent distribution.

For the KITTI-Masks dataset, [18] found that the transition between ground truth latents was sparse. The transition distributions are related to the conditional distributions. Two sampling methods were utilised to compare a transition that agrees with the knowledge from [18] that the transitions are sparse (a Laplace conditional distribution) or not (a Uniform conditional distribution).

The Laplace distribution takes advantage of the sparsity in transitions between the ground truth latent states across nearby frames [116] [18]. This distribution features a sharper peak and heavier tails compared to a Gaussian or Uniform distribution, favouring zero as a result.

An overview of the sampling methods as outlined in [18] is provided below for the different conditional distributions:

- **Uniform**: The work of [18] builds upon the creation of uniform datasets described in [169]. In [169], the authors explain that they sample from discrete, independent factors of variation based on the ground-truth generative model in order to create data (see [169] for more details). K independent factors of variation are sampled to enable denser changes, ensuring that the two images do not share these factors. The coordinates are then resampled to generate new factors of variation. Exactly k factors change as stated in [18], with the new factors of variation sampled uniformly. As an aside, note that a uniform transition does not describe natural transitions well, such as in the case of KITTI-Masks [18].

- **Laplace**: To produce pairs of images, for every ground truth factor, the first value in the pair is selected from a uniform distribution across all possible values in the latent space. The second value is chosen by weighting nearby values in the latent

$\overline{\Delta t}$	Conditional distribution	Model Space	Loss	MCC (%)
0.05	Laplace	Unbounded	InfoNCE	77.1±1.0
0.05	Laplace	Box	InfoNCE	74.1±4.4
0.05	Laplace	Unbounded	InfoNCE (DC)	**77.2±5.3**
0.05	Laplace	Box	InfoNCE (DC)	74.4±4.7
0.05	Laplace	Unbounded	W_{MSE}	65.1±6.3
0.05	Laplace	Box	W_{MSE}	72.2±3.5
0.05	Laplace	Unbounded	$W_{DistCorr}$	59.2±4.5
0.05	Laplace	Box	$W_{DistCorr}$	67.0±4.5
0.15	Laplace	Unbounded	InfoNCE	79.4±1.9
0.15	Laplace	Box	InfoNCE	80.9±3.8
0.15	Laplace	Unbounded	InfoNCE (DC)	**81.5±2.7**
0.15	Laplace	Box	InfoNCE (DC)	78.4±5.6
0.15	Laplace	Unbounded	W_{MSE}	63.9±6.2
0.15	Laplace	Box	W_{MSE}	63.2±2.9
0.15	Laplace	Unbounded	$W_{DistCorr}$	57.0±4.1
0.15	Laplace	Box	$W_{DistCorr}$	60.7±2.5

Table 7.3: Mean ± standard deviation MCC scores are presented for whitening-based and InfoNCE-based representation learning techniques. A sparse Laplacian conditional distribution is employed in all cases. $\overline{\Delta}$t represents the average temporal distance of the frames utilised.

space using probabilities that follow a Laplacian distribution. This approach aligns naturally with the sparsity of the KITTI-Masks transitions identified in [18].

From Tables 7.3 and 7.4, the whitening methods underperform compared to the traditional contrastive learning method InfoNCE in the Laplace case but can outperform InfoNCE in the Uniform case. However, as previously mentioned, the Uniform conditional does not represent the transition between the frames in the KITTI-Masks dataset. This is exemplified by the MCC scores of the InfoNCE method with a Laplace underperforming with this conditional distribution.

The sparseness of the Uniform case reduces the advantages that the traditional InfoNCE method has over the whitening techniques observed in the Laplace example. In the whitening methods, the similar MCC scores for both the Uniform and Laplace cases suggest that scattering, rather than learned uniformity, may not be adequate for the KITTI-Masks case.

Regarding the whitening methods themselves, similar MCC scores for the Uniform and Laplace cases suggest that scattering through whitening, rather than learned unifor-

$\overline{\Delta t}$	Conditional	Model Space	Loss	MCC (%)
0.05	Uniform	Unbounded	InfoNCE	58.3±5.4
0.05	Uniform	Box	InfoNCE	59.9±5.5
0.05	Uniform	Unbounded	InfoNCE (DC)	74.9±2.9
0.05	Uniform	Box	InfoNCE (DC)	76.4±3.0
0.05	Uniform	Unbounded	W_{MSE}	66.2±7.5
0.05	Uniform	Box	W_{MSE}	72.1±5.0
0.05	Uniform	Unbounded	$W_{DistCorr}$	57.1±6.7
0.05	Uniform	Box	$W_{DistCorr}$	65.6±4.3
0.15	Uniform	Unbounded	InfoNCE	60.2±8.7
0.15	Uniform	Box	InfoNCE	68.4±6.7
0.15	Uniform	Unbounded	InfoNCE (DC)	76.0±1.7
0.15	Uniform	Box	InfoNCE (DC)	79.0±4.7
0.15	Uniform	Unbounded	W_{MSE}	65.5±4.4
0.15	Uniform	Box	W_{MSE}	63.6±3.3
0.15	Uniform	Unbounded	$W_{DistCorr}$	58.7±5.0
0.15	Uniform	Box	$W_{DistCorr}$	62.3±2.9

Table 7.4: Mean ± standard deviation MCC scores are presented for whitening-based and InfoNCE-based representation learning techniques. A Uniform conditional distribution is employed in all cases. $\overline{\Delta}$t represents the average temporal distance of the frames utilised.

mity, may not be adequate for the KITTI-Masks case. I believe that that the performance ceiling of W_{MSE} and W_{DCorr} on the KITTI-Masks dataset stems from their assumption of a continuous whitened latent space, which misrepresents the inherently binary nature of the KITTI-Masks data and leads to poor compatibility between the model and the latent space geometry.

[116] demonstrate that encoders trained with InfoNCE effectively invert the data-generating process, utilising the box normalisation on the latent representations for KITTI-Masks in specific experiments. By selecting a latent geometry that reflected the hyper-rectangle vertices of the binary mask inputs and without degrading the quality of the representations, the assumed latent geometry was empirically validated. Indeed, the L_2 normalisation of W_{MSE} imposes by definition a spherical geometry, distorting inter-mask-representation distances. However, an L_∞ normalisation focuses on the largest coordinate, making the contrastive objective sensitive to important, infrequent, large-magnitude representation values that binary mask inputs and sparse, heavy-tailed latents are likely to produce [170].

Similarly to the results of [116], the InfoNCE results with a $\overline{\Delta t}$ of 0.15 slightly outperform frames taken from closer time-steps. However both InfoNCE methods, unlike SlowVAE [116], do not see a degradation in performance due to the limited expressiveness of the decoder and therefore do not appear to have as marked a decrease for similar frames, i.e. those closer by $\overline{\Delta t} = 0.05$.

As previously mentioned, the Laplace model outperforms the Uniform model in the KITTI-Masks dataset as the transition between the latents of nearby frames is sparse, allowing for the Laplacian distribution to describe it better. The best average MCC score is 81.5% and is related to the updated version of InfoNCE that I introduced which used an inputted double-centered matrix. This method exceeds the Box InfoNCE Laplace version by 0.6%.

Moreover, it was observed that the double-centered InfoNCE (DC) loss function yielded a higher MCC score compared to the other loss functions when using the Uniform model. This indicates that the double-centering in the InfoNCE (DC) loss makes the inputted representation sparse, even when the Uniform model is used for the conditional distribution.

To reiterate, there are two significant points. Firstly, the best MCC score is related to the InfoNCE (DC) loss function that was introduced in this thesis. Secondly, that the double-centering step adds an element of sparesness in the KITTI-Masks case, which reduces the difference in results between the Laplace and the Uniform cases, caused by the inherent sparseness of the transitions between nearby frames.

7.5 Conclusions

In this chapter, the use of distance correlation has been introduced for use in established machine learning tasks, such as fine-tuning a SepFormer model for the separation of audio data in the Libri2Mix dataset and for representation learning in the CIFAR-10 and KITTI-Masks datasets examples.

The results from various test cases offer insights into the performance of distance correlation across different scenarios. These insights are crucial for advancing machine learning and signal processing, as they can help to select appropriate methods for specific tasks.

In the synthetic problem involving the mixing of three underlying audio signals from the LibriSpeech dataset using a known mixing matrix, the average SI-SDR values for each signal were found to be above 29 dB, which is considered a good result. The effectiveness of source separation using different signal lengths and sampling methods varied. The Resampling method, which better preserves the information contained within each of the three signals during the sampling process, yielded the best results in terms of robustness. This finding opens new avenues for future applications of distance correlation in the source separation of longer signals.

In a study where SI-SNR and distance correlation were utilised to fine-tune a pre-trained SepFormer model for separating underlying speech signals from the Libri2Mix dataset, it was observed that using the SI-SNR value as an objective function for fine-tuning did not improve the performance of distance correlation compared to the initial pre-trained model. Conversely, employing distance correlation as an objective function also did not enhance SI-SNR.

In the CIFAR-10 case, the distance correlation method with whitening exhibited noticeable underperformance compared to state-of-the-art techniques, with a difference of around 2.4%. This result highlights the necessity for further research and improvements in applying distance correlation. The MSE whitening methods are more likely to yield imperfect representations when the learning rate hyperparameter is varied, leading to either divergence or convergence to a poor local minimum. The distance correlation loss is particularly susceptible to settling in bad local minima due to its complex data representation, causing training to fail. Although this is suboptimal and can result in training failures, it was observed that using a narrower range of learning rates allowed the classifiers to achieve a better performance with less deviation in values. This indicates that while distance correlation can be challenging to train, successful training can lead to identifying worthwhile data representations.

In the KITTI-Masks case, the InfoNCE (DC) loss with double centring, serving as a proxy for distance correlation, showed promise by outperforming the leading InfoNCE loss (Laplace Unbounded model) by 0.6%. It is important to note that the means of the InfoNCE and InfoNCE (DC) methods are within one standard deviation of each other. This suggests that a degree of caution should be taken with respect to the representativeness

of the sample, and the mean of InfoNCE (DC) outperforming the original InfoNCE [1]. In comparing the Laplace and Uniform conditional distributions, it was observed that the InfoNCE (DC) model introduced an element of sparsity, which reduced the difference in the MCC score between the two distributions. This finding indicates the potential advantages of using distance correlation-based metrics for analysing non-sparse data.

[1]The one-tailed two-sample t-test to determine if 81.5 ± 2.7 is significantly greater than 80.9 ± 3.8, with 10 samples, produces a one-tailed p-value of 0.3447. This suggests that 81.5 ± 2.7 is not significantly greater than 80.9 ± 3.8.

CHAPTER 8

Conclusions

In this chapter, I will provide an overview of the work contained within this thesis, followed by the conclusions relating to each research chapter, overall thoughts, and thoughts on future work.

8.1 Overview

In recent years, the volume of GNSS data has grown significantly, opening up new avenues for research that leverage big data to solve geodetic problems.

This thesis aimed to investigate how to separate geodetic signals from GNSS displacement time series, which record pseudorange data instead of the actual position of a receiver station. To address the challenge of geodetic signals being obscured by non-geodetic signals that are of a much higher magnitude, I have defined the problem of extracting geodetic signals as a blind source separation problem. In this work, I assume that the underlying components of the GNSS pseudorange are independent.

Several independence metrics, such as mutual information-based MINE, negentropy-based FastICA and distance correlation-based methods, were investigated to determine whether they would be suitable loss functions for machine learning-based source separa-

tion or representation learning. This problem was split into two problems:

- Is the metric capable of effectively describing the independence of sources?
- How well can it optimise to good extrema?

In terms of my original aims and objectives (Section 1.3), I will provide an overview of my results and how they follow from my aim:

1. **To compare distance correlation, non-Gaussianity and mutual information as measures of independence, especially regarding their suitability as loss functions**: In Chapter 3, I addressed these tasks using the non-parametric BPSK over an AWGN channel, which is well-established in communication theory. An investigation was conducted on the use of distance correlation applied to different colours of noise. This study also examined the independence of a binary signal and the average of multiple binary signals. It was observed that as the number of signals included in the average increased, the combined signal tended towards white noise. The BPSK problem over an AWGN signal addressed the first problem. It was found that both the negentropy and the distance correlation were highly correlated to the ground truth mutual information, identifying for this task that each metric describes independence well.

   In the context of BPSK over an AWGN channel, it was observed that as the variance increased, the negentropy tended to approach zero. Despite this, a linear relationship was still evident. In this particular scenario, the distance correlation did not approach zero, indicating that it might provide better description of source independence. Additionally, when considering coloured noise examples, the distance correlation exhibited a steeper gradient at higher variances compared to negentropy. Higher variances were used as an approximation for independence, as the noise component of the signal becomes dominant. Consequently, it was suggested that distance correlation might offer improved optimisation for gradient descent in such cases.

2. **To investigate the relationship between natural signal whitening procedures and independence as determined by distance correlation:** In Chapter 4, I exam-

ined five natural forms of whitening. In this task, no form of whitening was found to optimise distance correlation in source separation problems specifically. While not unexpected, this finding provided an element of verification for my research. It was also not surprising to find that no single whitening method emerged as superior; independence demands stricter conditions than decorrelation, considering both linear and non-linear relationships. However, in most instances, the best distance correlation between the input and its whitened equivalent was with one of the forms of ZCA whitening.

Regarding where in the pipeline whitening should be applied, it was tested as a pre-processing step and a step after the separation of sources. I found that whitening as a pre-processing step before separation limited the information explained by a given number of sources. Therefore, whitening should be implemented before the loss calculation after unmixing the mixtures into their underlying sources.

3. **Assess the performance of these methods regarding the separation of a synthetic earthquake signal embedded in a known GNSS signal, and then on blind source separation of an actual seismic event (without a known ground truth)**: In Chapter 5, the Synthetic problem and hybrid GNSS and InSAR problems were used to compare distance correlation, mutual information computed by MINE, and negentropy as measures of signal independence. For the Synthetic problem, using a linear layer as an unmixing function, distance correlation gave the best results on average, with SI-SDR values of 19.3±0.1, 57.4±3.0 and 19.1±0.2, for the sine, square and sawtooth waves, respectively. The SI-SDR values for the gold-standard baseline FastICA method were 21.9 ± 0.0, 30.9 ± 0.0, and 21.6 ± 0.0 for the sine, square, and sawtooth waves. Thus, this thesis proposes that distance correlation as a loss function may be at least as effective as the widely used FastICA method and can be more easily integrated into a machine learning context, opening an exciting new avenue for research.

   When a mixing matrix in the form of a linear layer was applied to learned parametrised sources, it produced outputs that were trained to be as close as possible to the known mixtures by minimising a reconstruction loss between the known and outputted

mixtures. A Restart algorithm was employed to identify optimal extrema results for distance correlation and reconstruction loss, focusing on minimising reconstruction loss. This prioritisation is crucial because it constrains the possible values of the outputted sources. For instance, while three white noise signals can be independent, they are unlikely to reconstruct the known mixtures. I found the best SI-SDR values for the mixing method to be 13.2 ± 1.1, 31.0 ± 0.5, and 14.6 ± 1.6 for sine, square, and sawtooth waves, respectively. It is important to note that the Reconstruction method underperformed the Separation method, highlighting the trade-offs involved. However, the mixing method offers greater flexibility in real-world GNSS scenarios when the number of sources varies from the number of mixtures.

In Chapter 6, blind source separation using distance correlation was applied to a standard 2-mix-2-source GNSS problem on data from the G119 and G025 stations. As there was no known ground truth, I used a proxy ground truth by taking an element of the trend and its residual at the time of a known seismic event instead. The SI-SDRs were 26.0 ± 0.0 and 29.9 ± 0.0 for distance correlation and 28.2 ± 0.0 and 30.0 ± 0.0 for FastICA, making the latter the apparent best method. I note, however, that the SI-SDR values are close, and the ground truth is a proxy. Both methods were found to be promising. The slight differences in the SI-SDR may be related to the method used to determine the ground-truth seismic signal; thus, the results should be taken cautiously, but with a sense of optimism for the potential of both methods.

The Reconstruction algorithm using a distance correlation objective function was applied to a blind source separation case study for a region in Southern California. Ten underlying sources were extracted using the reconstruction method with a distance correlation loss and the benchmark Variational Bayesian ICA. Though the most obvious conclusion is that vbICA is less robust to outliers, the geodetic meanings of the underlying signals are more subjective. In the distance correlation and the vbICA case, afterslip and post-seismic deformation could describe some of the extracted sources. The reconstruction method with Principal Component initialisation also extracted a source potentially representative of the Brawley swarm deformation. However, I was unable to separate the step function that represents

a seismic event into a single source. This limitation restricted the use of source separation as a preprocessing step, preventing the creation of a single input time series that contains the seismic information for training an earthquake prediction algorithm.

4. **To determine the applicability of a distance correlation-based loss on other machine learning tasks (fine-tuning the source separation of the Libri2Mix dataset; whitening for self-supervised representation learning; and the disentanglement task using the KITTI-masks dataset)**: In Chapter 7, I applied distance correlation to several established machine learning tasks and enhanced a form of representation learning by incorporating elements of distance correlation. For the CIFAR-10 representation task, I modified the W-MSE algorithm to include distance correlation instead of MSE. The Top 1 accuracy achieved with distance correlation was 88.82%. However, this distance correlation version underperformed compared to the original whitening MSE method, which had a Top 1 accuracy of 91.2%. As discussed previously, the $W_{DistCorr}$ method, whilst more challenging to train, did have a smaller range of accuracy scores, suggesting it is more likely to produce better representations consistently, when training is successful.

   In the disentanglement of the KITTI-Masks dataset, the previously mentioned whitening representation methods did not perform well. Nevertheless, when I updated the InfoNCE loss (Laplace, Unbounded) using double-centered inputs, drawing on principles from the distance correlation calculation, the state-of-the-art MCC score improved by 0.6%.

   Some principles of distance correlation can also be applied to other machine learning tasks, such as representation learning. However, distance correlation and the related work are particularly well-suited for source separation tasks because of their connection to independence.

8.2 Discussions

On concluding the research chapters, it is important to note that this work was not without its challenges. The overarching research objectives were to compare distance correlation to other measures of independence and assess its suitability as a loss function for blind source separation. I have found that the extrema of distance correlation describes independence between variables well. However, the optimisation for these extrema proved to be a complex task, often leading to suboptimal local minima. This highlights the complexity of the research process.

My research has uncovered both new insights and potential applications. For instance, when it comes to separating known synthetic signals, a simple neural network with a distance correlation loss performed well against the gold-standard FastICA algorithm. This finding opens up new possibilities for practical applications in the field. When the reconstruction method was applied to the SoCal case study, it was found that both it and the vbICA method extracted the step function representing a seismic event into multiple signals. This decreases their applicability in separating one source that can then be used to train for 'mainshock' events. However, in the synthetic data problem a 'seismic' and 'non-seismic' signal were able to be extracted, which could be used in a binary representation learning task.

In the final task, distance correlation was applied to benchmark representation tasks. In the case of CIFAR-10, the baseline W_{MSE} outperformed its distance correlation counterpart. This underperformance indicates that, while distance correlation is a powerful loss function, it may not be the most appropriate optimisation function for representation learning, as it inherently describes independence. According to [158], a tighter bound on mutual information between positive samples can lead to worse learned representations in contrastive representation learning.

For the KITTI-Masks representation learning task, the whitening representation learning methods developed on the CIFAR-10 task performed worse than the state-of-the-art loss function for this specific task, InfoNCE. When double-centered distances of the latent space random variables were used as input for the InfoNCE loss instead of the random variable itself, the MCC score improved by 0.6% for the Laplace model. Furthermore, double-centered InfoNCE effectively managed the sparsity of the features, enhancing the

MCC score in the Uniform case compared to other loss functions. Therefore, although distance correlation with whitening may not yield the best results, due to the nature of the dataset, the foundational concepts behind distance correlation can still provide valuable insights for representation learning.

8.3 Future work

In the future, an interesting research avenue could involve using temporal GANs (Generative Adversarial Networks) to impute missing data in GNSS time series. In this context, a conditional GAN trained with the station's location data would be particularly relevant, as different faults exhibit distinct seismic features.

After employing a temporal conditional GAN to impute the data, predicting seismic events before they occur can be framed as a binary classification problem. This can utilise data from the UNR master step file [2], which includes information about when historical seismic events happened. By shifting the event label to a time before the actual event, the neural network may learn to predict when seismic events are likely to occur.

However, seismic events are infrequent and do not always produce well-defined step functions, even if the station is within a certain radius of a significant event. Therefore, exploring a GAN-based anomaly detection method instead of relying solely on supervised binary classification or event-timing anomaly detection could be a promising direction for future research.

In [135], the chaotic behaviour of slow earthquakes, that produce little slip and therefore can repeatedly occur over a relatively short period, were studied. Deterministic chaos is a process governed by deterministic laws but highly sensitive to initial conditions [135]. The analysis revealed a low-dimensional, non-linear chaotic system rather than a stochastic system. Therefore, the machine learning problem can be reduced to a local area with slow slip events, such as the Cascadia fault in Canada, and the machine learning can learn the non-linear dynamics to assess its viability. In [135], the onset of significant slip events can be correctly forecasted by high values of the instantaneous dimension, allowing for a gold standard for such a comparison.

Regarding future optimisation research, particularly in the context of distance corre-

lation for the representation learning or source separation of GNSS data, two additional optimisation techniques should be explored, curriculum learning and meta-learning:

- Curriculum learning can effectively manage a challenging distance-correlation loss by guiding the model through a sequence of progressively complex feature extractions. Starting, for example, with synthetic mixtures that allow the network to learn the fundamental mappings. As one introduces real GNSS or seismic signals with unknown parameters, the model can utilise its already developed robust feature extractor to avoid becoming trapped in poor local minima. However, if the easy data does not accurately reflect the actual underlying statistics of the features, such as labquake representations that fail to generalise to real seismic measurements, the network may overfit to simple tasks and struggle with the GNSS signals.

- Meta-learning, particularly the optimisation of hyperparameters, provides an effective approach to navigating complex loss landscapes by leveraging experience across various tasks for quicker adaptation. In this process, a meta-learner is trained on a diverse set of problems with different parameters. This training helps the meta-learner develop an initialisation or hyperparameter policy that enables rapid tuning of the model for new separation tasks using minimal data. This automated tuning process enhances the robustness of the pipeline to unfamiliar mixing scenarios and decreases the reliance on manual trial-and-error methods. However, meta-learning can be computationally demanding, as it often involves nested loops of task sampling and adaptation. Furthermore, its effectiveness depends heavily on the diversity and representativeness of the tasks used for meta-training. If these tasks do not adequately cover a range of real-world conditions, the meta-learner may struggle when faced with highly novel or noisy data.

Meta-learning and curriculum learning offer complementary ways to tame complex GNSS representation tasks. With meta-learning, one trains across many station scenarios, varying noise levels, environmental loading, fault proximity, so the model learns an initialisation or hyperparameter policy that adapts instantly to a new site with only a few samples. Curriculum learning, by contrast, orders training from simple displacement patterns (e.g. pure tectonic drift) to seasonal cycles and finally to full-complexity time series,

smoothing the highly non-convex loss surface and reducing gradient noise. Together, they can yield faster convergence, more robust feature extractors across stations, and deeper insights into which signal components drive generalisable GNSS representations. Future research could include whether and how their implementation improves machine learning training that uses GNSS, or other geodetic, datasets.

In addition, by collaborating with seismologists, we could integrate high-resolution waveform recordings, seismometer and GNSS measurements, and plate-motion rates (proxies for fault energy buildup) into a unified ML pipeline. Moreover, fusing GNSS displacement data with SAR/InSAR and optical imagery could yield rich, multimodal feature sets for blind-source separation or representation learning of surface deformations. In these instances, applying graph neural networks over the spatial network of stations could be investigated, as it can exploit geospatial adjacency and tectonic feature information.

Creating ground-truth catalogues, which include magnitude, depth, focal mechanism, and locations, could then enable us to benchmark extracted sources against physics-based rupture models, potentially predicting event occurrence with categorical labels or step-function station displacements as a function of distance from the epicentre. Moreover, physics-informed and hybrid frameworks could embed core constraints (mass conservation, elastic dislocation theory, plate-boundary stress accumulation) directly into network architectures or loss functions, ensuring data-driven insights remain geophysically sound.

Unlike mature domains such as medical AI, geological datasets lack standardised benchmarks and uniform usage. Establishing an open suite that combines seismic waveforms, GNSS, InSAR, and agreed-upon evaluation protocols (SI-SDRi, MCC, event labels) would accelerate AI/ machine learning progress in the field of geodesy. By integrating expertise from seismology, geospatial AI, earth-system modelling, and hazard management, we can refine core algorithms and fast-track their translation into societally relevant applications.

Bibliography

[1] World Health Organisation, "Earthquakes: Overview," https://www.who.int/health-topics/earthquakes, Accessed: 2024-05-07. (document)

[2] Geoffrey Blewitt, William Hammond, and Corn Kreemer, "Harnessing the GPS data explosion for interdisciplinary science," *Eos*, vol. 99, 09 2018. (document), 1, 1.3, 1.2, 1.5, 2.3.1, 8.3

[3] Adriano Gualandi, Zhen Liu, and Chris Rollins, "Post-large earthquake seismic activities mediated by aseismic deformation processes," *Earth and Planetary Science Letters*, vol. 530, pp. 115870, 10 2019. (document), 2.2, 4.1, 4.2, 6.3, 6.6

[4] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe, "Whitening for self-supervised representation learning," 2021. (document), 1.4, 2.6.2, 4.2.3, 4.3.1, 4.4, 5.2, 5.3, 5.2.2, 5.2.2, 7.1, 7.2, 7.3.2

[5] Sylvain Michel, Adriano Gualandi, and JeanPhilippe Avouac, "Interseismic coupling and slow slip events on the Cascadia megathrust," *Pure and Applied Geophysics*, vol. 176, 09 2019. (document), 1.1, 1.2, 2.2, 5.2, 5.5, 6.2.1, 6.2

[6] Agnan Kessy, Alex Lewin, and Korbinian Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, Oct. 2018. (document), 2.6.2, 4.1, 4.1, 4.1, 4.1, 4.1.1, 1, 2, 3, 4, 5, 4.2, 4.2.2, 4.1, 4.2, 4.2.3

[7] Shannon Doocy, Amy Daniels, Catherine Packer, Anna Dick, and Thomas Kirsch, "The human impact of earthquakes: A historical review of events 1980-2009 and systematic literature review," *PLoS currents*, vol. 5, 04 2013. 1

[8] Yasuhide Okuyama, Geoffrey J. D. Hewings, and Michael Sonis, *Measuring Economic Impacts of Disasters: Interregional Input-Output Analysis Using Sequential Interindustry Model*, pp. 77–101, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. 1

[9] International GNSS Service, "IGS14 reference frame," http://www.igs.org/article/igs14-reference-frame-transition, 2020. 1.1

[10] Herb Dragert, Kelin Wang, and Thomas S James, "A silent slip event on the deeper cascadia subduction interface," *Science*, vol. 292, no. 5521, pp. 1525–1528, 2001. 1.1

[11] A Gualandi, E Serpelloni, and M. E. Belardinelli, "Blind source separation problem in GPS time series," *Journal of Geodesy*, vol. 90, pp. 323–341, Apr. 2016. 1.1, 1.1.2, 1.2, 1.2, 1.2, 2, 2.2, 2.2, 5.2, 5.5

[12] B. Hofmann-Wellenhof, H. Lichtenegger, and E. Wasle, *GNSS – Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and more*, Springer Vienna, 2007. 1.2

[13] J. Sickle, "GPS and GNSS for Geospatial Professionals," `https://e-education.psu.edu/geog862/node/1759`, Accessed: 2020-05-25. 1.2

[14] Peter Cederholm, "Statistical characteristics of l1 carrier phase observations from four low-cost gps receivers," *Nordic Journal of Surveying and Real Estate Research*, vol. 7, no. 1, Oct. 2010. 1.2

[15] Jiaji Wu, Jinguang Jiang, Yanan Tang, and Jianghua Liu, "Gaussian–student's t mixture distribution-based robust kalman filter for global navigation satellite system/inertial navigation system/odometer data fusion," *Remote Sensing*, vol. 16, no. 24, 2024. 1.2

[16] S.V. Hum, "Atmospheric effects," `http://www.waves.utoronto.ca/prof/svhum/ece422/notes/20a-atmospheric-refr.pdf`, Accessed: 2020-05-25. 1.2

[17] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," 2020. 1.3, 2.5.2, 2.6.1, 7.1

[18] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton, "Towards nonlinear disentanglement in natural data with temporal sparse coding," 2020. 1.3, 2.5.3, 7.1, 7.3.2, 7.4.2

[19] Andreas Geiger, Philip Lenz, Christian Stiller, and Raquel Urtasun, "Kitti-masks dataset," Zenodo, Version 1.0, Jul 2020, Based on the original KITTI Segmentation challenge at https://www.vision.rwth-aachen.de/page/mots. 1.3

[20] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville, "Mutual information neural estimation," in *ICML*, 2018, pp. 530–539. 1.4, 2.3.1, 2.3.1, 3.2.3, 5.1

[21] Xingjian Zhen, Zihang Meng, Rudrasis Chakraborty, and Vikas Singh, "On the versatile uses of partial distance correlation in deep learning," in *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds., Cham, 2022, pp. 327–346, Springer Nature Switzerland. 1.5, 2.3.2

[22] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001. 2.1

[23] Aapo Hyvärinen and Erkki Oja, "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 07 1997. 2.1, 5.2

[24] A. Hyvarinen, "One-unit contrast functions for independent component analysis: a statistical analysis," in *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, 1997, pp. 388–397. 2.1

[25] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000. 2.1, 2, 5.2

[26] ChiaHsiang Yang, Yi-Hsin Shih, and Herming Chiueh, "An 81.6 FastICA processor for epileptic seizure detection," *IEEE transactions on biomedical circuits and systems*, vol. 9, 06 2014. 2.1

[27] Bruce A. Draper, Kyungim Baek, Marian Stewart Bartlett, and J.Ross Beveridge, "Recognizing faces with PCA and ICA," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 115–137, 2003, Special Issue on Face Recognition. 2.1

[28] Mohamed Farhat, Yasser Gritli, and Mohamed Benrejeb, "Fast–ICA for mechanical fault detection and identification in electromechanical systems for wind turbine applications," *International Journal of Advanced Computer Science and Applications*, vol. 8, 01 2017. 2.1

[29] Scott C Douglas, "A statistical convergence analysis of the FastICA algorithm for two-source mixtures," in *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.* IEEE, 2005, pp. 335–339. 2.1

[30] Petr Tichavsky, Zbyněk Koldovský, and Erkki Oja, "Performance analysis of the FastICA algorithm and cramér-rao bounds for linear independent component analysis," *Signal Processing, IEEE Transactions on*, vol. 54, pp. 1189 – 1203, 05 2006. 2.1

[31] Esa Ollila, "The deflation-based fastica estimator: Statistical analysis revisited," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1527–1541, 2010. 2.1

[32] Hao Shen, Martin Kleinsteuber, and Knut Hueper, "Local convergence analysis of FastICA and related algorithms," *Neural Networks, IEEE Transactions on*, vol. 19, pp. 1022 – 1032, 07 2008. 2.1

[33] Tianwen Wei, "On the spurious solutions of the Fastica algorithm," in *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, 2014, pp. 161–164. 2.1

[34] J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998. 2.1

[35] Z. Koldovsky, P. Tichavsky, and E. Oja, "Cramer-rao lower bound for linear independent component analysis," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, vol. 3, pp. iii/581–iii/584 Vol. 3. 2.1

[36] Ella Bingham and Aapo Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *International Journal of Neural Systems*, vol. 10, no. 01, pp. 1–8, 2000, PMID: 10798706. 2.1

[37] Hualiang Li and Tülay Adali, "Algorithms for complex ML ICA and their stability analysis using wirtinger calculus," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 6156 – 6167, 01 2011. 2.1

[38] Yang Zhang and Saleem Kassam, "Optimum nonlinearity and approximation in complex FastICA," *2012 46th Annual Conference on Information Sciences and Systems, CISS 2012*, pp. 1–6, 03 2012. 2.1

[39] Benedikt Loesch and Bin Yang, "Cramér-rao bound for circular complex independent component analysis," in *Latent Variable Analysis and Signal Separation*, Fabian Theis, Andrzej Cichocki, Arie Yeredor, and Michael Zibulevsky, Eds., Berlin, Heidelberg, 2012, pp. 42–49, Springer Berlin Heidelberg. 2.1

[40] Francis R Bach and Michael I Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002. 2.1

[41] Shunichi Amari, Andrzej Cichocki, and Howard Yang, "A new learning algorithm for blind signal separation," *Advances in neural information processing systems*, vol. 8, 1995. 2.1, B.0.2

[42] EG Learned-Miller and JW Fisher, "ICA using spacings estimates of entropy," *Journal of Machine Learning Research*, vol. 4, pp. 1271–1295, 10 2004. 2.1, 5.4

[43] Aapo Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," *Advances in neural information processing systems*, vol. 10, 1997. 2.1

[44] Petr Tichavsky, Zbyněk Koldovský, and Erkki Oja, "Speed and accuracy enhancement of linear ICA techniques using rational nonlinear functions," in *International Conference on Independent Component Analysis and Signal Separation*, 01 2007, pp. 285–292. 2.1

[45] Jih Cheng Chao and Scott C. Douglas, "Using piecewise linear nonlinearities in the natural gradient and FastICA algorithms for blind source separation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1813–1816. 2.1

[46] Vicente Zarzoso, Pierre Comon, and Kallel Mariem, "How fast is FastICA," *European Signal Processing Conference*, 01 2006. 2.1

[47] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Transactions on Neural Networks*, vol. 21, no. 2, pp. 248–261, Feb. 2010. 2.1

[48] Elena Issoglio, Paul Smith, and Jochen Voss, "On the estimation of entropy in the FastICA algorithm," 2018. 2.1

[49] M. Bottiglieri, M. Falanga, U. Tammaro, F. Obrizzo, P. De Martino, C. Godano, and F. Pingue, "Independent component analysis as a tool for ground deformation analysis," *Geophysical Journal International*, vol. 168, no. 3, pp. 1305–1310, 03 2007. 2.2

[50] Ehsan Forootan and Jürgen Kusche, "Separation of global time-variable gravity signals into maximally independent components," *Journal of Geodesy*, vol. 86, pp. 477 – 497, 07 2012. 2.2

[51] Ehsan Forootan and Jürgen Kusche, "Separation of deterministic signals using independent component analysis (ICA)," *Studia Geophysica et Geodaetica*, vol. 57, pp. 17–26, 2013. 2.2

[52] J.F. Cardoso and Antoine Souloumiac, "Blind beamforming for non gaussian signals," *Radar and Signal Processing, IEE Proceedings F*, vol. 140, pp. 362 – 370, 01 1994. 2.2

[53] Rowena Lohman and Jessica Murray, "The scec geodetic transient-detection validation exercise," *Seismological Research Letters*, vol. 84, pp. 419–425, 05 2013. 2.2

[54] Rizwan A Choudrey, *Variational methods for Bayesian independent component analysis*, Ph.D. thesis, University of Oxford Oxford, UK, 2002. 2.2, 2.2

[55] David J. C. MacKay, *Bayesian Non-Linear Modeling for the Prediction Competition*, pp. 221–234, Springer Netherlands, Dordrecht, 1996. 2.2

[56] Adriano Gualandi and Zhen Liu, "Variational bayesian independent component analysis for InSAR displacement time-series with application to central California, USA," *Journal of Geophysical Research: Solid Earth*, vol. 126, no. 4, pp. e2020JB020845, 2021. 2.2, 6.3

[57] Eugenio Mandler, Francesco Pintori, Adriano Gualandi, Letizia Anderlini, Enrico Serpelloni, and Maria Elina Belardinelli, "Post-seismic deformation related to the 2016 Central Italy seismic sequence from GPS displacement time-series," *Journal of Geophysical Research: Solid Earth*, vol. 126, no. 9, pp. e2021JB022200, 2021. 2.2

[58] Krittanon Sirorattanakul, Zachary E Ross, Mostafa Khoshmanesh, Elizabeth S Cochran, Mateo Acosta, and JeanPhilippe Avouac, "The 2020 Westmorland, California earthquake swarm as aftershocks of a slow slip event sustained by fluid flow," *Journal of Geophysical Research: Solid Earth*, vol. 127, no. 11, 2022. 2.2

[59] Warren M. Lord, Jie Sun, and Erik M. Bollt, "Geometric k-nearest neighbor estimation of entropy and mutual information," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 3, pp. 033114, mar 2018. 2.3.1

[60] Yanan Fan, Pierre Lafaye de Micheaux, Spiridon Penev, and Donna Salopek, "Multivariate nonparametric test of independence," *Journal of Multivariate Analysis*, vol. 153, pp. 189–210, 2017. 2.3.1

[61] Gabor Székely and Maria Rizzo, "Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method," *Journal of Classification*, vol. 22, pp. 151–183, 02 2005. 2.3.1

[62] GJ Székely, ML Rizzo, and NK Bakirov, "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007. 2.3.2, 2.3.2

[63] Gábor J. Székely and Maria L. Rizzo, "Partial distance correlation with methods for dissimilarities," *The Annals of Statistics*, vol. 42, no. 6, pp. 2382 – 2412, 2014. 2.3.2, 2.3.2

[64] Arin Chaudhuri and Wenhao Hu, "A fast algorithm for computing distance correlation," *Computational statistics & data analysis*, vol. 135, pp. 15–24, 2019. 2.3.2

[65] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola, "A kernel statistical test of independence," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. 2007, vol. 20, Curran Associates, Inc. 2.3.2

[66] Arthur Gretton and László Györfi, "Consistent nonparametric tests of independence," *The Journal of Machine Learning Research*, vol. 11, pp. 1391–1423, 2010. 2.3.2

[67] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu, "Equivalence of distance-based and rkhs-based statistics in hypothesis testing," *The annals of statistics*, pp. 2263–2291, 2013. 2.3.2

[68] Cencheng Shen and Joshua T Vogelstein, "The exact equivalence of distance and kernel methods in hypothesis testing," *AStA Advances in Statistical Analysis*, vol. 105, pp. 385–403, 2021. 2.3.2

[69] Thomas R. Knapp, "Canonical correlation analysis: A general parametric significance-testing system," *Psychologica Bulletin*, p. 410–416, 1978. 2.3.3

[70] Harold Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992. 2.3.3

[71] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004. 2.3.3

[72] Wolfgang Karl Härdle, Léopold Simar, and Wolfgang Karl Härdle, "Canonical correlation analysis," *Applied multivariate statistical analysis*, pp. 443–454, 2015. 2.3.3

[73] Y Escoufier, "Le traitement des variables vectorielles. biometrics 29 751–760," *Mathematical Reviews (MathSciNet): MR334416 Digital Object Identifier: doi*, vol. 10, pp. 2529140, 1973. 2.3.3

[74] Paul Robert and Yves Escoufier, "A unifying tool for linear multivariate statistical methods: the rv-coefficient," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 25, no. 3, pp. 257–265, 1976. 2.3.3

[75] Ruth Heller, Yair Heller, and Malka Gorfine, "A consistent multivariate test of association based on ranks of distances," *Biometrika*, vol. 100, no. 2, pp. 503–510, 2013. 2.3.3

[76] Ruth Heller and Yair Heller, "Multivariate tests of association based on univariate tests," *Advances in Neural Information Processing Systems*, vol. 29, 2016. 2.3.3

[77] Sambit Panda, Cencheng Shen, and Joshua T. Vogelstein, "Learning interpretable characteristic kernels via decision forests," 2023. 2.3.3

[78] Cencheng Shen and Yuexiao Dong, "High-dimensional independence testing via maximum and average distance correlations," 2024. 2.3.3

[79] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006. 2.4, 2.4

[80] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR - half-baked or well done?," 2018. 2.4, 5.3.1

[81] Geoffrey Blewitt, "An improved equation of latitude and a global system of graticule distance coordinates," *Journal of Geodesy*, vol. 98, no. 1, pp. 6, 2024. 2.5.1

[82] Geoffrey Blewitt, Corné Kreemer, William C Hammond, and Julien Gazeaux, "Midas robust trend estimator for accurate GPS station velocities without step detection," *Journal of Geophysical Research: Solid Earth*, vol. 121, no. 3, pp. 2054–2068, 2016. 2.5.1

[83] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35. 2.5.2

[84] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. 2.5.2

[85] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech 2019*, 2019, pp. 1368–1372. 2.5.2, 2.6.1

[86] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe, "Mots: Multi-object tracking and segmentation," 2019. 2.5.3

[87] Alex Krizhevsky, "Learning multiple layers of features from tiny images," *University of Toronto*, 05 2012. 2.5.4, 7.1, 7.3.2

[88] ChaoLing Hsu and JyhShing Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 310 – 319, 03 2010. 2.6.1

[89] PoSen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566. 2.6.1

[90] Joan Bruna, Pablo Sprechmann, and Yann LeCun, "Source separation with scattering non-negative matrix factorization," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1876–1880. 2.6.1

[91] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2015. 2.6.1

[92] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, IEEE. 2.6.1

[93] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27(8), pp. 1256–1266, 2019. 2.6.1, 7.4.1

[94] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50. 2.6.1

[95] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," 2020. 2.6.1

[96] Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu, "Sandglasset: A Light Multi-Granularity Self-attentive Network For Time-Domain Speech Separation," 2021. 2.6.1

[97] Dong Yu, Morten Kolbæk, ZhengHua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," 2017. 2.6.1

[98] GenePing Yang, ChaoI Tuan, HungYi Lee, and Linshan Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," 2019. 2.6.1

[99] SungFeng Huang, ShunPo Chuang, DaRong Liu, YiChen Chen, GenePing Yang, and Hungyi Lee, "Stabilizing label assignment for speech separation by self-supervised pre-training," 2021. 2.6.1

[100] Kai Li, Runxuan Yang, and Xiaolin Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," 2023. 2.6.1

[101] Shahar Lutati, Eliya Nachmani, and Lior Wolf, "Separate and diffuse: Using a pretrained diffusion model for improving source separation," 2023. 2.6.1

[102] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jiaqi Yip, Dianwen Ng, and Bin Ma, "MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation," 2023. 2.6.1

[103] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25. 2.6.1, 5.2, 7.4.1

[104] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive multiview coding," 2020. 2.6.2, 2.6.2

[105] Tongzhou Wang and Phillip Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," 2022. 2.6.2, 7.2

[106] Kihyuk Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. 2016, vol. 29, Curran Associates, Inc. 2.6.2

[107] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, IEEE. 2.6.2

[108] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," 2017. 2.6.2

[109] Xiaolong Wang and Abhinav Gupta, "Unsupervised learning of visual representations using videos," 2015. 2.6.2

[110] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," 2016. 2.6.2

[111] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio, "Learning deep representations by mutual information estimation and maximization," 2019. 2.6.2, 2.6.2

[112] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," 2019. 2.6.2, 2.6.2, 2.6.2

[113] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304. 2.6.2

[114] Geoffrey Roeder, Luke Metz, and Diederik P. Kingma, "On linear identifiability of learned representations," 2020. 2.6.2

[115] Aapo Hyvarinen, Hiroaki Sasaki, and Richard E. Turner, "Nonlinear ica using auxiliary variables and generalized contrastive learning," 2019. 2.6.2

[116] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel, "Contrastive learning inverts the data generating process," 2022. 2.6.2, 7.1, 7.3.2, 7.4.2, 7.4.2

[117] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," 2020. 2.6.2

[118] Philip Bachman, R Devon Hjelm, and William Buchwalter, "Learning representations by maximizing mutual information across views," 2019. 2.6.2

[119] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988. 2.6.2

[120] Takuya Nishimura, "Time-independent forecast model for large crustal earthquakes in southwest japan using gnss data," *Earth, Planets and Space*, vol. 74, no. 1, pp. 58, 2022. 2.6.3

[121] Xuechuan Li, Changyun Chen, Hongbao Liang, Yu Li, and Wei Zhan, "Earthquake source parameters estimated from high-rate multi-gnss data: a case study of the 2022 m 6.9 menyuan earthquake," *Acta Geophysica*, vol. 71, no. 2, pp. 625–636, 2023. 2.6.3

[122] Jiun-Ting Lin, Diego Melgar, Valerie J Sahakian, Amanda M Thomas, and Jacob Searcy, "Real-time fault tracking and ground motion prediction for large earthquakes with hr-gnss and deep learning," *Journal of Geophysical Research: Solid Earth*, vol. 128, no. 12, pp. e2023JB027255, 2023. 2.6.3

[123] Claudia Quinteros-Cartaya, Jonas Köhler, Wei Li, Johannes Faber, and Nishtha Srivastava, "Exploring a cnn model for earthquake magnitude estimation using hr-gnss data," *Journal of South American earth sciences*, vol. 136, pp. 104815, 2024. 2.6.3

[124] Diego Brum, Mauricio Roberto Veronez, Eniuce Menezes de Souza, Ismael Érique Koch, Luiz Gonzaga, Ivandro Klein, Marcelo Tomio Matsuoka, Vinicius Francisco Rofatto, Ademir Marques Junior, Graciela Eliane dos Reis Racolte, et al., "A proposed earthquake warning system based on ionospheric anomalies derived from GNSS measurements and artificial neural networks," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 9295–9298. 2.6.3

[125] Shivani Joshi, Suresh Kannaujiya, and Utkarsh Joshi, "Analysis of gnss data for earthquake precursor studies using ionolab-tec in the himalayan region," *Quaternary*, vol. 6, no. 2, pp. 27, 2023. 2.6.3

[126] Zheng Liu, Keliang Zhang, Weijun Gan, and Shiming Liang, "Refined coseismic slip and afterslip distributions of the 2021 mw 6.1 yangbi earthquake based on gnss and insar observations," *Remote Sensing*, vol. 16, no. 21, pp. 3996, 2024. 2.6.3

[127] Clayton MJ Brengman and William D Barnhart, "Identification of surface deformation in INSAR using machine learning," *Geochemistry, Geophysics, Geosystems*, vol. 22, no. 3, pp. e2020GC009204, 2021. 2.6.3

[128] Nantheera Anantrasirichai, Juliet Biggs, Fabien Albino, Paul Hill, and David Bull, "Application of machine learning to classification of volcanic deformation in routinely generated INSAR data," *Journal of Geophysical Research: Solid Earth*, vol. 123, no. 8, pp. 6592–6606, 2018. 2.6.3

[129] A Novellino, M Cesarano, P Cappelletti, D Di Martire, M Di Napoli, M Ramondini, A Sowter, and D Calcaterra, "Slow-moving landslide risk assessment combining machine learning and INSAR techniques," *Catena*, vol. 203, pp. 105317, 2021. 2.6.3

[130] Seyed Amir Naghibi, Behshid Khodaei, and Hossein Hashemi, "An integrated INSAR-machine learning approach for ground deformation rate modeling in arid areas," *Journal of Hydrology*, vol. 608, pp. 127627, 2022. 2.6.3

[131] Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell, "Learning to navigate in cities without a map," 2019. 2.6.3

[132] Akpojoto Siemuri, Kannan Selvan, Heidi Kuusniemi, Petri Valisuo, and Mohammed S Elmusrati, "A systematic review of machine learning techniques for GNSS use cases," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 6, pp. 5043–5077, 2022. 2.6.3

[133] Fabio Corbi, Jonathan Bedford, Laura Sandri, Francesca Funiciello, Adriano Gualandi, and Mathias Rosenau, "Predicting imminence of analog megathrust earthquakes with machine learning: Implications for monitoring subduction zones," *Geophysical Research Letters*, vol. 47, no. 7, pp. e2019GL086615, 2020. 2.6.3

[134] Matthias Rosenau, Fabio Corbi, and Stephane Dominguez, "Analogue earthquakes and seismic cycles: Experimental modelling across timescales," *Solid Earth*, vol. 8, no. 3, pp. 597–635, 2017. 2.6.3

[135] A. Gualandi, J.P. Avouac, S. Michel, and D. Faranda, "The predictable chaos of slow earthquakes," *Science Advances*, vol. 6, no. 27, pp. eaaz5548, 2020. 2.6.3, 8.3

[136] A. Gualandi, D. Faranda, C. Marone, M. Cocco, and G. Mengaldo, "Deterministic and stochastic chaos characterize laboratory earthquakes," *Earth and Planetary Science Letters*, vol. 604, pp. 117995, 2023. 2.6.3

[137] Robert G. Gallager, *Low-Density Parity-Check Codes*, MA: MIT Press, 1963. 3.1

[138] Francis D. Natali and William J. Walbesser, "Phase-locked-loop detection of binary psk signals utilizing decision feedback," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-5, no. 1, pp. 83–90, 1969. 3.1

[139] John B. Anderson, Tor Aulin, and Carl-Erik Sundberg, *Digital Phase Modulation*, Plenum Press, 1986. 3.1

[140] Alireza Nooraiepour and Sina Rezaei Aghdam, "Learning end-to-end codes for the BPSK-constrained gaussian wiretap channel," *Physical Communication*, vol. 46, pp. 101282, 2021. 3.1

[141] Alvaro Santamaría-Gómez, MarieNoëlle Bouin, Xavier Collilieux, and Guy Wöppelmann, "Correlated errors in GPS position time series: Implications for velocity estimates," *Journal of Geophysical Research: Solid Earth*, vol. 116, no. B1, 2011. 3.1, 3.3

[142] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy, "Deep variational information bottleneck," 2019. 3.2.3, 3.2.3

[143] Jens Timmer and Michel Koenig, "On generating power law noise.," *Astronomy and Astrophysics, v. 300, p. 707*, vol. 300, pp. 707, 1995. 3.3, A

[144] Sebastian Schlecht, Benoit Alary, Vesa Välimäki, and Emanüel A P Habets, "Optimized velvet-noise decorrelator," in *Proceedings of the International Conference on Digital Audio Effects*, Portugal, Sept. 2018, Proceedings of the International Conference on Digital Audio Effects, pp. 87–94, University of Aveiro, International Conference on Digital Audio Effects, DAFX ; Conference date: 04-09-2018 Through 08-09-2018. 3.3

[145] Anthony J Bell and Terrence J Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision research*, vol. 37, no. 23, pp. 3327–3338, 1997. 1

[146] Bruno A Olshausen and David J Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997. 1

[147] Jerome H Friedman, "Exploratory projection pursuit," *Journal of the American statistical association*, vol. 82, no. 397, pp. 249–266, 1987. 2

[148] Mohsen Pourahmadi, "Covariance estimation: The glm and regularization perspectives," *Statistical Science*, pp. 369–387, 2011. 5

[149] Hlynur Davíð Hlynsson and Laurenz Wiskott, "Learning gradient-based ICA by neurally estimating mutual information," in *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 2019, pp. 182–187. 4.2.1, 5.1, 5.2, 5.2.1, B.0.2

[150] Y.C. Eldar and A.V. Oppenheim, "MMSE whitening and subspace whitening," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1846–1851, 2003. 4.2.3

[151] Shiro Ikeda, "Ica on noisy data: A factor analysis approach," in *Advances in independent component analysis*, pp. 201–215. Springer, 2000. 4.2.3

[152] Nail K Bakirov, Maria L Rizzo, and Gábor J Székely, "A multivariate nonparametric test of independence," *Journal of multivariate analysis*, vol. 97, no. 8, pp. 1742–1756, 2006. 5.1, 5.4.1, 7.1

[153] Naoya Takahashi and Yuki Mitsufuji, "D3Net: Densely connected multidilated DenseNet for music source separation," 2020. 5.2

[154] R. A. Choudrey and S. J. Roberts, "Variational Mixture of Bayesian Independent Component Analyzers," *Neural Computation*, vol. 15, no. 1, pp. 213–252, 01 2003. 5.2

[155] Susanna Ebmeier, "Application of independent component analysis to multitemporal INSAR data with volcanic case studies," *Journal of Geophysical Research. Solid Earth*, vol. 121, 11 2016. 5.2

[156] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 5.2

[157] Lampros Mouselimis, *CopernicusDEM: Copernicus Digital Elevation Models*, 2024, R package version 1.0.3 produced using Copernicus WorldDEMTM-90 DLR e.V. 2010-2014 and Airbus Defence and Space GmbH 2014-2018 provided under COPERNICUS by the European Union and ESA; all rights reserved. 5.3.2

[158] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic, "On mutual information maximization for representation learning," *arXiv preprint arXiv:1907.13625*, 2019. 5.4.2, 8.2

[159] Zhetao Zhang, Bofeng Li, and Yunzhong Shen, "Comparison and analysis of unmodelled errors in GPS and BeiDou signals," *Geodesy and Geodynamics*, vol. 8, no. 1, pp. 41–48, 2017. 6.1

[160] Zhanhong Huang, Lei Xie, Lei Zhao, and Wenbin Xu, "Spatiotemporal Distribution of Afterslip following the 2014 Yutian Mw 6.9 Earthquake Using COSMO-SkyMed and Sentinel-1 InSAR Data," *Remote Sensing*, vol. 15, no. 9, 2023. 6.2.1

[161] Jianwen Xie, Pamela K. Douglas, Ying Nian Wu, Arthur L. Brody, and Ariana E. Anderson, "Decoding the Encoding of Functional Brain Networks: an fMRI Classification Comparison of Non-negative Matrix Factorization (NMF), Independent Component Analysis (ICA), and Sparse Coding Algorithms," 2016. 7.1

[162] Chippy Jayaprakash, Bharath Bhushan Damodaran, Sowmya Viswanathan, and Kutti Padannayil Soman, "Randomized independent component analysis and linear discriminant analysis dimensionality reduction methods for hyperspectral image classification," *Journal of Applied Remote Sensing*, vol. 14, no. 03, pp. 1, July 2020. 7.1

[163] Rabel Guharoy, Nanda Dulal Jana, Suparna Biswas, and Lalit Garg, "Empirical analysis of different dimensionality reduction and classification techniques for epileptic seizure detection," 2023. 7.1

[164] Hermanni Hälvä and Aapo Hyvärinen, "Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series," 2020. 7.1

[165] Alexander Schell and Harald Oberhauser, "Nonlinear independent component analysis for discrete-time and continuous-time signals," 2023. 7.1

[166] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, JuChieh Chou, SungLin Yeh, SzuWei Fu, ChienFeng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "Speechbrain: A general-purpose speech toolkit," 2021. 7.1, 7.3.1

[167] Neil Zeghidour and David Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021. 7.4.1

[168] Ui-Hyeop Shin, Sangyoun Lee, Taehan Kim, and Hyung-Min Park, "Separate and reconstruct: Asymmetric encoder-decoder for speech separation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 52215–52240, 2024. 7.4.1

[169] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen, "Weakly-supervised disentanglement without compromises," *CoRR*, vol. abs/2002.02886, 2020. 7.4.2

[170] Bruno A Olshausen and David J Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996. 7.4.2

APPENDIX A

Additional colours of noise

In this appendix, various colours of noise added to the BPSK signal are explored as an extension to Section 3.3.

Pink noise is a signal with a power per frequency interval that is inversely proportional to its frequency. The colourednoise Python module is used to generate the coloured noise, which is based on the algorithm in [143]. In the pink noise case, the exponent is 1. Figure A.1 shows the distance correlation and negentropy for transmission of a binary signal through a pink noise channel.

Blue noise has a power law spectrum with an exponent of -1, in relation to Equation 3.6. In this case the PSDgenerator function was used in this case to produce a random signal with the necessary power signal of -1. The distance correlation and negentropy for the transmission of a binary signal through an added blue noise channel can be seen in Figure A.2.

Violet noise is a signal with a frequency spectrum with a power per frequency interval that is proportional to f^2 of the signal over a finite range. In this case the PSDgenerator function was used in this case to produce a random signal with the necessary power signal of -2.

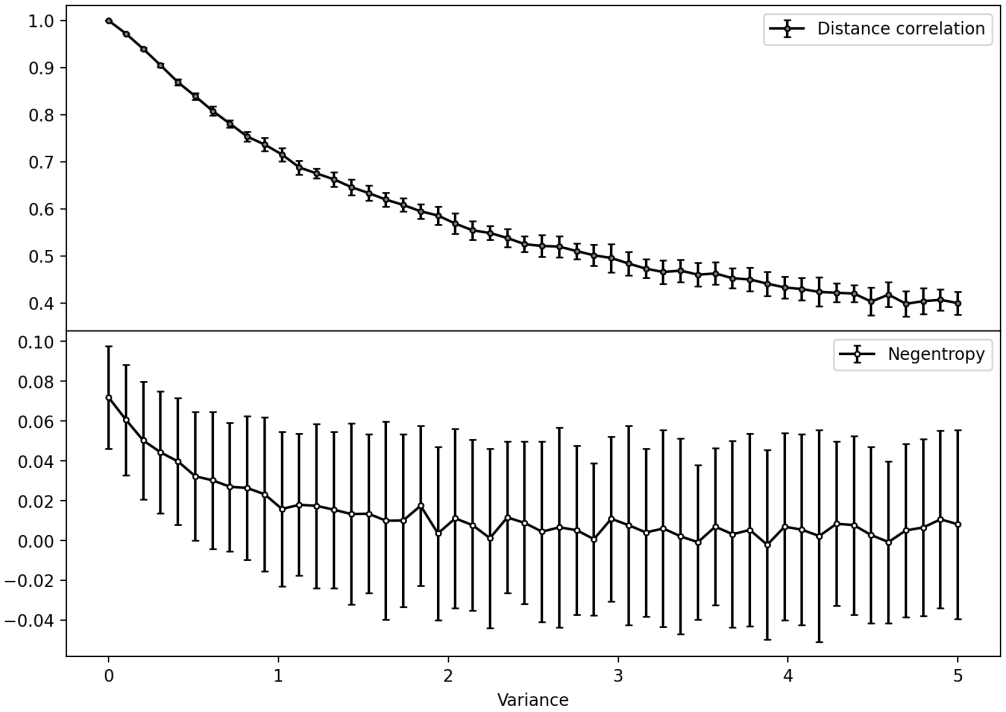


Figure A.1: Distance correlation and negentropy for transmission of a binary signal through an added pink noise channel, with assorted variances of the added noise signal.

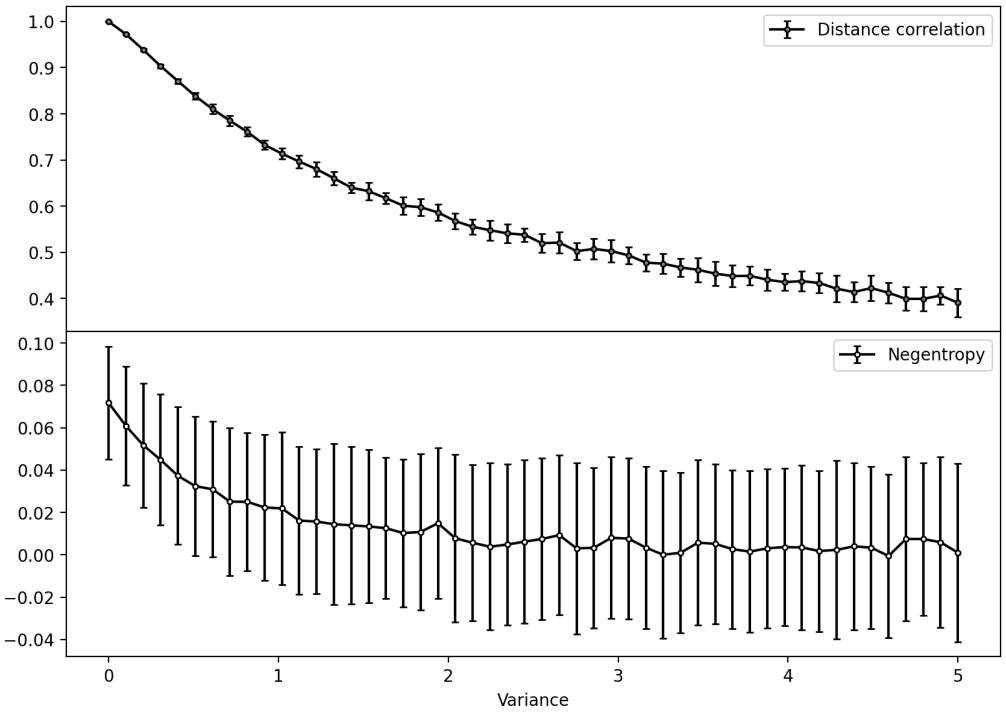


Figure A.2: Distance correlation and negentropy for transmission of a binary signal through an added blue noise channel, with assorted variances of the added noise signal.

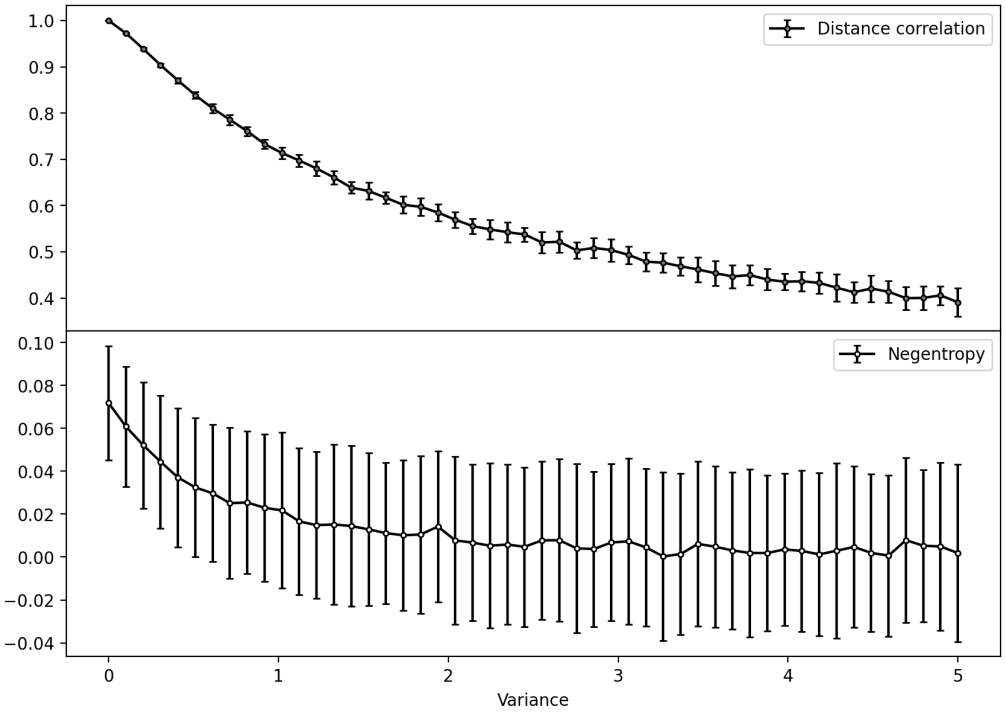


Figure A.3: Distance correlation and negentropy for transmission of a binary signal through an added violet noise channel, with assorted variances of the added noise signal.

APPENDIX B

A detailed analysis of the synthetic problem

This appendix contains work that is separate from the main thesis to preserve the narrative flow of the main text. In this section, I examine the small synthetic example, used in the main text, to determine whether machine learning, using a distance correlation loss function, effectively addresses BSS problems. This example features a limited set of simple signals that may have less complex content compared to real-world data. The results in this appendix are designed to provide evidence to answer several research questions on the effectiveness of distance correlation as a loss function in source separation.

I formulated several specific questions based on the findings from the chapters in the main text to explore fundamental aspects of distance correlation as a loss function, such as its capacity to describe independence and its optimisation. This appendix aims to handle these questions systematically. The four questions handled are:

1. Does the initial distance correlation correlate with the final distance correlation?
2. Does minimising distance correlation lead to the extraction of better sources?
3. Are the results of source separation using a distance correlation loss consistent?
4. What impact does whitening have on the training of the machine learning model?

To investigate these questions, I will use the 3-mix-3-source synthetic problem described in Section 4.2.1. I conducted an experiment to separate a sine, a square and a sawtooth wave from three mixtures using the Separation architecture (Figure 5.2a) and the average pairwise distance correlation as the loss function.

I did not use the Restart Algorithm (Algorithm 1) because it would not allow for an accurate evaluation of how well the loss function optimises for a simple problem. The Restart Algorithm compares the distance correlation for random initialisations with gradient descent to escape local minima. As a result, the number of times the global minimum is reached would be distorted and not reflective of training conducted without additional assistance.

The linear layer architecture used random weight initialisations for each repeat and was trained over 2,000 epochs with a learning rate of 0.001. This process was repeated 5,000 times, each time using a different random initialisation. The training was conducted under two scenarios: with and without ZCA whitening. In the first scenario, the whitening transformation was applied to the outputs generated by the linear layer architecture.

While this experiment is simple, the analysis of distance correlation as a loss function is thorough. Although it may not represent larger architectures or more complex problems, it enables an investigation into the behaviour of distance correlation as a loss function. The relationship between the previous questions and the behaviour of distance correlation as a loss function will be discussed in the following sections.

B.0.1 Correlation between initial and final distance correlations

In addressing Question 1, I examined whether there is a correlation between the initial and final distance correlation. Hypothetically, if the initial and final average distance correlations are strongly positively correlated, achieving global minima is only feasible if the initial source separation is sufficiently good. This analysis would reveal a limited ability of gradient descent to learn effective source separations when randomly initialised.

To examine the relationship between the initial distance correlation at epoch zero and the final distance correlation at the last epoch for a specific random weight initialisation, I created scatter plots illustrating the relationship between initial and final distance correlations. These can be found in Figure B.1, without (left) and with whitening (right).

Although the non-whitened and whitened plots in the appendix are often displayed side by side, examining these cases to understand the effects of whitening will be addressed later, specifically when tackling Question 4 in Section B.0.3.

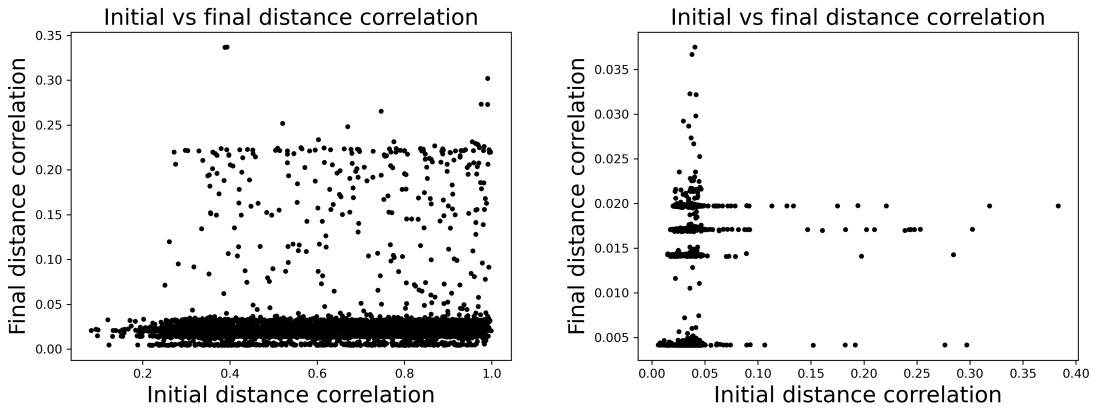


Figure B.1: Scatter plot of the initial vs final distance correlation values **without** (left) and **with** (right) the whitening step.

To address the first research question, one should refer to Figure B.1. In both the non-whitened and whitened examples, a low final distance correlation does not necessarily indicate that the initial distance correlation was also low. The correlation between the average initial and final distance correlations for the non-whitened examples was 0.131, while for the whitened examples, it was 0.0418. It is important to note that the final distance correlation is always lower than the initial distance correlation due to the optimisation step involved in the loss function.

Upon examining Figure B.1, it is evident that there is no positive slope in the data, indicating a lack of positive correlation between the initial and final distance correlations in both the whitened and non-whitened cases. This suggests that lower initial distance correlation values do not necessarily result in achieving the global minimum.

In Figure B.1, both the whitened and non-whitened scatter plots reveal the presence of clusters. The reader will observe this phenomenon in the whitened scatter plot due to its smaller range of final distance correlation values. There are several final distance correlation values associated with a wide variety of initial distance correlation values, which can be most clearly observed from the broader horizontal range for a limited number of final distance correlation values. These groupings will be referred to as clusters and will be discussed in more detail in the next section.

The only certainty is that the loss will decrease due to the optimisation definition (unless divergence occurs). Therefore, if the initial loss is below the second-lowest cluster's value, it is likely that training will reach the global minimum.

This complexity highlights the nature of the loss landscape, indicating that multiple runs of an algorithm (or the application of the Restart algorithm - Algorithm 1) may be necessary to reach the global minimum effectively.

To explore the relationship between initial and final distance correlations more thoroughly, I will introduce HiPlots, a visualisation method that creates parallel plots to represent high-dimensional data. I used HiPlots to analyse if the elements of the unmixing matrix at the first epoch (when randomly initialised) and at the final epoch (after training was completed) exhibited a clear relationship with the final distance correlation value. This study was mainly focused on the lowest cluster case (the global minimum). In Figure B.1, the reader can see that the lowest cluster value, present in both the non-whitened and whitened cases, occurs just below 0.005 (their related extracted time series are in Figures B.8 and B.14). I will create a HiPlot to analyse the final distance correlation values of this range. This analysis will help me to more effectively compare the patterns in the initial and final weights of the linear layer and the initial distance correlation associated with the global minimum. The cluster below 0.005 is the global minimum achieved across 5,000 repetitions of this experiment. If distance correlation is a valid metric, the extracted sources should be a sine, square, and sawtooth wave at the global minimum—similar to the known underlying signals. This extraction can occur with inversions, scaling, and changes in sign affecting the sources, with an equivalently permuted, scaled and sign-inverted unmixing matrix. If I normalise the sources and include the scaling factor in the unmixing matrix, there are 12 possible formulations for this matrix. Suppose the HiPlot displays repeated final weight values for every third component, or their negative counterparts, to represent permutations and sign changes. Specifically, there are six permutations of source order (3! = 6) and two possible sign values. This convergence would imply that the distance correlation solution is unique at the global maximum.

The four cases presented are the non-whitened case for all final distance correlation losses, those below 0.005 (representing the global minimum) and the same two sets of final distance correlations in the whitened case. The results can be seen in Figures B.2,

B.3, B.4, and B.5, respectively. If the reader wishes to view the outputted sources for the lowest cluster case, jump ahead to Figures B.8 and B.14 for the not whitened and the whitened source outputs.

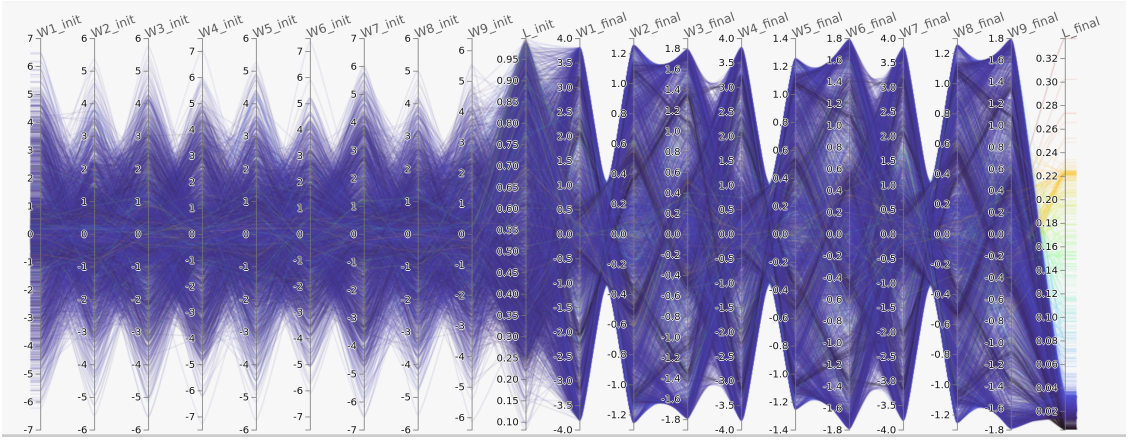


Figure B.2: HiPlot of the initial and final weights and the initial and final distance correlations for the case without whitening and for all final distance correlations.

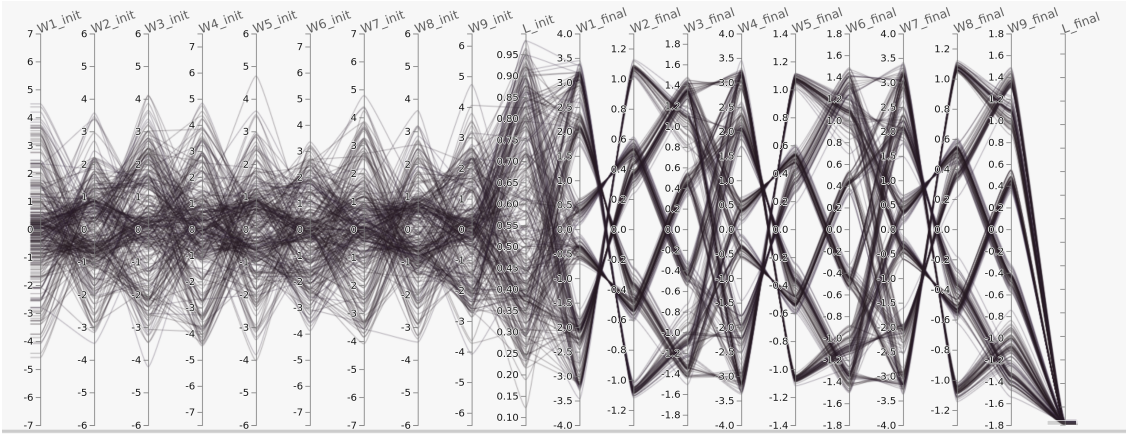


Figure B.3: HiPlot of the initial and final weights and the initial and final distance correlations for the case without whitening and for the range of final distance correlations below 0.005. This HiPlot displays the weight data for 179 initial random weight configurations.

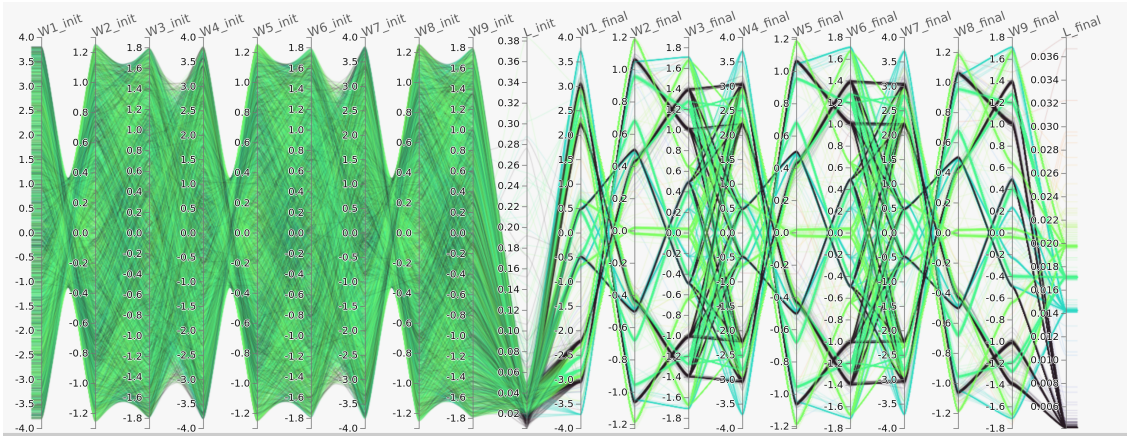


Figure B.4: HiPlot of the initial and final weights and the initial and final distance correlations for the whitened case and for all final distance correlations.

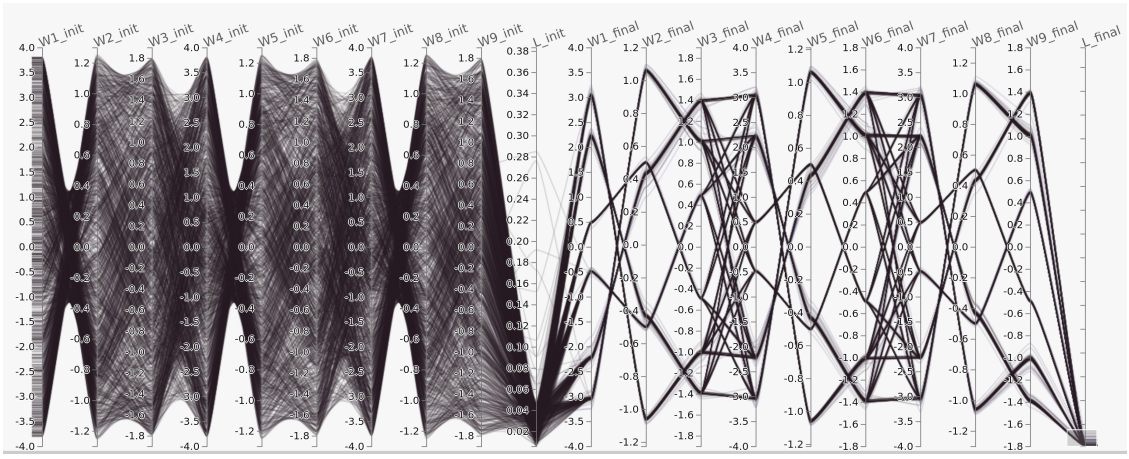


Figure B.5: HiPlot of the initial and final weights and the initial and final distance correlations for the whitened case and for the range of final distance correlations below 0.005. This HiPlot displays the weight data for 1,292 initial random weight configurations.

When comparing the initial and final average distance correlation weights shown in Figure B.3 and Figure B.5, there is no clear pattern of weight initialisations that corresponds to the lowest range group of the final average distance correlation. This lack of correlation is expected, as these initial weights relate to the initial average distance correlation loss, which does not align well with the final distance correlation loss.

The final weights associated with the lowest loss converge on a single solution when accounting for mirroring and permutation of the outputted sources. For instance, in Figure B.5, consider the weights $W1\_final$, $W4\_final$ and $W7\_final$, whose values are around 0.5, 2.2, and 3 in each instance. The three weight values allowed for the different

positions in the unmixing matrix to correspond to the various permutations of the outputted sources. The negative equivalents of these weights are included, as they indicate the possibility of mirroring of the outputted sources. Therefore, Figure B.5 addresses the uniqueness of the unmixing matrix at the global minimum. It suggests that the underlying sources are equivalent if scaling, mirroring, and permutations are considered. The findings indicate that the training converges on a single solution within the limitations of ICA methods. Next, I will assess whether the optimal distance correlation, used as an independence metric, at the global minimum accurately reflects the underlying sine, square, and sawtooth waves.

The relationship between the initial and final weights and how they correspond to the data associated with the lowest final distance correlation in both the non-whitened and whitened examples will be further explored in Sections B.0.2 and B.0.3.

To conclude this section, there was no correlation between the initial and final distance correlation values. This finding could simplify the training process by allowing users to select the lowest initial distance correlation. Additionally, if the initial weights are generally close to one of the final weight solutions for the lowest cluster, the training tends to converge to the global minimum, even though these weights are not known in advance. Hence, the benefit of the Restart Algorithm is to escape local minima. Furthermore, the final weights for the lowest cluster of distance correlations represent one set of output sources within the limitations of ICA.

B.0.2 Effectiveness of distance correlation in source separation

In this section, I will discuss Question 2. I aim to determine whether optimising a loss function based on distance correlation, used as a measure of pairwise source independence, results in sources that are also closer to the ground-truth sine, square, or sawtooth waves. To investigate this, I will use a test bed consisting of the small, known case to assess the effectiveness of distance correlation as a loss function for independent source separation.

As briefly mentioned in Section B.0.1 and illustrated in Figure B.1, the data clusters around several different final distance correlation values in both the non-whitened and whitened cases. These clusters display a range of initial distance correlation values.

By comparing the outputted sources at the end of training among these clusters, given that many data points correspond to only a few final distance correlations, I can explore whether a lower distance correlation associated with a cluster corresponds to weights that are closer to the inverse of the mixing matrix, measured using Amari distance. The Amari distance serves as a measure for how well the outputted sources align with the ground truth.

For the remaining questions, I will refer to the samples from the outputted sources grouped by similar final distance correlation values to facilitate interpretation. The selection of clusters was somewhat arbitrary, yet it was guided by kernel density estimation (KDE) plots.

These plots were generated using the *scipy.stats.kde.gaussian_kde* function for both whitened and non-whitened outputted source versions. Gaussian KDE is a technique to estimate the probability density function of a random variable by employing Gaussian probability distributions. In this context, the *gaussian_kde* function from the *scipy.stat* module offers a method for performing KDE using Gaussian kernels.

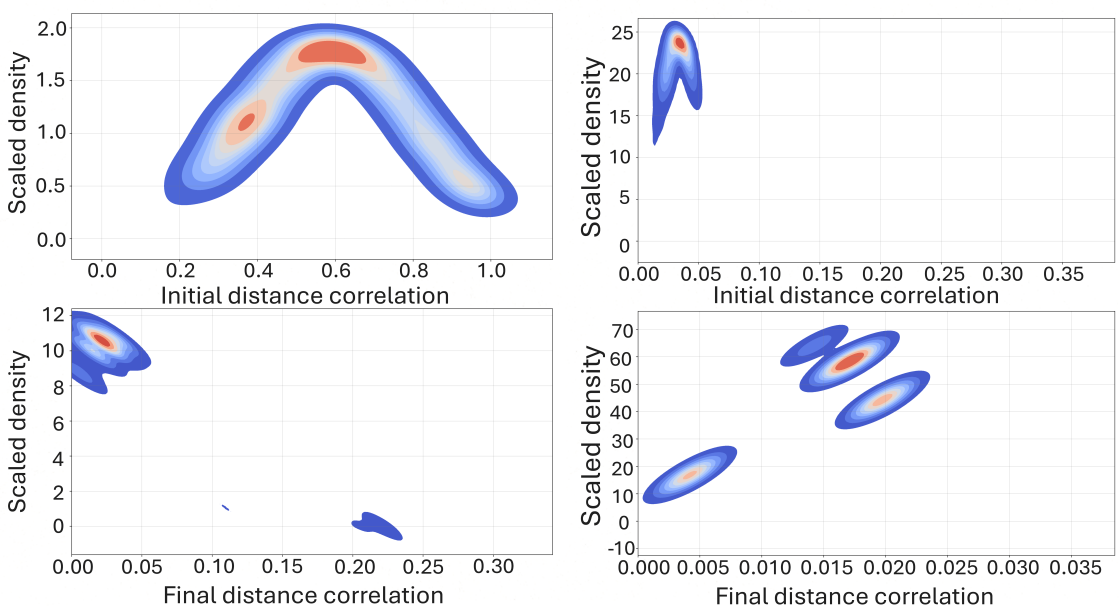


Figure B.6: **Non-Whitened** (left) and **whitened** (right) KDE plots of the initial (upper) and final (lower) probability densities for the average distance correlation between the three outputted sources. Note that the probability density has been scaled by the reciprocal of the standard deviation of the distance correlation data to allow for easier viewing.

There are two important observations to make from the left-hand side of Figure B.6. First, in the whitened case, the initial distance correlation has a smaller range, with most values being less than 0.05, when compared to the non-whitened case. Second, there are four distinct clusters of final distance correlations. Consequently, the five output sample's sources provided for comparison in the whitened case will be drawn from these four clusters. In the non-whitened case, the number of clusters is less clear; therefore, four clusters are selected for values below 0.05 to align with the whitened case, while an additional higher cluster is chosen for values above 0.05 based on the KDE plot.

The five selected cluster ranges are represented in a replot of the left-hand side of Figure B.1 as coloured bands. The replot is displayed in Figure B.7. It is important to note that the outputted sources are colour-coded according to their respective ranges, and these colours were chosen to resemble those associated with the final losses in the Hiplot, depicted in Figure B.2.

The five randomly selected output trios of sources for the non-whitened case with final distance correlation values less than 0.005 or final distance correlation ranges of 0.013 to 0.015, 0.019 to 0.023, 0.031 to 0.034 and 0.21 to 0.24 are shown in Figures B.8, B.9, B.10, B.11 and B.12, respectively.

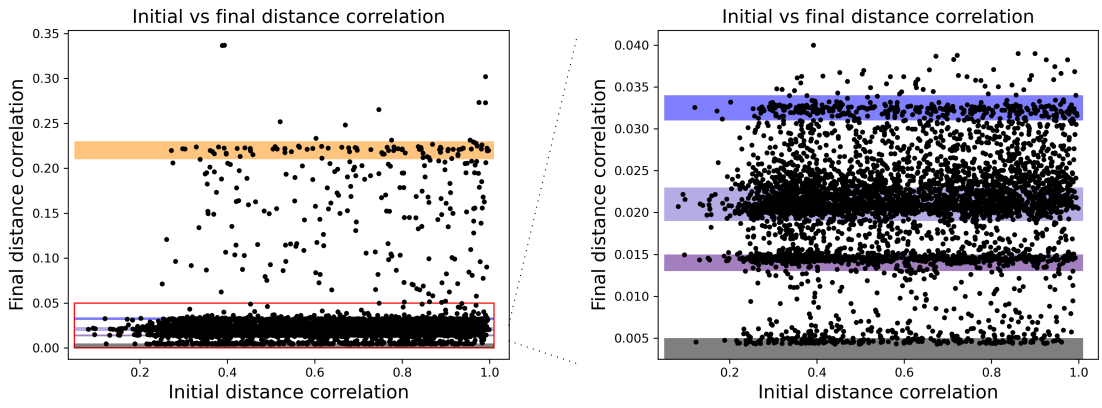


Figure B.7: Scatter plot depicting the initial vs final distance correlation values without a whitening step. The left side displays the complete range of final distance correlation values, while the right focuses on the lower range highlighted in the red box.

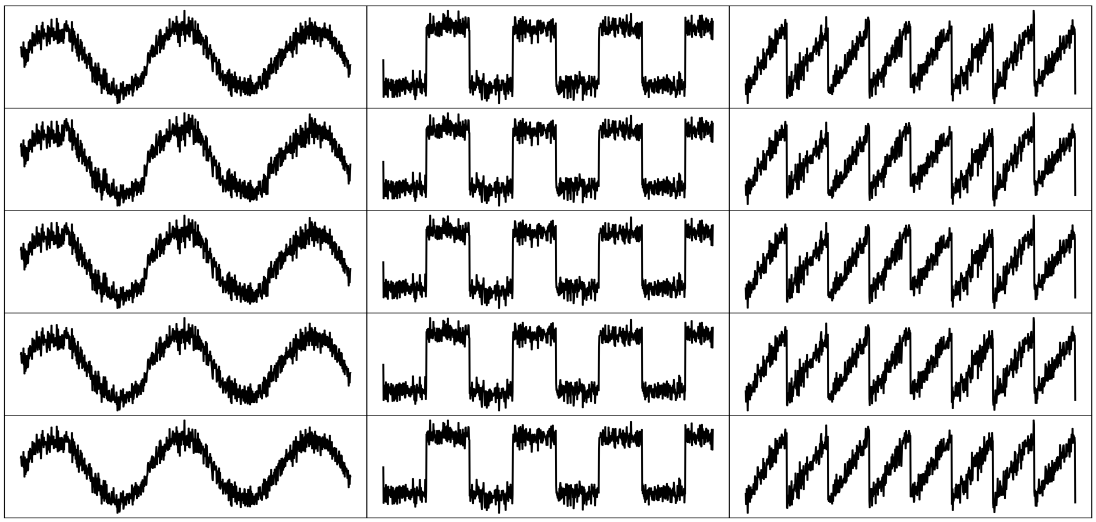


Figure B.8: Example of 5 out of 179 outputs for the synthetic problem, where the sources have not been whitened. This represents the outputs from the lowest plateau in the data data as seen in B.7. The range of output distance correlations were less than 0.005.

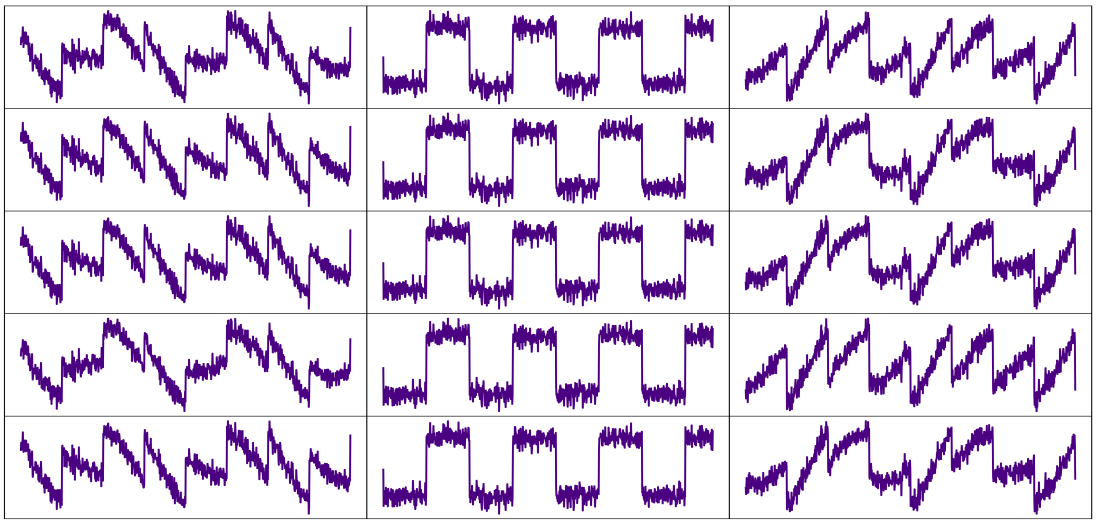


Figure B.9: Example of 5 out of 682 outputs for the synthetic problem, where the sources have not been whitened. This represents the outputs from the second lowest plateau in the data data as seen in B.7. The range of output distance correlations were between 0.013 and 0.015.

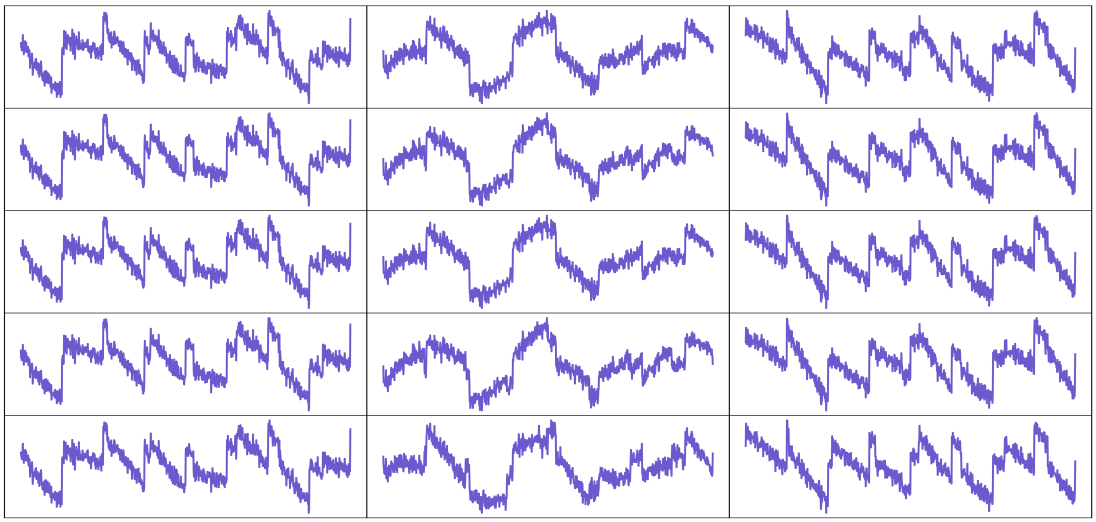


Figure B.10: Example of 5 out of 1,887 outputs for the synthetic problem, where the sources have not been whitened. This represents the outputs from the third lowest plateau in the data data as seen in B.7. The range of output distance correlations were between 0.019 and 0.023.

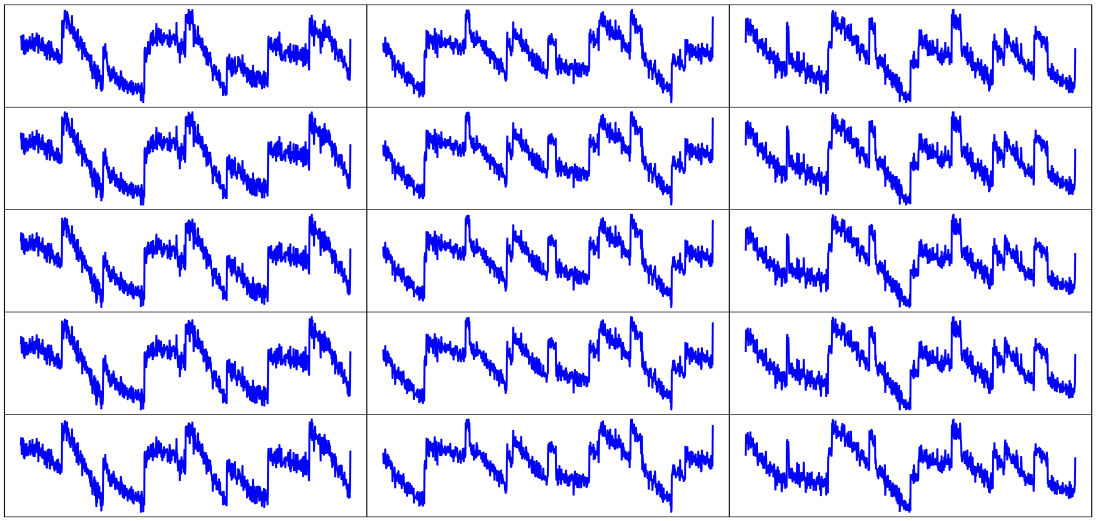


Figure B.11: Example of 5 out of 328 outputs for the synthetic problem, where the sources have not been whitened. This represents the outputs from the second highest plateau in the data data as seen in B.7. The range of output distance correlations were between 0.031 and 0.034.

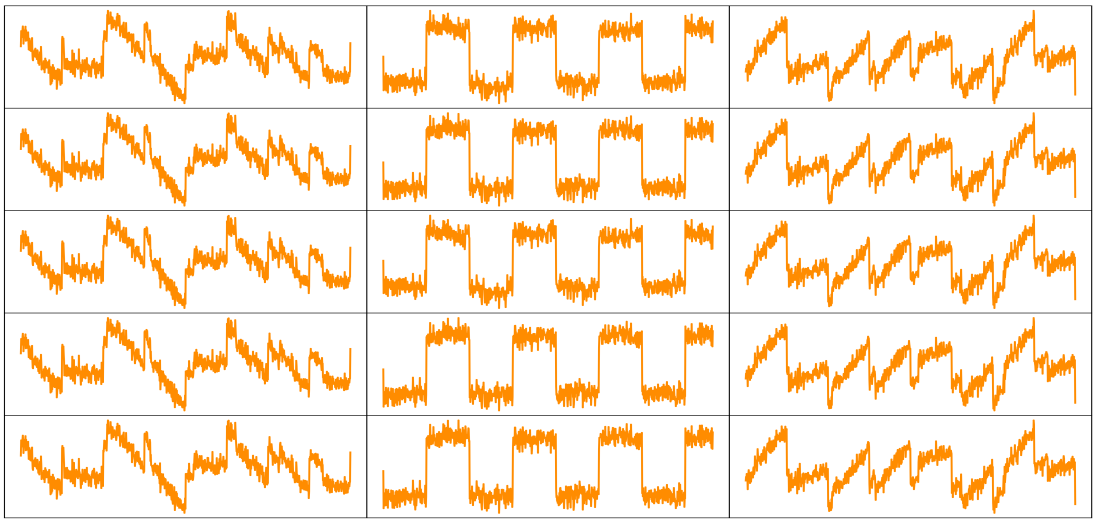


Figure B.12: Example of 5 out of 85 outputs for the synthetic problem, where the sources have not been whitened. This represents the outputs from the highest plateau in the data data as seen in B.7. The range of output distance correlations were between 0.21 and 0.23.

Furthermore, five randomly selected examples of three extracted sources in the ZCA whitened case are displayed for four ranges represented by Figure B.1: less than 0.005 in Figure B.14, 0.013 to 0.015 in Figure B.15, 0.015 to 0.017 in Figure B.16, and 0.019 to 0.021 in Figure B.17.

Figure B.13 is a replot of the right side of Figure B.1, featuring coloured bands linked to the colours in the Hiplot shown in Figure B.4. The colours of these bands are used as plot colours for the outputted source time series.

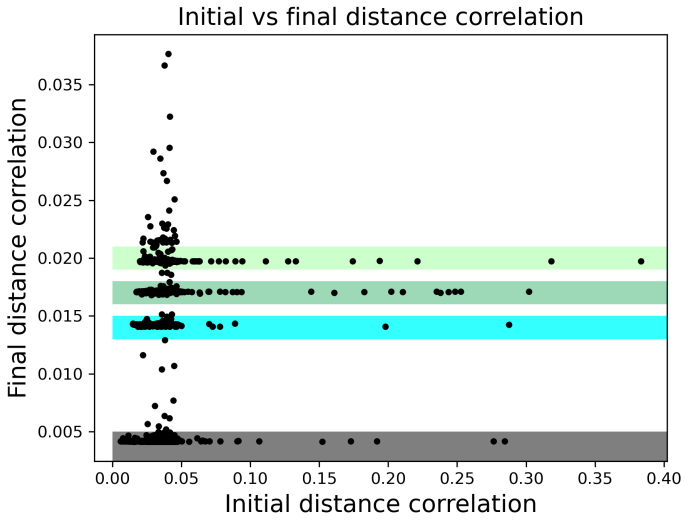


Figure B.13: Scatter plot of the initial vs final distance correlation values for the example where ZCA whitening is applied to the sources at each epoch.

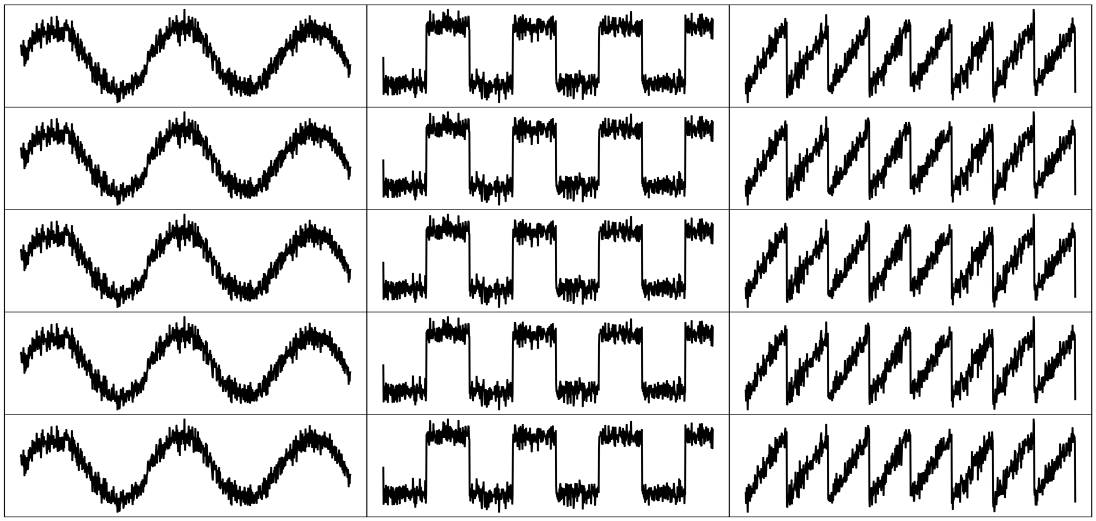


Figure B.14: Example of 5 out of 1,292 outputs for the synthetic problem, where the sources have been whitened. This represents the outputs from the lowest cluster of data as seen in B.13. The range of output distance correlations were less than 0.005.

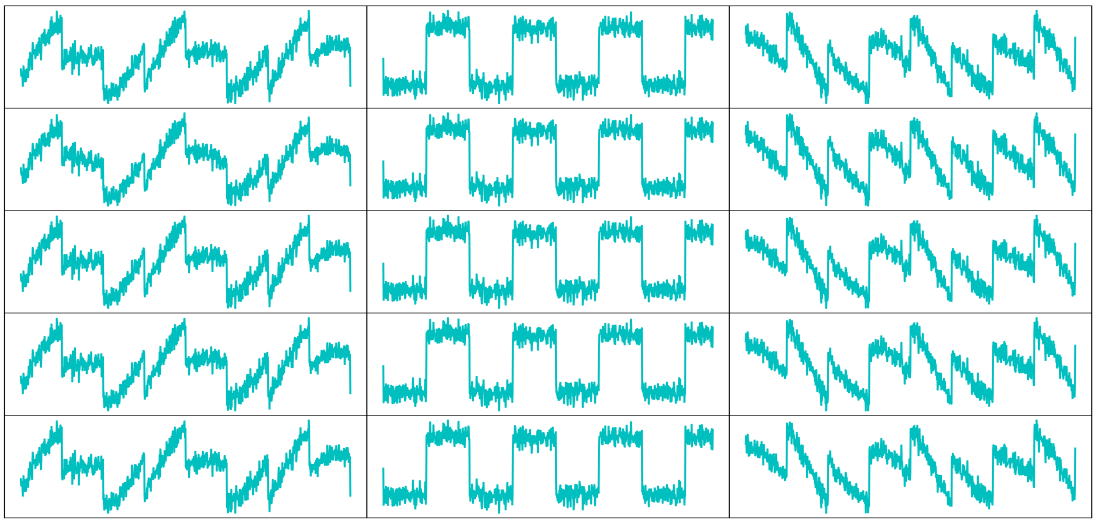


Figure B.15: Example of 5 out of 274 outputs for the synthetic problem, where the sources have been whitened. This represents the outputs from the second lowest cluster of data as seen in B.13. The range of output distance correlations were between 0.013 and 0.015.

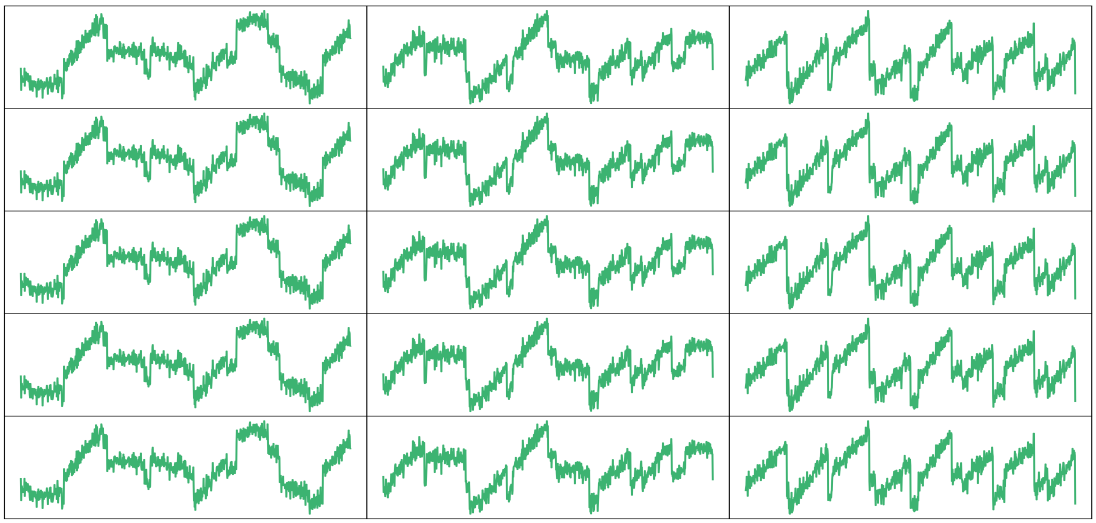


Figure B.16: Example of 5 out of 1,982 outputs for the synthetic problem, where the sources have been whitened. This represents the outputs from the second highest cluster of data as seen in B.13. The range of output distance correlations were between 0.016 and 0.018.

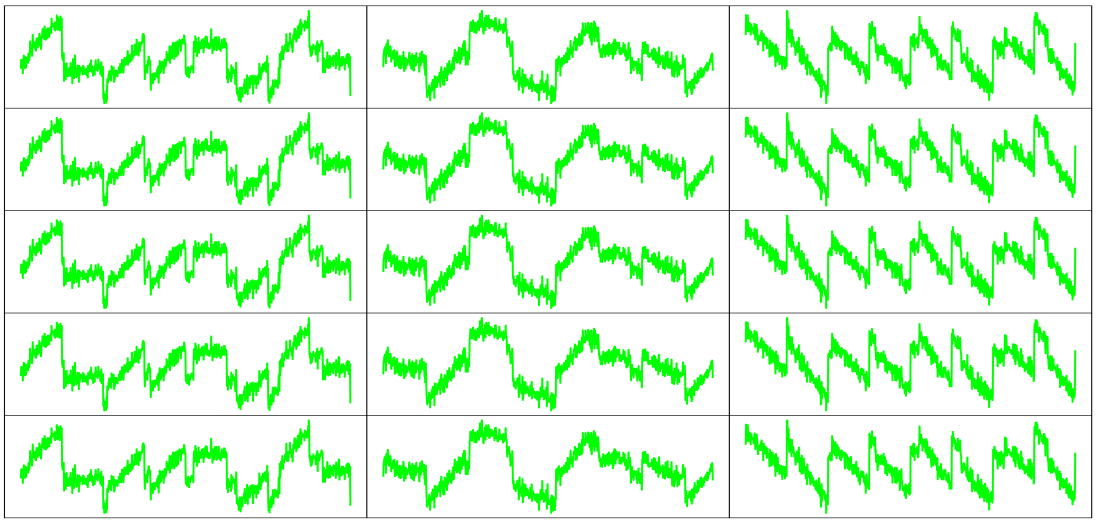


Figure B.17: Example of 5 out of 1,400 outputs for the synthetic problem, where the sources have been whitened. This represents the outputs from the highest cluster of data as seen in B.13. The range of output distance correlations were between 0.019 and 0.021.

As a reminder, the second research question investigates whether reducing the distance correlation will yield better sources when compared to the known input sources in this test case. It is evident that the five examples with the lowest final distance correlation, in both the whitened and non-whitened cases, produce results that are closest to the true sine, square, and sawtooth waves. Therefore, identifying the lowest distance correlation, which indicates the most independent sources, optimises source separation in this test case.

I will use a quantitative approach by employing the Amari distance, which measures the similarity between two invertible matrices. This metric is valuable for comparing different ICA solutions and determining how sufficiently an algorithm converges. The Amari distance was introduced as a performance measure for BSS in [41].

The Amari distance serves as a reflection- and permutation-invariant method to quantitatively assess how closely the unmixing matrix approximates the inverse of the true mixing matrix. A smaller distance indicates that the unmixing matrix is closer to its optimal value.

For the case without whitening but using normalisation, the Amari distances, ordered from the lowest to the highest final distance correlation range, are as follows:

Distance correlation	<0.005	0.013-0.015	0.019-0.023	0.031-0.034	0.21-0.23
Amari distance	0.09 ± 0.02	0.42 ± 0.04	0.9 ± 0.1	0.80 ± 0.04	10 ± 50

In the case of the whitened version, the Amari distances are as follows:

Distance correlation	<0.005	0.013-0.015	0.016-0.018	0.019-0.021
Amari distance	0.056 ± 0.006	0.43 ± 0.01	0.8 ± 0.2	0.855 ± 0.006

In both cases, the mean Amari distance increases as the final average distance correlation increases, with one exception. The exception involves the middle and second-highest clusters in the non-whitened example. This suggests that, although not always, finding a better local minimum usually leads to a superior set of output sources, or at least a closer unmixing matrix to the ground truth, compared to clusters with higher final average distance correlation values.

I want to discuss a detail of the time series extraction. A square wave is identified for the lowest and second-lowest final distance correlation for the whitened and non-whitened cases and the highest distance correlation for the non-whitened example.

First, focus on the non-whitened case with the highest distance correlation (see Figure B.12 for its related time series). In this scenario, the absolute correlations between the square wave and the other two extracted waves were 0.49 ± 0.05 and 0.46 ± 0.05. The correlation is higher than between the ground truth square signal and the sine and sawtooth, of which the absolute correlation between the ground truth square wave and the sine or sawtooth waves were 0.0860 and 0.0496, respectively. The correlations being further from 0 for the local minimum indicate that some characteristics of the square wave can be identified in the other two separated waves, resulting in higher mutual information for this separation. The non-zero correlation also explains why a similar solution does not exist in the whitened case; decorrelation through whitening precludes this possibility.

Next, in the case of the second-lowest distance correlation range for the non-whitened case (whose time series are represented by Figures B.9), the distance correlations for the pairs of sources containing the square wave are 0.003 ± 0.001 or 0.0047 ± 0.0006 (comparing the furthest left and the furthest right sources to the square wave, respectively). In contrast, the distance correlation for the pairing that excludes the square wave is 0.036 ± 0.002.

The ground truth distance correlations between the sine and square waves, the sine and sawtooth waves, and the sawtooth and square waves are 0.00945, 0.0132, and 0.00309,

	Sin-Sq	Sin-Saw	Sq-Saw
MI regression	0.015	0.123	0.015
MINE	0.009±0.002	0.074±0.018	0.003±0.001
Distance correlation	0.088	0.103	0.065

Table B.1: Pairwise comparison of independence metrics computed on the ground truth signals from the synthetic source separation example. The MINE computation is an average over the final 100 mutual information estimates out of a total of 1000 epochs.

respectively. Note that the pairing of the sine and sawtooth waves has the highest ground truth distance correlation.

Hlynsson et al. [149] adopted this synthetic problem to determine how effective blind source separation using MINE was. If one looks at Table B.1, for a number of metrics, the pairs of sources show forms of dependence.

In both the second-lowest case and the ground truth distance correlations, the fact that the distance correlations are lower when a square wave is involved when compared to the sine-sawtooth pairing suggests that the distance correlation loss is influenced by the pairs that include square waves, as these waves exhibit lower entropy compared to the sine or sawtooth waves. Consequently, this lower entropy is due to the square wave being a slim-peaked bimodal distribution exhibiting minimal uncertainty. The low entropy makes it more likely that including a square wave would reduce mutual information ($I(X;Y) = H(X)+H(Y)-H(X,Y)$, where H is the entropy and I is the mutual information), and be extracted during pairwise source separation.

Additionally, in the case of the second lowest plateau distance correlation case, at least one of the wave pairings must become more dependent on the other signals to lower the average distance correlations.

In Section B.0.1, I determined that the sources generated at the global minimum, corresponding to the lowest final distance correlation cluster, were equivalent when considering the limitations of ICA. In this section, I explored whether the sources obtained at this minimum are related to the underlying sources used to create the mixtures in the synthetic problem. I have concluded that, for this synthetic problem, the sources identified at the global minimum are the best separated. Visual inspections of Figures B.8 and B.14, along with their corresponding Amari scores, support this conclusion.

B.0.3 Consistency and the effect of whitening

The third question assesses the consistency of the results of source separation using distance correlation. Specifically, it explores the similarity of the outputted source signals in both local and global minima.

To answer this question, one can first visually compare the five examples from each specific cluster of final distance correlations, see Figures B.8-B.12 and B.14-B.17. In all nine cases, the five outputted example's sources, coming from the same cluster representing either a local or global minimum, look very similar. Additionally, using the quantitative Amari distance, in most instances, the standard deviations are low, suggesting that the source separation results for the selected cluster are consistent.

Framing the third question slightly differently, I want to explore the likelihood that the results converge to the same minimum when multiple random initialisations are employed and whether the predominantly chosen minimum is the global minimum.

In the non-whitened case, convergence to the global minimum occurred 179 times (3.58%). In the whitened case, convergence to the global minimum occurred 1,292 times (25.84%). However, this does not reflect the highest number of cluster occurrences. In the non-whitened scenario, the only cluster with fewer occurrences than the global minimum was the 0.21 to 0.23 cluster. Only the 0.013 to 0.015 cluster had fewer occurrences in the whitened case. Therefore, the random initialisations do not primarily converge on the global minimum. Additionally, as indicated by the counts for each case in the captions of the respective output figures, there is no minimum where a majority convergences occur.

Finally, the fourth question is concerned with the effect of whitening, which here was applied after the linear layer but before calculating the distance correlation loss. To investigate this question, I will compare the results of the ZCA-whitened and non-whitened cases.

It is important to note that the whitening transformation, which decorrelates the data, affects the distance correlation. At epoch 0, the distance correlation is lower with whitening than without. As illustrated in Figure B.6, the initial distance correlation without whitening can reach values as high as 1.0, while the initial distance correlation with whitening falls below 0.4.

In both cases, the lowest distance correlation cluster range is below 0.005, indicating

that both whitened and non-whitened methods ultimately minimise to the same global minimum in the best-case scenario, also compare Figures B.8 and B.14.

Whitening increases the number of distinct local minima and removes the highest final distance correlation minimum, seen in Section B.0.2 to produce outputted sources with linear relationships. However, in both the whitened and non-whitened cases, most initialisations fail to reach the global minimum, requiring multiple runs for each approach. The whitened method is more likely to converge to the global minimum than the non-whitened method (25.84% to 3.58%). Nonetheless, as previously noted, when the global minimum of distance correlation is reached, the resulting source separation, aimed at achieving independence, yields the best-outputted sources compared to the ground truth.

In conclusion, the outputs from each cluster are consistent with one another. The introduction of whitening reduced the number of minima by eliminating one that had linear relationships among its outputs. Whitening also increases the likelihood that training will converge on the global minimum.

B.0.4 Discussion

This appendix systematically outlines the fundamentals of using distance correlation as a loss function. From the data analysis, considering both the weight and the comparison between the initial and final distance correlation, it is evident that the loss landscape, even for this straightforward problem, is quite complex. This complexity means there isn't a clear subset of weights that can be used to initialise a run to converge on the global minimum reliably without a priori knowledge.

However, once a local minimum is reached, the outputs are consistent. Furthermore, the global minimum aligns with the best source separation when assessed through visual inspection and Amari distance in relation to the ground truths.

Additionally, while whitening was found to be unnecessary, it does help mitigate some of the more extreme local minima and enhance the convergence rate toward the global minimum.

To conclude, all of the experimental results from this appendix indicate that employing distance correlation as a loss function for independent component analysis through machine learning is a promising approach. The main limitation is that it presents a com-

plex loss landscape, diminishing the chances of successfully converging on the global minimum. Although whitening improves convergence toward the global minimum, this appendix also highlights the rationale for utilising the Restart Algorithm (Algorithm 1). This method's comparison of learned and reinitialised weights helps gradient descent escape local minima.

APPENDIX C

PCA-ICA approach

In PCA, the eigenvectors of the dataset's covariance matrix are chosen based on their eigenvalues. The goal is to select eigenvectors, or principal components, that account for as much variance in the data as possible.

In contrast, I have modified the approach for ICA to utilise distance covariance instead of the traditional covariance matrix. This adjustment allows the eigenvectors to represent independent components rather than principal components. As a result, projecting the dataset onto these independent components can help extract the underlying independent sources.

I found that the MSE between the double-centred distances, A and B, corresponds the distance correlation squared of their underlying vectors, X and Y. The distance covariance matrix is equivalent to the covariance matrix used in the PCA context. The two methods are illustrated by Equations 5.8 for the PCA case and 5.4 for the ICA case.

Algorithm 2 outlines a detailed explanation of the ICA version of PCA proposed, or the PCA-ICA method, using an inputted $n \times k$ matrix $\mathbf{X}$, where each row represents a k-dimensional mixture.

The PCA-ICA method can be used to calculate the percentage of distance variance described by different numbers of sources. This adaptation was applied to two source

Algorithm 2 Temporal algorithm

1: **Input:** An $n \times k$ matrix $\mathbf{X}$ consists of rows representing k-dimensional random vectors $X_1, X_2, \ldots, X_n$, which each correspond to a mixture.

2: Calculate the $n \times n$ matrix C with elements the pairwise distance correlations between the rows of $\mathbf{X}$:

$$C = \begin{bmatrix} R_n^2(\mathbf{X_1}, \mathbf{X_1}) & R_n^2(\mathbf{X_1}, \mathbf{X_2}) & \ldots & R_n^2(\mathbf{X_1}, \mathbf{X_n}) \\ R_n^2(\mathbf{X_2}, \mathbf{X_1}) & R_n^2(\mathbf{X_2}, \mathbf{X_2}) & \ldots & R_n^2(\mathbf{X_2}, \mathbf{X_n}) \\ \ldots & \ldots & \ldots & \ldots \\ R_n^2(\mathbf{X_n}, \mathbf{X_1}) & R_n^2(\mathbf{X_n}, \mathbf{X_2}) & \ldots & R_n^2(\mathbf{X_n}, \mathbf{X_n}) \end{bmatrix}$$

The matrix C is an equivalent of the covariance matrix in PCA, where instead of covariance one has distance covariance.

3: Calculate the eigenvalues λ_i, and the eigenvectors $\mathbf{v_i}$ of C. For an appropriately chosen $n' \leq n$, let C' be the $n' \times n$ matrix with rows the eigenvectors $\mathbf{v_i}$ corresponding to the highest eigenvalues, i.e. the principal directions of C. The eigenvalues are then used to compute the percentage of distance variance described by each direction.

4: For each mixture X_i, consider the double centered distance covariance matrices A_i in Eq. 2.15, flatten them into $1 \times k^2$ matrices, and concatenate them into a $n \times k^2$ matrix D.

5: Project the flattened double centered distance covariance matrices onto the principal directions of C, by computing the $n' \times k^2$ matrix $C'D$.

6: **Output:** Finally, one uses PCA to project the $n' \times k^2$ matrix $C'D$, onto a lower dimensional subspace of dimension $n' \times k$.

Mixtures	DistCov(%)	DistCov(%)	DistCov(%)	DistCov(%)	Total of top 3 (%)
3	38.0	33.1	28.9	-	100
4	39.6	26.6	21.9	11.8	88.2

Table C.1: The percentage of the distance covariance accounted for by each IC, highest to lowest from left to right, for the 4- and 3-mix cases with 3 underlying sources. The total distance covariance described by the top 3 ICs is also given.

Fourth signal	DistCov(%)	DistCov(%)	DistCov(%)	DistCov(%)	Total of top 3 (%)
Sine	41.3	25.2	24.1	9.4	90.6
Square	39.8	26.3	20.7	13.3	86.7
Sawtooth	45.5	25.9	23.8	4.7	95.3

Table C.2: The percentage of the distance covariance accounted for by each IC, highest to lowest from left to right, for the 4-mix-4-source cases, with the fourth signal a scaled version of one of the other 3 underlying sources. The total distance covariance described by the top 3 ICs is also given.

separation tasks, the 3-mix-3-source problem, described in Section 4.2, and with a 4-mix-3-source problem, created by linearly combining a sine, square and sawtooth wave using the mixing matrix:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.5 & 2 & 1 & 2 \\ 1.5 & 1 & 2 & 0.5 \\ 2 & 2 & 1 & 1.5 \end{bmatrix}$$

and adding white noise.

From Table C.1, it can be observed that when the number of mixtures is three, the total percentage of dependence shown by the distance correlation squared of ZCA whitened vectors accounts for 100% (which it will by definition), with each source describing approximately a third of total percentage of distance variance. In contrast, in the four-mixture case, the total distance variance described by 3 sources was 88.2%. Although the value is lower in the four-mixture scenario, three sources still explain most of the distance covariance. If an appropriate threshold is chosen, this could effectively guide the user in selecting the optimal number of sources to extract.

Moreover, I investigated a scenario where the PCA-ICA method was tasked with a

	Sine	Square	Sawtooth
Restart	19.3±0.1	57.4±3.0	19.1±0.2
PCA	1.2±0.0	2.1±0.0	-2.6 ±0.0
PCA-ICA	6.3±0.0	9.1±0.0	7.7±0.0

Table C.3: Comparison of parametric Restart method compared to the results of PCA and the PCA-ICA version.

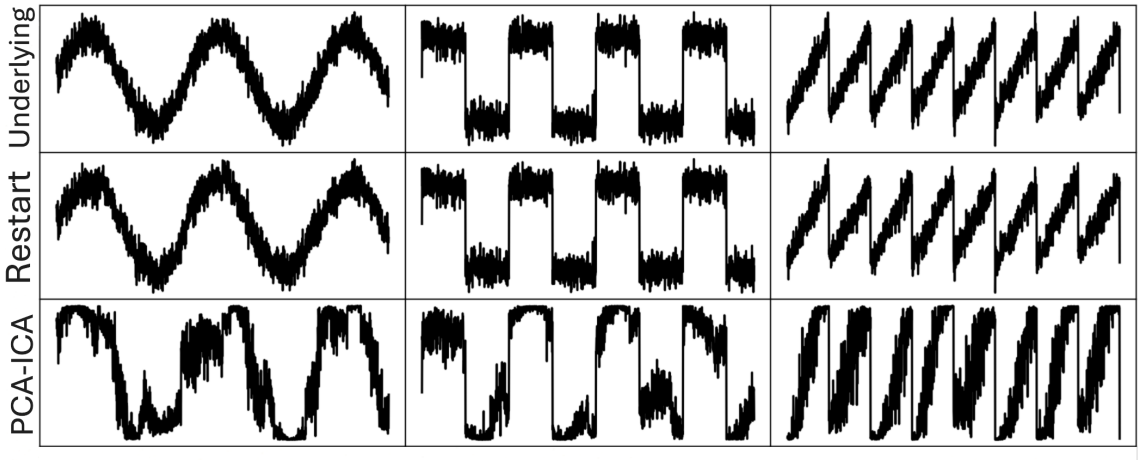


Figure C.1: Comparison of the underlying sine, square, and sawtooth signals with the results of the Restart algorithm using a linear layer neural network and distance correlation loss, applying PCA and implementing the PCA-ICA method.

four-mixture, four-source problem, including one source that was a scaled version of one of the other underlying sources. This was a test of the method's ability to distinguish between identical underlying sources. As shown in Table C.2, the total for the top three distance covariances, with sine, square, and sawtooth waves duplicated, were 90.6%, 86.7%, and 95.3%, respectively. These results suggest a promising potential of the method in accounting for duplicated underlying sources.

This serves as a good illustration of how to count the potential number of underlying sources using the percentage of distance covariance represented by the independent components (ICs). However, it is important to note that this illustration relies solely on a synthetic problem and is not exhaustive, though this method has been implemented to identify the number of sources to extract from the GNSS data in Chapter 4.

Now, I will provide an example of PCA-ICA being implemented on the synthetic problem. Table C.3 shows the average SI-SDR values over ten repeats for the best distance

correlation method, the Restart method, along with PCA-ICA.The parametric method far outperforms the non-parametric techniques and that the PCA-ICA method slightly outperforms PCA. In Fig. C.1, I show the Restart results for the synthetic problem with distance correlation as the objective function and a linear layer architecture, along with the PCA-ICA method and the ground-truth signals. In this case, it can be seen that the Restart method outperforms the PCA-ICA methods, agreeing with the quantitative results from Table C.3.

It is important to note that, according to Algorithm 2, the distance matrix projected onto the dimensions that capture the most significant dependence, as defined by distance correlation, requires a dimension reduction step. This step is necessary to obtain vectors representing the given pairwise distances of the projected data.

First, dimension reduction may not preserve the pairwise distances in the projected data. Even in cases where the distances are preserved, the resulting vectors may not combine linearly to form the known mixtures. The synthetic problem assumes that the known mixtures are linear combinations of the underlying sources. Therefore, the PCA-ICA method may not be the most effective approach since the linear constraint is not strongly enforced and PCA dimension reduction does not strictly preserve pairwise distances.

A more effective strategy for enforcing linearity is using distance correlation as a loss function for a neural network that incorporates a linear layer. In this context, the PCA-ICA variant performs worse than methods that impose a strict linearity constraint.