University of Alberta

Robust matrix rank reduction methods for seismic data processing

by

Ke Chen

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Geophysics

Department of Physics

Fall 2013
Edmonton, Alberta

Examining Committee

Mauricio Sacchi, Physics

Claire Currie, Physics

Mathieu Dumberry, Physics

Yu Gu, Physics

To my parents and all my teachers

Abstract

An important step of seismic data processing entails signal de-noising. Traditional de-noising methods assume Gaussian noise model and their performance degrades in the presence of erratic (non-Gaussian) noise. This thesis examines the problem of designing reduced-rank noise attenuation algorithms that are resistant to erratic noise.

I first introduce a robust matrix factorization based on M-estimate and incorporate it into the formulation of the classical Singular Spectrum Analysis (SSA) algorithm. This new algorithm (Robust SSA) permits to de-noise seismic data that have been contaminated by non-Gaussian noise.

I also propose a second Robust SSA algorithm that attacks the data de-noising and reconstruct problems as low-rank matrix recovery problem that is solved by a convex optimization algorithm. The NP-hard rank minimization problem is replaced by its tightest convex relaxation, the nuclear-norm minimization. An augmented Lagrangian method is used to numerically look for the solution that minimizes the cost function.

Acknowledgements

First and foremost, I thank my supervisor, Dr. Mauricio Sacchi. His passion for geophysics research has deeply influenced me. His patience and encouragement are important for me for completing this MSc degree. I feel honoured to be a part of the Signal Analysis and Imaging Group (SAIG) consortium for my MSc study and research. I thank my supervisor for forming this group and the consortium sponsors for providing financial support. I would like to thank my colleagues at the SAIG for scientific discussions as well as friendships. I thank the friends I made in the University of Alberta for making life here enjoyable. In particular, I would like to thank my examining committee, Dr. Claire Currie, Dr. Mathieu Dumberry and Dr. Yu Gu, for taking their time to read and providing valuable suggestions on my thesis. Finally, I thank my parents for their sustaining encouragement and support on my journey of overseas study.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Background

Exploration geophysics is an interdisciplinary science involved with physics, mathematics, and geology. It utilizes geophysical data observed on the surface of the earth to measure and then invert for physical properties of the subsurface. The aim is to explore minerals, hydrocarbons, groundwater reservoirs without the need to directly penetrate the earth's interior. Different geophysical methods can be used for "imaging" subsurface structures, e.g. seismic methods, gravitational methods, electrical methods, magnetic and electromagnetic methods. The seismic method is often used in the exploration of hydrocarbons. In this dissertation, I will mainly focus on problems in exploration seismology.

In a *seismic experiment*, a controlled seismic source excites the earth and generates impulsive sound waves that travel in the earth's interior. The waves propagate through the earth, part of them are attenuated in the earth's interior, part of them are reflected back when they reach geological boundaries. The reflected data are recorded by the receivers (geophones) deployed on the earth's surface. This is the first step of exploration seismology, know as *data acquisition*. The simplest acquisition configuration is to deploy a source and receivers along a line (Figure 1.1 (a)). The latter is referred to as 2-D seismic survey. The source is fired and the seismograms are recorded by the receivers. The seismograms recorded at different receivers are grouped together in one *common shot gather*. Then, the whole acquisition system is moved along the line to the next position to repeat experiment, and so on. The data acquisition is in shot-receiver coordinate system. While, many seismic processing sequences are applied in midpoint-offset coordinate system (Figure 1.1 (b)). The recorded seismic traces with the same mid-point location can be grouped together as a common mid-point (CMP) gather. A shot gather from Yilmaz's data set (shot 25 in the

book *Seismic Data Analysis*) is shown in Figure 1.2. This shot gather is from a 2-D land survey in Alberta, Canada. The survey uses a split-spread geometry where the source is located in the center of receiver cable. Different waves are observed including direct waves, refractions and reflections as indicated in Figure 1.2. The exploration method that uses refractions is known as the *refraction method* and it is often used for near surface studies and for crustal studies. The exploration, development and monitoring of reservoirs of oil and gas is mainly carried out via reflected waves with the *reflection method*.

Recorded seismic wavefields are often contaminated by coherent and incoherent noise. Several types of noise are presented in Figure 1.2, coherent noise such as ground roll, high-amplitude erratic (non-Gaussian) noise and the ambient random noise presents in the whole section.

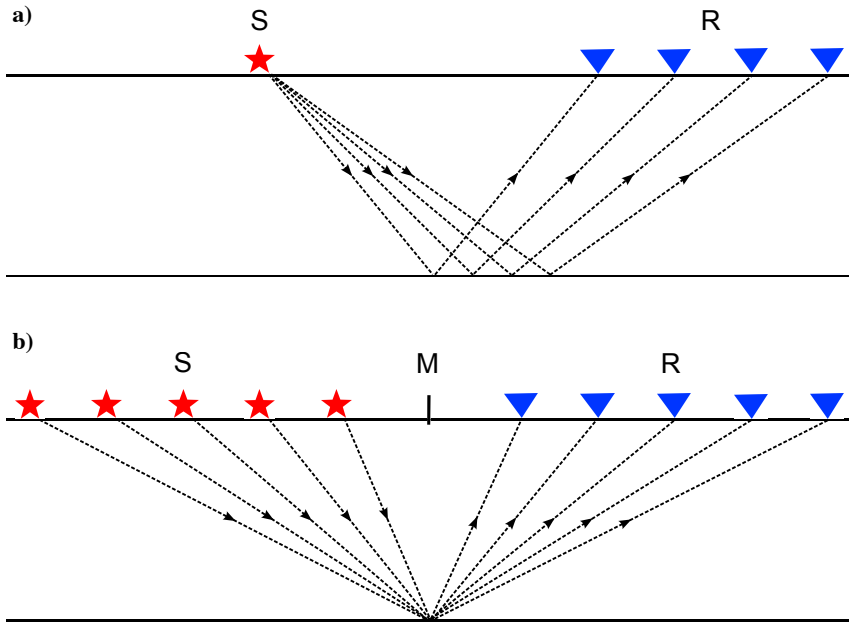

Figure 1.1: A simple sketch map of 2-D seismic survey (one flat layer). Red star represents source, blue triangle represents receivers, dash line represents ray path, and M is midpoint. a) Shot-receiver coordinates. b) Midpoint-offset coordinates.

The second step in the reflection seismology is called *data processing*. In this step the collected data are processed and analyzed via mathematical and physical principles. In the *interpretation* step, the processed data and images are interpreted as geological structures to indicate the location of potential oil and gas reservoir. The seismic data processing can be divided to three principal steps: *deconvolution*, *CMP stacking*, and *migration*. Deconvolution removes the seismic wavelet from the seismic data to broaden the frequency band

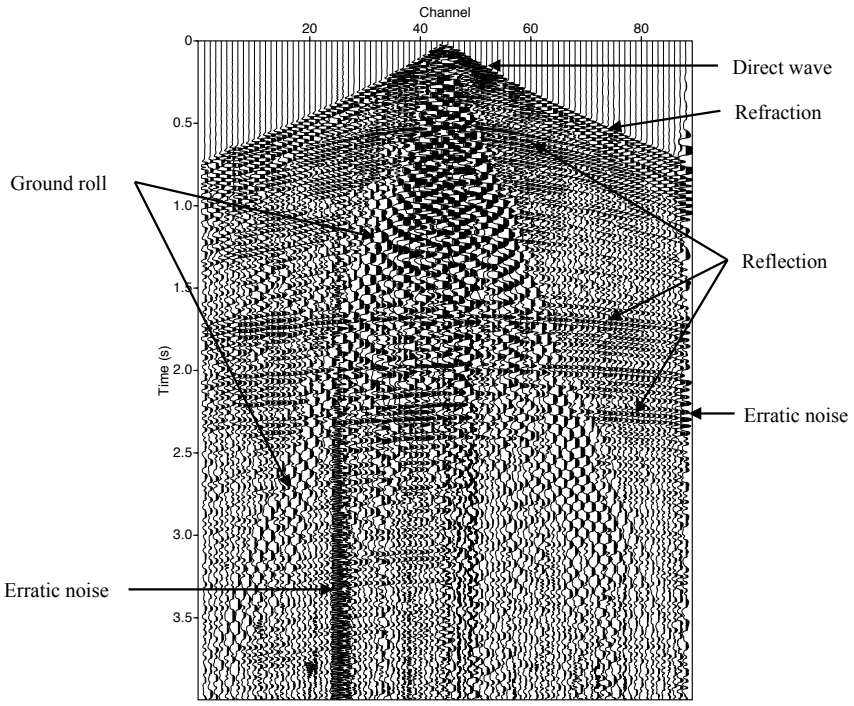

Figure 1.2: A shot gather from a 2-D seismic survey in Alberta.

of the data. In other words, deconvolution is used to improve temporal resolution. CMP stacking averages NMO corrected seismic traces in each CMP gather along offset dimension to estimate a zero-offset seismic section. It can suppress both random noise and coherent noise to improve the signal-to-noise ratio of seismic data. Migration is an imaging process that moves dipping events to their true subsurface positions. There are many other auxiliary processes that are often run in between each one of the three main processes. This thesis will study noise suppression and missing data reconstruction via reduce-rank methods. This is a step that should be carried out before or after stacking and prior to migration. Noise suppression can also be carried out in different domains.

1.2 Seismic noise

In exploration seismology, one wants to keep reflections and eliminate coherent and incoherent noise (Yilmaz, 2001). *Incoherent noise*, as its name implies, is not correlated from trace to trace. That is to say, the phase of the noise is independent between adjacent traces. Incoherent noise is also known as *random noise*. It can be caused by a variety of factors such as wind, human and animal activity, rain drops, instrumental noise, etc.

Coherent noise is energy correlated in both spatial and temporal direction, or just correlated in time. I wound like to divide the coherent noise into two categories: *spatially coherent noise* and *temporally coherent noise*. Spatially coherent noise includes *ground roll*, *multiples*, *reverberations*, *side-scattered noise*, *guided waves* and *air waves*. They are correlated in space and time. "Being correlated" means that the phases of the coherent noise in adjacent samples have some particular relationships. Spatially coherent noises are also called *source-generated noise*. Among them, ground roll, side-scattered noise, guided waves, and air waves are called *linear noise* because they are almost linear in common shot gather. Ground roll is a very common type of coherent noise in land data. Ground roll is the vertical component of the Rayleigh wave. It has the properties of being dispersive, high-amplitude, low frequency and low group velocity. Multiples correspond to energy that is reflected more than once in the travel path in the subsurface. Multiples have the properties of large moveout and periodicity. Temporally coherent noise is not generated by the seismic source, it is coherent in the temporal direction but incoherent in spatial direction. Temporally coherent noise includes *noisy trace* and *noise burst*. Noisy trace means that most part of the trace is corrupted with high-amplitude noise. The examples of noisy trace are trace with power line noise and/or traffic noise in land seismic data, trace with swell noise in marine seismic data, noisy trace caused by electronic transients and glitches in the recording instrument. Power line noise usually presents itself as monochromatic sinusoidal wave with frequency of 50 or 60 Hz. Swell noise is caused by bad weather during marine acquisition. It has the properties of high-amplitude and low frequency. Its frequency ranges from 2 Hz to 10 Hz (Elboth et al., 2010). Noise burst is high-amplitude noise lasting ten to several hundred milliseconds (Anderson and McMechan, 1989). There is a kind of noise named noise spike that is high-amplitude noise lasting only a few time samples. Generally, the noisy trace, noise burst and noise spike have high-amplitude and are not modeled well by the Gaussian distribution. They can be classified as the *erratic noise* (Trickett et al., 2012). We can regard the *erratic data* (Claerbout and Muir, 1973) as clean data corrupted with erratic noise. In the field of robust statistics, erratic data are referred to as *outliers* (Maronna et al., 2006). The methods developed in this thesis is most suitable for noisy traces suppression.

1.3 Seismic noise attenuation methods

This section reviews some popular methods for suppressing spatially incoherent noise including random noise and erratic (non-Gaussian) noise. Signal processing methods for noise attenuation generally exploit differences between the signal and noise. They represent the data in a particular domain where the signal and noise are more easily distinguished from each other. For example, methods based on transforms map the seismic data to domains where signal and noise can be better separated (Ulrych et al., 1999). In this particular do-

main, the coefficients corresponding to noise are eliminated while the remaining coefficients are inverted back to the original domain. One can also eliminate the coefficients corresponding to the signal and keep those associated with noise. The latter permits an estimation of a model of the noise that can be subtracted from the original data.

1.3.1 Random seismic noise attenuation

Many methods for random noise attenuation have been developed in the past several decades. First, *CMP stacking* (Mayne, 1962) was proposed to reduce random noise by averaging NMO corrected seismic traces with different offsets in each common mid-point gather. There are processes that require suppressing noise prior to stacking or to apply noise suppression after stacking. This is why our arsenal of seismic processing algorithms often contains various methods for noise attenuation that can operate in different domains and with pre-stack and post-stack data.

Frequency band-pass filtering can be used for suppressing ambient noise by restricting the amplitude spectrum of the seismic data. However, signal and noise often overlap in the frequency domain and therefore, one might be eliminating a portion of the signal when applying frequency domain band-pass filtering.

Random noise reduction via *spatial prediction filtering* has been proposed as an alternative to frequency band-pass filtering. The prediction filters can be estimated and applied in the f-x domain or t-x domain. The principle in this type of filters resides on the lateral predictability of signals. Canales (1984) firstly proposed the f-x prediction technique for seismic random noise reduction. This method assumes that noise-free seismic signal is composed of linear or nearly linear events in t-x domain. For one particular frequency slice in f-x domain, the signal is the superposition of a finite number of complex exponentials. The Fourier coefficients of one particular frequency from different traces are linearly dependent with each other. Therefore, some of the Fourier coefficients can be predicted from others, i.e. f-x signal is linearly predictable in space. This method implicitly assumes that the f-x domain seismic data can be represented by the autoregressive (AR) model, i.e. the linearly predictable part is the signal and the unpredictable portion is the white noise. The prediction error filter (PEF) is firstly estimated from the data, and then the noise is estimated via applying the PEF on the data. Gulunay (1986) named this technique as f-x deconvolution. The f-x deconvolution is known to damage original signal if the noise level is high, i.e. it cannot separate the signal and noise perfectly. One of the reasons is that it uses the autocorrelation method that assumes the data outside the implicit window are zero. Harris and White (1997) propose to use the transient-free data matrix to alleviate this problem (Ulrych and Clayton, 1976), which is known as covariance method. The other

reason is that, the data in f-x actually are not modeled well by the AR model because of the additional observation noise in the data. This problem is more serious when the signal-to-noise ratio is low. In real application, a large order AR model is used for better representing the data (Ulrych and Sacchi, 2005). Harris and White (1997) suggests to "clean up" the data matrix before estimating the prediction filter (Kay and Marple, 1981; Tufts and Kumaresan, 1982). Later, Soubaras (1994, 1995) proposed the f-x projection filtering technique, which utilizes the additive noise model and the concept of quasi-predictability. This technique has the advantage of preserving the signal. It estimates the noise via applying the autodeconvolved PEF (projection filter) on the data. Sacchi and Kuehl (2001) pointed out that the appropriate model for linear events in f-x domain is the autoregressive/moving-average (ARMA) model. The approach to estimate the noise from the data in their paper is equivalent to the estimation approach in the f-x projection filter (Soubaras, 1994). While, their result is derived from ARMA model and they find the closed-form solution of the PEF via solving an eigendecomposition problem. The f-x prediction filtering techniques have been widely and successfully used in oil industry for random noise reduction. On the other hand, Hornbostel (1991) introduced the t-x prediction filtering techniques for signal-to-noise ratio enhancement, which is more suitable when signal or noise is non-stationary in temporal or spatial direction. It's a 2-D adaptive least-mean-square filter modified from Widrow et al. (1967). In t-x domain, a given sample is firstly predicted via weighting the samples in a rectangular window. Then, the filter coefficients (weights) are updated from the prediction error. The new prediction filter is applied on the next sample and the described two steps are repeated: predicting sample and updating weights. This technique does not need to divide the entire data into windows, i.e. it's adaptive. Abma and Claerbout (1995) propose a t-x prediction filtering method that the prediction filter is estimated in t-x domain using conjugate-gradient method. They pointed out the f-x prediction filter is equivalent to a t-x prediction filter that is as long as the data.

The fourth kind of methods is based on *matrix rank reduction*. It assumes that the matrix formed in some particular way from the noise-free seismic signal is low rank, i.e. the *singular spectrum* of this matrix is sparse. The presence of random noise in seismic data will increase the rank of the formed matrix, but they only present as small singular values. They conclude that rank reduction on the matrix via Truncated Singular Value Decomposition (TSVD) can remove the random noise from the data. The Karhunen-Loeve (K-L) transform was firstly introduced for seismic data processing by Hemon and Mace (1978). Jones and Levy (1987) extended the K-L transform technique for incoherent and dipping coherent noise suppression in the stacked seismic section. The assumption is that the shallowly dipping events (linearly horizontal) are more strongly correlated from trace to trace and exhibit as large eigenvalues in eigenspectrum of the covariance matrix. The steeply dipping events (e.g. ground roll or marine streamer noise) and random noise are less coherent and

display as smaller eigenvalues. This allows the use of zero-lag K-L transform to filter out the noise. They also propose the slant-KL transform that modifies the covariance matrix with time lags determined by the dips in seismic section. It can handle dipping events situation. This method is equivalent to using zero-lag covariance matrix after flattening the events via linear moveout correction. Finally, a reverse linear moveout correction is applied after the zero-lag K-L filtering. Besides, they discuss about using K-L transform to remove multiples in CMP gather. The multiples are flattened by NMO correction and primary events are under-corrected or over-corrected. The multiples correspond to the largest eigenvalues and primaries correspond to smaller eigenvalues. Marchisio et al. (1988) applied the full K-L transform that has all lags (temporal and spatial), and also the partial K-L transform that has fewer time lags in the covariance matrix for random noise attenuation and VSP wavefield separation. It works for data with dipping events. Al-Yahya (1991) proposes the partial K-L transform for incoherent noise attenuation in the situation that there are several conflict dips in the seismic section. It is an extension of Jones and Levy (1987)'s method. For each dipping event, it is flattened by the linear moveout correction and a zero-lag K-L filtering is followed. The inverse linear moveout correction is applied on the filtered data. This procedure is repeated for all the dipping events and the results are summed. Freire and Ulrych (1988) applied singular value decomposition (SVD) in t-x domain to separate the upgoing and downgoing wavefield in vertical seismic profiling (VSP) data. They also discussed the relationship between the SVD and the K-L transform. Ulrych et al. (1988) discussed several applications of SVD for reflection seismic data processing such as signal to noise enhancement, dip filtering, separation of upgoing and downgoing wavefield in VSP data and residual static correction. They referred to this technique as eigenimage reconstruction. The t-x domain eigenimage approach has the advantages that the regular sampling in time or space direction is not necessary and it's free of aliasing problems. Liu (1999) and Chiu and Howell (2008) and Cary and Zhang (2009) applied the K-L transform or SVD for ground roll attenuation. The above t-x domain rank reduction methods need linear move-out correction when dealing with dipping events. Mars et al. (1987) applied the spectral matrix filtering for seismic random noise attenuation in CDP gather. They also used a synthetic example to show that spectral matrix filtering can be used for wavefield separation. Mari and Glangeaud (1990) applied the spectral matrix filtering for signal-to-noise ratio enhancement in VSP seismic data, and for wavefield separation in VSP seismic data. It works in f-x domain and the K-L transform is applied on a constructed spectral matrix. Trickett (2003) proposed the f-xy eigenimage filtering for random noise reduction in stacked 3D seismic volumes. It conducts rank reduction on each 2-D constant-frequency slice, which works well for dipping events. Trickett (2002) introduced Cadzow's algorithm for random seismic noise attenuation on 2-D seismic section as an alternative to f-x prediction methods. He called it f-x eigenimage filtering. Trickett (2008) extended Cadzow's algorithm for

seismic denoising to three or more dimensional seismic data. He modified the name of the algorithm as Cadzow filtering. Trickett et al. (2010) proposed to use Cadzow's algorithm for seismic data reconstruction. Sacchi (2009) introduced the f-x domain Singular Spectrum Analysis (SSA) method for seismic noise suppression and discussed the relationship between SSA and Cadzow algorithms. SSA/Cadzow filtering can efficiently remove Gaussian noise in the presence of dipping events with good preservation of signal. Oropeza and Sacchi (2011) applied and integrated the Multichannel Singular Spectrum Analysis (MSSA) into the projection onto convex sets (POCS) framework, which results in a simultaneous denoising and data reconstruction algorithm. When the MSSA is applied on multi-dimensional data, the size of the block structured matrix will be large, the computation of the SVD (Golub and Van Loan, 1996) is very expensive. Oropeza and Sacchi (2011) also adopted a randomized singular value decomposition to accelerate the rank reduction in MSSA. Gao et al. (2013) extended MSSA algorithm for 5-D seismic data denoising and reconstruction. They proposed to use the Lanczos bidiagonalization method combined with fast Toeplitz matrix-vector multiplication to reduce the computation cost. SSA/Cadzow reconstruction algorithms can not handle regularly decimated and aliased data. Naghizadeh and Sacchi (2013) proposed a MSSA/Cadzow based reconstruction algorithm for interpolating regularly sampled seismic data. It extract information from low frequencies to recover the regularly missing information at high frequencies. Chiu (2013) proposed the multichannel singular spectrum analysis in the randomized domain for simultaneous coherent and random noise attenuation in 3D data volume. It works well in the situation that the coherent noise is spatially aliased. NMO correction is applied on the primary events to flatten them. For each frequency, the randomizing operator randomly rearranges the order of data in frequency domain. Then, primary events remains coherent and coherent noise changes to incoherent noise. Then, MSSA is applied on the randomized data and followed by the inverse randomizing operator.

The spatial prediction filtering methods and matrix rank reduction methods are efficient for random Gaussian noise attenuation. However, they do not perform well when the seismic data are corrupted with erratic (non-Gaussian) noise. The reason is that they are based on least-squares minimization that are optimal when the noise is Gaussian but are seriously degraded when the noise is non-Gaussian.

1.3.2 Erratic seismic noise attenuation

There are different types of methods for erratic noise reduction. Conventional methods including trace editing, CMP stacking and band-pass filtering. (a) Trace editing can be used for removing noisy traces with high-amplitude erratic noise (Yilmaz, 2001). The noisy traces corrupted by monofrequency signals (e.g. powerline noise) and noisy traces caused by

transient glitches are deleted in trace editing. It's done by human interpreters. However, the data volume of modern seismic survey is usually too large that makes manual trace editing inefficient. Noisy trace editing, first-break refraction picking and velocity analysis are three seismic data processing steps that need a large amount of labor work. (b) CMP stacking can suppresses the erratic noise. While, the stack is usually the mean estimate of traces in CMP gathers. Mean estimate is sensitive to outliers. Also as pointed out before, some processes may require suppressing noise prior to or after stacking. (c) Band-pass filtering can be used for erratic noise attenuation. For example, low-cut filtering utilizes the low frequency property of swell noise to remove it. Notch filter can be used to remove powerline noise. However, the signal and erratic noise are not separated perfectly in frequency domain.

More automatic and robust methods for erratic noise attenuation have been proposed. (a) Outlier diagnostic and rejection. Erratic noise is detected first and followed by damping or interpolation. Neff and Wyatt (1986) proposed the amplitude rejection method for noise spikes and noise burst attenuation. For each seismic trace, data values greater than a given threshold are deleted. The threshold set by user is very crucial for both removing noise and preserving signal. They also proposed the slope rejection method that trace regions having slopes (difference between data values) greater than the preset slope threshold value are regarded as noise spikes or noise bursts and removed. They suggested a radial amplitude-slope rejection method that combines the normalized amplitude and slope for thresholding. The gaps resulting from the removing of noise spikes and noise bursts are interpolated or zero-filled. Berni (1987) proposed an automatic method for burst noise editing. The seismic section is divided into time gates with equal time interval. A value describing the energy of the data samples in each gate of each trace is calculated. It can be the average absolute amplitude, root mean square or sum of squares of the data samples in each gate. For each time interval, the threshold is defined as a scalar times the smoothed value or median of the energy of nearby gates. If the energy of a particular gate in this time interval is greater than this threshold value, it is surgically blanked. Anderson and McMechan (1989) proposed a method for automatic editing of noisy seismic data using relative amplitude decay rates of traces as criteria. The relative amplitude decay rate of a particular trace can reflect its signal-to-noise ratio because the signal amplitudes decrease with time but the spatially incoherent noise has constant amplitude. Mavko (1988) proposed to apply multivariate statistics on each seismic trace instead of univariate statistics to detect noise spikes and noise burst. The generalized squared distance (Mahalanobis distance) is used instead of the simple amplitude of data sample. However, the using of covariance matrix in the generalized square distance makes it non-robust and one also needs to define a threshold distance. Soubaras (1995) proposed a strategy to use the f-x projection filter to attenuate erratic noise. In each frequency slice or frequency band, the traces containing impulsive noise are detected, invalidated and then interpolated as missing traces. Cambois and Frelet (1995) applied this

technique for swell noise attenuation. With similar framework, Schonewille et al. (2008) implemented an iterative FX prediction filtering for swell noise attenuation via repeatedly applying FX filter on data to improve noise attenuation. Elboth et al. (2010) designed a time-frequency de-noising algorithm that is also based on detecting and then attenuating. Bekara and van der Baan (2010) proposed an expectation-maximization algorithm for high-amplitude noise detection. After the noisy samples been detected, it is rescaled with a constant factor. The detection procedure is automatic. Noisy-trace editing is a kind of *outlier rejection* procedure (Anderson and McMechan, 1989). The performance of rejection procedures is not as good as robust estimation procedures.

(b) Robust CMP stacking. A robust estimation method the α-trimmed mean is applied in stacking to deal with bad seismic traces (Watt and Bednar, 1983). Elston (1990) applied robust M-estimator method for stacking. He used different loss functions, the ℓ_1, Huber and Biweight functions, in the robust stacking. Trickett (2007) proposed maximum-likelihood-estimation stacking.

(c) The noise can be estimated first and then subtracted from the data. Linville and Meek (1992) proposed a Wiener filter approach for stationary sinusoidal noise canceling. The frequency of the sinusoidal noise is known in advance or can be automatically searched if it's unknown. A reference sinusoidal trace with this particular frequency is synthesized. A Wiener filter is used to match the reference sinusoidal trace with the sinusoidal noise in the data trace. Then, the convolution of the Wiener filter and the reference sinusoidal trace is subtracted from the data trace. Butler and Russell (1993) proposed an alternative estimation-subtraction procedure for harmonic noise cancellation. The harmonic noise is modeled by a linear combination of sinusoids. The coefficients of the combination are estimated from least-squares minimization. Then, the modeled noise is subtracted from the raw data. Butler and Russell (2003) extended their method for multiple harmonic noise cancellation. Dondurur and Karsl (2012) applied Wiener filter for swell noise suppression. For each trace, the Wiener filter is derived via matching the data with the low-pass filtered data (initially estimated swell noise). Then, the modeled swell noise is substracted from the raw data. Dragoset (1995) used Widrow-Hoff LMS method for adaptive noise canceling.

Erratic noise has high-amplitude noise and does not follow Gaussian distribution. All the processing steps can be seriously affected by erratic noise. Erratic noise suppression before stacking can improve multiple removal, velocity analysis, residual statics analysis, AVO analysis, prestack migration, etc. Sometimes, poststack data may also contain erratic noise in it. Erratic noise suppression after stacking can provide data with higher quality and benefit interpretation of geological structure.

1.4 Seismic data reconstruction methods

The wavefield excitated by seismic source is continuous. It's sampled by receivers deployed on the earth surface to discrete wavefield. The economic and logistic reasons will pose problems to the acquired seismic data. First, the spatial sampling interval may be too large that the signal or coherent noise is *spatially aliased*. Second, the seismic data may be *irregularly sampled* in space, or there may be *large gap* between seismic traces.

Some data processing steps rely on the well sampled condition, e.g. coherent noise attenuation and seismic imaging. Therefore, the acquired seismic data need to be reconstructed to denser and regular data. Lots of methods for seismic data reconstruction have been proposed. The seismic data reconstruction methods can be divided into two major categories: wave-equation methods and signal processing methods. Wave-equation methods reconstruct seismic data base on wave propagation principles (Stolt, 2002; Trad, 2003; Kaplan et al., 2010). This kind of methods generally requires the subsurface velocity structure.

The signal processing methods focus on the acquired data. They do not need the prior information of velocity structure. The signal processing seismic data reconstruction methods can be divided into three types. The first kind of methods relies on transformation, which assumes that the noise-free signal is sparse in the transformation domain. The signal and noise is well separated in this domain. Sacchi et al. (1998) and Liu and Sacchi (2004) proposed Fourier transform method for data reconstruction, which assumes that the seismic data is sparse in f-k domain (time series in sparse in Fourier domain). The sparse inversion is used to induce the sparsity of model. Darche (1990) proposed parabolic Radon transform for interpolating missing seismic traces in shot-gather. Trad et al. (2002) applied hyperbolic and elliptical time domain Radon transforms for data interpolation. It assumes that the seismic data is sparse in Radon domain and missing information is retrieved via sparse inversion. Herrmann and Hennenfent (2008) adopted curvelet transform for seismic data reconstruction.

The second important category of data reconstruction methods is based on f-x prediction based filtering, which assume that the f-x noise-free seismic data is predictable in space. The filters are data-adaptive. Spitz (1991) proposed an innovative method based on f-x prediction for interpolating seismic traces of regularly missing pattern on regular grids. The prediction filter is derived from low frequency data (no aliasing) and is used to interpolate (predict) high frequency data. The f-x projection filtering was proposed for interpolating aliased seismic data (Soubaras, 1997). The f-x projection filter is equivalent to data-adaptive f-k filter. Gulunay (2003) proposed f-k domain algorithm for interpolating aliased seismic data that similar with f-x interpolation operators.

The third category of methods are the rank-reduction based methods that arise in recent

years. Trickett et al. (2010) proposed a method based on Cadzow's algorithm for seismic trace interpolation. Multichannel singular spectrum analysis is integrated into the projection onto convex sets (POCS) framework (Oropeza and Sacchi, 2011), which results in a simultaneous seismic data denoising and reconstruction algorithm. The algorithm is applied on 3D prestack seismic volume. Gao et al. (2013) extended MSSA for 5D seismic data reconstruction. SSA/Cadzow reconstruction algorithms can not handle regularly decimated and aliased data. Naghizadeh and Sacchi (2013) proposed a MSSA/Cadzow based reconstruction algorithm for interpolating regularly sampled seismic data. It extract information from low frequencies to recover the regularly missing information at high frequencies. Tensor algebra for seismic data reconstruction was investigated by Kreimer and Sacchi (2012). The higher-order singular value decomposition (HOSVD) is applied for reducing the rank of the seismic tensor.

1.5 Motivations

The motivations of this thesis are summarized as follows:

- Propose matrix rank reduction based algorithms for simultaneou removal of Gaussian noise and non-Gaussian noise.
- Propose matrix rank reduction based algorithm for robust eismic data interpolation techniques that can resist non-Gaussian noise.

1.6 Organization of this thesis

This dissertation develops robust data-driven techniques for erratic noise attenuation. The thesis is organized as follows

- Chapter 2 reviews the basic concepts of random variables, principal component analysis and the theory of traditional singular spectrum analysis method. A real data example of using SSA for time series analysis, and a synthetic example of applying SSA for seismic data denoising are presented.
- Chapter 3 reviews concepts of robust statistics, and proposes a new robust singular spectrum analysis (R-SSA) algorithm for Gaussian and non-Gaussian seismic noise attenuation. The robust SSA adopts a robust matrix factorization based on M-estimators for rank reduction. The low-rank component is obtained via the process of iteratively reweighted least-squares and alternating minimization. Synthetic and field data examples are presented to examine the effectivity of the proposed algorithm.

- Chapter 4 introduces a robust SSA algorithm for Gaussian, non-Gaussian seismic noise reduction and data reconstruction based on convex optimization. The proposed robust SSA involves a low-rank matrix recovery problem. It looks for the low-rank component from the incomplete and Gaussian and non-Gaussian (impulsive) noise corrupted data matrix. The NP-hard rank minimization problem is approximated by its tightest convex relaxation, the nuclear norm minimization problem. It changes the non-convex optimization problem to convex one. An augmented Lagrangian method is used for numerically solving the optimization problem. A synthetic example is shown to evaluate the performance of the proposed algorithm.
- Chapter 5 gives the summary and conclusions of the thesis. Moreover, future work is discussed.

CHAPTER 2

Singular spectrum analysis and its applications in seismic data processing

2.1 Introduction

Singular Spectrum Analysis (SSA) is a tool for signal-to-noise ratio (SNR) enhancement, time series analysis and forecasting. It is a non-parametric method (Golyandina and Zhigljavsky, 2013). The origin of SSA can be even traced back to Prony's method (de Prony, 1795) in the 18th century (Golyandina and Zhigljavsky, 2013). Broomhead and King (1986a) developed the SSA technique for non-linear dynamical system analysis. SSA is used to extract qualitative and quantitative information that describes the underlying dynamical system from the observed experimental data. Their approach is based on the method of delays (Takens, 1981) and the theory of singular system analysis (Bertero and Pike, 1982). They illustrated the methodology by analyzing a time series generated from a Lorenz model (chaotic dynamical system). Almost at the same time, Fraedrich (1986), in an independent fashion, developed SSA for estimating the degrees of freedom (number of independent variables) of climate system from observations. Later, Vautard and Ghil (1989), Vautard et al. (1992) and Allen and Smith (1996) further developed and extended the method and theory of SSA.

In 1988, a subspace-based algorithm used for signal-to-noise ratio enhancement similar to SSA arose in the field of signal processing (Cadzow, 1988). It is referred to as Cadzow's algorithm. The SSA denoising algorithm is equivalent to one iteration of Cadzow's algorithm. Trickett (2002) introduced Cadzow's algorithm for seismic random noise attenuation in 2-D seismic noise attenuation as an alternative to f-x prediction methods. Trickett (2008) and Trickett and Burroughs (2009) extended Cadzow's algorithm for seismic denoising of

three or more dimensional seismic data. Sacchi (2009) introduced the f-x domain Singular Spectrum Analysis (SSA) method for seismic random noise suppression and pointed out the relationship between SSA and Cadzow algorithms. SSA/Cadzow method operates in the frequency-space domain (f-x) by embedding spatial data at a given monochromatic temporal frequency into a Hankel matrix. Then, the ideal Hankel matrix that one would have formed in the absence of noise is found via the low rank approximation (truncated SVD) of the Hankel matrix of the noisy observations. The elements on the anti-diagonals of the rank-reduced matrix are averaged to get the filtered f-x data. It can efficiently remove Gaussian noise in the presence of dipping events with good preservation of signal. Another important advantage of the matrix rank-reduction based method is that it is easy to extend to multi-dimensional situation. In the field of nonlinear dynamics, the multichannel singular spectrum analysis (MSSA) is an extension of SSA that applied on multivariate time series (Broomhead and King, 1986b). In seismic data processing, MSSA is integrated into the projection onto convex sets (POCS) framework (Oropeza and Sacchi, 2011), which results in a simultaneous seismic data denoising and reconstruction algorithm. When the MSSA is applied on multi-dimensional data, the size of the block Hankel (or Teoplitz) matrix can be very large, the computation of the singular value decomposition (Golub and Van Loan, 1996) is very expensive. Oropeza and Sacchi (2011) adopted a randomized singular value decomposition algorithm to accelerate the rank reduction in MSSA. Gao et al. (2013) proposed to use the Lanczos bidiagonalization method combined with fast Toeplitz matrix-vector multiplication to reduce the computation cost. SSA/Cadzow reconstruction algorithms can not handle regularly decimated and aliased data. Naghizadeh and Sacchi (2013) proposed a MSSA/Cadzow based reconstruction algorithm for interpolating regularly sampled seismic data. It extract information from low frequencies to recover the regularly missing information at high frequencies. SSA based methods are also developed for coherent noise attenuation (Oropeza and Sacchi, 2010; Chiu, 2013; Nagarajappa, 2012).

2.2 Review of multivariate statistics

SSA arises from the analysis of dynamical systems and relies on multivariate statistics. This section reviews some basic concepts of multivariate statistics.

2.2.1 Random variable (univariate)

A *random variable* is defined as a number $\mathbf{x}(\zeta)$ assigned to every outcome ζ of an experiment (Papoulis and Pillai, 2002). $\mathbf{x}(\zeta)$ is a function of the outcome ζ. The domain of this function is the set of all outcomes of the experiment, and the range of this function is the set of the

numbers assigned to the outcomes. The random variable is a mapping from the set of outcomes to the set of numbers. For example, the weight of a randomly chosen apple from a pile of apples is a random variable. An outcome is a chosen apple and the value of the random variable for this outcome is the weight of this apple. The set of all possible outcomes of an experiment is defined as sample space Ω. A set of outcomes of an experiment is called event, say event A. The probability $P(A)$ of event A is a number that measures how likely the event A will happen. Random variables are usually described by *cumulative distribution function (c.d.f.)* or *probability density function (p.d.f.)*. The cumulative distribution function of the random variable $\mathbf{x}$ is defined as

$$F(x) = P\{\mathbf{x} \leq x\}, \tag{2.1}$$

where $\{\mathbf{x} \leq x\}$ is an event that denotes the set of all outcomes ζ such that $\mathbf{x}(\zeta) \leq x$, x is a given number, $P\{\mathbf{x} \leq x\}$ indicates the probability associated with this event. There are two categories of random variables, *continuous random variable* and *discrete random variable*. A random variable $\mathbf{x}$ is continuous if its distribution function $F(x)$ is a continuous function. Otherwise, $F(x)$ is a piecewise constant function, $\mathbf{x}$ is called discrete random variable. The probability density function is defined as the derivative of the cumulative distribution function

$$f(x) = \frac{dF(x)}{dx}. \tag{2.2}$$

Because the distribution function of a discrete random variable $\mathbf{x}$ is discontinuous (piecewise constant), the density function of discrete random variable $\mathbf{x}$ can also be expressed as

$$f(x) = \sum_i p_i \delta(x - x_i), \tag{2.3}$$

where x_i is the discontinuous point of $F(x)$, $\delta(\cdot)$ is the Dirac delta function, $p_i = P\{\mathbf{x} = x_i\}$ is the probability that event $\{\mathbf{x} = x_i\}$ happens. Expression (2.3) is often named the *probability mass function (p.m.f.)* of the discrete random variable.

2.2.2 Random vector (multivariate)

A p dimensional random vector consists of p scalar random variables

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)^T, \tag{2.4}$$

where $\mathbf{x}_i$ is a random variable, $\mathbf{X}$ is organized as a column vector. Each random variable $\mathbf{x}_i$ has its own distribution function $F_i(x_i)$, which is referred to as *marginal distribution*. The marginal distributions can only describe the marginal statistical behaviour of the random variables but not the joint statistical behaviour. The joint statistics of the random variables

is determined by the *joint distribution*

$$F(X) = F(x_1, x_2, \ldots, x_p) = P\{\mathbf{x}_1 \leq x_1, \mathbf{x}_2 \leq x_2, \ldots, \mathbf{x}_p \leq x_p\}, \tag{2.5}$$

where $X = (x_1, x_2, \ldots, x_p)^T$ is column vector of real numbers in p dimensional space. $P\{\mathbf{x}_1 \leq x_1, \mathbf{x}_2 \leq x_2, \ldots, \mathbf{x}_p \leq x_p\}$ is the probability such that event $\{\mathbf{x}_1 \leq x_1, \mathbf{x}_2 \leq x_2, \ldots, \mathbf{x}_p \leq x_p\}$ happens. Considering continuous random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, the *joint density* of them is given by

$$f(X) = f(x_1, x_2, \ldots, x_p) = \frac{\partial^p F(x_1, x_2, \ldots, x_p)}{\partial x_1 \partial x_2 \ldots \partial x_p}. \tag{2.6}$$

Each random variable $\mathbf{x}_i$ has its own density function $f_i(x_i)$, which is referred to as *marginal density.*

Independence

Events A and B are said *independent* if and only if $P(A \cap B) = P(A)P(B)$. Independence means that the occurrence of one event does not affect the probability of occurrence of the other. Mutual independence of p events $A_i, i = 1, 2, \ldots, p$, is an inductive generalization of the two events case. If any $k < p$ events are mutually independent and $P(A_1 \cap A_2 \cdots \cap A_p) = P(A_1)P(A_2) \cdots P(A_p)$, events $A_i, i = 1, 2, \ldots, p$, are said to be mutual independent (Papoulis and Pillai, 2002). If events $\{\mathbf{x}_1 \leq x_1\}, \{\mathbf{x}_2 \leq x_2\}, \ldots, \{\mathbf{x}_p \leq x_p\}$ are mutually independent, random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are also mutually independent. Then, the joint statistics and marginal statistics have the following relationship

$$\begin{aligned} F(X) = F(x_1, x_2, \ldots, x_p) &= F_1(x_1)F_2(x_2) \cdots F_p(x_p), \\ f(X) = f(x_1, x_2, \ldots, x_p) &= f_1(x_1)f_2(x_2) \cdots f_p(x_p). \end{aligned} \tag{2.7}$$

Consider that $\mathbf{x}$ is a random variable with distribution $F(x)$ defined in an experiment. If the experiment is performed p times, there will be p random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. These random variables have the same distribution $F(x)$ and they are named *independent and identically distributed (i.i.d.)* random variables.

2.2.3 Population mean, population variance, population covariance and population correlation coefficient

The *expected value* or *population mean* of a random variable $\mathbf{x}$ is a function of random variable $\mathbf{x}$, say $E\{\mathbf{x}\}$. It estimates the "center" of the random variable. For continuous

random variable $\mathbf{x}$, the expected value is defined as

$$E\{\mathbf{x}\} = \int_{-\infty}^{\infty} xf(x)dx, \tag{2.8}$$

where $f(x)$ is the p.d.f. of the random variable $\mathbf{x}$. For discrete random variable $\mathbf{x}$, the expected value is defined as

$$E\{\mathbf{x}\} = \sum_i p_i x_i, \tag{2.9}$$

where $p_i = P\{\mathbf{x} = x_i\}$ is the probability that the random variable $\mathbf{x}$ takes the value x_i. In other words, the mean of a discrete random variable is actually a weighted average of all the values of the discrete random variable, with weights as the probabilities.

The *population variance* measures the "dispersion" of random variable around its mean

$$\text{var}\{\mathbf{x}\} = \sigma^2 = E\{(\mathbf{x} - \mu)^2\}, \tag{2.10}$$

where μ is the expected value of the random variable, one can demonstrate the following

$$\begin{aligned} \sigma^2 &= E\{(\mathbf{x} - \mu)^2\} = E\{\mathbf{x}^2 - 2\mu\mathbf{x} + \mu^2\}, \\ &= E\{\mathbf{x}^2\} - (E\{\mathbf{x}\})^2, \\ &= E\{\mathbf{x}^2\} - \mu^2. \end{aligned} \tag{2.11}$$

Furthermore, if $\mathbf{x}$ is a continuous random variable

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx. \tag{2.12}$$

If $\mathbf{x}$ is a discrete random variable, its variance is

$$\sigma^2 = \sum_i p_i (x_i - \mu)^2, \tag{2.13}$$

where $p_i = P\{\mathbf{x} = x_i\}$. The variance of a discrete random variable is a weighted average of the squared distance between the value of the random variable and the mean.

The *population covariance* c_{ij} of two random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$\text{cov}\{\mathbf{x}_i, \mathbf{x}_j\} = c_{ij} = E\{(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)\}, \tag{2.14}$$

where μ_i and μ_j are the expected values of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Covariance of two random variable can describe how they are related. Using the linearity of the mean, the

covariance can be expressed as follows

$$\begin{aligned} c_{ij} &= E\{\mathbf{x}_i\mathbf{x}_j - \mu_j\mathbf{x}_i - \mu_i\mathbf{x}_j + \mu_i\mu_j\}, \\ &= E\{\mathbf{x}_i\mathbf{x}_j\} - \mu_j E\{\mathbf{x}_i\} - \mu_i E\{\mathbf{x}_j\} - \mu_i\mu_j, \\ &= E\{\mathbf{x}_i\mathbf{x}_j\} - E\{\mathbf{x}_i\}E\{\mathbf{x}_j\}, \\ &= E\{\mathbf{x}_i\mathbf{x}_j\} - \mu_i\mu_j. \end{aligned} \tag{2.15}$$

If $\mathbf{x}_i$ and $\mathbf{x}_j$ are continuous random variables, the population covariance can be calculated by

$$c_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)(x_j - \mu_j)f_{ij}(x_i, x_j)dx_i dx_j, \tag{2.16}$$

where $f_{ij}(x_i, x_j)$ is the joint density function of the random variables $\mathbf{x}_i$ and $\mathbf{x}_j$.

If $\mathbf{x}_i$ and $\mathbf{x}_j$ are discrete random variables, the population covariance can be calculated by

$$c_{ij} = \sum_i \sum_j p_{ij}(x_i - \mu_i)(x_j - \mu_j), \tag{2.17}$$

where p_{ij} is the joint probability function of the discrete random variables $\mathbf{x}_i$ and $\mathbf{x}_j$.

The population covariance matrix of a random vector $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)^T$ is given by

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{pmatrix}, \tag{2.18}$$

where the entry c_{ij} is the covariance between the random variables $\mathbf{x}_i$ and $\mathbf{x}_j$. The covariance matrix is a symmetric, nonnegative definite matrix. Using the definition (2.14), the above expression is equivalent to

$$\mathbf{C} = E\{(\mathbf{X} - E\{\mathbf{X}\})(\mathbf{X} - E\{\mathbf{X}\})^T\}. \tag{2.19}$$

The *population correlation coefficient* of two random variables is the normalized version of the covariance.

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i\sigma_j} = \frac{\mathrm{E}\{(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)\}}{\sigma_i\sigma_j}, \tag{2.20}$$

where σ_i and σ_j are the standard deviation of the random variables $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. The advantage of the correlation coefficient is that it is dimensionless, i.e. it does not depend on physical units. Covariance and correlation coefficient have the same signs. If they are zero, the two random variables are said to be uncorrelated or independent. It is easy to

notice that the covariance of random variables $\mathbf{x}_i$, $\mathbf{x}_j$ and the covariance of random variables $\mathbf{x}_i - \mu_i$, $\mathbf{x}_j - \mu_j$ are the same. This is because the means of $\mathbf{x}_i - \mu_i$ and $\mathbf{x}_j - \mu_j$ are zeros. The correlation matrix is given by

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}, \tag{2.21}$$

where ρ_{ij} is the correlation coefficient between random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ (Definition (2.28)). The correlation matrix is a symmetric, nonnegative definite matrix.

2.2.4 Complex random variable

A complex random variable $\mathbf{x} = \mathbf{a} + i\mathbf{b}$ is a function with real part $\mathbf{a}$ and imaginary part $\mathbf{b}$ are both real random variables (Papoulis and Pillai, 2002). A Complex random variable $\mathbf{x}$ is statistically described by the joint distribution $F(a, b)$ (Equation (2.5)) of the real random variables $\mathbf{a}$ and $\mathbf{b}$. In this case, the complex random variable $\mathbf{x}$ is seen as a function of two real random variables $\mathbf{a}$ and $\mathbf{b}$. For example, if $\mathbf{a}$ and $\mathbf{b}$ are independent real zero mean Gaussian random variables with the same variance, $\mathbf{x}$ is a complex zero mean Gaussian random variable. The amplitude of a complex Gaussian random variable $\mathbf{x}$ is $|\mathbf{x}| = \sqrt{\mathbf{a}^2 + \mathbf{b}^2}$ follows Rayleigh distribution. The phase of complex Gaussian random variable $\mathbf{x}$ is $\boldsymbol{\theta} = \tan^{-1}(\mathbf{a}/\mathbf{b})$ follows uniform distribution (Papoulis and Pillai, 2002).

The mean or expected value of complex random variable $\mathbf{x}$ is given as

$$E\{\mathbf{x}\} = E\{\mathbf{a}\} + iE\{\mathbf{b}\}. \tag{2.22}$$

The variance of complex random variable $\mathbf{x}$ is given by

$$\begin{aligned} \sigma^2 &= E\{(\mathbf{x} - \mu)(\mathbf{x}^* - \mu^*)\}, \\ &= E\{|\mathbf{x} - \mu|^2\}, \\ &= E\{|\mathbf{x}|^2\} - |E\{\mathbf{x}\}|^2, \\ &= E\{|\mathbf{x}|^2\} - |\mu|^2. \end{aligned} \tag{2.23}$$

where $\mu = E\{\mathbf{x}\}$, $*$ represents complex conjugate.

A complex random vector consists of several complex random variables, i.e. $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)^T$. Obviously, each complex random variable is $\mathbf{x}_i = \mathbf{a}_i + i\mathbf{b}_i$. A complex random vector does not have a joint distribution or a joint density. The statistical properties of a complex ran-

dom vector $\mathbf{X}$ are described by the joint density $f(a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_p)$ of the $2p$ real random variables $\mathbf{a}_i$ and $\mathbf{b}_i$. Complex random variables $\mathbf{x}_i$ are said to be independent if

$$f(a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_p) = f(a_1, b_1) f(a_2, b_2) \cdots f(a_p, b_p). \tag{2.24}$$

The covariance of two complex random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$\begin{aligned} c_{ij} &= E\{(\mathbf{x}_i - \mu_i)(\mathbf{x}_j^* - \mu_j^*)\}, \\ &= E\{\mathbf{x}_i \mathbf{x}_j^*\} - \mu_j^* E\{\mathbf{x}_i\} - \mu_i E\{\mathbf{x}_j^*\} + \mu_i \mu_j^*, \\ &= E\{\mathbf{x}_i \mathbf{x}_j^*\} - E\{\mathbf{x}_i\} E\{\mathbf{x}_j^*\}, \\ &= E\{\mathbf{x}_i \mathbf{x}_j^*\} - \mu_i \mu_j^*. \end{aligned} \tag{2.25}$$

The *covariance matrix* of a complex random vector $\mathbf{X}$ is given by

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{pmatrix}, \tag{2.26}$$

where the entry c_{ij} is the covariance between complex random variable $\mathbf{x}_i$ and $\mathbf{x}_j$.

$$\mathbf{C} = E\{(\mathbf{X} - E\{\mathbf{X}\})(\mathbf{X} - E\{\mathbf{X}\})^H\}. \tag{2.27}$$

The correlation coefficient between the complex random variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j} = \frac{\mathrm{E}\{(\mathbf{x}_i - \mu_i)(\mathbf{x}_j^* - \mu_j^*)\}}{\sigma_i \sigma_j}. \tag{2.28}$$

Similarly, the correlation matrix is given by

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}. \tag{2.29}$$

The covariance and correlation matrices of complex random vector are both symmetric and nonnegative definite matrix.

2.2.5 Sample mean, sample variance, sample covariance, sample correlation coefficient

In many cases, the distribution of random variables is not known in advance. The population mean, variance, covariance and correlation coefficient are not known as well. In a practical application, they can be approximately estimated by the *sample* mean, variance, covariance and correlation coefficient, respectively (Johnson and Wichern, 2007). For example, the experiment is repeatedly performed n times, n values $x_1, x_2, \ldots, x_n$ (realizations) corresponds to random variable $\mathbf{x}$ are observed. This procedure is referred to as *sampling*. Samples $x_1, x_2, \ldots, x_n$ are independent observations from a common density function $f(x)$, or joint density function $f(a, b)$ for the complex random variable case. The variables $x_1, x_2, \ldots, x_n$ can actually be regarded as n independent and identically distributed random variables.

The *sample mean* of the observations can be expressed as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.30}$$

The sample mean $\bar{x}$ is actually the arithmetic average of the observed values x_i. When $n \to \infty, \bar{x} \to E\{\mathbf{x}\}$ (Papoulis and Pillai, 2002).

Similarly, the *sample variance* is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^*, \tag{2.31}$$

where $\bar{x}$ is the sample mean from equation (2.30), s is called *sample standard deviation (SD)*. Sample variance is a measure of spread of the observed values.

If every random variable in random vector $\mathbf{X}$ has n observations, the n by p data matrix can be organized as

$$\mathbf{X}_{obs} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix}, \tag{2.32}$$

where x_{ik} represents ith observation of the kth random variable. Each row of data matrix $\mathbf{X}_{obs}$ is a measurement with p components. In other words, all the values in the data matrix are unknown before the occurrence of the measurement. Then every element

x_{ik} can be regarded as a random variable, every row of $\mathbf{X}_{obs}$ can be regarded as a random vector, and matrix $\mathbf{X}_{obs}$ can be regarded as a random matrix (Johnson and Wichern, 2007). If the rows of $\mathbf{X}_{obs}$ are observations from a common joint density function $f(a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_p)$ and are mutual independent, they are called a *random sample* from the distribution $f(a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_p)$ (Johnson and Wichern, 2007).

The *sample covariance* between complex random variables $\mathbf{x}_j$ and $\mathbf{x}_k$ is given by

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)^*, \tag{2.33}$$

where $\bar{x}_j$ and $\bar{x}_k$ are the sample mean of random variables $\mathbf{x}_j$ and $\mathbf{x}_k$, respectively.

The *sample correlation coefficient* is a normalized version of sample covariance

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \tag{2.34}$$

where c_{jk} is the sample covariance between jth and kth variables, s_{jj} and s_{kk} are the sample variances of jth and kth variables, respectively.

To summarize the descriptive statistics, the sample means of a random vector $\mathbf{X}$ is given by

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}. \tag{2.35}$$

Sample covariance matrix of the random vector $\mathbf{X}$ is given by

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \tag{2.36}$$

Sample correlation coefficient matrix of the random vector $\mathbf{X}$ is given by

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} \tag{2.37}$$

2.2.6 Eigendecomposition

Eigenvectors of a n by n square matrix $\mathbf{S}$ are vectors that do not change their direction after they are multiplied by the matrix $\mathbf{S}$ (Strang, 1993)

$$\mathbf{S}\mathbf{u} = \lambda\mathbf{u}, \tag{2.38}$$

where λ is a scalar called *eigenvalue* whereas $\mathbf{u}$ is a $n \times 1$ vector called *eigenvector*. We should note that only a square matrix has eigenvectors and eigenvalues. Zero is one of the eigenvalues of the matrix $\mathbf{S}$ if it is singular. The multiplication of any nonzero scalar with an eigenvector of $\mathbf{S}$ is still an eigenvector of $\mathbf{S}$. The eigenvectors are usually normalized to unit vectors. If matrix $\mathbf{S}$ has n linearly independent eigenvectors

$$\begin{aligned} \mathbf{S}\mathbf{u}_1 &= \lambda_1\mathbf{u}_1, \\ \mathbf{S}\mathbf{u}_2 &= \lambda_2\mathbf{u}_2, \\ &\vdots \\ \mathbf{S}\mathbf{u}_n &= \lambda_n\mathbf{u}_n. \end{aligned} \tag{2.39}$$

The above equations can be organized into matrix form

$$\begin{aligned} \mathbf{S}\left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\right) &= \left(\lambda_1\mathbf{u}_1, \lambda_2\mathbf{u}_2, \ldots, \lambda_n\mathbf{u}_n\right), \\ \mathbf{S}\left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\right) &= \left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\right) \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}. \end{aligned} \tag{2.40}$$

Noting that $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$, $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, equation (2.40) is expressed as

$$\mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}. \tag{2.41}$$

All the n eigenvectors are assumed to be linearly independent, i.e. all column vectors of the matrix $\mathbf{U}$ are linearly independent. The matrix $\mathbf{U}$ is invertible

$$\mathbf{U}^{-1}\mathbf{S}\mathbf{U} = \mathbf{\Lambda}. \tag{2.42}$$

This process is called *matrix diagonalization* and matrix $\mathbf{S}$ is said to be *diagonalizable*. Some other matrices are *nondiagonalizable* because at least one of their eigenvalues is not a simple root of the characteristic polynomial (multiple eigenvalue). They are called *defective matrix*. From equation (2.42), the eigendecomposition or spectral decomposition of matrix $\mathbf{S}$ is given

by

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}. \tag{2.43}$$

Hermitian matrix ($\mathbf{S} = \mathbf{S}^H$) is always diagonalizable. All eigenvalues of Hermitian matrix are real numbers. The eigenvectors of Hermitian matrix are orthonormal. They constitute an unitary matrix, i.e. $\mathbf{U}^{-1} = \mathbf{U}^H$ and $\mathbf{U}^H\mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathbf{I}$. The eigendecomposition (2.43) changes to (Strang, 2006)

$$\begin{aligned} \mathbf{S} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \\ \mathbf{S} &= \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^H. \end{aligned} \tag{2.44}$$

The outer product $\mathbf{u}_i\mathbf{u}_i^H$ is a rank 1 matrix. The eigendecomposition decomposes an n by n Hermitian matrix into a weighted combination of n rank 1 matrices.

An n by n matrix $\mathbf{S}$ is defined as *positive definite matrix* if

$$\mathbf{u}^H\mathbf{S}\mathbf{u} > 0, \tag{2.45}$$

for any nonzero complex vector $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$. All the eigenvalues of a positive definite matrix are positive numbers. If $\mathbf{S}$ is both symmetric and positive definite, it is called *a symmetric positive definite matrix*. Similarly, if

$$\mathbf{u}^H\mathbf{S}\mathbf{u} \geq 0, \tag{2.46}$$

for all complex vector $\mathbf{u}$, $\mathbf{S}$ is said to be a *nonnegative definite matrix* or *positive semi-definite matrix*. All eigenvalues of nonnegative definite matrix are nonnegative numbers. The population covariance/correlation matrix and sample covariance/correlation matrix are symmetric nonnegative definite matrices.

2.2.7 Principal component analysis

Principal component analysis (PCA) is a very important tool in multivariate statistic analysis. It is widely used for data compression and interpretation. PCA is closely related with Karhunen-Loeve transform (Karhunen, 1947; Loeve, 1948) in signal processing, singular value decomposition (Golub and Van Loan, 1996) in numerical analysis, Hotelling transformation in image analysis, eigendecomposition in physical sciences, and empirical orthogonal functions in meteorology. The variables in a large data set are actually correlated in some way. PCA transforms the original variables to a new set of uncorrelated variables, the principal components (PCs). The first few new variables account for as much as possible of variability in the original data variables (Jolliffe, 2010). In this way, the underlying

independent variables are extracted and the dimensionality of the data set is reduced by PCA.

Population principal components

Consider a random vector $\mathbf{X}$, which consists of p scalar random variables $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)^T$. The realizations (sample observations) of the random variable $\mathbf{x}_i$ can be a subseries $\mathbf{m}_i$ windowed from a time series (SSA) or a seismic trace (Eigenimage analysis) (Freire and Ulrych, 1988). PCA linearly transforms the random variables to a set of new variables with maximum variance. The new variables are called principal components (PCs). Every PC is a linear function of all the original random variables. More specially, PCA looks for a linear combination of the original random variables, which has maximum variance subject to the constraint that the sum of squares of the transformation coefficients equals 1

$$\mathbf{w}_1 = \mathbf{u}_1^T \mathbf{X} = \sum_{j=1}^{p} u_{j1} \mathbf{x}_j, \tag{2.47}$$

where $\mathbf{u}_1^T \mathbf{u}_1 = 1$, $\mathbf{w}_1$ is the first PC. Then, PCA looks for the second PC $\mathbf{u}_2^T \mathbf{X}$ with maximum variance, which is *uncorrelated* with the first PC and fulfills the constraint that $\mathbf{u}_2^T \mathbf{u}_2 = 1$. Similarly, the ith PC should have maximum variance under the constraint that it is uncorrelated with all previous $i-1$ PCs. The problem of computing ith PC ($i > 1$) can be summarized as

$$\begin{aligned} &\hat{\mathbf{u}}_i = \underset{\mathbf{u}_i}{\operatorname{argmax}} \operatorname{var}\{\mathbf{u}_i^T \mathbf{X}\}, \\ &\text{subject to} \quad \mathbf{u}_i^T \mathbf{u}_i = 1, \\ &\operatorname{cov}\{\mathbf{u}_i^T \mathbf{X}, \mathbf{u}_l^T \mathbf{X}\} = 0 \quad \text{for} \quad l < i, \end{aligned} \tag{2.48}$$

where $\operatorname{var}\{\mathbf{u}_i^T \mathbf{X}\} = \mathbf{u}_i^T \operatorname{var}\{\mathbf{X}\} \mathbf{u}_i = \mathbf{u}_i^T \mathbf{C} \mathbf{u}_i$ and $\operatorname{cov}\{\mathbf{u}_i^T \mathbf{X}, \mathbf{u}_l^T \mathbf{X}\} = \mathbf{u}_i^T \mathbf{C} \mathbf{u}_l$. Here, $\mathbf{C}$ is the population covariance matrix of the random vector $\mathbf{X}$. There are in total p principal components to account for all the variability information in the original data set. However, usually most of the variability can be explained by a small number k of principal components. $\mathbf{u}_i$ is called the vector of coefficients or loadings for the ith PC. Elements of $\mathbf{u}_i$ are termed as PC coefficients or PC loadings.

The optimization problem (2.48) can be solved via *the Lagrange multiplier method* (Jolliffe, 2010). The problem of computing principal components is finally addressed by the eigendecomposition of the covariance or correlation matrix of the original random variables. The transformed coefficient vector $\mathbf{u}_i$ of the ith PC is actually the normalized eigenvector of the

covariance matrix $\mathbf{C}$ corresponding to the ith largest eigenvalue

$$\begin{aligned} \mathbf{C}\mathbf{u}_1 &= \lambda_1 \mathbf{u}_1, \\ \mathbf{C}\mathbf{u}_2 &= \lambda_2 \mathbf{u}_2, \\ &\vdots \\ \mathbf{C}\mathbf{u}_p &= \lambda_p \mathbf{u}_p, \end{aligned} \tag{2.49}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are eigenvalues of the covariance matrix $\mathbf{C}$ in non-increasing order, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ are the corresponding eigenvectors. The above equations can be arranged as

$$\mathbf{C}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \tag{2.50}$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p)$ and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$. Because $\mathbf{C}$ is a symmetric matrix, matrix $\mathbf{U}$ is an orthogonal matrix. It follows that the covariance matrix $\mathbf{C}$ has the eigendecomposition

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T. \tag{2.51}$$

The variance of ith PC is actually the ith largest eigenvalue of the covariance matrix $\mathbf{C}$

$$\text{var}\{\mathbf{u}_i^T \mathbf{X}\} = \mathbf{u}_i^T \mathbf{C}\mathbf{u}_i = \mathbf{u}_i^T \lambda_i \mathbf{u}_i = \lambda_i \mathbf{u}_i^T \mathbf{u}_i = \lambda_i. \tag{2.52}$$

The principal components can be organized as a random vector

$$\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p)^T = \mathbf{U}^T \mathbf{X}, \tag{2.53}$$

where $\mathbf{U}$ is an orthogonal matrix that consists of the eigenvectors of the population covariance matrix. Then, the original random vector $\mathbf{X}$ can be expressed as a transformation from the PCs

$$\mathbf{X} = \mathbf{U}\mathbf{W}. \tag{2.54}$$

There is a *geometrical interpretation* for population PCA (Jolliffe, 2010; Johnson and Wichern, 2007). The random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ represent the coordinate axes of the original Cartesian coordinate system. The following quadratic form represents an ellipsoid centered at the origin in the p dimensional Cartesian coordinate system

$$\mathbf{X}^T \mathbf{C}^{-1} \mathbf{X} = c^2, \tag{2.55}$$

where $\mathbf{X}$ is random vector, $\mathbf{C}$ is the population covariance matrix, and c is a constant. The principal axes of the ellipsoid are $\pm c\sqrt{\lambda_i}\mathbf{u}_i$, where λ_i and $\mathbf{u}_i$ are the ith eigenvalue and eigenvector of $\mathbf{C}$, respectively. That is to say, the directions of the axes are determined by

$\mathbf{u}_i$, and half-length of the principal axes are $c\sqrt{\lambda_i}$ in the original coordinate system. From relationship (2.54) and decomposition (2.51), equation of p dimensional ellipsoid (2.55) can be changed to

$$\begin{aligned}(\mathbf{UW})^T\mathbf{C}^{-1}(\mathbf{UW}) &= c^2,\\ \mathbf{W}^T(\mathbf{U}^T\mathbf{CU})^{-1}\mathbf{W} &= c^2,\\ \mathbf{W}^T\mathbf{\Lambda}^{-1}\mathbf{W} &= c^2,\end{aligned} \tag{2.56}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with a diagonal composed of the eigenvalues of $\mathbf{C}$ in decreasing order.

$$\frac{1}{\lambda_1}\mathbf{w}_1^2 + \frac{1}{\lambda_2}\mathbf{w}_2^2 + \ldots + \frac{1}{\lambda_p}\mathbf{w}_p^2 = c^2, \tag{2.57}$$

where $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$ are the principal components. It is clear that equation (2.57) represents a p dimensional ellipsoid (Strang, 1993; Jolliffe, 2010) with the principal components defining the direction of its principal axes. The half-lengths of the axes are $c\sqrt{\lambda_i}$. The principal components are the coordinate axes of the new Cartesian coordinate system. In fact, PCA involves a linear transformation of coordinate axes. The old coordinate axes are the original variables and the new coordinate axes are given by the principal components. The transformation matrix $\mathbf{U}$ indicating *rotation* and *stretch* from the original coordinate axes to the new coordinate axes. Figure 2.1 uses a 2-D random vector to show the geometric interpretation of population PCA. $\mathbf{x}_1$ and $\mathbf{x}_2$ are original random variables. The population covariance matrix is $\mathbf{C} = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 3 \end{pmatrix}$. The equation of the ellipse in Figure 2.1 is given by equation (2.55) with the above population covariance matrix and constant $c = 2$. $\mathbf{w}_1$ and $\mathbf{w}_2$ are the principal components. If the random vector follows a multivariate Gaussian distribution, the ellipsoid (Equation (2.55)) represents the constant probability density contour of $\mathbf{X}$. The directions of principal axes of constant probability density are determined by the principal components. The leading principal axes determine the directions of greatest statistical variations. Note that the PCA can also be performed with the correlation matrix. In some situations, e.g. the units used for different variables are different, the random variables are standardized using the standard deviation of each variable. The covariance matrix of the standardized variables equals to the correlation matrix of the original variables. However, the relationship between the eigenvalues/eigenvectors of the covariance matrix and eigenvalues/eigenvectors of the correlation matrix is not simple (Jolliffe, 2010).

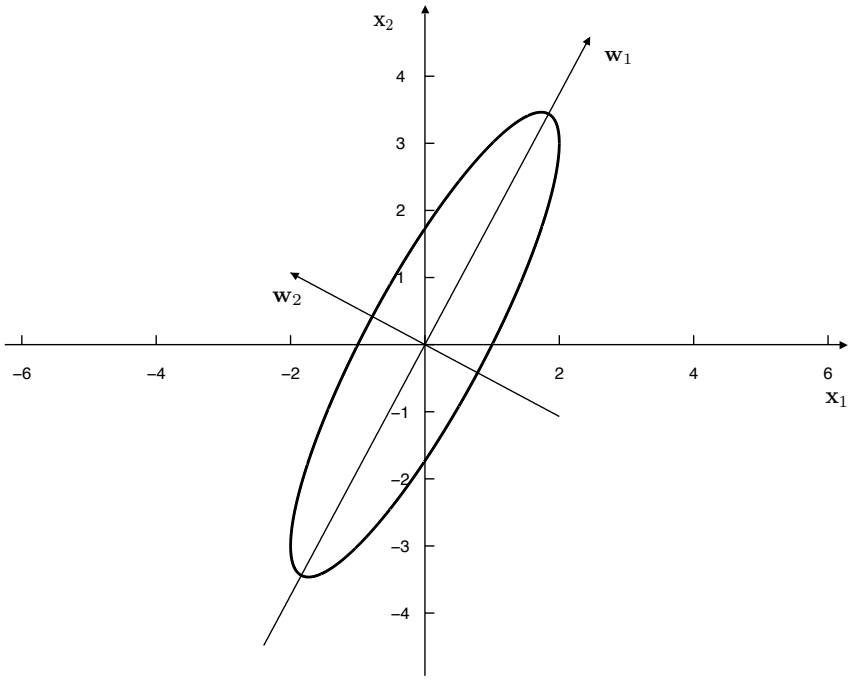

Figure 2.1: Geometric interpretation of population PCA, a 2-variate example. $\mathbf{x}_1$, $\mathbf{x}_2$ are the original random variable, also the original coordinate axes; $\mathbf{w}_1$ and $\mathbf{w}_2$ are the principal components, also the new coordinate axes. $\mathbf{w}_1 = \mathbf{u}_1^T\mathbf{X} = u_{11}\mathbf{x}_1 + u_{21}\mathbf{x}_2, \mathbf{w}_2 = \mathbf{u}_2^T\mathbf{X} = u_{12}\mathbf{x}_1 + u_{22}\mathbf{x}_2$. The ellipse is defined by the population covariance matrix $\mathbf{C}$ and constant c, which can represent a constant probability density contour of $\mathbf{X}$.

Sample principal components

If the p variate random vector $\mathbf{X}$ is sampled n times independently, a n by p data matrix $\mathbf{X}_{obs}$ is available.

$$\mathbf{X}_{obs} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}. \tag{2.58}$$

The n rows of $\mathbf{X}_{obs}$, $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$, are n p-dimensional column vectors representing n independent observations. The sample mean, sample covariance matrix and sample correlation matrix are given by equations (2.35)(2.36)(2.37). PCA looks for uncorrelated linear combinations of original variables that as much of variations in the original data can be explained by the first few linear combinations. These linear combinations are named as *sample principal components*. They are derived from the sample covariance matrix or sample correlation

matrix. The n observed values of the first linear combination are

$$\begin{aligned} w_{11} &= \mathbf{u}_1^T\mathbf{x}^1 = u_{11}x_{11}+u_{21}x_{12}+\ldots+u_{p1}x_{1p}, \\ w_{21} &= \mathbf{u}_1^T\mathbf{x}^2 = u_{11}x_{21}+u_{21}x_{22}+\ldots+u_{p1}x_{2p}, \\ &\vdots \\ w_{n1} &= \mathbf{u}_1^T\mathbf{x}^n = u_{11}x_{n1}+u_{21}x_{n2}+\ldots+u_{p1}x_{np}. \end{aligned} \tag{2.59}$$

or

$$\mathbf{w}_1 = \mathbf{X}_{obs}\mathbf{u}_1. \tag{2.60}$$

The vector of coefficients $\mathbf{u}_1$ maximizes the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(w_{i1}-\bar{w}_1)^2$ under the constraint that $\mathbf{u}_1^T\mathbf{u}_1 = 1$. $\bar{w}_1$ is the sample mean of w_{i1}. The linear combination of random variables $\mathbf{u}_1^T\mathbf{X}$ is defined as the 1st sample principal component (PC). w_{i1} is called the score of the ith observation $\mathbf{x}^i$ on the 1st PC. Similarly, PCA looks for second linear combination of variables $\mathbf{u}_2^T\mathbf{X}$ with the optimal vector of coefficients $\mathbf{u}_2$. The vector $\mathbf{u}_2$ maximizes the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(w_{i2}-\bar{w}_2)^2$ subject to $\mathbf{u}_2^T\mathbf{u}_2 = 1$ and $\mathbf{w}_2$ is uncorrelated with $\mathbf{w}_1$. At the jth step, the problem of computing the optimal vector of coefficients $\hat{\mathbf{u}}_j$ of the jth sample PC ($j > 1$) is

$$\begin{aligned} &\text{maximizes } \frac{1}{n-1}\sum_{i=1}^{n}(w_{ij}-\bar{w}_j)^2, \\ &\text{subject to } \mathbf{u}_j^T\mathbf{u}_j = 1, \\ &\frac{1}{n-1}\sum_{i=1}^{n}(w_{ij}-\bar{w}_j)(w_{il}-\bar{w}_l) = 0, \quad l < j, \end{aligned} \tag{2.61}$$

where $\mathbf{w}_j = \mathbf{X}_{obs}\mathbf{u}_j$ are the scores of observations on the jth sample PC. Linear combination $\mathbf{u}_j^T\mathbf{X}$ is the jth sample PC. Although there are p sample PCs in total, a small number k of sample PCs account for much of the variability in the original data set.

In problem (2.61), the sample variance of jth sample PC $\mathbf{u}_j^T\mathbf{X}$ is given by

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^{n}(w_{ij}-\bar{w}_j)^2 &= \frac{1}{n-1}(\mathbf{w}_j-\bar{w}_j\mathbf{1}_n)^T(\mathbf{w}_j-\bar{w}_j\mathbf{1}_n) \\ &= \frac{1}{n-1}(\mathbf{X}_{obs}\mathbf{u}_j-\bar{\mathbf{x}}^T\mathbf{u}_j\mathbf{1}_n)^T(\mathbf{X}_{obs}\mathbf{u}_j-\bar{\mathbf{x}}^T\mathbf{u}_j\mathbf{1}_n) \\ &= \frac{1}{n-1}(\mathbf{X}_{obs}\mathbf{u}_j-\mathbf{1}_n\bar{\mathbf{x}}^T\mathbf{u}_j)^T(\mathbf{X}_{obs}\mathbf{u}_j-\mathbf{1}_n\bar{\mathbf{x}}^T\mathbf{u}_j) \\ &= \frac{1}{n-1}\mathbf{u}_j^T(\mathbf{X}_{obs}-\mathbf{1}_n\bar{\mathbf{x}}^T)^T(\mathbf{X}_{obs}-\mathbf{1}_n\bar{\mathbf{x}}^T)\mathbf{u}_j \\ &= \mathbf{u}_j^T\mathbf{S}\mathbf{u}_j, \end{aligned} \tag{2.62}$$

where $\bar{\mathbf{x}}$ is the sample mean vector of $\mathbf{X}$, $\mathbf{1}_n$ is a $n \times 1$ vector with all elements equal to 1, and $\mathbf{S} = \frac{1}{n-1}(\mathbf{X}_{obs} - \mathbf{1}_n\bar{\mathbf{x}}^T)^T(\mathbf{X}_{obs} - \mathbf{1}_n\bar{\mathbf{x}}^T)$ is the sample covariance matrix of $\mathbf{X}$. The sample covariance between jth and lth sample PC is given by

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)(w_{il} - \bar{w}_l) &= \frac{1}{n-1}(\mathbf{w}_j - \bar{w}_j\mathbf{1}_n)^T(\mathbf{w}_l - \bar{w}_l\mathbf{1}_n) \\
&= \frac{1}{n-1}(\mathbf{X}_{obs}\mathbf{u}_j - \bar{\mathbf{x}}^T\mathbf{u}_j\mathbf{1}_n)^T(\mathbf{X}_{obs}\mathbf{u}_l - \bar{\mathbf{x}}^T\mathbf{u}_l\mathbf{1}_n) \\
&= \frac{1}{n-1}(\mathbf{X}_{obs}\mathbf{u}_j - \mathbf{1}_n\bar{\mathbf{x}}^T\mathbf{u}_j)^T(\mathbf{X}_{obs}\mathbf{u}_l - \mathbf{1}_n\bar{\mathbf{x}}^T\mathbf{u}_l) \\
&= \frac{1}{n-1}\mathbf{u}_j^T(\mathbf{X}_{obs} - \mathbf{1}_n\bar{\mathbf{x}}^T)^T(\mathbf{X}_{obs} - \mathbf{1}_n\bar{\mathbf{x}}^T)\mathbf{u}_l \\
&= \mathbf{u}_j^T\mathbf{S}\mathbf{u}_l.
\end{aligned} \tag{2.63}$$

Following the similar Lagrange multiplier method as in population PCA (Jolliffe, 2010), the optimal solution of PC coefficient corresponds to the jth PC is the eigenvector of sample covariance matrix corresponding to the jth largest eigenvalue. The sample variance of PC scores of the jth sample PC equals to the jth largest eigenvalue of the sample covariance matrix

$$\begin{aligned}
\mathbf{S}\mathbf{u}_j &= \lambda_j\mathbf{u}_j, \\
\mathbf{u}_j^T\mathbf{S}\mathbf{u}_j &= \lambda_j.
\end{aligned} \tag{2.64}$$

Grouping them together

$$\begin{aligned}
\mathbf{S}\mathbf{U} &= \mathbf{U}\mathbf{\Lambda}, \\
\mathbf{S} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T.
\end{aligned} \tag{2.65}$$

The matrix of PC scores is linear orthonormal transformation of the data matrix

$$\mathbf{W} = \mathbf{X}_{obs}\mathbf{U}. \tag{2.66}$$

In general, the means of each column of $\mathbf{X}$ are not zero. It is convenient to center the data matrix, i.e. subtract corresponding column mean from each column of $\mathbf{X}_{obs}$.

$$\tilde{\mathbf{X}}_{obs} = \mathbf{X}_{obs} - \mathbf{1}_n\bar{\mathbf{x}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}. \tag{2.67}$$

Then, the sample covariance matrix is given by

$$\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}_{obs}^T\tilde{\mathbf{X}}_{obs}. \tag{2.68}$$

The new matrix of PC scores is given by

$$\tilde{\mathbf{W}} = \tilde{\mathbf{X}}_{obs}\mathbf{U}. \tag{2.69}$$

Now, PC scores $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_p$ have zero sample mean and same sample variances and covariances as PC scores $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$.

There is also a *geometric interpretation* for the sample PCA. Data matrix $\mathbf{X}_{obs}$ can be represented by a group of n points in p dimensional space, which is referred to as *scatter plot*. Consider an ellipsoid, the coordinates $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ of the point on this ellipsoid fulfill the following equation

$$(\mathbf{X} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{X} - \bar{\mathbf{x}}) = c^2, \tag{2.70}$$

where $\mathbf{S}$ is the sample covariance matrix, $\bar{\mathbf{x}}$ is the sample mean vector, c is a constant representing the Mahalanobis distance from data samples to the sample mean. The equation means that the distance between the points on the ellipsoid and the center point $\bar{\mathbf{x}}$ is a constant. Figure 2.2 shows the geometric interpretation of sample PCA. There are 200 data samples drawn from multivariate Gaussian distribution with population covariance matrix $\mathbf{C}$ same as the one in Figure 2.1. The sample covariance matrix $\mathbf{S}$ can be derived from these data samples. Ellipse in Figure 2.2 is determined by equation (2.70) with sample covariance matrix $\mathbf{S}$, sample mean vector $\bar{\mathbf{x}} = \mathbf{0}$ and constant $c = 2$. $\mathbf{w}_1$ and $\mathbf{w}_2$ are sample principal components. The directions of principal axes of the ellipse are determined by the eigenvectors of sample covariance matrix. Similar with the situation in population PCA, the sample PCA can also be based on sample correlation matrix rather than sample covariance matrix.

2.3 Theory of SSA

This section briefly review the theory and algorithm of SSA. Some basic concepts of dynamical system are also mentioned, which are just used to show the origin of SSA.

2.3.1 Dynamical system

Basically, the dynamical system can be described by a system of ordinary differential equations (ODEs). That is to say, the ODEs govern how the variables of the dynamical system change with respect to continuous time (Broomhead and King, 1986a)

$$\frac{dz_i}{dt} = f_i(z_1, z_2, \ldots, z_q), \quad i = 1, 2, \ldots, q, \tag{2.71}$$

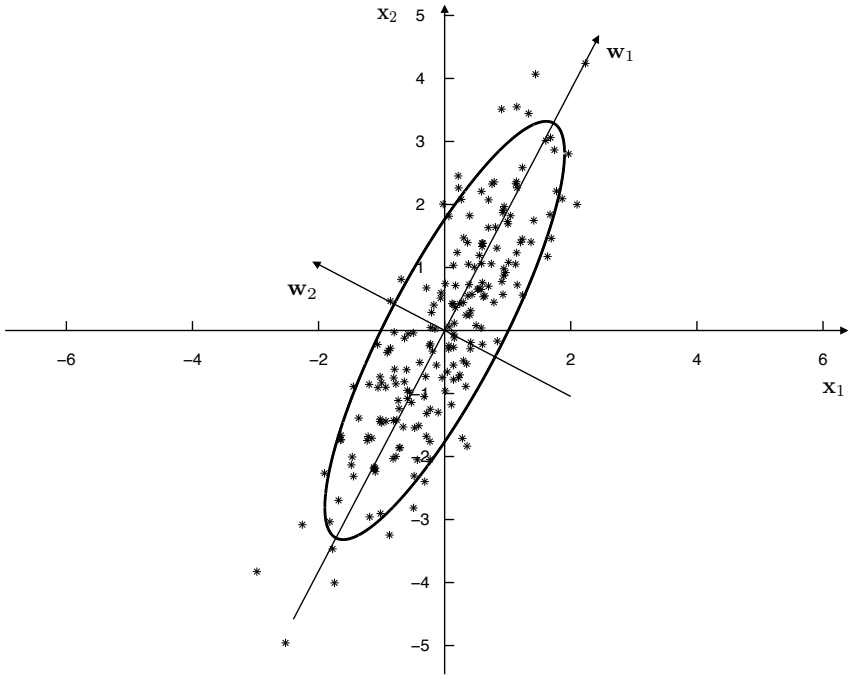

Figure 2.2: Geometric interpretation of sample PCA, a 2-variate example. $\mathbf{x}_1$, $\mathbf{x}_2$ are the original random variable, also the original coordinate axes; $\mathbf{w}_1$ and $\mathbf{w}_2$ are the sample principal components, also the new coordinate axes. $\mathbf{w}_1 = \mathbf{u}_1^T\mathbf{X} = u_{11}\mathbf{x}_1 + u_{21}\mathbf{x}_2, \mathbf{w}_2 = \mathbf{u}_2^T\mathbf{X} = u_{12}\mathbf{x}_1 + u_{22}\mathbf{x}_2$. $\mathbf{u}_1$ and $\mathbf{u}_2$ are the eigenvectors of the sample covariance matrix $\mathbf{S}$. The asterisks represent 200 data samples drawn from multivariate Gaussian distribution with covariance matrix $\mathbf{C}$ and zero mean. The ellipse is defined by the sample covariance matrix $\mathbf{S}$ and constant c.

or

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}), \tag{2.72}$$

where dynamical variables $\mathbf{z} := (z_1, z_2, \ldots, z_q)$ indicate a *state* of the dynamical system, $\dot{\mathbf{z}} := \frac{d\mathbf{z}}{dt} = (\frac{dz_1}{dt}, \frac{dz_2}{dt}, \ldots, \frac{dz_q}{dt})$. $\mathbf{f} := (f_1, f_2, \ldots, f_q)$ is referred as a *vector field*. If $\mathbf{f}$ is a non-linear operator, the dynamical system is a non-linear dynamical system. Otherwise, the dynamical system is a linear one. The space spanned by variables $z_i, i = 1, 2, \ldots, q$ is referred to as *phase space*, $S \in \mathbb{R}^q$. Each state can be regarded as a point in the phase space. Given an initial value $\mathbf{z}_0$, the solution $\mathbf{z}(t)$ of the ODEs corresponds to a curve in the phase space. This curve is named *trajectory* or *orbit* of the dynamical system, which portraits the evolution of the dynamical system with time under the initial condition. The ensemble of solutions to all possible initial conditions corresponds to a collection of trajectories in the phase space, which is called the *flow*.

2.3.2 Embedding into a trajectory matrix

Each measurement d_i of a time series is a function of a state of the system at the particular time. The N data points time series $\mathbf{d} = (d_1, d_2, \ldots, d_N) \in \mathbb{R}^N$ is windowed into vectors $\mathbf{m}_i \in \mathbb{R}^L, i = 1, 2, \ldots, N - L + 1$ in the *embedding space* $\mathbb{R}^L$

$$\begin{aligned} \mathbf{m}_1 &= (d_1, d_2, \ldots, d_L)^T, \\ \mathbf{m}_2 &= (d_2, d_3, \ldots, d_{L+1})^T, \\ &\vdots \\ \mathbf{m}_{N-L+1} &= (d_{N-L+1}, d_{N-L+2}, \ldots, d_N)^T, \end{aligned} \tag{2.73}$$

where L is the dimension of embedding space referred to as *embedding dimension* or *window length*, $N-L+1$ vectors $\mathbf{m}_i$ are called *snapshots* (Elsner and Tsonis, 1996). This procedure is referred as *embedding*, it is a mapping from manifold to the embedded vector space $\mathbb{R}^L$. The snapshots represent a discrete trajectory in the phase space of the system. These vectors are then constructed to a so-called *trajectory matrix* (Golyandina and Zhigljavsky, 2013)

$$\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{N-L+1}) \tag{2.74}$$

$$= \mathcal{H}[\mathbf{d}] = \begin{pmatrix} d_1 & d_2 & \cdots & d_{N-L+1} \\ d_2 & d_3 & \cdots & d_{N-L+2} \\ \vdots & \vdots & \ddots & \vdots \\ d_L & d_{L+1} & \cdots & d_N \end{pmatrix}, \tag{2.75}$$

where operator $\mathcal{H}$ constructs a vector to a *Hankel matrix*, which is named Hankel operator. Hankel matrix means that the element on ith row and jth column is $m_{ij} = d_{i+j-1}$, i.e. the elements on the anti-diagonals ($i + j$ is constant) is the same. All the column vectors and row vectors of $\mathbf{M}$ are subseries windowed from the original time series $\mathbf{d}$. The trajectory matrix relates the single time series to the multivariate statistics analysis. $\mathbf{M}$ can be seen as $N-L+1$ observations on a L-variate random vector. To make the notation simpler, we will note $L = m$ and $N - L + 1 = n$ in the following discussions ($\mathbf{M} \in \mathbb{R}^{m \times n}$). The rank of the trajectory matrix is the dimensionality of subspace that contains the embedding manifold (attractors) if there is no measurement noise (Broomhead and King, 1986a). .

2.3.3 Trajectory matrix decomposition

PCA is used to reveal the underlying structure of the trajectory matrix. The PC coefficients are calculated by singular value decomposition of the trajectory matrix. Before going further, I would like to point out that the times series should be centered (subtracting mean)

and normalized before the application of SSA. If the mean value is not subtracted first, it will be contained in the first principal component. Because the time series is normalized by the same scalar, all elements in the trajectory matrix are normalized by the same scalar. The sample covariance matrix before and after normalization will have the same eigenvectors but different eigenvalues. The ratio of corresponding eigenvalues of the covariance matrix before and after normalization is the normalization scalar.

Singular value decomposition

As I mentioned before, principal component analysis is an analysis tool in multivariate statistics for analyzing the concealed independent variables in the original data set. The PC coefficients can be calculated via eigendecomposition of the sample covariance matrix. It also can be calculated via Singular Value Decomposition (SVD) of the centered data matrix. SVD is a matrix decomposition tool. Eigendecomposition is used for square matrices. While, SVD can factor an arbitrary rectangular or square matrix. Any m by n matrix $\mathbf{M}$ can be diagonalized as (Golub and Van Loan, 1996)

$$\mathbf{U}^T\mathbf{M}\mathbf{V} = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_p\} \quad \in \mathbb{R}^{m\times n},\ p = \min\{m, n\} \tag{2.76}$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m) \in \mathbb{R}^{m\times m}$ and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \in \mathbb{R}^{n\times n}$ are orthogonal matrices, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_{m\times m}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_{n\times n}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$. σ_i is the ith largest *singular value* of $\mathbf{M}$, $\mathbf{u}_i$ is the ith *left singular vector* of $\mathbf{M}$ and $\mathbf{v}_i$ is the ith *right singular vector* of $\mathbf{M}$, corresponding to singular value σ_i. The collection of singular values ordered from large to small is named *singular spectrum*. Because of $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, $\mathbf{M}$ has the decomposition

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{2.77}$$

where $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_p\} \in \mathbb{R}^{m\times n}$.

Suppose that $m > n$

$$\mathbf{M} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2n} & \cdots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} & \cdots & u_{nm} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} & \cdots & u_{mm} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix}^T . \tag{2.78}$$

There are zero rows in $\mathbf{\Sigma}$, some information is redundant. A more economical version of SVD is available as

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{2.79}$$

where $\mathbf{U} \in \mathbb{R}^{m\times n}(m > n)$, $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_n\} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$. This version of SVD is called *thin SVD* (Golub and Van Loan, 1996). Express above equation in the form of elements

$$\mathbf{M} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix}^T, \tag{2.80}$$

where $\mathbf{V}$ is still an orthogonal matrix, $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_{n\times n}$. However, $\mathbf{U}$ is not an orthogonal matrix. The columns of $\mathbf{U}$ are orthonormal but the rows of $\mathbf{U}$ are not, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{n\times n}$ and $\mathbf{U}\mathbf{U}^T \neq \mathbf{I}_{m\times m}$.

The *rank* of a matrix is defined as the number of linearly independent rows or columns of the matrix. It equals to the number of nonzero singular values of the matrix (Hansen, 1998). If rank of matrix $\mathbf{M}$ is r $(r < p)$

$$\mathbf{M} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \mathbf{I}_1 + \mathbf{I}_2 + \ldots + \mathbf{I}_r, \tag{2.81}$$

because $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_p = 0$, outer product $\mathbf{u}_i\mathbf{v}_i^T$ is a rank 1 matrix, it's called ith eigenimage (Andrews and Hunt, 1977; Ulrych et al., 1988). Each eigenimage is weighted by the corresponding singular value. The matrix $\mathbf{I}_i = \sigma_i\mathbf{u}_i\mathbf{v}_i^T$ is referred to as the ith weighted eigenimage. The SVD provides an efficient way for computing the principal components in PCA. SVD of a matrix is closely related with the eigendecomposition of the sample covariance matrix. Remember that $\mathbf{M} \in \mathbb{R}^{m\times n}$ represents a set of n observations of a m-variate variable vector (Section 2.3.2).

In SSA, the *covariance matrix* is defined as $\mathbf{S} = \mathbf{M}\mathbf{M}^T = \sum_{j=1}^{n} \mathbf{m}_j\mathbf{m}_j^T$. There is a scalar difference $1/n$ with the one defined in (Broomhead and King, 1986a), it does not influence the eigenvectors and only has a scalar effect on the singular values. Covariance matrix is a symmetric, nonnegative definite matrix. From the SVD decomposition formula, $\mathbf{M}\mathbf{M}^T$ is

given by

$$\begin{aligned}\mathbf{S} = \mathbf{M}\mathbf{M}^T &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Lambda}_{m\times m}\mathbf{U}^T,\end{aligned} \tag{2.82}$$

where $\mathbf{\Lambda}_{m\times m} = \mathbf{\Sigma}\mathbf{\Sigma}^T$. Remember that $\mathbf{\Sigma}$ is a m by n rectangular diagonal matrix. It is easy to see that $\mathbf{\Lambda}$ is a square diagonal matrix, and the element on the diagonal has the relationship $\lambda_i = \sigma_i^2$. Moreover, $\mathbf{U}$ is an orthogonal matrix. Examining equation (2.44), we can find that (2.82) is indeed the eigendecomposition of $\mathbf{S}$. In conclusion, the ith largest singular value σ_i of $\mathbf{M}$ is the square root of ith largest eigenvalue λ_i of $\mathbf{S}$. The ith left singular vector $\mathbf{u}_i$ of $\mathbf{M}$ is the eigenvector of $\mathbf{S}$ corresponding to the ith largest eigenvalue λ_i. Here, the principal components are given by (equation (2.69))

$$\mathbf{W} = \mathbf{U}^T\mathbf{M}, \tag{2.83}$$

or

$$\begin{pmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_m^T \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_m^T \end{pmatrix} \begin{pmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_n \end{pmatrix}, \tag{2.84}$$

where $\mathbf{w}_i^T = \mathbf{u}_i^T\mathbf{M}$ is the ith principal component, i.e. the trajectory matrix projected onto the ith eigenvector $\mathbf{u}_i$. Applying SVD decomposition, the principal components can be expressed as (Freire and Ulrych, 1988)

$$\mathbf{W} = \mathbf{U}^T\mathbf{M} = \mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Sigma}\mathbf{V}^T, \tag{2.85}$$

and

$$\mathbf{M} = \mathbf{U}\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{m}\mathbf{u}_i\mathbf{w}_i^T = \sum_{i=1}^{m}\sigma_i\mathbf{u}_i\mathbf{v}_i^T, \tag{2.86}$$

where the ith principal component is $\mathbf{w}_i = \sigma_i\mathbf{v}_i$, and $m \leq n$ is assumed. The covariance matrix of principal components is

$$\mathbf{W}\mathbf{W}^T = (\mathbf{\Sigma}\mathbf{V}^T)(\mathbf{\Sigma}\mathbf{V}^T)^T = \mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T = \mathbf{\Sigma}\mathbf{\Sigma}^T = \mathbf{\Lambda}. \tag{2.87}$$

This demonstrates that the principal components are uncorrelated.

Similarly, $\mathbf{T} = \mathbf{M}^T\mathbf{M}$ is referred to as *structure matrix* of the trajectory in SSA (Broomhead

and King, 1986a). Applying SVD decomposition, we can get the following relationship

$$\begin{aligned}\mathbf{T} = \mathbf{M}^T\mathbf{M} &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Lambda}_{n\times n}\mathbf{V}^T,\end{aligned} \tag{2.88}$$

where the matrix $\mathbf{\Lambda}_{n\times n} = \mathbf{\Sigma}^T\mathbf{\Sigma}$, with diagonal elements $\lambda_i = \sigma_i^2$. In conclusion, the ith right singular vector $\mathbf{v}_i$ of $\mathbf{M}$ is the eigenvector of $\mathbf{M}^T\mathbf{M}$ corresponding to the ith largest eigenvalue. The trajectory are confined in a subspace of embedding space, which is spanned by the singular vectors $\mathbf{u}_i, i = 1, 2, \ldots, r$ (Broomhead and King, 1986a). The dimension of the subspace is the rank of trajectory matrix. Moreover, $\text{rank}(\mathbf{M}) = \text{rank}(\mathbf{M}^T\mathbf{M}) = \text{rank}(\mathbf{M}\mathbf{M}^T)$.

2.3.4 Rank reduction and Eigenimage grouping

While, the presence of white noise will increase the rank of the matrix $\mathbf{X}$ to full rank. The observed time series consists of deterministic component and stochastic component.

$$d_i = \bar{d}_i + n_i, \quad i = 1, 2, \ldots, N, \tag{2.89}$$

where $\bar{d}_i$ is the underlying true signal, i.e. the deterministic part, and n_i is the zero mean *white noise* with standard deviation ϵ. The trajectory matrix has the form

$$\mathbf{M} = \bar{\mathbf{M}} + \mathbf{N}, \tag{2.90}$$

$$\mathbf{M}\mathbf{M}^T = \bar{\mathbf{M}}\bar{\mathbf{M}}^T + \bar{\mathbf{M}}\mathbf{N}^T + \mathbf{N}\bar{\mathbf{M}}^T + \mathbf{N}\mathbf{N}^T. \tag{2.91}$$

Because of the white noise assumption,

$$E\{\mathbf{M}\mathbf{M}^T\} = \bar{\mathbf{M}}\bar{\mathbf{M}}^T + \epsilon^2\mathbf{I}. \tag{2.92}$$

Thus, the additive noise component $\mathbf{N}$ increase all the eigenvalues of $E\{\mathbf{M}\mathbf{M}^T\}$ by ϵ^2, and does not change the eigenvectors. There is a group of small singular values and a gap between large and small singular values in the singular value spectrum. In this case, the matrix is called *rank deficient* (Hansen, 1998). Large singular values indicates the direction of highly correlation in trajectory matrix, which correspond to oscillatory components in the original times series. Small singular values account for random noise in the original time series. SVD provides a way to estimate the "nearest" low rank approximation to a matrix. The "nearest" is defined in the Euclidean metric sense. The problem of low rank

approximation with respect to Euclidean metric is summarized as

$$\begin{aligned} \mathbf{M}_k =& \mathcal{R}_k(\mathbf{M}) = \underset{\hat{\mathbf{M}}}{\operatorname{argmin}} \|\mathbf{M} - \hat{\mathbf{M}}\|_F^2, \\ & \text{subject to } \operatorname{rank}(\hat{\mathbf{M}}) = k, \end{aligned} \tag{2.93}$$

where $k < r$, $\|\cdot\|_F$ is the matrix Frobenius norm, $\|\mathbf{E}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|e_{ij}|^2}$ for matrix $\mathbf{E} \in \mathbb{R}^{m\times n}$. The optimal rank k approximation to $\mathbf{M}$ is given by the *truncated SVD(TSVD)* (Eckart and Young, 1936)

$$\begin{aligned} \mathbf{M}_k &= \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T = \sum_{i=1}^{k}\sigma_i\mathbf{u}_i\mathbf{v}_i^T \\ &= \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \ldots + \sigma_k\mathbf{u}_k\mathbf{v}_k^T. \end{aligned} \tag{2.94}$$

This is the well-known Eckart-Young theorem. From equation (2.86) and (2.94), rank reduction of $\mathbf{M}$ has the alternative formulation (Freire and Ulrych, 1988)

$$\mathbf{M}_k = \mathbf{U}_k\mathbf{U}_k^T\mathbf{M}, \tag{2.95}$$

where $\mathcal{R}_k := \mathbf{U}_k\mathbf{U}_k^T$ is referred to as a rank reduction opertator or a "rank filter". In conclusion, TSVD can reveal the underlying k-dimensional subspace, which contains the "useful" information, buried in the embedding space $\mathbb{R}^m$.

Note that it is not necessary to group the first k weighted eigenimages together. It is possible to group arbitrary several weighted eigenimages together if they represent meaningful information (Golyandina and Zhigljavsky, 2013).

2.3.5 Time series reconstruction

The rank-reduced matrix $\mathbf{M}_k$ is not a Hankel matrix any more, i.e. the rank-reduction operator does not preserve the Hankel structure. The elements on the anti-diagonals of $\mathbf{M}_k$ are averaged to estimate a reconstructed time series $\hat{\mathbf{d}} = (\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_{m+n-1})$. Remember that $\mathbf{M}_k$ is a m by n matrix. To make the discussion more convenient, we assume that $m < n$. The following operator conduct the anti-diagonal averaging operation (Golyandina and Zhigljavsky, 2013). If $m > n$, the matrix can be transposed first, and the following operation can still be applied on the transposed matrix. Here, $\mathbf{M}_k(i,j)$ represents the

element of $\mathbf{M}_k$ on ith row and jth column.

$$\hat{d}_s = \begin{cases} \dfrac{1}{s}\displaystyle\sum_{t=1}^{s} \mathbf{M}_k(t, s-t+1) & \text{for } 1 \le s < m, \\ \dfrac{1}{m}\displaystyle\sum_{t=1}^{m} \mathbf{M}_k(t, s-t+1) & \text{for } m \le s \le n, \\ \dfrac{1}{m+n-s}\displaystyle\sum_{t=s-n+1}^{m} \mathbf{M}_k(t, s-t+1) & \text{for } n < s \le m+n-1, \end{cases} \tag{2.96}$$

where s indicates the index of anti-diagonals and t indicates the index of elements on the anti-diagonal. It's obvious that the reconstructed times series $\hat{\mathbf{d}}$ depends on the choice of eigenimages.

2.4 SSA for time series analysis

In this section, the basic SSA algorithm is demonstrated on a well-known climatic time series: the Southern Oscillation Index (SOI). Rasmusson et al. (1990) used SSA on SOI data from 1950-1987 for studying the time scales of El Nino-Southern Oscillation (ENSO) variability. Ghil et al. (2002) applied SSA on SOI data from year 1940 to 2000 for signal-to-noise ratio enhancement, data compression and dynamical system interpretation. Here, I use a more complete SOI data from year 1876 to 2013 to illustrate the procedure of SSA for time series decomposition. The SOI data were downloaded from the Bureau of Meteorology of Australia (http://www.bom.gov.au/climate/current/soihtm1.shtml). It is a complete SOI data from January 1876 to July 2013. The "Troup SOI" is defined as 10 times the *standardized* and *centered* differences between the monthly means of the sea level pressures at Tahiti and Darwin, i.e. SOI $= 10\dfrac{\Delta P - \text{Mean}(\Delta P)}{\text{SD}(\Delta P)}$. In the given data, the mean (for centering) and standard deviation (for standardizing) were calculated from SOI data over the period 1933 to 1992, not the whole time period. For constructing the trajectory matrix, the window length is chosen to be 72. It represents the time interval of 72 months or 6 years. The singular spectrum of the trajectory matrix is shown in Figure 2.4. There are six large singular values and a group of smaller singular values in the singular spectrum. The six leading eigenvectors of covariance matrix $\mathbf{S}$ are plotted in Figure 2.5. We can find that the eigenvectors are behaved in pairs. Eigenvector 1 and 2 are similar in shape but have a shift in phases. They are called in quadrature (Rasmusson et al., 1990; Ghil et al., 2002). Eigenvector 3 and 4 are another eigenvector pair and have similar behaviour. The eigenvectors of covariance matrix are the transformation bases of SSA. They are derived from data itself and not necessarily harmonic functions (Ghil et al., 2002). This property gives SSA a better chance over other transformations to represent anharmonic oscillations.

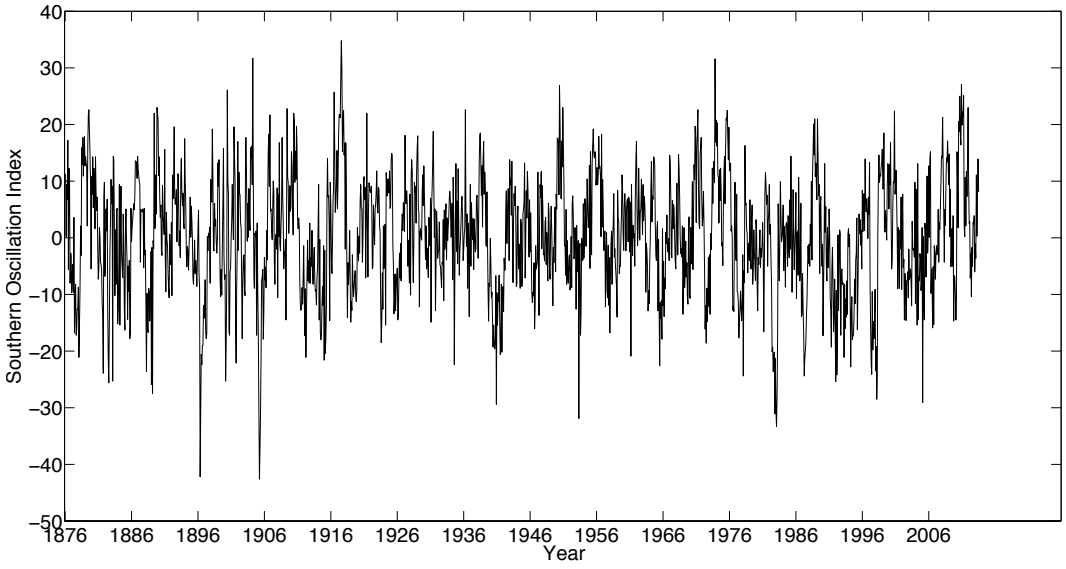

Figure 2.3: Southern Oscillation Index (SOI) from January 1876 to July 2013.

That is to say, some anharmonic oscillations may need many Fourier bases (harmonics) to be represented but only a few eigenvectors in SSA. Figure 2.6 shows the principal components 1 to 6 of the trajectory matrix. They are were computed by projecting the trajectory matrix onto eigenvectors 1 to 6, respectively. That is to say, the ith principal component is given by $\mathbf{w}_i^T = \mathbf{u}_i^T\mathbf{M}$. $\mathbf{u}_i^T$ is the ith eigenvector of the covariance matrix $\mathbf{S}$ and $\mathbf{M}$ is the trajectory matrix. Principal components 1 and 2 are similar in shape, they have a period of about 4 years. Principal components 3 and 4 are similar in shape, their period are about 2 years. Figure 2.7 shows the time series reconstructed from eigenimages 1 to 6, respectively. That is to say, the ith reconstructed time series is given by Equation (2.96) with eigenimage $\mathbf{M}_k = \mathbf{M}_i = \sigma_i\mathbf{u}_i\mathbf{v}_i^T$. Reconstructed time series 5 displays a nonlinear trend and an oscillatory component. The oscillation patterns in the reconstructed time series 6 (or principal component 6) is dominated by oscillatory noise. The signal-to-noise ratio enhanced SOI time series are reconstructed by adding the first 4 reconstructed time series (Figure 2.8). We can find that the reconstructed time series captures most of the oscillation behaviour of the time series with a reduction of noise.

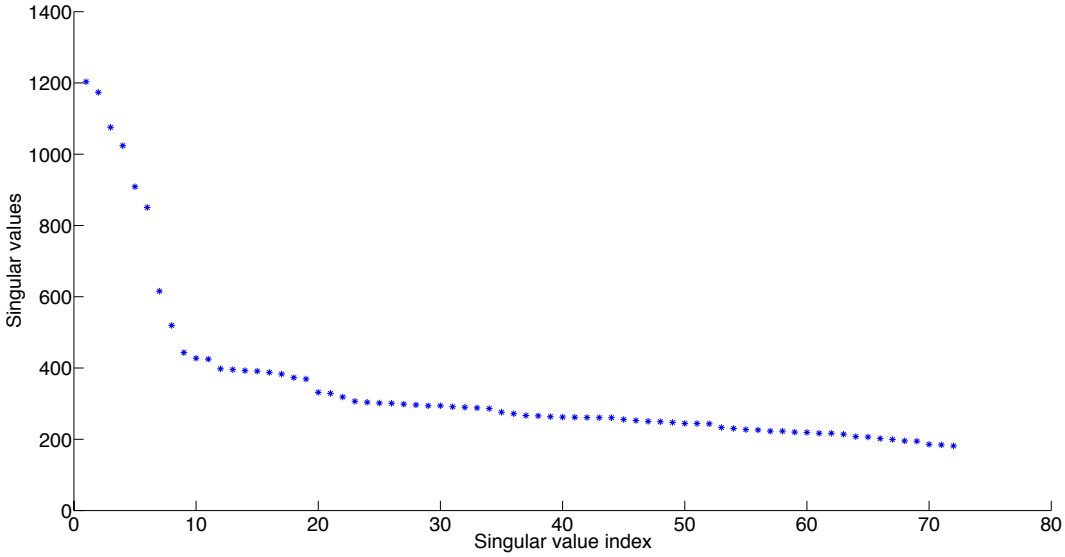

Figure 2.4: The singular spectrum of trajectory matrix of Southern Oscillation Index. There are six leading singular values and remaining smaller singular values.

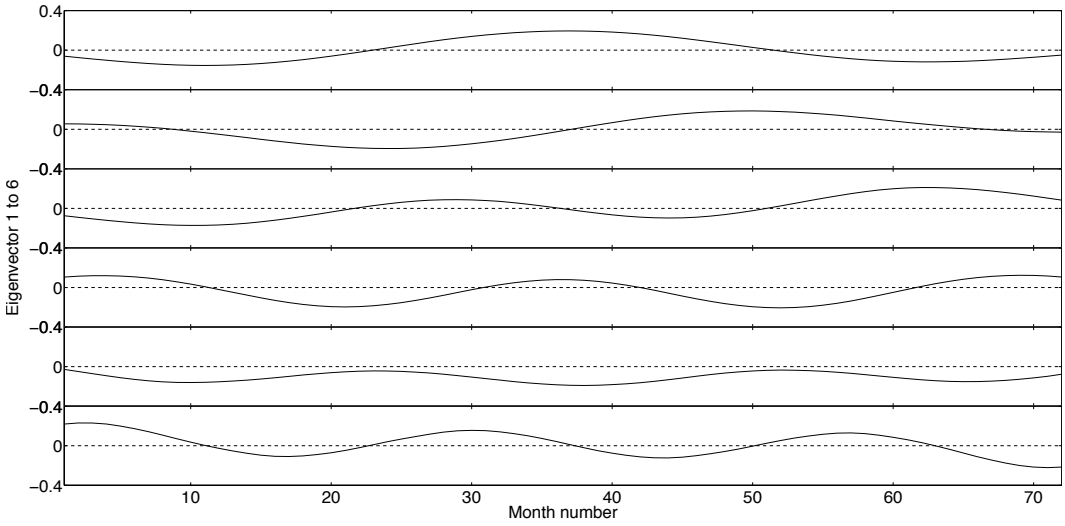

Figure 2.5: Six leading eigenvectors of the covariance matrix. From top to bottom, the eigenvalues corresponding the eigenvectors decrease. These transformation bases are derived from the data itself.

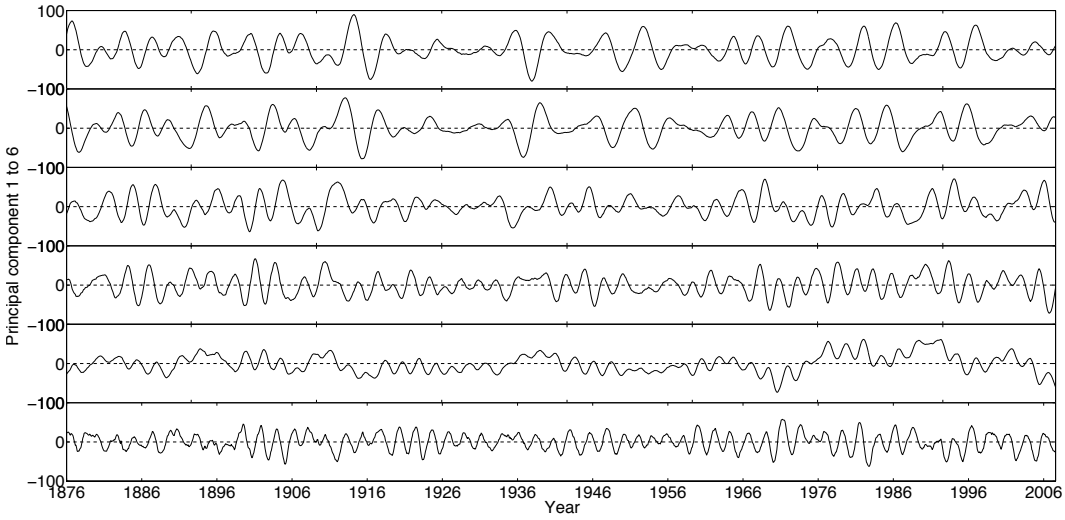

Figure 2.6: Six principal components. They are the projection of trajectory matrix onto eigenvectors 1 - 6, respectively.

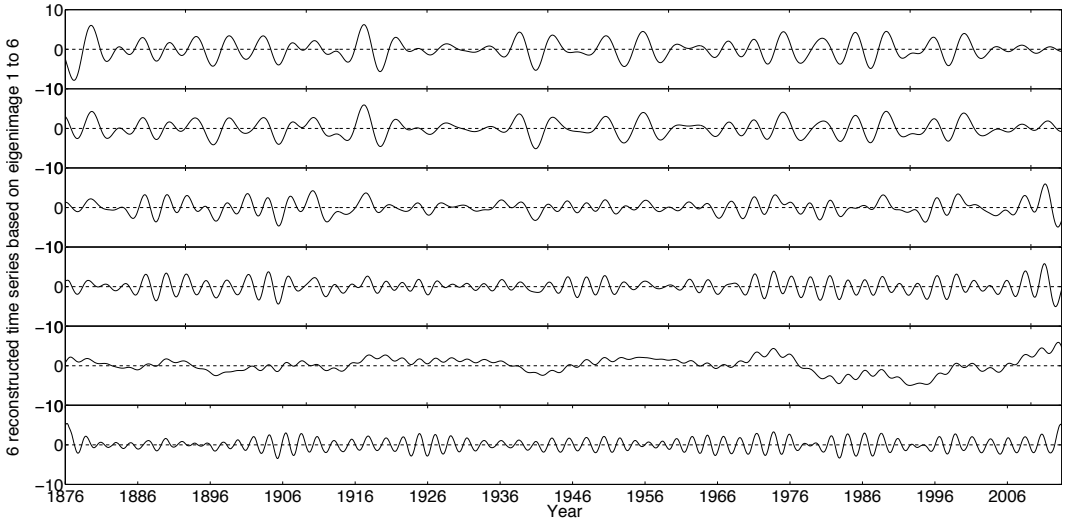

Figure 2.7: Six reconstructed time series by eigenimages 1-6, respectively.

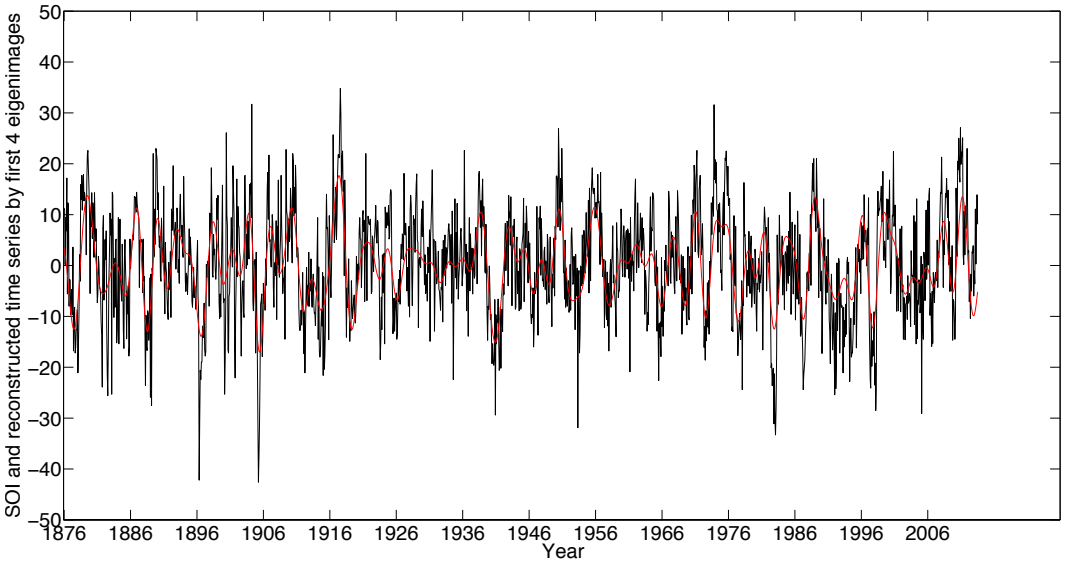

Figure 2.8: Original SOI time series (black line) and reconstructed time series by first 4 eigenimages (red line).

2.5 Applications of SSA in seismic data processing

In seismic data processing, SSA is applied in *frequency-space* domain (Sacchi, 2009). Each frequency slice of seismic data is a complex valued "spatial series", similar to time series. SSA algorithm can be extended to complex-valued time series by replacing the transpose operator T by complex conjugate operator H (Golyandina and Zhigljavsky, 2013).

2.5.1 Signal model in Fourier domain

Consider a noise-free seismic data section consists of one single linear event, the data can be expressed as

$$d(t,x) = a(t - px), \tag{2.97}$$

where $a(t)$ is a source wavelet, d is the propagated waveform, t is time, x is spatial position and ray parameter $p = 1/V = \Delta t/\Delta x$ is the slope or dip of the linear event. The Fourier transform changs the time delay to phase shift

$$D(\omega,x) = A(\omega)e^{-i\omega px}, \tag{2.98}$$

where ω is frequency, complex coefficient $A(\omega)$ is the Fourier transform of the wavelet $a(t)$. Assume that the spatial interval between traces is regular, $x = (j-1)\Delta x, j = 1, 2, \ldots, N$ is the trace index in the spatial axis and Δx is the spatial interval between two adjacent traces. Here, we use $j-1$ to indicate that the first trace has zero offset. The Fourier coefficient at frequency ω of jth trace is

$$D(\omega, (j-1)\Delta x) = A(\omega)e^{-i\omega p(j-1)\Delta x}. \tag{2.99}$$

To be more convenient, we note $D_j(\omega) = D(\omega, (j-1)\Delta x)$

$$D_j(\omega) = A(\omega)e^{-i\omega p(j-1)\Delta x}, \tag{2.100}$$

The fourier coefficient at frequency ω at trace $j-1$ is

$$D_{j-1}(\omega) = A(\omega)e^{-i\omega p(j-2)\Delta x} = A(\omega)e^{-i\omega p(j-1)\Delta x}e^{i\omega p\Delta x} = e^{i\omega p\Delta x}D_j(\omega), \tag{2.101}$$

In other words, there is a linear recursion relationship between adjacent traces

$$D_j(\omega) = PD_{j-1}(\omega), \tag{2.102}$$

where $P = e^{-i\omega p\Delta x}$. The linear predictable property of one single linear event in f-x domain is shown in Figure 2.9. The frequency slice of Fourier transform of seismic section consists of one single linear event is a complex harmonic. The real part of the frequency slice is a sinusoid (Figure 2.9(c)).

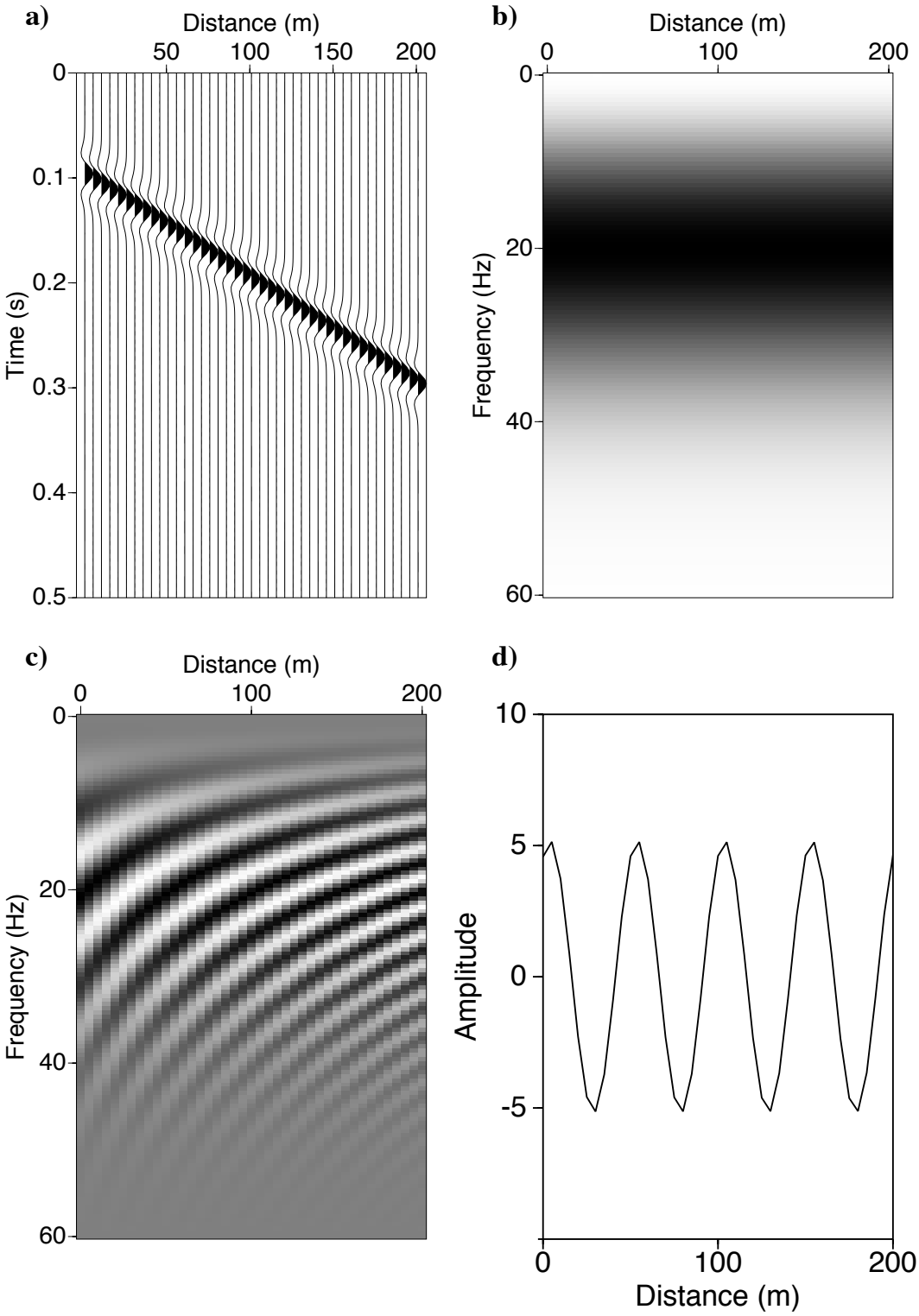

Figure 2.9: The predictable property of linear events in f-x domain. a) A seismic section consists of one single linear event in t-x domain. b) Amplitude spectra of the t-x data from 0 Hz to 60 Hz. c) The real part of the data in f-x domain from 0 Hz to 60 Hz. d) The real part of the complex Fourier coefficient at 20 Hz.

Now, we consider more general case. If a noise-free seismic section consists of K linear events with Kdistinct ray parameters p_k, it can be represented in the frequency-space domain via the superposition of plane waves

$$D_j(\omega) = \sum_{k=1}^{K} A_k(\omega) e^{-i\omega p_k (j-1)\Delta x}, \tag{2.103}$$

where complex coefficient $A_k(\omega)$ is the Fourier transform of the wavelet of event k. The number of events should be smaller than the number of traces. Arrange the expression of superposition of plane waves at N traces together

$$\begin{pmatrix} D_1(\omega) \\ D_2(\omega) \\ \vdots \\ D_N(\omega) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{-i\omega p_1 \Delta x} & e^{-i\omega p_2 \Delta x} & \cdots & e^{-i\omega p_K \Delta x} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-i\omega p_1 (N-1)\Delta x} & e^{-i\omega p_2 (N-1)\Delta x} & \cdots & e^{-i\omega p_K (N-1)\Delta x} \end{pmatrix} \begin{pmatrix} A_1(\omega) \\ A_2(\omega) \\ \vdots \\ A_K(\omega) \end{pmatrix}, \tag{2.104}$$

or note as

$$\mathbf{d}(\omega) = \mathbf{S}(\omega)\mathbf{a}(\omega), \tag{2.105}$$

where $\mathbf{S}(\omega)$ is a Vandermonde matrix. Because all the events have distinct ray parameters p_i and we assume that there is no *aliasing*, the rank of $\mathbf{S}(\omega)$ is K. Any row is a linear combination of K other rows, i.e. any complex coefficient at one trace can be predicted from coefficients in K other traces. For example, the jth row of $S(\omega)$ can be written as the linear combination of previous K rows.

$$\begin{aligned} &\left(e^{-i\omega p_1 (j-1)\Delta x}, e^{-i\omega p_2 (j-1)\Delta x}, \cdots, e^{-i\omega p_K (j-1)\Delta x}\right), \\ &= (P_K(\omega), P_{K-1}(\omega), \ldots, P_1(\omega)) \begin{pmatrix} e^{-i\omega p_1 (j-K-1)\Delta x} & e^{-i\omega p_2 (j-K-1)\Delta x} & \cdots & e^{-i\omega p_K (j-K-1)\Delta x} \\ e^{-i\omega p_1 (j-K)\Delta x} & e^{-i\omega p_2 (j-K)\Delta x} & \cdots & e^{-i\omega p_K (j-K)\Delta x} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-i\omega p_1 (j-2)\Delta x} & e^{-i\omega p_2 (j-2)\Delta x} & \cdots & e^{-i\omega p_K (j-2)\Delta x} \end{pmatrix} \end{aligned} \tag{2.106}$$

Multiply the vector $\mathbf{a}(\omega)$ to two sides of above equation, we can get the linear recursion relationship

$$D_j(\omega) = (P_K(\omega), P_{K-1}(\omega), \ldots, P_1(\omega)) \begin{pmatrix} D_{j-K}(\omega) \\ D_{j-K+1}(\omega) \\ \vdots \\ D_{j-1}(\omega) \end{pmatrix}, \tag{2.107}$$

or

$$D_j(\omega) = \sum_{k=1}^{K} P_k(\omega) D_{j-k}(\omega) = P_1(\omega) D_{j-1}(\omega) + P_2(\omega) D_{j-2}(\omega) + \ldots + P_K(\omega) D_{j-K}(\omega). \tag{2.108}$$

This linear recursion relationship is the basis of f-x SSA, f-x prediction and f-x projection seismic data processing methods. f-x prediction and f-x projection methods derive the prediction filter $(P_K(\omega), P_{K-1}(\omega), \ldots, P_1(\omega))$ from the data. f-x SSA embeds the frequency slice into Hankel matrix and derive the subspace which contains the signal.

2.5.2 Embedding

The SSA method constructs a trajectory matrix by embedding spatial data at one frequency, i.e. $\mathbf{D}(\omega) = [D_1(\omega), D_2(\omega), \cdots, D_N(\omega)]^T$ into the following Hankel matrix

$$\mathbf{M}(\omega) = \mathcal{H}[\mathbf{D}(\omega)] = \begin{pmatrix} D_1(\omega) & D_2(\omega) & \cdots & D_{N-L+1}(\omega) \\ D_2(\omega) & D_3(\omega) & \cdots & D_{N-L+2}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ D_L(\omega) & D_{L+1}(\omega) & \cdots & D_N(\omega) \end{pmatrix},$$

where the symbol $\mathcal{H}$ is Hankel operator. I will use $\mathbf{M}$ to represent trajectory matrix in SSA for seismic data processing. In SSA for time series analysis, the window length L will influence the singular spectrum and singular vectors. It is relatively important to choose a suitable window length in SSA for time series analysis. While, the results of SSA for noise attenuation are not very sensitive to the selection of window length. For convenience, we choose $L = \lfloor \frac{N}{2} \rfloor + 1$ to make the Hankel matrix approximately square (Trickett, 2008). $\mathbf{M}(\omega) \in \mathbb{C}^{L \times (N-L+1)}$ is a complex matrix and we will omit the symbol ω and understand that the analysis is carried out for all frequencies. We also use $L = m$ and $N - L + 1 = n$ in the following discussions. Because of the linear recursive relationship for exponentials (2.108), we can prove that the rank of the matrix $\mathbf{M}$ equal to the number of linear events

in the seismic section.

$$\begin{aligned}\mathbf{M} &= \begin{pmatrix} D_1 & \cdots & D_K & D_{K+1} & \cdots & D_{N-L+1} \\ D_2 & \cdots & D_{K+1} & D_{K+2} & \cdots & D_{N-L+2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ D_L & \cdots & D_{K+L-1} & D_{K+L} & \cdots & D_N \end{pmatrix} \\ &= \begin{pmatrix} D_1 & \cdots & D_K & \sum_{k=1}^{K} P_k D_{K+1-k} & \cdots & \sum_{k=1}^{K} P_k D_{N-L+1-k} \\ D_2 & \cdots & D_{K+1} & \sum_{k=1}^{K} P_k D_{K+2-k} & \cdots & \sum_{k=1}^{K} P_k D_{N-L+2-k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ D_L & \cdots & D_{K+L-1} & \sum_{k=1}^{K} P_k D_{K+L-k} & \cdots & \sum_{k=1}^{K} P_k D_{N-k} \end{pmatrix}\end{aligned} \tag{2.109}$$

It means that each column can be written as the linear combination of previous K columns. Therefore, $\text{rank}(\mathbf{M}) = K$. The presence of random noise will increase the rank of $\mathbf{M}$ because they are not linearly predictable and there is no exact linear relationship between the columns.

2.5.3 Decomposition

Singular value decomposition can extract the subspace that contains signal from the whole space. The SVD of a complex matrix is very similar to the SVD of a real matrix except that we replace the transpose operator with the complex conjugate operator.

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{2.110}$$

where $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_p\} \in \mathbb{R}^{m\times n}(p = \min\{m, n\})$ is the matrix containing singular values of $\mathbf{M}$ on its diagonal. $\mathbf{U} \in \mathbb{C}^{m\times m}$ is an unitary matrix such that $\mathbf{U}^H\mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathbf{I}_m$. $\mathbf{V} \in \mathbb{C}^{n\times n}$ is an unitary matrix such that $\mathbf{V}^H\mathbf{V} = \mathbf{V}\mathbf{V}^H = \mathbf{I}_n$.

2.5.4 Rank reduction

If the seismic section is composed of K linear events and small amplitude random noise, the singular spectrum of $\mathbf{M}$ will have K relative large singular values and $p - K$ relative small singular values. Similar with rank reduction in real matrix case, the rank K approximation of the complex matrix $\mathbf{M}$ can be found by solving the following problem

$$\begin{aligned}&\mathbf{M}_K = \mathcal{R}_K[\mathbf{M}] = \underset{\hat{\mathbf{M}}}{\text{argmin}} \|\mathbf{M} - \hat{\mathbf{M}}\|_F^2, \\ &\text{subject to } \text{rank}(\hat{\mathbf{M}}) = K,\end{aligned} \tag{2.111}$$

where $\|\cdot\|_F$ is the Frobenius norm of complex matrix, $\|\mathbf{E}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|e_{ij}|^2}$ of the matrix $\mathbf{E} \in \mathbb{C}^{m\times n}$. The rank reduction problem has an unique analytic solution, the truncated singular value decomposition (TSVD)

$$\begin{aligned}\mathbf{M}_K =& \mathcal{R}_K[\mathbf{M}] = \mathbf{U}_K\mathbf{\Sigma}_K\mathbf{V}_K^H \\ =& \mathbf{U}_K\mathbf{U}_K^H\mathbf{M},\end{aligned} \tag{2.112}$$

where $\mathbf{U}_K \in \mathbb{C}^{m\times K}$ and $\mathbf{V}_K \in \mathbb{C}^{n\times K}$ are matrices containing singular vectors associated to the first K-largest singular values $\sigma_j, j = 1\ldots K$ which are also the diagonal elements of the matrix $\mathbf{\Sigma}_K \in \mathbb{R}^{K\times K}$.

2.5.5 Anti-diagonal averaging

The rank-reduced matrix $\mathbf{M}_K$ does not have Hankel structure, an anti-diagonal averaging procedure is needed to recover the filtered frequency slice. It is done with the same anti-diagonal averaging equation (2.96) in SSA for time series analysis. The filtered frequency slice at frequency ω is $\hat{\mathbf{D}}(\omega) = \mathcal{A}[\mathbf{M}_K(\omega)]$, where $\mathcal{A}$ is the anti-diagonal averaging operator. For each frequency, the f-x SSA filters the complex data by $\hat{\mathbf{D}}(\omega) = \mathcal{A}[\mathcal{R}_K[\mathcal{H}[\mathbf{D}(\omega)]]]$.

2.5.6 Inverse Fourier transform

After the filtering, the complex coefficient in *frequency-space* domain are transformed back to *time-space* domain via applying inverse Fourier transform to each channel of the data.

2.5.7 Examples

Figure 2.10 demonstrates the relationship between the number of distinct dips and the rank of the trajectory matrix if the seismic data is noise-free. Figure 2.10 (a) is a noiseless seismic section consists of three linear events. The analyzing frequency band ranges from 1 to 40 Hz. Figure 2.10 (b) shows the singular spectrum for each frequency. We can find that each of the singular spectrum only has three nonzero singular values. I would like to say that the signal is *sparse* in the "singular spectrum domain". Figure 2.10 (c) is the real part of the frequency slice at 10 Hz, which is the superposition of three monochromatic sinusoids. Each frequency slice is a superposition of three complex exponentials. Figure 2.10 (d) is the singular spectrum of the trajectory matrix constructed from the frequency slice at 10 Hz, which has three nonzero singular values. In other words, the rank of the trajectory matrix is 3. But note that, the number of nonzero singular values corresponding to the real part of frequency slice is 6 because that each sinusoid is the superposition of two complex

exponentials. Each complex exponential can be represented by one nonzero singular value in singular spectrum of trajectory matrix. Figure 2.11 (a) is the result of applying SSA filtering on noise-free seismic section 2.10 (a). The rank for rank-reduction in SSA is chosen as 3. Figure 2.11 (b) is the data after f-x deconvolution. The length of prediction filter and regularization parameter in f-x deconvolution are 10 and 0.001, respectively. Both the two methods recovered the signal. Figure 2.11 (c) and (d) are the error panels of SSA and f-x deconvolution, respectively. There is almost no energy in the error panels, both the two methods do not damage the original signal.

As discussed in previous section, the presence of random noise will increase the rank of the trajectory matrix. This is demonstrated in Figure 2.12. Figure 2.12 (a) is a seismic section consists of three linear events that is same with the events in Figure 2.10 (a). However, the seismic section here is also corrupted with random Gaussian noise with signal-to-noise ratio (SNR) equals to 1. The SNR is defined as the maximum amplitude of the signal divided by the maximum amplitude of the noise. The analyzing frequency band ranges from 1 to 40 Hz. Figure 2.12 (b) shows the singular spectrum for each frequency slice. For each singular spectrum, we can find that the random Gaussian noise presents as small singular values in the whole spectrum. Figure 2.12 (c) is the real part of the frequency slice at 10 Hz of data in 2.12 (a), which is a superposition of sinusoids but corrupted with random noise. Figure 2.12 (d) is the singular spectrum of the trajectory matrix constructed from frequency slice at 10 Hz. We can find that there are 3 largest singular values and 17 small singular values in the singular spectrum. Note that, if the noise is strictly white, i.e. uncorrelated from trace to trace, it will enlarge all the 20 singular values as discussed in equation (2.92). However, we find that the first three singular values in Figure 2.12 (d) is smaller that the first three singular values in Figure 2.10 (d). This is caused by the fact that the random noise we added to the seismic section is smoothed (Hanning window) white Gaussian noise. It is actually not strictly uncorrelated. Figure 2.13 (a) is the data after SSA filtering. The rank for reconstruction in SSA is chosen to be 3. Figure 2.13 (b) is the data after f-x deconvolution filtering. The length of prediction filter and regularization parameter in f-x deconvolution are 10 and 0.001, respectively. SSA and f-x have relatively similar capability for random noise reduction. However, SSA has the advantage of preserving the original signal (Figure 2.13(c)). While, f-x deconvolution tends to damage the original signal (Figure 2.13(d)). It is caused by the fact it uses AR model to approximate the ARMA model.

In this example, we show that the performances of SSA and f-x deconvolution degenerate drastically when the seismic data is corrupted with erratic noise. We added a sinusoid noise with frequency 10 Hz to the seismic data to simulate the erratic noise (Figure 2.14(a)). The erratic noise presents as a large spike in the f-x domain (Figure 2.14(c)). Therefore, the trajectory matrix also contains elements with very large values, i.e. outliers (Chapter 3). The least-squares estimation TSVD breaks down with the presence of outliers. We

can find that there are several very large singular values in the singular spectrum, which is caused by the outliers (Figure 2.14(d)). Figure 2.15(a) is the result after SSA filtering (rank = 3). We can see that the high-amplitude erratic noise spreads out in the adjacent traces (Figure 2.15(a)). The f-x deconvolution also can't completely remove the erratic noise (Figure 2.15(b)), and it damages original signal (Figure 2.15(d)). The prediction filter length is 10 and regularization parameter is 0.001 in f-x deconvolution. Both the SSA and f-x deconvolution belong to least-squares estimations, which are very sensitive to outliers (non-Gaussian). Robust estimation procedure is needed to cope with outliers. Figure 2.16 (a) shows the result of robust SSA applied on seismic section corrupted with Gaussian noise and erratic noise. Both the erratic noise and Gaussian noise are effectively suppressed. Moreover, the robust SSA algorithm preserves the original signal (Figure 2.16(b)). More details about robust statistics and the robust SSA algorithm are given in next chapter.

2.6 Summary

In this Chapter, we reviewed basic concepts of multivariate random variable and principal component analysis. We also reviewed the method of singular spectrum analysis. The observed time series is centered and normalized. Then it is embedded into a Hankel matrix. The singular value decomposition is used to decompose the Hankel matrix into different weighted eigenimages. A new reconstructed low rank matrix can be estimated by combining several particular weighed eigenimages. The reconstructed time series is computed by averaging the elements of the anti-diagonals of the reconstructed matrix. SSA allows decomposition, analysis and SNR enhancement of the original time series. The example of SOI time series is used to show the ability of SSA for time series analysis. The application of SSA for seismic data random noise attenuation is also reviewed in this chapter. SSA is applied on each frequency slice of seismic data in the f-x domain. The SSA was tested on a synthetic seismic section with random noise. The result is compared with f-x deconvolution. SSA has the advantage of preserving the amplitude of the desired signal.

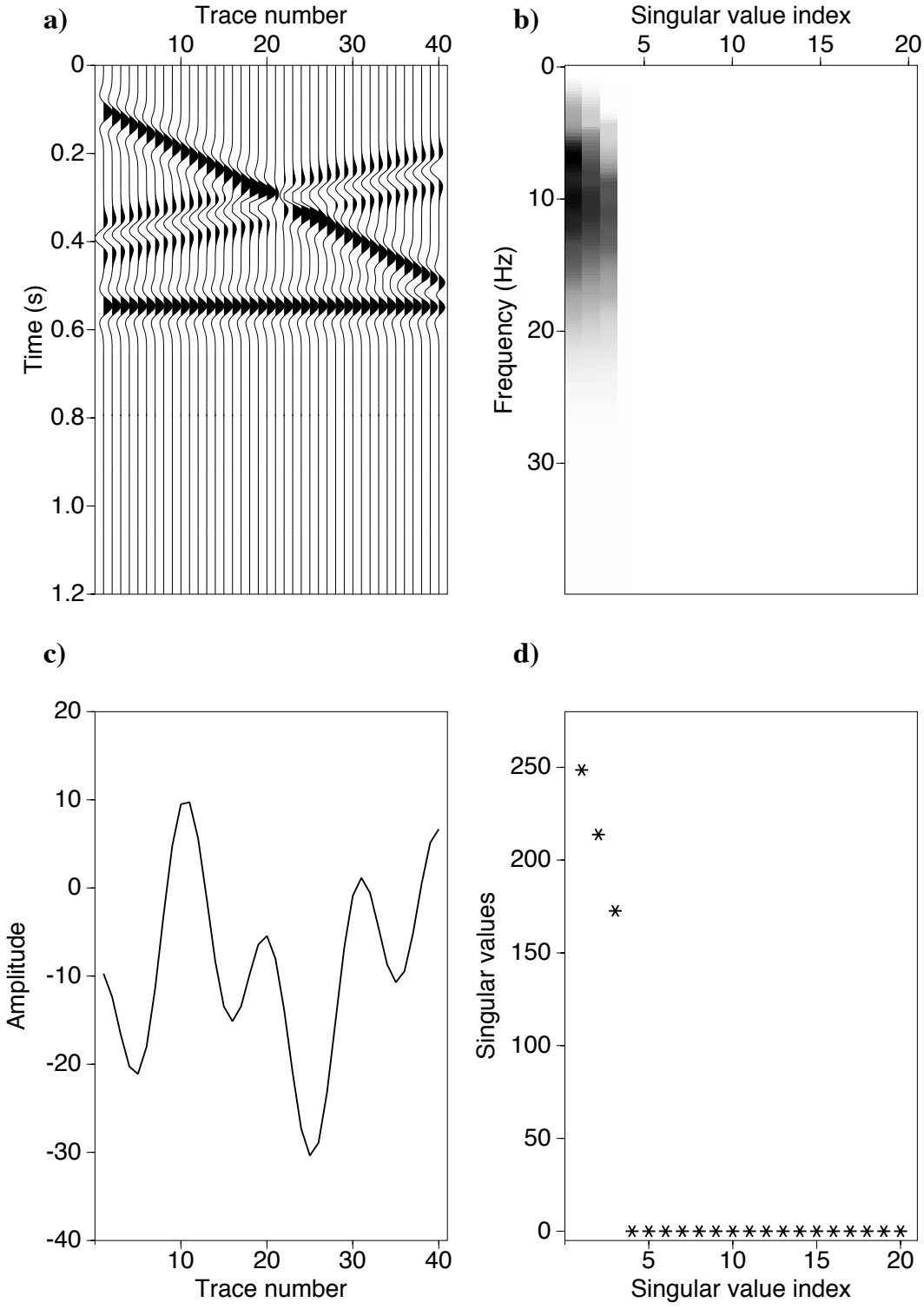

Figure 2.10: a) Noise-free seismic data section consists of three linear events. b) The singular spectra of the trajectory matrices constructed from different frequency slices. c) The real part of the frequency slice at 10 Hz. d) The singular spectrum of the trajectory matrix constructed from frequency slice at 10 Hz.

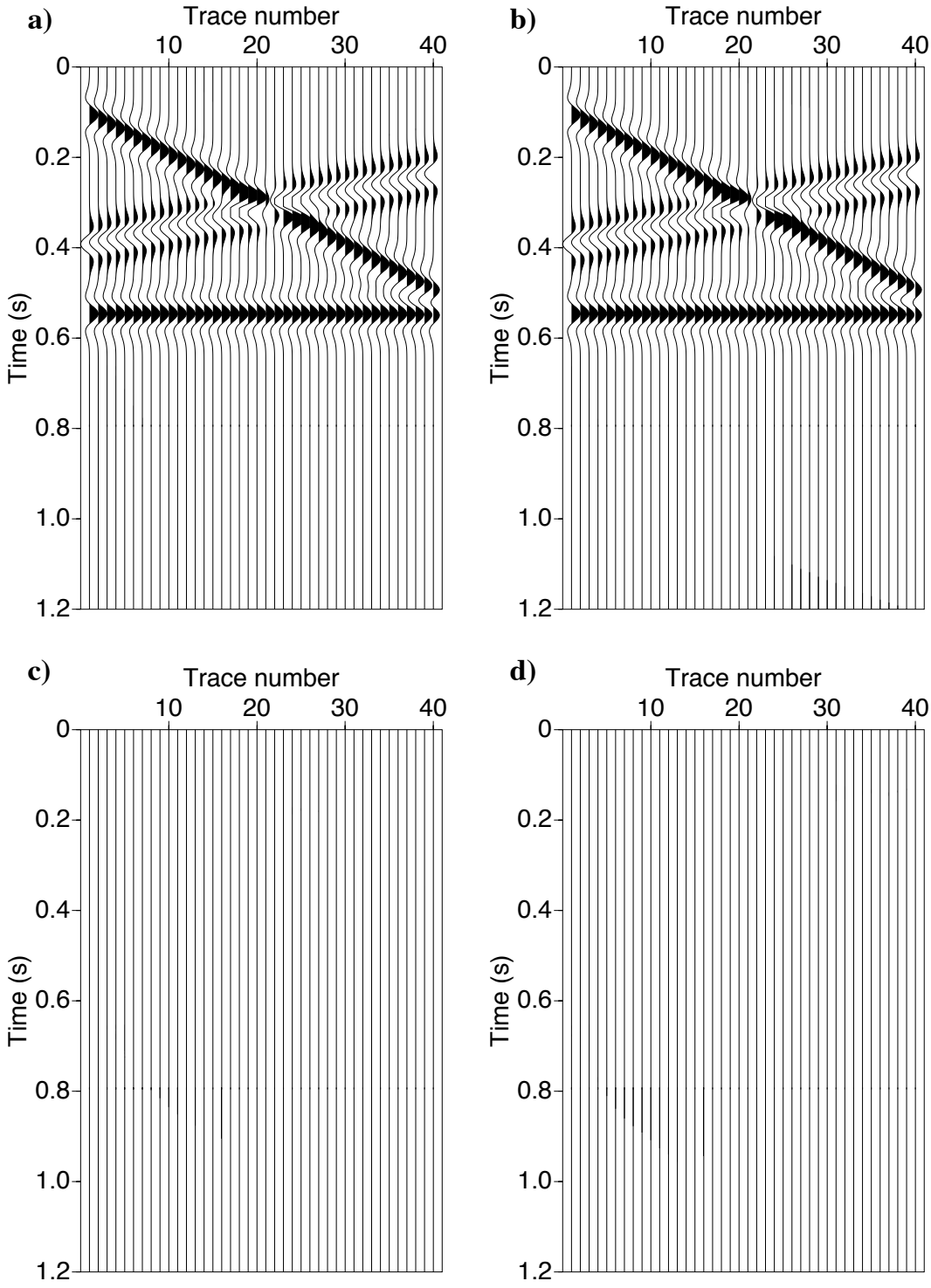

Figure 2.11: a) Noise-free data after SSA filtering. b) Noise-free data after f-x deconvolution filtering. c) Difference between noise-free data and SSA filtered data. d) Difference between noise-free data and f-x deconvolution filtered data.

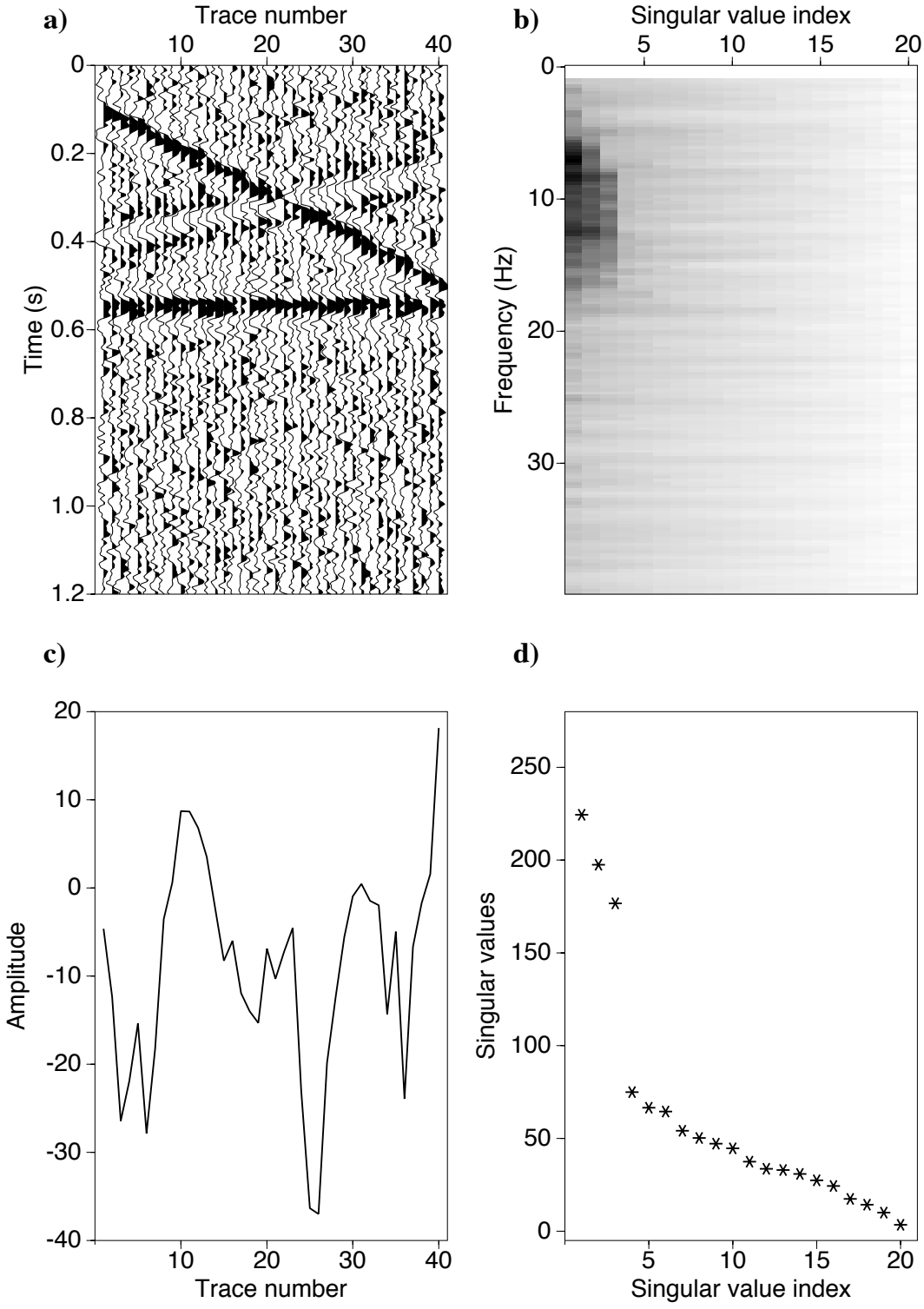

Figure 2.12: a) Seismic data section consists of three linear events, corrupted with Gaussian noise (SNR=1). b) The singular spectra of the trajectory matrices constructed from different frequency slices. c) The real part of the frequency slice at 10 Hz. d) The singular spectrum of the trajectory matrix constructed from frequency slice at 10 Hz.

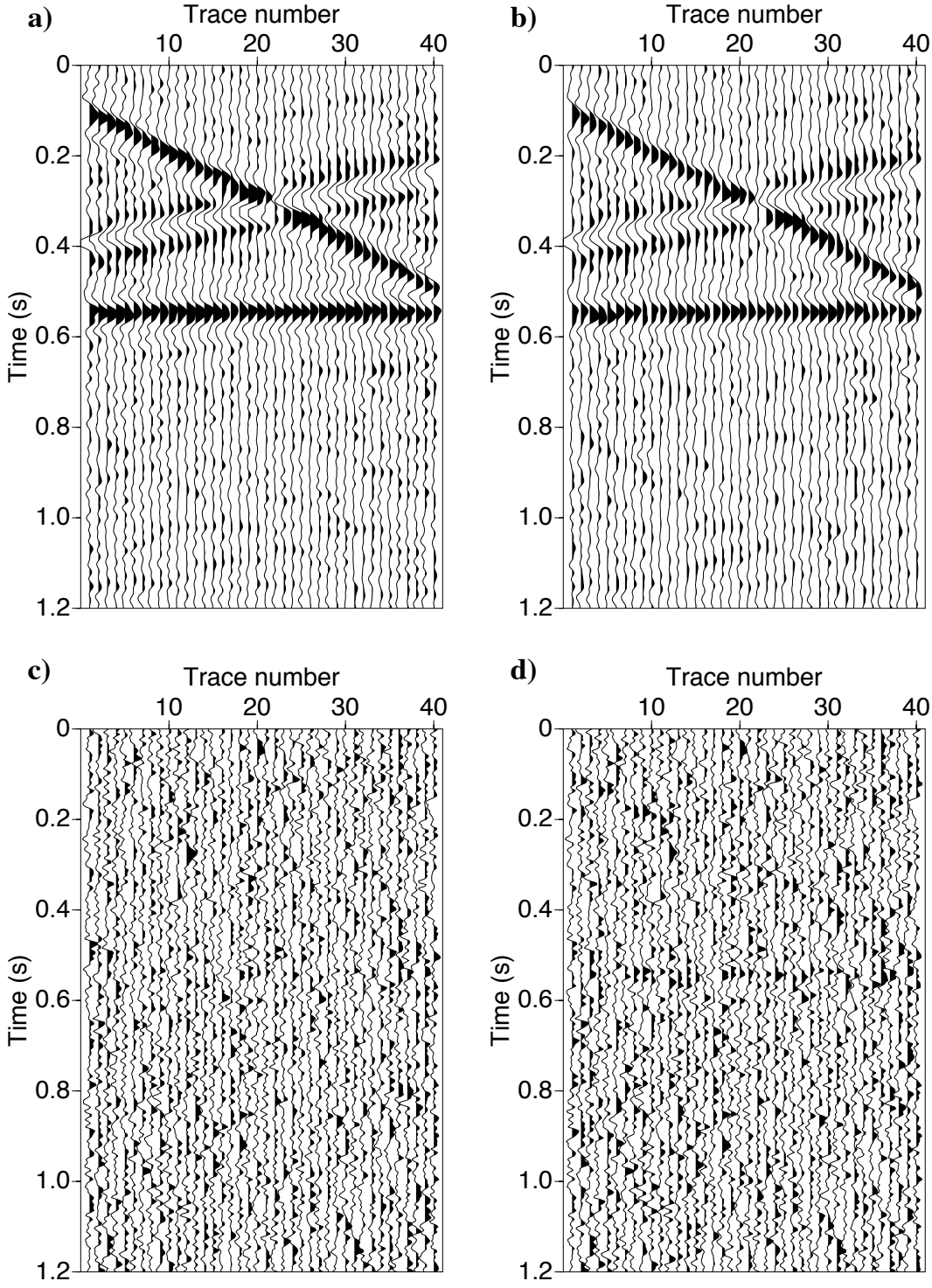

Figure 2.13: a) Data corrupted with Gaussian noise after SSA filtering. b) Data corrupted with Gaussian noise after f-x deconvolution filtering. c) Difference between noisy data and SSA filtered data. d) Difference between noisy data and f-x deconvolution filtered data.

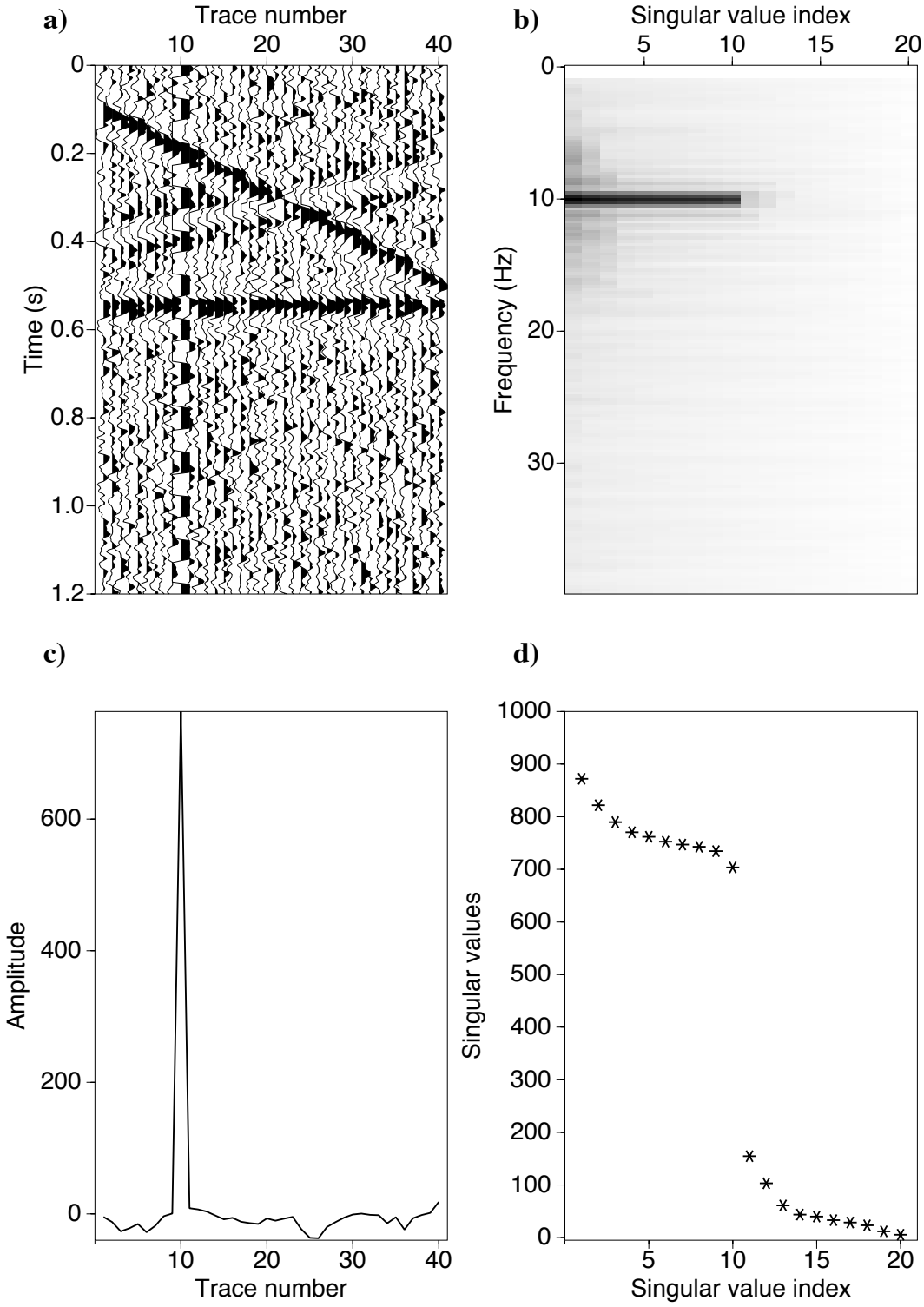

Figure 2.14: a) Seismic data section consists of three linear events, corrupted with Gaussian noise (SNR=1) and erratic noise. b) The singular spectra of the trajectory matrices constructed from different frequency slices. c) The real part of the frequency slice at 10 Hz. d) The singular spectrum of the trajectory matrix constructed from frequency slice at 10 Hz.

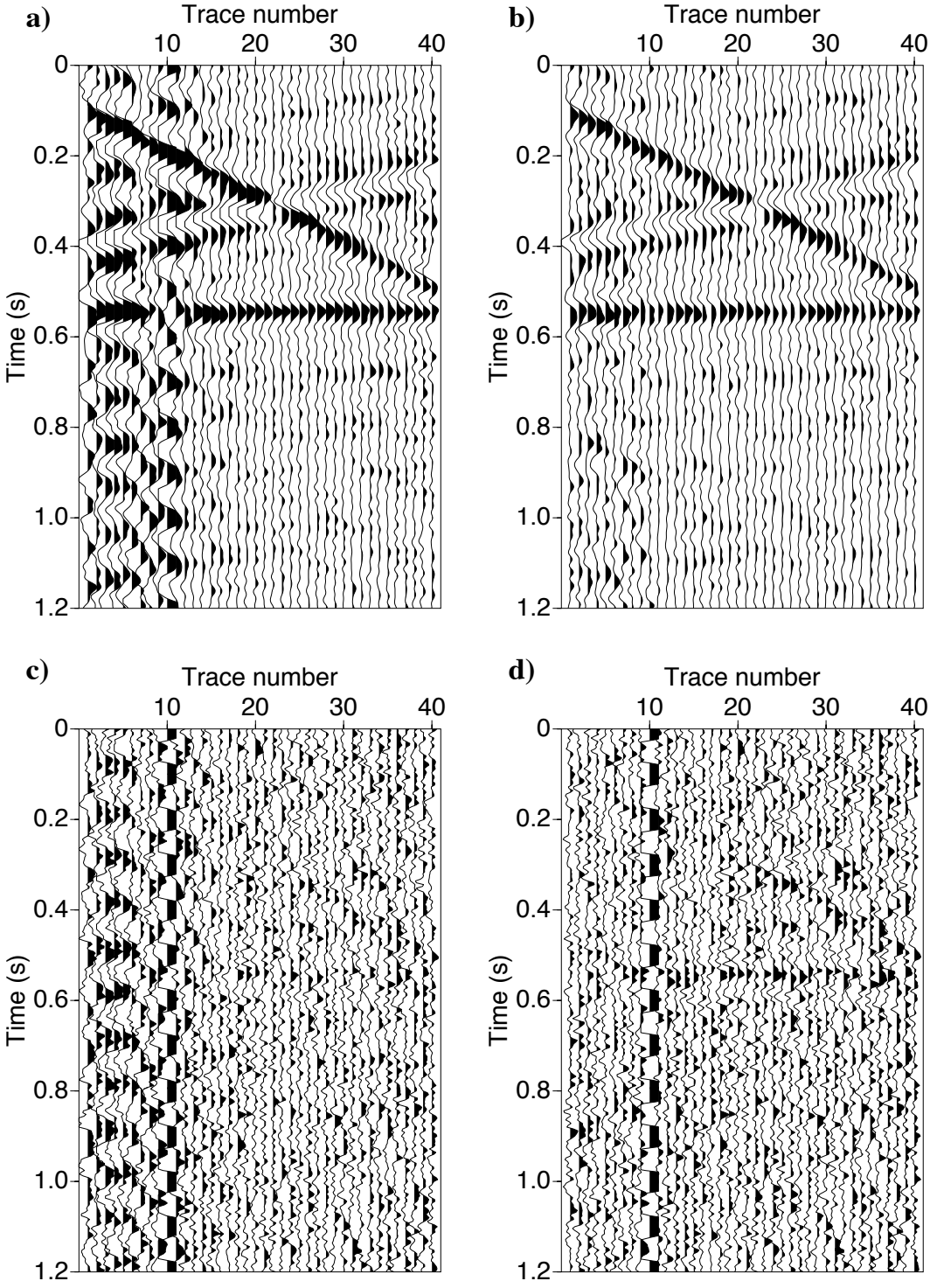

Figure 2.15: a) Data corrupted with Gaussian noise and erratic noise after SSA filtering. b) Data corrupted with Gaussian noise and erratic noise after f-x deconvolution filtering. c) Difference between noisy data and SSA filtered data. d) Difference between noisy data and f-x deconvolution filtered data.

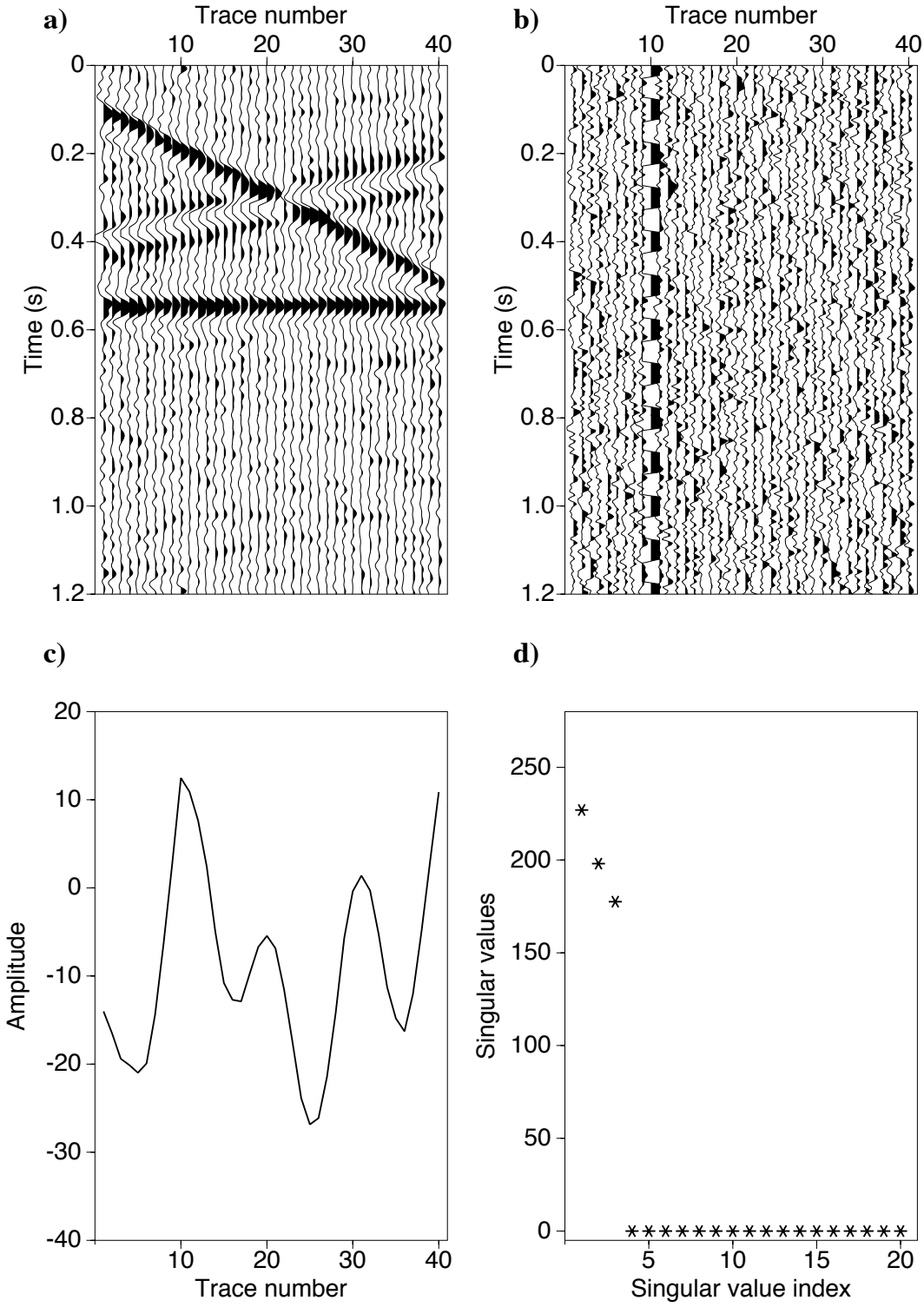

Figure 2.16: a) Data corrupted with Gaussian noise and erratic noise after robust SSA filtering. b) Difference between noisy data and robust SSA filtered data. c) Real part of the frequency slice at 10 Hz of the robust SSA filtered data. d) Singular spectrum of the frequency slice at 10 Hz of the robust SSA filtered data.

CHAPTER 3

Robust singular spectrum analysis for erratic noise attenuation

3.1 Introduction

The matrix rank-reduction techniques used in previous SSA algorithms, e.g. truncated SVD (Sacchi, 2009; Golub and Van Loan, 1996), rank-reduction based on randomized SVD (Oropeza and Sacchi, 2011; Halko et al., 2011), rank-reduction based on Lanczos bidiagonalization (Gao et al., 2013; Golub and Van Loan, 1996; Simon and Zha, 2000), all adopt the quadratic error criterion (Equation (2.93)). Least-squares estimation is optimal when the observed data (or say the noise in the data) follow *Gaussian distribution*. It is suboptimal when the noise is non-Gaussian. This chapter proposes a new robust SSA algorithm based on M-estimate for simultaneous Gaussian and erratic (non-Gaussian) seismic noise attenuation. In the field of statistics, the Gaussian distribution assumption has been used for almost two centuries. It is the base for regression analysis and multivariate analysis. The statistical methods based on Gaussian distribution assumption are referred to as *classical statistical methods* (Maronna et al., 2006). They are widely used in many fields because the derivation of optimal estimators is simple and the Gaussian assumption is relatively reasonable for many data sets.

In reality, not all observed data follow the Gaussian distribution. There may be a group of *atypical data* that are far away from the majority of data. Atypical data are referred to as *outliers* or *gross errors*, which follow other distributions or there is *no clear distribution* to describe them. The probability distribution describing the data set containing outliers has a nearly Gaussian shape in the center and heavier tail than the one in Gaussian distribution. It is referred as *heavy-tailed distribution*. Classical statistical methods are very sensitive to

outliers. Even one single outlier drastically degrades the estimated results. More *robust* estimates are needed such that they are acceptable even when the data do not strictly follow the given distribution. Robust methods provide almost the same estimation results as classical methods when no outliers are present in the data, and should get almost the same results as the classical methods applied on the available "clean" data when the data contain outliers. Compared with classical statistics, this kind of statistics is called *robust statistics*. Robust statistics combines the parametric and nonparametric approaches. It uses parametric models for deriving information from the data, but this procedure does not critically depend on the assumptions in the parametric models (Hampel et al., 1986). In late 19th and early 20th century, scientists did leading preliminary work on robust estimates resisting outliers, e.g. geophysicist Harold Jefferys, astronomer Simon Newcomb and astrophysicist Arthur Stanley Eddington (Huber, 1981).

As for geophysics, atypical data (outliers) are also often contained in the seismic data such as noise bursts, incoherent signals arising from improper geophone coupling, and source generated noise. Claerbout and Muir (1973) explored the application of absolute value error criteria (ℓ_1 norm) for different kinds of robust geophysical data fitting when the data are contaminated with outliers. In that paper, they referred to these data as *erratic data*. They also proposed to use ℓ_1 norm for sparse promoting, which is the first application of ℓ_1 norm for sparsity (Candès et al., 2008). Taylor et al. (1979) borrowed idea from Claerbout and Muir (1973) and applied ℓ_1 norm for data fitting and model constraint in seismic deconvolution. Chave et al. (1987) proposed to use M-estimator method for computing robust power spectra, coherences, and transfer functions, and illustrated it on electromagnetic data. Scales and Gersztenkorn (1988) and Scales et al. (1988) investigated the method of least-absolute deviation (LAD or ℓ_1 norm) for robust inversion, and applied it on inverse scattering and traveltime tomography. The approximate solution is computed by iteratively reweighted least squares (IRLS) with conjugate gradient (CG) as linear system solver. Bube and Langan (1997) proposed hybrid ℓ_1/ℓ_2 minimization with the IRLS as solver and applied it on tomography. Guitton and Symes (2003) proposed to use Huber norm for robust inversion, which is solved via a quasi-Newton method. Their algorithm is tested for velocity analysis. Trickett et al. (2012) present a robust Cadzow filtering, which is similar with the POCS framework used in MSSA for denoising and interpolation (Oropeza and Sacchi, 2011).

3.2 Review of robust statistics

The estimates in classical statistics, e.g. sample mean, sample variance, sample covariance, sample correlation, linear regression via least-squares and principal component analysis are

seriously adversely influenced by outliers. Robust estimates are needed in robust statistics that fit the major part of "clean" data and are not (or not very much) influenced by outliers.

There is a branch of methods called *outlier diagnostic*. The outliers are firstly detected and then deleted or modified from the data. Then, classical statistical estimation methods are applied on the cleaned data. The detection of outliers is based on an initial classical statistical fit to the data. For example, the residuals between the data and classical statistical fit is examined and analyzed. There are some drawbacks of this method. First, this is a more subjective way for estimation because the detection procedure depends on the user's decision. Second, the initial statistical fit for detection is based on classical statistics that is that robust. Instead of diagnostic, there should be a robust estimation method that fits the bulk of the "clean" data and not sensitive to the outliers. The robust method we will discuss is a more *automatic* or *semi-automatic* and *data-adaptive* approach in the sense that it detects outliers by examining the departure between the data and a robust fit to it.

3.2.1 Location estimation

Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ are n observations of a random variable. For example, the data from measuring the weight of an apple n times. $x_1, x_2, \ldots, x_n$ can also be considered as n random variables because they are unknown before the measurements. The *location model* is given by

$$x_i = \mu + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{3.1}$$

where μ is the true value or say deterministic term, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are the measurement errors or say stochastic term. If the measurements are repeatedly conducted under same conditions, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ can be assumed to be independent and follow the same probability density function $f_\epsilon(x)$, i.e. they are independent and identically distributed (i.i.d.). In other words, $x_1, x_2, \ldots, x_n$ are also i.i.d.. The aim of location estimation is to derive an estimate from the observations that approaches μ.

In classical statistics, the sample mean provides an estimate of the center position of the observations (Definition 2.30)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.2}$$

Sample mean $\bar{x}$ is an optimal estimate (MLE) when the observed data $x_1, x_2, \ldots, x_n$ follow Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$.

Sample median is a robust alternative for location estimation when the data contain outliers. The sample median is defined as the numerical value such that the number of observations larger than it equals to the number of observations smaller than it. If the n real observations

$x_1, x_2, \ldots, x_n$ are ordered from small to large

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}, \tag{3.3}$$

$x_{(i)}$ is named the ith order statistic. If n is odd, $x_{(\frac{n+1}{2})}$ is the sample median. If n is even, any value between $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$ can be regarded as sample median. Usually, the $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$ is taken. Sample mean is an optimal estimate (MLE) if the observed data $x_1, x_2, \ldots, x_n$ follow Laplace distribution. Figure 3.1 demonstrates the robustness of sample median over sample mean.

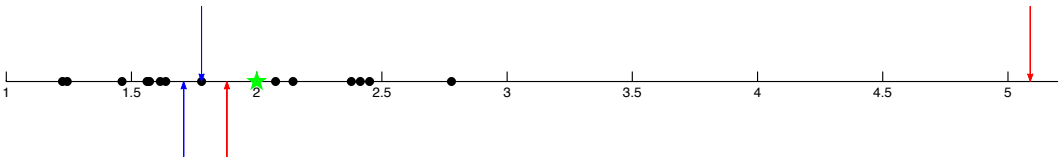

Figure 3.1: Gaussian distributed samples (black solid circles), population mean (green five-pointed star), the sample mean (red arrow below axis) and sample median (blue arrow below axis) of 14 "clean" samples, the sample mean (red arrow above axis) and sample median (blue arrow above axis) of 15 samples containing one outlier.

There are 15 data samples in the example of Figure 3.1. In the data set, 14 samples are drawn from Gaussian distribution $\mathrm{N}(2, 0.5^2)$. They are represented by black solid circles in the figure. The 15th sample is an outlier with the value of 50. It is not plotted in the figure because it is too large. The green five-point star indicates the true center of the variable, i.e. the population mean. The sample mean (red arrow below the horizontal axis) of the samples with outlier removed is 1.8813. The sample median (blue arrow blow horizontal axis) of samples with outlier removed is 1.7085. The sample mean (red arrow above horizontal axis) of samples containing the outlier is 5.0892. We can also see from the figure that the sample mean is far away from the bulk of data samples, and it can not describe the appropriate center of the samples any more. While, the sample median (blue arrow above horizontal axis) of samples with outlier is 1.7798. It is still valid and not influenced by the outlier.

3.2.2 Scale estimation

Except for the center position of samples, the spread, dispersion or variability of the samples should also be described by some quantities. In classical statistics, the description of spread

of the observations is given by sample standard deviation (SD).

$$s = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{\frac{1}{2}}, \tag{3.4}$$

where $\bar{x}$ is the sample mean, s is called unbiased estimate of standard deviation. For the example in last section, the sample standard deviation of 14 "clean" samples is 0.4897. While, the sample standard deviation of 15 samples is 12.4331. The SD breaks down with the presence of one single outlier in data.

The *mean absolute deviation (MD) about mean* is given by

$$\mathrm{MD}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|, \tag{3.5}$$

where $\bar{x}$ is the sample mean. Unfortunately, the mean absolute deviation about the mean is not robust to outliers because of the non-robustness of $\bar{x}$. In the above example, the MD for samples without and with outlier are 0.4219 and 5.9881, respectively.

A robust alternative and important estimate for scale is the *median absolute deviation (MAD)*

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{med}|\mathbf{x} - \mathrm{med}(\mathbf{x})|, \tag{3.6}$$

where "med" represents taking the median. In other words, MAD takes the median value of the absolute residuals about the sample median $\mathrm{Med}(\mathbf{x})$. The median absolute deviation of the clean 14 samples is 0.4019, and it is 0.3656 for the 15 samples with one outlier. It's interesting to note that adding one outlier may sometimes reduce the MAD because the sample median changes after adding one sample. The MAD of a standard normal distributed ($\sim \mathrm{N}(0,1)$) random variable is 0.6745 (Maronna et al., 2006). The MAD of a normal distributed ($\sim \mathrm{N}(\mu, \sigma^2)$) random variable is 0.6745σ. The *normalized MAD (NMAD)*

$$\mathrm{NMAD}(\mathbf{x}) = \frac{\mathrm{MAD}(\mathbf{x})}{0.6745}, \tag{3.7}$$

is used more often in order to be consistent with the standard deviation. That is to say, NMAD of a Gaussian distributed ($\sim \mathrm{N}(\mu, \sigma^2)$) random variable is its standard deviation σ.

3.2.3 Linear regression

Linear regression aims at modeling the relationship between a response variable and one or more explanatory variables. It is named simple linear regression if there is only one explanatory variable. It is called multiple linear regression if there are more than one

explanatory variable. The general *multiple linear regression model* is given by

$$y_i = \sum_{j=1}^{p} x_{ij}\alpha_j + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.8}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{3.9}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and y_i is the ith observation on response variable, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ and $(x_{i1}, x_{i2}, \ldots, x_{ip})$ is the ith observation on p explanatory variables, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ are the p regression coefficients. Sometimes, there is also an *intercept* term α_0 included in the regression model

$$y_i = \alpha_0 + \sum_{j=1}^{p} x_{ij}\alpha_j + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.10}$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{3.11}$$

Consequently, it can still be noted as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{3.12}$$

where the elements of the first column of $\mathbf{X}$ are 1 and the first element of $\boldsymbol{\alpha}$ is α_0.

In classical statistics, the estimate of regression coefficients $\boldsymbol{\alpha}$ is via least-squares estimate

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|_2^2. \tag{3.13}$$

The optimal solution is the stationary point where the derivative of cost function with respect to parameters $\boldsymbol{\alpha}$ equals to zero

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} = \mathbf{X}^T\mathbf{y}. \tag{3.14}$$

This system of equations is referred to as *normal equations*. If $\mathbf{X}$ has full column rank, the analytic solution of normal equations is given by

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.15}$$

Least-squares result is the maximum likelihood estimate (MLE) if the noise follows multivariate Gaussian distribution. A robust alternative for linear regression when data contain

outliers is the Least Absolute Deviations (LAD) regression

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|_1^1. \tag{3.16}$$

It is applied by Claerbout and Muir (1973) for robust modeling in geophysical problems. For LAD regression estimation, there is no analytic solution as the Least-squares regression estimation. It can be solved by linear programming (LP) or iteratively reweighted least-squares (IRLS).

When there is only one explanatory variable, multiple regression model reduces to *simple linear regression model* as follows

$$y_i = \mu + x_i\alpha + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.17}$$

or

$$\mathbf{y} = \mu\mathbf{1} + \alpha\mathbf{x} + \boldsymbol{\epsilon}, \tag{3.18}$$

where $y_1, y_2, \ldots, y_n$ are response variable values, $x_1, x_2, \ldots, x_n$ are explanatory variable values, α is regression coefficient, μ is intercept, and $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are noise terms. The simple linear regression analysis is actually a straight-line fitting problem with α as slope and μ as intercept. If considered in least-squares sense, the solution of simple linear regression is given by (3.14), as

$$\begin{aligned} \hat{\alpha} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\mu} &= \bar{y} - \hat{\alpha}\bar{x}, \end{aligned} \tag{3.19}$$

where $\bar{x}$ is the sample mean of explanatory variable values and $\bar{y}$ is the sample mean of response variable values.

3.3 M-estimate method

Even though the derived results of least-squares (ℓ_2 norm) estimation and least absolute deviations (ℓ_1 norm) are relatively simple, ℓ_2 norm estimator is not robust to outliers and ℓ_1 norm estimator has low efficiency (statistical efficiency, not the computation efficiency) at Gaussian distribution (Maronna et al., 2006). In location model, the variance of sample median is larger than the variance of sample mean if the samples follow Gaussian distribution. Here, sample mean and sample median are considered as random variables and have probability distributions because they are functions of samples (random variables). If $x_i \sim \mathrm{N}(\mu, \sigma^2)$, the sample mean has distribution $\mathrm{N}(\mu, \frac{\sigma^2}{n})$ and the sample median has distribution $\mathrm{N}(\mu, 1.57\frac{\sigma^2}{n})$. ℓ_1 norm considers small residuals and large residuals in the same

way. In many real data sets, the majority of data contain only Gaussian noise (i.e. the residuals are small) and a small fraction of data are outliers. The data approximately follow Gaussian distribution but has heavy tails in the probability distribution. There should be some kinds of estimations that are both *efficient* for Gaussian noise and *robust* to outliers. This is the basic motivation of robust estimation methods. There are three basic classes of robust estimates, M-estimates, L-estimates and R-estimates (Huber, 1981). L-estimates are based on linear combination of order statistics. R-estimates are derived from rank tests. M-estimate method is a generalization of *maximum likelihood estimate (MLE)* method. MLEs are special cases of M-estimates. The MLE method estimates the parameters of a statistical model from samples by maximizing the likelihood function. It is very important in statistical inference. Least-squares (ℓ_2 norm) and least absolute deviations (ℓ_1 norm) estimates are special cases of M-estimate. The M-estimates are the most flexible among the three classes of robust estimators. They also can be straightforwardly generalized to multiparameter problems, e.g. regression model and multivariate analysis. I applied M-estimate method for robust rank reduction in this dissertation.

3.3.1 M-estimates of location

In location model, error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent, identically distributed with the common density $f_\epsilon(x)$. In other words, samples $x_1, x_2, \ldots, x_n$ are also independent, identically distributed with the common density $f_x(x)$, $f_x(x) = f_\epsilon(x - \mu)$. Because the observations are independent random variables, the joint probability density function is given by

$$\mathcal{L}(\mu|x_1, x_2, \ldots, x_n) = f_x(x_1, x_2, \ldots, x_n|\mu) = \prod_{i=1}^{n} f_x(x_i|\mu) = \prod_{i=1}^{n} f_\epsilon(x_i - \mu), \tag{3.20}$$

which is referred to as *likelihood function*. The maximum likelihood estimate of μ maximizes the likelihood function $\mathcal{L}$

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\ \mathcal{L} = \underset{\mu}{\operatorname{argmax}} \prod_{i=1}^{n} f_\epsilon(x_i - \mu), \tag{3.21}$$

where $\hat{\mu}$ is referred to as *maximum likelihood estimate*. It is the optimal solution if the probability distribution is known exactly. It is more convenient to use the logarithm of $\mathcal{L}$

$$\log \mathcal{L} = \sum_{i=1}^{n} \log\ f_\epsilon(x_i - \mu). \tag{3.22}$$

Then, the maximum likelihood estimate $\hat{\mu}$ is given by

$$\hat{\mu} = \underset{\mu}{\text{argmax}} \ \log \mathcal{L} = \underset{\mu}{\text{argmax}} \sum_{i=1}^{n} \log \ f_\epsilon(x_i - \mu), \tag{3.23}$$

Noting function $\rho = -\log \ f_\epsilon$, equation (3.23) reduces to

$$\hat{\mu} = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho(x_i - \mu), \tag{3.24}$$

where ρ is called loss function. If the error terms strictly follow f_ϵ and f_ϵ is exactly known, solutions of equation (3.21), equation (3.23) and equation (3.24) are the same (the MLE). When f_ϵ is only approximately known, more attention is paid on solving optimization problem (3.24) instead of the probability distribution. The solution of (3.24) is referred to as *maximum likelihood type estimate* or *M-estiamte* (Huber, 1981). If function ρ is differentiable with respect to μ, the optimal solution of equation (3.24) can be obtained by setting the derivative equal to zero

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0, \tag{3.25}$$

where $\psi(u) = \frac{\partial \rho(u)}{\partial u}$. Now, the M-estimation problem is reduced to solving the *M-estimating equation* (3.25). Sample mean and sample median are two special cases of M-estimate. Their corresponding probability model are exactly known: Gaussian and Laplace distribution. Their derivation from the distribution model are used as examples to show the derivation procedure of M-estimates. If residuals follow a common standard normal (Gaussian) distribution, say $\epsilon_i \sim \mathrm{N}(0, 1)$ and samples $x_i \sim \mathrm{N}(\mu, 1)$. We have

$$\begin{aligned} f_\epsilon(x) =& \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}, \\ \rho = -\log \ f_\epsilon &= \frac{x^2}{2} + \log(\sqrt{2\pi}). \end{aligned} \tag{3.26}$$

In location estimation, the cost function is minimized with respect to location parameter μ

$$\begin{aligned} \hat{\mu} &= \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2} + n\log(\sqrt{2\pi}), \\ &= \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} (x_i - \mu)^2. \end{aligned} \tag{3.27}$$

The ψ function corresponds to this ρ function is $\psi(u) = u$. It is not hard to prove that the solution of (3.27) is the sample mean $\bar{x}$. If the residuals ϵ_i follow a common Laplace

distribution,

$$\begin{aligned} f_\epsilon(x) = \frac{1}{2}e^{-|x|}, \quad f_x(x) = \frac{1}{2}e^{-|x-\mu|}, \\ \rho = -\log\, f_\epsilon = |x| + \log\, 2. \end{aligned} \tag{3.28}$$

The maximum likelihood estimate $\hat{\mu}$ is

$$\begin{aligned} \hat{\mu} &= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - \mu| + \log\, 2, \\ &= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - \mu|. \end{aligned} \tag{3.29}$$

ψ function is $\psi(u) = \frac{\partial \rho(u)}{\partial u} = \operatorname{sgn}(u)$. The solution of (3.29) is given by sample median.

3.3.2 A weighted least-squares view

Robust location M-estimate can be interpreted as a weighted mean with weights given by

$$w(u) = \frac{\psi(u)}{u}. \tag{3.30}$$

Using $\psi(u) = \frac{\partial \rho(u)}{\partial u}$, we can get

$$w(u) = \frac{\partial \rho(u)}{\partial u}\frac{1}{u}. \tag{3.31}$$

Now, the M-estimating equation (3.25) changes to

$$\sum_{i=1}^{n} w(x_i - \hat{\mu})(x_i - \hat{\mu}) = 0. \tag{3.32}$$

Note that this is a non-linear equation that the weights $w(x_i - \hat{\mu})$ depend on the model parameter $\hat{\mu}$. Solution of (3.32) can be expressed as

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{3.33}$$

where $w_i = w(x_i - \hat{\mu})$. It suggests an iteratively reweighted least-squares (IRLS) algorithm for calculating the M-estimate.

3.3.3 Scale equivariant M-estimate of location

Suppose that $\hat{\mu}$ is the M-estimate of samples $(x_1, x_2, \ldots, x_n)$. If $\hat{\mu} + b$ is the M-estimate of samples $(x_1 + b, x_2 + b, \ldots, x_n + b)$, M-estimate $\hat{\mu}$ is said to be *shift equivariant*. If $a\hat{\mu}$ is the

M-estimate of sample $(ax_1, ax_2, \ldots, ax_n)$, the M-estimate $\hat{\mu}$ is said to be *scale equivariant*. Sample mean and sample median are both shift equivariant and scale equivariant. General M-estimates are shift equivariant but not necessary scale equivariant. A scale parameter is usually needed in M-estimation procedure to fix this problem.

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right), \tag{3.34}$$

where σ is a scale parameter. Result (3.34) also comes from the generalizing maximum likelihood estimate that will be shown as follows. Now, consider the location model

$$x_i = \mu + \xi_i = \mu + \sigma\epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.35}$$

where ξ_i are the new error terms. That is to say, the variance of noise inflated by σ times. The pdf of ϵ_i, ξ_i and x_i are f_ϵ, f_ξ and f_x, respectively.

$$f_x(x) = f_\xi(x - \mu) = \frac{1}{\sigma} f_\epsilon\left(\frac{x - \mu}{\sigma}\right), \tag{3.36}$$

where $\frac{1}{\sigma}$ is the normalization value to make $\int_{-\infty}^{+\infty} f_x(x) = 1$. There are three different situations for scale σ: (1) σ is known a priori (not the case for many real data set), (2) σ is computed a priori (e.g. by NMAD, often used) and (3) σ is derived by M-estimation procedure (Huber, 1981) (more complicated way).

First situation, the scale parameter σ is known, and the likelihood function is

$$\mathcal{L}(\mu | x_1, x_2, \ldots, x_n) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_\epsilon\left(\frac{x_i - \mu}{\sigma}\right). \tag{3.37}$$

Remember that $\rho = -\log f_\epsilon$, with leads to the maximum likelihood estimate $\hat{\mu}$

$$\begin{aligned} \hat{\mu} &= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right) + n\log\,\sigma \\ &= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right). \end{aligned} \tag{3.38}$$

The optimal solution which minimizes the cost function is the stationary point, the M-estimating equation is

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}}{\sigma}\right) = 0, \tag{3.39}$$

where $\psi(u) = \frac{\partial \rho(u)}{\partial u}$.

In the second situation, the scale parameter σ is approximated by a priori calculated scale $\hat{\sigma}$, e.g. the robust scale estimate normalized median absolute deviation

$$\hat{\sigma} = \text{NMAD}(\mathbf{x}) = \frac{1}{0.6745}\text{Med}\left(|\mathbf{x} - \text{Med}(\mathbf{x})|\right). \tag{3.40}$$

With the computed scale, the likelihood function is

$$\mathcal{L}(\mu|x_1, x_2, \ldots, x_n) = \frac{1}{\hat{\sigma}^n}\prod_{i=1}^{n} f_\epsilon\left(\frac{x_i - \mu}{\hat{\sigma}}\right). \tag{3.41}$$

Using $\rho = -\log f_\epsilon$, the maximum likelihood estimate $\hat{\mu}$ is given by

$$\begin{aligned}\hat{\mu} &= \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}}\right) + n\log\,\hat{\sigma} \\ &= \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}}\right).\end{aligned} \tag{3.42}$$

The optimal solution is the stationary point, the M-estimating equation is

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0, \tag{3.43}$$

where $\psi(u) = \frac{\partial\rho(u)}{\partial u}$.

In the third situation, σ is unknown and is calculated using the M-estimation method. The likelihood function is a function of two parameters μ and σ.

$$\mathcal{L}(\mu, \sigma|x_1, x_2, \ldots, x_n) = \frac{1}{\sigma^n}\prod_{i=1}^{n} f_\epsilon\left(\frac{x_i - \mu}{\sigma}\right). \tag{3.44}$$

Two unknown parameters μ and σ have to be optimized

$$(\hat{\mu}, \hat{\sigma}) = \underset{\mu,\sigma}{\text{argmax}}\ \mathcal{L}(\mu, \sigma|x_1, x_2, \ldots, x_n) = \underset{\mu,\sigma}{\text{argmax}}\ \frac{1}{\sigma^n}\prod_{i=1}^{n} f_\epsilon\left(\frac{x_i - \mu}{\sigma}\right). \tag{3.45}$$

With $\rho = -\log f_\epsilon$

$$(\hat{\mu}, \hat{\sigma}) = \underset{\mu,\sigma}{\text{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right) + n\log\,\sigma \tag{3.46}$$

Parameters μ and σ are coupled together, so an alternating minimization framework is required. From the previous discussion, the parameter μ is the solution of the M-estimating

equation

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0, \tag{3.47}$$

where $\psi(u) = \frac{\partial \rho(u)}{\partial u} = -f'_\epsilon(u)/f_\epsilon(u)$. Parameter σ is also estimated via the M-estimation procedure, given by

$$\frac{1}{n}\sum_{i=1}^{n} \theta\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = \kappa,\ 0 < \kappa < \theta(\infty), \tag{3.48}$$

where $\theta(u) = -uf'_\epsilon(u)/f_\epsilon(u) = \psi(u)u$. The details of M-estimate of scale σ are described as follows in section 3.3.4. The simultaneous estimation of μ and σ is an alternating procedure between equation (3.47) and equation (3.48). Note that the location parameter estimated by simultaneous M-estimation of location and scale is not as robust as the location estimate via M-estimators with a previously computed scale. Therefore, I will choose to compute the approximated scale a priori via the normalized absolute deviation.

3.3.4 Auxiliary step: M-estimate of scale

$$\mathcal{L}(\sigma|x_1, x_2, \ldots, x_n) = \frac{1}{\sigma^n}\prod_{i=1}^{n} f_\epsilon\left(\frac{x_i - \hat{\mu}}{\sigma}\right). \tag{3.49}$$

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}}\ \mathcal{L}(\sigma|x_1, x_2, \ldots, x_n) = \underset{\sigma}{\operatorname{argmax}} \frac{1}{\sigma^n}\prod_{i=1}^{n} f_\epsilon\left(\frac{x_i - \hat{\mu}}{\sigma}\right). \tag{3.50}$$

Taking the logarithm of $\mathcal{L}$

$$\log\ \mathcal{L} = \sum_{i=1}^{n} \log\ f_\epsilon\left(\frac{x_i - \hat{\mu}}{\sigma}\right) - n\log\ \sigma. \tag{3.51}$$

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}}\ \log\ \mathcal{L} = \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^{n} \log\ f_\epsilon\left(\frac{x_i - \hat{\mu}}{\sigma}\right) - n\log\ \sigma. \tag{3.52}$$

Setting derivative $\partial \log\ \mathcal{L}/\partial\sigma = 0$, we can get

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{1}{f_\epsilon(\frac{x_i-\hat{\mu}}{\hat{\sigma}})} \frac{\mathrm{d}f_\epsilon(\frac{x_i-\hat{\mu}}{\hat{\sigma}})}{\mathrm{d}(\frac{x_i-\hat{\mu}}{\hat{\sigma}})} \left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 1. \tag{3.53}$$

Set $\theta(u) = -uf'_\epsilon(u)/f_\epsilon(u)$, and the above equation reduces to

$$\frac{1}{n}\sum_{i=1}^{n} \theta\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 1. \tag{3.54}$$

Scale estimates of Gaussian and Laplace distribution will be illustrated as examples. If

$\epsilon_i \sim \mathrm{N}(0,1)$, $\xi_i \sim \mathrm{N}(0,\sigma)$ and $x_i \sim \mathrm{N}(\mu,\sigma)$.

$$f_\epsilon(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{3.55}$$

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{3.56}$$

$$\theta(x) = x^2 \tag{3.57}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right)^2 = 1 \tag{3.58}$$

$$\hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right]^{1/2}, \tag{3.59}$$

where $\hat{\mu}$ is given by sample mean $\bar{x}$, $\hat{\sigma}$ is the MLE of standard deviation under Gaussian distribution assumption. Similarly, when errors follow Laplace distribution

$$f_\epsilon(x) = \frac{1}{2} e^{-|x|} \tag{3.60}$$

$$f_x(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma} \tag{3.61}$$

$$\theta(x) = |x| \tag{3.62}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right| = 1 \tag{3.63}$$

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{\mu}|, \tag{3.64}$$

where $\hat{\mu}$ can be given by sample median, $\hat{\sigma}$ is called *mean absolute deviation about median*. Any solution of the following equation is an M-estimate of scale (Maronna et al., 2006).

$$\frac{1}{n}\sum_{i=1}^{n}\theta\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = \kappa,\ 0 < \kappa < \theta(\infty), \tag{3.65}$$

where κ is a constant. Scale estimate is said to be equivariant if $\hat{\sigma}(ax_1, ax_2, \ldots, ax_n) = a\hat{\sigma}(x_1, x_2, \ldots, x_n)$ for any $a > 0$. Scale M-estimate is equivariant.

3.3.5 Iteratively reweighted least-squares

The weighted least-square view of location M-estimate suggests an iteratively reweighted least-squares method to compute it. The scale σ is computed a priori by the normalized median absolute deviation. The algorithm is summarized as

Algorithm 1 M-estimate of location via IRLS with previously computed scale

1: Compute scale $\hat{\sigma} = \text{NMAD}(\mathbf{x})$
2: **Initialization:** $\hat{\mu}^0 = \text{Med}(\mathbf{x})$
3: **While** $\frac{|\hat{\mu}^{k+1} - \hat{\mu}^k|}{\hat{\sigma}} > \varepsilon$ or $k < kmax$ **do**:
4: $w_i^k = w\left(\frac{x_i - \hat{\mu}^k}{\hat{\sigma}}\right),\ i = 1, 2, \ldots, n$
5: $\hat{\mu}^{k+1} = \frac{\sum_{i=1}^n w_i^k x_i}{\sum_{i=1}^n w_i^k}$
6: $k \leftarrow k + 1$
7: **End**

3.3.6 Loss function ρ, ψ function and weight function

It is very important to point out that robust statistics actually focus on the robust estimates in (3.24) with a given loss function ρ, the estimates are *not necessary* the MLEs of *any distributions* (Maronna et al., 2006). Holland and Welsch (1977) discussed several different loss functions ρ. There are two frequently used loss functions: Huber function and biweight functions. Huber function (Huber, 1964) is as follows

$$\rho_H(x) = \begin{cases} \dfrac{x^2}{2} & |x| \le \tau \\ \tau|x| - \dfrac{\tau^2}{2} & |x| > \tau, \end{cases} \tag{3.66}$$

where τ is a tuning constant in Huber function. The corresponding ψ-function is

$$\psi_H(x) = \begin{cases} x & |x| \le \tau \\ \tau \text{sgn}(x) & |x| > \tau, \end{cases} \tag{3.67}$$

Huber estimate belongs to monotone M-estimates because that $\psi_H(u)$ is a monotonic function (Figure 3.3). The weight function of Huber-estimate is

$$w_H(x) = \begin{cases} 1 & |x| \le \tau \\ \dfrac{\tau}{|x|} & |x| > \tau, \end{cases} \tag{3.68}$$

The biweight function proposed by Beaton and Tukey (1974) is

$$\rho_B(x) = \begin{cases} \dfrac{1}{6}\tau^2 \left\{1 - \left[1 - \left(\dfrac{x}{\tau}\right)^2\right]^3\right\} & |x| \le \tau \\ \dfrac{1}{6}\tau^2 & |x| > \tau, \end{cases} \tag{3.69}$$

where τ is a tuning constant in biweight function. The corresponding ψ-function is

$$\psi_B(x) = \begin{cases} x\left[1-\left(\frac{x}{\tau}\right)^2\right]^2 & |x| \le \tau \\ 0 & |x| > \tau. \end{cases} \tag{3.70}$$

Biweight estimate belongs to redescending M-estimates because that $\psi_B(u)$ is not monotonic function (Figure 3.3). Redescending M-estimates are more suitable than monotone M-estimates when the data contains extreme outliers. Biweight M-estimate completely rejects outliers, while Huber M-estimate limits the influence of outliers. Biweight estimate is not MLE for any distribution (Maronna et al., 2006). The weight function of biweight estimate is

$$w_B(x) = \begin{cases} \left[1-\left(\frac{x}{\tau}\right)^2\right]^2 & |x| \le \tau \\ 0 & |x| > \tau. \end{cases} \tag{3.71}$$

Biweight estimation gives zero weight to outliers, it completely removes outliers with the fitting of the bulk of the "clean" data. Remember that the loss function, ψ-function and weight function of least-squares estimation are $\rho_Q(x) = \frac{1}{2}x^2$, $\psi_Q(x) = x$ and $w_Q(x) = 1$. Three different loss functions are plotted in Figure (3.2). Quadratic function increases fastest among the three, Huber function takes the second place and biweight function increases slowest. Quadratic and Huber functions are convex functions. While, biweight function is non-convex.

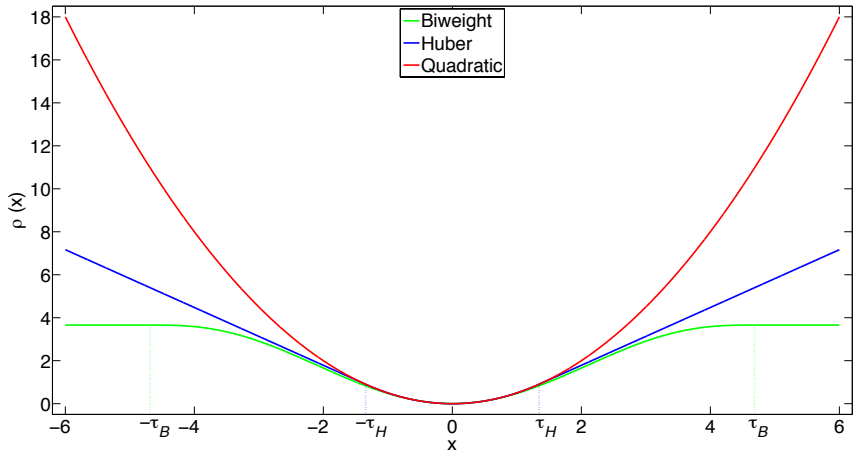

Figure 3.2: Three loss functions. Quadratic function, Huber function and Biweight function. τ_H is the tuning constant in Huber function, τ_B is the tuning constant in biweight function.

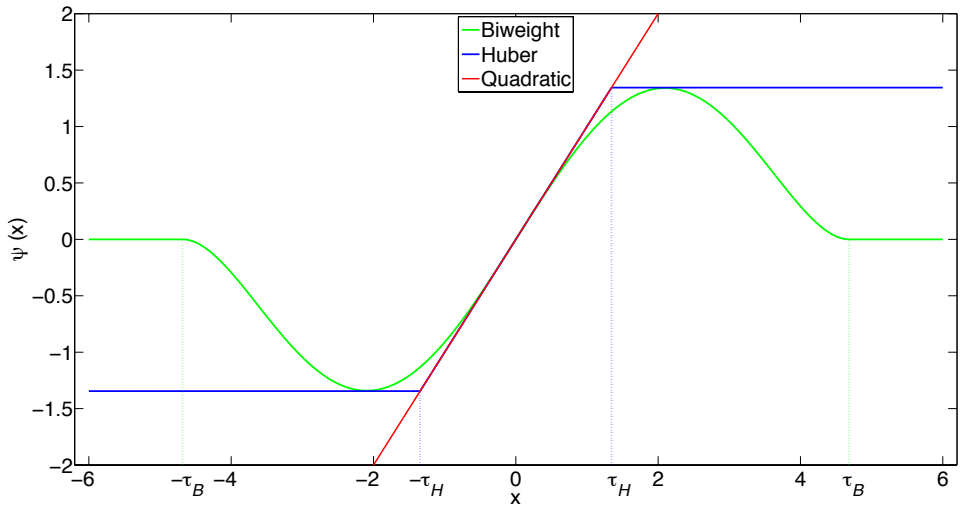

Figure 3.3: Three ψ functions corresponding to Quadratic function, Huber function and Biweight function. τ_H is the tuning constant in ψ function of Huber function, τ_B is the tuning constant in ψ function of biweight function.

The ψ functions are shown in Figure 3.3. ψ_Q and ψ_H are monotone functions, ψ_B is a "redescending" function. ψ_Q is a linear function, ψ_H and ψ_B are nonlinear.

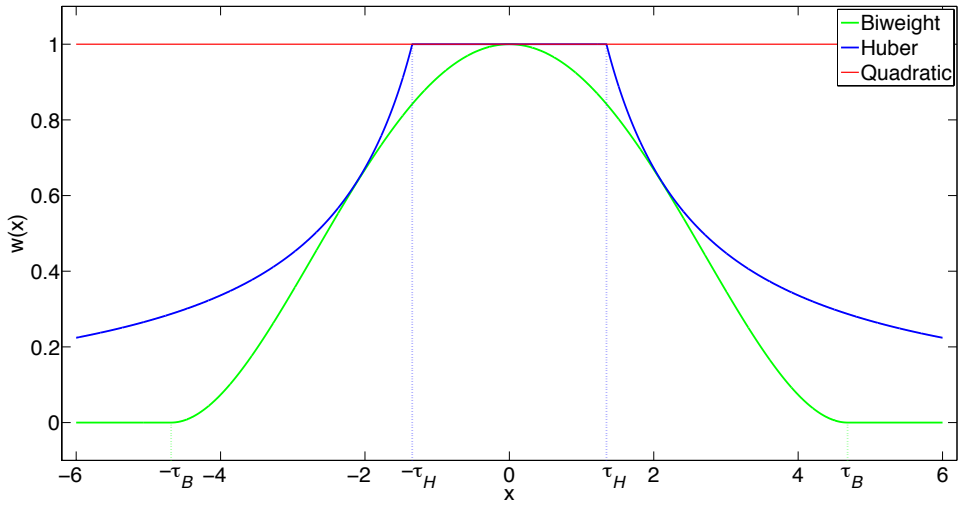

Figure 3.4: Weight functions corresponding to Quadratic function, Huber function and biweight function. τ_H is the tuning constant in weight function of Huber function, τ_B is the tuning constant in weight function of biweight function.

Figure 3.4 show the comparison of weight functions for quadratic, Huber and biweight estimates. Quadratic function gives equal weights to small and large errors, Huber function gives less weight to large errors and biweight function gives zero weights to outliers. The important tuning constant τ controls the robustness and statistical efficiency (at Gaussian distribution) of the M-estimate. The *asymptotic efficiency* of an M-estimate is defined as the ratio of the *asymptotic variance* of the maximum likelihood estimate at target distribution (e.g. Gaussian distribution) and the *asymptotic variance* of the M-estimate. It measures how near the M-estimate is to the optimum. For example, sample mean has 100% efficiency at Gaussian distribution, sample median has 63.69% efficiency at Gaussian distribution. In M-estimation, the smaller value τ is, the more data with large residuals are considered as outliers, and the more robust the estimation procedure is. But smaller τ can result in lower statistical efficiency of the estimate. There is a trade-off between robustness and efficiency. Holland and Welsch (1977) gives the tuning constants of several different estimation functions for 95% asymptotic efficiency at the Gaussian distribution. They suggests $\tau = 1.345$ for Huber function and $\tau = 4.685$ for biweight function for 95% asymptotic efficiency at the Gaussian distribution.

Figure 3.5 show the results of biweight M-estimate of location with previously computed scale parameter. The iteratively reweighted least-squares (Algorithm 1) was adopted for this example. The green arrows above and below the horizontal axis are the biweight M-estimator of location of 15 samples and 14 "clean" samples, respectively. The presence of outlier does not change the M-estimate of location too much.

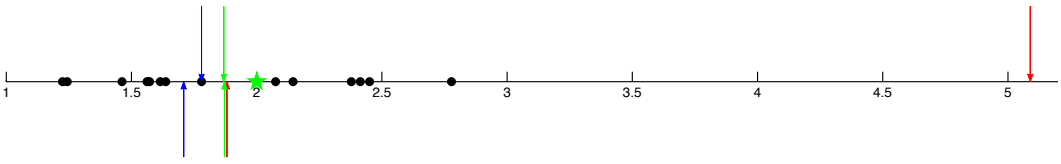

Figure 3.5: Gaussian distributed samples (black solid circles), population mean (green five-pointed star), the sample mean (red arrow below axis), sample median (blue arrow below axis) and M-estimate of location using biweight function (green arrow below axis) of 14 "clean" samples, the sample mean (red arrow above axis), sample median (blue arrow above axis) and M-estimate of location using biweight function (green arrow above axis) of 15 samples containing one outlier.

3.4 Robust SSA

This section proposes a robust singular spectrum analysis algorithm that adopts the M-estimate procedure. The truncated SVD in traditional SSA is based on least-squares minimization. It is quite sensitive to outliers. As seen in Figure 2.15, even one trace with erratic noise will degrade the performance of SSA. Instead of TSVD, we propose to use a robust low rank approximation based on M-estimate in the SSA framework. Robust low rank approximation, or called robust rank-reduction or robust matrix factorization is an application of robust M-estimation in multivariate analysis.

3.4.1 Robust low rank approximation

We now propose to replace the Frobenius metric for distance between two matrices in equation (2.93) by a robust metric (Verboon and Heiser, 1994; De la Torre and Black, 2003; Maronna and Yohai, 2008). The new problem becomes

$$\begin{aligned} \mathbf{M}_K = \mathcal{R}_K(\mathbf{M}) = \underset{\hat{\mathbf{M}}}{\operatorname{argmin}} \|\mathbf{M} - \hat{\mathbf{M}}\|_\rho \\ \text{subject to } \operatorname{rank}(\hat{\mathbf{M}}) = K, \end{aligned} \tag{3.72}$$

where $||\mathbf{M} - \hat{\mathbf{M}}||_\rho = \sum_{i=1}^{m} \sum_{j=1}^{n} \rho(\frac{m_{ij} - \hat{m}_{ij}}{\sigma})$, m_{ij} is the element at i-th row and j-th column of $\mathbf{M}$, σ is a scale parameter for function ρ. When ρ is not quadratic, problem 3.72 is a non-convex optimization problem. No closed-form solution exists for this problem in general and further more the non-convex cost function has local minima. These make solving for global minimum not an easy task.

Matrix Factorization

The low rank approximation problem 3.72 can be addressed in a matrix factorization view

$$\begin{aligned} &\min_{\hat{\mathbf{M}}} ||\mathbf{M} - \hat{\mathbf{M}}||_\rho, \\ &s.t. \quad \hat{\mathbf{M}} = \mathbf{U}\mathbf{V}^H, \end{aligned} \tag{3.73}$$

where $\mathbf{U} \in \mathbb{C}^{m \times K}$, $\mathbf{V} \in \mathbb{C}^{n \times K}$ are the two factor matrices. K is the dimension of the approximated subspace. The reason for replacing the constraint of rank function by the product of two factor matrices is that approximating a matrix by rank K is equivalent to fitting the matrix by a matrix product $\mathbf{UV}$, where the size of $\mathbf{U}$ and $\mathbf{V}$ is $m \times K$ and $K \times n$, respectively (Gabriel and Zamir, 1979). Or rewrite the above problem 3.73 to an

unconstrained one

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{M} - \mathbf{U}\mathbf{V}^H||_\rho. \tag{3.74}$$

Robust M-estimate

The goal is to find the approximation matrix $\mathbf{U}\mathbf{V}^H$ to minimize the following cost function

$$E(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \rho\left(\frac{m_{ij} - \sum_{q=1}^{K} u_{iq}v_{jq}^*}{\sigma}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} \rho\left(\frac{r_{ij}}{\sigma}\right), \tag{3.75}$$

where σ is a robust measurement of the noise variation (dispersion estimate). We use the same scale parameter σ for all the elements in the residual matrix because the noise in $\mathbf{D}$ is assumed to be independent and identically distributed (i.i.d.). The selection of σ is given in the *Parameter Selection* section. The function $\rho\left(\frac{r_{ij}}{\sigma}\right)$ is not an analytic function with respect to r_{ij} in the complex domain. According to Wirtinger's Caculus (Brandwood, 1983), $\rho\left(\frac{r_{ij}}{\sigma}, \frac{r_{ij}^*}{\sigma}\right)$ is regarded as a function of both r_{ij} and r_{ij}^* and the partial complex-variable derivative is applied here. More details are given in appendix. By taking the derivative of equation 3.75 with respect to $\mathbf{U}^*$ and $\mathbf{V}^*$, we get the following M-estimate equations

$$\begin{aligned}
&\sum_{j=1}^{n} \psi_1\left(\frac{r_{aj}}{\sigma}\right) v_{jb} = 0, \quad a = 1,\ldots,n \quad b = 1,\ldots,p, \\
&\sum_{i=1}^{m} \psi_2\left(\frac{r_{ic}}{\sigma}\right) u_{id} = 0, \quad c = 1,\ldots,n \quad d = 1,\ldots,p
\end{aligned} \tag{3.76}$$

where $\psi_1(x) = \dfrac{\partial\rho(x)}{\partial x^*}$ and $\psi_2(x) = \dfrac{\partial\rho(x)}{\partial x}$ are the ψ-functions. The above M-estimate equations can be reformed as the weighted least squares problem

$$\begin{aligned}
&\sum_{j=1}^{n} w\left(\frac{r_{aj}}{\sigma}\right) r_{aj} v_{jb} = 0, \quad a = 1,\ldots,n \quad b = 1,\ldots,p, \\
&\sum_{i=1}^{m} w\left(\frac{r_{ic}}{\sigma}\right) r_{ic}^* u_{id} = 0, \quad c = 1,\ldots,n \quad d = 1,\ldots,p,
\end{aligned} \tag{3.77}$$

where $w(x) = \dfrac{\psi_1(x)}{x} = \dfrac{\partial\rho(x)}{\partial x^*}\dfrac{1}{x} = \dfrac{\psi_2(x)}{x^*} = \dfrac{\partial\rho(x)}{\partial x}\dfrac{1}{x^*} = \dfrac{1}{2}\dfrac{\partial\rho(x)}{\partial|x|}\dfrac{1}{|x|}$ is the weight function. Note that Equation (3.77) is non-linear, the weights w depends on the model and the model depends on the weights. This non-linear problem can be solved by the iteratively reweighted least squares (IRLS) method. The weights are approximately obtained from the residual of

the previous iteration. The cost function 3.75 is approximated by

$$E_W(\mathbf{U},\mathbf{V}) = ||\mathbf{W}^{\frac{1}{2}} \odot (\mathbf{M} - \mathbf{U}\mathbf{V}^H)||_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}|m_{ij} - \sum_{q=1}^{K} u_{iq}v_{jq}^*|^2, \tag{3.78}$$

where $\frac{1}{2}$ on the upper right of the matrix indicates elementwise square root of the matrix. $\mathbf{W} \in \mathbb{R}^{+m\times n}$ is the weighting matrix calculated from the residuals of previous IRLS iteration. The symbol $\odot$ represents the Hadamard product (elementwise product). w_{ij} is the ith row and jth column of $\mathbf{W}$ given by $w_{ij} = w(\frac{r_{ij}}{\sigma})$. The alternating minimization algorithm (Gabriel and Zamir, 1979) can be approximated by

$$\begin{aligned} E_W(\mathbf{V}) &= ||\mathbf{W}^{\frac{1}{2}} \odot (\mathbf{M} - \mathbf{U}\mathbf{V}^H)||_F^2 = \sum_{j=1}^{n}(\mathbf{m}^j - \mathbf{U}\mathbf{v}_j)^H\mathbf{W}^j(\mathbf{m}^j - \mathbf{U}\mathbf{v}_j), \\ E_W(\mathbf{U}) &= ||\mathbf{W}^{\frac{1}{2}} \odot (\mathbf{M} - \mathbf{U}\mathbf{V}^H)||_F^2 = \sum_{i=1}^{m}(\mathbf{m}_i - \mathbf{V}\mathbf{u}_i)^H\mathbf{W}_i(\mathbf{m}_i - \mathbf{V}\mathbf{u}_i), \end{aligned} \tag{3.79}$$

where

$$\mathbf{M} = \begin{pmatrix} \mathbf{m}^1 & \mathbf{m}^2 & \cdots & \mathbf{m}^n \end{pmatrix} = \begin{pmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_m \end{pmatrix}^H, \tag{3.80}$$

here all the vectors are column vectors; $\mathbf{m}^j$ is the jth column of $\mathbf{M}$; $\mathbf{m}_i$ is the conjugate transpose of ith row of $\mathbf{M}$. Similarly,

$$\begin{aligned} \mathbf{U} &= \begin{pmatrix} \mathbf{u}^1 & \mathbf{u}^2 & \cdots & \mathbf{u}^K \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{pmatrix}^H, \\ \mathbf{V} &= \begin{pmatrix} \mathbf{v}^1 & \mathbf{v}^2 & \cdots & \mathbf{v}^K \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}^H, \end{aligned} \tag{3.81}$$

$\mathbf{W}^j = diag\{\mathbf{w}^j\} \in \mathbb{R}^{+n\times n}$ is the diagonal weighting matrix containing the jth column of $\mathbf{W}$. $\mathbf{W}_i = diag\{\mathbf{w}_i\} \in \mathbb{R}^{+m\times m}$ is diagonal weighting matrix containing the ith row of $\mathbf{W}$. Equation 3.79 can be broken up into smaller optimization problems

$$for\ i = 1, 2..., m \qquad \min_{\mathbf{u}_i} (\mathbf{m}_i - \mathbf{V}\mathbf{u}_i)^H\mathbf{W}_i(\mathbf{m}_i - \mathbf{V}\mathbf{u}_i), \tag{3.82a}$$

$$for\ j = 1, 2, ..., n \qquad \min_{\mathbf{v}_j} (\mathbf{m}^j - \mathbf{U}\mathbf{v}_j)^H\mathbf{W}^j(\mathbf{m}^j - \mathbf{U}\mathbf{v}_j). \tag{3.82b}$$

The updating of $\mathbf{U}$ and $\mathbf{V}$ is an alternating procedure. The closed-form solutions of the above equations are given by

$$for\ i = 1, 2..., m \qquad \mathbf{V}^H\mathbf{W}_i\mathbf{V}\mathbf{u}_i = \mathbf{V}^H\mathbf{W}_i\mathbf{m}_i, \tag{3.83a}$$

$$for\ j = 1, 2, ..., n \qquad \mathbf{U}^H\mathbf{W}^j\mathbf{U}\mathbf{v}_j = \mathbf{U}^H\mathbf{W}^j\mathbf{m}^j. \tag{3.83b}$$

For the examples in this dissertation, QR factorization (Golub and Van Loan, 1996) is used for solving the weighted least-squares minimization problems (Equation (3.82)).

The choice of the robust function ρ depends on the how many outliers are there in the data or how robust the algorithm is desired. Redescending M-estimate is more robust than monotone M-estimate with the price of non-convexity. In this thesis, we use the Tukey's bisquare function (Beaton and Tukey, 1974). The bisquare function in the complex domain is given by

$$\rho_B(x) = \begin{cases} \frac{1}{6}\tau^2 \left\{ 1 - \left[1 - \left(\frac{|x|}{\alpha} \right)^2 \right]^3 \right\} & |x| \leq \tau \\ \frac{1}{6}\tau^2 & |x| > \tau \end{cases} . \tag{3.84}$$

The weighting function for bisquare function is

$$w_B(x) = \begin{cases} \left[1 - \left(\frac{|x|}{\tau} \right)^2 \right]^2 & |x| \leq \tau \\ 0 & |x| > \tau \end{cases} , \tag{3.85}$$

where the tuning constant α is chosen to get both high efficiency for attenuating Gaussian noise and robustness for eliminating outliers. The selection of α will be discussed in the *Parameter Selection* section.

Iterative Algorithm

The robust low rank approximation algorithm can summarized as follows
(1) Start with initial model $\mathbf{U}$ and $\mathbf{V}$.
(2) Calculate residual matrix $\mathbf{R} = \mathbf{M} - \mathbf{U}\mathbf{V}^H$.
(3) Calculate weighting matrix $\mathbf{W}$ using equation 3.85.
(4) Update factor matrix $\mathbf{U}$ by solving least-squares minimization problem (3.82a) with QR factorization.
(5) Update factor matrix $\mathbf{V}$ by solving least-squares minimization problem (3.82b) with QR factorization.
(6) Iterate steps (4) - (5) until convergence or reach a maximum iteration number.
(7) Iterate steps (2) - (6) until convergence or reach a maximum iteration number.

Parameter Selection and Initialization

We adopt the normalized MAD as the robust scale

$$\sigma = 1.4826 \text{ MAD } = 1.4826 \text{ med } | \mathbf{r} - \text{med } | \mathbf{r} ||, \tag{3.86}$$

where $\mathbf{r}$ is the residual vector obtained by reshaping the misfit matrix $\mathbf{R}$ from the previous iteration. The multiplication of 1.4826 is used for adjusting the bias between MAD and standard deviation (SD) at Gaussian noise distribution. Holland and Welsch (1977) recommend to fix the scale σ because the convergence is supported by known theory. There is only convergence theory for iterating scale when Huber loss function is used (Huber, 1981). Holland and Welsch (1977) states that if the scale is iterated and/or least-squares initialization is used, it is better to iterate the algorithm with Huber loss function with scale iterated using equation 3.86 until convergence and use a non-convex loss function with the scale fixed using Huber scale. However, the least-squares initialization TSVD is quite skewed by the outliers. We prefer to use a non-convex loss function directly with the scale iterated in each iteration using equation (3.86). For initialization, I adopt the TSVD of the matrix $\text{med}(\text{vec}(\mathbf{M})) \times \text{rand}(m, n)$. Here, $\text{vec}(\cdot)$ is a vectorization operator, $\text{med}(\cdot)$ is used to get the median of a vector and $\text{rand}(m, n)$ indicates a m by n matrix whose elements are random numbers between 0 and 1.This initialization strategy is used throughout this work. More robust and complicated initialization strategies are also possible, but generally they are more computationally demanding. For the tuning constant α, Holland and Welsch (1977) recommend to take $\alpha = 4.685$ for the bisquare function to get 95% asymptotic efficiency at standard normal distribution. Maronna et al. (2006) gives different α values for different asymptotic efficiency at the standard normal distribution. The value $\alpha\sigma$ performs as the threshold to distinct outliers and inlierss. Smaller $\alpha\sigma$ will penalize the outliers more heavily which results in a robust estimation.

3.5 Examples

We present a synthetic example in t-x domain to test the performance of robust low rank approximation. The results of TSVD and robust low rank approximation are compared. Then, we present a synthetic example and also two field data examples to illustrate the proposed robust SSA algorithm. We compare the performance of the robust SSA, classical SSA and f-x deconvolution for erratic and Gaussian noise attenuation.

3.5.1 t-x domain robust rank reduction

Figure 3.6 shows the comparison of TSVD and robust low rank approximation on 2D t-x seismic data section. Figure 3.6 (d) is a synthetic seismic section with four flat events. High-amplitude non-Gaussian noise is added to this data as shown in Figure 3.6 (a). Figure 3.6 (b) is the result after TSVD filtering. The rank for reconstruction is chosen to be 2. There is still high-amplitude non-Gaussian noise left in the result. Figure 3.6 (c) is the result

after robust low rank approximation filtering. The rank chosen for rank reduction is also 2. The tuning constant α for bisquare estimator is 3. The numbers of external iterations (for updating weights) and internal iterations (for alternating minimization) equal to 10 and 5, respectively. More non-Gaussian noise is suppressed by robust low rank approximation. Both TSVD and robust low rank approximation (Figure 3.6 (e) and (f)) preserve the signal.

3.5.2 Synthetic Example

Figure 3.7 (b) shows a 2-D synthetic t-x data set, which has 40 traces and a total time of 1.2 s with sampling interval 0.004 s. It contains Gaussian noise with signal-to-noise ratio (SNR) equal to 1, and isolated noisy traces. The SNR here is defined as the ratio between the maximum amplitude of the clean data and the maximum amplitude of the Gaussian noise. The amplitude of the erratic noise traces is 3 and 2 times of the maximum amplitude of the uncorrupted data. The wiggles have been clipped (clip=1 in Seismic Unix) when plotted and it's the same for other wiggle plots in this synthetic example. The processing frequency band ranges from 1 to 40 Hz. We select the size of subspace of the reconstructed data in SSA and robust SSA methods to be $K = 3$. We choose the number of external iterations (for updating weights) equal to 10 and number of internal iterations (for alternating minimization) equal to 5. The tuning constant α for bisquare estimator is set to be 4.685. The length of prediction filter in f-x deconvolution is 10, the trade-off parameter is 0.001. The results of f-x deconvolution, SSA and robust SSA are compared. Figure 3.7 (a) is the noise free data, Figure 3.7 (b) is the contaminated noisy data and Figure 3.7 (c) is the added noise. Figure 3.8 (a) shows the result of f-x deconvolution, we can see that the result is not very good because large amplitude noise leaks over several traces in the output panel. Shorter prediction filter can remove more noise but it also distorts the signal more seriously. No matter what parameters are chosen, the f-x deconvolution cannot eliminate the high-amplitude erratic noise. The outlier (atypical observations) affects the estimation of correct prediction filter in f-x deconvolution. Figure 3.8 (b) shows the result of the classical non-robust SSA implemented via the TSVD. Again, we observe that the erratic noise has not been properly removed and noticeable artifacts are present in the output gather. The TSVD used in classical SSA cannot correctly extract the singular vectors/values which properly explain the variances of signal from the outliers corrupted Hankel matrix. The result of robust SSA method is shown in Figure 3.8 (c). In this case, the Gaussian and erratic noise were successfully suppressed. By examining the error panels (input noisy data minus filtered data) of the three methods (Figure 3.9), we can see the obvious energy leakage of the f-x deconvolution (Figure 3.9(a)). However, robust SSA preserves the original signal (Figure 3.9(c)). We also compare the result of robust SSA

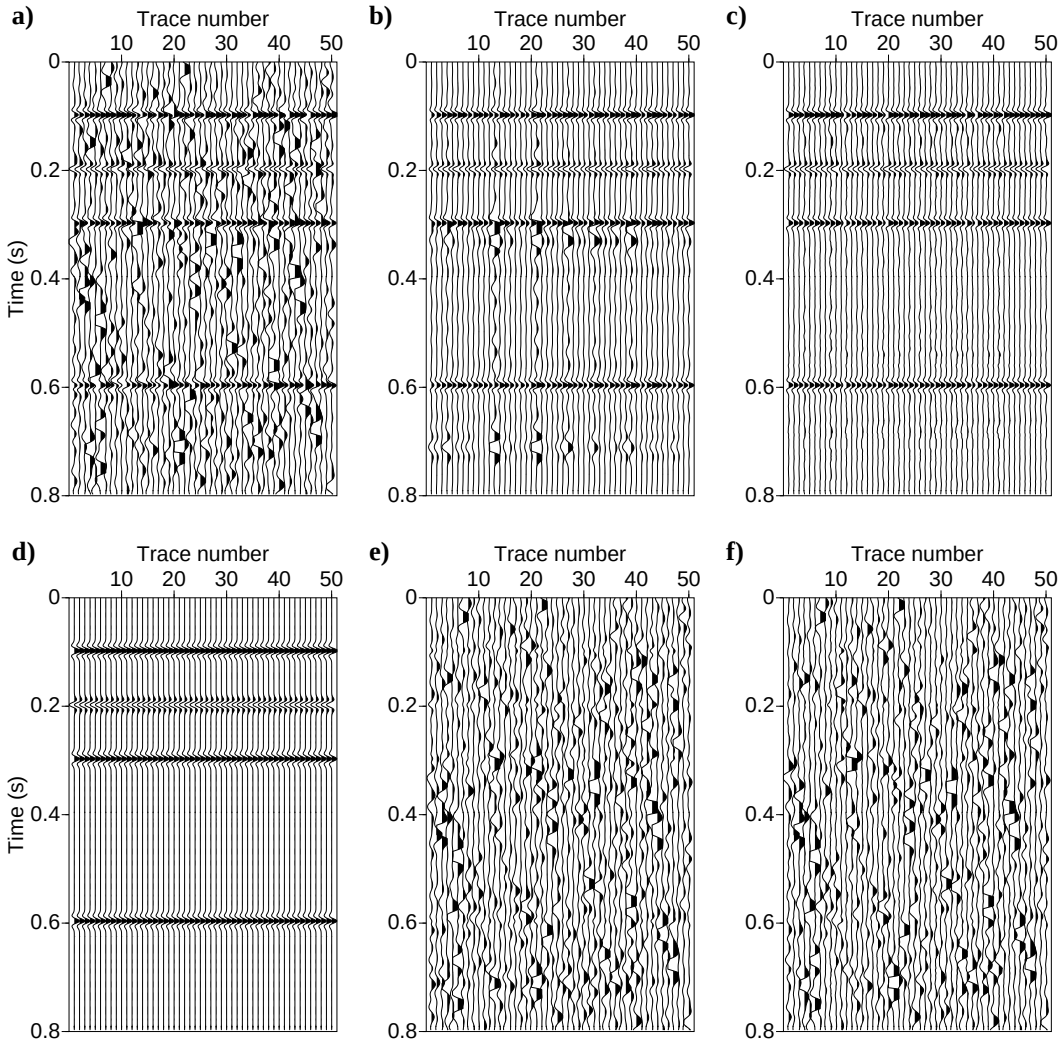

Figure 3.6: a) 2D synthetic data with four flat events and non-Gaussian noise. b) Data after TSVD filtering (rank=2). c) Data after robust low rank approximation filtering (rank=2). d) 2D noise-free synthetic data. e) Difference between noisy data and result of TSVD filtering. f) Difference between noisy data and result of robust low rank approximation filtering.

on data corrupted with erratic noise and Gaussian noise (Figure 3.8(c)), with the result of classical SSA on data with only Gaussian noise (Figure 3.10(b)). Note that the Gaussian noise in Figure 3.10(a) is the same with Gaussian noise in Figure 3.7(b). We can find that the two results are quite similar to each other. We evaluate the denoising performance by evaluating the factor $Q = 10 \log \frac{||d^0||_F^2}{||d^0-\hat{d}||_F^2}$, where d^0 is the noise free data, $\hat{d}$ is the

reconstructed data. Larger value of Q means better denoising performance. The Q value of f-x deconvolution is $Q_{fx} = 7.7$, the Q value for SSA is $Q_{ssa} = -2.8$ and the Q value of robust SSA is $Q_{rssa} = 12.8$. The Q value of the classical SSA on data with only Gaussian noise (Figure 3.10) is $Q_{ssaG} = 13.1$. These values indicate that the robust SSA method offers a good alternative to SSA and f-x deconvolution when the data are contaminated by erratic noise.

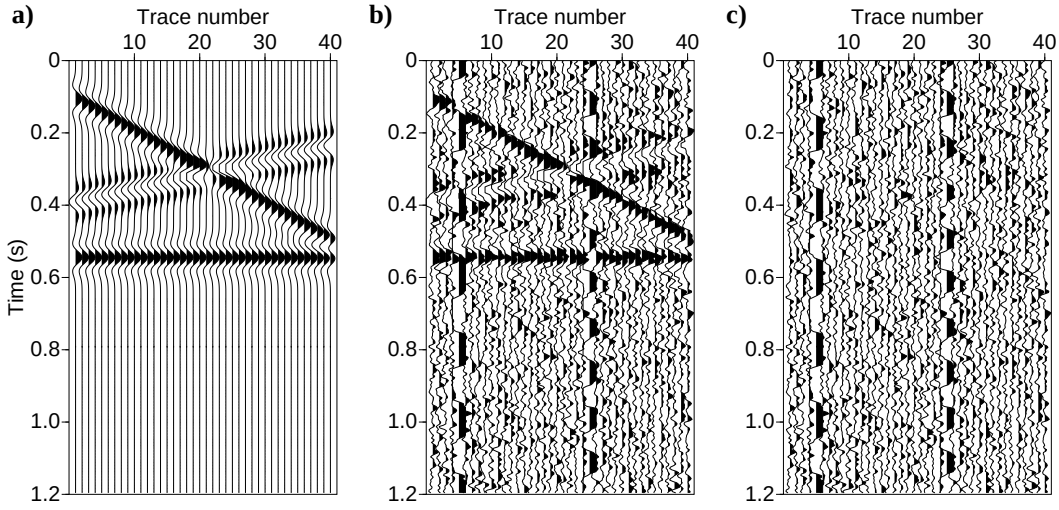

Figure 3.7: The synthetic data with three linear events. (a) Clean data. (b) Data with Gaussian noise and erratic spatial noise. (c) The noise added to the data.

3.5.3 Field Data Example

Western Canadian Sedimentary Basin

Figure 3.11(a) is a poststack data section from a survey in Western Canadian Sedimentary Basin. It has 800 traces and 1500 time samples with the time sample interval equals to 2 ms. Figure 3.11(b) and Figure 3.11(c) are the zoomed data section in the left and right rectangular windows highlighted in Figure 3.11(a), respectively. We can see high-amplitude noise in this data set. The whole data are divided into overlapped windows with suitable size. Then, all windows are filtered and added back to recover the clean data. In spatial direction, each window has 50 traces and the overlap between two adjacent windows is 25 traces. In temporal direction, each window has 300 samples (0.6 s) and the overlap between two adjacent windows is 100 samples (0.2 s). All the three filtering methods are applied for frequencies in the band of 1-80 Hz. The size of the reconstructed subspace in both SSA and

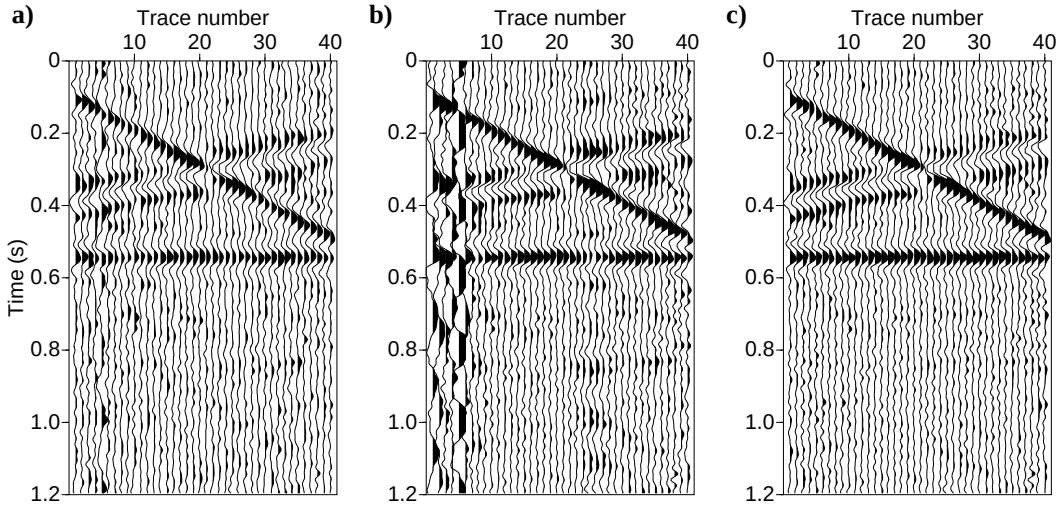

Figure 3.8: (a) Data in Figure 3.7(b) after f-x deconvolution. (b) Data after classical SSA filtering. (c) Data after robust SSA filtering.

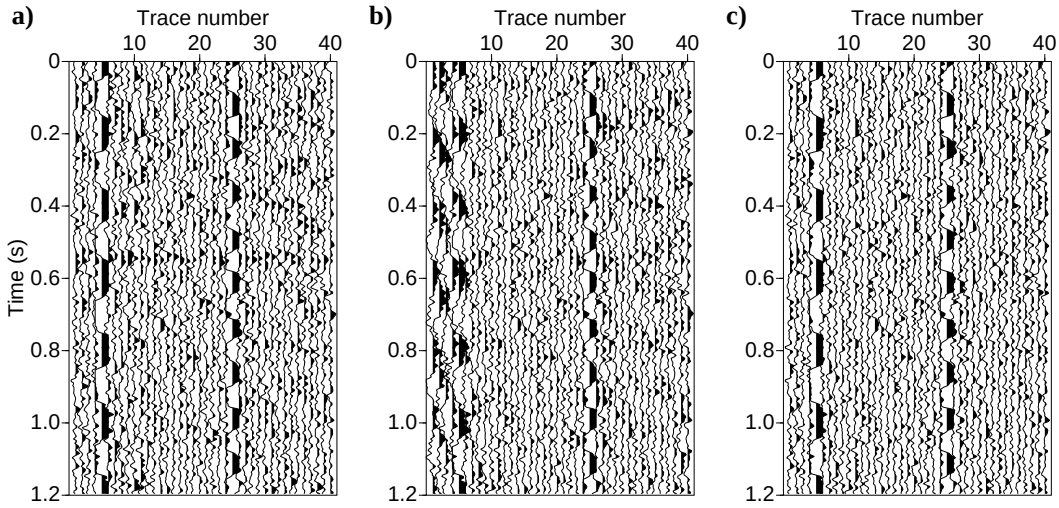

Figure 3.9: Error panels of f-x deconvolution (a), SSA (b), and robust SSA (c).

robust SSA methods is set to be 2. The reason for choosing this rank is that the data are only a two dimensional data set and also the events in each small window are relatively flat. In robust SSA, the external iterations (for updating weights) is 10 and number of internal iterations (for alternating minimization) is 5. We set the tuning constant α for bisquare

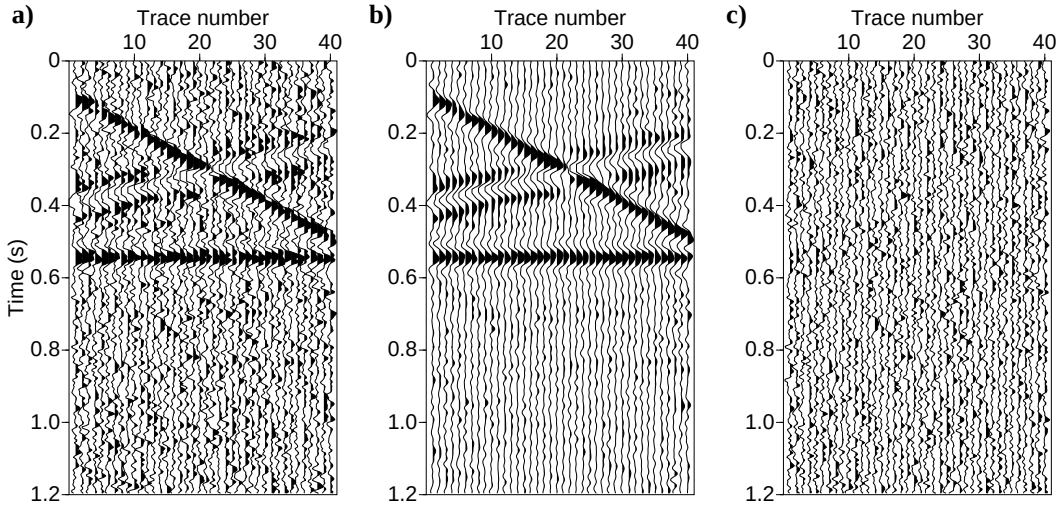

Figure 3.10: (a) Data corrupted with only Gaussian noise. (b) Data after classical SSA filtering. (c) Error panel of classical SSA.

estimator as 3.3. We set the length of f-x prediction filter as 6 and the trade-off parameter as 0.001. We use the same parameters for the whole data set. Again, we compare the performance of f-x deconvolution, SSA and robust SSA on noise attenuation. To compare the results of three methods objectively, all the image plots (Figure 3.11(a), Figure 3.12, Figure 3.13) have been clipped to the same value. The wiggle plots corresponding to the left rectangular window (Figure 3.11(b), Figure 3.14) have been clipped to the same value. The error panels (Figure 3.15) have been clipped to another same value to better compare the details of the estimated noise of the three methods. Similarly, wiggle plots corresponding to the right rectangular window (Figure 3.11(c), Figure 3.16) have been clipped to the same value. The error panels (Figure 3.17) have been clipped to another same value. The results of the three methods applied on the whole data set are shown in Figure 3.12. Robust SSA suppresses much more high-amplitude erratic noise than f-x deconvolution and classical SSA. The comparison of error panels (Figure 3.13) shows that the f-x deconvolution leaks more signal energy into the noise section than robust SSA. We show the zoomed results for window to the left in Figure 3.11(a) as Figure 3.14. The results for window to the right of Figure 3.11(a) are shown in Figure 3.16. Robust SSA is more effective than f-x deconvolution and SSA. We can find that there are more details for deep layers appear after robust SSA filtering than the other two. The error panels (Figure 3.15, Figure 3.17) of the two particular windows highlighted in Figure 3.11(a) do not show obvious energy leakage of the signal. Note that the using of patching technique makes the estimation of f-x filters or least-squares singular vectors in one particular patch (e.g. shallower windows) not

influenced by the erratic noise in other patches (e.g. deeper windows). This causes the fact that the performance of f-x deconvolution and classical SSA in the shallow part of data is not that bad.

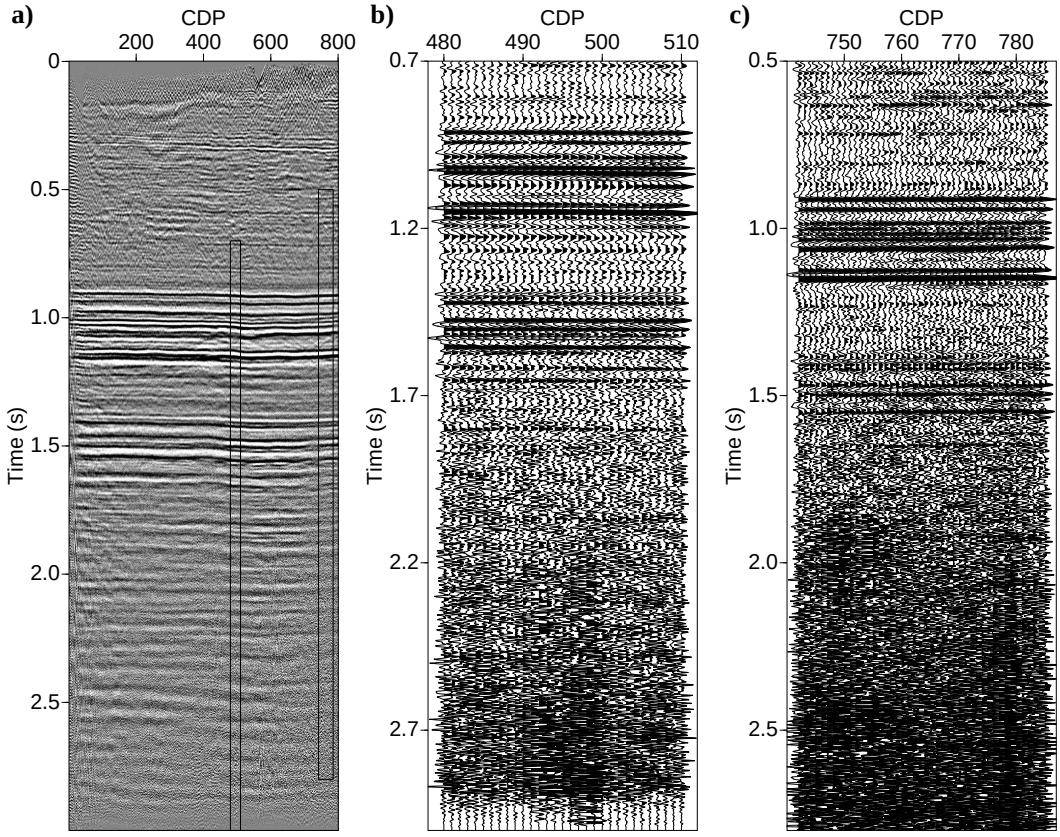

Figure 3.11: Poststack field data. (a) The whole data set. (b) The data in the left rectangular window. (c) The data in the right rectangular window.

Alaska data set

We also test the algorithms on a data set which has more complex geological structure. It is a poststack data section from a 2-D land survey in Alaska (Figure 3.18(a)). There is low amplitude random noise in the data. High-amplitude sinusoids with various frequencies (1-60 Hz) and amplitudes (maximum amplitude is as large as 4 times the signal) are added to the data to simulate erratic noise. In total, ten percent of the traces are corrupted with this kind of noise. Again, the patching technique is used for processing here. Each small window is composed of 36 traces with 16 overlapping traces between adjacent windows in spatial

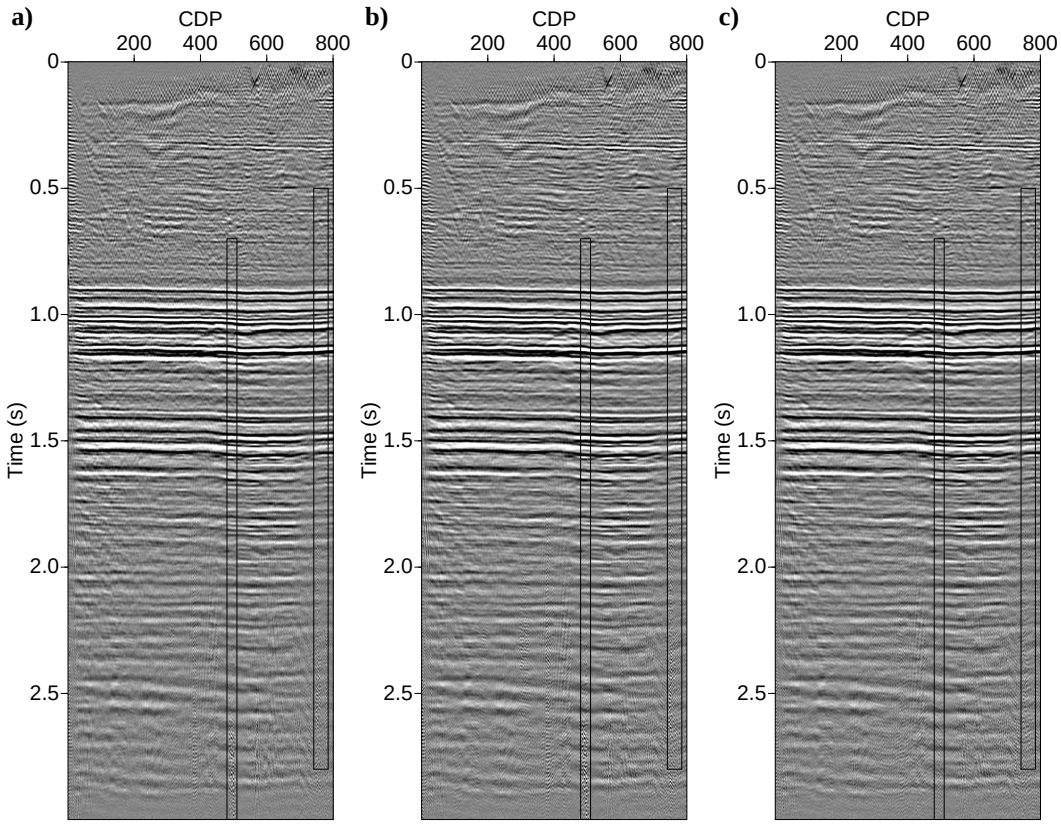

Figure 3.12: The comparison of the results of three different methods. (a) Data after f-x deconvolution filtering. (b) Data after classical SSA filtering. (c) Data after robust SSA filtering.

direction and, 300 time samples (0.6 s) with 75 overlapping samples (0.15 s) in temporal direction. The processing frequency band for all the three methods ranges from 1 to 60 Hz. The size of subspace of the reconstructed data is chosen to be $K = 3$ in both robust SSA and SSA. In robust SSA, the outer iteration (for reweighting) number is set to be 10 and the inner iteration (for alternating minimization) number is set to be 5. The tuning constant α for bisquare estimator used here is 4.2. The prediction length and trade-off parameter of f-x deconvolution are 8 and 0.001, respectively. The results of the f-x deconvolution, SSA and robust SSA are compared. All the image plots and wiggle plots are clipped to the same value to make the results are better compared. Robust SSA (Figure 3.18 (d)) thoroughly removes the high amplitude erratic noise. While, both f-x deconvolution (Figure 3.18 (b)) and SSA (Figure 3.18 (c)) are not efficient for erratic noise attenuation. Figure 3.19 show the error panels of the three methods. Still, we can find that robust SSA preserves signal

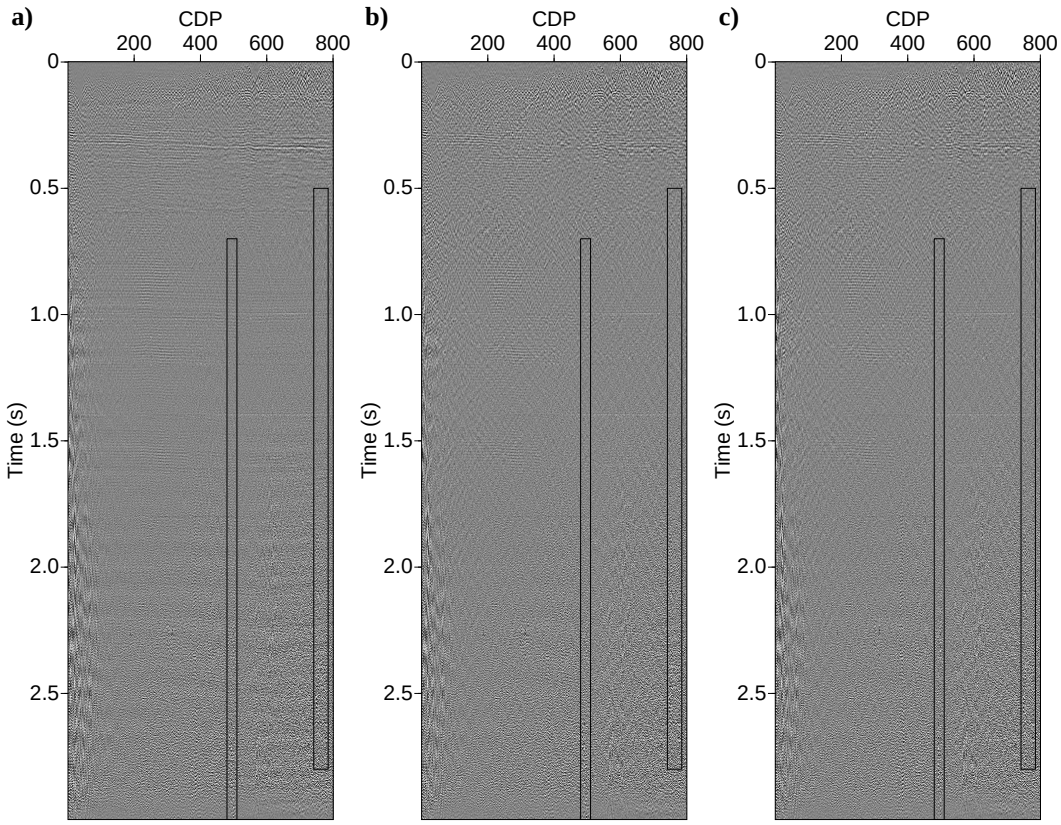

Figure 3.13: The comparison of error panels of three different methods. Error panels of f-x deconvolution (a), SSA (b), and robust SSA (c).

and f-x deconvolution damages signal. To show the details better, we display the zoomed results and error panels in the rectangular window as Figure 3.20.

3.6 Summary

In this Chapter, we first reviewed the basic concepts for robust statistics with an emphasis on M-estimators. We propose a robust version of the SSA method which can remove Gaussian and non-Gaussian (erratic) noise. The robust matrix factorization is used in the new method instead of the truncated SVD. A t-x domain synthetic example shows the advantage of robust matrix factorization over TSVD with the presence of non-Gaussian noise in the data matrix. Another synthetic example shows that the proposed robust SSA can remove Gaussian and erratic noise, while the least-squares minimization based methods SSA and

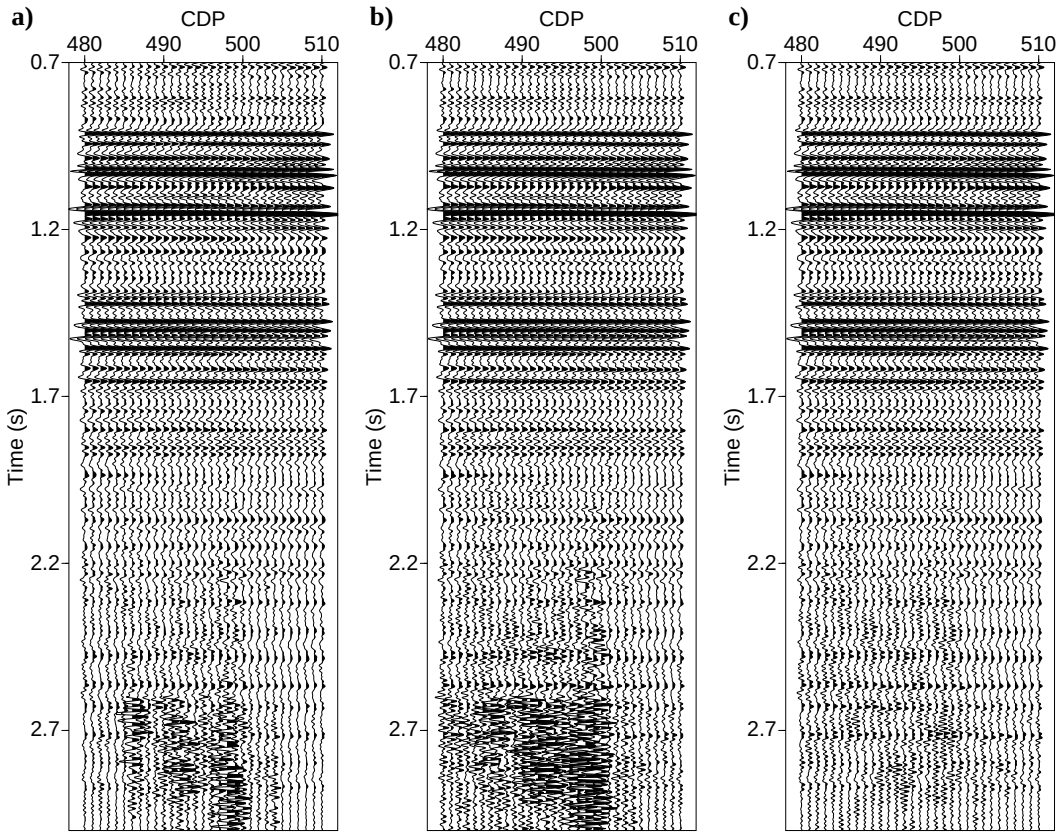

Figure 3.14: The comparison of results of the data in the left rectangular window by three different methods. (a) Data after f-x deconvolution filtering. (b) Data after classical SSA filtering. (c) Data after robust SSA filtering.

f-x deconvolution do not perform well. The field data examples from Western Canadian Sedimentary Basin and Alaska are used to analyze the performance of the new algorithm on real data. One possible concern is the computation cost of the robust algorithm. Computational time can be reduced by adopting windowing strategies to minimize the size of the Hankel matrices to factorize. Another strategy is to truncate the number of iterations of the alternating minimization algorithm and IRLS solvers in a way that an inexact factorization is estimated. We have noticed that an inexact factorization can yield better results than conventional non-robust Rank-reduction via the truncated SVD.

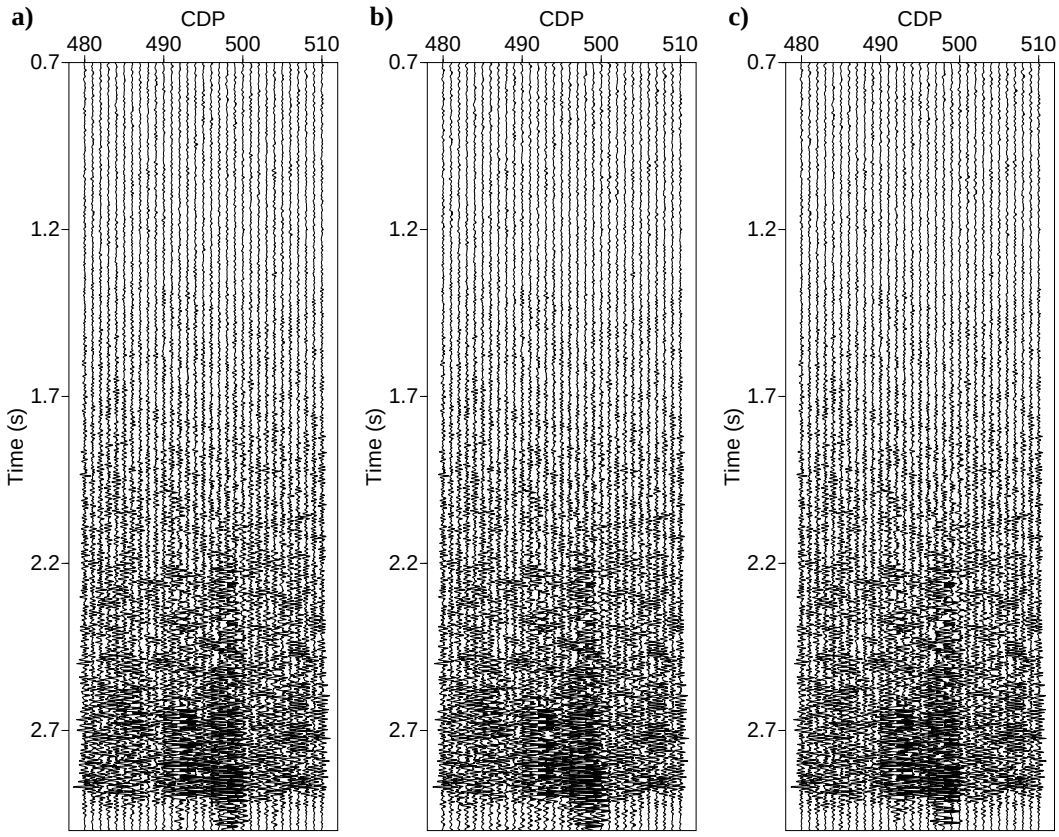

Figure 3.15: The comparison of error panels of three different methods in the left rectangular window. Error panels of f-x deconvolution (a), SSA (b), and robust SSA (c).

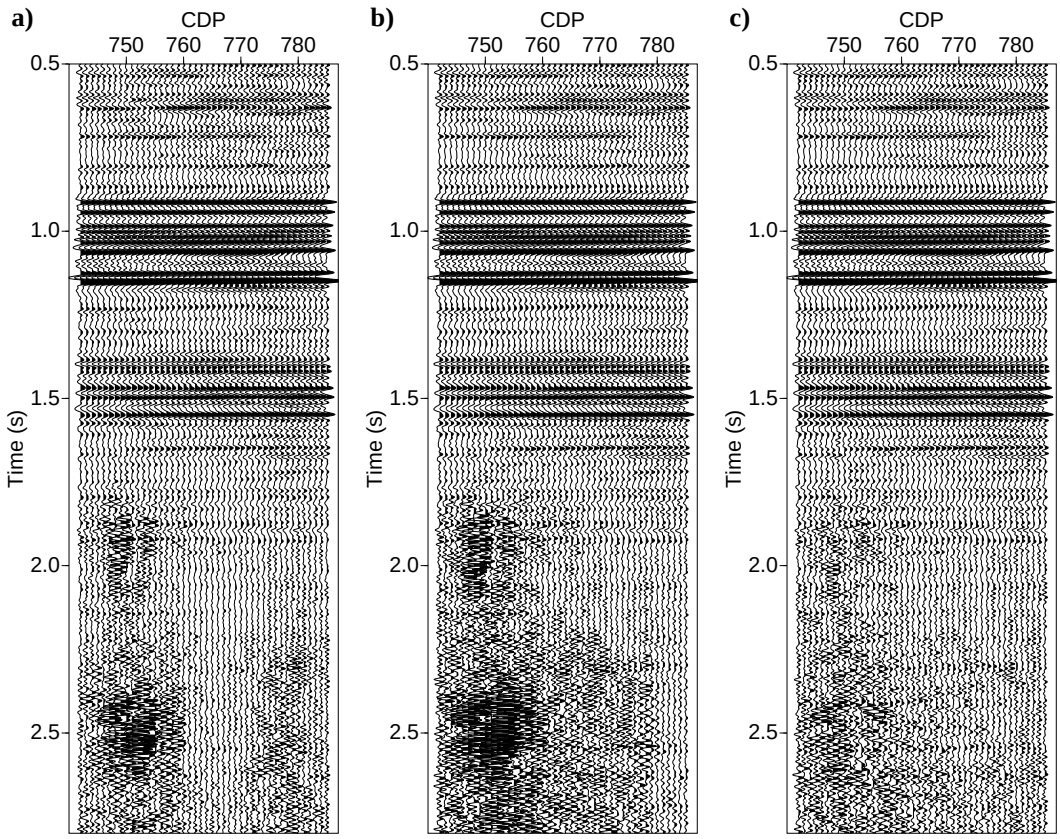

Figure 3.16: The comparison of results of the data in the right rectangular window by three different methods. (a) Data after f-x deconvolution filtering. (b) Data after classical SSA filtering. (c) Data after robust SSA filtering.

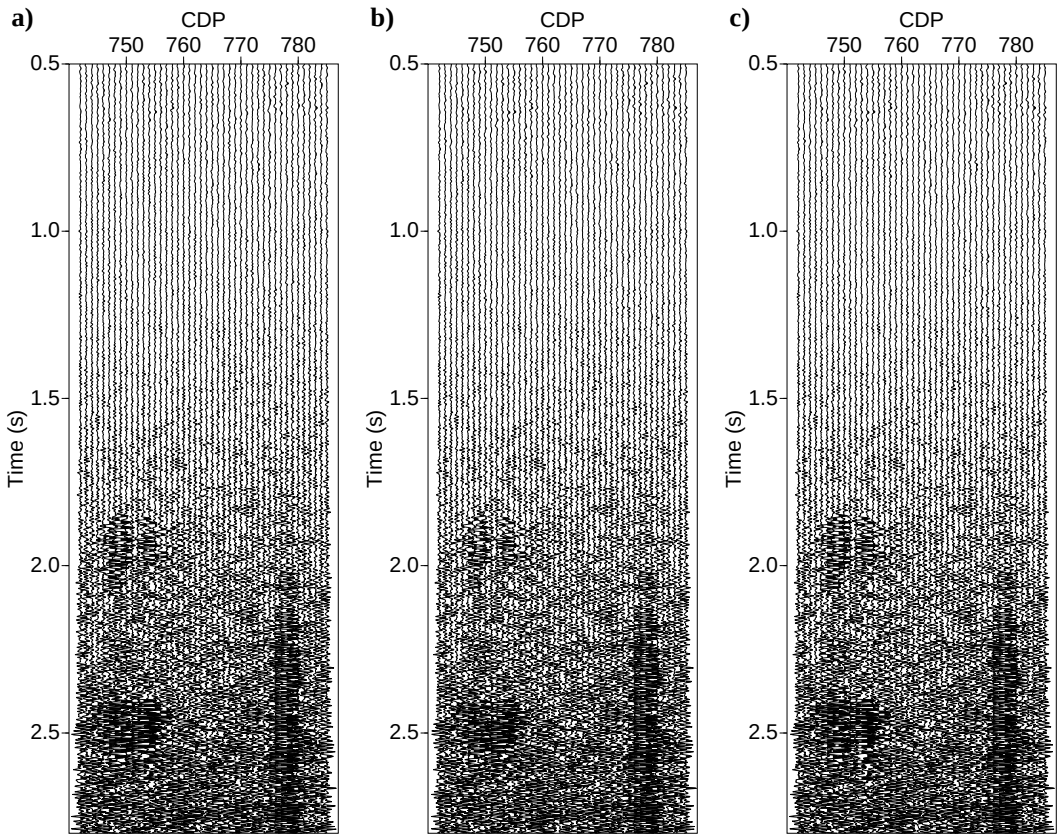

Figure 3.17: The comparison of error panels of three different methods in the right rectangular window. Error panels of f-x deconvolution (a), SSA (b), and robust SSA (c).

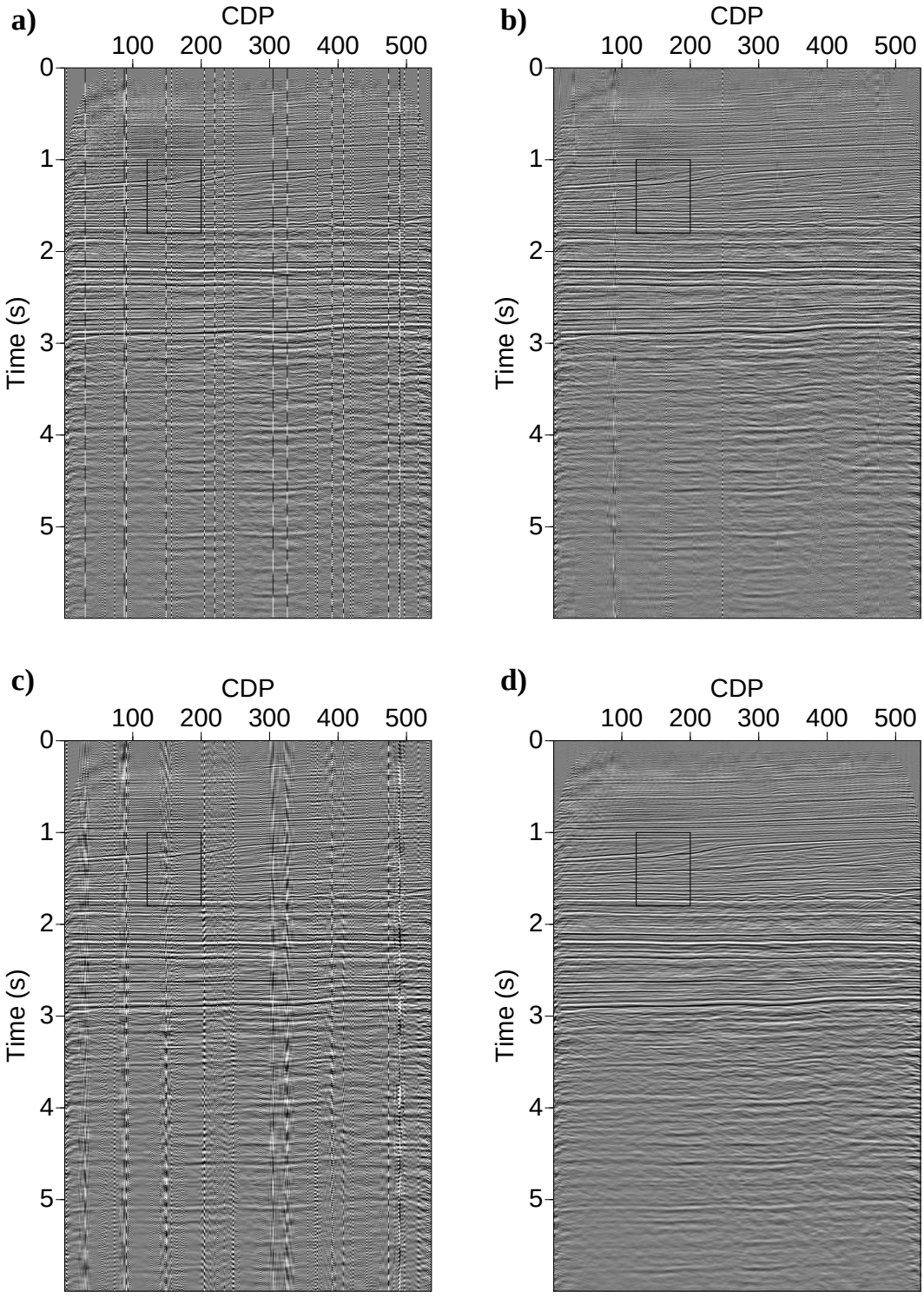

Figure 3.18: Field data example from Alaska. (a) Poststack data with erratic noise. (b) Data filtered by f-x deconvolution. (c) Data filtered by SSA. (d) Data filtered by robust SSA.

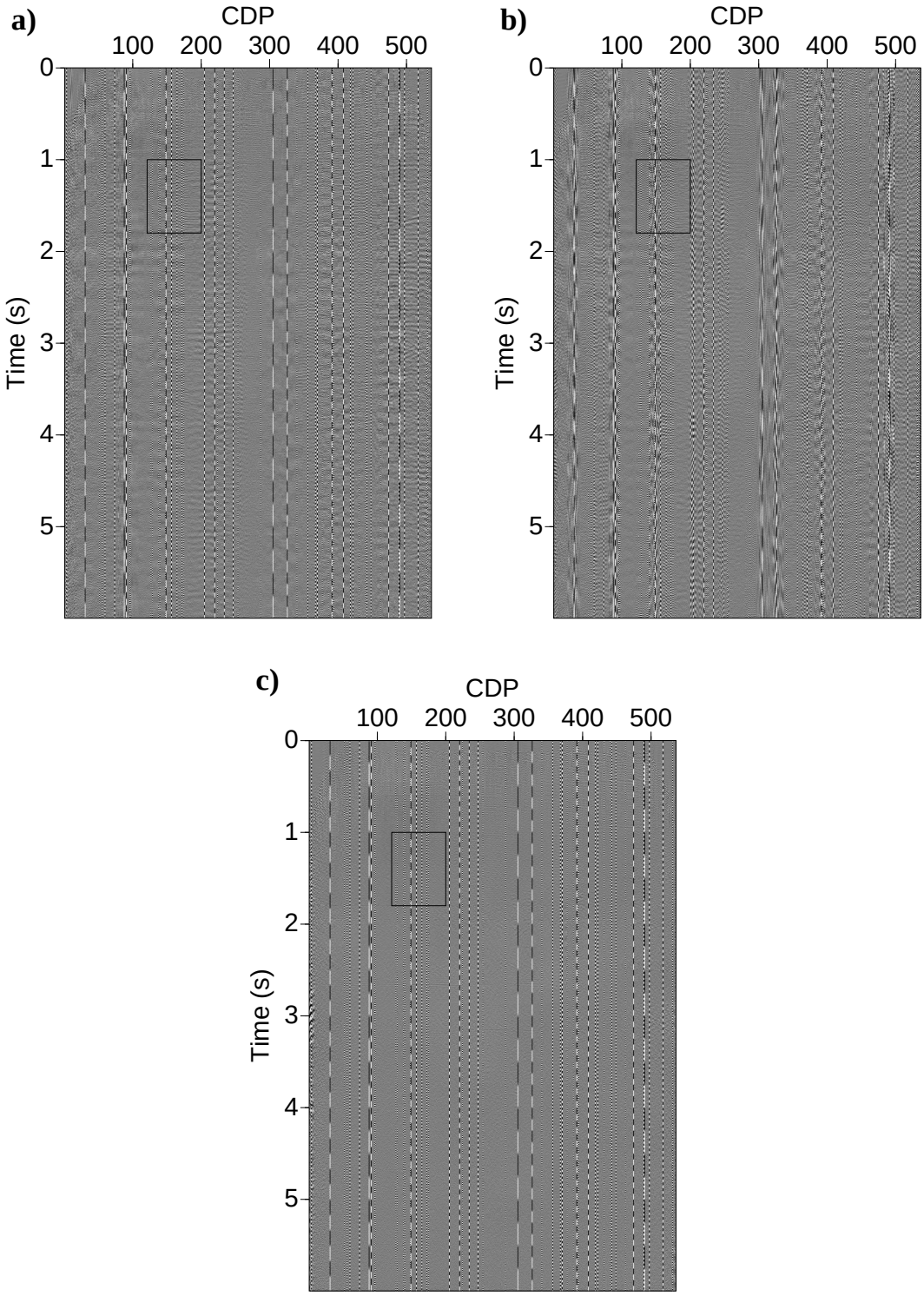

Figure 3.19: Error panels of (a) f-x deconvolution, (b) SSA, (c) robust SSA.

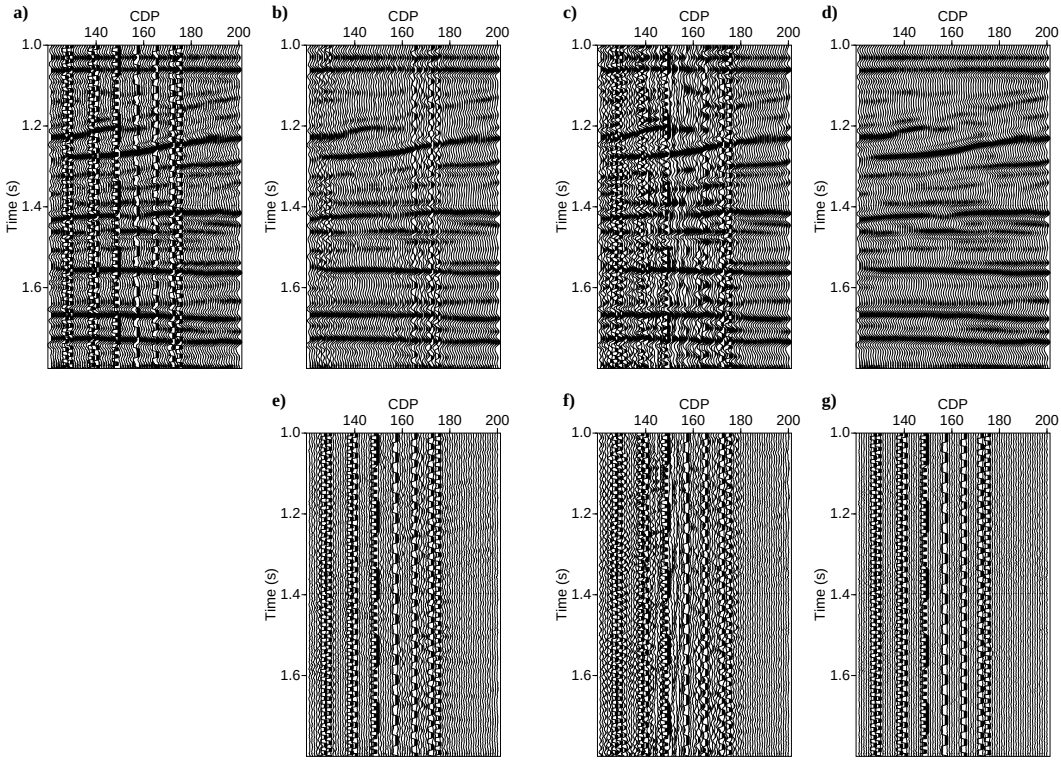

Figure 3.20: Zoomed sections correspond to the rectangular window. (a) Original data with erratic noise. (b) Data filtered by f-x deconvolution. (c) Data filtered by SSA. (d) Data filtered by robust SSA. (e) Error panel of f-x deconvolution. (f) Error panel of SSA. (g) Error panel of robust SSA.

CHAPTER 4

Matrix rank reduction approximated by nuclear-norm minimization

4.1 Introduction

As mentioned in the introduction, seismic data acquired in field may have too large spatial sampling interval, may have large gap between traces, or may be irregularly sampled in space. Cadzow/SSA based methods have been applied for irregularly decimated seismic data reconstruction (Trickett et al., 2010; Oropeza and Sacchi, 2011; Gao et al., 2013). In this case, the constructed Hankel matrix is not only perturbed by random Gaussian noise but also incomplete. Many elements of the matrix are missing in an irregular pattern. Trickett et al. (2010) proposed a Cadzow filtering based method for multidimensional trace interpolation. It applies an algorithm of matrix completion instead of the direct TSVD on the incomplete Hankel matrix. Oropeza and Sacchi (2011) proposed a seismic data reconstruction method similar with the projection onto convex sets (POCS) (Abma and Kabir, 2006) that replaces the frequency spectrum thresholding in POCS by MSSA filtering. The rank reduction is based on a randomized SVD instead of naive SVD for acceleration. Gao et al. (2013) extended the MSGEO-2013-0350SA reconstruction to 5D. The rank reduction is based on Lanczos bidiagonalization, and the Toeplitz matrix-vector multiplication is accelerated by the Fast Fourier Transform (FFT). These methods works for regular data grid with irregular missing pattern. They can not deal with aliasing problem. Naghizadeh and Sacchi (2013) proposed a MSSA/Cadzow based reconstruction algorithm for interpolating *regularly* sampled seismic data. It extract information from low frequencies to recover the regularly missing information at high frequencies.

The methods mentioned above do not consider erratic data (Claerbout and Muir, 1973).

The least-squares minimization performs poorly in this situation because its breakdown point is zero. Even one outlier will destroy the resulting fit. Other robust error criterions are required for the data fitting to robustify the inversion procedure. In this chapter, we propose a robust singular spectrum analysis method for removing the Gaussian, erratic noise and interpolating missing data simultaneously. First, the frequency domain data is embedded into a Hankel matrix, which contains dense Gaussian error, sparse outliers and missing elements. Then, a robust low rank approximation of this corrupted and incomplete matrix is achieved by solving a low-rank matrix recovery problem (Candès et al., 2009; Zhou et al., 2010). It minimizes a weighted combination of nuclear-norm, ℓ_1 norm and ℓ_2 norm terms. In this Robust SSA algorithm, the size of the reconstructed subspace is detected automatically. There are several fast first order algorithms that exist to solve the low-rank matrix recovery problem. We choose the alternating splitting augmented Lagrangian method (Tao and Yuan, 2011) to retrieve the low-rank component. Our 2D synthetic examples show that the new robust singular spectrum analysis method performs well.

4.2 Theory

4.2.1 Notation

Several notations are introduced here. The Frobenius norm of matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ is $||\mathbf{X}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}|^2}$, which is the ℓ_2 norm of the vector of singular values and the ℓ_1 norm is $||\mathbf{X}||_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}|$. $\|\mathbf{X}\|_0$ is the ℓ_0 norm indicating the number of non-zero elements of matrix $\mathbf{X}$. The nuclear-norm of a rank r matrix $\mathbf{X}$ is defined as the sum of the singular values $||\mathbf{X}||_* = \sum_{k=1}^{r}\sigma_k$, with $\mathbf{X} = \sum_{k=1}^{r}\sigma_k\mathbf{u}_k\mathbf{v}_k^H$ is the singular value decomposition of $\mathbf{X}$. Nuclear-norm is the ℓ_1 norm of the vector of singular values. The nuclear-norm minimization is the tightest convex relaxation of the rank minimization problem, which is similar as the fact that ℓ_1 norm minimization is the tightest convex relaxation of the ℓ_0 norm minimization problem (Candès and Plan, 2009). Nuclear-norm measures the 2-D sparsity of the matrix. The inner product of two matrix in Euclidean space is denoted as $\langle\mathbf{X},\mathbf{Y}\rangle = \text{trace}(\mathbf{X}^H\mathbf{Y})$. In this chapter, the linear mapping operators are denoted by calligraphic letters, e.g. $\mathcal{O}(\mathbf{X})$.

4.2.2 Singular spectrum analysis

Remember that the SSA algorithm is summarized as, for each frequency slice, $\hat{\mathbf{D}} = \mathcal{A}(\mathcal{R}(\mathcal{H}(\mathbf{D})))$, where $\mathcal{A}$ is anti-diagonal averaging operator, $\mathcal{R}$ is the truncated SVD filtering operator and $\mathcal{H}$ is the Hankel operator. If we consider missing observations, $\tilde{\mathbf{D}}$ denotes the completely observed data and Γ denotes the support of the index set of observations in $\mathbf{D}$, i.e. only

the entries $\{\tilde{\mathbf{D}}_i, i \in \Gamma\}$ are recorded. Then, the $\mathcal{P}_\Gamma$ indicates the projection onto the space of vectors supported on Γ. We call it the sampling operator, $\mathbf{D} = \mathcal{P}_\Gamma(\tilde{\mathbf{D}})$. Most of the singular spectrum analysis methods apply least-squares minimization in the rank reduction, e.g. the truncated SVD (Trickett, 2008), the method based on Random SVD (Oropeza and Sacchi, 2011) and the method based on Lanczos bidiagonalization (Gao et al., 2013). It is well known that the least-squares process is not robust. Even one single outlier will result in an erroreous solution. The erratic noise in the seismic data need robust algorithm to suppress. The extraction of low-rank component from the incomplete and corrupted matrix is achieved by solving a low-rank matrix recovery problem (Candès et al., 2009).

4.2.3 Low-rank matrix recovery

We utilize a robust low rank approximation other than the truncated SVD to recover the low-rank component from the partly observed and grossly corrupted matrix $\mathbf{M}$. Our Robust SSA algorithm is summarized as follows

$$\begin{aligned} &For\ each\ frequency\ slice: \\ &\hat{\mathbf{D}} = \mathcal{A}(\mathcal{M}_{\mathcal{R}}(\mathcal{H}(\mathbf{D}))), \end{aligned} \tag{4.1}$$

where $\mathcal{M}_{\mathcal{R}}$ denotes the operator for solving the low-rank matrix recovery problem, $\hat{\mathbf{D}}$ is the reconstructed frequency slice. Let $\tilde{\mathbf{M}}$ denotes the Hankel matrix from the completely sampled data $\tilde{\mathbf{D}}$, i.e. $\tilde{\mathbf{M}} = \mathcal{H}(\tilde{\mathbf{D}})$. $\mathcal{H}$ is a Hankel operator. Matrix $\tilde{\mathbf{M}}$ can be decomposed to three components as:

$$\tilde{\mathbf{M}} = \mathbf{L} + \mathbf{S} + \mathbf{N}, \tag{4.2}$$

where $\mathbf{L}$ is the low rank matrix embedded from the f-x signal, $\mathbf{S}$ is a sparse matrix corresponding to impulsive noise and $\mathbf{N}$ is a dense perturbation matrix representing Gaussian noise. If there is only impulsive noise ($\mathbf{N} = \mathbf{0}$), the problem is recovering low-rank component from completely observed but impulsive noise corrupted matrix. It is also referred to as the robust principal component analysis problem (Candès et al., 2009). The solution can be obtained from solving the matrix rank minimization problem

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{S}}\ \operatorname{rank}(\mathbf{L}) + \gamma\|\mathbf{S}\|_0, \\ &\text{subject to}\ \ \tilde{\mathbf{M}} = \mathbf{L} + \mathbf{S}, \end{aligned} \tag{4.3}$$

where γ is a trade-off parameter balancing the low-rank of $\mathbf{L}$ and the sparsity of $\mathbf{S}$. Unfortunately, both rank function and ℓ_0 function are non-convex. This low-rank matrix recovery problem is NP-hard because the combinational nature of rank function and ℓ_0 norm. Candès et al. (2009) proved that the low-rank and sparse component can be solved by a relaxed

convex program, the Principal Component Pursuit

$$\begin{aligned} \min_{\mathbf{L},\mathbf{S}} \ & \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1, \\ \text{subject to} \ & \tilde{\mathbf{M}} = \mathbf{L} + \mathbf{S}, \end{aligned} \tag{4.4}$$

the rank minimization is relaxed to the nuclear norm minimization and the ℓ_0 norm minimization is relaxed to ℓ_1 norm minimization, λ is a trade-off parameter to balance the sparsity and low rank. The nuclear-norm guarantees the low rank of component $\mathbf{L}$, the ℓ_1 norm induces the sparsity of component $\mathbf{S}$, i.e. the robustness of this recovery algorithm with respect to outliers. However, the seismic data are usually corrupted with dense Gaussian noise, i.e. $\mathbf{N} \neq \mathbf{0}$. The problem changes to the recovery of low-rank matrix from a matrix that is corrupted with sparse impulsive noise and small dense Gaussian noise. Zhou et al. (2010) proved that this problem can be solved by the convex program Stable Principal Component Pursuit

$$\begin{aligned} \min_{\mathbf{L},\mathbf{S}} \quad & ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1, \\ \text{subject to} \quad & ||\tilde{\mathbf{M}} - \mathbf{L} - \mathbf{S}||_F \leq \delta, \end{aligned} \tag{4.5}$$

where the Frobenius norm induce the stability towards Gaussian noise perturbation, δ is the Gaussian noise level.

When the observed data is under sampled ($\mathbf{D} = \mathcal{P}_\Gamma(\tilde{\mathbf{D}})$), the constructed Hankel matrix has missing elements. Suppose that Ω is the support of nonzero elements of matrix $\mathbf{M}$, i.e. entries $\{\tilde{\mathbf{M}}_{ij}, (i,j) \in \Omega\}$ are the recorded elements. $\mathcal{P}_\Omega$ denotes the sampling operator acted on the complete observed matrix $\tilde{\mathbf{M}}$, i.e. $\mathbf{M} = \mathcal{P}_\Omega(\tilde{\mathbf{M}})$. Now, the problem changes to recover the low-rank component from a fraction of the grossly corrupted and randomly perturbed entries of the matrix. It can be expressed as

$$\begin{aligned} \min_{\mathbf{L},\mathbf{S}} \quad & ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1, \\ \text{subject to} \quad & ||\mathcal{P}_\Omega(\tilde{\mathbf{M}} - \mathbf{L} - \mathbf{S})||_F \leq \delta. \end{aligned} \tag{4.6}$$

When $\mathcal{P}_\Omega$ is the identity operator, 4.6 recovers low-rank matrix from completely observed, Gaussian and impulsive noise corrupted matrix 4.5. When $\mathcal{P}_\Omega$ is the identity operator and $\sigma = 0$, it recovers low-rank matrix from completely observed, impulsive noise corrupted matrix 4.4. When $\sigma = 0$, $\lambda = 0$ and $\mathbf{S} = \mathbf{0}$, it solves the matrix completion problem (Candès and Recht, 2009). When $\lambda = 0$ and $\mathbf{S} = \mathbf{0}$, it solves the matrix completion with Gaussian noise problem (Candès and Plan, 2009).

The penalized version of convex program 4.6 is

$$\min_{\mathbf{L},\mathbf{S}} \quad ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 + \frac{1}{2\mu}||\mathcal{P}_\Omega(\tilde{\mathbf{M}} - \mathbf{L} - \mathbf{S})||_F^2, \tag{4.7}$$

where μ is a parameter balancing the Frobenius norm and the other two terms in cost function. The cost function is a non-smooth convex function, several first order algorithms exists to solve it. The augmented Lagrangian based Method (Lin et al., 2010; Tao and Yuan, 2011) is adopted in this thesis.

4.2.4 Augmented Lagrangian method

Here, we simply describe the general augmented Lagrangian method designed for solving the following equality constrained optimization problem (Bertsekas, 1982)

$$\begin{aligned} &\min_{\mathbf{x}} \quad f(\mathbf{x}), \\ &\text{subject to} \quad g(\mathbf{x}) = \mathbf{0}, \end{aligned} \tag{4.8}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are given functions. The augmented Lagrangian is defined as

$$\mathcal{L}_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \beta) = f(\mathbf{x}) + \langle \mathbf{y}, g(\mathbf{x}) \rangle + \frac{\beta}{2} ||g(\mathbf{x})||_F^2. \tag{4.9}$$

where $\mathbf{y}$ is the dual variable or Lagrange multiplier, and $\beta > 0$ is the penalty parameter. Note, the augmented Lagrangian has an additional term $\frac{\beta}{2}||g(\mathbf{x})||_F^2$ comparing with the standard Lagrangian. The algorithm of general augmented Lagrangian method is summarized as

Algorithm 2 General augmented Lagrangian method

1: $\rho \geq 1$
2: **while** not converge **do**
3: Compute $\mathbf{x}^{k+1}$: $\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}}\ \mathcal{L}_{\mathcal{A}}(\mathbf{x}, \mathbf{y}^k, \beta^k)$.
4: Update $\mathbf{y}^{k+1}$: $\mathbf{y}^{k+1} = \mathbf{y}^k + \beta^k g(\mathbf{x}^{k+1})$.
5: Update β^{k+1}: $\beta^{k+1} = \rho\beta^k$.
6: **end while**

The update of the Lagrange multiplier $\mathbf{y}^{k+1}$ is obtained from the maximizing the dual function. This is the so-called gradient ascent update. ρ is a parameter for updating the penalty parameter β in each iteration.

To use the augmented Lagrangian method, Equation 4.7 is firstly transformed to the equality constrained optimization problem (Tao and Yuan, 2011). Recall that $\mathbf{M} = \mathcal{P}_\Omega(\tilde{\mathbf{M}})$ is the real observations, a part of $\tilde{\mathbf{M}}$ is unknown in advance. A new variable matrix is introduced $\mathbf{Z} = \mathbf{M} - \mathbf{L} - \mathbf{S} = \mathcal{P}_\Omega(\tilde{\mathbf{M}}) - \mathbf{L} - \mathbf{S}$. Therefore, $\mathcal{P}_\Omega(\mathbf{Z}) = \mathcal{P}_\Omega(\mathcal{P}_\Omega(\tilde{\mathbf{M}}) - \mathbf{L} - \mathbf{S}) = \mathcal{P}_\Omega(\tilde{\mathbf{M}} - \mathbf{L} - \mathbf{S})$.

Equation 4.7 is modified as

$$\begin{aligned} \min_{\mathbf{L},\mathbf{S},\mathbf{Z}} \quad & ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 + \frac{1}{2\mu}||\mathcal{P}_\Omega(\mathbf{Z})||_F^2, \\ \text{subject to} \quad & \mathbf{L}+\mathbf{S}+\mathbf{Z} = \mathbf{M}. \end{aligned} \tag{4.10}$$

The elements outside the set Ω in $\mathbf{L}+\mathbf{S}$ are compensated by corresponding elements in $\mathbf{Z}$. Then the augmented Lagrangian function of Problem 4.10 is

$$\mathcal{L}_\mathcal{A}(\mathbf{L},\mathbf{S},\mathbf{Z},\mathbf{Y},\beta) = ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 + \frac{1}{2\mu}||\mathcal{P}_\Omega(\mathbf{Z})||_F^2 - \langle \mathbf{Y}, \mathbf{L}+\mathbf{S}+\mathbf{Z}-\mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L}+\mathbf{S}+\mathbf{Z}-\mathbf{M}||_F^2, \tag{4.11}$$

where $\mathbf{Y}$ is the Lagrange multiplier, β is the penalty parameter. The general augmented Lagrangian method optimizes all the variables together ($\mathbf{L}^{k+1}$, $\mathbf{S}^{k+1}$, $\mathbf{Z}^{k+1}$) at step 3 in Algorithm 2 and then update the Lagrange Multiplier $\mathbf{Y}^{k+1}$ and penalize parameter β^{k+1}. Tao and Yuan (2011) further explore the separable structure of the cost function and constraint in 4.11 and propose a more efficient alternating splitting augmented Lagrangian method (ASALM), which compute the three components separately, i.e. $\mathbf{L}^{k+1}$, $\mathbf{S}^{k+1}$ and then $\mathbf{Z}^{k+1}$. The idea of splitting is similar with the inexact augmented Lagrangian method (Lin et al., 2010) and the alternating direction method (ADM) (Yuan and Yang, 2009). The three variables are updated via solving three sub-problems

$$\begin{aligned} \mathbf{Z}^{k+1} &:= \arg\min_{\mathbf{Z}} \ \frac{1}{2\mu}||\mathcal{P}_\Omega(\mathbf{Z})||_F^2 - \langle \mathbf{Y}^k, \mathbf{L}^k+\mathbf{S}^k+\mathbf{Z}-\mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L}^k+\mathbf{S}^k+\mathbf{Z}-\mathbf{M}||_F^2, \\ \mathbf{S}^{k+1} &:= \arg\min_{\mathbf{S}} \ \lambda||\mathbf{S}||_1 - \langle \mathbf{Y}^k, \mathbf{L}^k+\mathbf{S}+\mathbf{Z}^{k+1}-\mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L}^k+\mathbf{S}+\mathbf{Z}^{k+1}-\mathbf{M}||_F^2, \\ \mathbf{L}^{k+1} &:= \arg\min_{\mathbf{L}} \ ||\mathbf{L}||_* - \langle \mathbf{Y}^k, \mathbf{L}+\mathbf{S}^{k+1}+\mathbf{Z}^{k+1}-\mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L}+\mathbf{S}^{k+1}+\mathbf{Z}^{k+1}-\mathbf{M}||_F^2, \\ \mathbf{Y}^{k+1} &:= \mathbf{Y}^k - \beta(\mathbf{L}^{k+1}+\mathbf{S}^{k+1}+\mathbf{Z}^{k+1}-\mathbf{M}), \end{aligned} \tag{4.12}$$

where the penalize parameter β is fixed in the ASALM method. The most important benefit is that each subproblem has closed form solution. They can be obtained by setting the gradient (Frobenius norm) or subgradient (ℓ_1 norm, nuclear-norm) to zero. It's easy to see that the update in Equation 4.12 is the same as the following problems (Tao and Yuan, 2011)

$$\begin{aligned} \mathbf{Z}^{k+1} &:= \arg\min_{\mathbf{Z}} \ \frac{1}{2\mu}||\mathcal{P}_\Omega(\mathbf{Z})||_F^2 + \frac{\beta}{2}||\mathbf{L}^k+\mathbf{S}^k+\mathbf{Z}-\frac{1}{\beta}\mathbf{Y}^k-\mathbf{M}||_F^2, \\ \mathbf{S}^{k+1} &:= \arg\min_{\mathbf{S}} \ \lambda||\mathbf{S}||_1 + \frac{\beta}{2}||\mathbf{L}^k+\mathbf{S}+\mathbf{Z}^{k+1}-\frac{1}{\beta}\mathbf{Y}^k-\mathbf{M}||_F^2, \\ \mathbf{L}^{k+1} &:= \arg\min_{\mathbf{L}} \ ||\mathbf{L}||_* + \frac{\beta}{2}||\mathbf{L}+\mathbf{S}^{k+1}+\mathbf{Z}^{k+1}-\frac{1}{\beta}\mathbf{Y}^k-\mathbf{M}||_F^2, \\ \mathbf{Y}^{k+1} &:= \mathbf{Y}^k - \beta(\mathbf{L}^{k+1}+\mathbf{S}^{k+1}+\mathbf{Z}^{k+1}-\mathbf{M}), \end{aligned} \tag{4.13}$$

where the update of $\mathbf{S}^{k+1}$ and $\mathbf{L}^{k+1}$ can be obtained from two well-known *shrinkage operators*. Usually, the soft shrinkage operator (Chen et al., 1998) $\mathcal{S}_\tau : \mathbb{R} \to \mathbb{R}$ is defined as $\mathcal{S}_\tau(x) = \text{sgn}(x)\max(|x| - \tau, 0)$. Because we apply the method in frequency domain, we use the soft shrinkage operator defined in the complex domain (Sardy, 2000): $\mathcal{S}_\tau : \mathbb{C} \to \mathbb{C}$ is defined as $\mathcal{S}_\tau(x) = \frac{x}{|x|}\max(|x| - \tau, 0)$. It is extended to the case of matrices

$$(\mathcal{S}_\tau(\mathbf{X}))_{ij} := \frac{\mathbf{X}_{ij}}{|\mathbf{X}_{ij}|}\max(|\mathbf{X}_{ij}| - \tau, 0), \qquad \mathbf{X} \in \mathbb{C}^{m\times n}. \tag{4.14}$$

$\mathcal{S}_\tau(\mathbf{X})$ is the solution of the following minimization problem

$$\min_{\mathbf{Y}} \quad \tau||\mathbf{Y}||_1 + \frac{1}{2}||\mathbf{Y} - \mathbf{X}||_F^2. \tag{4.15}$$

Suppose the rank r matrix $\mathbf{X} \in \mathbb{C}^{m\times n}$ has the following singular value decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{4.16}$$

where $\mathbf{U} \in \mathbb{C}^{m\times r}$ is the matrix containing left singular vectors, $\mathbf{V} \in \mathbb{C}^{n\times r}$ is the matrix containing right singular vectors and $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\}$ is the matrix containing singular values. The singular value shrinkage operator $\mathcal{D}_\tau$ applied on $\mathbf{X}$ is defined as following (Cai et al., 2008)

$$\mathcal{D}_\tau(\mathbf{X}) = \mathbf{U}\mathcal{S}_\tau(\mathbf{\Sigma})\mathbf{V}^H. \tag{4.17}$$

$\mathcal{D}_\tau(\mathbf{X})$ is the solution of the following minimization problem

$$\min_{\mathbf{Y}} \quad \tau||\mathbf{Y}||_* + \frac{1}{2}||\mathbf{Y} - \mathbf{X}||_F^2. \tag{4.18}$$

So the solutions of Equation 4.13 can be expressed using the above shrinkage operator and singular value shrinkage operator. The alternating splitting augmented Lagrangian method (ASALM) (Tao and Yuan, 2011) is summarized as following Algorithm 3

The proof of convergence of the ASALM algorithm is given in Tao and Yuan (2011). The dominant cost of the algorithm is given by the cost of the singular value shrinkage operator.

4.2.5 Parameter Selection and Stopping Criterion

Candès et al. (2009) and Zhou et al. (2010) proved that the selection of $\lambda = 1/\sqrt{\max(m, n)}$ can guarantee good recovery result in the (Stable) Principal Component Pursuit program. It is fixed in the algorithm. μ and β are tuning parameters that depend on the data. Tao

Algorithm 3 Alternating Splitting Augmented Lagrangian Method (ASALM) for Low Rank Matrix Recovery

1: **Initialization**: $\mathbf{L}^0 = \mathbf{0}$, $\mathbf{S}^0 = \mathbf{0}$, $\mathbf{Y}^0 = \mathbf{0}$, $\lambda = 1/\sqrt{\max(m,n)}$, μ, β.
2: **while** not converge **do**
3: Compute $\mathbf{T}^k = \frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{L}^k - \mathbf{S}^k$.
4: Compute $\mathbf{Z}^{k+1}$:

$$\mathbf{Z}_{ij}^{k+1} = \begin{cases} \mathbf{T}_{ij}^k & \text{when } (i,j) \notin \Omega; \\ \dfrac{\mu\beta}{1+\mu\beta}\mathbf{T}_{ij}^k & \text{when } (i,j) \in \Omega. \end{cases}$$

5: Compute $\mathbf{S}^{k+1}$: $\mathbf{S}^{k+1} = \mathcal{S}_{\lambda/\beta}(\frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{L}^k - \mathbf{Z}^{k+1})$.
6: Compute $\mathbf{L}^{k+1}$: $\mathbf{L}^{k+1} = \mathcal{D}_{1/\beta}(\frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{S}^{k+1} - \mathbf{Z}^{k+1})$.
7: Update $\mathbf{Y}^{k+1}$: $\mathbf{Y}^{k+1} = \mathbf{Y}^k + \beta(\mathbf{M} - \mathbf{L}^{k+1} - \mathbf{S}^{k+1} - \mathbf{Z}^{k+1})$.
8: **end while**

and Yuan (2011) recommend to choose the parameters based on the following strategy

$$\begin{aligned} \mu &= \frac{1}{10}\sqrt{\min(m,n) + \sqrt{8\min(m,n)}}\sigma, \\ \beta &= \eta\frac{|\Omega|}{||\mathbf{M}||_1}, \end{aligned} \tag{4.19}$$

where σ is the standard derivation (SD) of Gaussian noise. The coefficient η depends on the percentage of the outliers in the matrix, $|\Omega|$ is the cardinality of the set Ω, i.e. the number of elements of set Ω. The stopping criterion is chosen to be (Tao and Yuan, 2011)

$$SP = \frac{||(\mathbf{L}^{k+1}, \mathbf{S}^{k+1}) - (\mathbf{L}^k, \mathbf{S}^k)||_F}{||(\mathbf{L}^k, \mathbf{S}^k)||_F + 1} \leq \xi\sigma, \tag{4.20}$$

where ξ is a coefficient which is tunable. The stopping criterion measures the change of the low-rank component and sparse component in two consecutive iterations.

4.3 Examples

We present two synthetic examples to test the proposed algorithm.

4.3.1 Synthetic Example 1

We first test a simple synthetic data that is incomplete and corrupted with large amplitude erratic noise (coherent in temporal direction). Because the outliers do not appear in all the frequencies, we analyze the behavior of the Algorithm 3 in different situations. For

the frequency slices which contain outliers and missing elements, it is actually a Gaussian noiseless ($\mathbf{N} = \mathbf{0}$) low-rank matrix recovery problem ($\sigma = 0$ in convex program 4.6).

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{S}} ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1, \\ &\text{subject to} \quad \mathcal{P}_\Omega(\mathbf{L}+\mathbf{S}) = \mathcal{P}_\Omega(\tilde{\mathbf{M}}) = \mathbf{M}. \end{aligned} \tag{4.21}$$

Similarly, with the replacement $\mathbf{L} + \mathbf{S} + \mathbf{Z} = \mathbf{M}$, it can be changed to

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{S}} ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1, \\ &\text{subject to} \quad \mathbf{L} + \mathbf{S} + \mathbf{Z} = \mathbf{M},\ \mathcal{P}_\Omega(\mathbf{Z}) = \mathbf{0}, \end{aligned} \tag{4.22}$$

In this case, the augmented Lagrangian is

$$\mathcal{L}_\mathcal{A}(\mathbf{L}, \mathbf{S}, \mathbf{Z}, \mathbf{Y}, \beta) = ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 - \langle \mathbf{Y}, \mathbf{L} + \mathbf{S} + \mathbf{Z} - \mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L} + \mathbf{S} + \mathbf{Z} - \mathbf{M}||_F^2, \tag{4.23}$$

with $\mathcal{P}_\Omega(\mathbf{Z}) = 0$. The algorithm for convex program 4.21 is a special case of Algorithm 3 with parameter $\mu = 0$.

For the frequency slices which do not contain outliers, they only contain missing elements. The constructed Hankel matrix is incomplete but noise free. The problem changes to a noiseless matrix completion problem

$$\begin{aligned} &\min_{\mathbf{L}} ||\mathbf{L}||_*, \\ &\text{subject to} \quad \mathcal{P}_\Omega(\mathbf{L}) = \mathcal{P}_\Omega(\tilde{\mathbf{M}}) = \mathbf{M}. \end{aligned} \tag{4.24}$$

with the replacement $\mathbf{L} + \mathbf{Z} = \mathbf{M}$, it is changed to

$$\begin{aligned} &\min_{\mathbf{L}} ||\mathbf{L}||_*, \\ &s.t. \quad \mathbf{L} + \mathbf{Z} = \mathbf{M}, \quad \mathcal{P}_\Omega(\mathbf{Z}) = \mathbf{0}. \end{aligned} \tag{4.25}$$

The augmented Lagrangian is

$$\mathcal{L}_\mathcal{A}(\mathbf{L}, \mathbf{Z}, \mathbf{Y}, \beta) = ||\mathbf{L}||_* - \langle \mathbf{Y}, \mathbf{L} + \mathbf{Z} - \mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L} + \mathbf{Z} - \mathbf{M}||_F^2, \tag{4.26}$$

with $\mathcal{P}_\Omega(\mathbf{Z}) = 0$. The algorithm for noiseless matrix completion problem is (Lin et al., 2010)

Algorithm 4 is a particular case of Algorithm 3 with $\mu = 0$ and $\mathbf{S} = \mathbf{0}$. With a good thresholding parameter λ/β in Algorithm 3, the elements in $\mathbf{S}$ does not contain the signal. That is to say, the matrix $\mathbf{S}$ will stay as zero matrix during the iterations. Then, it is

Algorithm 4 Alternating Splitting Augmented Lagrangian Method (ASALM) for Matrix Completion

1: **Initialization**: $\mathbf{L}^0 = \mathbf{0}$,$\mathbf{Y}^0 = \mathbf{0}$, β.
2: **while** not converge **do**
3: Compute $\mathbf{T}^k = \frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{L}^k$.
4: Compute $\mathbf{Z}^{k+1}$:

$$\mathbf{Z}_{ij}^{k+1} = \begin{cases} \mathbf{T}_{ij}^k & \text{when } (i,j) \notin \Omega; \\ 0 & \text{when } (i,j) \in \Omega. \end{cases}$$

5: Compute $\mathbf{L}^{k+1}$: $\mathbf{L}^{k+1} = \mathcal{D}_{1/\beta}(\frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{Z}^{k+1})$.
6: Update $\mathbf{Y}^{k+1}$: $\mathbf{Y}^{k+1} = \mathbf{Y}^k + \beta(\mathbf{M} - \mathbf{L}^{k+1} - \mathbf{Z}^{k+1})$.
7: **end while**

reasonable to use Algorithm 3 with setting $\mu = 0$ for all the frequency slices in this synthetic example.

The synthetic data has 80 traces with time sampling interval 0.004 s. There are 5 traces corrupted with large amplitude (1 to 7 times the amplitude of wavelet) time coherent noise and 50% traces randomly missing. Figure 4.1 (d) shows the noise free data, Figure 4.1 (a) shows the data with missing traces and corrupted with erratic noise. We use Algorithm 3 to recover the low-rank component $\mathbf{L}$ and then the reconstructed data $\hat{\mathbf{D}}$. The frequency band to be processed is 1-40 Hz. For parameter selection, we simply set $\mu = 0$ and $\beta = 0.15\frac{|\Omega|}{||\mathbf{M}||_1}$. The stopping criterion is chosen as: $SP \leq 10^{-7}$. Figure 4.1 (b) is the reconstructed result, Figure 4.1 (e) is the difference between noise free data and the reconstructed data. Figure 4.1 (c) is the recovered sparse component. We can see that the robust SSA via low-rank matrix recovering exactly reconstructed the signal and also the time coherent noise. We use the quality factor $Q = 10 \log \frac{||d^0||_F^2}{||d^0 - \hat{d}||_F^2}$ to evaluate the result. The $Q = 114$ in this synthetic example.

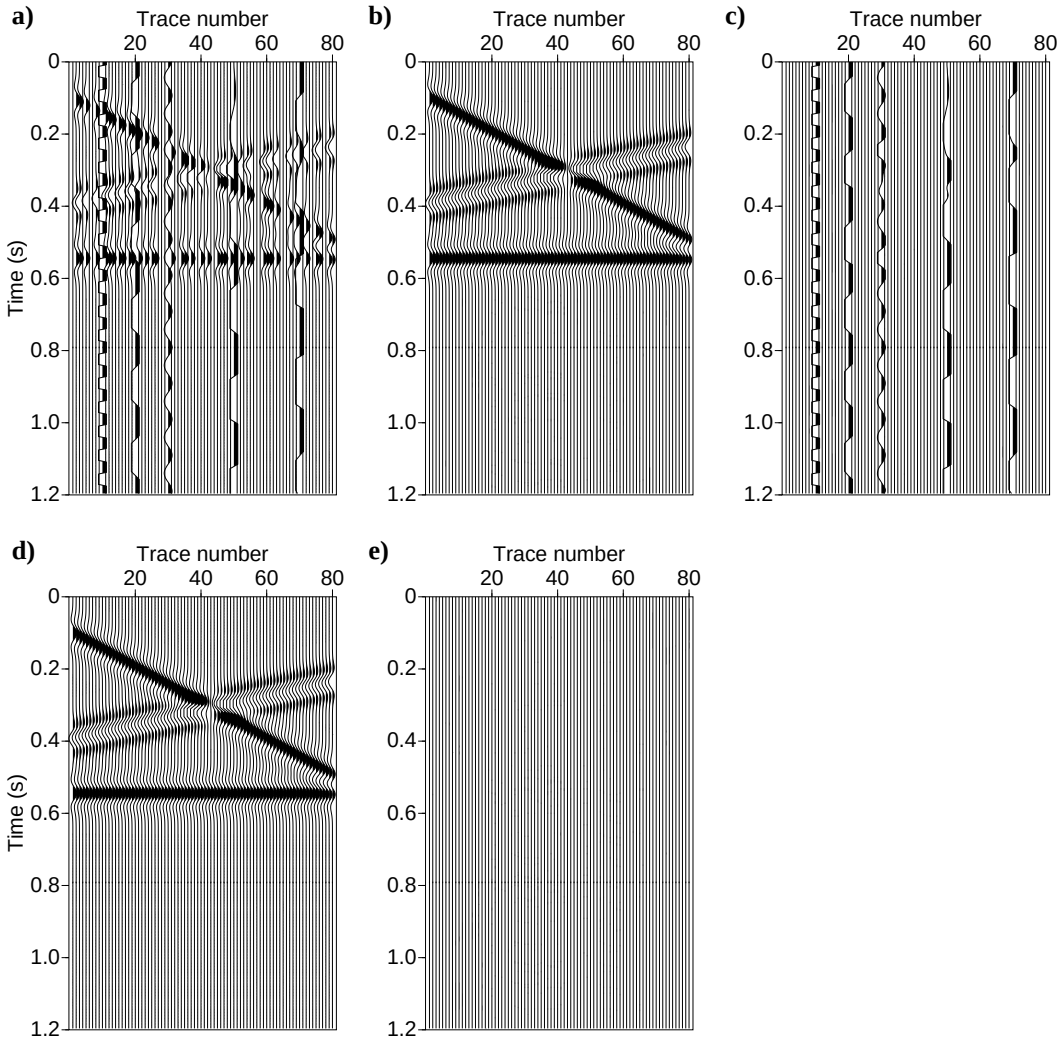

Figure 4.1: (a) Synthetic seismic data with 50% traces missing and 5 traces corrupted with erratic noise. b) Data filtered by robust SSA via low-rank matrix recovering. (c) Sparse erratic noise obtained from robust SSA. (d) Noise-free synthetic data. (e) Difference section between noise-free synthetic data and data filtered by robust SSA.

4.3.2 Synthetic Example 2

The synthetic data is both undersampled and corrupted with large amplitude time coherent noise and smoothed Gaussian noise. If frequency slice $\mathbf{D}$ has outliers, the recovering of $\mathbf{L}$ is achieved by solving convex program 4.6.

If there is no outlier in frequency slice $\mathbf{D}$, it changes to a Gaussian-noisy matrix completion problem

$$\begin{aligned} \min_{\mathbf{L},\mathbf{Z}} \quad & ||\mathbf{L}||_* + \frac{1}{2\mu}||\mathcal{P}_\Omega(\mathbf{Z})||_F^2, \\ \text{subject to} \quad & \mathbf{L}+\mathbf{Z}=\mathbf{M}. \end{aligned} \tag{4.27}$$

The augmented Lagrangian function is

$$\mathcal{L}_\mathcal{A}(\mathbf{L},\mathbf{Z},\mathbf{Y},\beta) = ||\mathbf{L}||_* + \frac{1}{2\mu}||\mathcal{P}_\Omega(\mathbf{Z})||_F^2 - \langle \mathbf{Y}, \mathbf{L}+\mathbf{Z}-\mathbf{M}\rangle + \frac{\beta}{2}||\mathbf{L}+\mathbf{Z}-\mathbf{M}||_F^2 \tag{4.28}$$

The algorithm of alternating splitting augmented Lagrangian method for Gaussian-noisy matrix completion problem is

Algorithm 5 Alternating Splitting Augmented Lagrangian Method (ASALM) for Matrix Completion with Noise

1: **Initialization**: $\mathbf{L}^0=\mathbf{0}$, $\mathbf{Y}^0=\mathbf{0}$, μ, β.
2: **while** not converge **do**
3: Compute $\mathbf{T}^k = \frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{L}^k$.
4: Compute $\mathbf{Z}^{k+1}$:

$$\mathbf{Z}_{ij}^{k+1} = \begin{cases} \mathbf{T}_{ij}^k & \text{when } (i,j)\notin\Omega; \\ \dfrac{\mu\beta}{1+\mu\beta}\mathbf{T}_{ij}^k & \text{when } (i,j)\in\Omega. \end{cases}$$

5: Compute $\mathbf{L}^{k+1}$: $\mathbf{L}^{k+1} = \mathcal{D}_{1/\beta}(\frac{1}{\beta}\mathbf{Y}^k + \mathbf{M} - \mathbf{Z}^{k+1})$.
6: Update $\mathbf{Y}^{k+1}$: $\mathbf{Y}^{k+1} = \mathbf{Y}^k + \beta(\mathbf{M} - \mathbf{L}^{k+1} - \mathbf{Z}^{k+1})$.
7: **end while**

Algorithm 5 is a special case of Algorithm 3 with $\mathbf{S}=\mathbf{0}$. In this situation, if a suitable threshold parameter λ/β is selected in Algorithm 3, matrix $\mathbf{S}$ will remain a zero matrix. Algorithm 3 can be applied on all the frequency slices. Figure 4.2 (d) is the noise-free data, Figure 4.2 (a) is the incomplete and corrupted data. The erratic noise added is the same with the one in Example 1. There are 25% traces missing in the original data set. The SNR of the Gaussian noise is 2. The frequency band to be processed is 1-40 Hz like previous examples. As to the parameters, $\mu = 0.1\sqrt{\min(m,n)+\sqrt{8\min(m,n)}}\sigma$, $\beta = 0.1\dfrac{|\Omega|}{||\mathbf{M}||_1}$. The stopping criterion is selected to be $SP \le 10^{-7}\sigma$. Figure 4.2 (b) is the reconstructed result. We can find that the algorithm removed the Gaussian noise and erratic noise and also recovered the missing traces. The quality factor of the reconstructed result is $Q=24$. Figure 4.2 (c) shows the erratic noise recovered by the sparse matrix $\mathbf{S}$. However, there is a little energy in the difference section Figure 4.2 (e).

4.4 Summary

In this chapter, we propose a robust SSA method for erratic noise suppression and missing data interpolation via solving low-rank matrix recovery problem. The low-rank matrix recovery is achieved by solving a convex program, i.e. minimizing the weighted combination of nuclear-norm ℓ_1 norm and Frobenius norm. We present some preliminary synthetic results. It shows that the proposed algorithm can remove Gaussian and erratic noise and interpolating missing traces simultaneously. There are several possible directions for the future works. First, the soft thresholding of the singular values allows the processing of curvature events situation. It is natural to extend the method to 3D and 5D situations.

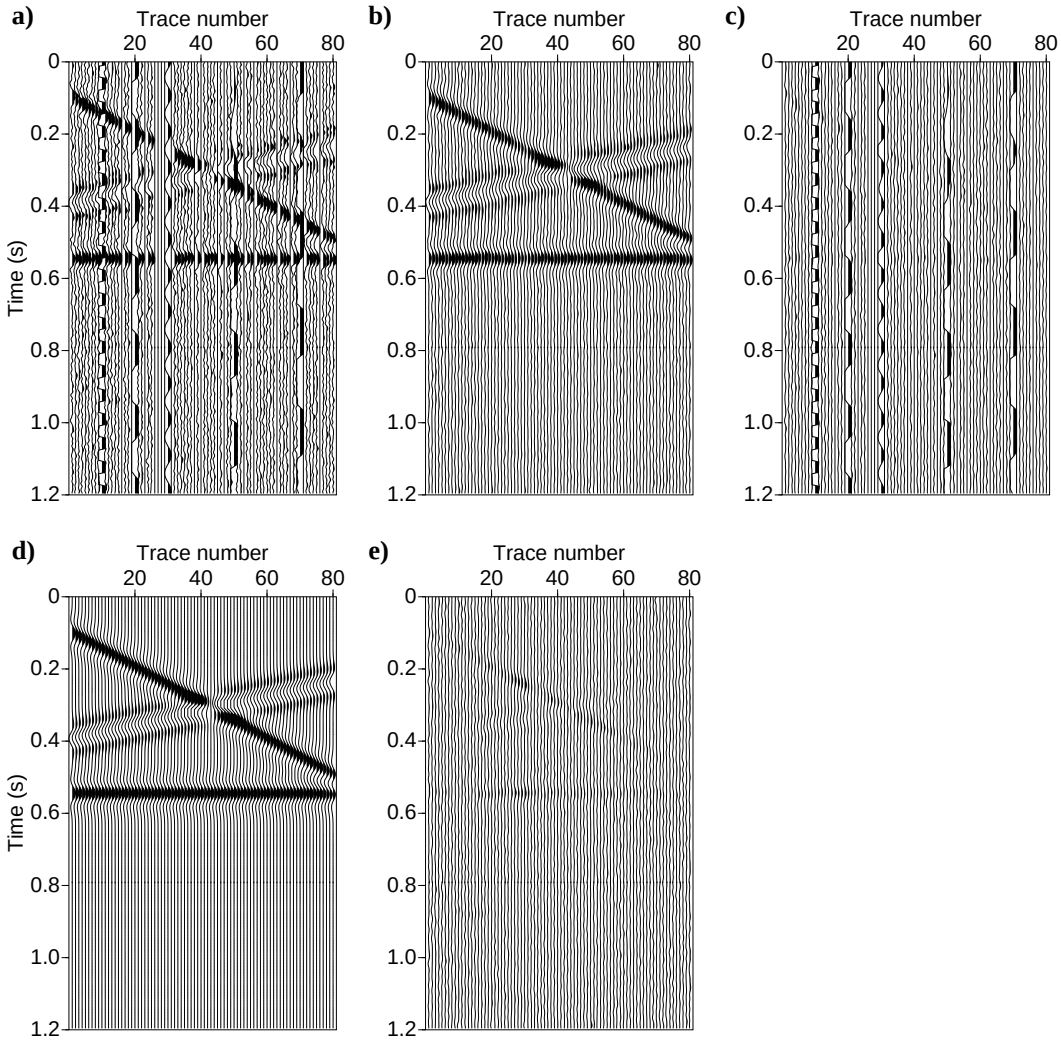

Figure 4.2: (a) Synthetic seismic data with 25% traces missing, Gaussian noise (SNR=2) and 5 traces corrupted with erratic noise. b) Data filtered by robust SSA via low-rank matrix recovering. (c) Sparse erratic noise obtained from robust SSA. (d) Noise-free synthetic data. (e) Difference section between noise-free synthetic data and data filtered by robust SSA.

CHAPTER 5

Conclusions

This thesis focuses on the subject of non-Gaussian seismic noise suppression in seismic data processing. More specially, we propose robust matrix rank-reduction methods for simultaneous Gaussian and non-Gaussian seismic noise attenuation.

I have proposed a new method that permits to robustify the SSA denoising method. This is achieved by introducing robust matrix factorization and nuclear norm minimization into the formulation of the SSA algorithm.

In Chapter 1, I briefly described the seismic data processing sequences in reflection seismology. Different types of seismic noise are described and different kinds of seismic noise attenuation methods were reviewed. I also reviewed methods for seismic data reconstruction. In Chapter 2, I reviewed concepts in multivariate statistics and linear algebra. The principal component analysis, singular value decomposition and eigendecomposition are described and their relationships are discussed. I described the algorithm of singular spectrum analysis for time series analysis. It is applied to the Southern Oscillation Index (SOI). The f-x SSA for seismic random noise attenuation is shown in this chapter as well. It is an alternative to the f-x deconvolution for random seismic noise attenuation and it has the advantage of preserving the amplitudes of seismic signals. I have recognize that SSA cannot cope with erratic noise and consequently chapter 3 gives a overview of the robust statistics with an emphasis on the M-estimators method. Loss functions other than the quadratic function (ℓ_2 norm) were described and I proposed a robust singular spectrum analysis method for Gaussian and non-Gaussian seismic noise attenuation that utilize robust loss functions. The TSVD in traditional SSA is replaced with the robust low rank approximation that is based on the M-estimate method. The biweight function was chosen as the loss function for our algorithm and iteratively reweighed least-squares and alternating minimization was used for numerically computing the robust factorization. Both

synthetic and field data examples show the superiority of the proposed robust SSA over the traditional SSA and f-x deconvolution for erratic seismic noise attenuation. Moreover, the robust SSA inherits the merit of traditional SSA that preserves the amplitude of the original signal. In Chapter 4, I proposed an algorithm for simultaneously removing Gaussian and non-Gaussian noise and interpolating missing traces. The recovering of low-rank matrix from incompletely observed and Gaussian and impulsive noise corrupted data matrix is solved through a convex optimization program. It minimizes a weighted combination of nuclear-norm, ℓ_1 norm and Frobenius norm. The augmented Lagrangian method is applied for the optimization. Preliminary synthetic examples are given to evaluate the performance of the proposed algorithm.

I conclude that incorporating robust statistics in reduced-rank (SSA) noise attenuation algorithms enables us to design algorithms that are resistant to outliers and erratic noise. This idea is important for processing land and marine seismic data that are often corrupted by noise that does not obey the Gaussian distribution.

There are several possible directions for future work: Extend the proposed robust SSA to the multidimensional case and improve the computational efficiency of the proposed robust algorithm via accelerating the robust low rank approximation step.

Bibliography

Abma, R. and J. Claerbout. "Lateral prediction for noise attenuation by t-x and f-x techniques." *Geophysics* 60 (1995): 1887–1896.

Abma, R. and N. Kabir. "3D interpolation of irregular data with a POCS algorithm." *Geophysics* 71 (2006): E91–E97.

Al-Yahya, K. "Application of the partial Karhunen-Loeve transform to suppress random noise in seismic sections." *Geophysical Prospecting* 39 (1991): 77–93.

Allen, M. and L. Smith. "Monte Carlo SSA: Detecting irregular oscillations in the Presence of Colored Noise." *Journal of Climate* 9 (1996).

Anderson, R. and G. McMechan. "Automatic editing of noisy seismic data." *Geophysical Prospecting* 37 (1989): 875–892.

Andrews, H. and B. Hunt. *Digital image restoration*. Prentice-Hall signal processing series. Prentice-Hall, 1977.

Beaton, A. and J. Tukey. "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data." *Technometrics* 16 (1974): pp. 147–185.

Bekara, M. and M. van der Baan. "High-amplitude noise detection by the expectation-maximization algorithm with application to swell-noise attenuation." *Geophysics* 75 (2010): V39–V49.

Berni, A. "Automatic surgical blanking of burst noise in marine seismic data." *57th Annual International Meeting, SEG, Expanded Abstracts* (1987): 477–478.

Bertero, M. and E. Pike. "Resolution in Diffraction-limited Imaging, a Singular Value Analysis." *Optica Acta: International Journal of Optics* 29 (1982): 727–746.

Bertsekas, D. *Constrained optimization and Lagrange multiplier methods*. Athena Scientific, 1982.

Brandwood, D. "A complex gradient operator and its application in adaptive array theory." *Communications, Radar and Signal Processing, IEE Proceedings F* 130 (1983): 11–16.

Broomhead, D. and G. King. "Extracting qualitative dynamics from experimental data." *Physica D: Nonlinear Phenomena* 20 (1986): 217 – 236.

Broomhead, D. and G. King. "On the qualitative analysis of experimental dynamical systems." *Nonlinear Phenomena and Chaos*. Ed. S. Sarkar. Bristol, England: Adam Hilger, 1986. 113–144.

Bube, K. and R. Langan. "Hybrid minimization with applications to tomography." *Geophysics* 62 (1997): 1183–1195.

Butler, K. and R. Russell. "Subtraction of powerline harmonics from geophysical records." *Geophysics* 58 (1993): 898–903.

Butler, K. and R. Russell. "Cancellation of multiple harmonic noise series in geophysical records." *Geophysics* 68 (2003): 1083–1090.

Cadzow, J. "Signal enhancement-a composite property mapping algorithm." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 36 (1988): 49–62.

Cai, J., E. Candès, and Z. Shen. "A Singular Value Thresholding Algorithm for Matrix Completion." *arXiv preprint arXiv:0810.3286* (2008).

Cambois, G. and J. Frelet. "Can we surgically remove swell noise?." *65th Annual International Meeting, SEG, Expanded Abstracts* (1995): 1381–1384.

Canales, L. "Random noise reduction." *54th Annual International Meeting, SEG, Expanded Abstracts* (1984): 525–527.

Candès, E., X. Li, Y. Ma, and J. Wright. "Robust Principal Component Analysis?." *CoRR* abs/0912.3599 (2009).

Candès, E. and Y. Plan. "Matrix Completion With Noise." *CoRR* abs/0903.3131 (2009).

Candès, E. and B. Recht. "Exact Matrix Completion via Convex Optimization." *Foundations of Computational Mathematics* 9 (2009): 717–772.

Candès, E., M. Wakin, and S. Boyd. "Enhancing Sparsity by Reweighted ℓ_1 Minimization." *Journal of Fourier Analysis and Applications* 14 (2008): 877–905.

Cary, P. and C. Zhang. "Ground roll attenuation with adaptive eigenimage filtering." *79th Annual International Meeting, SEG, Expanded Abstracts* (2009): 3302–3306.

Chave, A., D. Thomson, and M. Ander. "On the robust estimation of power spectra, coherences, and transfer functions." *Journal of Geophysical Research: Solid Earth* 92 (1987): 633–648.

Chen, S., D. Donoho, and M. Saunders. "Atomic Decomposition by Basis Pursuit." *SIAM Journal on Scientific Computing* 20 (1998): 33–61.

Chiu, S. "Coherent and random noise attenuation via multichannel singular spectrum analysis in the randomized domain." *Geophysical Prospecting* 61 (2013): 1–9.

Chiu, S. and J. Howell. "Attenuation of coherent noise using localized-adaptive eigenimage filter." *78th Annual International Meeting, SEG, Expanded Abstracts* (2008): 2541–2545.

Claerbout, J. and F. Muir. "Robust modeling with erratic data." *Geophysics* 38 (1973): 826–844.

Darche, G. "Spatial interpolation using a fast parabolic transform." (1990): 1647–1650.

Torre, F.De la and M. Black. "A Framework for Robust Subspace Learning." *International Journal of Computer Vision* 54 (2003): 117–142.

de Prony, G. "Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool à différentes températures." *J de l'Ecole Polytechnique* 1(2) (1795): 24–76.

Dondurur, D. and H. Karsl. "Swell Noise Suppression by Wiener Prediction Filter." *Journal of Applied Geophysics* 80 (2012): 91 – 100.

Dragoset, B. "Geophysical applications of adaptive-noise cancellation." *65th Annual International Meeting, SEG, Expanded Abstracts* (1995): 1389–1392.

Eckart, C. and G. Young. "The approximation of one matrix by another of lower rank." 10.1007/BF02288367. *Psychometrika* 1 (1936): 211–218.

Elboth, T., I. Presterud, and D. Hermansen. "Time-frequency seismic data de-noising." *Geophysical Prospecting* 58 (2010): 441–453.

Elsner, J. and A. Tsonis. *Singular Spectrum Analysis: A New Tool in Time Series Analysis.* Plenum Press, 1996.

Elston, S. "Use of robust estimators in multichannel stacking." *60th Annual International Meeting, SEG, Expanded Abstracts* (1990): 1693–1696.

Fraedrich, K. "Estimating the Dimensions of Weather and Climate Attractors." *Journal of the Atmospheric Sciences* 43 (1986): 419–432.

Freire, S. and T. Ulrych. "Application of singular value decomposition to vertical seismic profiling." *Geophysics* 53 (1988): 778–785.

Gabriel, K. and S. Zamir. "Lower Rank Approximation of Matrices by Least Squares with Any Choice of Weights." *Technometrics* 21 (1979): 489–498.

Gao, J., M. Sacchi, and X. Chen. "A fast reduced-rank interpolation method for prestack seismic volumes that depend on four spatial dimensions." *Geophysics* 78 (2013): V21–V30.

Ghil, M., M. Allen, M. Dettinger, K. Ide, D. Kondrashov, M. Mann, A. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. "Advanced spectral methods for climatic time series." *Reviews of Geophysics* 40 (2002): 3–1–3–41.

Golub, G. and C. Van Loan. *Matrix Computations*. Third edition. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996.

Golyandina, N. and A. Zhigljavsky. *Singular Spectrum Analysis for Time Series*. SpringerBriefs in Statistics. Springer London, Limited, 2013.

Guitton, A. and W. Symes. "Robust inversion of seismic data using the Huber norm." *Geophysics* 68 (2003): 1310–1319.

Gulunay, N. "FXDECON and complex wiener prediction filter." *56th Annual International Meeting, SEG, Expanded Abstracts* (1986): 279–281.

Gulunay, N. "Seismic trace interpolation in the Fourier transform domain." *Geophysics* 68 (2003): 355–369.

Halko, N., P. Martinsson, and J. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions." *SIAM Review* 53 (2011): 217–288.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, Inc, 1986.

Hansen, P. *Rank-Deficient and Discrete Ill-Posed Problems*. Society for Industrial and Applied Mathematics, 1998.

Harris, P. and R. White. "Improving the performance of f-x prediction filtering at low signal-to-noise ratios." *Geophysical Prospecting* 45 (1997): 269–302.

Hemon, CH. and D. Mace. "Use of the Karhunen-Loeve transformation in seismic data processing." *Geophysical Prospecting* 26 (1978): 600–626.

Herrmann, F. and G. Hennenfent. "Non-parametric seismic data recovery with curvelet frames." *Geophysical Journal International* 173 (2008): 233–248.

Holland, P. and R. Welsch. "Robust regression using iteratively reweighted least-squares." *Communications in Statistics - Theory and Methods* 6 (1977): 813–827.

Hornbostel, S. "Spatial prediction filtering in the t-x and f-x domains." *Geophysics* 56 (1991): 2019–2026.

Huber, P. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics* 35 (1964): pp. 73–101.

Huber, P. *Robust Statistics*. John Wiley & Sons, Inc, 1981.

Johnson, R. and D. Wichern. *Applied multivariate statistical analysis*. Sixth edition. Pearson Prentice-Hall, 2007.

Jolliffe, I. *Principal Component Analysis*. Second edition. Springer Series in Statistics. Springer, 2010.

Jones, I. and S. Levy. "Signal-to-noise ratio enhancement in multichannel seismic data via the Karhunen-Loeve Transform." *Geophysical Prospecting* 35 (1987): 12–32.

Kaplan, S., M. Naghizadeh, and M. Sacchi. "Data reconstruction with shot-profile least-squares migration." *Geophysics* 75 (2010): WB121–WB136.

Karhunen, K. "Ueber lineare Methoden in der Wahrscheinlichkeitsrechnung." *Annales Academiae scientiarum Fennicae. Series A* 137 (1947).

Kay, S. and S. Marple. "Spectrum analysis - A modern perspective." *Proceedings of the IEEE* 69 (1981): 1380–1419.

Kreimer, N. and M. Sacchi. "A tensor higher-order singular value decomposition for prestack seismic data noise reduction and interpolation." *Geophysics* 77 (2012): V113–V122.

Lin, Z., M. Chen, and Y. Ma. "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices." *ArXiv e-prints* (sep 2010).

Linville, A. and R. Meek. "Canceling stationary sinusoidal noise." *Geophysics* 57 (1992): 1493–1501.

Liu, B. and M. Sacchi. "Minimum weighted norm interpolation of seismic records." *Geophysics* 69 (2004): 1560–1568.

Liu, X. "Ground roll supression using the Karhunen-Loeve transform." *Geophysics* 64 (1999): 564–566.

Loeve, M. "Fonctions alatoires de seconde ordre." *Processes Stochastiques et Mouvement Brownien* (1948).

Marchisio, G., J. Pendrel, and B. Mattocks. "Applications of full and partial Karhunen-Loeve transformation to geophysical image enhancement." *58th Annual International Meeting, SEG, Expanded Abstracts* (1988): 1266–1269.

Mari, J. and F. Glangeaud. "Spectral Matrix Filtering Applied to VSP Processing." *Oil & Gas Science and Technology - Rev. IFP* 45 (1990): 417–434.

Maronna, R., R. Martin, and V. Yohai. *Robust Statistics*. John Wiley & Sons, Inc, 2006.

Maronna, R. and V. Yohai. "Robust Low-Rank Approximation of Data Matrices With Elementwise Contamination." *Technometrics* 50 (2008): 295–304.

Mars, J., F. Glangeaud, J. Lacoume, J. Fourmann, and S. Spitz. "Separation of seismic waves." *57th Annual International Meeting, SEG, Expanded Abstracts* (1987): 489–492.

Mavko, G. "Spectra-consistent automatic noise editing." *58th Annual International Meeting, SEG, Expanded Abstracts* (1988): 1275–1277.

Mayne, W. "Common reflection point horizontal data stacking techniques." *Geophysics* 27 (1962): 927–938.

Nagarajappa, N. "Coherent noise estimation by adaptive Hankel matrix rank reduction." *82nd Annual International Meeting, SEG, Expanded Abstracts* (2012): 1–5.

Naghizadeh, M. and M. Sacchi. "Multidimensional de-aliased Cadzow reconstruction of seismic records." *Geophysics* 78 (2013): A1–A5.

Neff, D. and S. Wyatt. "Noise suppression by the radial amplitude-slope rejection method." *Geophysics* 51 (1986): 844–850.

Oropeza, V. and M. Sacchi. "Application of singular spectrum analysis to ground roll attenuation." *CSPG CSEG CWLS Convention* (2010).

Oropeza, V. and M. Sacchi. "Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis." *Geophysics* 76 (2011): V25–V32.

Papoulis, A. and S. Pillai. *Probability, Random Variables And Stochastic Processes*. Fourth edition. McGraw-Hill series in electrical engineering. McGraw-Hill, 2002.

Rasmusson, E., X. Wang, and C. Ropelewski. "The biennial component of ENSO variability." *Journal of Marine Systems* 1 (1990): 71 – 96.

Sacchi, M. "FX Singular Spectrum Analysis." *CSPG CSEG CWLS CONVENTION* (2009).

Sacchi, M. and H. Kuehl. "ARMA formulation of FX prediction error filters and projection filters." *Journal of Seismic Exploration* 3 (2001): 185–197.

Sacchi, M., T. Ulrych, and C. Walker. "Interpolation and extrapolation using a high-resolution discrete Fourier transform." *Signal Processing, IEEE Transactions on* 46 (1998): 31–38.

Sardy, S. "Minimax threshold for denoising complex signals with Waveshrink." *Signal Processing, IEEE Transactions on* 48 (apr 2000): 1023 –1028.

Scales, J. and A. Gersztenkorn. "Robust methods in inverse theory." *Inverse Problems* 4 (1988): 1071.

Scales, J., A. Gersztenkorn, and S. Treitel. "Fast l_p solution of large, sparse, linear systems: Application to seismic travel time tomography." *Journal of Computational Physics* 75 (1988): 314 – 333.

Schonewille, M., A. Vigner, and A. Ryder. "Swell-noise attenuation using an iterative FX prediction filtering approach." *78th Annual International Meeting, SEG, Expanded Abstracts* (2008): 2647–2651.

Simon, H. and H. Zha. "Low-Rank Matrix Approximation Using the Lanczos Bidiagonalization Process with Applications." *SIAM Journal on Scientific Computing* 21 (2000): 2257–2274.

Soubaras, R. "Signal-preserving random noise attenuation by the f-x projection." *64th Annual International Meeting, SEG, Expanded Abstracts* (1994): 1576–1579.

Soubaras, R. "Prestack random and impulsive noise attenuation by f-x projection filtering." *65th Annual International Meeting, SEG, Expanded Abstracts* (1995): 711–714.

Soubaras, R. "Spatial interpolation of aliased seismic data." (1997): 1167–1170.

Spitz, S. "Seismic trace interpolation in the F-X domain." *Geophysics* 56 (1991): 785–794.

Stolt, R. "Seismic data mapping and reconstruction." *Geophysics* 67 (2002): 890–908.

Strang, G. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 1993.

Strang, G. *Linear Algebra and Its Applications*. Fourth edition. Thomson Brooks/Cole, 2006.

Takens, F. "Detecting strange attractors in turbulence." *Dynamical Systems and Turbulence, Warwick 1980.* . Lecture Notes in Mathematics. Springer Berlin Heidelberg, 1981. 366–381.

Tao, M. and X. Yuan. "Recovering Low-Rank and Sparse Components of Matrices from Incomplete and Noisy Observations." *SIAM Journal on Optimization* 21 (2011): 57–81.

Taylor, H., S. Banks, and J. McCoy. "Deconvolution with the ℓ_1 norm." *Geophysics* 44 (1979): 39–52.

Trad, D. "Interpolation and multiple attenuation with migration operators." *Geophysics* 68 (2003): 2043–2054.

Trad, D., T. Ulrych, and M. Sacchi. "Accurate interpolation with highresolution timevariant Radon transforms." *Geophysics* 67 (2002): 644–656.

Trickett, S. "F-x eigenimage noise suppression." *72nd Annual International Meeting, SEG, Expanded Abstracts* (2002): 2166–2169.

Trickett, S. "F-xy eigenimage noise suppression." *Geophysics* 68 (2003): 751–759.

Trickett, S. "Maximum-likelihoodestimation stacking." *77th Annual International Meeting, SEG, Expanded Abstracts* (2007): 2640–2643.

Trickett, S. "F-x Cadzow noise suppression." *78th Annual International Meeting, SEG, Expanded Abstracts* (2008): 2586–2590.

Trickett, S. and L. Burroughs. "Prestack rankreducing noise suppression: Theory." *79th Annual International Meeting, SEG, Expanded Abstracts* (2009): 3332–3336.

Trickett, S., L. Burroughs, and A. Milton. "Robust rank-reduction filtering for erratic noise." *82nd Annual International Meeting, SEG, Expanded Abstracts* (2012): 1–5.

Trickett, S., L. Burroughs, A. Milton, L. Walton, and R. Dack. "Rank-reduction-based trace interpolation." *80th Annual International Meeting, SEG, Expanded Abstracts* (2010): 3829–3833.

Tufts, D. and R. Kumaresan. "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood." *Proceedings of the IEEE* 70 (1982): 975–989.

Ulrych, T. and R. Clayton. "Time series modelling and maximum entropy." *Physics of the Earth and Planetary Interiors* 12 (1976): 188 – 200.

Ulrych, T., S. Freire, and P. Siston. "Eigenimage processing of seismic sections." *58th Annual International Meeting, SEG, Expanded Abstracts* (1988): 1261–1265.

Ulrych, T. and M. Sacchi. *Information-Based Inversion and Processing with Applications*. Handbook of Geophysical Exploration: Seismic Exploration. Elsevier Science, 2005.

Ulrych, T., M. Sacchi, and J. Graul. "Signal and noise separation: Art and science." *Geophysics* 64 (1999): 1648–1656.

Vautard, R. and M. Ghil. "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series." *Physica D: Nonlinear Phenomena* 35 (1989): 395 – 424.

Vautard, R., P. Yiou, and M. Ghil. "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals." *Physica D: Nonlinear Phenomena* 58 (1992): 95 – 126.

Verboon, P. and W. Heiser. "Resistant lower rank approximation of matrices by iterative majorization." *Computational Statistics & Data Analysis* 18 (1994): 457 – 467.

Watt, T. and J. Bednar. "Role of the alpha-trimmed mean in combining and analyzing seismic commondepthpoint gathers." *53rd Annual International Meeting, SEG, Expanded Abstracts* (1983): 276–277.

Widrow, B., P. Mantey, L. Griffiths, and B. Goode. "Adaptive antenna systems." *Proceedings of the IEEE* 55 (1967): 2143–2159.

Yilmaz, Oz. *Seismic Data Analysis: Processing, Inversion, and Interpretation of Seismic Data*. Second edition. Society of Exploration Geophysicists, 2001.

Yuan, X. and J. Yang. "Sparse and low-rank matrix decomposition via alternating direction methods." (2009).

Zhou, Z., X. Li, J. Wright, E. Candès, and Y. Ma. "Stable Principal Component Pursuit." *CoRR* abs/1001.2363 (2010).

APPENDIX A

Gradient in Complex Domain

Here, I describe how to use the partial complex derivative to compute the weighting function of the cost function (3.75):

$$E(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\rho\left(\frac{m_{ij} - \sum_{q=1}^{K} u_{iq}v_{jq}^{*}}{\sigma}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n}\rho\left(\frac{r_{ij}}{\sigma}\right). \tag{A.1}$$

If the matrix $\mathbf{V}$ is fixed, E is a function of $\mathbf{U}$. Similarly, E is a function of $\mathbf{V}$ if $\mathbf{U}$ is fixed. However, even though $\mathbf{V}$ is fixed, E is not a complex analytic function with respect to $\mathbf{U}$. The complex derivative does not exist. By using Wirtinger's Caculus (Brandwood, 1983), we can regard $\mathbf{U}$ and $\mathbf{U}^{*}$ as independent variables. Either setting the partial derivative $\frac{\partial E}{\partial \mathbf{U}}$ or $\frac{\partial E}{\partial \mathbf{U}^{*}}$ to zero lead to stationary points. Usually, $\frac{\partial E}{\partial \mathbf{U}^{*}}$ is preferred because it gives the direction where the cost function E has the maximum rate of change with respect to $\mathbf{U}$

(Brandwood, 1983).

$$\begin{aligned}
\frac{\partial E}{\partial u_{ab}^*} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial \rho\left(\frac{r_{ij}}{\sigma},\frac{r_{ij}^*}{\sigma}\right)}{\partial u_{ab}^*} = \sum_{j=1}^{n}\frac{\partial \rho\left(\frac{r_{aj}}{\sigma},\frac{r_{aj}^*}{\sigma}\right)}{\partial u_{ab}^*} \\
&= \sum_{j=1}^{n}\left(\frac{\partial \rho\left(\frac{r_{aj}}{\sigma},\frac{r_{aj}^*}{\sigma}\right)}{\partial r_{aj}}\frac{\partial r_{aj}}{\partial u_{ab}^*} + \frac{\partial \rho\left(\frac{r_{aj}}{\sigma},\frac{r_{aj}^*}{\sigma}\right)}{\partial r_{aj}^*}\frac{\partial r_{aj}^*}{\partial u_{ab}^*}\right) \\
&= \sum_{j=1}^{n}\frac{\partial \rho\left(\frac{r_{aj}}{\sigma},\frac{r_{aj}^*}{\sigma}\right)}{\partial r_{aj}^*}\frac{\partial r_{aj}^*}{\partial u_{ab}^*} = -\sum_{j=1}^{n}\frac{\partial \rho\left(\frac{r_{aj}}{\sigma}\right)}{\partial r_{aj}^*}v_{jb} \\
&= -\sum_{j=1}^{n}\frac{\partial \rho\left(\frac{r_{aj}}{\sigma}\right)}{\partial r_{aj}^*}\frac{1}{r_{aj}}r_{aj}v_{jb} \\
&= -\frac{1}{\sigma^2}\sum_{j=1}^{n} w\left(\frac{r_{aj}}{\sigma}\right) r_{aj}v_{jb},
\end{aligned} \tag{A.2}$$

where $w(x) = \frac{\partial \rho(x)}{\partial x^*}\frac{1}{x}$ with $x = \frac{r_{aj}}{\sigma}$ is the weighting function that is different from real value case. In the above equations, ρ is not an analytic function of r_{aj} or r_{aj}^* therefore, we applied the chain rule to the complex partial derivative of r_{aj} and r_{aj}^*. Due to the relationship $\frac{\partial |x|}{\partial x^*} = \frac{1}{2}\frac{x}{|x|}$, we have that $w(x) = \frac{\partial \rho(x)}{\partial x^*}\frac{1}{x} = \frac{\partial \rho(x)}{\partial x}\frac{1}{x^*} = \frac{1}{2}\frac{\partial \rho(x)}{\partial |x|}\frac{1}{|x|}$.

Similarly, we can compute $\frac{\partial E}{\partial \mathbf{V}^*}$

$$\begin{aligned}
\frac{\partial E}{\partial v_{cd}^*} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial \rho\left(\frac{r_{ij}}{\sigma},\frac{r_{ij}^*}{\sigma}\right)}{\partial v_{cd}^*} = \sum_{i=1}^{m}\frac{\partial \rho\left(\frac{r_{ic}}{\sigma},\frac{r_{ic}^*}{\sigma}\right)}{\partial v_{cd}^*} \\
&= \sum_{i=1}^{m}\left(\frac{\partial \rho\left(\frac{r_{ic}}{\sigma},\frac{r_{ic}^*}{\sigma}\right)}{\partial r_{ic}}\frac{\partial r_{ic}}{\partial v_{cd}^*} + \frac{\partial \rho\left(\frac{r_{ic}}{\sigma},\frac{r_{ic}^*}{\sigma}\right)}{\partial r_{ic}^*}\frac{\partial r_{ic}^*}{\partial v_{cd}^*}\right) \\
&= \sum_{i=1}^{m}\frac{\partial \rho\left(\frac{r_{ic}}{\sigma}\right)}{\partial r_{ic}}\frac{\partial r_{ic}}{\partial v_{cd}^*} = \sum_{i=1}^{m}\frac{\partial \rho\left(\frac{r_{ic}}{\sigma}\right)}{\partial r_{ic}}(-u_{id}) \\
&= -\sum_{i=1}^{m}\frac{\partial \rho\left(\frac{r_{ic}}{\sigma}\right)}{\partial r_{ic}}\frac{1}{r_{ic}^*}r_{ic}^*u_{id} \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{m} w\left(\frac{r_{ic}}{\sigma}\right) r_{ic}^*u_{id},
\end{aligned} \tag{A.3}$$

where $w\left(x\right) = \frac{\partial \rho(x)}{\partial x}\frac{1}{x^*} = \frac{\partial \rho(x)}{\partial x^*}\frac{1}{x} = \frac{1}{2}\frac{\partial \rho(x)}{\partial |x|}\frac{1}{|x|}$ with $x = \frac{r_{ic}}{\sigma}$.