# Principal and Independent Component Analysis for Seismic Data

by

Sam T. Kaplan
B. Sc., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in
THE FACULTY OF GRADUATE STUDIES
(Department of Earth and Ocean Sciences)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
April 2003

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Earth and Ocean Sciences
The University of British Columbia
2219 Main Mall
Vancouver, BC, Canada
V6T 1Z4

Date: April 22, 2003

# Abstract

Principal and Independent component analysis (PCA and ICA) are two ideas which are very much related; both employing a statistical understanding of data to achieve their goals. Whereas PCA exploits statistical correlation, ICA uses statistical independence to glean useful information from data. Seismic data is inherently noisy, and is complicated by the presence of an unknown seismic wavelet. Analysis of the data is aided by, both, noise suppression and blind deconvolution techniques.

First, consider the subject of noise suppression. If the data are organized into several sequences where, from one sequence to the next, the signal is correlated while the noise is uncorrelated, then PCA has the ability to separate noise and signal. Here, PCA is analyzed from three points of view, variance maximization, the singular value decomposition and neural networks. The resulting theory is used to filter noise from a set of common midpoint seismic gathers by exploiting correlations which exist from one gather to the next.

To further simplify analysis of these data, the Earth is often approximated as a linear system; thus, the seismic trace is subject to the convolutional model. Convolution is a linear operation, and consequently, can be formulated as a linear system of equations. If only the output of the system (the convolved signal) is known, then the problem is blind so that given one equation, two unknowns are sought. This problem is well suited for ICA which has the ability to find some estimate of the two unknowns, and here the blind deconvolution problem is solved using ICA. To facilitate this, several time-lagged versions of the convolved signal are extracted and used to construct realizations of a random vector. For ICA, this random vector is the, so called, mixture vector, created by the matrix-vector multiplication of the two unknowns, the mixing matrix and the source vector. Due to the properties of convolution, the mixing matrix is banded with its nonzero elements containing the convolution's filter. This banded property is incorporated into the ICA algorithm as prior information, giving rise to a banded ICA algorithm (B-ICA) which is, in turn, used in a new blind deconvolution algorithm. This algorithm is considered for both noiseless and noisy data.

# Contents

# Appendices

# List of Figures

# Acknowledgments

First and foremost, I would like to thank my supervisor, Dr. Tadeusz Ulrych. His enthusiasm for science is infectious and inspiring. I am very appreciative for the time he took to listen to my ideas, as well as his efforts to inspire new ones; which he did with great success. Thank-you, also, to professors Felix Herrmann and Michael Bostock for taking time from your busy schedules to critically read my thesis. It is a pleasure to be a graduate student at UBC. This is, in large part, due to the people in the department; thanks for Friday afternoon soccer, organizing hockey matches, providing doughnuts and discussing science. I would, particularly, like to thank Daniel Trad who on many occasions found himself listening to my, often incorrect, ideas. His feedback was always valuable and is an important contribution to this thesis. This thesis was written while working as a research assistant with the CDSST. Thanks to my supervisor for forming this group, and to the consortium sponsors for their support. Lastly, I would like to thank my family for their continuing support; while never, quite, understanding what I was working on, they always encouraged me to continue doing so.

# CHAPTER 1

# Introduction

## 1.1 Motivation and Tasks

This thesis is concerned with processing data. Reflection seismic data are of particular interest, and methods to both clarify and simplify these data are presented. Reflection seismic data analysis is a mature field of research. It is a means to, more efficiently, harvest the Earth's natural resources. Thus, the funding and attention that it has received are hardly surprising. The abundance of research is apparent in the multitude of available literature, including a two thousand page treatment of the subject by Yilmaz [2001].

While this thesis is, ultimately, about data, its focus is processing. In particular, two concepts are considered: First principal component analysis (PCA), and second independent component analysis (ICA). These two ideas are very much related; both employing statistical understandings of the available data to achieve their goals. PCA exploits statistical correlation, while ICA considers statistical independence, as such, the relation between PCA and ICA is revealed in the equations that bind independence and correlation.

Two tasks, stemming from seismic data, are pondered while considering PCA and ICA. Namely, the tasks of noise suppression and blind deconvolution. As with all real data, seismic data are inherently noisy. PCA is used as a means for suppressing random noise. This is by no means a new concept. However, the methods are adapted, in this thesis, for collections of two dimensional seismic gathers; seismic gathers which provide some representation of the Earth. To further simplify

analysis of these data, the Earth is often approximated as a linear system; thus, the seismic trace is expressed as the convolutional model which, because of the nature of the reflection seismology experiment, is a single equation in two unknowns. The unknowns are the reflectivity of the Earth and the seismic wavelet; the known quantity is the seismic trace. Solving for the two unknowns in the one equation, allowed for by the convolutional model, is blind deconvolution. In this thesis, a new algorithm is devised, using a modified ICA algorithm, providing a solution to the blind deconvolution problem for both noiseless and noisy data.

## 1.2 Thesis Overview

In Chapter 2, PCA is analyzed from three points of view. First variance maximization, second using the singular value decomposition and third using neural networks and ordinary differential equations. These analyses give insight into how and why PCA can be used for the attenuation of random noise. Both synthetic and real data examples illustrate its effectiveness.

Whereas PCA exploits correlation, ICA exploits independence. In Chapter 3, concepts from higher order statistics and information theory allow ICA to solve for two unknowns in one equation. In Chapter 4, a new blind deconvolution algorithm utilizing ICA is described. For this purpose, the ICA algorithm is modified to match properties inherent in the convolutional model. In Chapter 5, the blind deconvolution problem is treated with the additional complication of random noise. Again, this requires modification to ICA to account for the noise and its presence in the convolutional model.

## 1.3 Related Methods and Algorithms

As mentioned, one of the two tasks of this thesis, blind deconvolution, is solved by way of ICA, and ICA can be likened to other approaches described in the literature. Most prominent of these is projection pursuit [e.g. Jones and Sibson, 1987] which looks for interesting features in data. It so happens, that one of these *measures of interest* is non-Gaussianity. As will be shown in Chapter 3, this same measure is used for ICA. Further, a relation can be found between ICA and self organizing neural networks. Indeed, in this thesis, the relationship between PCA and neural networks is explicitly shown in Chapter 2. A similar relation holds between ICA and neural

networks, and is well summarized in Bell and Sejnowski [1995].

Blind deconvolution is by no means a new subject. In geophysics, it was first introduced by Wiggins [1978] and has received much attention [Haykin, 1994, 2000]. The result is an abundance of work to draw from. It is not the intent of this thesis to provide an overview. Rather, the intention is to utilize recent advances, stemming from ICA, to produce a new blind deconvolution algorithm.

# CHAPTER 2

# Principal Component Analysis

## 2.1 Introduction

Consider a multivariate data set, $\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$, where $\mathbf{x}$ is a random vector and $x_i(t_j)$ is the $j^{th}$ realization of the $i^{th}$ random variable.[1] In seismic data, for example, the realizations of $x_i$ might be a seismic trace, and a single realization of $\mathbf{x}$ could be a time slice from a seismic section. Principal component analysis (PCA) transforms $\mathbf{x}$ via

$$\zeta_i = \mathbf{u}_i^T \mathbf{x} \quad , \quad i = 1 \ldots m \tag{2.1}$$

where the random variables, $\zeta_i$, are principal components and $\mathbf{u}_i^T = \begin{bmatrix} u_{i1} & u_{i2} & \cdots & u_{im} \end{bmatrix}$ are chosen to explain the data with few dimensions (principal components).

This chapter describes PCA from three points of view. Each is somewhat similar as they satisfy the same constraints. However, each approach has unique motivation; thus, providing further understanding of the method. First, PCA is derived by maximizing the variance of the principal components [Cooley and Lohnes, 1971, Ch. 4]. Second, a connection is drawn between explaining variance and explaining data using the singular value decomposition (SVD). Lastly, neural networks are used such that the principal components can be updated as more information (realizations) becomes available [Oja, 1982]. The chapter concludes with an example of PCA in signal processing. In particular, an application to noise suppression for seismic data is considered

[1] All vectors in this thesis are column vectors. i.e. $\mathbf{x}$ is a column vector, and $\mathbf{x}^T$ is a row vector.

where an original (to seismic data processing) extension of this method to three dimension is explained [Kaplan and Ulrych, 2002]. While this chapter can be read independently of the rest of the thesis, the methods developed herein are used in the remaining chapters.

## 2.2 PCA by Variance Maximization

In the approach of variance maximization, $\mathbf{u}_i$ are found such that $\mathrm{var}(\zeta_i)$ are maximized subject to some constraints. Namely that $\mathbf{u}_i^T\mathbf{u}_i = 1$, that the second principal component is uncorrelated with the first, the third uncorrelated with both the first and second, and so on. The first of these constraints is built explicitly into the cost function using a Lagrange multiplier. Constraining the principal components to be uncorrelated is implicit in the formulation and, as will be shown, falls nicely out of the mathematics. Hence, the appropriate cost functions (for maximization) are

$$\begin{aligned}
\phi(\mathbf{u}_i) &= \mathrm{var}(\zeta_i) + \lambda_i\left(1 - \mathbf{u}_i^T\mathbf{u}_i\right) \\
&= \mathrm{E}\left(\zeta_i^2\right) + \lambda_i\left(1 - \mathbf{u}_i^T\mathbf{u}_i\right) \qquad (2.2) \\
&= \mathrm{E}\left[\left(\mathbf{u}_i^T\mathbf{x}\right)\left(\mathbf{u}_i^T\mathbf{x}\right)^T\right] + \lambda_i\left(1 - \mathbf{u}_i^T\mathbf{u}_i\right) \\
&= \mathrm{E}\left(\mathbf{u}_i^T\mathbf{x}\mathbf{x}^T\mathbf{u}_i\right) + \lambda_i\left(1 - \mathbf{u}_i^T\mathbf{u}_i\right) \\
&= \mathbf{u}_i^T\mathbf{C}_x\mathbf{u}_i + \lambda_i\left(1 - \mathbf{u}_i^T\mathbf{u}_i\right) \qquad (2.3)
\end{aligned}$$

where $\lambda_i$ are Lagrange multipliers and $\mathbf{C}_x$ is the correlation matrix of $\mathbf{x}$. Equation (2.2) assumes $\mathrm{E}(\zeta_i) = 0$; that is, $\mathrm{var}(\zeta_i) = \mathrm{E}\left(\zeta_i^2\right) - \mathrm{E}(\zeta_i)^2$ . This assumption is trivial since the mean of $\mathbf{x}$ is easily set to zero and $\mathrm{E}(\zeta_i) = \mathrm{E}\left(\mathbf{u}_i^T\mathbf{x}\right) = \mathbf{u}_i^T\mathrm{E}(\mathbf{x})$. Taking the gradient of equation (2.3) gives

$$\nabla\phi(\mathbf{u}_i) = 2\mathbf{C}_x\mathbf{u}_i - 2\lambda_i\mathbf{u}_i,$$

and setting this result to zero yield the extrema of the cost functions,

$$\mathbf{C}_x\mathbf{u}_i = \lambda_i\mathbf{u}_i. \qquad (2.4)$$

Equation (2.4) is easily recognized as an eigen problem where $\mathbf{u}_i$ are the eigenvectors of the symmetric matrix $\mathbf{C}_x$, and are therefore mutually orthonormal. Hence,

$$\mathrm{E}\left(\zeta_i\zeta_j\right) = \mathrm{E}\left[\left(\mathbf{u}_i^T\mathbf{x}\right)\left(\mathbf{u}_j^T\mathbf{x}\right)^T\right] = \mathrm{E}\left(\mathbf{u}_i^T\mathbf{x}\mathbf{x}^T\mathbf{u}_j\right) = \mathbf{u}_i^T\mathbf{C}_x\mathbf{u}_j = \lambda_j\mathbf{u}_i^T\mathbf{u}_j = \begin{cases} 0 & , \quad i \neq j \\ \lambda_j = \mathrm{var}\left(\zeta_j\right) & , \quad i = j \end{cases}. \tag{2.5}$$

Equation (2.5) illustrates two ideas. First, it confirms that the principal components are uncorrelated; and second, it demonstrates that the variance of the $i^{th}$ principal component, $\zeta_i = \mathbf{u}_i^T\mathbf{x}$, is the $i^{th}$ eigenvalue, $\lambda_i$. Hence, ordering the pairs of eigenvectors and eigenvalues in the usual fashion so that $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ completes the solution.

This derivation clarifies the role of PCA in terms of variance. In particular, the method attempts to explain the variance in the data with few dimensions (principal components). However, the connection between *explaining variance* and *explaining data* is, at best, mysterious. To illuminate this relation it behooves us to consider PCA in terms of the SVD.

## 2.3 PCA and the SVD

The SVD decomposes a $m \times n$ matrix, $\mathbf{A}$, into the product of three matrices,

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2.6}$$

where

$$\begin{aligned} \mathbf{U} &= \left[\begin{array}{c|c|c|c} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{array}\right], \\ \mathbf{V} &= \left[\begin{array}{c|c|c|c} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{array}\right], \\ \mathbf{\Sigma} &= \mathrm{diag}\left(\sigma_1, \sigma_2, \ldots, \sigma_p, 0, \ldots, 0\right) \\ &= \mathrm{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p}, 0, \ldots, 0\right) \end{aligned}$$

and $\mathbf{u}_i$, $i = 1 \ldots m$ and $\mathbf{v}_i$, $i = 1 \ldots n$ are the eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ respectively, and $\sigma_i$ and $\lambda_i$, $i = 1 \ldots p = \mathrm{rank}(\mathbf{A})$ are, respectively, the nonzero singular values of $\mathbf{A}$ and the correpsonding eigenvalues of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ [e.g. Strang, 1988, p. 443].

The connection between PCA derived in the previous section and the SVD is readily found. If

$$\mathbf{A} = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_1(t_2) & \cdots & x_1(t_n) \\ x_2(t_1) & x_2(t_2) & \cdots & x_2(t_n) \\ \cdots & \cdots & \cdots & \cdots \\ x_m(t_1) & x_m(t_2) & \cdots & x_m(t_n) \end{bmatrix}, \tag{2.7}$$

then $\mathbf{A}\mathbf{A}^T = n\mathbf{C}_x$ where $\mathbf{A}$ contains $n$ realizations of $\mathbf{x}$ and $\mathbf{C}_x$ is, as in Section 2.2, the correlation matrix of $\mathbf{x}$. Hence, $\mathbf{u}_i$ are the eigenvectors of $\mathbf{C}_x$. Further, noting that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, equation (2.6) gives

$$\zeta = \mathbf{U}^T\mathbf{A} = \mathbf{\Sigma}\mathbf{V}^T \tag{2.8}$$

where

$$\zeta = \begin{bmatrix} \zeta_1(t) \\ \zeta_2(t) \\ \vdots \\ \zeta_m(t) \end{bmatrix} = \begin{bmatrix} \zeta_1(t_1) & \zeta_1(t_2) & \cdots & \zeta_1(t_n) \\ \zeta_2(t_1) & \zeta_2(t_2) & \cdots & \zeta_2(t_n) \\ \cdots & \cdots & \cdots & \cdots \\ \zeta_m(t_1) & \zeta_m(t_2) & \cdots & \zeta_m(t_n) \end{bmatrix}$$

and $\zeta_i(t)$ are, as in Section 2.2, principal components.

In addition to decomposing $\mathbf{A}$ into the product of three matrices (equation (2.6)), the SVD, equivalently, decomposes $\mathbf{A}$ into a sum of matrices,

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{p} \sigma_i \mathbf{E}_i \tag{2.9}$$

where $p = \text{rank}(\mathbf{A})$ and $\mathbf{E}_i = \mathbf{u}_i\mathbf{v}_i^T$ are eigenimages. The eigenimages are significant since if

$$\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{E}_i \quad , \quad k < p,$$

then

$$\min||\mathbf{A} - \mathbf{B}||_2 = ||\mathbf{A} - \mathbf{A}_k||_2 = \sigma_{k+1} \tag{2.10}$$

where $\mathbf{B}$ is any rank$(k)$ matrix. That is, with respect to the $l_2$ norm, $\mathbf{A}_k$ is the closest approximation

to $\mathbf{A}$ for any rank($k$) matrix [Golub and Van Loan, 1996, pp. 72-73]. Additionally, the rows of $\mathbf{E}_i$ are scalar multiples of $\mathbf{v}_i$. A fact which is easily seen by examining an eigenimage in its matrix form,

$$\mathbf{E}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{im} \end{bmatrix} \begin{bmatrix} v_{i1} & v_{i2} & \cdots & v_{in} \end{bmatrix} = \begin{bmatrix} u_{i1}\mathbf{v}_i^T \\ u_{i2}\mathbf{v}_i^T \\ \vdots \\ u_{im}\mathbf{v}_i^T \end{bmatrix}.$$

Therefore, the structural information of $\mathbf{E}_i$, and thus also $\mathbf{A}$, can be expressed using only the vectors $\mathbf{v}_i$; and hence, using only principal components (see equation (2.8)). Combined with equation (2.10), eigenimages imply that information, which is coherent across the rows of $\mathbf{A}$, is represented by the first few eigenimages (principal components). In other words, a matrix containing mainly coherent information is synonymous with a matrix of small rank.

The relation between eigenimages and principal components along with equation (2.10) connect the ideas of *explaining variance* and *explaining data*, namely that the two concepts are equivalent. Hence, PCA attempts to explain the data with few dimensions (principal components), and in doing so, extracts coherent information from the data.

## 2.4 PCA by Neural Networks

In Sections 2.2 and 2.3 principal components, $\zeta_i$, are computed with full knowledge of the correlation matrix, $\mathbf{C}_x$, which, in turn, requires some fixed number of realizations of $\mathbf{x}(t)$. A neural network formulation of PCA allows for online computation of principal components. That is, as more realizations of $\mathbf{x}(t)$ are made available, the principal components are updated accordingly. Interestingly, the derivation of PCA in a neural network framework is motivated through a learning rule designed to mimic the human brain. In particular, a modified version of a learning rule postulated by Hebb [1949] is used which, as it happens, allows the neural network to learn principal components.

Figure 2.1 is a schematic of a neural network consisting of $m$ input neurons and one output neuron. It finds the first principal component of some data, $\mathbf{x}(t)$. The data are passed through the input neurons of the network, and subsequently through the weights of the network, producing the

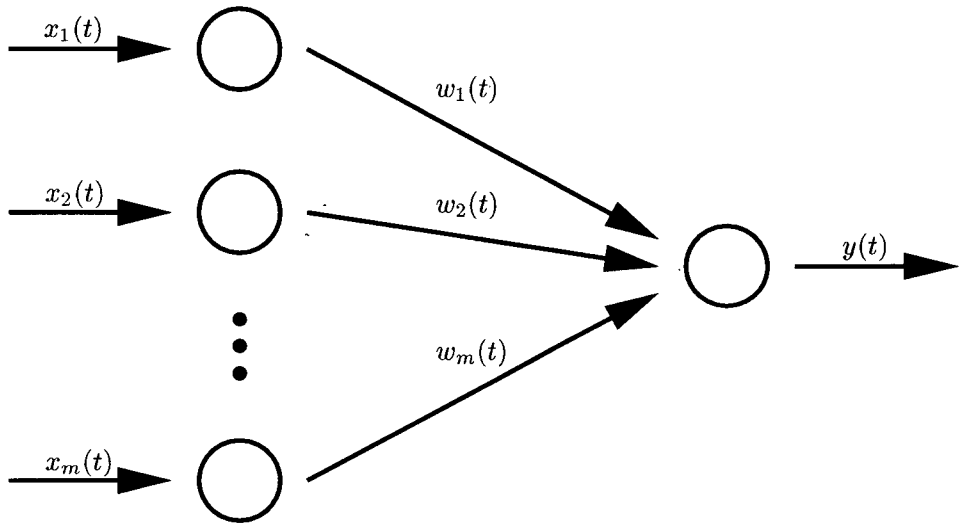


Figure 2.1: A schematic of a neural network for computing the first principal component.

output, $y(t)$, such that

$$y(t) = \sum_{i=1}^{m} w_i(t)x_i(t).$$

For each realization of $\mathbf{x}(t)$ presented to the network, the weights, $\mathbf{w}(t)$, are updated according to a simple heuristic, called Hebb's learning rule, stating that if both an input neuron and the output neuron show activity simultaneously, then the weight connecting those two neurons should be increased. In other words,

$$w_i(t_{j+1}) = w_i(t_j) + \eta y(t_j) x_i(t_j)$$

where $\eta$ is some small scalar value. In this form the Hebbian learning rule, under certain circumstances, is a non-convergent algorithm. To circumvent this difficulty a modified Hebbian rule is used where competition between the weights is introduced through a normalization term [Oja, 1982],

$$w_i(t_{j+1}) = \frac{w_i(t_j) + \eta y(t_j) x_i(t_j)}{[\sum_{i=1}^{m} (w_i(t_j) + \eta y(t_j) x_i(t_j))]^{\frac{1}{2}}}. \tag{2.11}$$

Expanding equation (2.11) in a truncated power series gives [Oja, 1982]

$$w_i(t_{j+1}) \approx w_i(t) + \eta y(t_j)(x_i(t_j) - y(t_j) w_i(t_j)). \tag{2.12}$$

Amazingly, as Oja [1982] shows, Hebb's simple heuristic provides an algorithm which computes the first principal component of the data. That is, $\mathbf{w}(t_j) \to \mathbf{u}_1$ as $j \to \infty$ where $\mathbf{u}_1$ is the first eigenvector of $\mathbf{C}_x = \mathrm{E}\left(\mathbf{x}\mathbf{x}^T\right)$; and consequently, $y(t)$ converges to the first principal component. The proof follows from associating the learning rule with a set of ordinary differential equations (ODEs). Hence, the convergence analysis of the learning rule is transfered to the stability analysis of a set of ODEs. If $\delta\mathbf{w} = \mathbf{w}(t_{j+1}) - \mathbf{w}(t_j)$, then from equation (2.12),

$$\begin{aligned} \delta\mathbf{w} &= \eta y(t_j)\left(\mathbf{x}(t_j) - y(t_j)\,\mathbf{w}(t_j)\right) \\ &= \eta\left[\mathbf{x}(t_j)\,\mathbf{x}(t_j)^T\mathbf{w}(t_j) - \left(\mathbf{w}(t_j)^T\mathbf{x}(t_j)\,\mathbf{x}(t_j)^T\mathbf{w}(t_j)\right)\mathbf{w}(t_j)\right] \end{aligned} \tag{2.13}$$

where $y(t_j) = \mathbf{w}(t_j)^T\mathbf{x}(t_j) = \mathbf{x}(t_j)^T\mathbf{w}(t_j)$. Dividing equation (2.13) through by $\delta t = t_{j+1} - t_j$, and letting $\delta t \to 0$ and $\eta \to 0$ at comparable rates gives

$$\begin{aligned} \lim_{\delta t\to 0}\frac{\delta\mathbf{w}}{\delta t} &= \lim_{\delta t,\eta\to 0}\frac{\eta}{\delta t}\left[\mathbf{x}(t_j)\,\mathbf{x}(t_j)^T\mathbf{w}(t_j) - \left(\mathbf{w}(t_j)^T\mathbf{x}(t_j)\,\mathbf{x}(t_j)^T\mathbf{w}(t_j)\right)\mathbf{w}(t_j)\right] \\ \frac{d\mathbf{w}}{dt} &= \mathbf{x}(t)\,\mathbf{x}(t)^T\mathbf{w}(t) - \left(\mathbf{w}(t)^T\mathbf{x}(t)\,\mathbf{x}(t)^T\mathbf{w}(t)\right)\mathbf{w}(t). \end{aligned} \tag{2.14}$$

Taking the expectation of equation (2.14) with respect to the random vector $\mathbf{x}(t)$ yields

$$\begin{aligned} \frac{d\mathbf{w}}{dt} &= \mathrm{E}\left(\mathbf{x}(t)\,\mathbf{x}(t)^T\right)\mathbf{w}(t) - \left[\mathbf{w}(t)^T\,\mathrm{E}\left(\mathbf{x}(t)\,\mathbf{x}(t)^T\right)\mathbf{w}(t)\right]\mathbf{w}(t) \\ &= \mathbf{C}_x\mathbf{w} - \left(\mathbf{w}^T\mathbf{C}_x\mathbf{w}\right)\mathbf{w}. \end{aligned} \tag{2.15}$$

It is easy to see that the stability points of equation (2.15) are given by $\mathbf{w} = \mathbf{0}$ and $\mathbf{w} = \mathbf{u}_i$, $i = 1 \ldots m$ where $\mathbf{u}_i$ are the eigenvector of $\mathbf{C}_x$. That is,

$$\left.\frac{d\mathbf{w}}{dt}\right|_{\mathbf{w}=\mathbf{0}} = \mathbf{0}$$

and

$$\left.\frac{d\mathbf{w}}{dt}\right|_{\mathbf{w}=\mathbf{u}_i} = \lambda_i\mathbf{u}_i - \left(\mathbf{u}_i^T\lambda_i\mathbf{u}_i\right)\mathbf{u}_i = \lambda_i\mathbf{u}_i - \lambda_i\mathbf{u}_i = \mathbf{0}.$$

It can be shown, under certain conditions [e.g. Haykin, 1999, Ch. 8], that the stability point, $\mathbf{u}_1$, is the only one which exhibits local convergence. To illustrate this consider a two dimensional

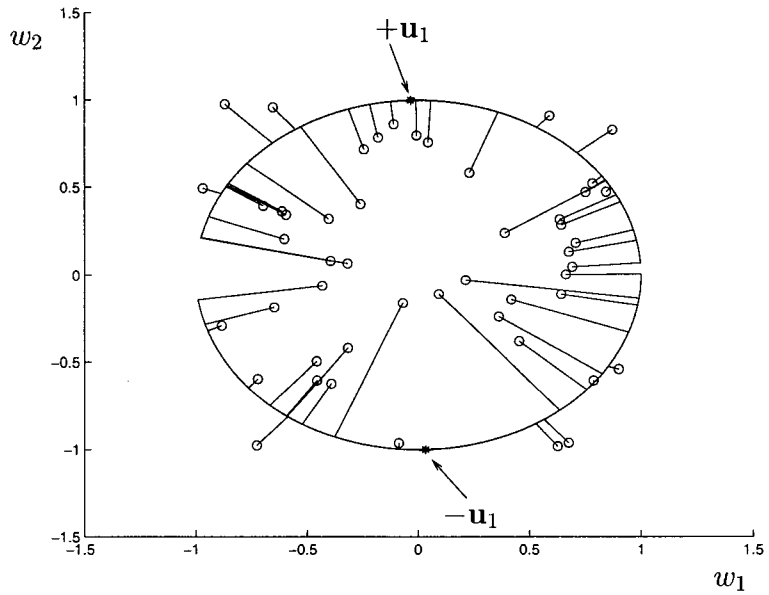


Figure 2.2: Solutions to the ODEs in equation (2.15) in phase space. Regardless of initial value (denoted by small circles), $\mathbf{w}(t_0)$, the solution converges to the first eigenvector of the correlation matrix of the data.

example where $x_1(t) = \sin(t)$ and $x_2(t) = \cos(t)$. Figure 2.2 plots trajectories of solutions, $\mathbf{w}(t)$, for various initial values. The initial values of the trajectories are denoted by small circles in the plot. Without fail the algorithm converges to $\pm\mathbf{u}_1$.

## 2.5 Application to Noise Suppression

Consider a hypothetical sub-surface consisting of perfectly horizontal and flat reflectors. A seismic survey is performed where the receiver spacing is kept perfectly constant. From such a survey seismic traces could, of course, be gathered into common midpoint sections (CMPs), each of which would contain the same signal but different realizations of random noise. For this situation a clever processing step would be to simply stack the CMP gathers, thus preserving the consistent signal while attenuating the unwanted random noise. However, the point of performing the survey in the first place is to find the nature of the sub-surface. We cannot add together the CMP gathers with the hope of reducing the noise in the prestack domain without first knowing something about the geometry of the sub-surface. Doing so would, of course, attenuate random noise; but more importantly, doing so would destroy signal.

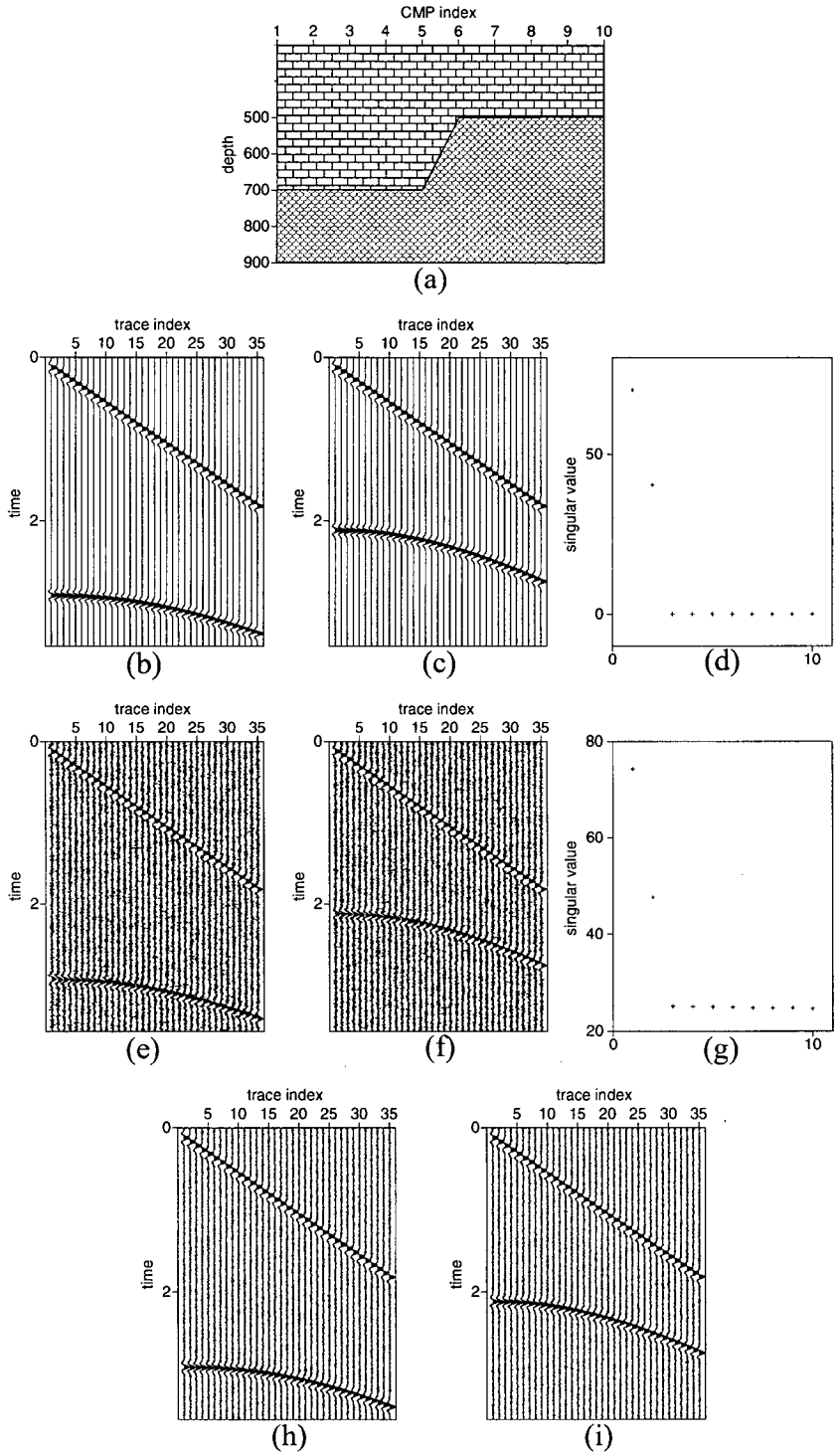


Figure 2.3: The synthetic example. (a) The reflector. Two of the CMPs (b)(c) without incoherent noise and (e)(f) with incoherent noise. (d) The singular values corresponding to the noiseless data, and (g) the noisy data. (h) The CMP in (e) projected onto the first two eigensections. (i) The CMP in (f) projected onto the first two eigensections.

Here, PCA is used to increase the signal to noise ratio in the prestack domain while respecting lateral variations in the subsurface. The method extends the work of Freire and Ulrych [1988] and is similar to methods studied in multispectral satellite imagery [Ready and Wintz, 1973; Richards, 1993, pp. 133-148] and face recognition software [Kirby and Sirovich, 1987; Pentland and Turk, 1991; Turk and Pentland, 1991]. From a collection of two dimensional common midpoint gathers the SVD computes a vector basis, the components of which are called eigensections[2] and which are trivially related to principal components. Projecting the seismic sections onto these eigensections attenuates the noise in the data.

In terms of PCA each seismic section is assigned to a random variable, $x_i(t)$, $i = 1 \ldots m$ where the realizations of each random variable are data in a seismic section (CMP). Hence, the data form the matrix, $\mathbf{A}$, as in equation (2.7) where each row, $x_i(t)$, holds data from a lexicographic reordered two dimensional seismic section. Computing the SVD of $\mathbf{A}$, as in equation (2.6), yields the orthonormal vector basis $\mathbf{v}_i$, $i = 1 \ldots m$ where $\mathbf{v}_i$ are called eigensections which are, in turn, trivially related to principal components (equation (2.8)). The projection of a row of $\mathbf{A}$ (a seismic section) onto the subspace spanned by the eigensections maps the vector representation of the section, $x_i(t)$, to a new vector of dimension $m$ via the relation

$$c_{ij} = \mathbf{v}_j^T x_i(t) \tag{2.16}$$

giving a new set of coordinates, $\mathbf{c}_i$, for each seismic section. Projecting the seismic sections, $\mathbf{A}$, onto the first $k$ eigensections gives the rank($k$) matrix, $\mathbf{A}_k$. Through elimination of the last $m - k$ eigensections from the projection, $\mathbf{A}_k$ is the approximation to $\mathbf{A}$ with the most incoherent information (the random noise) between the rows of $\mathbf{A}$ (the seismic sections) removed.

Here, two examples are considered. First, a simple but instructive synthetic (toy) example shows the removal of incoherent noise in the CMP domain. Second, the lessons learned in the synthetic example are applied to real data.

Consider ten CMP gathers recorded at evenly spaced points where the sub-surface topography, shown in Figure 2.3a, consists of a whole space over a half space. Because of the simplicity of this synthetic example (complications caused by the sudden change in depth of the impedance bound-

[2] A similar vector basis is computed in face recognition systems, the components of which are called eigenfaces.

ary, multiple reflections, and so on are neglected), the two CMP gathers shown in Figures 2.3b-c represent all variations in the signal. Applying the eigensection technique to these data produces ten eigensections ($\mathbf{v}_i$, $i = 1 \ldots 10$) along with ten singular values ($\sigma_i$, $i = 1 \ldots 10$) shown in Figure 2.3d. Notice that there are only two non-zero singular values. Thus, the first two eigensections are the only significant ones (see equation (2.9)). These two eigensections span a subspace that contains all ten of the CMP gathers. Thus, the CMP gathers can be projected onto these two eigensections without loss of signal. In other words,

$$x_i(t) = c_{i1}\mathbf{v}_1 + c_{i2}\mathbf{v}_2$$

where $c_{i1}$ and $c_{i2}$ are given by equation (2.16).

Next, Gaussian distributed random noise is added to the data (Figures 2.3e-f) and the eigensections are computed. The random noise is distributed throughout all of the eigensections. Thus, while the singular values do not decay to zero, they do decay to some horizontal asymptote (Figure 2.3g). The singular values that fall close to this asymptote represent incoherent noise. Therefore, this noise can be filtered by eliminating these undesired eigensections from the basis and projecting the CMP sections onto this reduced basis,

$$\hat{x}_i(t) = c_{i1}\mathbf{v}_1 + c_{i2}\mathbf{v}_2$$

where, again, $c_{i1}$ and $c_{i2}$ are given by equation (2.16). Figures 2.3h-i show the result of this filtering.

The extension of the eigensection technique to real data is straight forward. Figure 2.5a plots nine CMPs to be considered in this real data example. Each CMP consists of twenty traces. For dramatic effect, Gaussian distributed random noise is added to the data. As already mentioned, the goal is to increase the signal to noise ratio of these data in the prestack domain without making assumptions about the consistency of the sub-surface. The eigensection technique allows exactly this.

Figure 2.5b plots $\zeta_i(t) = \sigma_i \mathbf{v}_i$, $i = 1 \ldots 9$ where $\mathbf{v}_i$ are eigensections computed from the data in Figure 2.5a and $\sigma_i$ are the corresponding singular values. Of course, for plotting purposes, the vectors, $\mathbf{v}_i$, are re-organized into their original two dimensional form via the inverse of the lexicographic reordering. The singular values are plotted in Figure 2.4.

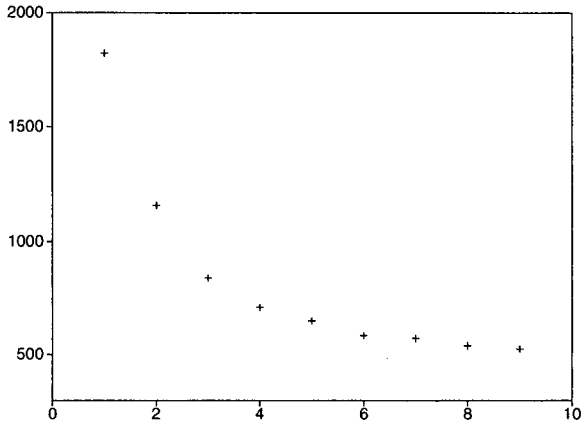


Figure 2.4: Real data example. Singular values associated with the respective eigensections of Figure 2.5b.

Though examination of Figures 2.4 and 2.5b, signal is contained in more than just the first eigensection. This means that there is variation amongst all the CMPs and simply stacking them to enhance the signal to noise ratio in the prestack domain would not be appropriate. However, PCA can increase the signal to noise ratio by considering the projection of a CMP onto a subset of the eigensections.

Figure 2.6 plots the fourth CMP of Figure 2.5a along with a stack of the CMPs and the approximation to the fourth CMP using projections onto various combinations of eigensections. In particular, plotted are

$$\hat{\mathbf{A}}_i' = \sum_{j=1}^{K} c_{ij}\mathbf{v}_j$$

for $K = 1$ through $K = 4$ in Figures 2.6c-f respectively where $c_{ij}$ are given by equation (2.16).

As in the previous synthetic example, the singular values (Figure 2.4) indicate how many eigensections should be included in the basis. Figure 2.6 shows that as more eigensections are used in the reconstruction, signal that was destroyed in the stack reappears. In particular, two hyperbolic events, delineated by arrows, illustrate this point. Of course, one cannot get something for nothing. When more eigensections are used in the reconstruction of the original CMP both signal and noise are added to the final sum.

A key point in the analysis outlined in this section is that each CMP has its own set of coordinates, $\mathbf{c}_i$, and so when reconstructed from the eigensections, any and all of the CMPs in Figure 2.5a are recovered uniquely.

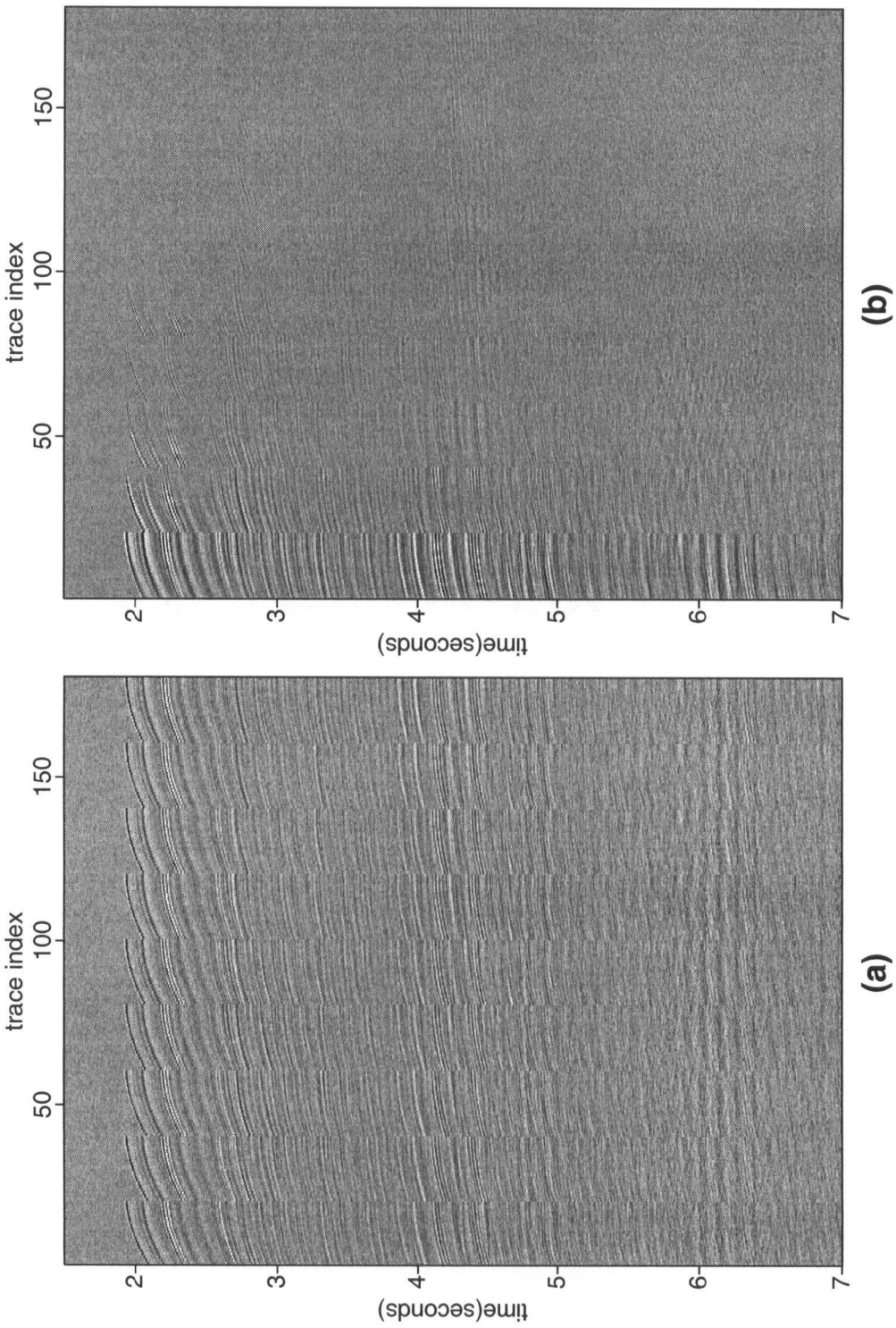


Figure 2.5: Real data example. (a) CMPs and (b) eigensections computed from the data in (a) and weighted by appropriate singular values. The singular values are plotted in Figure 2.4.

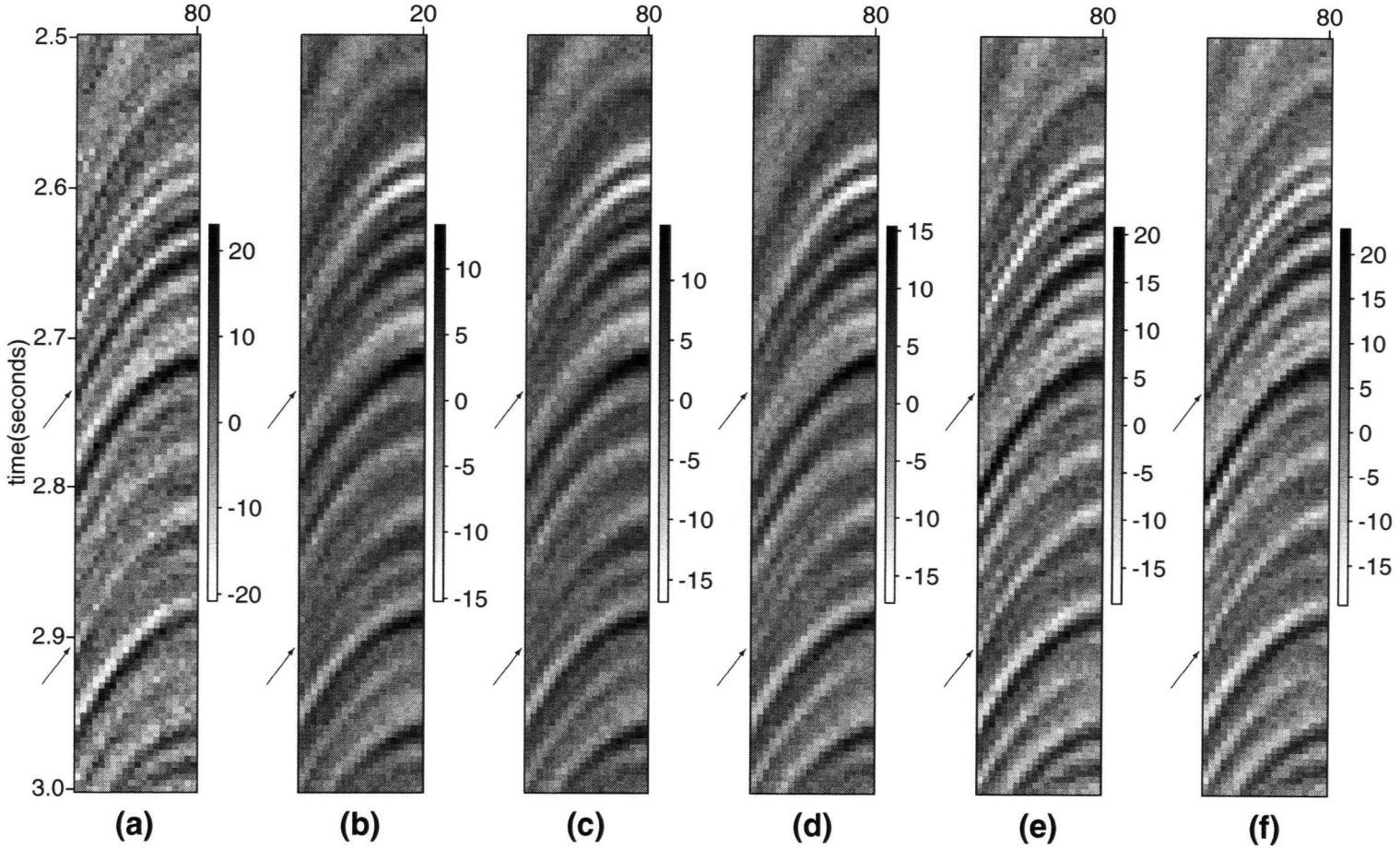


Figure 2.6: Real data example. (a) The fourth CMP, (b) the stack, and the CMPs reconstruction using (c) one eigensection, (d) two eigensections, (e) three eigensections and (f) four eigensections. Notice the events identified by arrows.

## 2.6 Summary

This chapter presented PCA from three points of view. First PCA was derived using variance, illustrating a principal component's disposition for explaining variance in the data and proving that principal components are uncorrelated. Second, principal components were computed using the SVD, showing that *explaining variance* and *explaining data* are conceptually equivalent. Lastly, principal components were computed using a neural network with a Hebbian learning rule allowing for the online computation of principal components and an understanding of them in terms of ODEs. Regardless of the derivation, the result is a tool which is beneficial to signal and image processing techniques. Here, PCA was used as a coherency filter for the purpose of attenuating noise. In particular, a novel extension to three dimensions (time, offset and CMP gather) in seismic data processing was presented which allows for signal to noise enhancement in the seismic prestack domain.

In this thesis, the benefits of PCA extend beyond the scope of this chapter. In Chapter 3, PCA and its ability to find uncorrelated random variables play an essential role in independent component analysis (ICA). ICA uses this property of PCA to extract, from data, useful features.

# CHAPTER 3

# Independent Component Analysis

## 3.1 Introduction

Consider the linear system

$$\mathbf{As} = \mathbf{x} \tag{3.1}$$

such that $\mathbf{x}$ is data generated by applying the forward operator, $\mathbf{A}$, to a model, $\mathbf{s}$. Inverse theory provides methods for finding $\mathbf{s}$ given both $\mathbf{A}$ and $\mathbf{x}$. That is, the forward operator and the data allow for, one way or another, the reconstruction of a model. However, if only $\mathbf{x}$ is given while $\mathbf{A}$ and $\mathbf{s}$ are unknown, the problem becomes insolvable without additional information. Robinson [1957] introduced a solution to one such problem. Namely seismic deconvolution where the model is the seismic reflectivity, the data is the seismic trace and the forward operator is a circulant matrix generated from the seismic wavelet. To compensate for a lack of information (only the seismic trace is known), Robinson postulated a white reflectivity and a minimum phase wavelet. This extra information allows for the simultaneous reconstruction of the reflectivity and the minimum phase wavelet given only the trace. Thus, Robinson found a method for solving the linear inverse problem in equation (3.1) when both $\mathbf{A}$ and $\mathbf{s}$ are unknown, and the wavelet is minimum phase.

Now, consider an alternate set of prior knowledge, namely that the elements of the model are mutually independent, and that, at least, all but one of the components of $\mathbf{x}$ follow non-Gaussian statistics. No assumptions are made about $\mathbf{A}$. These assumptions lead, indirectly, to independent component analysis (ICA) [Common, 1994] which, once again, given only $\mathbf{x}$ in equation (3.1), allows

for the simultaneous reconstruction of both **A** and **s**. This chapter describes ICA and illustrates it with a simple (toy) example. In particular, the concepts of independence and Gaussianity are related though the central limit theorem (CLT). Entropy, an indirect measure of Gaussianity and independence, is explained and efficient methods for its computation are described. The use of entropy in the ICA problem gives rise to an objective function whose extrema are trivially related to the inverse of **A**, and which is further constrained through the use of principal component analysis (PCA) (see Chapter 2).

## 3.2 The ICA Model

In what follows, the components of $\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$ will be referred to as mixtures produced by applying a square and nonsingular mixing matrix, **A**, to sources, $\mathbf{s}^T = \begin{bmatrix} s_1 & s_2 & \cdots & s_m \end{bmatrix}$. As in Chapter 2, the components of both the source and mixture vectors are treated as random variables so that, for example, $x_i(t_j)$ is the $j^{th}$ realization of the $i^{th}$ mixture. Hence, the ICA model is

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1m}s_m(t), \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + \cdots + a_{2m}s_m(t), \\ &\vdots \\ x_m(t) &= a_{m1}s_1(t) + a_{m2}s_2(t) + \cdots + a_{mm}s_m(t) \end{aligned}$$

with the mixing matrix,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mm}. \end{bmatrix}. \tag{3.2}$$

Additionally, define a matrix, **B**, so that

$$\mathbf{y} = \mathbf{B}\mathbf{x} \tag{3.3}$$

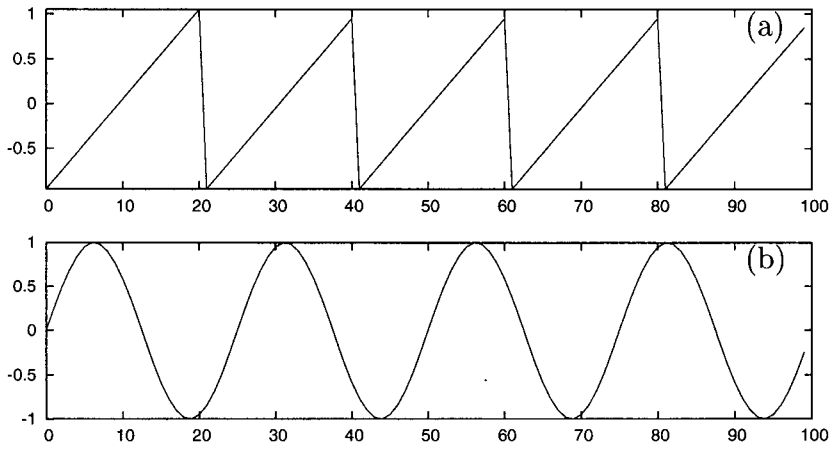


Figure 3.1: Independent and non-Gaussian sources. (a) $s_1(t)$ and (b) $s_2(t)$.

where $\mathbf{y}^T = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}$, $y_i = \mathbf{b}_i^T \mathbf{x}$, $i = 1 \ldots m$ and $\mathbf{b}_i^T$ is the $i^{th}$ row of $\mathbf{B}$. The random variable, $y_i$, is an independent component exactly when $\mathbf{b}_i$ is found such that $y_i \propto s_j$ for some $j$. In other words, independent components are sought such that they are proportional to the sources. The vagueness in the proportionality between a source and an independent component is resolved by, arbitrarily, setting $\text{var}(y_i) = 1$. Additionally, for reasons of simplicity, which will become clear shortly, the ICA model assumes $\text{E}(y_i) = 0$. This assumption is trivial to apply since the mean of the mixture vector, $\mathbf{x}$, is easily set to zero, and, as in Chapter 2, $\text{E}(y_i) = \text{E}\left(\mathbf{b}_i^T \mathbf{x}\right) = \mathbf{b}_i^T \text{E}(\mathbf{x})$.

To further illustrate the ICA model consider Figures 3.1a-b which plot one hundred realizations of $s_1(t)$ and $s_2(t)$ respectively for some $s_1$ and $s_2$. Applying a mixing matrix to these sources produces, for example,

$$\begin{aligned} \mathbf{x} &= \mathbf{A}\mathbf{s} \\ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 1.0 & 1.1 \\ 1.2 & 1.3 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \end{aligned}$$

where $x_1(t)$ and $x_2(t)$ are plotted in Figures 3.2a-b respectively. How $\mathbf{s}$ is found from just $\mathbf{x}$ is the subject of ICA and this chapter.

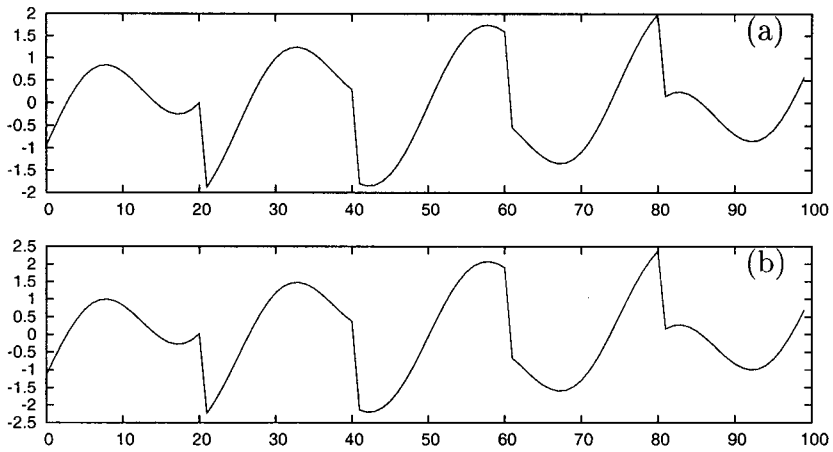


Figure 3.2: Mixtures produced by taking linear combinations of the sources, $s_i(t)$, in Figure 3.1. (a) $x_1(t)$ and (b) $x_2(t)$.

## 3.3 The CLT, Non-Gaussianity and Independence

The CLT plays an essential role in understanding the workings of ICA. In particular, it behooves us to understand the relation between the CLT, Gaussianity and independence. Doing so illustrates the basic principals of ICA and leads, indirectly, to an appropriate algorithm.

The CLT is stated as follows. Let $s_1, s_2, \ldots, s_m$ be independent and identically distributed (iid) random variables with variance $\sigma^2$ and mean 0. If

$$y = \sum_{i=1}^{m} s_i,$$

then

$$\lim_{m\to\infty} P\left(\frac{y}{\sigma\sqrt{m}} \leq y'\right) = F_Y\left(Y < y'\right) \quad , \quad -\infty < y' < \infty$$

where $F_Y$ is the cumulative distribution function for a standard Gaussian random variable [Rice, 1995, pp. 166-173]. Hence, if $s_i$ are non-Gaussian and iid random variables, then the sum, $y$, is more Gaussian then the parts, $s_i$.

Further, from the ICA model presented in equations (3.3) and (3.1),

$$\mathbf{y} = \mathbf{Bx} = \mathbf{BAs} = \mathbf{Ds}$$

where $\mathbf{D} = \mathbf{BA}$. Hence, the $i^{th}$ independent component is

$$y_i = \mathbf{d}_i^T \mathbf{s} \tag{3.4}$$

where $\mathbf{d}_i^T$ is the $i^{th}$ row of $\mathbf{D}$. Given equation (3.4), iid and zero mean sources, and the CLT, the following two points relate Gaussianity and independent components.

- If an independent component is sought such that $y_i \propto s_j$, then $\mathbf{d}_i$ must only have one nonzero component.
- Further, if $y_i$ is sought such that it is as non-Gaussian as possible, then $\mathbf{d}_i$ must have only one nonzero component; otherwise, at least two random variables are summed producing a more Gaussian result.

Hence, $y_i = \mathbf{b}_i^T \mathbf{x} = \mathbf{d}_i^T \mathbf{s}$ is an independent component exactly when it is maximally non-Gaussian. This, in turn, means that ICA requires some measure of Gaussianity.

The CLT presented above is the Lévy Theorem. It requires iid sources and, although providing a CLT with an easily understood proof, is therefore rather restrictive. However, the Lévy Theorem is a specific case of the more general Lindeberg Theorem which provides a CLT for a sequence of independent random variables with finite variances. Even more general forms of the CLT exist for independent random variables which require no assumptions about the existence of moments [Petrov, 2000]. Additionally, the assumption of independence can be weakened, leading to the aptly named *weak dependence conditions* [Sunklodas, 2000]. These conditions drop the requirement of mutual independence in favor of independence between sets of the random variables. Regardless, for the purpose of this thesis, one can assume that ICA seeks out independent components which are, indeed, mutually independent. However, the *weak dependence conditions* imply that the assumption of mutual independence, for ICA, is sufficient but not absolute.

## 3.4 Entropy and Gaussianity

In the previous section Gaussianity was used as a measure of independence. Here, a measurement central to information theory, called entropy for discrete random variables and differential entropy for continuous random variables [Shannon, 1948], is considered. It is well known that if only the

mean and variance of a continuous random variable, $y$, are given, then $y$ has maximum differential entropy exactly when it has Gaussian statistics.

Entropy, $H(p_1, p_2, \ldots)$, measures the randomness (disorder) of a discrete random variable, $Y$, where $p(Y = Y_i) = p_i$ is a probability mass function. As such, entropy must satisfy the following conditions [Jaynes, 1995, Ch. 11].

- $H(p_1, p_2, \ldots)$ exists and provides a relation between real numbers and uncertainty such that if there are many possibilities, entropy is *large*; and conversely, if there are few possibilities, entropy is *small.*
- $H(p_1, p_2, \ldots)$ is a continuous function of $p_i$.
- $H(p_1, p_2, \ldots)$ is consistent such that if there are multiple derivations, each arrives at the same measure.

It can be shown that [Jaynes, 1995, Ch. 11]

$$H(p_i) = -\sum_i p_i \log_b p_i \tag{3.5}$$

fulfills these requirements.[1]

For example consider, as Cover and Thomos [1991, pp. 14-15] do, an experiment that has two possible outcomes with corresponding probabilities $p$ and $1-p$ (a Bernoulli distribution). Following equation (3.5) the entropy of this experiment,

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p), \tag{3.6}$$

is plotted in Figure 3.3 for $p \in [0, 1]$. If $p = 1/2$, then the experiment is in a state of maximum disorder or maximum uncertainty, and so, entropy (a measure of uncertainty) is maximum. Conversely, if the outcome of the experiment is more certain, then $p$ is either closer to 0 or closer to 1, and entropy is smaller.

For continuous random variables an analogous measure called differential entropy, $h(p_Y)$, is

[1] When the base two logarithm is used ($b = 2$), the units of entropy are bits. So called because of its relation with coding lengths and binary numbers.

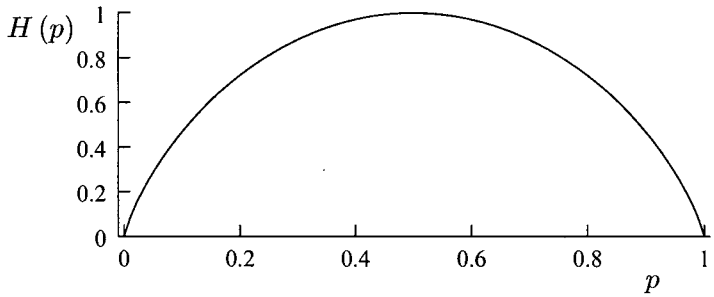


Figure 3.3: $H(p)$ as in equation (3.6) plotted with $p \in [0, 1]$.

used such that

$$h\left(p_Y\right) = -\int_{-\infty}^{\infty} p_Y(y) \ln p_Y(y) dy \tag{3.7}$$

where, in the context of ICA, $y$ is an independent component. The subscript $i$ is dropped for the sake of clarity.

It follows that distributions can be found which maximize entropy. Indeed, for the simple discrete example presented above, a Bernoulli distribution with $p = 1/2$ maximizes entropy. Of course, this is for a rather limited scenario where the experiment has only two possible outcomes. Consider, instead, maximizing the differential entropy of a continuous random variable, $y \sim p_Y(y)$, satisfying the usual conditions,

$$p_Y(y) \geq 0 \quad , \quad y \in \mathcal{R}, \tag{3.8}$$

$$\int_{-\infty}^{\infty} p_Y(y) dy = 1 \tag{3.9}$$

and the moment constraints provided by $r_i(y)$ and $c_i$ such that

$$\int_{-\infty}^{\infty} r_i(y) p_Y(y) dy = c_i \quad , \quad i = 1 \ldots l. \tag{3.10}$$

Hence, the appropriate cost function (for maximization) is

$$\begin{aligned} \phi\left(p_Y\right) &= h\left(p_Y\right) + \lambda_\circ \left( \int_{-\infty}^{\infty} p_Y\left(y'\right) dy' - 1 \right) + \sum_{i=1}^{l} \lambda_i \left( \int_{-\infty}^{\infty} r_i\left(y'\right) p_Y\left(y'\right) dy' - c_i \right) \\ &= -\int_{-\infty}^{\infty} p_Y\left(y'\right) \ln p_Y\left(y'\right) dy' + \lambda_\circ \left( \int_{-\infty}^{\infty} p_Y\left(y'\right) dy' - 1 \right) + \sum_{i=1}^{l} \lambda_i \left( \int_{-\infty}^{\infty} r_i\left(y'\right) p_Y\left(y'\right) dy' - c_i \right) \end{aligned}$$

where $\lambda_i$ are Lagrange multipliers. Differentiating with respect to the $y^{th}$ component of $p_Y$ gives

$$\frac{\partial \phi}{\partial p_Y} = -\ln p_Y(y) - 1 + \lambda_\circ + \sum_{i=1}^{p} \lambda_i r_i(y),$$

and setting this result to zero yields the extreme point of the cost function,[2]

$$p_Y(y) = \exp\left(-1 + \lambda_\circ + \sum_{i=1}^{l} \lambda_i r_i(y)\right). \tag{3.11}$$

In Section 3.2, independent components are assigned a mean of 0 and a variance of 1. Hence, $l = 2$ and equation (3.11) becomes

$$p_Y(y) = e^{\lambda_\circ - 1} e^{\lambda_1 y + \lambda_2 y^2}.$$

Setting $\lambda_\circ = \ln(2\pi)^{-1/2} + 1$, $\lambda_1 = 0$ and $\lambda_2 = 1/2$ yields a Gaussian distribution which satisfies the constraints in equations (3.8)-(3.10) and, hence, maximizes entropy.

Entropy gives a measure of Gaussianity in that minimizing entropy maximizes non-Gaussianity. Hence, $y$ is an independent component exactly when it has minimum entropy. Unfortunately, as is evident from equation (3.7), this creates the rather difficult task of estimating integrals of probability density functions (pdfs). Indeed, when only a finite sampling of the random variable is given to constrain the governing pdf, the task seems daunting. In the next two sections, two solutions to this problem are described, both, giving attainable approximations of entropy.

## 3.5 Entropy and Polynomial Expansions of pdfs

Here, to approximate entropy, a pdf is expanded on a set of basis functions, called Chebyshev-Hermite polynomials, derived from the Gaussian distribution. The resultant series, known as Gram-Charlier and Edgeworth expansions, are used in the definition of entropy, replacing the integration operator with expectations, and thus, allowing for efficient computations. In particular, equation (3.21), which estimates a measure called negentropy, will be derived such that maximizing negentropy is equivalent to minimizing entropy.

[2]Proof that the extreme point is a maximum is left for Appendix A.

Letting

$$\alpha(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{3.12}$$

define Chebyshev-Hermite polynomials, $H_i$, such that

$$(-1)^i \frac{d^i}{dy^i} \alpha(y) = H_i(y)\, \alpha(y). \tag{3.13}$$

Equations (3.12) and (3.13) define $H_j$ and allow for its explicit formulation. From equation (3.12),

$$\begin{cases} \frac{d}{dy}\alpha(y) &= -y\alpha(y) \\ \frac{d^2}{dy^2}\alpha(y) &= \left(y^2-1\right)\alpha(y) \\ \frac{d^3}{dy^3}\alpha(y) &= \left(-y^3+3\right)\alpha(y) \\ \frac{d^4}{dy^4}\alpha(y) &= \left(y^4-6y^2+3\right)\alpha(y) \\ \frac{d^5}{dy^5}\alpha(y) &= \left(-y^5+10y^3-15y\right)\alpha(y) \end{cases},$$

and hence, it follows from equation (3.13) that

$$\begin{cases} H_\circ &= 1 \\ H_1 &= y \\ H_2 &= y^2-1 \\ H_3 &= y^3-3y \\ H_4 &= y^4-6y^2+3 \\ H_5 &= y^5-10y^3+15y \end{cases},$$

yielding a pattern so that, in general [Kendall and Stuart, 1977, p. 167],

$$H_i(y) = y^i - \frac{i^{[2]}}{2\,(1!)} y^{i-2} + \frac{i^{[4]}}{2^2\,(2!)} y^{i-4} - \frac{i^{[6]}}{2^3\,(3!)} y^{i-6} + \cdots \tag{3.14}$$

where

$$i^{[n]} = (i)\,(i-1)\cdots(i-(n-1)).$$

Additionally, it can be shown [Kendall and Stuart, 1977, p. 168] that

$$\int_{-\infty}^{\infty} H_i(y) H_j(y) \alpha(y) dy = \begin{cases} 0 & , \quad i \neq j \\ j! & , \quad i = j \end{cases} . \tag{3.15}$$

Expanding an arbitrary pdf, $p_Y(y)$, on a basis of functions comprised of Chebyshev-Hermite polynomials yields the Gram-Charlier series of type A,

$$p_Y(y) = \sum_{i=0}^{\infty} c_i H_i(y) \alpha(y) \tag{3.16}$$

where, due to the orthogonality properties in equation (3.15),

$$c_i = \frac{1}{i!} \int_{-\infty}^{\infty} p_Y(y) H_i(y) dy. \tag{3.17}$$

Substituting equation (3.14) into equation (3.17) gives

$$\begin{aligned} c_i &= \frac{1}{i!} \int_{-\infty}^{\infty} p_Y(y) \left( y^i - \frac{i^{[2]}}{2(1!)} y^{i-2} + \frac{i^{[4]}}{2^2(2!)} y^{i-4} - \frac{i^{[6]}}{2^3(3!)} y^{i-6} + \cdots \right) dy \\ &= \frac{1}{i!} \left[ \mathrm{E}(y^i) - \frac{i^{[2]}}{2(1!)} \mathrm{E}(y^{i-2}) + \frac{i^{[4]}}{2^2(2!)} \mathrm{E}(y^{i-4}) - \frac{i^{[6]}}{2^3(3!)} \mathrm{E}(y^{i-6}) + \cdots \right], \end{aligned}$$

and combining this result with the first four terms ($j = 0 \ldots 3$) in equation (3.16) yields

$$p_Y(y) = \alpha(y) \left\{ H_\circ + \mathrm{E}(y) H_1 + \frac{1}{2} \left[ \mathrm{E}(y^2) - 1 \right] H_2(y) + \frac{1}{6} \left[ \mathrm{E}(y^3) - 3\mathrm{E}(y) \right] H_3(y) + \frac{1}{24} \left[ \mathrm{E}(y^4) - 6\mathrm{E}(y^2) + 3 \right] H_4(y) + O(y^5) \right\}.$$

However, if $y$ is an independent component, $\mathrm{E}(y) = 0$ and $\mathrm{E}(y^2) = 1$; hence, noting that $H_\circ = 1$,

$$p_Y(y) = \alpha(y) \left( 1 + \frac{1}{6} \kappa_3 H_3(y) + \frac{1}{24} \kappa_4 H_4(y) + O(y^5) \right) \tag{3.18}$$

where $\kappa_3 = \mathrm{E}(y^3)$ and $\kappa_4 = \mathrm{E}(y^4) - 3\left[\mathrm{E}(y^2)\right]^2$ are, respectively, the skewness and kurtosis of the zero mean random variable, $y$ [e.g. Nikias and Mendel, 1993].

Equation (3.18) is used in the definition of differential entropy (equation (3.7)) yielding an

estimate of a related measure called negentropy, $J(p_Y(y))$, such that

$$J(p_Y(y)) = h(p_Y(y)) - h(p_\xi(\xi)) \tag{3.19}$$

where $\xi \sim N(0,1)$. That is, $\xi$ has Gaussian statistics with the same mean and variance as $y$. Hence, as entropy does, negentropy gives a measure of Gaussianity; in particular, the distance from a Gaussian distributed random variable. Substituting equation (3.18) into the definition of differential entropy (equation (3.7)), and neglecting higher order terms gives

$$h(p_Y(y)) \approx -\int_{-\infty}^{\infty} \alpha(y)\left[1+\frac{1}{6}\kappa_3 H_3(y)+\frac{1}{24}\kappa_4 H_4(y)\right]\cdot \left[\ln\alpha(y)+\ln\left(1+\frac{1}{6}\kappa_3 H_3(y)+\frac{1}{24}\kappa_4 H_4(y)\right)\right]dy. \tag{3.20}$$

Further, applying the first two terms in the Taylor series, $\ln(\epsilon) = \epsilon - \epsilon^2/2 + \cdots$, to equation (3.20) yields

$$h(p_Y(y)) \approx -\int_{-\infty}^{\infty} \alpha(y)\left(1+\frac{1}{6}\kappa_3 H_3(y)+\frac{1}{24}\kappa_4 H_4(y)\right)\cdot \left[\ln\alpha(y)+\frac{1}{6}\kappa_3 H_3(y)+\frac{1}{24}\kappa_4 H_4(y)-\frac{1}{2}\left(\frac{1}{6}\kappa_3 H_3(y)+\frac{1}{24}\kappa_4 H_4(y)\right)^2\right]dy$$

which, using the orthogonality constraints in equation (3.15), becomes

$$h(p_Y(y)) \approx -\int_{-\infty}^{\infty} \alpha(y)\ln\alpha(y)\,dy + \frac{\kappa_3^2(y)}{12}+\frac{\kappa_4^2(y)}{48}.$$

Therefore, it follows from equation (3.19) that [Jones and Sibson, 1987]

$$J(p_Y(y)) \approx \frac{\kappa_3^2(y)}{12}+\frac{\kappa_4^2(y)}{48} \tag{3.21}$$

which is relatively simple to compute.

To review, the random variable $y$ is an independent component exactly when it is maximally non-Gaussian, or equivalently, when its negentropy is maximum. However, while the approximation in equation (3.21) is useful (as will be illustrated shortly), it can also be problematic in that

any outliers in a sample of $y$ adversely effects the approximation of the cumulants (kurtosis and skewness) [Hyvärinen et al., 2001, p. 182]. A more accurate measure of negentropy, described anon, greatly improves the robustness of ICA.

## 3.6 Entropy and Nonpolynomial Expansions of pdfs

Hyvärinen [1998] introduced an alternative to the polynomial expansions, using a basis of nonpolynomial functions, which greatly reduces the effect of outliers in the approximation of negentropy. In Section 3.4 the maximum entropy distribution,

$$p_Y(y) = \exp(-1 + \lambda_\circ) \exp\left(\sum_{i=1}^{l} \lambda_i r_i(y)\right), \tag{3.22}$$

was derived and required to satisfy the moment constraints,

$$\int_{-\infty}^{\infty} r_i(y)\, p_Y(y) dy = c_i \quad , \quad i = 1 \ldots l. \tag{3.23}$$

Here, equations (3.22) and (3.23) are used to find an estimate of negentropy. The end result is equation (3.35).

Since $y$ is an independent component, $\mathrm{E}(y) = 0$ and $\mathrm{E}(y^2) = 1$; consequently, appropriate moment constraints are obtained by letting $r_1 = y$, $r_2 = y^2$, $c_1 = 0$ and $c_2 = 1$. Hence, equation (3.22) becomes

$$\begin{aligned} p_Y(y) &= \exp(-1 + \lambda_\circ) \exp\left(\lambda_1 y + \lambda_2 y^2 + \sum_{i=3}^{l} \lambda_i r_i(y)\right) \\ &= \exp(-1 + \lambda_\circ) \exp\left[-\frac{y^2}{2} + \lambda_1 y + \left(\lambda_2 + \frac{1}{2}\right) y^2 + \sum_{i=3}^{l} \lambda_i r_i(y)\right] \\ &= \sqrt{2\pi}\alpha(y) \exp(-1 + \lambda_\circ) \exp\left[\lambda_1 y + \left(\lambda_2 + \frac{1}{2}\right) y^2 + \sum_{i=3}^{l} \lambda_i r_i(y)\right] \end{aligned} \tag{3.24}$$

where $\alpha(y)$ has the form of a standard Normal distribution as in equation (3.12). Applying the

first two terms in the Taylor series, $e^{\epsilon} = 1 + \epsilon + \cdots$, to equation (3.24) gives

$$p_Y(y) \approx A\left[1 + \lambda_1 y + \left(\lambda_2 + \frac{1}{2}\right) y^2 + \sum_{i=3}^{l} \lambda_i r_i(y)\right] \tag{3.25}$$

where $A = \sqrt{2\pi}\alpha(y)\exp(-1+\lambda_\circ)$. Additionally, it is useful to assume that $r_i(y)$ follow the orthogonality constraints [Hyvärinen, 1998],

$$\int_{-\infty}^{\infty} \alpha(y) r_i(y) r_j(y) dy = \begin{cases} 0 & , \quad i \neq j \\ 1 & , \quad i = j \end{cases} \tag{3.26}$$

and

$$\int_{-\infty}^{\infty} \alpha(y) r_i(y) y^k dy = 0 \quad , \quad k = 0, 1, 2 \quad , \quad i = 1 \ldots l. \tag{3.27}$$

Inserting equation (3.25) into $\int p_Y dy = 1$ and the prescribed moment constraints in equation (3.23); and using the orthogonality constraints in equations (3.26) and (3.27), a system of algebraic equations is derived which, when solved, greatly simplifies equation (3.25). In particular, letting $A = \tilde{A}\alpha(y)$, this scheme yields

$$\int_{-\infty}^{\infty} p_Y(y) dy = 1 = \tilde{A}(1 + \lambda_2 + 1/2), \tag{3.28}$$

and for the zero mean constraint ($i = 1$ in equation (3.23)),

$$\int_{-\infty}^{\infty} p_Y(y) y dy = 0 = \tilde{A}\lambda_1. \tag{3.29}$$

Third, using the unit variance constraint ($i = 2$ in equation (3.23)),

$$\int_{-\infty}^{\infty} p_Y(y) y^2 dy = 1 = \tilde{A}\left[1 + (\lambda_2 + 1/2)\,\mathrm{E}\left(\xi^4\right)\right]$$

where $\xi \sim N(0,1)$. Further, since $\kappa_4(\xi) = 0 = \mathrm{E}\left(\xi^4\right) - 3\mathrm{E}\left(\xi^2\right)$, $\mathrm{E}\left(\xi^4\right) = 3\mathrm{E}\left(\xi^2\right) = 3$; and, as a result,

$$\tilde{A}\left[1 + 3(\lambda_2 + 1/2)\right] = 1. \tag{3.30}$$

Again, inserting equation (3.25) into the moment constraints in equation (3.23) for $i = 3 \ldots l$,

$$\int_{-\infty}^{\infty} p_Y(y)\, r_j(y) dy = c_j = \tilde{A} \sum_{i=3}^{l} \lambda_i \int_{-\infty}^{\infty} \alpha(y)\, r_i(y)\, r_j(y) dy = \tilde{A}\lambda_j. \tag{3.31}$$

Equations (3.28)-(3.31) provide a system of algebraic equations which are easily solved yielding $\tilde{A} = 1$, $\lambda_1 = 0$, $\lambda_2 = -1/2$ and $\lambda_j = c_j$. Hence, from equation (3.25),

$$p_Y(y) = \alpha(y) \left( 1 + \sum_{i=3}^{l} c_i r_i(y) \right).$$

Combining this result with differential entropy gives

$$\begin{aligned}
h(p_Y(y)) &= -\int_{-\infty}^{\infty} p_Y(y) \ln p_Y(y) dy \\
&= -\int_{-\infty}^{\infty} \alpha(y) \left( 1 + \sum_{i=3}^{l} c_i r_i(y) \right) \ln \left[ \alpha(y) \left( 1 + \sum_{i=3}^{l} c_i r_i(y) \right) \right] dy \\
&= -\int_{-\infty}^{\infty} \alpha(y) \ln \alpha(y) dy \qquad (3.32) \\
&\quad -\int_{-\infty}^{\infty} \alpha(y) \sum_{i=3}^{l} c_i r_i(y) \ln \alpha(y) dy \qquad (3.33) \\
&\quad -\int_{-\infty}^{\infty} \alpha(y) \left( 1 + \sum_{i=3}^{l} c_i r_i(y) \right) \ln \left( 1 + \sum_{i=3}^{l} c_i r_i(y) \right) dy. \qquad (3.34)
\end{aligned}$$

The first term (3.32) is the differential entropy of a Gaussian random variable, $h\left(p_\xi(\xi)\right)$. The second term (3.33) is eliminated by noting that $\ln \alpha(y) = -\ln\left(\sqrt{2\pi}\right) - y^2/2$; hence,

$$\begin{aligned}
&-\int_{-\infty}^{\infty} \alpha(y) \sum_{i=3}^{l} c_i r_i(y) \ln \alpha(y) dy \\
&\qquad = \sum_{i=3}^{l} \left[ \frac{1}{2} c_i \int_{-\infty}^{\infty} \alpha(y)\, y^2 r_i(y) dy + c_i \ln\left(\sqrt{2\pi}\right) \int_{-\infty}^{\infty} \alpha(y)\, r_i(y) dy \right] = 0.
\end{aligned}$$

The third term (3.34) is simplified using the first two terms in the Taylor series, $(1+\epsilon)\ln(1+\epsilon) =$

$\epsilon + \epsilon^2/2 + \cdots$, such that

$$
\begin{aligned}
&-\int_{-\infty}^{\infty} \alpha(y) \left(1 + \sum_{i=3}^{l} c_i r_i(y)\right) \ln\left(1 + \sum_{i=3}^{l} c_i r_i(y)\right) dy \\
&\approx -\int_{-\infty}^{\infty} \alpha(y) \left[\sum_{i=3}^{l} c_i r_i(y) + \frac{1}{2}\left(\sum_{i=3}^{l} c_i r_i(y)\right)^2\right] dy \\
&= -\sum_{i=3}^{l} c_i \int_{-\infty}^{\infty} \alpha(y) r_i(y) dy - \frac{1}{2}\int_{-\infty}^{\infty} \left(\sum_{i=3}^{l} c_i^2 \alpha(y) r_i(y)^2 - 2\sum_{i=3}^{l-1}\sum_{j=4}^{l} c_i c_j \alpha(y) r_i(y) r_j(y)\right) dy \\
&= -\frac{1}{2}\sum_{i=3}^{l} c_i^2.
\end{aligned}
$$

Recombining these results produces

$$h(p_Y(y)) \approx h(p_\xi(\xi)) - \frac{1}{2}\sum_{i=3}^{l} c_i^2,$$

and so [Hyvärinen, 1998],

$$J(p_Y(y)) = h(p_\xi(\xi)) - h(p_Y(y)) \approx \frac{1}{2}\sum_{i=3}^{l} c_i^2 \tag{3.35}$$

where $c_i = \mathrm{E}(r_i(y))$.

The preceding derivation is long; but, given the simplicity and utility of the result, well worth the effort. What is left is to choose appropriate nonlinearities, $r_i(y)$, the choice imposing a distribution on $y$ (see equation (3.22)).

## 3.7 ICA and its Cost Function

Due to the relation between negentropy and independent components, the ICA problem is reduced to one in optimization with an associated cost function measuring negentropy. Two such measures are presented in Sections 3.5 and 3.6. Here, this optimization problem is explicitly defined, and the utility of PCA in terms of ICA is explained.

PCA, used as a pre-processor, allows for the derivation of much needed constraints for the

optimization problem. Given zero mean mixtures, $\mathbf{x}$, let

$$\mathbf{z} = \mathbf{\Gamma x}$$

where $\mathbf{z}^T = \begin{bmatrix} z_1 & z_2 & \cdots & z_m \end{bmatrix}$ are whitened mixtures such that $\mathrm{E}(\mathbf{z}) = \mathbf{0}$, $\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right) = \mathbf{C}_z = \mathbf{I}$ and $\mathbf{I}$ is the identity matrix. That is, the random variables, $z_i$, $i = 1 \ldots m$ are mutually uncorrelated. The utility of $\mathbf{z}$ is illustrated by understanding the relation between *uncorrelated* and *independent*. Namely that independent implies uncorrelated. Consider, for example, two random variables, $y_1$ and $y_2$, that follow the bivariate pdf, $p_{Y_1,Y_2}(y_1, y_2)$, with marginal pdfs, $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$. Also let $g_1(y_1)$ and $g_2(y_2)$ be arbitrarily defined functions. The random variables, $y_1$ and $y_2$, are, said to be, uncorrelated if

$$\mathrm{E}(y_1 y_2) = \mathrm{E}(y_1)\,\mathrm{E}(y_2).$$

Further, if $y_1$ and $y_2$ are independent, then $p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)\, p_{Y_2}(y_2)$. Thus,

$$\begin{aligned} \mathrm{E}[g_1(y_1)\, g_2(y_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(y_1)\, g_2(y_2)\, p_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 \\ &= \int_{-\infty}^{\infty} g_1(y_1)\, p_{Y_1}(y_1) dy_1 \int_{-\infty}^{\infty} g_2(y_2)\, p_{Y_2}(y_2) dy_2 \\ &= \mathrm{E}[g_1(y_1)]\,\mathrm{E}[g_2(y_2)]. \end{aligned} \tag{3.36}$$

Therefore, uncorrelated is a special case of independent where $g_1(y_1) = y_1$ and $g_2(y_2) = y_2$; and hence, independent implies uncorrelated but uncorrelated does not imply independent. Since the goal of ICA is to produce components that are independent, they are also uncorrelated and under orthogonal transformations they stay that way. Therefore, an appropriately chosen rotation transforms uncorrelated components into independent components. This immediately drops the degrees of freedom in the optimization problem by one.

From Chapter 2, and in particular equation (2.5), an appropriate choice for $\mathbf{\Gamma}$ is easily found such that

$$\mathbf{\Gamma} = \mathbf{\Sigma}^{-1}\mathbf{U}^T \tag{3.37}$$

where, as in Chapter 2, $\mathbf{\Sigma} = \mathrm{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_m}\right)$ and $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & | & \mathbf{u}_2 & | & \cdots & | & \mathbf{u}_m \end{bmatrix}$ where $\mathbf{u}_i$ and $\lambda_i$ are, respectively, the eigenvectors and eigenvalues of the covariance matrix of the

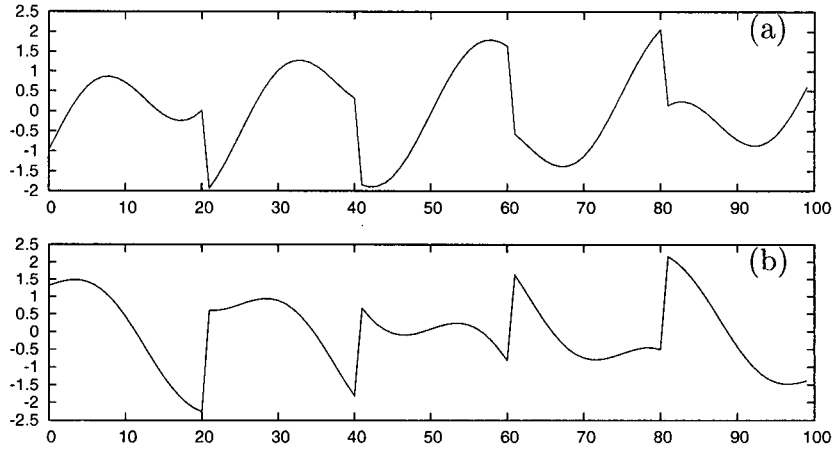


Figure 3.4: A whitened version of the mixtures in Figure 3.2. (a) $z_1(t)$ and (b) $z_2(t)$.

mixtures, $\mathbf{C}_x$. Recalling that $\mathbf{u}_i$ are mutually orthonormal,

$$\mathrm{E}\left(z_i z_j\right) = \mathrm{E}\left[\left(\frac{\mathbf{u}_i^T\mathbf{x}}{\sqrt{\lambda_i}}\right)\left(\frac{\mathbf{u}_j^T\mathbf{x}}{\sqrt{\lambda_j}}\right)^T\right] = \frac{\mathrm{E}\left(\mathbf{u}_i^T\mathbf{x}\mathbf{x}^T\mathbf{u}_j\right)}{\sqrt{\lambda_i\lambda_j}} = \frac{\mathbf{u}_i^T\mathbf{C}_x\mathbf{u}_j}{\sqrt{\lambda_i\lambda_j}} = \frac{\lambda_j\mathbf{u}_i^T\mathbf{u}_j}{\sqrt{\lambda_i\lambda_j}} = \begin{cases} 0 & , \quad i \neq j \\ 1 & , \quad i = j \end{cases}$$

confirms that equation (3.37) is a good choice for $\mathbf{\Gamma}$. The result of whitening the data in Figure 3.2 is shown in Figure 3.4.

With whitening, the ICA model becomes $\mathbf{\Gamma A s} = \mathbf{\Gamma x}$ or more succinctly,

$$\mathbf{W s} = \mathbf{z}$$

where $\mathbf{W} = \mathbf{\Gamma A}$. Additionally, define $\mathbf{Q}$ such that $\mathbf{y} = \mathbf{Q z}$, $\mathbf{q}_i^T$ is the $i^{th}$ row of $\mathbf{Q}$ and $y_i = \mathbf{q}_i^T\mathbf{z}$ is an independent component exactly when $\mathbf{q}_i$ is chosen such $y_i$ has maximum negentropy. Hence, an appropriate cost function (for minimization) is

$$\phi\left(\mathbf{q}_i\right) = -J\left(p_Y\left(y_i\right)\right) = -J\left(p_Y\left(\mathbf{q}_i^T\mathbf{z}\right)\right). \tag{3.38}$$

Figure 3.5 plots equation (3.38) for the whitened data in Figure 3.4 using the measure of negentropy defined in equation (3.21).

As already mentioned, whitening the data further constrains the cost function. In particular, recalling that $\mathrm{var}\left(y_i\right) = 1$, $\mathrm{E}\left(y_i\right) = 0$ and that the independent components are uncorrelated such

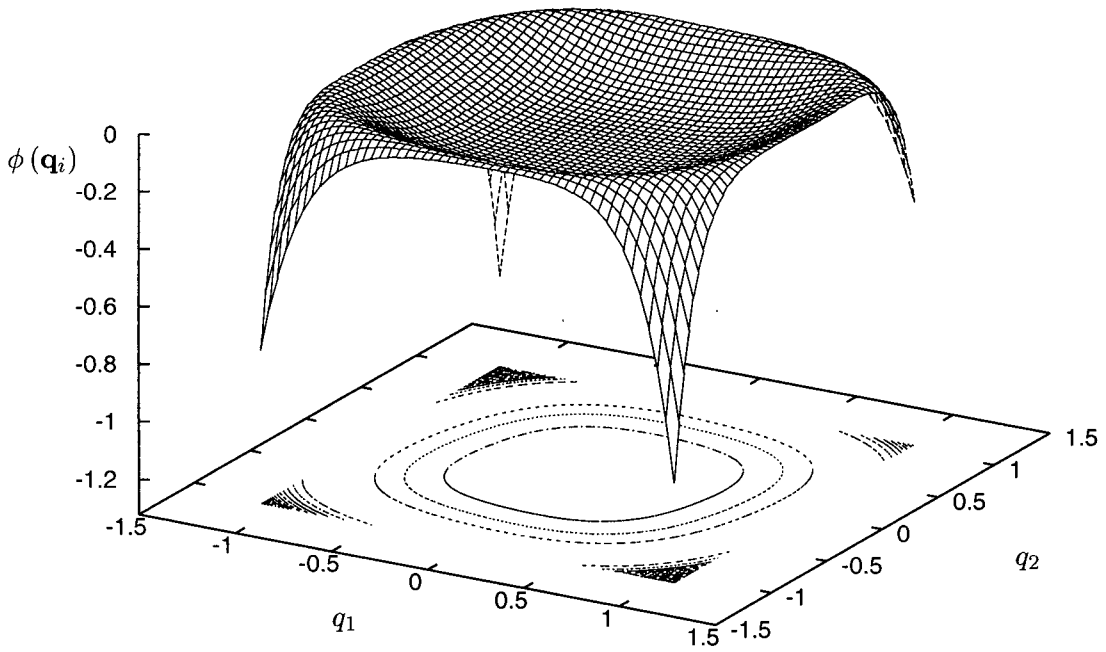


Figure 3.5: The cost function for ICA (negative negentropy) computed from the whitened mixtures in Figure 3.4 using the Gram-Charlier series.

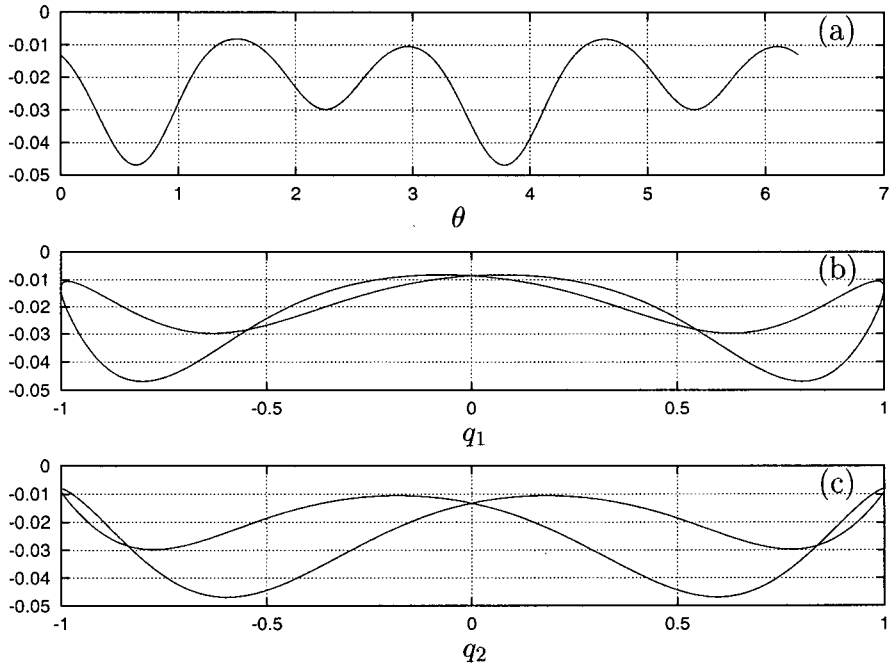


Figure 3.6: The cost function in Figure 3.5 for the unit circle plotted versus (a) $\theta$, (b) $q_1$ and (c) $q_2$.

that $\mathrm{E}(y_i y_j) = 0$, $i \neq j$ gives

$$\mathrm{E}(y_i y_j) = \mathrm{E}\left[\left(\mathbf{q}_i^T \mathbf{z}\right)\left(\mathbf{q}_j^T \mathbf{z}\right)^T\right] = \mathrm{E}\left(\mathbf{q}_i^T \mathbf{z}\mathbf{z}^T \mathbf{q}_j\right) = \mathbf{q}_i^T \mathbf{C}_z \mathbf{q}_j = \mathbf{q}_i^T \mathbf{q}_j = \begin{cases} 0 & , \quad i \neq j \\ 1 & , \quad i = j \end{cases}.$$

The result, $\mathbf{q}_i^T \mathbf{q}_i = 1$, means that the cost function in Figure 3.5 need only be considered on the unit circle of its domain. That is, the two degrees of freedom in $\phi\left(\mathbf{q}_i^T\right)$ for $\mathbf{q}_i^T = \begin{bmatrix} q_1 & q_2 \end{bmatrix}$ is reduced to one variable, $\theta$, such that $q_1 = \sin(\theta)$ and $q_2 = \cos(\theta)$. The cost function traced out along this unit circle is plotted in Figure 3.6. Notice that there are four distinct local minima in Figure 3.6a. Each one corresponds to an independent component. However, for this example there are only two sources and four independent components. Recalling that $y_i \propto s_j$ and remembering the constraint $\mathrm{var}(y_i) = 1$, it is clear that both $y_i$ and $-y_i$ satisfy the ICA definition. Hence, generalizing, the cost function always provides twice as many local minima as sources. Multiple local minima are found through consideration of the constraint, $\mathbf{q}_i^T \mathbf{q}_j = 0$, $i \neq j$. Indeed, using Gram-Schmidt orthogonalization it is trivial to find a $\mathbf{q}_j$ which is orthogonal to $\mathbf{q}_i$ [e.g. Bretscher, 1997, p. 201].

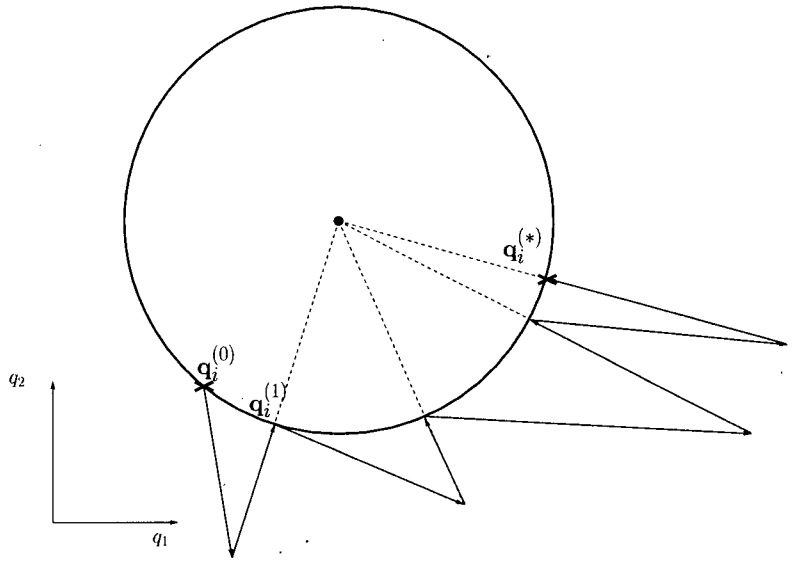


Figure 3.7: A simple optimization scheme for ICA. Given a point, $\mathbf{q}_i^{(0)}$, on the unit circle, a step is taken following the gradient of the cost function producing a new point which is projected back onto the unit circle giving $\mathbf{q}_i^{(1)}$. This is repeated until $||\mathbf{q}_i^{(k+1)} - \mathbf{q}_i^{(k)}||_2 < \epsilon$ where $\epsilon$ is some prescribed tolerance.

## 3.8 ICA Optimization Algorithms

The problem presented in the previous section requires, of course, a scheme for finding the extreme points of the cost function given the constraints. Here, two such methods are presented. The first using a gradient descent type algorithm and the second using a Newton type algorithm.

Figure 3.7 illustrates a simple gradient algorithm for finding one local minimum, and hence one independent component, of the cost function in equation (3.38) subject to the constraint, $\mathbf{q}_i^T \mathbf{q}_i = 1$. It is a modified gradient descent algorithm such that in the transition from iteration $k$ to $k+1$

$$\begin{aligned} \mathbf{q}_i^{(k+1)} &= \mathbf{q}_i^{(k)} - \alpha \nabla \phi \left( \mathbf{q}_i^{(k)} \right) \\ \mathbf{q}_i^{(k+1)} &\leftarrow \frac{\mathbf{q}_i^{(k+1)}}{||\mathbf{q}_i^{(k+1)}||_2} \end{aligned} \tag{3.39}$$

where $\alpha$ is some specified constant which governs the rate of convergence. Hence, each iteration produces a new vector, $\mathbf{q}_i^{(k+1)}$, by following the negative gradient of the cost function away from the unit circle and projecting this result back onto the unit circle. At a local minimum, $\mathbf{q}_i^{(*)}$, the gradient of the cost function is orthogonal to the unit circle [e.g. Nocedal and Wright, 1999, p. 320]; hence,

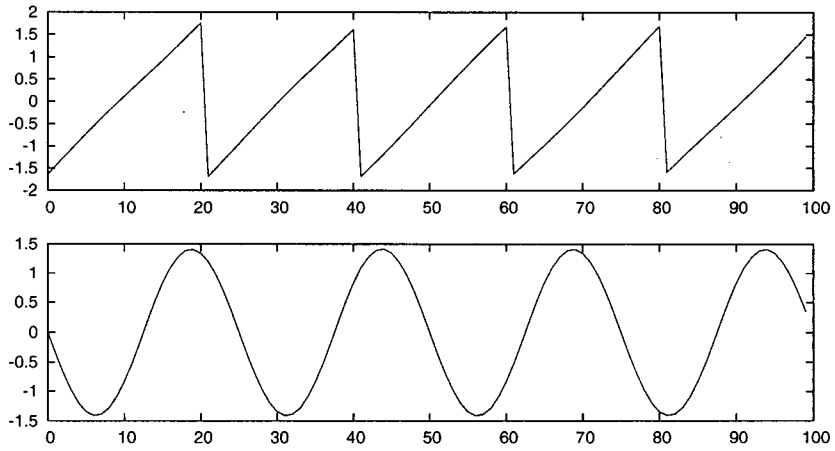


Figure 3.8: The independent components computed from the whitened mixtures in Figure 3.4. The solution is found using the polynomial approximation of negentropy in equation (3.21) and the gradient optimization scheme in equation (3.39).

$\mathbf{q}_i^{(k+1)} = \mathbf{q}_i^{(k)} = \mathbf{q}_i^{(*)}$ and the algorithm converges. Multiple independent components are readily found using Gram-Schmidt orthogonalization and choosing $\mathbf{q}_j^{(0)}$ such that it is orthogonal to $\mathbf{q}_i^{(*)}$ for $i = 1 \ldots j-1$. Figure 3.8 plots the independent components found from the mixtures in Figure 3.2 using the polynomial approximation of negentropy in equation (3.21) and the gradient descent optimization scheme in equation (3.39). It is obvious that the algorithm has found independent components. That is, it is obvious that $y_i \propto s_i$, $i = 1, 2$.

Hyvärinen [1999a] presents an alternative scheme to equation (3.39) which employs approximative Newton steps in the iterative scheme. In particular, the nonpolynomial approximation of negentropy in equation (3.35) is considered using only one term in its series expansion which, combined with the constraint $\mathbf{q}_i^T \mathbf{q}_i = 1$, gives, for minimization,

$$\phi(\mathbf{q}_i) = -\frac{1}{2}\left[\mathrm{E}\left(r\left(y_i\right)\right)\right]^2 + \lambda\left(\mathbf{q}_i^T \mathbf{q}_i - 1\right) \tag{3.40}$$

where $y_i = \mathbf{q}_i^T \mathbf{z}$ and $\lambda$ is a Lagrange multiplier. The gradient of $\phi$ is

$$\nabla\phi(\mathbf{q}_i) = -\mathrm{E}\left(r\left(y_i\right)\right)\mathrm{E}\left(r'\left(y_i\right)\mathbf{z}\right) + 2\lambda\mathbf{q}_i,$$

and ignoring the scalar value, $-\mathrm{E}\left(r\left(y_i\right)\right)$, allows for computation of an approximative Hessian, $\mathbf{H}$,

such that

$$\mathbf{H} \approx \mathrm{E}\left(r''\left(y_i\right)\mathbf{z}\mathbf{z}^T\right) + 2\lambda\mathbf{I} \approx \mathrm{E}\left(r''\left(y_i\right)\right)\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right) + 2\lambda\mathbf{I} = \left[\mathrm{E}\left(r''\left(y_i\right)\right) + 2\lambda\right]\mathbf{I}.$$

This approximation gives a Hessian which is easily inverted, leading to the approximative Newton step (from iteration $k$ to $k+1$) given by

$$\mathbf{q}_i^{(k+1)} = \mathbf{q}_i^{(k)} - \frac{\mathrm{E}\left(r'\left(y_i\right)\mathbf{z}\right) + 2\lambda\mathbf{q}_i^{(k)}}{\mathrm{E}\left(r''\left(y_i\right)\right) + 2\lambda}. \tag{3.41}$$

Multiplying equation (3.41) through by the denominator in its third term yields

$$\begin{aligned}\left[\mathrm{E}\left(r''\left(y_i\right)\right) + 2\lambda\right]\mathbf{q}_i^{(k+1)} &= \left[\mathrm{E}\left(r''\left(y_i\right)\right) + 2\lambda\right]\mathbf{q}_i^{(k)} - \mathrm{E}\left(r'\left(y_i\right)\mathbf{z}\right) - 2\lambda\mathbf{q}_i^{(k)} \\ &= \mathrm{E}\left(r''\left(y_i\right)\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(r'\left(y_i\right)\mathbf{z}\right).\end{aligned}$$

Hence, an appropriate algorithm is

$$\begin{aligned}\mathbf{q}_i^{(k+1)} &= \mathrm{E}\left(r''\left(y_i\right)\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(r'\left(y_i\right)\mathbf{z}\right) \\ \mathbf{q}_i^{(k+1)} &\leftarrow \frac{\mathbf{q}_i^{(k+1)}}{||\mathbf{q}_i^{(k+1)}||_2}\end{aligned}. \tag{3.42}$$

The projection back onto the unit circle compensates for the approximations made which neglect scalar values in both the gradient and the Newton step.

While the simple gradient scheme works well for small examples. It is found that the algorithm of Hyvärinen [1999a] in equation (3.42) is advantageous in both its efficiency and robustness. Therefore, in the remainder of the thesis, it is used extensively.

## 3.9 Summary

This chapter introduced ICA and described algorithms for computing independent components. ICA considers mixtures of random variables such that its goal is, given only the mixtures and an assumption of independence, to recover the corresponding sources. It was shown that the CLT allows for exactly this such that the independent components are obtained exactly when their statistics are maximally non-Gaussian. Measures of Gaussianity, entropy and negentropy, were

derived and approximated using series expansions of the pdf of the independent component. Both polynomial (Chebyshev-Hermite polynomials) and nonpolynomial [Hyvärinen, 1998] basis functions were used yielding two different approximations of negentropy and, hence, two measures for the computation of independent components.

This chapter presented a simple, but instructive, example of ICA. For this example the polynomial expansion was sufficient for finding independent components. However, for geophysical data sets the sources are, more often than not, super-Gaussian. For such data sets, the nonpolynomial expansion is essential, the polynomial expansion suffering from outliers in the sampling of the associated random variable.

# CHAPTER 4

# Blind Deconvolution by ICA

## 4.1 Introduction

Consider two time sequences, $h(t)$ and $\rho(t)$, and their convolution,

$$\begin{aligned} \chi(t) &= h(t) * \rho(t) \\ &= \int_{-\infty}^{\infty} h(t-\tau)\,\rho(\tau)d\tau. \end{aligned} \tag{4.1}$$

Neglecting noise, equation (4.1) is often used to model seismic data where $\chi(t)$ is a seismic trace generated by convolving a wavelet, $h(t)$ (the filter), with the Earth's reflectivity, $\rho(t)$. This linear representation of the truth is useful but, for real data, introduces an equation with two unknowns (the wavelet and the reflectivity). A blind deconvolution algorithm must find the unknowns given only the trace; hence, the problem is ill-posed and requires additional constraints. Here, independent component analysis (ICA) is used to develop a blind deconvolution algorithm such that the reflectivity is constrained to follow the ICA model.

Chapter 3 introduced and explained ICA. In this chapter the convolution problem is presented and adapted to an ICA framework yielding a new blind deconvolution algorithm. The convolutional model yields a mixing matrix which is banded, and this information is incorporated into the ICA algorithm as prior information. This banded ICA algorithm (B-ICA) is then used to simultaneously recover the seismic wavelet and reflectivity for a noise free trace, $\chi(t)$.

Wiggins [1978] introduced a blind deconvolution algorithm called minimum entropy deconvolu-

tion where the statistics of the reflectivity were constrained using the varimax criterion (a measure of kurtosis). This algorithm, for a short time, was popular and has received much attention [e.g. Donoho, 1981; Ooe and Ulrych, 1979; Sacchi et al., 1994; Walden, 1985]. Similar methods were, independently, derived by Shalvi and Weinstein [1990]. More recently, Kaaresen and Taxt [1998] derived an algorithm which explicitly incorporates the sparseness of the reflectivity by using a spike train as a model where the location, amplitude and number of spikes are considered. With the exception of Kaaresen and Taxt [1998], all of these methods employ higher order statistics. As such, while the method presented in this chapter is derived from ICA, it has roots reaching a wider scope of literature.

## 4.2 Discrete Convolution and the ICA model

As is evident in equation (4.1), convolution is, of course, linear and can be expressed as a linear system of equations. This lends itself to an ICA formulation of blind deconvolution which given only one time sequence, $\chi(t)$, allows for the reconstruction of both $h(t)$ and $\rho(t)$.

For discretely sampled signals, the convolutional model is modified, such that

$$\chi(t_i) = \sum_j h(t_{i-j})\,\rho(t_j),$$

or equivalently

$$\mathbf{As} = \mathbf{x} \tag{4.2}$$

where

$$\mathbf{s}^T = \begin{bmatrix} \rho(t_1) & \rho(t_2) & \cdots & \rho(t_n) \end{bmatrix},$$
$$\mathbf{x}^T = \begin{bmatrix} \chi(t_1) & \chi(t_2) & \cdots & \chi(t_n) \end{bmatrix}$$

and $\mathbf{A}$ is an $n \times n$ banded matrix. The columns of $\mathbf{A}$ are constructed from delayed versions of the wavelet, $\mathbf{h}$, such that

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{a}_1 & | & \mathbf{a}_2 & | & \mathbf{a}_3 & | & \cdots & | & \mathbf{a}_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{N}_1\mathbf{h} & | & \mathbf{N}_2\mathbf{h} & | & \mathbf{N}_3\mathbf{h} & | & \cdots & | & \mathbf{N}_n\mathbf{h} \end{bmatrix}, \end{aligned} \tag{4.3}$$

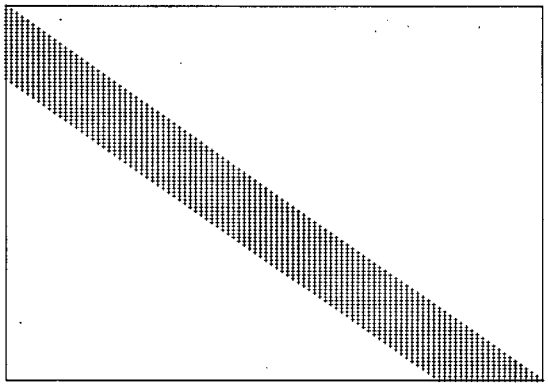

Figure 4.1: The nonzero bits of a banded mixing matrix, $\mathbf{A}$, for convolution ($n = 100$ and $nw = 20$).

$$\mathbf{h}^T = \begin{bmatrix} h(t_1) & h(t_2) & \cdots & h(t_{nw}) \end{bmatrix}$$

and $\mathbf{N}_i$ are zero padding matrices [e.g. Claerbout, 1992, p. 107] where the $i^{th}$ element of $\mathbf{a}_i$ is $h(t_1)$, element $(i+1)$ is $h(t_2)$ and so on. Figure 4.1 illustrates $\mathbf{A}$, showing the nonzero bits of the matrix when $n = 100$ and $nw = 20$.

Equation (4.2) is recognized as the ICA model from Chapter 3 where $\mathbf{s}$ are sources, $\mathbf{x}$ are mixtures and $\mathbf{A}$ is the mixing matrix. As before $\mathbf{s}$ and $\mathbf{x}$ are random vectors. However, the convolutional model in equation (4.2) provides only one realization of each. Obviously this is inadequate to characterize the corresponding statistics and, hence, is inadequate for ICA. However, the available information can be reorganized in a clever way providing several realizations. The trick is to consider time delayed versions of $\rho(t)$ and $\chi(t)$ [Hyvärinen et al., 2001, p. 360]. In particular, let

$$\mathbf{s}^T = \begin{bmatrix} z^{n-1}\rho(t) & z^{n-2}\rho(t) & \cdots & z\rho(t) & \rho(t) \end{bmatrix}$$

and

$$\mathbf{x}^T = \begin{bmatrix} z^{n-1}\chi(t) & z^{n-2}\chi(t) & \cdots & z\chi(t) & \chi(t) \end{bmatrix}$$

where $z$ is the unit delay operator. Hence, organizing the realizations of $\mathbf{s}$ into the columns of a

matrix gives

$$\mathbf{s} = \begin{bmatrix} z^{n-1}\rho(t) \\ z^{n-2}\rho(t) \\ \vdots \\ z\rho(t) \\ \rho(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \rho(t_1) \\ 0 & 0 & 0 & \cdots & \rho(t_1) & \rho(t_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \rho(t_1) & \rho(t_2) & \cdots & \rho(t_{n-2}) & \rho(t_{n-1}) \\ \rho(t_1) & \rho(t_2) & \rho(t_3) & \cdots & \rho(t_{n-1}) & \rho(t_n) \end{bmatrix} \tag{4.4}$$

and similarly

$$\mathbf{x} = \begin{bmatrix} z^{n-1}\chi(t) \\ z^{n-2}\chi(t) \\ \vdots \\ z\chi(t) \\ \chi(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \chi(t_1) \\ 0 & 0 & 0 & \cdots & \chi(t_1) & \chi(t_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \chi(t_1) & \chi(t_2) & \cdots & \chi(t_{n-2}) & \chi(t_{n-1}) \\ \chi(t_1) & \chi(t_2) & \chi(t_3) & \cdots & \chi(t_{n-1}) & \chi(t_n) \end{bmatrix} \tag{4.5}$$

where the $j^{th}$ realization of $\mathbf{x}$ is the convolution of the $j^{th}$ realization of $\mathbf{s}$ with the wavelet, $\mathbf{h}$. In other words,

$$\mathbf{x}(t_j) = \mathbf{h} * \mathbf{s}(t_j).$$

Thus, blind deconvolution is posed in a manner that can be solved using ICA. In particular, ICA computes some approximation to the rows of $\mathbf{s}$, each containing a portion of the reflectivity. However, as described in Chapter 3, ICA does not directly recover $\mathbf{h}$ but rather $\mathbf{q}_i$ which maps the whitened mixtures to the $i^{th}$ independent component. Additionally, recall that ICA relies on computing the statistics of the independent components. Clearly the first few rows of $\mathbf{s}$ and $\mathbf{x}$ provide few nonzero realizations; thus, doing little to define the statistics of their corresponding random variables.

## 4.3 Banded ICA

Recall the notation used, in Chapter 3, to describe ICA:

$$\mathbf{As} = \mathbf{x} \qquad (\mathbf{\Gamma As} = \mathbf{\Gamma x}) = (\mathbf{Ws} = \mathbf{z}) \qquad \mathbf{y} = \mathbf{Bx} \qquad \mathbf{y} = \mathbf{Qz} \tag{4.6}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}$ and $\mathbf{Q}$ are $m \times m$ matrices, $\mathbf{\Gamma}$ is chosen such that $\mathbf{z}$ is white and $\mathbf{Q}$ is chosen such that the elements of $\mathbf{y}$, $y_i$, are independent components. That is, $\mathbf{Q}$ is chosen such that $y_i \propto s_j$ for some $j$ where $s_j$ is the $j^{th}$ element of $\mathbf{s}$. Additionally, define a new matrix, $\mathbf{P}$, such that

$$\mathbf{Py} = \mathbf{x} \tag{4.7}$$

and

$$\mathbf{P} = \left[ \begin{array}{c|c|c|c} \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_m \end{array} \right].$$

By definition, $\mathbf{y}$ is a scaled and permuted version of $\mathbf{s}$; thus, $\mathbf{P}$ and $\mathbf{A}$ provide similar mappings. Here the ICA algorithm is modified such that instead of finding rows of $\mathbf{Q}$, it finds columns of $\mathbf{P}$. This, conveniently, allows for application of the given prior knowledge to ICA. Namely, the banded nature of $\mathbf{A}$ can be applied to $\mathbf{P}$, leading to the new B-ICA algorithm and a solution to the blind deconvolution problem.

Recall, from Chapter 3, that the ICA algorithm involves finding a minimum of some cost function, $\phi(\mathbf{q}_i)$, which measures the entropy of an independent component, $y_i = \mathbf{q}_i^T \mathbf{z}$, where $\mathbf{q}_i^T$ is the $i^{th}$ row of $\mathbf{Q}$. A relationship between $\mathbf{q}_i$ and $\mathbf{p}_i$ is readily found. Noting that the independent components, $y_i$, are zero mean and uncorrelated random variables with unit variance, and that $\mathbf{y} = \mathbf{Qz}$ (equation (4.6)), gives

$$\mathrm{E}\left(\mathbf{yy}^T\right) = \mathrm{E}\left(\mathbf{Qzz}^T\mathbf{Q}^T\right) = \mathbf{Q}\mathrm{E}\left(\mathbf{zz}^T\right)\mathbf{Q}^T = \mathbf{QQ}^T = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix. Therefore, assuming that $\mathbf{Q}^{-1}$ exists,

$$\begin{aligned} \mathbf{Q}^{-1}\mathbf{QQ}^T &= \mathbf{Q}^{-1} \\ \mathbf{Q}^T &= \mathbf{Q}^{-1}. \end{aligned}$$

Further, using equation (4.6),

$$\mathbf{z} = \mathbf{Q}^{-1}\mathbf{y} = \mathbf{Q}^T\mathbf{y} = \mathbf{\Gamma}\mathbf{x};$$

hence,

$$\mathbf{x} = \mathbf{\Gamma}^{-1}\mathbf{Q}^T\mathbf{y}. \tag{4.8}$$

Equations (4.7) and (4.8) allow for the explicit formulation of $\mathbf{P}$ in terms of $\mathbf{Q}$ such that

$$\mathbf{P} = \mathbf{\Gamma}^{-1}\mathbf{Q}^T \qquad\qquad \mathbf{Q}^T = \mathbf{\Gamma}\mathbf{P}$$

or more explicitly,

$$\mathbf{Q}^T = \left[\begin{array}{c|c|c|c} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_m \end{array}\right] = \mathbf{\Gamma}\left[\begin{array}{c|c|c|c} \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_m \end{array}\right] = \mathbf{\Gamma}\mathbf{P}.$$

Hence, $\mathbf{q}_i = \mathbf{\Gamma}\mathbf{p}_i$ and using equation (4.6) gives

$$y_i = \mathbf{q}_i^T\mathbf{z} = (\mathbf{\Gamma}\mathbf{p}_i)^T\mathbf{z} = (\mathbf{\Gamma}\mathbf{N}_i\mathbf{h})^T\mathbf{z} = \mathbf{h}^T\mathbf{N}_i^T\mathbf{\Gamma}^T\mathbf{z} = \mathbf{h}^T\tilde{\mathbf{x}} \tag{4.9}$$

where $\tilde{\mathbf{x}} = \mathbf{N}_i^T\mathbf{\Gamma}^T\mathbf{z}$ and $\mathbf{N}_i$ is a zero padding matrix which maps $\mathbf{h}$ to the $i^{th}$ column of $\mathbf{P}$, $\mathbf{p}_i$, which, in turn, corresponds to a particular column of the mixing matrix. In other words, $\mathbf{N}_i$ is the prior information. It enforces the banded property of the mixing matrix, $\mathbf{A}$, by explicitly choosing the number and location of zero entries in $\mathbf{p}_i$, and in doing so forces $\mathbf{p}_i$ to correspond to columns of $\mathbf{A}$ which have an equivalent sparse structure.

In equation (4.9), $\tilde{\mathbf{x}}$ can be thought of as a new set of mixtures ($nw$ in total) with corresponding independent components, $\tilde{\mathbf{y}}$, such that

$$\tilde{\mathbf{y}} = \tilde{\mathbf{B}}\tilde{\mathbf{x}}$$

where

$$\tilde{\mathbf{B}}^T = \left[\begin{array}{c|c|c|c} \tilde{\mathbf{h}}_1 & \tilde{\mathbf{h}}_2 & \cdots & \tilde{\mathbf{h}}_{nw} \end{array}\right]$$

is an $nw \times nw$ matrix; so, assuming an ICA model, equation (4.9) can be generalized such that

$$\tilde{\mathbf{A}}\tilde{\mathbf{s}} = \tilde{\mathbf{x}} \qquad \left(\tilde{\mathbf{\Gamma}}\tilde{\mathbf{A}}\tilde{\mathbf{s}} = \tilde{\mathbf{\Gamma}}\tilde{\mathbf{x}}\right) = \left(\tilde{\mathbf{W}}\tilde{\mathbf{s}} = \tilde{\mathbf{z}}\right) \qquad \tilde{\mathbf{y}} = \tilde{\mathbf{B}}\tilde{\mathbf{x}} \qquad \tilde{\mathbf{y}} = \tilde{\mathbf{Q}}\tilde{\mathbf{z}} \tag{4.10}$$

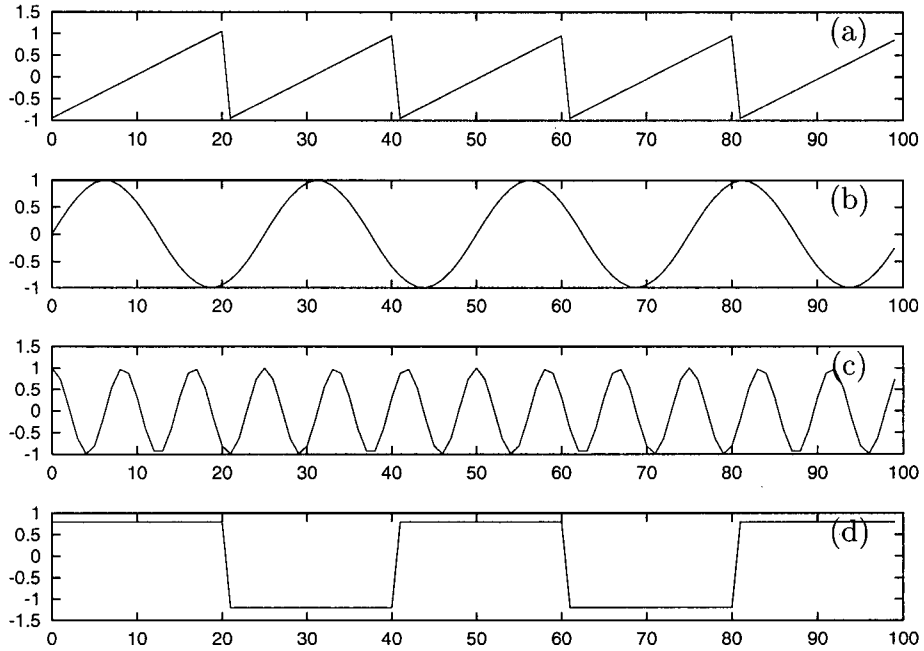


Figure 4.2: An example of B-ICA. The four sources (a)-(d) are mixed using equation (4.11) producing Figure 4.3.

where $\tilde{\mathbf{B}} = \tilde{\mathbf{Q}}\tilde{\mathbf{\Gamma}}$. Hence, given $\tilde{\mathbf{x}}$, the ICA algorithm is used to find $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{B}}$ where one element of $\tilde{\mathbf{y}}$ is the desired independent component, and one row of $\tilde{\mathbf{B}}$ is the nonzero bits of one column of $\mathbf{P}$. In other words, $\tilde{\mathbf{h}}_i \propto \mathbf{h}$ for some $i$.

The above algorithm can be further generalized so that $\mathbf{p}_i = \mathbf{N}_i\mathbf{h}_i$ where $\mathbf{h}_i$ are all of dimension $nw$, giving a more general form of equation (4.3) such that

$$\mathbf{A} = \left[ \begin{array}{c|c|c|c} \mathbf{N}_1\mathbf{h}_1 & \mathbf{N}_2\mathbf{h}_2 & \cdots & \mathbf{N}_m\mathbf{h}_m \end{array} \right].$$

For example, consider the mixing matrix,

$$\mathbf{A} = \begin{bmatrix} 1.0 & 0 & 0 & 0 \\ 1.1 & 1.2 & 0 & 0 \\ 0 & 1.3 & 1.4 & 0 \\ 0 & 0 & 1.5 & 1.6 \end{bmatrix} \tag{4.11}$$

which provides a mapping between the sources, $\mathbf{s}$, in Figure 4.2 and the mixtures, $\mathbf{x}$, in Figure 4.3.

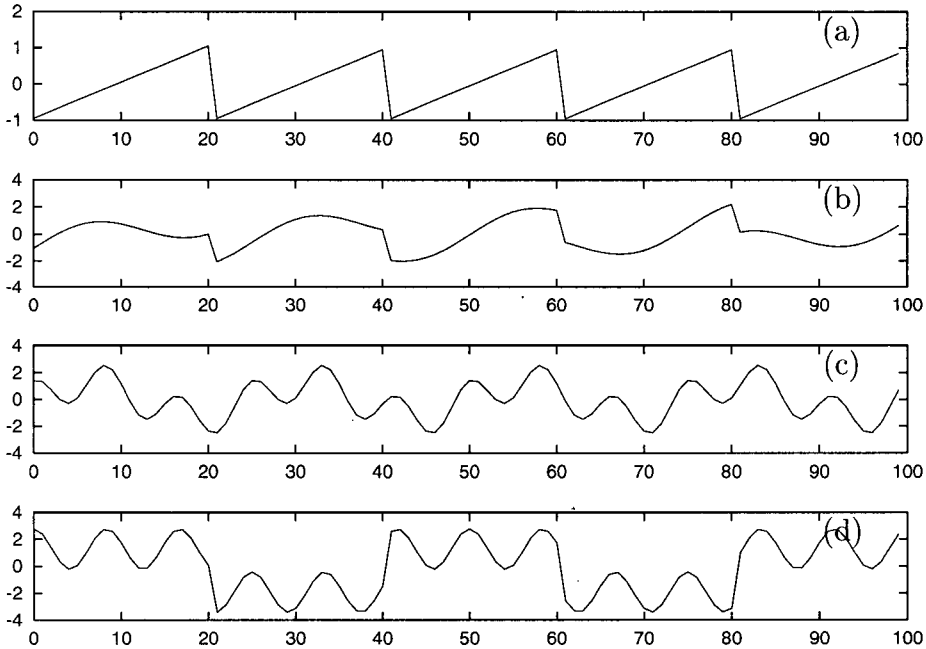


Figure 4.3: An example of B-ICA. The four mixtures (a)-(d) are produced by mixing the sources in Figure 4.2 according to the mixing matrix in equation (4.11).

New mixtures, $\tilde{\mathbf{x}}$, are computed according to equation (4.9) such that $\tilde{\mathbf{x}} = \mathbf{N}_2^T \mathbf{\Gamma}^T \mathbf{z}$. Figure 4.4a plots the cost function, $\phi(\tilde{\mathbf{q}}_i)$, for $||\tilde{\mathbf{q}}_1||_2 = 1$, where $\tilde{\mathbf{q}}_i^T$ is the $i^{th}$ row of $\tilde{\mathbf{Q}}$. The cost function is computed using the approximation to negentropy in equation (3.21). As expected, there are four local minima corresponding to two independent components in $\tilde{\mathbf{y}}$. These independent components, $\tilde{y}_1$ and $\tilde{y}_2$, are plotted in Figures 4.4b-c respectively. Clearly Figure 4.4b corresponds to the source in Figure 4.2b. For this example the prior information is $\mathbf{N}_2$ which constrains the ICA algorithm to find $\mathbf{p}_2$; consequently, it finds an independent component proportional to the second element of $\mathbf{s}$, $s_2$. This logic is echoed in Figure 4.4.

As a second example let $\tilde{\mathbf{x}} = \mathbf{N}_3^T \mathbf{\Gamma}^T \mathbf{z}$. The corresponding independent components are plotted in Figure 4.5. Through examination of equation (4.11), it is clear that this prior information, $\mathbf{N}_3$, allows for both the third and fourth columns of $\mathbf{A}$. Both have the same nonzero bits, so both obey the prior information imposed by $\mathbf{N}_3$. As such, both the third and fourth elements of $\mathbf{s}$, $s_3$ and $s_4$, are represented in Figures 4.5b-c respectively. The corresponding cost function is plotted in Figure 4.5a.

While B-ICA allows for application of the given prior knowledge, it still presents a difficulty in

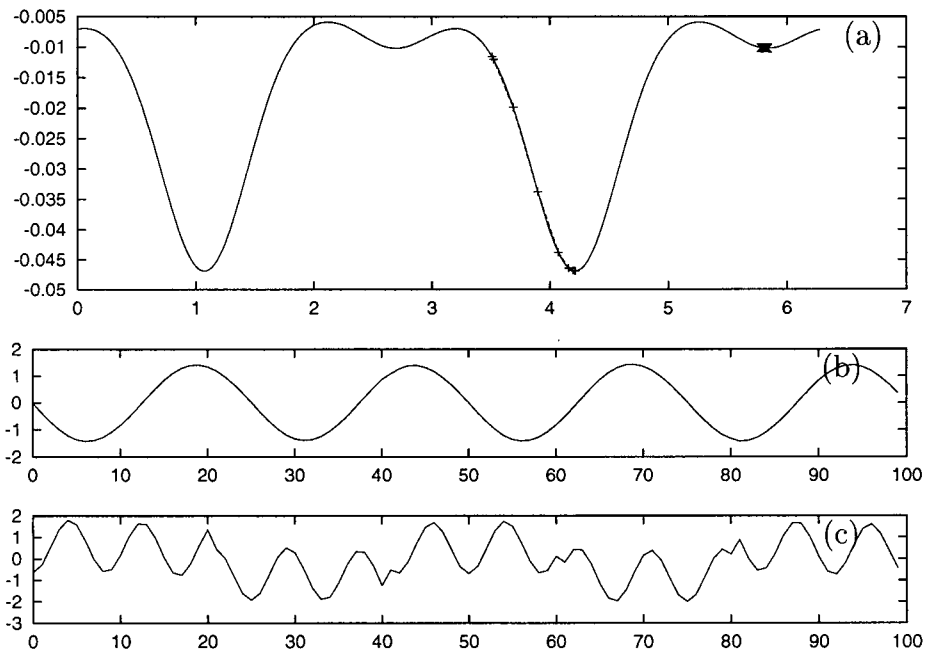


Figure 4.4: An example of B-ICA. (a) The cost function computed from the mixtures in Figure 4.3 using the prior information in $\mathbf{N}_2$ and plotted for $||\tilde{\mathbf{q}}||_2 = 1$. Superimposed on the cost function are the optimization paths which the algorithm followed to find the local minima. (b)-(c) The independent components, $\tilde{\mathbf{y}}$, corresponding to the local minima in (a). Notice that the independent component in (b) is representative of the second source (Figure 4.2b).

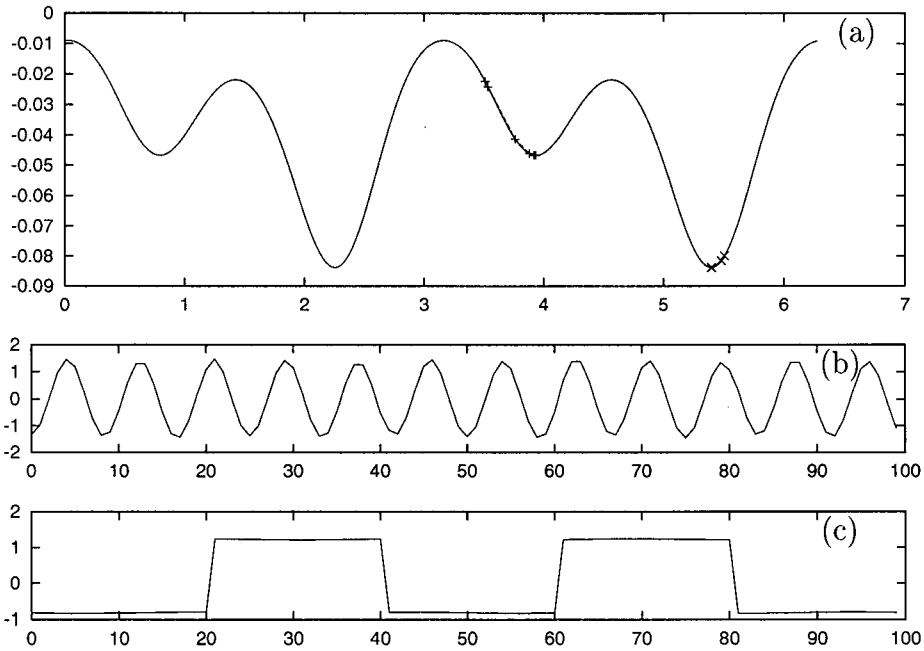


Figure 4.5: A second example of B-ICA. (a) The cost function computed from the mixtures in Figure 4.3 using the prior information in $\mathbf{N}_3$ and plotted for $||\tilde{\mathbf{q}}||_2 = 1$. Superimposed on the cost function are the optimization paths which the algorithm followed to find the local minima. (b)-(c) The independent components, $\tilde{\mathbf{y}}$, corresponding to the local minima in (a). Notice that the independent components are representative of the sources in Figures 4.2c-d.

the ambiguity of the result. Namely that the algorithm produces as many independent components as the dimension of $\mathbf{h}$ (or $\mathbf{h}_i$). As such, there is left the task of choosing one independent component and its corresponding row of $\tilde{\mathbf{B}}$. Fortunately, as will be shown, when B-ICA is used for blind deconvolution a solution presents itself.

## 4.4 B-ICA for Blind Deconvolution

In Section 4.2 the convolutional model was formulated as an ICA problem with a banded mixing matrix. Here, B-ICA, presented in Section 4.3, is used to solve for the filter, $\mathbf{h}$. Unfortunately, as illustrated in Section 4.3, B-ICA provides as many independent components as the dimension of $\mathbf{h}$. The best solution must be chosen from the pool of candidate solutions, yielding one approximation of both $\mathbf{h}$ and $\rho(t)$.

Further, the $\mathbf{s}$ and $\mathbf{x}$ proposed in equations (4.4) and (4.5) are inadequate in that the first few mixtures and sources provide few nonzero realizations; hence, doing little to constrain the statistics of their corresponding random variables. Therefore the algorithm must be modified to compensate for this lack of information. This modification produces an approximate convolutional model such that if $\mathbf{A}$ is an $m \times m$ matrix and $\chi(t_i)$ is $n$ points ($i = 1 \ldots n$), then modifying equation (4.4) such that $m < n$ gives

$$\mathbf{s} = \begin{bmatrix} z^{m-1}\rho(t) \\ z^{m-2}\rho(t) \\ \cdots \\ z\rho(t) \\ \rho(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \rho(t_1) & \rho(t_2) & \cdots & \rho(t_{n-m}) \\ 0 & 0 & \cdots & \rho(t_1) & \rho(t_2) & \rho(t_3) & \cdots & \rho(t_{n-m-1}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \rho(t_1) & \cdots & \cdots & \rho(t_{n-3}) & \rho(t_{n-2}) & \rho(t_{n-1}) & \\ \rho(t_1) & \rho(t_2) & \cdots & \cdots & \rho(t_{n-2}) & \rho(t_{n-1}) & \rho(t_n) & \end{bmatrix} . \tag{4.12}$$

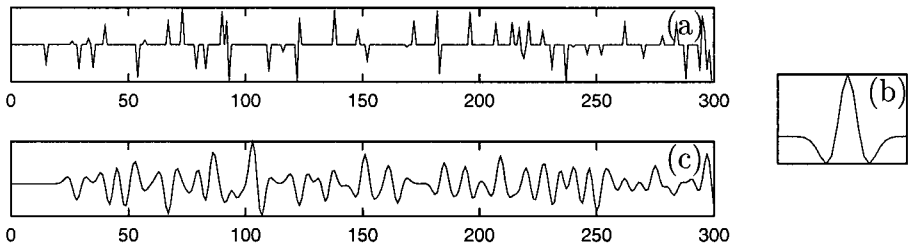


Figure 4.6: B-ICA used for blind deconvolution. (a) A sparse spike train convolved with (b) the twenty-five point filter, $\mathbf{h}$, produces (c) the signal, $\chi(t)$. Given the data in (c), B-ICA finds the information, $\tilde{\mathbf{B}}$, presented in Figure 4.7.

and, similarly from equation (4.5),

$$\mathbf{x} = \begin{bmatrix} z^{m-1}\chi(t) \\ z^{m-2}\chi(t) \\ \cdots \\ z\chi(t) \\ \chi(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \chi(t_1) & \chi(t_2) & \cdots & \chi(t_{n-m}) \\ 0 & 0 & \cdots & \chi(t_1) & \chi(t_2) & \chi(t_3) & \cdots & \chi(t_{n-m-1}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \chi(t_1) & \cdots & \cdots & \chi(t_{n-3}) & \chi(t_{n-2}) & \chi(t_{n-1}) & \\ \chi(t_1) & \chi(t_2) & \cdots & \cdots & \chi(t_{n-2}) & \chi(t_{n-1}) & \chi(t_n) & \end{bmatrix}. \tag{4.13}$$

Hence, each random variable has a number of nonzero realizations to constrain their statistics. Given equations (4.12) and (4.13), the mapping between $\mathbf{s}$ and $\mathbf{x}$, imposed by the convolutional mixing matrix (equation (4.3)), is not exact. Rather, it provides an approximate convolutional model such that $\mathbf{As} \approx \mathbf{x}$. In particular, through careful inspection of equations (4.12) and (4.13), it is clear that, given $\mathbf{A}$ and $\mathbf{s}$, $x_i(t_j)$ is incorrectly mapped for

$$(i \in \{1 \ldots nw\}) \cap (j \in \{(m-1) \ldots n\}).$$

However, for the remainder of $\mathbf{x}$ the mapping is correct which, as will be illustrated, given only knowledge of $\chi(t)$ (i.e. $\mathbf{x}$ in equation (4.13)), allows ICA to find a wavelet following the true convolutional model.

Consider the synthetic time sequence, $\chi(t)$, in Figure 4.6c which is the convolution of the sparse spike train, $\rho(t)$, in Figure 4.6a with the twenty-five point filter, $\mathbf{h}$, (a Ricker wavelet) in Figure 4.6b. Given only $\chi(t)$, B-ICA produces $\tilde{\mathbf{B}}$, the rows of which are plotted in Figure 4.7. The matrix, $\tilde{\mathbf{B}}$, is

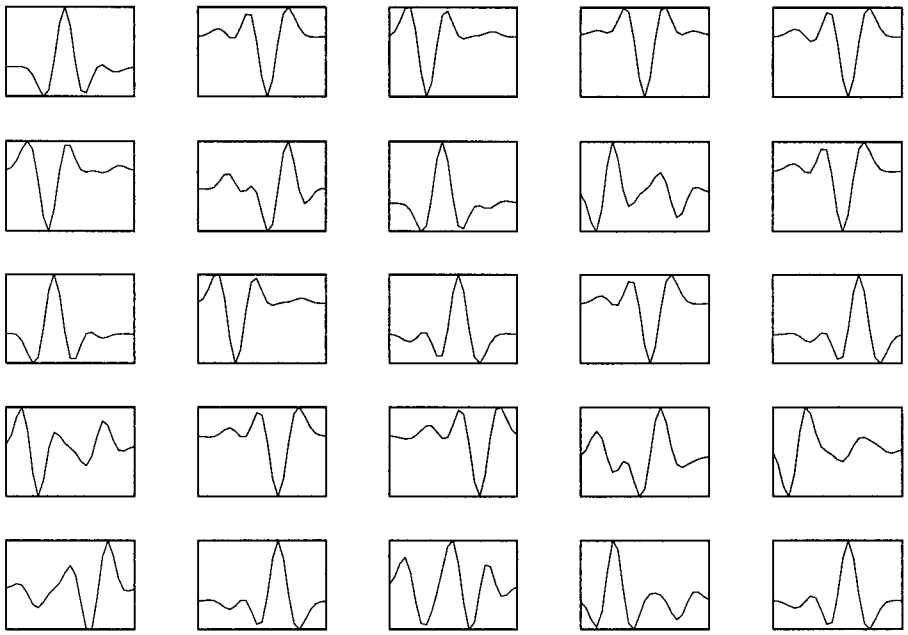

Figure 4.7: B-ICA used for blind deconvolution. Plotted are the twenty-five rows of $\tilde{\mathbf{B}}$ computed using B-ICA and the data in Figure 4.6c.

computed using the ICA model in equation (4.10) which is arrived at using the prior information in the zero padding matrix, $\mathbf{N}_{(m-nw-10)}$. The employed algorithm estimates negentropy using the nonpolynomial expansion of the corresponding pdf (as shown in equation (3.35)), and performs the optimization using the routine of Hyvärinen [1999a] (equation (3.42)). The forward model is approximated such that $m = 100$, and the nonlinearity used in the nonpolynomial expansion is $r(\tilde{y}_i) = \exp\left(-\frac{\tilde{y}_i^2}{2}\right)$. A quick search through the panels reveals good approximations to the wavelet.

While the $\tilde{h}_i$ in Figure 4.7 are an interesting result, their utility is not immediately obvious. In practice the filter is, of course, not known. Hence, simply presenting the choices in Figure 4.7 is nonsense. Instead, there is a sensible way to check for the best result. In particular, coefficients, $c_i$, can be calculated such that, given the prior $\mathbf{N}_k$,

$$\psi(c_i) = ||\mathbf{x}_k - c_i\tilde{\mathbf{h}}_i * \tilde{\mathbf{y}}_i||_2^2 \quad , \quad i = 1 \ldots nw \tag{4.14}$$

is, for each $i$, minimized where, here, $\mathbf{x}_k^T = \begin{bmatrix} x_k(t_1) & x_k(t_2) & \cdots & x_k(t_n) \end{bmatrix}$ are the realizations of the $k^{th}$ mixture, and $\tilde{\mathbf{y}}_i^T = \begin{bmatrix} \tilde{y}_i(t_1) & \tilde{y}_i(t_2) & \cdots & \tilde{y}_i(t_n) \end{bmatrix}$ are the realizations of the $i^{th}$ indepen-

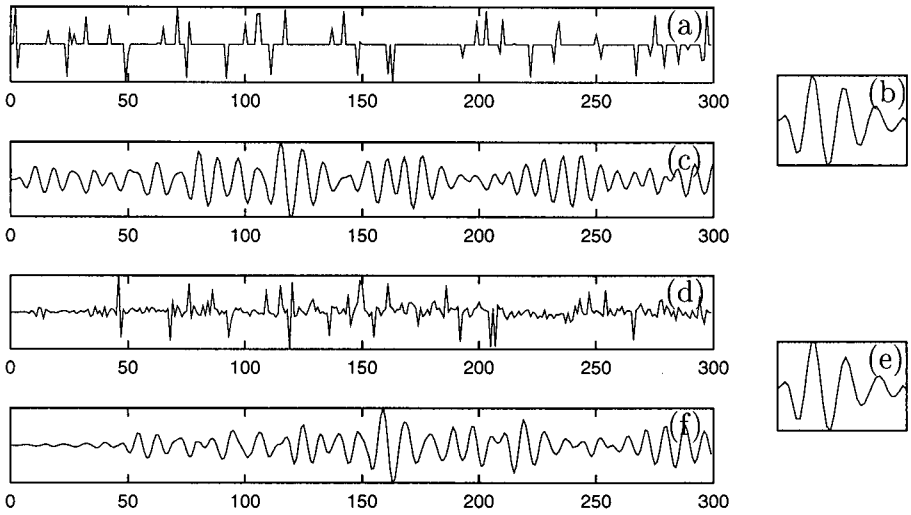


Figure 4.8: B-ICA used for blind deconvolution. (a) A sparse spike train convolved with (b) the thirty-five point filter, $\mathbf{h}$, produces (c) the signal, $\chi(t)$. (e) The recovered filter, $\tilde{\mathbf{h}}_*$, and (d) the independent component using B-ICA and the criterion in equation (4.14). (f) The convolution of (d) and (e).

dent component. It is easily verified that equation (4.14) has its extreme point when

$$c_i = c_i^{(*)} = \frac{\mathbf{x}_k^T \left(\tilde{\mathbf{h}}_i * \tilde{\mathbf{y}}_i\right)}{\left(\tilde{\mathbf{h}}_i * \tilde{\mathbf{y}}_i\right)^T \left(\tilde{\mathbf{h}}_i * \tilde{\mathbf{y}}_i\right)}.$$

The best solution, $(\tilde{\mathbf{y}}_*, \tilde{\mathbf{h}}_*)$, is chosen such that

$$\psi\left(c_*^{(*)}\right) = \min_i \left\{\psi\left(c_i^{(*)}\right)\right\} \quad , \quad i = 1 \ldots nw.$$

For example, consider the synthetic time sequence, $\chi(t)$, in Figure 4.8c generated by convolving the spike train, $\rho(t)$, in Figure 4.8a with the thirty-five point filter (a Berlage wavelet), $\mathbf{h}$, in Figure 4.8b. B-ICA is used to compute $\tilde{\mathbf{B}}$ such that the prior information is the zero padding matrix, $\mathbf{N}_{(m-nw-10)}$, $m = 75$ and, again, the algorithm of Hyvärinen [1999a] (equation (3.42)) is used with $r(\tilde{y}_i) = \exp\left(-\frac{\tilde{y}_i^2}{2}\right)$. As a result, thirty-five wavelets are recovered, $\tilde{\mathbf{h}}_1 \ldots \tilde{\mathbf{h}}_{35}$. The best result $(\tilde{\mathbf{h}}_*, \tilde{y}_*(t))$ is extracted from the pool of thirty-five candidate solutions according to the criterion in equation (4.14) (with $k = m - nw - 10$) and is plotted in Figures 4.8d-e. Figure 4.8d is, as expected, an approximation to the sparse spike series in Figure 4.8a with a linear phase shift.

The shift is due to how the mixtures, $\mathbf{x}$, are organized (see equation (4.5)). Finally, the convolution of the independent component in Figure 4.8d with the recovered wavelet in Figure 4.8e is plotted in Figure 4.8f. Clearly the algorithm has done a reasonable job in recovering both the wavelet and the reflectivity.

## 4.5 Summary

This chapter adapted the ICA algorithm to include prior information on the mixing matrix. In particular, its banded nature was accounted for by specifying its nonzero bits. Additionally, the deconvolution problem was coaxed into an ICA formulation with a banded mixing matrix computed from the filter, $\mathbf{h}$. Initially, this resulted in an ICA model with only one realization of both the source and mixture. To compensate for this lack of information, time delayed versions of $\rho(t)$ and $\chi(t)$ were considered. This resulted in a new blind deconvolution method, utilizing B-ICA, which in turn gave rise to a second complication. Namely that the first few mixtures and sources had few nonzero realizations, and so, had insufficient information to constrain their statistics. The solution was to use an approximate convolutional model which proved to be sufficient. B-ICA created a further complication. It produced as many candidate solutions as the dimension of $\mathbf{h}$. This problem was overcome by using the extra information provided by the convolutional model.

While useful, the model presented in this chapter neglected noise. In seismic data, as with all real data, noise plays an important factor, and the convolutional model in equation (4.1) must be modified such that

$$\chi(t) = \rho(t) * h(t) + n(t)$$

where $n(t)$ introduces additive random noise. Chapter 5 deals with this extra complication.

# CHAPTER 5

# Noisy ICA and Blind Deconvolution

## 5.1 Introduction

Chapter 4 described a blind deconvolution algorithm for a model devoid of noise. For real world applications data are inherently noisy, and the convolutional model in equation (4.1) requires modification such that

$$\chi(t) = h(t) * \rho(t) + n(t) \tag{5.1}$$

where $n(t)$ is additive random noise. In terms of ICA, the addition of noise augments equation (3.1) such that

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{n} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{5.2}$$

where $\mathbf{n}$ is a random noise vector and $\hat{\mathbf{x}}$ are noisy mixtures. In this chapter, B-ICA (see Chapter 4) is adapted such that the effect of $n(t)$ on the recovered wavelet, $h(t)$, is mitigated. As a means to this end, a modification, following the work of Hyvärinen [1999b], to the ICA algorithm is sought such that the estimated de-mixing matrix, $\mathbf{Q}$, is invariant to the noise term, $\mathbf{n}$.

The incorporation of noise into the ICA algorithm is a two-fold process. First, the noise must be accounted for in the whitening of the noisy mixtures. Second, the cost function measuring entropy requires modification such that it is invariant to noise. Further, two additional assumptions must be applied to the ICA model; the noise, $\mathbf{n}$, and the noise free mixtures, $\mathbf{x}$, are assumed to be independent, and the noise is assumed to follow a Gaussian distribution with a known (estimated)

covariance matrix. Subsequently, this modified algorithm can be applied to B-ICA, giving a noisy B-ICA algorithm, and a blind deconvolution algorithm for noisy signals.

## 5.2 Pre-Processing Noisy Mixtures

Section 3.7 described a pre-processing step for the ICA algorithm where, using principal component analysis (PCA), a whitening operator, $\mathbf{\Gamma}$, is extracted from the mixtures, $\mathbf{x}$, such that $\mathbf{z} = \mathbf{\Gamma x}$ and $\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right) = \mathbf{I}$. Here, the whitening algorithm is modified so that the whitening operator, $\mathbf{\Gamma}$, is invariant to the noise term in equation (5.2) [e.g. Douglas et al., 1998].

The whitening operator, $\mathbf{\Gamma}$, is sought such that

$$\hat{\mathbf{z}} = \mathbf{\Gamma}\hat{\mathbf{x}} = \mathbf{\Gamma}\left(\mathbf{x} + \mathbf{n}\right) = \mathbf{\Gamma x} + \mathbf{\Gamma n} = \mathbf{z} + \mathbf{\Gamma n}, \tag{5.3}$$

$\mathbf{z} = \mathbf{\Gamma x}$ and $\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right) = \mathbf{I}$. Recall, from section 3.7, that the whitening operator is computed using the covariance matrix of the noise free mixtures, $\mathbf{C}_x$. Fortunately, knowledge of the covariance matrix of the noise coupled with the, already mentioned, assumption of independence between the noise, $\mathbf{n}$, and noise free mixtures, $\mathbf{x}$, allows for computation of $\mathbf{C}_x$. In the case of zero-mean and noisy mixtures, $\hat{\mathbf{x}}$,

$$\begin{aligned}
\mathbf{C}_{\hat{x}} &= \mathrm{E}\left[\left(\mathbf{x} + \mathbf{n}\right)\left(\mathbf{x} + \mathbf{n}\right)^T\right] \\
&= \mathrm{E}\left(\mathbf{x}\mathbf{x}^T + \mathbf{x}\mathbf{n}^T + \mathbf{n}\mathbf{x}^T + \mathbf{n}\mathbf{n}^T\right) \\
&= \mathrm{E}\left(\mathbf{x}\mathbf{x}^T\right) + \mathrm{E}\left(\mathbf{n}\mathbf{n}^T\right) \\
&= \mathbf{C}_x + \mathbf{C}_n
\end{aligned}$$

where $\mathbf{C}_{\hat{x}}$ and $\mathbf{C}_n$ are, respectively, the covariance matrices of $\hat{\mathbf{x}}$ and $\mathbf{n}$. Hence, $\mathbf{C}_x = \mathbf{C}_{\hat{x}} - \mathbf{C}_n$, and given $\mathbf{C}_n$, $\mathbf{C}_x$ is obtained and used in the computation of $\mathbf{\Gamma}$; thus, allowing equation (5.3).

Recall, from Chapter 3, that the ICA algorithm consisted of an optimization problem; the corresponding cost function, $\phi\left(\mathbf{q}_i\right)$, providing a measure of entropy (independence) for $y_i = \mathbf{q}_i^T\mathbf{z}$. Additionally recall that $y_i$ is an independent component exactly when $\mathbf{q}_i$ is chosen such that $\phi\left(\mathbf{q}_i\right)$ is a local minimum and $||\mathbf{q}_i||_2 = 1$. Thus, if a cost function, $\phi\left(\mathbf{q}_i\right)$, measuring entropy can be

constructed such that

$$\phi(\mathbf{q}_i) = f\left(\mathbf{q}_i^T \hat{\mathbf{z}}\right) = f\left(\mathbf{q}_i^T \mathbf{\Gamma}(\mathbf{x}+\mathbf{n})\right) = f\left(y_i + \mathbf{q}_i^T \mathbf{\Gamma}\mathbf{n}\right) = f(y_i + n) = f(y_i), \tag{5.4}$$

then $\mathbf{q}_i$ can be found as in the noise free case. In other words, the desired cost function is invariant to noise.

For a first example, let $f(\cdot) = \kappa_4(\cdot)$ where $\kappa_4$ is kurtosis. Hence, given independence between the noise free mixtures and the noise,[1]

$$\phi(\mathbf{q}_i) = \kappa_4\left(\mathbf{q}_i^T \hat{\mathbf{z}}\right) = \kappa_4\left(\mathbf{q}_i^T \mathbf{z} + \mathbf{q}_i^T \mathbf{\Gamma}\mathbf{n}\right) = \kappa_4\left(\mathbf{q}_i^T \mathbf{z}\right) + \kappa_4\left(\mathbf{q}_i^T \mathbf{\Gamma}\mathbf{n}\right). \tag{5.5}$$

Further, if $\mathbf{n}$ follows a Gaussian distribution, then $\mathbf{q}_i^T \mathbf{\Gamma}\mathbf{n}$ also has Gaussian statistics and, therefore, zero kurtosis so that

$$\phi(\mathbf{q}_i) = \kappa_4\left(\mathbf{q}_i^T \mathbf{z}\right) = \kappa_4(y_i). \tag{5.6}$$

Hence, using kurtosis as a measure of entropy yields a cost function that is invariant to noise.

The incorporation of the noise covariance matrix in the whitening algorithm allows for equation (5.6). However, kurtosis is not a sufficiently robust measure of entropy for the application of blind deconvolution. Instead, the nonpolynomial expansion of the appropriate probability density function (pdf) (Section 3.6) is required. The task is to find an appropriate nonlinearity, for the nonpolynomial expansion, such that equation (5.4) is realized.

## 5.3 Gaussian Moments

In Chapter 3 a cost function, for ICA, was devised from an estimate of negentropy which employed a nonpolynomial expansion of the corresponding pdf (equation (3.40)). With additive noise, this cost function becomes

$$\phi(\mathbf{q}_i) = -\frac{1}{2}\left[\mathrm{E}\left(r(y_i + n)\right)\right]^2 + \lambda\left(\mathbf{q}_i^T \mathbf{q}_i - 1\right) \tag{5.7}$$

where $\lambda$ is a Lagrange multiplier and, as in equation (5.4), $n = \mathbf{q}_i^T \mathbf{\Gamma}\mathbf{n}$. Equation (5.7) can be adapted such that equation (5.4) is applicable and, hence, the cost function is invariant to noise.

[1] A proof of equation (5.5) is provided in Appendix A.

As a means to this end, the nonlinearity is chosen according to the work of Hyvärinen [1999b] who lets $r(\cdot)$ be a Gaussian pdf.

To understand the importance of the Gaussian pdf, it is necessary to contemplate its moments. In particular, given

$$\psi_c(y) = \frac{1}{\sqrt{2\pi}c} \exp\left(-\frac{y^2}{2c^2}\right)$$

and

$$\psi_d(y) = \frac{1}{\sqrt{2\pi}d} \exp\left(-\frac{y^2}{2d^2}\right), \tag{5.8}$$

it will be shown (and is shown by Hyvärinen [1999b]) that,

$$\mathrm{E}(\psi_c(y)) = \mathrm{E}(\psi_d(y+n)) \tag{5.9}$$

where $y$ and $n$ are random variables, $n \sim N(0, \sigma^2)$ and $d = \sqrt{c^2 - \sigma^2}$. A proof of equation (5.9) follows the derivation presented in Hyvärinen [1999b], and is included here. Let $p_{Y,N}(y,n)$ be the bivariate pdf for $n$ and $y$ with marginal pdfs, $p_N(n) = \psi_\sigma(n)$ and $p_Y(y)$. Further, let $y$ and $n$ be independent such that

$$\begin{aligned}
\mathrm{E}(\psi_d(y+n)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_d(y+n)\, p_{Y,N}(y,n)\, dydn \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_d(y+n)\, p_Y(y)\, \psi_\sigma(n)\, dydn.
\end{aligned}$$

Letting $y' = y+n$ so that $n = y'-y$ and $dn = dy'$, and noting that $\psi_\sigma(y)$ is an even function gives

$$\begin{aligned}
\mathrm{E}(\psi_d(y+n)) &= \int_{-\infty}^{\infty} p_Y(y) \left[\int_{-\infty}^{\infty} \psi_d(y')\, \psi_\sigma(y'-y)\, dy'\right] dy \\
&= \int_{-\infty}^{\infty} p_Y(y) \left[\int_{-\infty}^{\infty} \psi_d(y')\, \psi_\sigma(y-y')\, dy'\right] dy \\
&= \int_{-\infty}^{\infty} p_Y(y) \left[\psi_d(y) * \psi_\sigma(y)\right] dy.
\end{aligned} \tag{5.10}$$

It can be shown [e.g. Frieden, 1983, pp. 75-76] that

$$\psi_d(y) * \psi_\sigma(y) = \psi_c(y) \tag{5.11}$$

where $c = \sqrt{d^2 + \sigma^2}$.[2] Hence, equation (5.9) follows from equation (5.10), and the proof is complete.

Equation (5.9) enables the cost function in equation (5.7) to be adapted to account for the noise. In particular, choosing

$$r(y_i + n) = \psi_d(y_i + n) = \psi_c(y_i)$$

allows for a cost function which is invariant to noise. In other words,

$$\begin{aligned} \phi(\mathbf{q}_i) &= -\frac{1}{2}\left[\mathrm{E}\left(\psi_d(y_i + n)\right)\right]^2 + \lambda\left(\mathbf{q}_i^T\mathbf{q}_i - 1\right) \\ &= -\frac{1}{2}\left[\mathrm{E}\left(\psi_c(y_i)\right)\right]^2 + \lambda\left(\mathbf{q}_i^T\mathbf{q}_i - 1\right) \end{aligned} \tag{5.12}$$

where

$$d = \sqrt{c^2 - \mathrm{E}\left(nn^T\right)} = \sqrt{c^2 - \mathrm{E}\left[\left(\mathbf{q}_i^T\boldsymbol{\Gamma}\mathbf{n}\right)\left(\mathbf{q}_i^T\boldsymbol{\Gamma}\mathbf{n}\right)^T\right]} = \sqrt{c^2 - \mathbf{q}_i^T\boldsymbol{\Gamma}\mathbf{C}_n\boldsymbol{\Gamma}^T\mathbf{q}_i} = \sqrt{c^2 - \mathbf{q}_i^T\widetilde{\mathbf{C}}_n\mathbf{q}_i} \tag{5.13}$$

and $\widetilde{\mathbf{C}}_n = \boldsymbol{\Gamma}\mathbf{C}_n\boldsymbol{\Gamma}^T$.

## 5.4 A Noisy ICA Optimization Algorithm

In the preceding section, a cost function well suited for noisy ICA was found. The remaining task is to find an algorithm, analogous to equation (3.42), for finding the minima of equation (5.12). For readability, equation (3.42) is restated here:

$$\begin{aligned} \mathbf{q}_i^{(k+1)} &= \mathrm{E}\left(r''(y_i)\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(r'(y_i)\,\mathbf{z}\right) \\ \mathbf{q}_i^{(k+1)} &\leftarrow \frac{\mathbf{q}_i^{(k+1)}}{||\mathbf{q}_i^{(k+1)}||_2} \end{aligned} \tag{5.14}$$

As in Chapter 3, the optimization scheme, for noisy mixtures, is devised using a Newton type algorithm; thus, requiring expressions for both the gradient and Hessian of $\phi(\mathbf{q}_i)$. The end result of our efforts is the algorithm in equation (5.22).

[2] For a proof of equation (5.11), please refer to Appendix A.

First, the gradient is computed as

$$\nabla\phi\left(\mathbf{q}_i\right) = -\mathrm{E}\left(\psi_d\left(y_i+n\right)\right)\mathrm{E}\left(\nabla\psi_d\left(y_i+n\right)\right) + 2\lambda\mathbf{q}_i^{(k)} \tag{5.15}$$

where, after substituting in equation (5.8), and some calculus,

$$\nabla\psi_d\left(y_i+n\right) = \frac{1}{\sqrt{2\pi}d^3}\exp\left[-\frac{\left(y_i+n\right)^2}{2d^2}\right]\widetilde{\mathbf{C}}_n\mathbf{q}_i - \frac{\left(y+n\right)^2}{\sqrt{2\pi}d^5}\exp\left[-\frac{\left(y_i+n\right)^2}{2d^2}\right]\widetilde{\mathbf{C}}\mathbf{q}_i - \frac{y_i+n}{\sqrt{2\pi}d^3}\exp\left[-\frac{\left(y_i+n\right)^2}{2d^2}\right]\hat{\mathbf{z}}.$$

Further, making use of equation (5.8) and its derivative,

$$\psi_d'\left(y\right) = -\frac{y}{d^2}\psi_d\left(y\right), \tag{5.16}$$

allows for

$$\nabla\psi_d\left(y_i+n\right) = \widetilde{\mathbf{C}}_n\mathbf{q}_i\frac{1}{d^2}\left[\psi_d\left(y_i+n\right) + \left(y_i+n\right)\psi_d'\left(y+n\right)\right] + \hat{\mathbf{z}}\psi_d'\left(y+n\right). \tag{5.17}$$

Lastly, equation (5.17) is simplified by noting that

$$\psi_d''\left(y\right) = -\frac{1}{d^2}\psi_d\left(y\right) - \frac{y}{d^2}\psi_d'\left(y\right).$$

Therefore,

$$\nabla\psi_d\left(y_i+n\right) = -\widetilde{\mathbf{C}}_n\mathbf{q}_i\psi_d''\left(y_i+n\right) + \hat{\mathbf{z}}\psi_d'\left(y_i+n\right). \tag{5.18}$$

Second, consider the Hessian matrix, $\mathbf{H}$; however, rather than computing it explicitly, simply let

$$\mathbf{H} = \left(\alpha+2\lambda\right)\mathbf{I} \tag{5.19}$$

where $\alpha$ is some scalar value and $\mathbf{I}$ is the identity matrix. The choice made in equation (5.19) will be validated shortly, and a value for $\alpha$ will fall out of the mathematics.

Using the expressions for the Hessian and gradient in equations (5.15) and (5.19) respectively, and ignoring the scalar term, $-\mathrm{E}\left(\psi_d\left(y_i+n\right)\right)$, in the expression for the gradient, an approximative

Newton step (from iteration $k$ to $k+1$) is given by

$$\begin{aligned}\mathbf{q}_i^{(k+1)} &= \mathbf{q}_i^{(k)} - \mathbf{H}^{-1}\nabla\phi\left(\mathbf{q}_i^{(k)}\right) \\ &= \mathbf{q}_i^{(k)} - \frac{\mathrm{E}\left(\nabla\psi_d\left(y_i+n\right)\right) + 2\lambda\mathbf{q}_i^{(k)}}{\alpha+2\lambda}.\end{aligned}$$

Multiplying through by $\alpha+2\lambda$ gives

$$\begin{aligned}\left(\alpha+2\lambda\right)\mathbf{q}_i^{(k+1)} &= \left(\alpha+2\lambda\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(\nabla\psi_d\left(y_i+n\right)\right) + 2\lambda\mathbf{q}_i^{(k)} \\ &= \alpha\mathbf{q}_i^{(k)} - \mathrm{E}\left(\nabla\psi_d\left(y_i+n\right)\right). \qquad (5.20)\end{aligned}$$

Finally, recall the algorithm in equation (5.14) where the first part of the update rule, for the noise free case, can be written as

$$\mathbf{q}_i^{(k+1)} = \mathrm{E}\left(\psi_c''\left(y_i\right)\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(\nabla\psi_c\left(y_i\right)\right).$$

Hence, remembering equation (5.9), an appropriate choice for $\alpha$ is given by

$$\alpha = \mathrm{E}\left(\psi_d''\left(y_i+n\right)\right) = \mathrm{E}\left(\psi_c''\left(y_i\right)\right). \qquad (5.21)$$

Combining equations (5.20), (5.21) and (5.18) yields

$$\begin{aligned}\left[\mathrm{E}\left(\psi_d''\left(y_i+n\right)\right)\right]\mathbf{q}_i^{(k+1)} &= \mathrm{E}\left(\psi_d''\left(y_i+n\right)\right)\mathbf{q}_i^{(k)} + \widetilde{\mathbf{C}}_n\mathbf{q}_i^{(k)}\mathrm{E}\left(\psi_d''\left(y_i+n\right)\right) - \mathrm{E}\left(\hat{\mathbf{z}}\psi_d'\left(y_i+n\right)\right) \\ &= \left(\mathbf{I}+\widetilde{\mathbf{C}}_n\right)\mathrm{E}\left(\psi_d''\left(y_i+n\right)\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(\hat{\mathbf{z}}\psi_d'\left(y_i+n\right)\right).\end{aligned}$$

Hence, an appropriate algorithm for noisy ICA is [Hyvärinen, 1999b]

$$\begin{aligned}\mathbf{q}_i^{(k+1)} &= \left(\mathbf{I}+\widetilde{\mathbf{C}}_n\right)\mathrm{E}\left(\psi_d''\left(y_i+n\right)\right)\mathbf{q}_i^{(k)} - \mathrm{E}\left(\hat{\mathbf{z}}\psi_d'\left(y_i+n\right)\right) \\ \mathbf{q}_i^{(k+1)} &\leftarrow \frac{\mathbf{q}_i^{(k+1)}}{\|\mathbf{q}_i^{(k+1)}\|_2}\end{aligned} \qquad (5.22)$$

To review, consideration of the noise in the whitening algorithm allows for a cost function which is invariant to Gaussian noise. The result is an algorithm which, given noisy mixtures, computes an optimal model, $\mathbf{q}_i^{(*)}$, which is also invariant to noise. This is an appealing result which, in the

following section, is used to adapt the B-ICA and blind deconvolution algorithms of Chapter 4 so that they are applicable to noisy data.

## 5.5 Gaussian Moments, B-ICA and Blind Deconvolution

The incorporation of noise into B-ICA requires its marriage with the algorithm presented in the previous section which, in turn, requires appropriate modifications to the noise covariance matrix. The end result is a much improved approximation of the wavelet.

Recall equation (4.9) which provides a relation between the $i^{th}$ independent component, the whitened mixtures and the filter, $\mathbf{h}^T = \begin{bmatrix} h(t_1) & h(t_2) & \cdots & h(t_{nw}) \end{bmatrix}$. For noisy mixtures, this equation is modified such that

$$y_i + \tilde{n} = \mathbf{q}_i^T (\mathbf{z} + \mathbf{\Gamma n}) = (\mathbf{\Gamma N}_i \mathbf{h})^T (\mathbf{z} + \mathbf{\Gamma n}) = \mathbf{h}^T (\tilde{\mathbf{x}} + \tilde{\mathbf{n}}) \tag{5.23}$$

where $\tilde{\mathbf{x}} = \mathbf{N}_i^T \mathbf{\Gamma}^T \mathbf{z}$, $\tilde{\mathbf{n}} = \mathbf{N}_i^T \mathbf{\Gamma}^T \mathbf{\Gamma n}$, $\tilde{n} = \mathbf{h}^T \tilde{\mathbf{n}}$ and, as before, $\mathbf{\Gamma}$ is chosen such that $\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right) = \mathbf{I}$. As in Chapter 4, $\tilde{\mathbf{x}}$ can be thought of as mixtures for a new noisy ICA problem with an $nw \times nw$ mixing matrix so that

$$\tilde{\mathbf{A}}\tilde{\mathbf{s}} = \tilde{\mathbf{x}} + \tilde{\mathbf{n}} \qquad \tilde{\mathbf{\Gamma}}\tilde{\mathbf{A}}\tilde{\mathbf{s}} = \tilde{\mathbf{\Gamma}}(\tilde{\mathbf{x}} + \tilde{\mathbf{n}}) \qquad \tilde{\mathbf{y}} + \tilde{\mathbf{B}}\tilde{\mathbf{n}} = \tilde{\mathbf{B}}(\tilde{\mathbf{x}} + \tilde{\mathbf{n}}) \qquad \tilde{\mathbf{y}} + \tilde{\mathbf{Q}}\tilde{\mathbf{\Gamma}}\tilde{\mathbf{n}} = \tilde{\mathbf{Q}}\left(\tilde{\mathbf{z}} + \tilde{\mathbf{\Gamma}}\tilde{\mathbf{n}}\right) \tag{5.24}$$

and $\tilde{\mathbf{\Gamma}}$ is chosen such that $\mathrm{E}\left(\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T\right) = \mathbf{I}$. That is, $\mathbf{\Gamma}$ is computed, as usual, using the covariance matrix of the noise free mixtures,

$$\mathbf{C}_x = \mathbf{C}_{\hat{x}} - \mathbf{C}_n,$$

and, likewise, $\tilde{\mathbf{\Gamma}}$ is computed using the covariance matrix of $\tilde{\mathbf{x}}$,

$$\mathbf{C}_{\tilde{x}} = \mathbf{C}_{\tilde{x}+\tilde{n}} - \mathbf{C}_{\tilde{n}}$$

where

$$\mathbf{C}_{\tilde{n}} = \mathrm{E}\left(\tilde{\mathbf{n}}\tilde{\mathbf{n}}^T\right) = \mathrm{E}\left(\mathbf{N}_i^T \mathbf{\Gamma}^T \mathbf{\Gamma n} \mathbf{n}^T \mathbf{\Gamma}^T \mathbf{\Gamma N}_i\right) = \mathbf{N}_i^T \mathbf{\Gamma}^T \mathbf{\Gamma C}_n \mathbf{\Gamma}^T \mathbf{\Gamma N}_i.$$

If equation (5.24) is true; that is, if $\mathbf{\Gamma}$ and $\tilde{\mathbf{\Gamma}}$ are computed with proper consideration for the noise, then the algorithm in equation (5.22) is applicable, and $\tilde{\mathbf{Q}}$ is found such that it is invariant

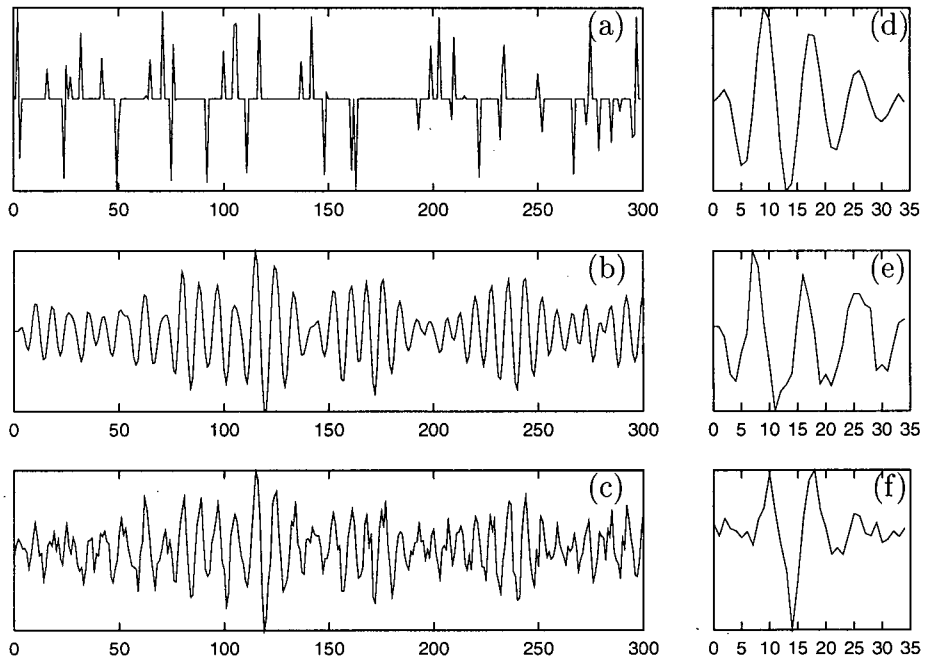


Figure 5.1: (b) The convolution of (a) with the filter (d). (c) The signal in (b) with additive, random and Gaussian noise. (e) The filter recovered from (c) using noisy B-ICA. (f) The filter recovered from (c) using B-ICA (without consideration for noise).

to noise. However, it is evident that the transformation of the noise occurring in equation (5.24) must be incorporated into the computation of $d$; hence, $d$ is re-evaluated as

$$d = \sqrt{c^2 - \mathrm{E}\left(\tilde{n}\tilde{n}^T\right)} = \sqrt{c^2 - \tilde{\mathbf{q}}_i^T \tilde{\mathbf{\Gamma}} \mathrm{E}\left(\tilde{\mathbf{n}}\tilde{\mathbf{n}}^T\right) \tilde{\mathbf{\Gamma}}^T \tilde{\mathbf{q}}_i};$$

and, in equation (5.22), $\widetilde{\mathbf{C}}_n$ becomes

$$\widetilde{\mathbf{C}}_n = \tilde{\mathbf{\Gamma}} \mathrm{E}\left(\tilde{\mathbf{n}}\tilde{\mathbf{n}}^T\right) \tilde{\mathbf{\Gamma}}^T = \tilde{\mathbf{\Gamma}} \mathbf{C}_{\tilde{n}} \tilde{\mathbf{\Gamma}}^T.$$

The end result of these modification to B-ICA is a new algorithm, noisy B-ICA, which, in turn, can be used in a blind deconvolution algorithm (see Chapter 4) where the convolved signal is corrupted with additive noise.

For illustration, consider the synthetic example in Figure 5.1. Figure 5.1b is the convolution of the signal in Figure 5.1a with the filter (a Berlage wavelet) in Figure 5.1d. Figure 5.1c is the same as Figure 5.1b, but with additive, random, Gaussian and zero-mean noise such that the signal

to noise ratio is 7 (about 14 percent noise)[3]. In applying noisy B-ICA to the noisy signal, $\mathbf{C}_n$ is chosen as a constant diagonal matrix; hence, to apply the algorithm, the user need only select one parameter. Figure 5.1e shows the wavelet, $h(t)$, recovered from the noisy signal using noisy B-ICA. For comparison, Figure 5.1e shows the wavelet recovered from the noisy signal using B-ICA (without consideration for the noise). By inspection, it is obvious that the incorporation of noise into the algorithm yields improved results.

## 5.6 Summary

This chapter adapted the ICA and B-ICA algorithms for noisy data. In particular, the noise covariance matrix was used to provide a whitening operator which is invariant to noise; and subsequently, its use produced cost functions, measuring entropy, which are invariant to noise. First, an example using kurtosis was provided. Second, the properties of Gaussian moments were exploited so that a more robust measure, using nonpolynomial expansions of the appropriate pdf, of entropy could be used. This resulted in a new noisy B-ICA algorithm which was used for blind deconvolution.

While the results of the blind deconvolution routine, presented in Figure 5.1, are interesting, for higher noise levels the quality of the results diminish. Further, it was found that the algorithm is very sensitive to the choice of $\mathbf{C}_n = \sigma_n^2\mathbf{I}$. A choice which must be made by the user of the algorithm. Despite these shortcomings, the method is sound in its mathematics and shows promising results.

[3]The signal to noise ratio is computed as the absolute maximum amplitude of the signal divided by the standard deviation of the noise.

# CHAPTER 6

# Conclusions

## 6.1 Summary

Principal and independent component analysis (PCA and ICA) use the statistical properties inherent in data, extracting useful information which, in this thesis, is used for noise suppression and blind deconvolution.

PCA and noise suppression was explained in Chapter 2. PCA was explained from three perspectives; variance, the singular value decomposition and ordinary differential equations. The subsequent derivations gave rise to an orthogonal basis consisting of, so called, eigensections which proved useful in their ability to separate coherent and incoherent information. Examples, illustrating the theory, were given for both synthetic and real seismic data, attenuating the random noise while conserving the coherent signal and, thus, increasing signal to noise ratios.

Whereas PCA uses correlation, ICA uses independence. The relation between independence and correlation was explained in Chapter 3 where uncorrelated was shown to be a special case of independent. In fact, it was shown that independent random variables are nonlinearly decorrelated. Chapter 3 developed ICA algorithms from the perspective of information theory. In particular, the concepts of independence and entropy were related through the central limit theorem, and entropy was used as a tool for building ICA algorithms. The estimation of entropy is not a trivial matter, and two approaches were considered. First, entropy was approximated using higher order moments and, second, using nonpolynomial functions. The algorithm employing nonpolynomials is preferred

for the applications in this thesis due to its robustness.

In Chapters 4 and 5, the ICA algorithm, employing the estimate of entropy using the non-polynomial expansion, was used for blind deconvolution. Chapter 4 considered the noise free case, and Chapter 5 used a convolutional model corrupted with additive, random and Gaussian noise. To facilitate a blind deconvolution algorithm using ICA, the properties of the convolutional model were associated with a banded ICA mixing matrix. Taking the banded nature of the mixing matrix into account, the ICA algorithm was modified, producing banded ICA (B-ICA) and a new blind deconvolution algorithm.

## 6.2 Future Work

Chapters 4 and 5 illustrated blind deconvolution using ICA. The result is some estimate of the wavelet and its corresponding independent component. For the noise free case, this independent component is, in turn, an estimate of the reflectivity with some linear phase shift. Moreover, when the trace is corrupted with noise, the recovered independent component need not be representative of the reflectivity. Hence, it would seem useful to have a deconvolution algorithm which, given the estimated wavelet, can recover the full reflectivity. When the wavelet is exactly known and the reflectivity is sufficiently sparse, this problem has a known solution [e.g. Walker and Ulrych, 1983; Oldenburg et al., 1983]. However, when the wavelet is estimated, and thus subject to error, the solution is more elusive. Deconvolution is an inverse problem where the forward operator is constructed from the wavelet. Thus, when the wavelet is estimated, the forward operator, inevitably, contains errors. This suggests that methods are needed which allow for errors in both the forward operator (e.g. seismic wavelet) and the data (e.g. seismic trace). One such method is total least squares (TLS) [e.g. Golub, 1973]. Future work could include a deconvolution algorithm which incorporates some version of TLS; thus, enabling a deconvolution algorithm applicable when the wavelet is estimated using the blind deconvolution algorithm presented in Chapters 4 and 5.

# Bibliography


Bell, Anthony J. and Terrence J. Sejnowski. "An Information-Maximisation Approach to Blind Separation and Blind Deconvolution." *Neural Computation* 7 (1995): 1129–1159.

Bracewell, Ronald M. *The Fourier Transform and Its Applications.* McGraw-Hill, Inc., 1978.

Bretscher, Otto. *Linear Algebra with Applications.* Prentice-Hall, Inc., 1997.

Claerbout, Jon. *Earth Soundings Analysis: Processing versus Inversion.* Blackwell Scientific Publications, Inc., 1992.

Common, P. "Independent component analysis, A new concept?." *Signal Processing* 36 (1994): 287–314.

Cooley, William W. and Paul R. Lohnes. *Multivariate Data Analysis.* John Wiley & Sons, Inc., 1971.

Cover, Thomas M. and Joy A. Thomos. *Elements of Information Theory.* Ed. Donald L. Schilling. Wiley Series in Telecommunications. John Wiley & Sons, Inc., 1991.

Donoho, David. "On Minimum Entropy Deconvolution." *Applied Time Series Analysis II.* Ed. D. Findley. Academic Press Inc., 1981.

Douglas, Scott C., Andrzej Cichocki, and Shun ichi Amari. "A Bias Removal Technique for Blind Source Separation with Noisy Measurements." *Electronic Letters* 34 (July 1998): 1379–1380.

Freire, S. L. and T. J. Ulrych. "Applications of Singular Value Decomposition to Vertical Seismic Profiling." *Geophysics* 53 (1988): 778–785.

Frieden, B. Roy. *Probability, Statistical Optics, and Data Testing: A Problem Solving Approach.* Ed. King sun Fu, Thomas S. Huang, and Manfred R. Schroeder. Springer Series in Information Sciences. Springer-Verlag, 1983.

Golub, Gene H. "Some Modified Matrix Eigenvalue Problems." *SIAM Review* 15 (April 1973): 318–334.

Golub, Gene H. and Charles F. Van Loan. *Matrix Computations.* 3rd edition. The Johns Hopkins University Press, 1996.

Haykin, S., editor. *Blind Deconvolution.* Prentice-Hall, 1994.

Haykin, Simon. *Neural Networks: A Comprehensive Foundation.* 2nd edition. Prentice-Hall, Inc., 1999.

Haykin, Simon, editor. *Unsupervised Adaptive Filtering: Blind Deconvolution.* Volume II . John Wiley & Sons, Inc., 2000.

Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory.* New York: Wiley, 1949.

Hyvärinen, Aapo. "New Approximation of Differential Entropy for Independent Component Analysis and Projection Pursuit." *Advances in Neural Information Procesing Systems 10.* Ed. Michael I. Jordan, Michael J. Kearns, and Sara A. Solla MIT Press, 1998, 273–279.

Hyvärinen, Aapo. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis." *Neural Networks, IEEE Transactions on* 10 (May 1999): 626–634.

Hyvärinen, Aapo. "Fast ICA for Noisy Data using Gaussian Moments." *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems.* IEEE, 1999, 57–61.

Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja. *Independent Component Analysis.* Ed. Simon Haykin. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., 2001.

Jaynes, E. T. *Probability Theory: The Logic of Science.* http://bayes.wustl.edu/etj/prob.html. 1995.

Jones, M. C. and Robin Sibson. "What is Projection Pursuit?." *Journal of the Royal Society. Series A (General)* 150 (1987): 1–37.

Kaaresen, Kjetil F. and Tofinn Taxt. "Multichannel blind deconvolution of seismic signals." *Geophysics* 63 (November-December 1998): 2093–2107.

Kaplan, Sam T. and T. J. Ulrych. "From Eigenfaces to Eigensections." *Journal of Seismic Exploration* 10 (2001-2002): 353–366.

Kendall, Maurice and Alan Stuart. *The Advanced Theory of Statistics.* Volume 1 . MacMillan Publishing Co., Inc., 1977.

Kirby, M. and L. Sirovich. "Low-Dimensional Procedure for the Characterization of Human Faces." *J. Opt. Soc. Am. A* 4 (March 1987): 519–524.

Nikias, Chrysostomos L. and Jerry M. Mendel. "Signal Processing with Higher-Order Spectra." *IEEE Signal Processing Magazine* (July 1993): 10–37.

Nocedal, Jorge and Stephen J. Wright. *Numerical Optimization.* Ed. Peter Glynn and Stephen M. Robinson. Springer Series in Operations Research. Springer-Verlag, 1999.

Oja, Erkki. "A Simplified Neuron Model as a Principal Component Analyzer." *Journal of Mathematical Biology* 15 (1982): 267–273.

Oldenburg, D. W., T. Scheuer, and S. Levy. "Recovery of the Acoustic Impedance from Reflection Seismograms." *Geophysics* 48 (October 1983): 1318–1337.

Ooe, M. and T. J. Ulrych. "Minimum Entropy Deconvolution with an Exponential Transformation." *Geophysical Prospecting* 27 (1979): 458–473.

Pentland, A. and M. Turk. "Eigenfaces for Recognition." *Journal of Cognitive Neuroscience* 3 (March 1991): 71–86.

Petrov, V. V. "Classical-Type Limit Theorems for Sums of Independent Random Variables." *Limit Theorems of Probability Theory.* Ed. Yu. V. Prokhorov and V. Statulevičius. Springer-Verlag, 2000. 1–24.

Ready, Patrick J. and Paul A. Wintz. "Information Extraction, SNR Improvement, and Data Compression in Multispectral Imagery." *IEEE Transactions on Communications* COM-21 (October 1973): 1123–1131.

Rice, John A. *Mathematical Statistics and Data Analysis.* 2nd edition. Wadsworth, Inc., 1995.

Richards, John A. *Remote Sensing Digital Image Analysis.* Springer-Verlag, 1993.

Robinson, E. A. "Predictive decomposition of seismic traces." *Geophysics* 22 (1957): 767–778.

Sacchi, Mauricio D., Danilo R. Velis, and Alberto H. Cominguez. "Minimum entropy deconvolution with frequency-domain constraints." *Geophysics* 59 (June 1994): 938–945.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics.* KTK Scientific Publishers, 1986.

Shalvi, Ofir and Ehud Weinstein. "New Criteria for Blind Deconvolution of Nonminimum Phase Systems (Channels)." *IEEE Transactions on Information Theory* 36 (March 1990).

Shannon, C. E. "A Mathematical Theory of Communications." *Bell System Technical Journal* 27 (July-October 1948): 379–423,623–656.

Strang, Gilbert. *Linear Algebra and its Applications.* 3rd edition. Harcourt Brace & Company, 1988.

Sunklodas, J. "Approximation of Distributions of Sums of Weakly Dependent Random Variables by the Normal Distribution." *Limit Theorems of Probability Theory.* Ed. Yu. V. Prokhorov and V. Statulevičius. Springer-Verlag, 2000. 1–24.

Turk, Mathew A. and Alex P. Pentland. "Face Recognition Using Eigenfaces." *Computer Vision and Pattern Recognition.* IEEE, 1991, 586–591.

Walden, A. T. "Non-Gaussian reflectivity, entropy, and deconvolution." *Geophysics* 50 (1985): 2862–2888.

Walker, Colin and T. J. Ulrych. "Autoregressive Recovery of the Acoustic Impedance." *Geophysics* 48 (October 1983): 1338–1350.

Wiggins, R. "Minimum entropy deconvolution." *Geoexploration* 16 (1978): 21–95.

Yilmaz, Öz. *Seismic Data Analysis: Processing, inversion and Interpretation of Seismic Data.* Society of Exploration Geophysicists, 2001.

# APPENDIX A

# Additional Proofs

*Proof that equation (3.11) is a maximum.* Consider the Kullback-Leibler divergence measure [e.g. Sakamoto et al., 1986, p. 38],

$$I(p_1, p_2) = \mathrm{E}\left[\ln\left(\frac{p_1(y)}{p_2(y)}\right)\right] = \int_{-\infty}^{\infty} p_1(y)\,\frac{p_1(y)}{p_2(y)}dy \tag{A.1}$$

where $p_1(y)$ and $p_2(y)$ are probability density functions (pdfs). Jensens's inequality [e.g. Cover and Thomos, 1991] states that if $f(y)$ is convex and $y$ is a random variable, then

$$\mathrm{E}(f(y)) \geq f(\mathrm{E}(y)). \tag{A.2}$$

Letting $y = p_2(y)/p_1(y)$, and defining the convex function, $f(y) = -\ln(y)$, equation (A.2) is applicable. In other words,

$$I(p_1(y), p_2(y)) = \mathrm{E}(f(y)) \geq f(\mathrm{E}(y)) = -\ln\left[\int_{-\infty}^{\infty} p_1(y)\,\frac{p_2(y)}{p_1(y)}dy\right] = -\ln\left[\int_{-\infty}^{\infty} p_2(y)\,dy\right] = 0.$$

Hence,

$$I(p_1(y), p_2(y)) \geq 0. \tag{A.3}$$

Recall the extreme point given in equation (3.11),

$$p_Y(y) = \exp\left(-1 + \lambda_\circ + \sum_{i=1}^{l} \lambda_i r_i(y)\right). \tag{A.4}$$

Let $g(y)$ be a pdf which obeys the same moment constraints as $p_Y(y)$. The differential entropy of $g(y)$ is

$$\begin{aligned} h(g(y)) &= -\int_{-\infty}^{\infty} g(y) \ln g(y)\, dy \\ &= -\int_{-\infty}^{\infty} g(y) \ln \left( \frac{g(y)}{p_Y(y)} p_Y(y) \right) dy \\ &= -\int_{-\infty}^{\infty} g(y) \ln \left( \frac{g(y)}{p_Y(y)} \right) dy - \int_{-\infty}^{\infty} g(y) \ln p_Y(y)\, dy. \end{aligned}$$

However, from equations (A.1) and (A.3),

$$\int_{-\infty}^{\infty} g(y) \ln \left( \frac{g(y)}{p_Y(y)} \right) dy = I(g(y), p_Y(y)) \geq 0.$$

Therefore, since $g(y)$ and $p_Y(y)$ share the same moment constraints,

$$\begin{aligned} h(g(y)) &\leq -\int_{-\infty}^{\infty} g(y) \ln p_Y(y)\, dy \\ &= -\int_{-\infty}^{\infty} p_Y(y) \ln p_Y(y)\, dy. \end{aligned}$$

Hence, $h(g(y)) \leq h(p_Y(y))$ and the proof is complete.

*Proof of equation (5.5).* It was shown in equation (3.36) that independent random variables are also nonlinearly uncorrelated. That is, given two independent and random variables, $y_1$ and $y_2$, and two arbitrary functions, $g_1(y_1)$ and $g_2(y_2)$,

$$\mathrm{E}[g_1(y_1) g_2(y_2)] = \mathrm{E}[g_1(y_1)] \mathrm{E}[g_2(y_2)]. \tag{A.5}$$

Additionally, recall, from Chapter 3, the definition of kurtosis: $\kappa_4 = \mathrm{E}(y_1^4) - 3[\mathrm{E}(y_1^2)]^2$. Thus, the kurtosis of the sum of $y_1$ and $y_2$ is

$$\kappa_4(y_1 + y_2) = \mathrm{E}\left[(y_1 + y_2)^4\right] - 3\left\{\mathrm{E}\left[(y_1 + y_2)^2\right]\right\}^2$$

where

$$(y_1 + y_2)^4 = y_1^4 + 4y_1^3 y_2 + 6y_1^2 y_2^2 + 4y_1 y_2^3 + y_2^4$$

and

$$(y_1 + y_2)^2 = y_1^2 + 2y_1y_2 + y_2^2.$$

If $y_1$ and $y_2$ are independent, then they are also nonlinearly uncorrelated. Hence, equation (A.5) is applicable, and after some algebra, assuming that $\mathrm{E}(y_1) = \mathrm{E}(y_2) = 0$,

$$\begin{aligned}\kappa_4(y_1 + y_2) &= \mathrm{E}(y_1^4) + \mathrm{E}(y_2^4) + 6\mathrm{E}(y_1^2)\,\mathrm{E}(y_2^2) - 3\left[\mathrm{E}(y_1^2)^2 + \mathrm{E}(y_2^2)^2 + 2\mathrm{E}(y_1^2)\,\mathrm{E}(y_2^2)\right] \\ &= \mathrm{E}(y_1^4) - 3\mathrm{E}(y_1^2)^2 + \mathrm{E}(y_2^4) - 3\mathrm{E}(y_2^2)^2;\end{aligned}$$

thus, allowing for equation (5.5) and completing the proof.

*Proof of equation (5.11).* Consider the moment generating function (mgf) [e.g. Rice, 1995, pp. 142-144],

$$M_{Y_1}(t) = \int_{-\infty}^{\infty} \exp(ity)\, p_{Y_1}(y_1)\, dy_1 = \mathcal{F}^{-1}\left(p_{Y_1}\right) \tag{A.6}$$

where $i = \sqrt{-1}$ and $y_1 \sim p_{Y_1}(y_1)$. $M_{Y_1}(t)$ defines the moments of $p_{Y_1}(y_1)$; hence, both $M_{Y_1}(t)$ and $p_{Y_1}(y_1)$ are equally valid representations of the random variable, $y_1$. Further, equation (A.6) is recognized as one half of a Fourier transform pair; hence,

$$p_{Y_1}(y_1) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-ity_1)\, M_{Y_1}(t)\, dt.$$

With a second random variable, $y_2$, the convolution theorem [e.g. Bracewell, 1978, pp. 108-111] gives,

$$M_{Y_1}(y_1)\, M_{Y_2}(y_2) = \mathcal{F}^{-1}\left[p_{Y_1}(y_1) * p_{Y_2}(y_2)\right] \tag{A.7}$$

where the operator, $\mathcal{F}^{-1}$, is defined by equation (A.6).

If $y_1$ and $y_2$ are independent, then their sum, $x = y_1 + y_2$, has the mgf,

$$\begin{aligned}M_X(t) &= \mathrm{E}[\exp(itx)] \\ &= \mathrm{E}\{\exp[it(y_1 + y_2)]\} \\ &= \mathrm{E}[\exp(ity_1)]\,\mathrm{E}[\exp(ity_2)] \\ &= M_{Y_1}(t)\, M_{Y_2}(t).\end{aligned}$$

Therefore, it follows, from equation (A.7), that $x$ has the pdf,

$$p_X(x) = p_{Y_1} * p_{Y_2}.$$

Equation (5.11), and thus completion of the proof, follows from the fact that the variance of the sum of two independent random variables is the sum of the variances of the random variables.