Reduced Order Linear System Identification

A dissertation presented by

Mohammad Khalil

to

The Department of Civil and Environmental Engineering

in partial fulfillment of the requirements
for the degree of
Master of Applied Science
in the subject of
Civil Engineering

Carleton University
Ottawa, Ontario, Canada
August 2006

The Master of Applied Science in Civil Engineering is a joint program
with the University of Ottawa, administered by
The Ottawa-Carleton Institute for Civil Engineering

Your file Votre référence
ISBN: 978-0-494-18318-2
Our file Notre référence
ISBN: 978-0-494-18318-2

Abstract

The investigation undertaken explores the feasibility of identifying a reduced order model of linear dynamical system operating on the mid-frequency regime. Proper Orthogonal Decomposition and Independent Component Analysis are used as vehicles for model reduction. Such reduced-order model circumvents the limitations of traditional modal analysis which, although well-adapted in the low-frequency range, is prone to computational and conceptual difficulties in the mid-frequency range.

The inverse problem involving identification of the system matrices (namely mass, damping and stiffness matrices) are posed in the framework of a least-squares estimation problem. To achieve this objective, Kronecker Algebra is aptly exploited to identify these matrix-valued variables. The concept of Tikhonov regularization permits additional physical constraints to be satisfied in terms of a symmetric property of the system matrices.

The usefulness of the proposed methodology is demonstrated using a simple discrete linear dynamical system.

Dedicated to my wife Samar,
my father Adnan,
and my mother Athra.

Acknowledgments

I have been very privileged to have the intuitive and supportive advisor Professor Abhijit Sarkar. I have been stimulated and excited by his constant flow of ideas. He has fostered a most open, friendly and collaborative research atmosphere. He also knew when to give me a little push in the forward direction when I needed it.

Throughout the last two years, I was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the High Performance Computing Virtual Laboratory of Canada (HPCVL), Carleton University, and through the generosity of my advisor.

Miss Xiaoli Zhu, Professor Sarkar's other student, has been a supportive and influential colleague. She has been a knowledgable research companion and she has certainly made my research experience more memorable.

On the foreign front, I must thank Professor Sondipon Adhikari of the University of Bristol (UK) for guiding me towards an understanding of damping modeling and identification in vibrating structures. The donation of some of his numerical algorithms is greatly appreciated.

I would like to thank my father, Adnan, and my mother, Athra, for their endless financial as well as emotional support. Their backing of my decision to leave the field of biological sciences to pursue a career in the applied sciences field was crucial for my success.

Finally, my wife Samar has been my guiding light over these last two years. I am indebted to her for the endless emotional support and encouragement.

Citations

The context of the Kronecker Algebra Section of Chapter 3 is obtained from the following book:

> "Kronecker Products and Matrix Calculus: With Applications", A. Graham, Ellis Horwood, 1981.

The fundamentals behind Chapter 4 were acquired from the following book:

> "Turbulence, Coherent Structures, Dynamical Systems and Symmetry", P. Holmes, J. L. Lumley, and G Berkooz, Cambridge University Press, 1996.

Finally, the context of large portions of Chapter 5, as well as Sections 3.1 and 3.2 of Chapter 3, are derived from the following books:

> "Independent Component Analysis", A. Hyvärinen, J. Karhunen, and E. Oja, Wiley, 2001.
>
> "Independent Component Analysis:Theory and Applications", T. Lee, Kluwer Academic Publishers, 1998.
>
> "Independent Component Analysis: Principles and Practice", S. Roberts and R. Everson, Cambridge University Press, 2001.

Contents

List of Figures

List of Acronyms

FEM Finite Element Method

FRF Frequency Response Function

ICA Independent Component Analysis

MLE Maximum Likelihood Estimation

POD Proper Orthogonal Decomposition

POM Proper Orthogonal Mode

ROM Reduced-Order Model

Chapter 1

Introduction

1.1 Motivation and Problem Statement

The general nature of an *inverse problem* is to deduce a cause from an effect. Consider a physical system depending on a set of parameters. If all of the parameters were known then for a given input we could predict the output. This is referred to as a *forward problem.* It may happen, however, that some of the parameters characterizing the system are not known, or are roughly known, being inaccessible to measurement. We infer the values of these parameters from the output of the system. Thus we seek the cause (i.e. the system parameters) given the effect (i.e. the output of the system for a given input).

An important example is the inverse problem of geophysics, in which we investigate the structure of the interior of the earth. Elastic waves may propagate through the earth in a manner which depends on the material properties of the earth. A concentrated source of energy at the surface causes waves to penetrate into the earth which are then partially reflected back to the surface. If the material properties of the earth's interior were known

completely then we could predict the nature of the reflected wave from knowledge of the source. Since in fact we cannot measure these properties directly we seek to infer them by observing the reflected waves, or outputs, in response to a collection of known sources, or inputs.

In mathematical formulation of such problems, we typically seek to estimate one or more coefficients in a system of differential equations, given partial knowledge of certain special solutions of the equation. In the seismology problem just discussed, the propagation of waves in the earth is governed by the equations of elasticity, a system of partial differential equations in which the material properties of the earth appear as coefficient functions.

Inverse problems in differential equations arise in a variety of important scientific areas, such as optics, quantum mechanics, astronomy, and medical imaging, as well as materials testing, structural analysis, geophysics, and hydrology, to name a few.

In general, inverse problems are classified as ill-posed, i.e. the solution obtained for a given set of data is unstable to small perturbations in input data or system parameters (Hadamard, [1]). There are techniques developed to deal with such ill-posed aspect of inverse problems. For example, Tikhonov regularization can be used along with the regular inversion technique to arrive at a well-posed minimization problem (Tikhonov and Arsenin, [2]).

System identification, a well-known form of an inverse problem, plays a crucial role in model-based prediction of dynamical systems. Normally, a finite representation of a continuous media described by a partial differential equation leads to a discrete model of the dynamical system. In the practical applications involving linear continuous operators, the discrete system is fully characterized by the so-called mass, stiffness, and damping

matrices. Once these matrices are estimated with reasonable confidence, such discrete model can reliably predict the behavior of the underlying dynamical system due to external disturbances.

In the context of structural dynamics, identification of these system matrices is achieved by traditional modal analysis techniques. In this approach, an efficient and accurate reduced-order model can be obtained in the so-called low-frequency range where the system exhibits distinct resonant modes. In this regime, only a handful of structural modes contribute to the total response. However, the situation changes significantly in the mid-frequency regime whereby a large number of structural modes contribute to the output. This renders the modal analysis impractical in constructing a reduced-order modal of the discrete dynamical system. Furthermore, the physical mechanism and mathematical form of damping plays a pivotal role which may lead to a complex eigenvalue problem. Additionally, the higher order eigenmodes are also susceptible to numerical error. This is partly due to the fact that a high resolution spatial discretization is necessary to capture short wavelength vibration features in the mid-frequency range [3].

On the other hand, the high-frequency vibration analysis involving wave and energy-based approaches, such as Statistical Energy Analysis (Lyon, [4]), gives unwieldy results due to conceptual difficulties. Such energy-based approach fails to capture the propagation effect emerging from the phase information of the oscillating system.

Consequently, it is important to construct a reduced-order model based on other approaches. To this effect, POD and ICA based methods appear themselves to be viable alternatives. The recent availability of data-acquisition hardware (such as laser-vibrometry [5]) makes it possible to acquire highly resolved spatio-temporal vibration data rendering

POD and ICA based methods amenable to practical application.

The current investigation explores the feasibility of reduced-order modeling to alleviate the limitations of traditional modal analysis to arrive at a low-dimensional model. To reduce computational effort, such a model will be exploited to identify the underlying dynamical system to be used as a predictive tool. This is the subject of this thesis.

1.2 Thesis Structure

The thesis is organized in the following manner. Chapter 1 presents an introduction and problem statement. A brief relevant literature review is presented in Chapter 2. Chapter 3 introduces the requisite mathematical preliminaries necessary for this thesis. The POD method and its application to reduced-order modeling is detailed in Chapter 4. The ICA method is delineated and then contrasted with POD in Chapter 5. In Chapter 6, the detailed formulation involving the system identification method is described. Chapter 7 reports the results from a numerical example elucidating the feasibility and usefulness of the mathematical formulation. The thesis concludes in Chapter 8 where a summary and findings of the current investigation are detailed including avenues of future investigation.

Chapter 2

Literature Review

In this Chapter, we provide a brief description of relevant literature. The review is by no means exhaustive, but relates only to the current investigation undertaken in the thesis. Note that each of the subtopics touched upon relates to a separate field of active research interest. For brevity, we restrict our attention only to a limited portion from the vast body of the existing literature.

2.1 System Modeling

In general, modeling of a system involves deriving mathematical approximation of certain spatially and/or temporally evolving physical phenomenon. Such an approximate mathematical model offers predictive tools to describe the evolutionary features of the system behavior. Broadly speaking, these mathematical models can be categorized as follows[6]:

- *Time and Frequency domain models*: Output of a dynamical systems can be represented either in time or frequency domain. The frequency domain representation is

normally achieved by use of the Laplace, Fourier and z-transformations.

- *Deterministic and Stochastic models*: Whenever output of the system exhibits significant random variablity, a stochastic model is generally adopted. Otherwise, a deterministic model can adequately predict the system behavior with sufficient accuracy.

- *Continuous-time and Discrete-time models*: Whenever the mathematical models describing the system permit closed-form (analytical) solutions, the continuous-time model is preferable. In absence of such convenient representation of the output, one has to take recourse to discrete-time model.

- *Parametric versus Non-parametric models*: A parametric model contains a finite set of parameters dictating the behavior of the system. On the contrary, whenever system behavior can not be described by a finite set of parameters, a non-parametric model is adopted.

A parametric linear deterministic system model is considered in this investigation. Both time and frequency domain representations are adopted for the mathematical formulation.

2.2 System Identification Methods

Numerous system identification methods are available in the existing literature. Each of these methods are well-suited to particular applications. A brief overview of some of the widely-used techniques follows.

2.2.1 Modal Analysis

Modal analysis determines the fundamental vibration mode shapes and corresponding frequencies of a vibrating system. In the low-frequency domain, only a few modes contribute to the total response of the system [7, 8, 9]. In general, modal analysis is used to arrive at a reduced order model of a system in the low-frequency regime. Mathematically speaking, this approach involves spectral representation of the original operator describing the system behavior. Experimental modal analysis permits extraction of such modal parameters through experimental measurement [10, 11]. The structure is set into motion by either a mechanical shaker or an impact hammer or to ambient vibration (such as wind or traffic load in a ridge structure), with the corresponding response measured at one or more points.

In one of the earliest works, Lancaster [12] had shown that the mass, damping and stiffness matrices can be obtained from the measured complex modes and frequencies. Ibrahim [13] used the higher order analytical modes together with the experimental set of complex modes to compute improved mass, stiffness and damping matrices. Subsequently, Adhikari [14] proposed a method to identify the system matrices using the residues and poles of the measured transfer functions. Roemer and Mook [15] have developed methods in the time domain for simultaneous identification of the mass, damping and stiffness matrices. Chen et al. [16] have proposed a direct frequency domain technique for identification of the system matrices. Zhao et al. [17] proposed a modal-data based identification method for a linear dynamical structure under unknown forcing. Yan et al. [18] proposed a wavelet-transform based method to identify modal parameters of a linear system.

Baruch [19] showed that simultaneous changes in the mass and stiffness matrices cannot be identified by using modal data only. The argument is that the same mode shapes and

natural frequencies can be obtained for an infinite number of different pairs of stiffness and mass matrices. Baruch also proposed a method to apply corrections to existing mass and stiffness matrices using collected modal data [20, 21, 22]. The methodology however does not consider systems with damping. Provasi et al. [23] proposed a modal parameter identification method in the frequency domain. The algorithm is based on the same formulation of the extended Kalman filter. The proposed method is recursive in its estimation of the modal parameters and relies on an initial "guess" of theses parameters. Furthermore, the method can only identify proportionally-damped systems.

Each of the above modal analysis based methods have their own advantages and disadvantages. The common issues regarding the identification of the system matrices using conventional modal analysis are:

- The accuracy of the identified modal parameters, and consequently the system matrices, relies on the presence of distinct 'peaks' in the measured FRFs.
- If the damping is non-proportional, the identification of complex modes poses a serious challenge [24, 25].

The first problem is inherent to conventional modal analysis. If the peaks in the measured FRFs are not distinct or are closely spaced, the modal parameter extraction procedure is difficult to apply [10]. As a consequence, the identified system matrices using the extracted modal parameters become erroneous. It is therefore difficult to extend the modal identification procedure in the mid-frequency range, or for periodic systems (such as bladed disks in turbomachineries) inherently containing closely spaced modes (in the form of pass-bands and stop-bands). The second problem arises for systems with high damping materials

such as a panel with viscoelastic damping.

2.2.2 Direct Identification

Direct identification of system parameters is another system analysis tool. In such an approach, the least-square estimation method is used to tackle deterministic systems [6, 26]. For stochastic systems, the maximum-likelihood method is applied [27, 28]. Another technique used for stochastic system analysis is the maximum entropy method [29].

Direct identification methods are an alternative to the modal analysis technique. Fritzen [30] and Mottershead [31] proposed direct system parameter identification methods. Apart from the comparative robustness of these identification procedures, the same authors state that the identified matrices are in general non-unique and likely to be faced with a set of incomplete data [32]. Furthermore, constraints satisfying certain physical properties of the system (i.e. symmetry of system matrices for example) are not dealt with in these methodologies. Software packages are also available for such system identification techniques (for example refer to [33]).

2.2.3 Nonlinear System Identification

In certain restrictive cases, the concept of modal analysis for linear systems is extended to non-linear systems using the idea of non-linear normal modes [34]. However, their application in non-linear structural dynamics is limited to single non-linear mode of vibration only.

One alternative to modal analysis is the Volterra and Wiener series approach, as it provides a relationship between the input and output of a non-linear system [35, 36]. For

example, Gifford and Tomlinson applied the Volterra series in the field of non-linear structural dynamics [37]. The method has been widely applied in subsequent investigations (for example see [38, 39]). However, one weakness of the method is that convergence is not always guaranteed and greatly depends on the extent of non-linearity.

2.3 Model Reduction Techniques

2.3.1 Proper Orthogonal Decomposition

POD provides a basis for the spectral decomposition of a spatio-temporal signal. Due to its attractive mathematical properties, POD has been used extensively in numerous applications. Perhaps the most compelling property is its mean-square optimality: it provides the most efficient way of capturing the dominant components of a high-dimensional signal with only a few dominant scales of fluctuations, namely Proper Orthogonal Modes (POMs).

The POD method has been used for the model reduction of linear as well as non-linear dynamical systems, for example see [40, 41, 42]. In the context of system identification, POD has been applied to measured displacements of a discrete undamped system with a known mass matrix leading to an estimation of the normal modes [43, 44]. The POD method is applied to the non-linear problem of vibro-impacting beams and rotors to create low-dimensional models, via a Galerkin projection using POMs as a basis [45, 46]. The other application of POD on model-reduction of non-linear mechanical systems are reported in [47, 48, 49, 50]. POD has been used for the identification of a non-linear dynamical system [51, 52], whereby the non-linear stiffness parameter was identified with a reduced-order model.

In various disciplines, the technique relating to POD has different nomenclature such as Karhunen-Loéve decomposition, principal component analysis, singular system analysis and singular value decomposition. The POD basis functions are also termed as empirical eigenfunctions, empirical basis functions, and empirical orthogonal functions. POD was introduced independently by numerous researchers (see [53]), including Kosambi [54], Loéve [55], Karhunen [56], Pougachev [57], and Obukhov [58]. The method has been applied in numerous disciplines including random process theory [29], image processing [59], signal analysis [60], and data compression [61].

2.3.2 Independent Component Analysis

The advantage of POD stems from the fact that resulting signals of the projected system response onto the POMs are uncorrelated. ICA achieves higher order decorrelation of the projected signals by applying further transformation of the resulting uncorrelated POD signals.

The technique of ICA, with a different nomenclature, was introduced in the early 1980s by Herault, Jutten, and Ans [62, 63, 64]. All through the 1980s, ICA was mostly known among researchers in the field of neural networks. In the field of higher order spectral analysis, ICA was introduced by Cardoso [65] and Comon [66]. In the field of signal processing, there had been ICA-related approaches to blind deconvolution [67, 68]. The mathematical framework of multichannel blind deconvolution bears similarity with ICA techniques. ICA gained wider attention and growing interest after Bell and Sejnowski published their approach based on the infomax principle [69, 70]. This algorithm was further refined by Amari [71], and its fundamental connections to maximum likelihood estimation was es-

tablished. Hyvärinen and Oja presented the fixed-point or FastICA algorithm [72, 73, 74] contributing to the application of ICA to large-scale problems due to its computational efficiency.

ICA has not been as widely applied as POD in terms of model reduction of dynamical systems, specially in the field of structural dynamics. However, ICA is widely used as a reduction tool in brain imaging [75], econometrics [76], and image feature extraction [77]. In the context of structural vibration, ICA has been used in system fault detection [78, 79].

Chapter 3

Mathematical Preliminaries

This chapter introduces some general mathematical concepts to be used in the subsequent chapters. Firstly, the essential concepts from gradient-based optimization theory needed to explain the ICA algorithms are presented. Statistical Estimation theory, involving the mean and covariance of random processes, is discussed next. Lastly, a brief introduction to certain identities of Kronecker Algebra from which the foundation of the newly proposed system identification method is provided.

3.1 Vector and Matrix Gradients

In some ICA methods, the gradient of the determinant of a matrix must be computed. This section will introduce the concept of vector and matrix gradients.

3.1.1 Vector Gradient

Consider a scalar valued function f of N variables

$$f = f(x_1, \ldots, x_N) = f(\mathbf{x}) \tag{3.1}$$

where $\mathbf{x} = (x_1, \ldots, x_N)^T$. Assuming f to be differentiable with respect to all N variables x_i, the vector gradient $\frac{\partial f}{\partial \mathbf{x}}$ of f with respect to $\mathbf{x}$ is given by

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{pmatrix} \tag{3.2}$$

3.1.2 Matrix Gradient

Consider a scalar valued function f of a square matrix $\mathbf{W}$ of order N

$$f = f(w_{11}, \ldots, w_{ij}, \ldots, w_{NN}) = f(\mathbf{W}) \tag{3.3}$$

Assuming f to be differentiable with respect to all N^2 variables w_{ij}, the matrix gradient $\frac{\partial f}{\partial \mathbf{W}}$ of f with respect to $\mathbf{W}$ is given by

$$\frac{\partial f}{\partial \mathbf{W}} = \begin{pmatrix} \frac{\partial f}{\partial w_{11}} & \cdots & \frac{\partial f}{\partial w_{1N}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial w_{N1}} & \cdots & \frac{\partial f}{\partial w_{NN}} \end{pmatrix} \tag{3.4}$$

Gradient of the Determinant of Matrices

The determinant of a matrix is a scalar valued function. The ICA method requires the computation of the gradient of this function with respect to the matrix itself.

Starting with the matrix algebra identity for the inverse of a matrix[80]

$$\mathbf{W}^{-1} = \frac{1}{\det(\mathbf{W})}\mathrm{adj}(\mathbf{W}) \tag{3.5}$$

where $\mathrm{adj}(\mathbf{W})$ and $\det(\mathbf{W})$ represent the adjoint and determinant of matrix $\mathbf{W}$, respectively.

The adjoint is given by

$$\mathrm{adj}(\mathbf{W}) = \begin{pmatrix} W_{11} & \dots & W_{N1} \\ \vdots & & \vdots \\ W_{1N} & \dots & W_{NN} \end{pmatrix} \tag{3.6}$$

where the scalars W_{ij} are the cofactors of the matrix $\mathbf{W}$. The cofactors W_{ij} are determined by first obtaining the $(N-1) \times (N-1)$ submatrix of $\mathbf{W}$ that remains after removing the ith row and jth column from matrix $\mathbf{W}$, then computing the determinant of that submatrix and finally multiplying by $(-1)^{i+j}$.

Furthermore, the determinant of matrix $\mathbf{W}$ can also be expressed in terms of the matrix cofactors as

$$\det(\mathbf{W}) = \sum_{j=1}^{N} w_{ij} W ij \tag{3.7}$$

invariant of the choice of row i. Note that a W_{ij} for a given row i and column j is independent of the element w_{ij}, thus the determinant of $\mathbf{W}$ is a linear function of these elements. Taking the partial derivative of the determinant (3.7) with respect to these elements, we obtain

$$\frac{\partial \det(\mathbf{W})}{\partial w_{ij}} = W ij \tag{3.8}$$

According to the definition (3.4), (3.8) becomes

$$\frac{\partial \det(\mathbf{W})}{\partial \mathbf{W}} = \begin{pmatrix} W_{11} & \dots & W_{1N} \\ \vdots & & \vdots \\ W_{N1} & \dots & W_{NN} \end{pmatrix} \tag{3.9}$$

which when compared to (3.6) yields

$$\frac{\partial \det(\mathbf{W})}{\partial \mathbf{W}} = \operatorname{adj}\left(\mathbf{W}^T\right). \tag{3.10}$$

Applying the transpose operator to (3.5) yields

$$\begin{aligned} \left(\mathbf{W}^T\right)^{-1} &= \frac{1}{\det\left(\mathbf{W}^T\right)} \operatorname{adj}\left(\mathbf{W}^T\right) \\ &= \frac{1}{\det(\mathbf{W})} \operatorname{adj}\left(\mathbf{W}^T\right). \end{aligned} \tag{3.11}$$

Finally, applying identity (3.11) to (3.10) yields the gradient of the determinant of matrix $\mathbf{W}$ with respect to itself:

$$\frac{\partial}{\partial \mathbf{W}} \det(\mathbf{W}) = \left(\mathbf{W}^T\right)^{-1} \det(\mathbf{W}). \tag{3.12}$$

Furthermore, the gradient of the natural logarithm of the absolute value of the determinant is

$$\begin{aligned} \frac{\partial}{\partial \mathbf{W}} \ln\left(\left|\det(\mathbf{W})\right|\right) &= \frac{1}{\left|\det(\mathbf{W})\right|} \frac{\partial \left|\det(\mathbf{W})\right|}{\partial \mathbf{W}} \\ &= \left(\mathbf{W}^T\right)^{-1}. \end{aligned} \tag{3.13}$$

Interestingly, the gradient of some basic scalar functions of matrices can be computed analytically using matrix differential calculus, as apparent from identities (3.12) and (3.13)

3.2 Parameter Estimation

POD and ICA are based on the theory of statistics. Some parameter estimation techniques involving random variables or signals are necessary for such methods. An important issue is how to estimate the parameters from a given finite set of measurements of the random quantities. Least-square estimation, maximum likelihood estimation, and weiner filtering are such estimation techniques [81]. The choice of a suitable estimation method depends on the quantified data model at hand.

Let X be a random variable quantified by a set of parameters $\boldsymbol{\theta}$ whose estimate is needed. Assume there are T measurements of X denoted by $x(1), \ldots, x(T)$. $\widehat{\boldsymbol{\theta}}$ denotes the estimate of $\boldsymbol{\theta}$. In that sense, we have

$$\widehat{\boldsymbol{\theta}} = f(x(1), \ldots, x(T)) \tag{3.14}$$

The functional form of the estimator f depends on the choice of estimation method. To that effect, one must first select a suitable model that describes the data properly in the statistical sense. If the model of the data is unknown, one can make certain assumptions of the model and proceed. The first question one needs to answer is whether the parameters to be estimated are themselves random or deterministic.

In the context of POD, one needs to estimate first and second order statistical moments, namely the mean and variance/covariance of a given set of random variables. To estimate the statistical moments, the maximum likelihood method is widely used.

3.2.1 MLE

The maximum likelihood estimator is the value of $\boldsymbol{\theta}$ that makes the observations $x(1)$, $x(2), \ldots, x(T)$ most likely [82]. The likelihood function $\mathcal{L}(\boldsymbol{\theta})$ is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = p(x(1), \ldots, x(T) \mid \boldsymbol{\theta}) \tag{3.15}$$

In the case that the observations $x(1)$, $\ldots$, $x(T)$ are independent of each other, the likelihood function simplifies to (3.16).

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{T} p(x(i) \mid \boldsymbol{\theta}) \tag{3.16}$$

As most probability density functions possess some exponential trend, it is more convenient to deal with the logarithm of the likelihood function, or *log likelihood function*, $\ln \mathcal{L}(\boldsymbol{\theta})$. The values of θ_i that maximize the log likelihood function are the same values that maximize the likelihood function. Thus the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is found from the solution to the log likelihood equation (3.17).

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0} \tag{3.17}$$

If the likelihood equation (3.17) has multiple solutions, i.e. several extrema exist for the log likelihood function (3.15), the value of $\widehat{\boldsymbol{\theta}}$ that corresponds to the absolute maximum is chosen. Note that in (3.17) the log likelihood function is differentiated with respect to a vector containing the parameters to be estimated. Furthermore, the resulting system of equations to be solved might be nonlinear, depending on the probability distribution of the variable. In that case, nonlinear optimization can be applied to solve for the estimates.

The maximum likelihood estimator has several asymptotic optimality properties. Most

importantly, the maximum likelihood estimator achieves minimum variance, given by the Cramer-Rao lower bound, in the limit when sample size tends to infinity[83].

Univariate Normal Distribution Parameter Estimation

Assuming T independent observations $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ of a gaussian random variable are made. The probability density function is

$$p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]. \tag{3.18}$$

The liklihood function (3.16) is thus equal to

$$\begin{aligned} \mathcal{L}(\mu, \sigma^2) &= \prod_{i=1}^{T} p(x(i)) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{T}[x(i)-\mu]^2\right]. \end{aligned} \tag{3.19}$$

The log likelihood function is

$$\ln \mathcal{L}(\mu, \sigma^2) = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}[x(i)-\mu]^2. \tag{3.20}$$

The first likelihood equation is obtained by differentiating the log likelihood function (3.20) with respect to μ

$$\begin{aligned} \frac{\partial}{\partial\mu}\ln\mathcal{L}(\mu, \sigma^2)\Big|_{\mu=\widehat{\mu}\ ,\ \sigma^2=\widehat{\sigma}^2} &= \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{T}[x(i)-\widehat{\mu}] \\ &= 0 \end{aligned} \tag{3.21}$$

which, when solved, leads to the maximum likelihood estimate $\widehat{\mu}$ of the mean

$$\widehat{\mu} = \frac{1}{T}\sum_{i=1}^{T} x(i). \tag{3.22}$$

The second likelihood equation is obtained by differentiating the log likelihood function (3.20) with respect to σ^2, the estimate of the variance

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \ln \mathcal{L}\left(\mu, \sigma^2\right)\Big|_{\mu=\widehat{\mu}\ ,\ \sigma^2=\widehat{\sigma}^2} &= -\frac{T}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{T}\left[x(i) - \widehat{\mu}\right]^2 \\ &= 0 \end{aligned} \tag{3.23}$$

which, when solved, leads to the maximum likelihood estimate $\widehat{\sigma}^2$ of the variance

$$\widehat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T}\left[x(i) - \widehat{\mu}\right]^2. \tag{3.24}$$

Equations (3.22) and (3.24) are the maximum likelihood estimators for the mean and variance of a gaussian random variable of which T observations are available.

Multivariate Normal Distribution Parameter Estimation

Generalizing the above derivation to a gaussian random vector, the maximum likelihood estimate for its covariance matrix can be obtained. Let us assume that T independent observations $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ of a gaussian random vector $\mathbf{x}$ of dimension n are available. The probability density function is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\left(\det(\boldsymbol{\Sigma})\right)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{3.25}$$

where $\boldsymbol{\Sigma}$ denotes the covariance matrix and $\boldsymbol{\mu}$ the mean vector.

The liklihood function (3.16) is thus equal to

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \prod_{i=1}^{T} p(\mathbf{x}(i)) \\
&= \frac{1}{(2\pi)^{Tn/2} (\det(\boldsymbol{\Sigma}))^{T/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{T} (\mathbf{x}(i) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}(i) - \boldsymbol{\mu})\right]. \quad (3.26)
\end{aligned}$$

The log likelihood function is

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{Tn}{2} \ln(2\pi) - \frac{T}{2} \ln|\det(\boldsymbol{\Sigma})| \\
&\quad - \frac{1}{2} \sum_{i=1}^{T} (\mathbf{x}(i) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}(i) - \boldsymbol{\mu}). \quad (3.27)
\end{aligned}$$

The first likelihood equation is obtained by determining the vector gradient (3.2) of the log likelihood function (3.27) with respect to the vector $\boldsymbol{\mu}$:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}} \ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\Big|_{\boldsymbol{\mu}=\widehat{\boldsymbol{\mu}},\, \boldsymbol{\Sigma}=\widehat{\boldsymbol{\Sigma}}} \\
&= +\frac{1}{2} \sum_{i=1}^{T} \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) + \frac{1}{2} \sum_{i=1}^{T} \left[(\mathbf{x}(i) - \widehat{\boldsymbol{\mu}})^T \widehat{\boldsymbol{\Sigma}}^{-1}\right]^T \\
&= +\frac{1}{2} \sum_{i=1}^{T} \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) + \frac{1}{2} \sum_{i=1}^{T} \left[\widehat{\boldsymbol{\Sigma}}^{-1}\right]^T (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) \\
&= +\frac{1}{2} \sum_{i=1}^{T} \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) + \frac{1}{2} \sum_{i=1}^{T} \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) \\
&= +\sum_{i=1}^{T} \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) \\
&= +\widehat{\boldsymbol{\Sigma}}^{-1} \sum_{i=1}^{T} (\mathbf{x}(i) - \widehat{\boldsymbol{\mu}}) \\
&= \mathbf{0}_{n \times 1} \quad (3.28)
\end{aligned}$$

where $\mathbf{0}_{n\times 1}$ denotes the column vector of order n whose elements are zero. The third line of (3.28) is obtained by making the assumption that the estimated covariance matrix and its inverse are symmetric. That is a reasonable assumption to make as the definition of the covariance matrix implies symmetry. Assuming that $\widehat{\boldsymbol{\Sigma}}^{-1}$ is not zero, the solution of (3.28) leads to the maximum likelihood estimate $\widehat{\boldsymbol{\mu}}$ of the mean

$$\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{i=1}^{T} \mathbf{x}(i)\,. \tag{3.29}$$

The second likelihood equation is obtained by determining the matrix gradient (3.4) of the log likelihood function (3.27) with respect to the matrix $\boldsymbol{\Sigma}$. Note that the last term of the log likelihood function (3.27) can be cast in another form as

$$\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{T} (\mathbf{x}(i)-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}(i)-\boldsymbol{\mu}) \\
&= -\frac{1}{2}\sum_{i=1}^{T} \operatorname{tr}\left((\mathbf{x}(i)-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}(i)-\boldsymbol{\mu})\right) \\
&= -\frac{1}{2}\sum_{i=1}^{T} \operatorname{tr}\left((\mathbf{x}(i)-\boldsymbol{\mu})(\mathbf{x}(i)-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\sum_{i=1}^{T} (\mathbf{x}(i)-\boldsymbol{\mu})(\mathbf{x}(i)-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\left(\sum_{i=1}^{T} (\mathbf{x}(i)-\boldsymbol{\mu})(\mathbf{x}(i)-\boldsymbol{\mu})^{T}\right)\boldsymbol{\Sigma}^{-1}\right)
\end{aligned} \tag{3.30}$$

The above expansion applies the identity that the trace of a scalar is the scalar itself and that $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ (see Section 3.3.3). Applying (3.30) to the log likelihood function (3.27), we have

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= -\frac{TN}{2}\ln(2\pi) - \frac{T}{2}\ln\left|\det(\boldsymbol{\Sigma})\right| \\
&\quad -\frac{1}{2}\operatorname{tr}\left(\left(\sum_{i=1}^{T} (\mathbf{x}(i)-\boldsymbol{\mu})(\mathbf{x}(i)-\boldsymbol{\mu})^{T}\right)\boldsymbol{\Sigma}^{-1}\right).
\end{aligned} \tag{3.31}$$

The second likelihood function is thus

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln \mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\Big|_{\boldsymbol{\mu}=\widehat{\boldsymbol{\mu}}\ ,\ \boldsymbol{\Sigma}=\widehat{\boldsymbol{\Sigma}}} \\
&= -\frac{T}{2}\widehat{\boldsymbol{\Sigma}}^{-1} + \frac{1}{2}\widehat{\boldsymbol{\Sigma}}^{-1}\left[\sum_{i=1}^{T}\left(\mathbf{x}\left(i\right)-\widehat{\boldsymbol{\mu}}\right)^{T}\left(\mathbf{x}\left(i\right)-\widehat{\boldsymbol{\mu}}\right)\right]\widehat{\boldsymbol{\Sigma}}^{-1} \\
&= -\frac{1}{2}\widehat{\boldsymbol{\Sigma}}^{-1}\left[T\mathbf{I}_{n} - \left[\sum_{i=1}^{T}\left(\mathbf{x}\left(i\right)-\widehat{\boldsymbol{\mu}}\right)^{T}\left(\mathbf{x}\left(i\right)-\widehat{\boldsymbol{\mu}}\right)\right]\widehat{\boldsymbol{\Sigma}}^{-1}\right] \\
&= \mathbf{0}_{n\times n}
\end{aligned} \tag{3.32}$$

where $\mathbf{0}_{n\times n}$ denotes the matrix of order $n \times n$ whose elements are zero, and $\mathbf{I}_n$ denotes the identity matrix of order n. The first term in the first line of (3.32) is obtained by applying Identity (3.13) for the gradient of the natural logarithm of the determinant of a matrix. The derivation of the second term can be found in [28]. Assuming that $\widehat{\boldsymbol{\Sigma}}^{-1}$ is not zero, the solution of (3.32) leads to the maximum likelihood estimate, $\widehat{\boldsymbol{\Sigma}}$, of the covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{i=1}^{T}\left(\mathbf{x}\left(i\right)-\widehat{\boldsymbol{\mu}}\right)^{T}\left(\mathbf{x}\left(i\right)-\widehat{\boldsymbol{\mu}}\right). \tag{3.33}$$

Equations (3.29) and (3.33) are the maximum likelihood estimators for the mean vector and covariance matrix of a random vector that follows a multivariate normal distribution, respectively.

3.3 Kronecker Algebra

This section introduces the concepts of the Kronecker matrix product and its properties. We closely follow the book by Graham [84].

3.3.1 Unit Vectors and Elementary Matrices

The unit vectors of order n are defined as

$$\mathbf{e}_1 = \begin{Bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}, \quad \mathbf{e}_2 = \begin{Bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}, \quad \ldots, \quad \mathbf{e}_n = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{Bmatrix}. \tag{3.34}$$

The identity vector $\mathbf{e}$ of order n is defined as

$$\mathbf{e} = \begin{Bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{Bmatrix}. \tag{3.35}$$

From (3.34) and (3.35), we obtain the relation

$$\mathbf{e} = \sum_{i=1}^{n} \mathbf{e}_i. \tag{3.36}$$

The elementary matrix $\mathbf{E}_{ij}$ is defined as the matrix of order $m \times n$ which has a unity in the (i, j)th position with all other elements being zero, i.e.

$$\mathbf{E}_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & otherwise \end{cases}. \tag{3.37}$$

The relation between $\mathbf{e}_i$, $\mathbf{e}_j$ and $\mathbf{E}_{ij}$ is as follows

$$\mathbf{E}_{ij} = \mathbf{e}_i \mathbf{e}_j^T. \tag{3.38}$$

where $\mathbf{e}_j^T$ denotes the transposed vector (that is, the row vector) of $\mathbf{e}_j$.

The Kronecker delta δ_{ij} is defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases} . \tag{3.39}$$

It can also be expressed as

$$\delta_{ij} = \mathbf{e}_i^T \mathbf{e}_j = \mathbf{e}_j^T \mathbf{e}_i. \tag{3.40}$$

We can now determine some relations betwen unit vectors and elementary matrices. Using (3.38) we have

$$\begin{aligned} \mathbf{E}_{ij}\mathbf{e}_r &= \mathbf{e}_i \mathbf{e}_j^T \mathbf{e}_r \\ &= \delta_{jr}\mathbf{e}_i \end{aligned} \tag{3.41}$$

and

$$\begin{aligned} \mathbf{e}_r^T \mathbf{E}_{ij} &= \mathbf{e}_r^T \mathbf{e}_i \mathbf{e}_j^T \\ &= \delta_{ri}\mathbf{e}_j^T. \end{aligned} \tag{3.42}$$

Also,

$$\begin{aligned} \mathbf{E}_{ij}\mathbf{E}_{rs} &= \mathbf{e}_i \mathbf{e}_j^T \mathbf{e}_r \mathbf{e}_s^T \\ &= \delta_{jr}\mathbf{e}_i \mathbf{e}_s^T \\ &= \delta_{jr}\mathbf{E}_{is}. \end{aligned} \tag{3.43}$$

In particular if $r = j$, we have

$$\begin{aligned} \mathbf{E}_{ij}\mathbf{E}_{js} &= \delta_{jj}\mathbf{E}_{is} \\ &= \mathbf{E}_{is}, \end{aligned} \tag{3.44}$$

and more generally

$$\begin{aligned} \mathbf{E}_{ij}\mathbf{E}_{js}\mathbf{E}_{sm} &= \mathbf{E}_{is}\mathbf{E}_{sm} \\ &= \mathbf{E}_{im}. \end{aligned} \tag{3.45}$$

Notice that from (3.43) that

$$\mathbf{E}_{ij}\mathbf{E}_{rs} = 0 \text{ if } j \neq r. \tag{3.46}$$

3.3.2 Decompositions of a Matrix

We consider a matrix $\mathbf{A}$ of order $m \times n$ have the following form

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = [a_{ij}]. \tag{3.47}$$

We denote the n columns of $\mathbf{A}$ by $\mathbf{A}_{\cdot 1}, \mathbf{A}_{\cdot 2}, \dots, \mathbf{A}_{\cdot n}$, so that

$$\mathbf{A}_{\cdot j} = \begin{Bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{Bmatrix} \quad (j = 1, 2, \dots, n) \tag{3.48}$$

and the m rows of $\mathbf{A}$ by $\mathbf{A}_{1\cdot}, \mathbf{A}_{2\cdot}, \dots, \mathbf{A}_{m\cdot}$, so that

$$\mathbf{A}_{i\cdot} = \begin{Bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{in} \end{Bmatrix} \quad (i = 1, 2, \dots, m) \tag{3.49}$$

Both the $\mathbf{A}_{\cdot j}$ and $\mathbf{A}_{i\cdot}$ are column vectors. In this notation we can write $\mathbf{A}$ as

$$\mathbf{A} = \left[\mathbf{A}_{\cdot 1}, \mathbf{A}_{\cdot 2}, \ldots, \mathbf{A}_{\cdot n}\right], \tag{3.50}$$

or as

$$\mathbf{A} = \left[\mathbf{A}_{1\cdot}, \mathbf{A}_{2\cdot}, \ldots, \mathbf{A}_{m\cdot}\right]^T. \tag{3.51}$$

The elements, the columns and the rows of $\mathbf{A}$ can be expressed in terms of the unit vectors as follows:

$$\mathbf{A}_{\cdot j} = \mathbf{A}\mathbf{e}_j \tag{3.52}$$

and

$$\mathbf{A}_{i\cdot}^T = \mathbf{e}_i^T\mathbf{A}. \tag{3.53}$$

So that

$$\begin{aligned} \mathbf{A}_{i\cdot} &= \left(\mathbf{e}_i^T\mathbf{A}\right) \\ &= \mathbf{A}^T\mathbf{e}_i. \end{aligned} \tag{3.54}$$

The (i,j)th element of $\mathbf{A}$ can now be rewritten as

$$\begin{aligned} a_{ij} &= \mathbf{e}_i^T\mathbf{A}\mathbf{e}_j \\ &= \mathbf{e}_j^T\mathbf{A}^T\mathbf{e}_i. \end{aligned} \tag{3.55}$$

We can express $\mathbf{A}$ as the sum

$$\mathbf{A} = \sum_i \sum_j a_{ij}\mathbf{E}_{ij} \tag{3.56}$$

where the $\mathbf{E}_{ij}$ are of the same order as $\mathbf{A}$ so that

$$\mathbf{A} = \sum_i \sum_j a_{ij}\mathbf{e}_i\mathbf{e}_j^T. \tag{3.57}$$

From (3.52) and (3.57), we have

$$\begin{aligned}
\mathbf{A}_{\cdot j} &= \mathbf{A}\mathbf{e}_j \\
&= \left(\sum_i \sum_k a_{ik}\mathbf{e}_i\mathbf{e}_k^T\right)\mathbf{e}_j \\
&= \sum_i \sum_k a_{ik}\mathbf{e}_i\left(\mathbf{e}_k^T\mathbf{e}_j\right) \\
&= \sum_i \sum_k a_{ik}\mathbf{e}_i\delta_{jk} \\
&= \sum_i a_{ij}\mathbf{e}_i. \qquad (3.58)
\end{aligned}$$

Similarly

$$\mathbf{A}_{i\cdot} = \sum_j a_{ij}\mathbf{e}_j \qquad (3.59)$$

so that

$$\mathbf{A}_{i\cdot}^T = \sum_j a_{ij}\mathbf{e}_j^T. \qquad (3.60)$$

It follows from (3.57), (3.58), and (3.60) that

$$\mathbf{A} = \sum_j \mathbf{A}_{\cdot j}\mathbf{e}_j^T \qquad (3.61)$$

and

$$\mathbf{A} = \sum_i \mathbf{e}_i\mathbf{A}_{i\cdot}^T. \qquad (3.62)$$

There exist interesting relations involving the elementary matrices operating on the matrix **A**. For example, using (3.51) and (3.53), we have

$$\begin{aligned}
\mathbf{E}_{ij}\mathbf{A} &= \mathbf{e}_i\mathbf{e}_j^T\mathbf{A} \\
&= \mathbf{e}_i\mathbf{A}_{j\cdot}^T. \qquad (3.63)
\end{aligned}$$

Similarly, using (3.52), we have

$$\begin{aligned} \mathbf{A}\mathbf{E}_{ij} &= \mathbf{A}\mathbf{e}_i\mathbf{e}_j^T \\ &= \mathbf{A}_{\cdot i}\mathbf{e}_j^T \end{aligned} \tag{3.64}$$

so that

$$\mathbf{A}\mathbf{E}_{jj} = \mathbf{A}_{\cdot j}\mathbf{e}_j^T, \tag{3.65}$$

and by (3.63) and (3.64) we have

$$\begin{aligned} \mathbf{A}\mathbf{E}_{ij}\mathbf{B} &= \mathbf{A}\mathbf{e}_i\mathbf{e}_j^T\mathbf{B} \\ &= \mathbf{A}_{\cdot i}\mathbf{B}_{j\cdot}^T. \end{aligned} \tag{3.66}$$

Furthermore, by (3.38) and (3.55) we have

$$\begin{aligned} \mathbf{E}_{ij}\mathbf{A}\mathbf{E}_{rs} &= \mathbf{e}_i\mathbf{e}_j^T\mathbf{A}\mathbf{e}_r\mathbf{e}_s^T \\ &= \mathbf{e}_i a_{jr}\mathbf{e}_s^T \\ &= a_{jr}\mathbf{e}_i\mathbf{e}_s^T \\ &= a_{jr}\mathbf{E}_{is}. \end{aligned} \tag{3.67}$$

In particular,

$$\mathbf{E}_{jj}\mathbf{A}\mathbf{E}_{rr} = a_{jr}\mathbf{E}_{jr}. \tag{3.68}$$

3.3.3 The Trace Function

The trace of a square matrix $\mathbf{A}$ of order $n \times n$ is the sum of the diagonal terms:

$$\mathrm{tr}\,(\mathbf{A}) = \sum_i a_{ii}. \tag{3.69}$$

Applying Identity (3.55), we have

$$\mathrm{tr}\,(\mathbf{A}) = \sum_i \mathbf{e}_i^T \mathbf{A} \mathbf{e}_i. \tag{3.70}$$

From (3.52) and (3.70), we find

$$\mathrm{tr}\,(\mathbf{A}) = \sum_i \mathbf{e}_i^T \mathbf{A}_{\cdot i} \tag{3.71}$$

and from (3.53) and (3.70) we have

$$\mathrm{tr}\,(\mathbf{A}) = \sum_i \mathbf{A}_{i\cdot}^T \mathbf{e}_i. \tag{3.72}$$

We can obtain similar expression for the trace of a product $\mathbf{AB}$ of matrices $\mathbf{A}$ and $\mathbf{B}$. Starting with (3.61) and (3.65), we have

$$\begin{aligned} \mathbf{A} &= \sum_j \mathbf{A}_{\cdot j} \mathbf{e}_j^T \\ &= \sum_j \mathbf{A} \mathbf{E}_{jj}, \end{aligned} \tag{3.73}$$

and thus

$$\begin{aligned} \mathbf{AB} &= \sum_j \mathbf{A} \mathbf{E}_{jj} \mathbf{B} \\ &= \sum_j \left(\mathbf{A} \mathbf{e}_j\right) \left(\mathbf{e}_j^T \mathbf{B}\right). \end{aligned} \tag{3.74}$$

From (3.70) and (3.74) we have

$$\mathrm{tr}\,(\mathbf{AB}) = \sum_i \mathbf{e}_i^T \mathbf{AB} \mathbf{e}_i \tag{3.75}$$

$$\begin{aligned} &= \sum_i \mathbf{e}_i^T \sum_j \left(\mathbf{A} \mathbf{e}_j\right) \left(\mathbf{e}_j^T \mathbf{B}\right) \mathbf{e}_i \\ &= \sum_i \sum_j \left(\mathbf{e}_i^T \mathbf{A} \mathbf{e}_j\right) \left(\mathbf{e}_j^T \mathbf{B} \mathbf{e}_i\right) \\ &= \sum_i \sum_j a_{ij} b_{ji}. \end{aligned} \tag{3.76}$$

Similarly,

$$\begin{aligned} \mathrm{tr}\left(\mathbf{BA}\right) &= \sum_i \mathbf{e}_i^T \mathbf{BA} \mathbf{e}_i \\ &= \sum_i \mathbf{e}_i^T \sum_j \left(\mathbf{Be}_j\right)\left(\mathbf{e}_j^T \mathbf{A}\right) \mathbf{e}_i \\ &= \sum_i \sum_j \left(\mathbf{e}_i^T \mathbf{Be}_j\right)\left(\mathbf{e}_j^T \mathbf{Ae}_i\right) \\ &= \sum_i \sum_j b_{ij} a_{ji}. \end{aligned} \tag{3.77}$$

From (3.76) and (3.77) we find that

$$\mathrm{tr}\left(\mathbf{AB}\right) = \mathrm{tr}\left(\mathbf{BA}\right). \tag{3.78}$$

From (3.52), (3.53), and (3.75) we have

$$\mathrm{tr}\left(\mathbf{AB}\right) = \sum_i \mathbf{A}_{i\cdot}^T \mathbf{B}_{\cdot i}. \tag{3.79}$$

Also, from (3.78) and (3.79) we have

$$\mathrm{tr}\left(\mathbf{AB}\right) = \sum_i \mathbf{B}_{i\cdot}^T \mathbf{A}_{\cdot i}. \tag{3.80}$$

Similarly,

$$\mathrm{tr}\left(\mathbf{AB}^T\right) = \sum_i \mathbf{A}_{i\cdot}^T \mathbf{B}_{i\cdot}, \tag{3.81}$$

and since $\mathrm{tr}\left(\mathbf{AB}^T\right) = \mathrm{tr}\left(\mathbf{A}^T\mathbf{B}\right)$,

$$\mathrm{tr}\left(\mathbf{AB}^T\right) = \sum_i \mathbf{A}_{\cdot i}^T \mathbf{B}_{\cdot i}. \tag{3.82}$$

Two important properties of the trace are

$$\mathrm{tr}\left(\mathbf{A} + \mathbf{B}\right) = \mathrm{tr}\left(\mathbf{A}\right) + \mathrm{tr}\left(\mathbf{B}\right) \tag{3.83}$$

and

$$\mathrm{tr}\,(\alpha\mathbf{A}) = \alpha\mathrm{tr}\,(\mathbf{A}) \tag{3.84}$$

where α is a scalar.

This shows the linearity property of the trace function.

3.3.4 The Vec Operator

We shall make use of a vector valued function denoted by $\mathrm{vec}\,(\mathbf{A})$ of a matrix $\mathbf{A}$ defined by Neudecker [85].

If $\mathbf{A}$ is of order $m \times n$, then

$$\mathrm{vec}\,(\mathbf{A}) = \begin{bmatrix} \mathbf{A}_{\cdot 1} \\ \mathbf{A}_{\cdot 2} \\ \cdots \\ \mathbf{A}_{\cdot n} \end{bmatrix}. \tag{3.85}$$

From the definition (3.85) it is clear that $\mathrm{vec}\,(\mathbf{A})$ is a vector of order mn.

3.3.5 The Kronecker Product

Consider a matrix $\mathbf{A} = [a_{ij}]$ of order $m \times n$ and a matrix $\mathbf{B} = [b_{ij}]$ of order $r \times s$. The Kronecker product of the two matrices, denoted by $\mathbf{A} \otimes \mathbf{B}$ is defined as the partitioned matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \tag{3.86}$$

The product $\mathbf{A} \otimes \mathbf{B}$ is seen to be a matrix of order $mr \times ns$. It has mn submatrices, the (i, j)th block is the matrix $a_{ij}\mathbf{B}$ of order $r \times s$.

Notice that the Kronecker product is defined irrespective of the dimension of the matrices involved. From this point of view it is a more general concept than the regular matrix multiplication.

3.3.6 Properties and Rules of Kronecker Products

The Kronecker product has the usual properties of a product:

I If α is a scalar, then

$$\mathbf{A} \otimes (\alpha\mathbf{B}) = \alpha (\mathbf{A} \otimes \mathbf{B}) . \tag{3.87}$$

Proof

The result follows from

$$\begin{aligned} (i, j) \text{ th block of } \mathbf{A} \otimes (\alpha\mathbf{B}) &= [a_{ij} (\alpha\mathbf{B})] \\ &= \alpha [a_{ij}\mathbf{B}] \\ &= \alpha [(i, j) \text{ th block of } \mathbf{A} \otimes \mathbf{B}] \end{aligned}$$

II The product is distributive with respect to addition, that is

(a)

$$(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C} \tag{3.88}$$

(b)

$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C} \tag{3.89}$$

Proof

We will consider the proof of Item IIa only. The (i, j)th block of $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C}$ is

$$(a_{ij} + b_{ij})\,\mathbf{C}.$$

The (i, j)th block of $\mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$ is

$$a_{ij}\mathbf{C} + b_{ij}\mathbf{C} = (a_{ij} + b_{ij})\,\mathbf{C}.$$

Since the two blocks are equal for every (i, j), the result follows.

III The product is associative:

$$\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}. \tag{3.90}$$

IV There exists a zero element $\mathbf{0}_{mn}$ and an identity element $\mathbf{I}_{mn}$ such that

$$\begin{aligned} \mathbf{0}_{mn} &= \mathbf{0}_m \otimes \mathbf{0}_n \\ \mathbf{I}_{mn} &= \mathbf{I}_m \otimes I_n \end{aligned} \tag{3.91}$$

The identity matrices are all square, for example $\mathbf{I}_m$ is the identity matrix of order $m \times m$.

V

$$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T. \tag{3.92}$$

Proof

The (i, j)th block of $(\mathbf{A} \otimes \mathbf{B})^T$ is

$$a_{ji}\mathbf{B}^T.$$

VI The mixed product rule:

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} \tag{3.93}$$

provided that the dimensions of the matrices are such that the various expressions exist.

Proof

The (i, j)th block of the left hand side is obtained by taking the product of the ith row block of $(\mathbf{A} \otimes \mathbf{B})$ and the jth column block of $(\mathbf{C} \otimes \mathbf{D})$, this is of the following form

$$\begin{bmatrix} a_{i1}\mathbf{B} & a_{i2}\mathbf{B} & \cdots & a_{in}\mathbf{B} \end{bmatrix} \begin{bmatrix} c_{1j}\mathbf{D} \\ c_{2j}\mathbf{D} \\ \cdots \\ c_{nj}\mathbf{D} \end{bmatrix} = \sum_r a_{ir}c_{rj}\mathbf{BD}.$$

The (i, j)th block of the right hand side is (by definition of the Kronecker product)

$$g_{ij}\mathbf{BD}$$

where g_{ij} is the (i, j)th element of the matrix $\mathbf{AC}$. But by the rule of matrix multiplications

$$g_{ij} = \sum_r a_{ir}c_{rj}.$$

Since the (i, j)th blocks are equal, the result follows.

VII Given a matrix $\mathbf{A}$ of order $m \times m$ and a matrix $\mathbf{B}$ of order $n \times n$ and subject to the existence of the various inverses,

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}. \tag{3.94}$$

Proof

Using (3.93), we have

$$\begin{aligned}(\mathbf{A} \otimes \mathbf{B})\left(\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}\right) &= \mathbf{A}\mathbf{A}^{-1} \otimes \mathbf{B}\mathbf{B}^{-}1 \\ &= \mathbf{I}_m \otimes \mathbf{I}_n \\ &= \mathbf{I}_{mn}.\end{aligned}$$

The result follows.

VIII

$$\operatorname{vec}(\mathbf{AYB}) = \left(\mathbf{B}^T \otimes \mathbf{A}\right) \operatorname{vec}(\mathbf{Y}) \tag{3.95}$$

Proof

We prove (3.95) for $\mathbf{A}$, $\mathbf{Y}$, and $\mathbf{B}$ each of order $n \times n$. The result is true for any $\mathbf{A}$ of order $m \times n$, $\mathbf{Y}$ of order $n \times r$, and $\mathbf{B}$ of order $r \times s$. Using (3.74), we have

$$\mathbf{AYB} = \sum_j (\mathbf{AYe}_j)\left(\mathbf{e}_j^T\mathbf{B}\right).$$

From (3.52), we have

$$\begin{aligned}(\mathbf{AYB})_{\cdot k} &= \mathbf{AYBe}_k \\ &= \sum_j (\mathbf{AYe}_j)\left(\mathbf{e}_j^T\mathbf{B}\right)\mathbf{e}_k \\ &= \sum_j (\mathbf{AYe}_j)\left(\mathbf{e}_j^T\mathbf{Be}_k\right).\end{aligned}$$

From (3.55), we have

$$\begin{aligned}(\mathbf{AYB})_{\cdot k} &= \sum_j (\mathbf{AYe}_j)\left(\mathbf{e}_j^T\mathbf{Be}_k\right) \\ &= \sum_j (\mathbf{AYe}_j)\, b_{jk} \\ &= \sum_j b_{jk}\,(\mathbf{AYe}_j)\,.\end{aligned}$$

From (3.52), we have

$$\begin{aligned}(\mathbf{AYB})_{\cdot k} &= \sum_j b_{jk}\,(\mathbf{AYe}_j) \\ &= \sum_j b_{jk}\mathbf{AY}_{\cdot j} \\ &= \sum_j (b_{jk}\mathbf{A})\,\mathbf{Y}_{\cdot j} \\ &= \begin{bmatrix} b_{1k}\mathbf{A} & b_{2k}\mathbf{A} & \cdots & b_{nk}\mathbf{A} \end{bmatrix} \begin{Bmatrix} \mathbf{Y}_{\cdot 1} \\ \mathbf{Y}_{\cdot 2} \\ \vdots \\ \mathbf{Y}_{\cdot n} \end{Bmatrix} \\ &= \left[\mathbf{B}_{\cdot k}^T \otimes \mathbf{A}\right] \mathrm{vec}\,(\mathbf{Y}) \\ &= \left[\left(\mathbf{B}^T\right)_{k\cdot}^T \otimes \mathbf{A}\right] \mathrm{vec}\,(\mathbf{Y})\end{aligned}$$

since the transpose of the kth column of $\mathbf{B}$ is the kth row of $\mathbf{B}^T$; the result follows.

IX If $\{\lambda_i\}$ and $\{\mathbf{x}_i\}$ are the eigenvalues and the corresponding eigenvectors of matrix $\mathbf{A}$ and $\{\mu_j\}$ and $\{\mathbf{y}_j\}$ are the eigenvalues and the corresponding eigenvectors of matrix $\mathbf{B}$, then $\mathbf{A} \otimes \mathbf{B}$ has eigenvalues $\{\lambda_i\mu_j\}$ and corresponding eigenvectors $\{\mathbf{x}_i \otimes \mathbf{y}_j\}$.

Proof

By (3.93), we have

$$\begin{aligned}(\mathbf{A}\otimes\mathbf{B})\left(\mathbf{x}_i\otimes\mathbf{y}_j\right) &= \left(\mathbf{A}\mathbf{x}_i\right)\otimes\left(\mathbf{B}\mathbf{y}_j\right)\\ &= \left(\lambda_i\mathbf{x}_i\right)\otimes\left(\mu_j\mathbf{y}_j\right)\\ &= \lambda_i\mu_j\left(\mathbf{x}_i\otimes\mathbf{y}_j\right).\end{aligned}$$

The result follows.

X Given the two matrices **A** and **B** or order $n\times n$ and $m\times m$ respectively, we have

$$\det\left(\mathbf{A}\otimes\mathbf{B}\right)=\left(\det\left(\mathbf{A}\right)\right)^m\left(\det\left(\mathbf{B}\right)\right)^n \tag{3.96}$$

Proof

Assuming that $\{\lambda_i\}_{i=1}^n$ and $\{\mu_j\}_{j=1}^m$ are the eigenvalues of **A** and **B** respectively. The proof relies on the fact that the determinant of a matrix is equal to the product of its eigenvalues [86]. Hence, from IX, we have

$$\begin{aligned}\det\left(\mathbf{A}\otimes\mathbf{B}\right) &= \prod_{i=1}^{n}\prod_{j=1}^{m}\lambda_i\mu_j\\ &= \left(\prod_{j=1}^{m}\lambda_1\mu_j\right)\left(\prod_{j=1}^{m}\lambda_2\mu_j\right)\cdots\left(\prod_{j=1}^{m}\lambda_n\mu_j\right)\\ &= \left(\lambda_1^m\prod_{j=1}^{m}\mu_j\right)\left(\lambda_2^m\prod_{j=1}^{m}\mu_j\right)\cdots\left(\lambda_n^m\prod_{j=1}^{m}\mu_j\right)\\ &= \left(\prod_{i=1}^{n}\lambda_i^m\right)\left(\prod_{j=1}^{m}\mu_j^n\right)\\ &= \left(\prod_{i=1}^{n}\lambda_i\right)^m\left(\prod_{j=1}^{m}\mu_j\right)^n\\ &= \left(\det\left(\mathbf{A}\right)\right)^m\left(\det\left(\mathbf{B}\right)\right)^n.\end{aligned}$$

XI If f is an analytic function, $\mathbf{A}$ is a matrix of order $n \times n$ and $f(\mathbf{A})$ exists, then

$$f(\mathbf{I}_m \otimes \mathbf{A}) = \mathbf{I}_m \otimes f(\mathbf{A}) \tag{3.97}$$

and

$$f(\mathbf{A} \otimes \mathbf{I}_m) = f(\mathbf{A}) \otimes \mathbf{I}_m \tag{3.98}$$

Proof

Since f is an analytic function it can be expressed as a power series such as

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots$$

so that

$$\begin{aligned} f(\mathbf{A}) &= a_0 \mathbf{I}_n + a_1 \mathbf{A} + a_2 \mathbf{A}^2 + \cdots \\ &= \sum_k a_k \mathbf{A}^k, \end{aligned}$$

where $\mathbf{A}^0 = \mathbf{I}_n$.

By the Cayley Hamilton theorem (see [86]), the right hand side of the equation for $f(\mathbf{A})$ is the sum of at most $n+1$ matrices.

By (3.87), (3.89), and (3.93), we now have

$$\begin{aligned} f(\mathbf{I}_m \otimes \mathbf{A}) &= \sum_k a_k (\mathbf{I}_m \otimes \mathbf{A})^k \\ &= \sum_k a_k \left(\mathbf{I}_m \otimes \mathbf{A}^k\right) \\ &= \sum_k \left(\mathbf{I}_m \otimes a_k \mathbf{A}^k\right) \\ &= \mathbf{I}_m \otimes \sum_k a_k \mathbf{A}^k \\ &= \mathbf{I}_m \otimes f(\mathbf{A}). \end{aligned}$$

We can also write

$$\begin{aligned} f(\mathbf{A} \otimes \mathbf{I}_m) &= \sum_k a_k (\mathbf{A} \otimes \mathbf{I}_m)^k \\ &= \sum_k a_k \left(\mathbf{A}^k \otimes \mathbf{I}_m\right) \\ &= \sum_k \left(a_k \mathbf{A}^k \otimes \mathbf{I}_m\right) \\ &= \left(\sum_k a_k \mathbf{A}^k\right) \otimes \mathbf{I}_m \\ &= f(\mathbf{A}) \otimes \mathbf{I}_m. \end{aligned}$$

XII We have

$$\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{tr}(\mathbf{A}) \operatorname{tr}(\mathbf{B}). \tag{3.99}$$

Proof Assuming that $\mathbf{A}$ is of order $n \times n$, we have

$$\begin{aligned} \operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) &= \operatorname{tr}(a_{11}\mathbf{B}) + \operatorname{tr}(a_{22}\mathbf{B}) + \cdots + \operatorname{tr}(a_{nn}\mathbf{B}) \\ &= a_{11}\operatorname{tr}(\mathbf{B}) + a_{22}\operatorname{tr}(\mathbf{B}) + \cdots + a_{nn}\operatorname{tr}(\mathbf{B}) \\ &= \left(\sum_i a_{ii}\right) \mathbf{B} \\ &= \operatorname{tr}(\mathbf{A}) \operatorname{tr}(\mathbf{B}). \end{aligned}$$

3.3.7 The Permutation Matrix Associating $\operatorname{vec}(\mathbf{X})$ and $\operatorname{vec}(\mathbf{X}^T)$

If $\mathbf{X}$ is a matrix of order $m \times n$, using (3.56) we have

$$\mathbf{X} = \sum_i \sum_j x_{ij} \mathbf{E}_{ij}$$

where $\mathbf{E}_{ij}$ is the elementary matrix of order $m \times n$. It follows that

$$\mathbf{X}^T = \sum_i \sum_j x_{ij} \mathbf{E}_{ij}^T$$

so that

$$\text{vec}\left(\mathbf{X}^T\right) = \sum_i \sum_j x_{ij} \text{vec}\left(\mathbf{E}_{ij}^T\right). \tag{3.100}$$

We can write (3.100) as

$$\text{vec}\left(\mathbf{X}^T\right) = \left[\begin{array}{cccc} \text{vec}\left(\mathbf{E}_{11}^T\right) & \text{vec}\left(\mathbf{E}_{21}^T\right) & \cdots & \text{vec}\left(\mathbf{E}_{mn}^T\right) \end{array}\right] \left\{\begin{array}{c} x_{11} \\ x_{21} \\ \cdots \\ x_{mn} \end{array}\right\},$$

that is

$$\text{vec}\left(\mathbf{X}^T\right) = \left[\begin{array}{cccc} \text{vec}\left(\mathbf{E}_{11}^T\right) & \text{vec}\left(\mathbf{E}_{21}^T\right) & \cdots & \text{vec}\left(\mathbf{E}_{mn}^T\right) \end{array}\right] \text{vec}\left(\mathbf{X}\right).$$

So the permutation matrix associating $vec\mathbf{X}$ and $\text{vec}\left(\mathbf{X}^T\right)$ is

$$\mathbf{U} = \left[\begin{array}{cccc} \text{vec}\left(\mathbf{E}_{11}^T\right) & \text{vec}\left(\mathbf{E}_{21}^T\right) & \cdots & \text{vec}\left(\mathbf{E}_{mn}^T\right) \end{array}\right]. \tag{3.101}$$

The transformation matrix $\mathbf{U}$ goes by multiple names, such as the vec-permutation matrix [87], the commutation matrix [85], or the tensor commutator [88]. This transformation matrix will be shown to be used to enforce symmetry condition upon the identified system matrices.

Chapter 4

Proper Orthogonal Decomposition

In many applications, the POD method is used to analyze experimental data with the aim to extract dominant features and trends, known as coherent structures. In this investigation, POD will provide a suitable set of basis functions by identifying a low-dimensional subspace in which the solution lies. The system of governing equations of a dynamical system is then projected onto this subspace to arrive at a low-dimensional model of the system. In the following sections, the theoretical framework of POD is presented and some extensions relevant to the present context of constructing low-dimensional models are then discussed. POD is an important tool in this context, since it generates an optimal sequence of finite-dimensional subspaces where the dominant dynamics of the system lie.

It is essential to note that POD is based on the premise that a given spatio-temporal process can be represented by a linear combination of POMs as basis functions. The assumption of linearity leads to the attractiveness as well as limitations of the POD method. Applying linear operator theory, a proof of the properties of the approximated system response via POD can be easily obtained. However, in proving optimality, for example, it is

important to keep in mind that optimality is implied only with respect to other linear representations of the output process. A linear representation herein relates to approximating the system output with a sum of basis functions scaled by appropriate coefficients. A well-known linear representation is the Fourier series. POD is an output-based representation of the system response. The output used to generate the POMs may emerge from either a linear or non-linear system. However, the POD method is equally applicable irrespective of the system linearity or nonlinearity.

The chapter will begin by introducing the POD in the framework of scalar fields. The subsequent section characterizes the properties of representations using POD basis functions, namely its reconstructive and optimality properties. The discrete formulation of POD is then presented. We also discuss the methodology used in determining the size of the optimal subspace necessary for POD. The chapter concludes with the specific example of obtaining a POD reduced-order linear dynamic model from a high-dimensional model.

4.1 Introduction

In this section, we present a brief theoretical framework of POD. Assuming an experiment provides an ensemble $\{u_t\}_{t=1}^{T}$ of real-valued spatio-temporal scalar fields, each a function $u_t = u_t(x)$ of a geometrical variable x defined on the domain $[0, 1]$. The aim of POD is to obtain an optimal representation of members of u_t, by a suitable set of basis functions φ_j. The subsequent remaining task is to determine the magnitude of the projections of u_t on each of these basis functions φ_j. We therefore assume that the u_t belong to an inner product space: the linear, infinite-dimensional Hilbert space $L^2([0, 1])$, of square integrable

functions with inner product defined as

$$(u_t, \varphi_j) = \int_0^1 u_t(x)\,\varphi_j(x)\,dx \tag{4.1}$$

The objective now is to determine the basis $\{\varphi_j\}_{j=1}^{\infty}$ for L^2 that is optimal for the set of data obtained. In other words, the finite-dimensional representation

$$u_t(x) = \sum_{j=1}^{m} a_j(t)\,\varphi_j(x) \qquad t = 1, \ldots, T \tag{4.2}$$

captures most of the energy of a characteristic member of the ensemble better than representations of the same dimension m using any other basis. The concept of a "characteristic" member implies the use of an averaging operator, which is denoted by $\langle\cdot\rangle$. The averaging operator is assumed to commute with the integral (4.1) of the L^2 inner product. In the next section, averaging is discussed in more detail. It is sufficient for now to consider the averaging operation as a time average over the ensemble with members $u_t(x) = u(x,t)$ obtained from consecutive measurements during a single experimental run.

Optimality is achieved by choosing φ so as to maximise the normalized ensemble average of the projection of u onto φ given by

$$\max_{\varphi \in L^2([0,1])} \frac{\left\langle |(u,\varphi)|^2 \right\rangle}{\|\varphi\|^2}, \tag{4.3}$$

where $|\cdot|$ represents the modulus and $\|\cdot\|$ the L^2-norm given by

$$\|\varphi\| = (\varphi, \varphi)^{1/2}. \tag{4.4}$$

As is stated, the solution of (4.3) would result in the best approximation to the ensemble members u_t by a single basis function φ. The optimization problem (4.3) actually has many

critical points, each representing a basis function. The critical points are all physically significant, and the set of critical points, taken together, provide the desired optimal basis.

The optimality condition (4.3) can be rewritten using the theory of the calculus of variations: the objective is to maximize $\left\langle |(u,\varphi)|^2 \right\rangle$ subject to the constraint $\|\varphi\|^2 = 1$ [89]. This constrained variational problem has the following functional

$$J[\varphi] = \left\langle |(u,\varphi)|^2 \right\rangle - \lambda \left(\|\varphi\|^2 - 1 \right), \tag{4.5}$$

and the extrema of this functional occurs at the point at which the functional derivative vanishes for all variations

$$\frac{\partial}{\partial \delta} J[\varphi + \delta\psi]\Big|_{\delta=0} = 0, \tag{4.6}$$

with the constraint $(\varphi + \delta\psi) \in L^2([0,1])$ and $\delta \in \mathbb{R}$.

Applying Identity (4.5), (4.6) becomes

$$\begin{aligned}
&\frac{\partial}{\partial\delta}J\left[\varphi+\delta\psi\right]\Big|_{\delta=0} \\
&= \frac{\partial}{\partial\delta}\left[\left\langle\left(u,\varphi+\delta\psi\right)\left(u,\varphi+\delta\psi\right)\right\rangle-\lambda\left(\varphi+\delta\psi,\varphi+\delta\psi\right)+\lambda\right]\Big|_{\delta=0} \\
&= \frac{\partial}{\partial\delta}\left[\left\langle\left[\int_0^1 u\left(\varphi+\delta\psi\right)dx\right]^2\right\rangle-\lambda\left[\int_0^1\left(\varphi+\delta\psi\right)^2dx\right]+\lambda\right]\Big|_{\delta=0} \\
&= \left\langle 2\int_0^1 u\left(\varphi+\delta\psi\right)\left(\frac{\partial}{\partial\delta}\int_0^1 u\left(\varphi+\delta\psi\right)dx\right)dx\right\rangle\Big|_{\delta=0} \\
&\quad -\lambda\left[\int_0^1 2\left(\varphi+\delta\psi\right)\left(\frac{\partial}{\partial\delta}\left(\varphi+\delta\psi\right)\right)dx\right]\Big|_{\delta=0} \\
&= \left\langle 2\int_0^1 u\left(\varphi+\delta\psi\right)dx\int_0^1 u\psi dx\right\rangle\Big|_{\delta=0} \\
&\quad -\lambda\left[\int_0^1 2\left(\varphi+\delta\psi\right)\psi dx\right]\Big|_{\delta=0} \\
&= 2\left\langle\int_0^1 u\varphi dx\int_0^1 u\psi dx\right\rangle-2\lambda\int_0^1\varphi\psi dx \\
&= 2\left[\left\langle\left(u,\varphi\right)\left(u,\psi\right)\right\rangle-\lambda\left(\varphi,\psi\right)\right] \\
&= 0. \qquad (4.7)
\end{aligned}$$

The term in brackets in (4.7) may be written as

$$\begin{aligned}
&\left\langle\left(u,\varphi\right)\left(u,\psi\right)\right\rangle-\lambda\left(\varphi,\psi\right) \\
&= \left\langle\int_0^1 u\left(x'\right)\varphi\left(x'\right)dx\int_0^1 u\left(x\right)\psi\left(x\right)dx'\right\rangle-\lambda\int_0^1\varphi\left(x\right)\psi\left(x\right)dx \\
&= \left\langle\int_0^1\int_0^1 u\left(x\right)\varphi\left(x'\right)u\left(x'\right)\psi\left(x\right)dx'dx\right\rangle-\lambda\int_0^1\varphi\left(x\right)\psi\left(x\right)dx \\
&= \int_0^1\int_0^1\left\langle u\left(x\right)u\left(x'\right)\right\rangle\varphi\left(x'\right)\psi\left(x\right)dx'dx-\lambda\int_0^1\varphi\left(x\right)\psi\left(x\right)dx \\
&= \int_0^1\left[\int_0^1\left\langle u\left(x\right)u\left(x'\right)\right\rangle\varphi\left(x'\right)dx'-\lambda\varphi\left(x\right)\right]\psi\left(x\right)dx \\
&= 0, \qquad (4.8)
\end{aligned}$$

where the integrals have been rearranged and brought the average inside applying the commutativity of $(\cdot)$ and $\int \cdot dx$. Finally, due to $\psi(x)$ being an arbitrary variation, the optimality condition (4.6) is equivalent to

$$\int_0^1 \langle u(x)\, u(x') \rangle \varphi(x')\, dx' = \lambda \varphi(x). \tag{4.9}$$

Thus, the optimal basis is the set of eigenfunctions φ_j of Equation (4.9) whose kernel is the time-averaged autocorrelation function $\langle u(x)\, u(x') \rangle = R(x, x')$. The set of eigenfunctions are thus sometimes referred to as the empirical eigenfunctions.

There is an attractive geometrical interpretation of the above representation, especially in the case of multi-dimensional observations $\mathbf{u}_t$, in which $\mathbf{u}_t$ are n-dimensional vectors, instead of functions, and the autocorrelation function is substituted by the $n \times n$ outer product matrix $\mathbf{R} = \langle \mathbf{u}\mathbf{u}^T \rangle$ referred to by the cross-correlation matrix. The eigenvectors of the resulting eigenvalue problem are then simply the fundamental axes of the cluster of observed data points $\mathbf{u}_t$ in the n-dimensional vector space.

Consequently, the operator in Equation (4.9) is denoted by $\mathcal{R}\varphi = \int R(x, x')\, \varphi dx$. $\mathcal{R}$ is a compact self-adjoint operator, and thus spectral theory [90] insures that the extremum in (4.3) exists and is equal to the eigenfunction corresponding to the largest eigenvalue of the integral equation (4.9), which can be rewritten in operatot form as

$$\mathcal{R}\varphi = \lambda \varphi. \tag{4.10}$$

Furthermore, Hilbert-Schmidt theory guarantees the existence of a continuous set of eigenvalues and their corresponding eigenfunctions, and thus the averaged autocorrelation function has a diagonal decomposition:

$$R(x, x') = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x)\, \varphi_j(x'), \tag{4.11}$$

Hilbert-Schmidt theory also assures us that the eigenfunctions φ_j are mutually orthogonal in L^2. This is a stronger result than the variational argument, since we are guaranteed a maximum for (4.3) instead of just a critical point.

The eigenvalues can be rearranged such that $\lambda_j \geq \lambda_{j+1}$. Additionally, the averaged autocorrelation function $R(x,x') = \langle u(x)\, u(x') \rangle$ is non-negative definite indicating that the integral operator $\mathcal{R}$ is also non-negative definite. This ensures that $\lambda_j \geq 0$ for all j. As will be see in Section 4.3, nearly every member (in a probabilistic sense) of the ensemble used in the averaging $\langle \cdot \rangle$ resulting in $R(x,x')$ can be represented by a spectral decomposition using the eigenfunctions $\{\varphi_j\}_{j=1}^{\infty}$ as in

$$u(x,t) = \sum_{j=1}^{\infty} a_j(t)\, \varphi_j(x). \tag{4.12}$$

The POD relates to Equation (4.12). Also, the diagonal representation (4.11) of the correlation tensor implies that

$$\langle a_j(t)\, a_k(t) \rangle = \delta_{jk} \lambda_j, \tag{4.13}$$

so that the coefficients $a_j(t)$ of the representation are uncorelated. In (4.13), δ_{jk} denotes the Kronecker delta given by (refer to Section 3.3.1)

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}. \tag{4.14}$$

If $u(x)$ is a system response (displacement or velocity), then the eigenvalues λ_j represent twice the average contributed by each mode φ_j. Thus, by choosing the subspace spanned by the modes $\{\varphi_j\}_{j=1}^{m}$ corresponding to the n-largest eigenvalues, the representation (4.2) contains the most energetic disturbances in the field, as claimed at the start. The λ_j are called empirical eigenvalues.

So far, only functions defined on a bounded interval were considered. This is applicable in the context of structural dynamics. The unbounded case, more natural in the contex of fluid flows for example, can be dealt with provided that inner product, now an infinite integral, is well defined, and that the space of functions to be decomposed still has a countable basis. In dealing with unbounded domains in practice, either a finite subdomain is selected and Sommerfield boundary conditions are applied at far field [91], or POD is restricted to functions rapidly decaying to zero outside a finite domain.

The derivation of the integral equation (4.9) also generalizes to functions of more than one spatial variable and to vector-valued fields such as the three-dimensional velocity fields $\mathbf{u}(\mathbf{x}, t)$ of turbulence, simply by using the appropriate Hilbert spaces and inner products [53]. For vector-valued functions, $R(x, x')$ is replaced by an autocorrelation tensor $\mathbf{R}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{u}(\mathbf{x}) \otimes \mathbf{u}(\mathbf{x}) \rangle$ and the eigenfunctions are vector valued.

The non-negative definiteness of $R(x, x')$ suggest that the empirical eigenvalues λ_j are non-negative themselves, but in general some eigenvalues might be zero. To produce a complete basis for L^2, we must include all those "additional" eigenfunctions φ_j with zero eigenvalues, although, in view of (4.13), they carry no information on the original data. It is therefore not necessary to examine those eigenfunctions with zero eigenvalues.

4.2 On Domains and Averaging

In the introduction, the focus was on field functions of a single spatial variable x. In the case of 3-dimensional structural dynamics, the fields depend on four variables, three spatial and one temporal. There is no reason to distinguish between space and time variables as the

multi-dimensional theory does not enforce such a distinction: for experiments performed in the one-dimensional spatial domain $0 \leq x \leq 1$ over times of duration T, one simply measures correlations with time lags as well as spatial separations limited in the space of $L^2([0,1] \times [0,T])$. Thus the inner, in this case, is a double integral over x and t. However, in the context of the intended use of POD to derive low dimensional models, only spatial basis functions are sought. The time-dependent modal coefficients $a_j(t)$ appear in POD representations as

$$u(x,t) = \sum_j a_j(t)\,\varphi_j(x)\,. \tag{4.15}$$

Their multi-dimensional analogues, will be determined subsequently via projection of the governing equations.

The nature of the physical problem dictates whether one should treat the data as stationary or non-stationary in time. If one seeks purely spatial representations of a space-time field $u(x,t)$ in a statistically stationary problem, the correlation functions between pairs of points in the physical space must be measured with no time difference. Assuming ergodicity, time is used to increase the ensemble size by including measurements taken at appropriately separated time slices during a single experimental run, in which case the ensemble members may be defined as $u(x,t_k)$. In that case, $\langle\cdot\rangle$ effectively becomes a time average. In the non-stationary case, one could still derive a purely spatial representation from the zero time lag correlation averaged over ensembles taken from different experiments. In the proposed methodology, stationarity in time is assumed.

In the case of structural dynamics applications involving a finite domain in physical space with prescribed Neumann and/or Dirichlet boundary conditions, examination of the mathematical details associated with the averaging operator is usually unnecessary. How-

ever, it is important to note that, when dealing with an infinite domain, difficulties may arise. The appropriate function space for time-stationary problems is $L^2(\Omega)$ where Ω is the (three-dimensional) spatial domain of the experiment, analogous to the interval $[0,1]$ in Section 4.1. The appropriate function space for time-dependent problems, is $L^2_{loc}(\Omega \times [0,\infty))$, where the subscript loc implies that the L^2-norm is finite on finite closed intervals in time. For certain problems, a finite L^2-norm is a reasonable assumption since it corresponds to finite energy.

4.3 Properties of POD

The following subsections describe properties that will be particularly important in our use of the POD to derive low-dimensional models. We first characterize the classes of functions that can be represented by empirical bases and then explain precisely how such representations preserve properties of the observations from which they are derived and how they are optimal. We then show how the rate of decay of the empirical eigenvalues determines geometrical properties of the attractor and how theoretical results on the regularity of solutions of the governing evolution equations are related to this.

4.3.1 Span of the Empirical Basis

The first step in understanding what can be done with representations using empirical eigenfunctions is characterizing the class of functions which can be accurately represented by the "relevant" elements of the basis: those containing spatial structures having finite energy on average. This is the set $S = \{\sum a_j\varphi_j | \sum |a_j|^2 < \infty, \lambda_j > 0\}$, or $\text{span}\{\varphi_j | j =$

$1, \ldots, \infty, \lambda_j > 0\}$. In this section, two functions u and v are equal in this sense if

$$\int_{\Omega} |u - v|^2 \, dx = 0. \tag{4.16}$$

Equation (4.16) is the mathematical definition of "accurately." We shall also frequently use a second notion of *almost every member of an ensemble* $\langle \cdot \rangle$. This is denoted by "a.e." In applications this average will typically be a finite sum over a set of realizations or an integral over a finite-time experimental run, but the theory is developed in the ideal case of infinite data sets.

A standing assumption in this section is that the averaged autocorrelation $R(x, x')$ is a continuous function. Discontinuities in R can lead to negative values in the power spectrum [53], the Fourier transform of R, and negative energies are unreasonable on physical grounds.

We first show that the empirical basis can reconstruct any function that is indistinguishable in the sense of (4.16) from a member of the original ensemble $\{u_t\}$. Let $u \in L^2(\Omega)$ be any such function and $\{\varphi_j\}$ be the orthonormal sequence of empirical eigenfunctions. The reconstruction of u will be a function $u_s(x) = \sum_j (u, \varphi_j) \varphi_j(x)$, belonging to S. We need to show that for a.e. u with respect to the ensemble average, we have $u = u_s$, that is

$$\left\langle \|u - u_s\|^2 \right\rangle. \tag{4.17}$$

We have

$$\begin{aligned} \left\langle \|u(x) - u_s(x)\|^2 \right\rangle &= \langle (u - u_s, u - u_s) \rangle \\ &= \langle (u, u) - 2(u, u_s) + (u_s, u_s) \rangle \end{aligned} \tag{4.18}$$

Since the functions u are the members of the original ensemble, the first term of (4.18)

is equivalent to

$$\begin{aligned} \langle (u,u) \rangle &= \left\langle \int_{\Omega} u(x)\, u(x)\, dx \right\rangle \\ &= \int_{\Omega} r(x,x)\, dx. \end{aligned} \tag{4.19}$$

The second term of (4.18) is equivalent to

$$\begin{aligned} \langle -2(u,u_s) \rangle &= -2 \left\langle \int_{\Omega} u(x) \left[\sum_j (u,\varphi_j)\, \varphi_j(x) \right] dx \right\rangle \\ &= -2 \left\langle \int_{\Omega} u(x) \sum_j \left[\int_{\Omega} u(x')\, \varphi_j(x')\, dx' \right] \varphi_j(x)\, dx \right\rangle \\ &= -2 \int_{\Omega} \sum_j \left[\int_{\Omega} \langle u(x)\, u(x') \rangle\, \varphi_j(x')\, dx' \right] \varphi_j(x)\, dx \\ &= -2 \int_{\Omega} \left(\sum_j \mathcal{R}\varphi_j \right) \varphi_j(x)\, dx \\ &= -2 \int_{\Omega} \sum_j \lambda_j \varphi_j(x)\, \varphi_j(x)\, dx \\ &= -2 \int_{\Omega} R(x,x)\, dx. \end{aligned} \tag{4.20}$$

Finally, the third term of (4.18) is equivalent to

$$\begin{aligned}
\langle (u_s, u_s) \rangle &= \left\langle \int_\Omega \left[\sum_i (u, \varphi_i)\, \varphi_i(x) \right] \left[\sum_j (u, \varphi_j)\, \varphi_j(x) \right] dx \right\rangle \\
&= \left\langle \sum_{i,j} (u, \varphi_i)\, (u, \varphi_j) \int_\Omega \varphi_i(x)\, \varphi_j(x)\, dx \right\rangle \\
&= \left\langle \sum_{i,j} (u, \varphi_i)\, (u, \varphi_j)\, \delta_{ij} \right\rangle \\
&= \left\langle \sum_j (u, \varphi_j)\, (u, \varphi_j) \right\rangle \\
&= \left\langle \sum_j \int_\Omega u(x)\, \varphi_j(x)\, dx \int_\Omega u(x')\, \varphi_j(x')\, dx' \right\rangle \\
&= \sum_j \int_\Omega \left[\int_\Omega \langle u(x)\, u(x') \rangle\, \varphi_j(x')\, dx' \right] \varphi_j(x)\, dx \\
&= \sum_j \int_\Omega \lambda_j \varphi_j(x)\, \varphi_j(x)\, dx
\end{aligned} \tag{4.21}$$

Using the continuity of R, we can apply Mercer's theorem for the uniform convergence of the series expression for R and interchange summation and integration, obtaining

$$\begin{aligned}
\langle (u_s, u_s) \rangle &= \int_\Omega \sum_j \lambda_j \varphi_j(x)\, \varphi_j(x)\, dx \\
&= \int_\Omega R(x, x)\, dx.
\end{aligned} \tag{4.22}$$

Combining (4.19), (4.20), and (4.22) we obtain (4.17).

We have shown that almost every member of the original ensemble can be reconstructed as a linear combination of empirical eigenfunctions having strictly positive eigenvalues. Now we want to show the converse: that each such eigenfunction can be expressed as a linear combination of observations. This will imply that any property of the ensemble members that is preserved under linear combination is inherited by the empirical basis

functions and hence by elements of S. basis functions and hence by elements of S.

Let X denote the set of functions (of full measure with respect to the averaging operation) for which reconstructions satisfying (4.17) are possible, and let θ be any function in S. We claim that there is a sequence $\{b_j\}_{j=1}^{\infty}$ of real numbers and a set of functions $u_j(x) \in X$ for $j = 1, \ldots, \infty$ such that

$$\theta(x) = \sum_{j=1}^{\infty} b_j u_j(x). \tag{4.23}$$

It immediately follows from (4.23) that, if $\mathcal{P}$ is a closed linear property of a subset of functions in $L^2(\Omega)$ and all the ensemble members u_k share that property, then the POD eigenfunctions also share the property. The converse holds too. Equation (4.23) and this remark characterize the "empirical subspace" S.

It remains to justify Equation (4.23). Let S' denote the set of all functions in $L^2(\Omega)$ with representations $\sum_i b_i u_i(x)$ with $u_i \in X$. We will show that $S'^{\perp} = S^{\perp}$, from which it follows that $S' = S$, and so the equation indeed holds.

Now $S^{\perp}$ is exactly the set of functions θ such that $(\theta, \varphi_i) = 0$ for every φ_i with eigenvalue $\lambda_i > 0$. From the first result of this section we have $u(x) = \sum_i b_i \varphi_i(x)$ where $\lambda_i > 0$, for a.e. (almost every) u. Thus we have $(\theta, u) = 0$ a.e. and so $(\theta, \sum_i b_i \varphi_i(x)) = 0$. This shows that $S^{\perp} \subset S'^{\perp}$.

To show $S'^{\perp} \subset S^{\perp}$ and hence, along with $S^{\perp} \subset S'^{\perp}$ already proved, conclude $S'^{\perp} = S^{\perp}$, assume that $(\theta, u) = \int_{\Omega} \theta(x') u(x') dx' = 0$ for a.e. u. Therefore for a.e. u we have $u(x) \int_{\Omega} u(x') \theta(x') dx' = 0$, and taking the average we get

$$\int_{\Omega} \langle u(x) u(x') \rangle \theta(x') dx' = 0, \tag{4.24}$$

which, from the eigenvalue equation (4.9), implies that $(\theta, \varphi_i) = 0$ for every i such that $\lambda_i > 0$.

The classic example of a property which passes from the data ensemble to the empirical basis is the satisfaction of the linear boundary conditions. This will be very useful when we project the system of equations describing the forced vibration of a linear system onto a subspace spanned by a collection of these eigenfunctions.

We now have characterization of the span of the eigenfunctions with strictly positive eigenvalues. This linear space S exactly coincides with that spanned by all realizations $u_t(x)$ of the original ensemble a.e. with respect to the measure induced by the averaging operation. For the case of having zero eigenvalues, we see that the set of empirical eigenfunctions $\{\varphi_j | \lambda_j > 0\}$ need not form a complete basis for $L^2(\Omega)$. While S may be infinite-dimensional, it is generally only a subset of the space $L^2(\Omega)$ in which we are working. It is complete only if one includes the kernel the operator $\mathcal{R}$, i.e. all the generalized) eigenfunctions with zero eigenvalues, but in doing so one loses the major advantage of POD, for in many applications one can argue on physical grounds that the realization $u(x,t)$ do not and should not span $L^2(\Omega)$. In such cases the discussion of this section highlights a strong property of the POD. Its use limits the space studied to the smallest linear subspace that is sufficient to describe the observed phenomena.

4.3.2 Optimality

A linear decomposition of a time-dependent, statistically stationary signal $u(x,t)$ with respect to any orthonormal basis $\{\varphi_j(x)\}_{j=l}^{\infty}$ is given by

$$u(x,t) = \sum_j b_j(t)\,\varphi_j(x) \tag{4.25}$$

If the $\varphi_j(x)$ are dimensionless, then the coefficients $b_j(t)$ carry the dimension of the quantity u. If $u(x,t)$ is a response quantity (i.e. displacement or velocity) in the conext of a linear dynamical system, for example, and $\langle\cdot\rangle$ is a time average, the average energy of the signal is given by

$$\begin{aligned}
&\frac{1}{2}\left\langle \int_\Omega u(x,t)\,u(x,t)\,dx \right\rangle \\
&= \frac{1}{2}\left\langle \int_\Omega \sum_i b_i(t)\,\varphi_i(x) \sum_i b_i(t)\,\varphi_i(x)\,dx \right\rangle \\
&= \frac{1}{2}\left\langle \sum_{i,j} b_i(t)\,b_j(t) \int_\Omega \varphi_i(x)\,\varphi_j(x)\,dx \right\rangle \\
&= \frac{1}{2}\left\langle \sum_{i,j} b_i(t)\,b_j(t)\,\delta_{ij} \right\rangle \\
&= \frac{1}{2}\left\langle \sum_i b_i(t)\,b_i(t) \right\rangle \\
&= \frac{1}{2}\sum_i \langle b_i(t)\,b_i(t)\rangle ,
\end{aligned} \tag{4.26}$$

and so the average energy in the ith mode is given by $\frac{1}{2}\langle b_i(t)\,b_i(t)\rangle$.

The following is a statement of the optimality of POD. Suppose that $u(x,t)$ in $L^2(\Omega)$ is a stationary random field and that $\{\varphi_i, \lambda_i | i = 1,\ldots,\infty; \lambda_i \geq \lambda_{i-1} > 0\}$ is the set of orthonormal empirical eigenfunctions with their associated eigenvalues obtained from time

averages of $u(x,t)$. Let

$$u(x,t) = \sum_i a_i(t)\,\varphi_i(x) \tag{4.27}$$

be the decomposition with respect to this basis and let $\{\varphi_i(x)\}_{i=1}^{\infty}$ be any other arbitrary orthonormal set such that

$$u(x,t) = \sum_i b_i(t)\,\psi_i(x) \tag{4.28}$$

then the following hold:

1. $\langle a_i(t)\,a_j(t)\rangle = \delta_{ij}\lambda_i$ (i.e. the POD random coefficients are uncorrelated).

2. For every m we have

$$\begin{aligned}\sum_{i=1}^{m}\langle a_i(t)\,a_i(t)\rangle &= \sum_{i=1}^{m}\lambda_i \\ &\geq \sum_{i=1}^{m}\langle b_i(t)\,b_i(t)\rangle\,,\end{aligned} \tag{4.29}$$

   i.e. POD is optimal on average in the class of linear representations: the first m POD basis functions capture more energy on average than the first m functions of any other basis.

The first assertion derives from the representation of $R(x,x')$, given in (4.11):

$$\begin{aligned}R(x,x') &= \langle u(x,t)\,u(x',t)\rangle \\ &= \left\langle \sum_i a_i(t)\,\varphi_i(x)\sum_j a_j(t)\,\varphi_j(x')\right\rangle \\ &= \sum_{i,j}\langle a_i(t)\,a_j(t)\rangle\,\varphi_i(x)\,\varphi_j(x')\,.\end{aligned} \tag{4.30}$$

But we know that from (4.11) that

$$R(x,x') = \sum_i \lambda_i\varphi_i(x)\,\varphi_i(x')\,,$$

and so, since the $\varphi_i(x)$ are an orthonormal family in $L^2(\Omega)$, we see that $\langle a_i(t)\, a_j(t)\rangle = \delta_{ij}\lambda_i$.

The second assertion relies on a result on linear operators. Let $\{\psi_j(x)\}_{j=1}^m$ be m arbitrary orthonormal vectors in $L^2(\Omega)$ that may be completed to form an orthonormal basis. Let Q denote projection onto $\mathrm{span}\{\psi_1, \ldots, \psi_m\}$. We can express the kernel R in terms of $\{\psi_j(x)\}_{j=1}^m$ as

$$\begin{aligned} R(x,x') &= \langle u(x,t)\, u(x',t)\rangle \\ &= \left\langle \sum_i b_i(t)\,\psi_i(x) \sum_j b_j(t)\,\psi_j(x') \right\rangle \\ &= \sum_{i,j} \langle b_i(t)\, b_j(t)\rangle\, \psi_i(x)\,\psi_j(x'). \end{aligned} \tag{4.31}$$

We can then write R in operator matrix notation as

$$\begin{bmatrix} \langle b_1(t)\, b_1(t)\rangle & \langle b_1 b_2(t)(t)\rangle & \langle b_1(t)\, b_3(t)\rangle & \ldots \\ \langle b_2(t)\, b_1(t)\rangle & \langle b_2 b_2(t)(t)\rangle & \langle b_2(t)\, b_3(t)\rangle & \ldots \\ \langle b_3(t)\, b_1(t)\rangle & \langle b_3 b_2(t)(t)\rangle & \langle b_3(t)\, b_3(t)\rangle & \ldots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{4.32}$$

and the product $R \circ Q$ yields

$$\begin{bmatrix} \langle b_1(t)\, b_1(t)\rangle & \langle b_1 b_2(t)(t)\rangle & \ldots & \langle b_1(t)\, b_m(t)\rangle & 0 & \ldots \\ \langle b_2(t)\, b_1(t)\rangle & \langle b_2 b_2(t)(t)\rangle & \ldots & \langle b_2(t)\, b_m(t)\rangle & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle b_m(t)\, b_1(t)\rangle & \langle b_m b_2(t)(t)\rangle & \ldots & \langle b_m(t)\, b_m(t)\rangle & 0 & \ldots \\ 0 & 0 & \ldots & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{4.33}$$

The proof is now completed by appeal to Remark 1.3 in Section V.1.2 of Temam [92], which states that the sum of the first m eigenvalues of a self-adjoint operator is greater than or equal to the sum of the diagonal terms in any m-dimensional projection of it:

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i &\geq \operatorname{tr}(R \circ Q) \\ &= \sum_{i=1}^{m} \langle b_i(t)\, b_i(t) \rangle . \end{aligned} \tag{4.34}$$

This characterization is the supports the claim that POD is optimal for reconstructing a signal $u(x,t)$. It implies that, from the set of all linear decompositions, POD is the most efficient so that for a certain number of modes, m, the projection on the subspace spanned by the dominant m POMs contains the greatest energy of the original signal. Moreover, the time series of the coefficients $a_i(t)$ are uncorrelated.

4.4 POD Discrete Formulation

Suppose that T linear snapshots $\mathbf{u}$ of dimension n are obtained at n locations (e.g., by use of piezoelectric accelerometers) of a system's response function $u(x,t)$. The response vector $\mathbf{u}(t)$ due to this forcing is normally stored in the discrete time format so that

$$\mathbf{u}(t_i) = \left\{ \begin{array}{c} u(x_1, t_i) \\ u(x_2, t_i) \\ \cdots \\ u(x_n, t_i) \end{array} \right\}, \tag{4.35}$$

where $t_i = i \times \delta t$ with δt is the uniform time step used in the data acquisition card, $i = 1, \ldots, T$, and T is the number of steps used in the measurement. Thus we have taken

snapshots of $u(x,t)$ in both time and physical space. We can form the matrix

$$\begin{aligned}\hat{\mathbf{U}} &= [\mathbf{u}(t_1), \ldots, \mathbf{u}(t_T)] \\ &= \begin{bmatrix} u_1(t_1) & \ldots & u_1(t_T) \\ \vdots & \vdots & \ldots \\ u_n(t_1) & \ldots & u_n(t_T) \end{bmatrix} \in \mathbb{R}^{n \times T}\end{aligned} \tag{4.36}$$

In the continuous time domain, we obtain the correlation matrix, $\mathbf{R}_{uu} \in \mathbb{R}^{n \times n}$, to be used in the proper orthogonal decomposition method, as

$$\mathbf{R}_{uu} = \left\langle \mathbf{u}(t)\, \mathbf{u}(t)^T \right\rangle \tag{4.37}$$

where $\langle \cdot \rangle$ is the time averaging operator. In the discrete time domain, under a limited number of observations of $\mathbf{u}(t)$, the maximum likelihood estimate of $\mathbf{R}_{uu}$ (refer to Section 3.2.1) is

$$\mathbf{R}_{uu} = \frac{1}{T} \hat{\mathbf{U}} \hat{\mathbf{U}}^T \tag{4.38}$$

The above matrix is symmetric and positive definite.

In the case that $\mathbf{u}(t)$ are snapshots of the output of a finite-dimensional MIMO (multi-input multi-output) linear system with $\mathbf{H}(\omega)$ as the system transfer matrix excited by incoherent band-limited (in frequency) stationary vector white noise of unit strength, the correlation matrix can also be expressed in the continuous frequency domain [93, 94] by

$$\mathbf{R}_{uu} = \int_B \Re \left\{ \mathbf{H}^\dagger(\omega) \mathbf{H}(\omega) \right\} d\omega \tag{4.39}$$

where $(\cdot)^\dagger$ is the is the complex conjugate transpose (Hermitian) operator, and B is the frequency bandwidth of interest. In the discrete frequency domain, the correlation matrix

can be obtained using

$$\mathbf{R}_{uu} = \frac{1}{K}\sum_{i=1}^{K} \Re\left\{\mathbf{H}^{\dagger}(\omega_i)\mathbf{H}(\omega_i)\right\} \tag{4.40}$$

where $\omega_1, \omega_2, \ldots, \omega_K \in B$ are the discrete frequencies at which $H(\omega)$ is evaluated.

Using the spectral decomposition of $\mathbf{R}_{uu}$ one obtains in matrix notation

$$\mathbf{R}_{uu} = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{4.41}$$

where $\mathbf{V}$ is the matrix containing the POD eigenvectors:

$$\mathbf{V} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n] \tag{4.42}$$

and $\mathbf{D}$ is the diagonal matrix whose entries are the corresponding POD eigenvalues:

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix}. \tag{4.43}$$

In summation notation, we have

$$\mathbf{R}_{uu} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \tag{4.44}$$

where the eigenvectors λ_i are arranged such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Due to the symmetry and positive definiteness of $\mathbf{R}_{uu}$, all eigenvalues are positive and the set of eigenvectors forms an orthonormal basis. For the POD method, as has already been shown, the first few modes are the most important, since they capture most of the systems energy, i.e. $\mathbf{R}_{uu}$ can be approximated by

$$\mathbf{R}_{uu} \approx \sum_{i=1}^{m} \lambda_i \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \tag{4.45}$$

where m is the number of dominant POD modes. As expected, the smaller the frequency band at which we are focusing, the smaller the number of dominant modes. Furthermore, the lower the frequencies that we use for order-reduction, the smaller the number of POD modes required to sufficiently describe the system behavior. The method for determining the optimal dimension, m, of the reduced-order system will be discussed in Section 4.5.

4.5 Determining the Number of Modes

Several methods for estimating the number of dominant POD modes have been used by researchers. There are several statistical and mathematical approaches to determining the number of modes, for example see [95]. However, the statistical and mathematical approaches have shortcomings in which they only determine the number of "trivial" modes forcing researchers to employ more objective techniques. For example, one mathematical approach is based on obtaining the rank of the correlation matrix $\mathbf{R}_{uu}$. However, in real data, the rank of the correlation matrix is almost always equal to its order, thus one has to apply more objective methods to determine the number of "nontrivial" modes.

Several methods have been suggested for estimating the number of nontrivial modes. The methods do not develop from specific rationales of either the statistical or mathematical varieties. Instead, a method usually develops from the researcher's experience with a series of POD analyses. The researcher happens to notice a common characteristic across these studies that appears to give the best number of modes and therefore suggests the use of this procedure in future research.

The two more important methods are based on computing the percentage of energy

extracted (Section 4.5.1) or plotting the POD eigenvalues (Section 4.5.2).

4.5.1 Percentage of Energy Extracted

Perhaps the oldest method that is of even limited current interest is the percentage of energy extracted. We have seen in Section 4.1 that each λ_i represents twice the average energy contributed by each mode φ_i. Thus by including more POD modes φ_i, one increases the total percentage of energy captured by the reduced order POD representation. The percentage of total energy extracted is computed by deviding the sum of the characteristic eigenvalues for the modes extracted by the sum of all the eigenvalues of the original correlation matrix, as in

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq \varepsilon \tag{4.46}$$

assuming that POD is required to capture ε of the energy of the measured displacements. Usually, investigators compute the cumulative percentage of energy extracted after each mode is removed from the correlation matrix and then stop the factoring process when $\varepsilon = 0.85, 0.90$, or 0.95 of the energy is accounted for. In general, m is much smaller than n.

A predetermined number of energy to be extracted may not be set; a table of the percentages may be examined instead. Usually factor extraction is stopped after a large portion of the energy has been extracted and when the next mode would add only a very small percentage to the total energy extracted. The procedure is an informal analysis of the information gained relative to the costs incurred. The information gained is reflected in the increased energy that is accounted for when an additional mode is extracted. The costs of that additional factor consist of increased complexity, more computational time, and greater

difficulties in arriving at such small factors. If a new mode does not add very much to the information already extracted, it would not be worth extracting and interpreting.

4.5.2 Plotting Eigenvalues

In accounting for an adequate amount of the energy while minimizing the number of modes, the percentage of energy extracted need not be computed. Because the devisor of (4.46) remains the same, only the cumulative eigenvalues need to be examined to obtain the same information. An even simpler procedure is to look at the raw eigenvalues themselves. When the eigenvalues drop dramatically in size, an additional mode would add relatively little to the information already extracted. That can be rationalized by examining the role of λ_i in (4.11). Thus the percentage of energy method has become a "break-in-eigenvalues" method.

The most detailed discussion of this procedure, along with additional theoretical rationale, can be found in [96]. Cattell refers to the procedure as the "scree" test because the term denotes the rubble at the bottom of a cliff. From an examination of a graph of characteristic eigenvalues, for example see Figure 7.3 on page 121 ,the analogy is obvious; the first few eigenvalues show the cliff and the rest show the rubble.

A basic rationale for the scree test is that the battery of variables is measuring a limited number of modes well and a larger number of trivial, specific, and error modes much less well. Therefore, the predominant modes account for most of the energy and are large, whereas the other modes are quite numerous but small. Because the principal mode solution extracts modes by size, the substantive modes will be extracted first and the smaller trivial modes will be removed later. Because the smaller modes are so numerous and are taken

out in order of size, it would be expected that plotting them on a graph would result in a straight line sloping downward. The dominant modes should not fall on this line because some of them will be much more dominant than others and they will be stronger than the trivial modes.

One complication that can occur is the presence of several breaks and several straight lines. For example, it is often apparent that the last few eigenvalues drop off rather sharply. Because no one is interested in extracting almost the same number of modes as variables, this drop is ignored and the straight line is based upon all the roots except those in the last drop.

A more serious problem occurs when there are two breaks among the first half of the eigenvalues. It is then difficult to decide which break reflects the proper number of modes, and other evidence may be needed to resolve the issue.

A problem also exists when there is no obvious break. Although the scree test can still provide a solution by noting where the last eigenvalues no longer form a straight line, the exact number of factors would not be conclusive. One should be sure that such a correlation matrix is significant because low-amplitude eigenvalues often have no major breaks in their curve.

4.6 System Model Reduction using POD modes

Previously, we have described the origins of POD and examined the attractive properties it possesses as a reduced order representation of a set of signals. In this section, we will describe how POD is applied towards achieving a reduced-order model of a system described

by a set of linear differential equations.

Let's examine the output vector $\mathbf{u}(t)$ of order n of a multi-input multi-output system. Having calculated the POD eigenspace, the output vector can be approximated by a linear representation involving the first m POD modes using the discrete form of (4.12) as

$$\mathbf{u}(t) = \sum_{i=1}^{m} a_i(t)\,\boldsymbol{\varphi}_i. \tag{4.47}$$

In matrix form, (4.47) becomes

$$\mathbf{u}(t) = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_m] \begin{Bmatrix} a_1(t) \\ \ldots \\ a_m(t), \end{Bmatrix} \tag{4.48}$$

which, can be rewritten as

$$\mathbf{u}(t) = \boldsymbol{\Sigma}\mathbf{a}(t). \tag{4.49}$$

where $\boldsymbol{\Sigma}$ is the matrix containing the first m dominant POD eigenvectors:

$$\boldsymbol{\Sigma} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_m] \in \mathbb{R}^{n \times m} \tag{4.50}$$

and $\mathbf{a}(t)$ is the vector containing the respective coefficients.

Thus, we can see that we can optimally approximate the system output by applying a transformation $\boldsymbol{\Sigma}$ upon the uncorrelated time-dependent modal coefficient vector $\mathbf{a}(t)$. To obtain the modal coefficients having calculated the eigenvector matrix $\boldsymbol{\Sigma}$, we multiply both sides of (4.49) on the left by $\boldsymbol{\Sigma}^T$:

$$\boldsymbol{\Sigma}^T\mathbf{u}(t) = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\mathbf{a}(t), \tag{4.51}$$

but we know that the POD modes are orthonormal, and thus we have $\mathbf{\Sigma}^T\mathbf{\Sigma} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix of order m. Applying this, we obtain

$$\begin{aligned} \mathbf{\Sigma}^T\mathbf{u}(t) &= \mathbf{I}\mathbf{a}(t) \\ &= \mathbf{a}(t), \end{aligned} \tag{4.52}$$

and thus we can calculate the modal coefficients from the knowledge of the transformation matrix containing the m-most dominant POD eigenvectors. Examining (4.52) closely, we see that each modal coefficient $a_i(t)$ is simply equal to the projection of the output vector $\mathbf{u}(t)$ onto the corresponding POD mode φ_i. This result was first stated in Section 4.2.

Starting with the system of equations describing the forced vibration of a viscously damped linear discrete system with n degrees of freedom:

$$\mathbf{M}_n\ddot{\mathbf{u}}_n(t) + \mathbf{C}_n\dot{\mathbf{u}}_n(t) + \mathbf{K}_n\mathbf{u}_n(t) = \mathbf{f}_n(t) \tag{4.53}$$

where $\mathbf{M}_n \in \mathbb{R}^{n\times n}$ is the mass matrix, $\mathbf{C}_n \in \mathbb{R}^{n\times n}$ is the damping matrix, $\mathbf{K}_n \in \mathbb{R}^{n\times n}$ is the stiffness matrix, $\mathbf{u}_n(t) \in \mathbb{R}^n$ is the displacement vector, and $\mathbf{f}_n(t) \in \mathbb{R}^n$ is the forcing vector at time t. Our aim is to arrive at a POD reduced order model of the above system of equations.

Using the representation (4.49) for $\mathbf{u}_n(t)$, we can obtain the first derivative of $\mathbf{u}$ with respect to time as

$$\begin{aligned} \dot{\mathbf{u}}_n(t) &= \frac{\partial}{\partial t}\mathbf{\Sigma}\mathbf{a}(t) \\ &= \mathbf{\Sigma}\frac{\partial}{\partial t}\mathbf{a}(t) \\ &= \mathbf{\Sigma}\dot{\mathbf{a}}(t). \end{aligned} \tag{4.54}$$

Similarly, the second derivative of $\mathbf{u}_n$ with respect to time is

$$\ddot{\mathbf{u}}_n(t) = \boldsymbol{\Sigma}\ddot{\mathbf{a}}(t). \tag{4.55}$$

Now using (4.49), (4.54), and (4.55), Equation (4.53) becomes

$$\mathbf{M}_n\boldsymbol{\Sigma}\ddot{\mathbf{a}}(t) + \mathbf{C}_n\boldsymbol{\Sigma}\dot{\mathbf{a}}(t) + \mathbf{K}_n\boldsymbol{\Sigma}\mathbf{a}(t) = \mathbf{f}_n(t). \tag{4.56}$$

Now, multiplying (4.56) on the left by the transpose of the POD transformation matrix $\boldsymbol{\Sigma}$ defined in (4.50), we obtain

$$\boldsymbol{\Sigma}^T\mathbf{M}_n\boldsymbol{\Sigma}\ddot{\mathbf{a}}(t) + \boldsymbol{\Sigma}^T\mathbf{C}_n\boldsymbol{\Sigma}\dot{\mathbf{a}}(t) + \boldsymbol{\Sigma}^T\mathbf{K}_n\boldsymbol{\Sigma}\mathbf{a}(t) = \boldsymbol{\Sigma}^T\mathbf{f}_n(t)(t). \tag{4.57}$$

The above system of equations can be rewritten in the reduced-order dimension as

$$\mathbf{M}_m\ddot{\mathbf{u}}_m(t) + \mathbf{C}_m\dot{\mathbf{u}}_m(t) + \mathbf{K}_m\mathbf{u}_m(t) = \mathbf{f}_m(t) \tag{4.58}$$

where

$$\mathbf{M}_m = \boldsymbol{\Sigma}^T\mathbf{M}_n\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m} \tag{4.59}$$

$$\mathbf{C}_m = \boldsymbol{\Sigma}^T\mathbf{C}_n\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m} \tag{4.60}$$

$$\mathbf{K}_m = \boldsymbol{\Sigma}^T\mathbf{K}_n\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m} \tag{4.61}$$

are the reduced order mass, damping, and stiffness matrices, respectively, and

$$\mathbf{u}_m(t) = \boldsymbol{\Sigma}^T\mathbf{u}_n(t) = \mathbf{a}(t) \tag{4.62}$$

$$\mathbf{f}_m(t) = \boldsymbol{\Sigma}^T\mathbf{f}_f(t) \tag{4.63}$$

are the reduced order displacement vector (equivalent to the POD modal coefficients) and reduced order forcing vector, respectively.

Transforming Equation (4.58) into the frequency domain, one has

$$\left[-\omega^2 \mathbf{M}_m + \mathrm{i}\omega \mathbf{C}_m + \mathbf{K}_m\right] \mathbf{U}_m(\omega) = \mathbf{F}_m(\omega) \tag{4.64}$$

where $\mathbf{U}_m(\omega) \in \mathbb{C}^m$ and $\mathbf{F}_m(\omega) \in \mathbb{C}^m$ are the Fourier transforms of $\mathbf{u}_m(t)$ and $\mathbf{f}_m(t)$, respectively.

Equation (4.58) is the reduced order model to be used for system identification. Thus, our objective is to identify the reduced order system matrices $\mathbf{M}_m$, $\mathbf{C}_m$, and $\mathbf{K}_m$. Recall that $\mathbf{U}_m(\omega)$ and $\mathbf{F}_m(\omega)$ are known from the original measurements.

Our proposed system matrix identification technique (see Chapter 6) can estimate, in theory, the system matrices of both the full-order system (4.53) as well as the matrices of the reduced-order system (4.58). In practice however, the order of a complex system can reach the tens of thousands, even hundreds of thousands for practical cases, if not more, depending on the problem at hand. However, having an optimal POD reduced-order model at hand to estimate, we are now trying to estimate matrices of orders in the tens or hundreds, allowing us to feasibly apply the proposed system identification method. The advantage in having a reduced order equation (4.58) is that the number of unknowns are now greatly reduced compared to the original system; each reduced-order system matrix has in the order of m^2 unknowns, a much smaller number than the number of unknowns in the full-order system matrices, in the order of n^2 unknowns.

Chapter 5

Independent Component Analysis

The aim of POD, as is described in Chapter 4, is to represent a spatio-temporal signal by optimal basis functions consisting of POMs. The basic advantage of POD stems from the fact that amplitudes of the projected signal on each of these basis functions are decorrelated. ICA achieves higher-order decorrelation of these projected amplitudes by choosing another transformation on POD basis. This Chapter outlines a reduced-order modeling technique using ICA and contrast its applicability with respect to POD-based model reduction strategy.

The chapter will begin by defining the basic concepts of ICA. The subsequent section discusses the similarities and differences between POD and ICA. Then two ICA methods are described in detail, namely the ICA by maximum likelihood estimation and ICA by tensorial methods. Finally, the chapter ends with the specific example of obtaining an ICA-based reduced-order model.

5.1 Introduction

5.1.1 ICA as Estimation of a Constructive Model

POD essentially relates to finding a linear representation in which the random amplitude coefficients are uncorrelated (see Section 4.3.2). ICA aims at finding a linear representation in which the components are statistically independent. In practical situations, we cannot in general find a representation where the components are really independent, but we can at least find components that are as independent as possible.

This leads to the following definition of ICA. Given a set of a vector random process $(x_1(t), x_2(t), \ldots, x_n(t))$, where t is the time or sample index, we assume they are generated as a linear combination of independent components (ICs):

$$\begin{Bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{Bmatrix} = \mathbf{A} \begin{Bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{Bmatrix} \tag{5.1}$$

where $\mathbf{A}$ is some unknown matrix. ICA relates to estimating both the matrix $\mathbf{A}$ and the $s_i(t)$, when we only observe $x_i(t)$. For simplicity, the number of independent components s_i is assumed to be equal to the number of observed variables. This simplifying assumption can be relaxed in the general case.

Alternatively, we could define ICA as follows: find a linear transformation

$$\mathbf{B}\begin{Bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{Bmatrix} = \begin{Bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{Bmatrix} \tag{5.2}$$

so that the random variables s_i, $i = 1, \ldots, n$, are independent, or as independent as possible. This formulation is not really very different from the previous one, since after estimating **A**, its inverse gives **B**.

It can be shown (see Section 5.3) that the problem is well-defined, that is, the model in (5.1) can be estimated if and only if the components s_i are nongaussian. This is a fundamental requirement that also explains the main difference between ICA and POD, as in the latter case the nongaussianity of the data is not taken into account. In fact, ICA could be interpreted as nongaussian POD, in which case we are also modeling the data as linear combination of some underlying components.

For simplicity, the time index t that was used in the previous definition will be ignored. This is because the basic ICA model assumes that the mixture x_i, as well as the ICs s_i, is a random process, instead of a measurable time signal or time series. The observed values $x_i(t)$ are then a sample of this random process.

ICA is very closely related to the method called blind source separation (BSS) or blind signal separation [97]. A source relates to an original signal, i.e. an IC. Blind means very little, if anything, is known of the mixing matrix, and that very weak assumptions are made on the source signals. ICA is one method for performing blind source separation.

It is usually more convenient to use a vector-matrix notation. Let us denote by **x** the

random vector whose elements are the mixtures $x_1, \ldots, x_n$, and likewise by **s** the random vector with elements $s_1, \ldots, s_n$. All vectors are column vectors; thus $\mathbf{x}^T$, or the transpose of **x**, is a row vector. Using this vector-matrix notation, the mixing model is written as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{5.3}$$

Sometimes we need the columns of matrix **A**; if we denote them by $\mathbf{a}_j$ the model can also be written as

$$\mathbf{x} = \sum_{i=1}^{n} \mathbf{a}_i s_i \tag{5.4}$$

The above definition for ICA is the most basic definition.

5.1.2 Restrictions in ICA

To make sure that the basic ICA model can be estimated with reasonable consistency, certain assumptions and restrictions are restrictions are made [98]:

I The independent components are assumed to be statistically independent.

This is the principle on which ICA is formulated. Surprisingly, this assumption all but guarantees that the model can be estimated. This is why ICA is such a powerful method with applications in many scientific areas.

Basically, random variables $y_1, y_2, \ldots, y_n$ are said to be independent if information on the value of y_i does not give any information on the value of y_j for $i \neq j$. Independence can be defined in respect to the probability densities. Let us denote by $p(y_1, y_2, \ldots, y_n)$ the joint probability density function (pdf) of the y_i, and by $p_i(y_i)$ the marginal pdf of y_i, i.e., the pdf of y_i when it is considered separately. In this case,

the y_i are independent if and only if the joint pdf is factorizable in the following way:

$$p(y_1, y_2, \ldots, y_n) = p_1(y_1)\, p_2(y_2) \cdots p_n(y_n). \tag{5.5}$$

II The independent components must be nongaussian.

Intuitively, one can say that the gaussian distributions are too elementary. The cumulants of order three and up are zero for gaussian distributions, but such higher-order information is essential for estimation of the ICA model, as will be seen in Section 5.2.2. Thus, ICA is essentially a POD representation if the observed variables have gaussian distributions. The case of gaussian components is treated in more detail in Section 5.3 below. Note that in the basic model no assumption on the nature of the nongaussian distributions of the ICs is made. If the distributions are known, the problem is considerably simplified.

III For simplicity, the unknown mixing matrix is assumed square.

In that case, the number of independent components is equal to the number of observed mixtures. This assumption simplifies the estimation procedure. As will be shown, POD is the first step in ICA. At the POD step the number of components m is chosen, as is discussed in Section 4.5. Consequently, ICA translates to determining the m independent components from the m POD components. In this case, after estimating the matrix $\mathbf{A}$, we can compute its inverse, say $\mathbf{B}$, and obtain the ICs simply by

$$\mathbf{s} = \mathbf{B}\mathbf{x}. \tag{5.6}$$

Under the preceding three assumptions, the ICA model is identifiable, meaning that the mixing matrix and the ICs can be estimated. The identifiability of the ICA model is not

proved here, since the proof is quite complicated. The proof can be found in [99, 100]. On the other hand, in the upcoming sections we present ICA estimation methods, and the developments there give a kind of intuitive proof of the identifiability without significant theoretical rigor.

5.1.3 indeterminacies of ICA

In the ICA model (5.3), the following ambiguities may arise:

I The variances (energies) of the independent components cannot be determined.

With both **s** and **A** unknown, any scalar multiplier in one of the sources s_i could always be canceled by dividing the corresponding column $\mathbf{a}_i$ of **A** by the same scalar, say α_i:

$$\mathbf{x} = \sum_{i=1}^{n} \left(\frac{1}{\alpha_i} \mathbf{a}_i \right) (\alpha_i s_i) . \tag{5.7}$$

As a consequence, we may fix the magnitudes of the independent components. Since they are assumed to be random processes, the most natural way to address this issue is to assume that each has unit variance: $E\{s_i^2\} = 1$. Then the matrix **A** will be adapted to take this restriction into account.

II The order of the independent components cannot be induced.

The reason is that, again both **s** and **A** being unknown, the order of the terms in the sum in (5.4) can be changed. A permutation matrix **P** and its inverse can be substituted in the model to give $\mathbf{x} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$. The elements of **Ps** are the original independent variables s_j, but in another order. The matrix $\mathbf{A}\mathbf{P}^{-1}$ is just a new unknown mixing matrix to be estimated by ICA.

5.1.4 Centering the Variables

Without loss of generality, we can assume that both the mixture variables and the independent components have zero mean. This assumption simplifies the theory and algorithms significantly and is pursued in the remainder of this chapter.

If the assumption of zero mean is not true, one can perform some preprocessing to make it hold. This is possible by centering the observed variables, i.e. subtracting their sample mean. This means that the original mixtures $\mathbf{x}'$ are preprocessed by

$$\mathbf{x} = \mathbf{x}' - \mathrm{E}\{\mathbf{x}'\} \tag{5.8}$$

before performing ICA. Thus the independent components are also a zero mean process, since

$$E\{\mathbf{s}\}\,\mathbf{x} = \mathbf{A}^{-1}E\{\mathbf{x}\}\,. \tag{5.9}$$

The mixing matrix, on the other hand, remains the same after this preprocessing, so we can always carry out this operation without affecting the estimation procedure of the mixing matrix. After estimating the mixing matrix and the independent components for the zero-mean data, the subtracted mean can be simply reconstructed by adding $\mathbf{A}^{-1}E\{\mathbf{x}'\}$to the zero-mean independent components.

5.2 ICA is stronger than Uncorrelation

Given some random variables or processes, it is straightforward to linearly transform them into their uncorrelated counterparts. Therefore, it would be tempting to try to estimate the independent components by such a method, which is often implemented by POD. In this

section, we show that this is not possible, and discuss the relation between ICA and decorrelation methods. It will be seen that decorrelation is, nevertheless, a useful preprocessing technique for ICA.

5.2.1 Uncorrelatedness

A weaker form of independence is uncorrelatedness. Two random variables y_1 and y_2 are said to be uncorrelated, if their covariance is zero:

$$\begin{aligned} \operatorname{cov}(y_1, y_2) &= \mathrm{E}\{y_1 y_2\} - \mathrm{E}\{y_1\}\mathrm{E}\{y_2\} \\ &= 0. \end{aligned} \tag{5.10}$$

In this chapter, all random variables are assumed to have zero mean, unless otherwise mentioned. Thus, covariance is equal to correlation, i.e. $\operatorname{corr}(y_1, y_2) = \mathrm{E}\{y_1 y_2\}$.

If random variables are independent, they are uncorrelated. This is because if the y_1 and y_2 are independent, then for any two functions, h_1 and h_2, we have [101]

$$\mathrm{E}\{h_1(y_1)\, h_2(y_2)\} = \mathrm{E}\{h_1(y_1)\}\mathrm{E}\{h_2(y_2)\}, \tag{5.11}$$

Taking $h_1(y_1) = y_1$ and $h_2(y_2) = y_2$, we see that this implies uncorrelatedness.

On the other hand, uncorrelatedness does not imply independence. For example, assume that (y_1, y_2) are discrete valued and follow such a distribution that the pair are with probability $\frac{1}{4}$ equal to any of the following values: $(0, 1)$, $(0, -1)$, $(1, 0)$, and $(-1, 0)$. Then y_1 and y_2 are uncorrelated, as can be simply calculated. On the other hand,

$$\mathrm{E}\{y_1^2 y_2^2\} = 0 \neq \frac{1}{4} = \mathrm{E}\{y_1^2\}\mathrm{E}\{y_2^2\} \tag{5.12}$$

so the condition in Equation (5.11) is violated, and the variables cannot be independent.

Uncorrelatedness of a zero-mean random vector, say $\mathbf{y}$, means that its components are uncorrelated. In other words, the covariance matrix (as well as the correlation matrix) of $\mathbf{y}$ is a diagonal matrix:

$$\mathrm{E}\left\{\mathbf{y}\mathbf{y}^T\right\} = \mathbf{D}. \tag{5.13}$$

Consequently, decorrelation means that we linearly transform the observed data vector $\mathbf{x}$ by linearly multiplying it with some matrix $\mathbf{V}$

$$\mathbf{z} = \mathbf{V}\mathbf{x} \tag{5.14}$$

so that we obtain a new vector $\mathbf{z}$ that is uncorrelated.

A decorrelation transformation is always possible. Decorrelation can be performed in connection with proper orthogonal decomposition (see Section 4.4), which gives a related decorrelation matrix. Assuming we have the eigenvalue decomposition (EVD) of the covariance matrix

$$\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T\right\} = \mathbf{E}\mathbf{D}\mathbf{E}^T \tag{5.15}$$

where $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T\right\}$ and $\mathbf{D}$ is the diagonal matrix of its eigenvalues, $\mathbf{D} = \mathrm{diag}\left(d_1, \ldots, d_n\right)$. Decorrelation can now be done by the matrix

$$\mathbf{V} = \mathbf{E}^T. \tag{5.16}$$

It is easy to show that the matrix $\mathbf{V}$ in (5.16) is indeed a decorrelation transformation.

Recalling that $\mathbf{E}$ is an orthogonal matrix satisfying $\mathbf{E}^T\mathbf{E} = \mathbf{E}\mathbf{E}^T = \mathbf{I}$, we have

$$\begin{aligned} \mathrm{E}\left\{\mathbf{z}\mathbf{z}^T\right\} &= \mathrm{E}\left\{\mathbf{V}\mathbf{x}\mathbf{x}^T\mathbf{V}^T\right\} \\ &= \mathbf{V}\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{V}^T \\ &= \mathbf{E}^T\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{E} \\ &= \mathbf{E}^T\mathbf{E}\mathbf{D}\mathbf{E}^T\mathbf{E} \\ &= \mathbf{I}\mathbf{D}\mathbf{I} \\ &= \mathbf{D}. \end{aligned} \tag{5.17}$$

The covariance of $\mathbf{z}$ is the diagonal matrix $\mathbf{D}$, hence $\mathbf{z}$ is uncorrelated.

The linear operator V of (5.16) is by no means the only unique decorrelation matrix. It is easy to see that any matrix $\mathbf{U}\mathbf{V}$, with $\mathbf{U}$ an orthogonal matrix, is also a decorrelation matrix. This is because for $\mathbf{z} = \mathbf{U}\mathbf{V}\mathbf{x}$ it holds:

$$\begin{aligned} \mathrm{E}\left\{\mathbf{z}\mathbf{z}^T\right\} &= \mathbf{U}\mathbf{V}\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{V}^T\mathbf{U}^T \\ &= \mathbf{U}\mathbf{D}\mathbf{U}^T \\ &= \mathbf{D}. \end{aligned} \tag{5.18}$$

5.2.2 Decorrelation is only half ICA

Now, suppose that the data in the ICA model is decorrelated, for example, by the matrix given in (5.16). Decorrelation transforms the mixing matrix into a new one, $\widetilde{\mathbf{A}}$. We have from (5.3) and (5.14)

$$\mathbf{z} = \mathbf{V}\mathbf{A}\mathbf{s} = \widetilde{\mathbf{A}}\mathbf{s}. \tag{5.19}$$

One could hope that decorrelation solves the ICA problem, since uncorrelatedness is related to independence. This is, however, not so. Uncorrelatedness is weaker than independence, and is not in itself sufficient for estimation of the ICA model. To see this, consider an orthogonal transformation $\mathbf{U}$ of $\mathbf{z}$:

$$\mathbf{y} = \mathbf{U}\mathbf{z}. \tag{5.20}$$

Due to the orthogonality of $\mathbf{U}$, we have

$$\begin{aligned} \mathrm{E}\left\{\mathbf{y}\mathbf{y}^T\right\} &= \mathrm{E}\left\{\mathbf{U}\mathbf{z}\mathbf{z}^T\mathbf{U}^T\right\} \\ &= \mathbf{U}\mathbf{D}\mathbf{U}^T \\ &= \mathbf{D}. \end{aligned} \tag{5.21}$$

In other words, $\mathbf{y}$ is uncorrelated as well. Thus, we cannot tell if the independent components are given by $\mathbf{z}$ or $\mathbf{y}$ using the uncorrelated property alone. Since $\mathbf{y}$ could be any orthogonal transformation of $\mathbf{z}$, decorrelation gives the ICs only up to an orthogonal transformation. This is not sufficient in most applications.

On the other hand, decorrelation is useful as a preprocessing step in ICA. The utility of decorelation resides in the fact that the new mixing matrix $\widetilde{\mathbf{A}} = \mathbf{V}\mathbf{A}$ is orthogonal. This can be seen from

$$\begin{aligned} \mathrm{E}\left\{\mathbf{z}\mathbf{z}^T\right\} &= \widetilde{\mathbf{A}}\mathrm{E}\left\{\mathbf{s}\mathbf{s}^T\right\}\widetilde{\mathbf{A}}^T \\ &= \widetilde{\mathbf{A}}\mathbf{D}\widetilde{\mathbf{A}}^T \\ &= \mathbf{D}. \end{aligned} \tag{5.22}$$

since $\mathbf{s}$ is independent and this is uncorrelated.

This means that we can restrict our search for the mixing matrix to the space of orthogonal matrices. Instead of having to estimate the n^2 parameters that are the elements of the original matrix $\mathbf{A}$, we only need to estimate an orthogonal mixing matrix $\widetilde{\mathbf{A}}$. An orthogonal matrix contains $n\left(n-1\right)/2$ degrees of freedom. For example, in two dimensions, an orthogonal transformation is determined by a single angle parameter. In larger dimensions, an orthogonal matrix contains only about half of the number of parameters of an arbitrary matrix.

Thus one can say that decorrelation solves half of the problem of ICA. Because decorrelation is a standard procedure, much simpler than any ICA algorithms, it is a good idea to reduce the complexity of the problem this way. The remaining half of the parameters has to be estimated by some other method as will be discussed in the upcoming sections.

In subsequent sections, we assume that the data has been preprocessed by decorrelation, in which case we denote the data by $\mathbf{z}$. Even in cases where decorrelation is not explicitly required, it is recommended, since it reduces the number of free parameters and considerably increases the performance of the methods, especially with high-dimensional data.

5.3 Why Gaussian Variables Are Forbidden

Decorrelation also helps us understand why gaussian variables are forbidden in ICA. Assume that the joint distribution of two ICs, s_1 and s_2, is gaussian. Furthermore, assume that the distribution has zero mean and identity covariance matrix, without loss of generality.

This means that their joint pdf is given by (refer to [101])

$$\begin{aligned} p(s_1, s_2) &= \frac{1}{2\pi}\exp\left(-\frac{s_1^2 + s_2^2}{2}\right) \\ &= \frac{1}{2\pi}\exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right). \end{aligned} \tag{5.23}$$

Now, assume that the mixing matrix $\mathbf{A}$ is orthogonal. For example, we could assume that that is so because the data has been decorrelated. Using the classic formula of transforming pdf [101], and noting that for an orthogonal matrix $\mathbf{A}^{-1} = \mathbf{A}^T$ holds, we get the joint density of the mixtures x_1 and x_2 as

$$p(x_1, x_2) = \frac{1}{2\pi}\exp\left(-\frac{\left\|\mathbf{A}^T\mathbf{x}\right\|^2}{2}\right)\left|\det\left(\mathbf{A}^T\right)\right|. \tag{5.24}$$

Due to the orthogonality of $\mathbf{A}$, we have $\left\|\mathbf{A}^T\mathbf{x}\right\|^2 = \|\mathbf{x}\|^2$ and $\left|\det\left(\mathbf{A}^T\right)\right| = 1$; note that if $\mathbf{A}$ is orthogonal, so is $\mathbf{A}^T$. Thus we have

$$p(x_1, x_2) = \frac{1}{2\pi}\exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right), \tag{5.25}$$

and we see that the orthogonal mixing matrix does not change the pdf, since it does not appear in this pdf at all. The original and mixed distributions are identical. Therefore, there is no way how we could infer the mixing matrix from the mixtures.

The phenomenon that the orthogonal mixing matrix cannot be estimated for gaussian variables is related to the property that uncorrelated jointly gaussian variables are necessarily independent [101]. Thus, the information on the independence of the components does not provide any further information than decorrelation.

Thus, in the case of gaussian independent components, we can only estimate the ICA model up to an orthogonal transformation. In other words, the matrix $\mathbf{A}$ is not identifiable for gaussian independent components. With gaussian variables, all we can achieve is

decorrelation of the data. There are some choices in the decorrelation procedure, however; POD is the classic choice.

In the case of estimating the ICA model where some of the components are gaussian and some nongaussian. one can estimate all the nongaussian components, but the gaussian components cannot be separated from each other. In other words, some of the estimated components will be arbitrary linear combinations of the gaussian components. Actually, this means that in the case of just one gaussian component, we can estimate the model, because the single gaussian component does not have any other gaussian components that it could be mixed with.

5.4 ICA by MLE

A very popular approach for estimating the ICA model is MLE. Maximum likelihood estimation is a fundamental method of statistical estimation; a short introduction was provided in Section 3.2.1. One interpretation of MLE is that we take those parameter values as estimates that give the highest probability for the observations. In this section, we show how to apply MLE to ICA estimation.

5.4.1 The Likelihood of the ICA Model

Deriving the Likelihood

The derivation of the likelihood of the ICA model is based on using the well-known result on the density of a linear transform [101]. According to this result, the density p_x of the

observed vector

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{5.26}$$

can be formulated as

$$\begin{aligned} p_x(\mathbf{x}) &= |\det(\mathbf{B})| \, p_s(\mathbf{s}) \\ &= |\det(\mathbf{B})| \prod_i p_{i=1}^n(s_i), \end{aligned} \tag{5.27}$$

where $\mathbf{B} = \mathbf{A}^{-1}$, and the p_i denote the densities of the independent components. This can be expressed as a function of $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ and $\mathbf{x}$, giving

$$p_x(\mathbf{x}) = |\det(\mathbf{B})| \prod_i p_i\left(\mathbf{b}_i^T \mathbf{x}\right). \tag{5.28}$$

Assume that we have T observations of $\mathbf{x}$, denoted by $\mathbf{x}(1), \ldots, \mathbf{x}(T)$. Then the likelihood can be obtained (see Section 3.2.1) as the product of this density evaluated at the T points. This is denoted by $\mathcal{L}$ and considered as a function of $\mathbf{B}$:

$$\begin{aligned} \mathcal{L}(\mathbf{B}) &= \prod_{t=1}^{T} |\det(\mathbf{B})| \prod_{i=1}^{n} p_i\left(\mathbf{b}_i^T \mathbf{x}(t)\right) \\ &= |\det(\mathbf{B})|^T \prod_{t=1}^{T} \prod_{i=1}^{n} p_i\left(\mathbf{b}_i^T \mathbf{x}(t)\right). \end{aligned} \tag{5.29}$$

Very often it is more practical to use the logarithm of the likelihood, since it is algebraically simpler. This does not make any difference here since the maximum of the logarithm is obtained at the same point as the maximum of the likelihood. The log-likelihood is given by

$$\ln \mathcal{L}(\mathbf{B}) = T \ln |\det(\mathbf{B})| + \sum_{t=1}^{T} \sum_{i=1}^{n} \ln p_i\left(\mathbf{b}_i^T \mathbf{x}(t)\right). \tag{5.30}$$

To simplify notation, we can denote the sum over the sample index t by an expectation operator, and divide the likelihood by T to obtain

$$\frac{1}{T}\ln \mathcal{L}\left(\mathbf{B}\right) = \ln\left|\det\left(\mathbf{B}\right)\right| + \mathrm{E}\left\{\sum_{i=1}^{n} \ln p_i\left(\mathbf{b}_i^T\mathbf{x}\right)\right\}. \tag{5.31}$$

The expectation here is not the theoretical expectation, but an average computed from the observed sample. Of course, in the algorithms the expectations are eventually replaced by sample averages, so the distinction is purely theoretical.

Estimation of the Densities

In the preceding section, we have expressed the likelihood as a function of the parameters of the model, which are the elements of the mixing matrix. For simplicity, we used the elements of the inverse **B** of the mixing matrix. This is allowed since the mixing matrix can be directly computed from its inverse.

There is, however, another aspect to the estimation in the ICA model. This is the densities of the independent components, $p_i\left(s_i\right)$. Actually, the likelihood is a function of these densities as well. This makes the problem much more complicated, because the estimation of densities is, in general, a nonparametric problem. Nonparametric means that it cannot be reduced to the estimation of a finite parameter set. In fact the number of parameters to be estimated is infinite, or in practice, very large. Thus the estimation of the ICA model has also a nonparametric part, which is why the estimation is sometimes called semiparametric.

Nonparametric estimation of densities is known to be a difficult problem. Many parameters are always more difficult to estimate than just a few; since nonparametric problems have an infinite number of parameters, they are the most difficult to estimate. This is why

we would like to avoid the nonparametric density estimation in the ICA. There are two ways to avoid it.

First, in some cases we might know the densities of the independent components in advance, using some prior knowledge on the data at hand. In this case, we could simply use these prior densities in the likelihood. Then the likelihood would really be a function of $\mathbf{B}$ only. If reasonably small errors in the specification of these prior densities have little influence on the estimator, this procedure will give acceptable results. In fact, it will be shown below that this is the case.

A second way to solve the problem of density estimation is to approximate the densities of the independent components by a family of densities that are specified by a limited number of parameters. If the number of parameters in the density family needs to be very large, we do not gain much from this approach, since the goal was to reduce the number of parameters to be estimated. However, if it is possible to use a very simple family of densities to estimate the ICA model for any densities p_i, we will get a simple solution. Fortunately, this turns out to be the case. We can use an extremely simple parameterization of the p_i, consisting of the choice between two densities, i.e., a single binary parameter.

It turns out that in maximum likelihood estimation, it is enough to use just two approximations of the density of an independent component. For each independent component, we just need to determine which one of the two approximations is better. This shows that, first, we can make small errors when we fix the densities of the independent components, since it is enough that we use a density that is in the same half of the space of probability densities. Second, it shows that we can estimate the independent components using very simple models of their densities, in particular, using models consisting of only two densities.

The validity of these approaches is shown in the following theorem, whose proof can be found in [82].

Theorem 1. *Denote by $\widetilde{p}_i$ the assumed densities of the independent components, and*

$$\begin{aligned} g_i(s_i) &= \frac{\partial}{\partial s_i} \ln \widetilde{p}_i(s_i) \\ &= \frac{1}{\widetilde{p}_i(s_i)} \frac{\partial \widetilde{p}_i(s_i)}{\partial s_i}. \end{aligned} \tag{5.32}$$

Constrain the estimates of the independent components $s_i = \mathbf{b}_i^T \mathbf{x}$ to be uncorrelated and to have unit variance. Then the MLE estimator is locally consistent, if the assumed densities $\widetilde{p}_i$ fulfill

$$\mathrm{E}\left\{ s_i g_i(s_i) - \frac{\partial g_i(s_i)}{\partial s_i} \right\} > 0 \tag{5.33}$$

for all i.

This theorem shows rigorously that small misspecifications in the densities p_i do not affect the local consistency of the MLE estimator, since sufficiently small changes do not change the sign in (5.33).

Moreover, the theorem shows how to construct families consisting of only two densities, so that the condition in (5.33) is true for one of these densities. For example, consider the following log-densities:

$$\ln \widetilde{p}_i^{+}(s_i) = \alpha_1 - 2 \ln \cosh(s_i) \tag{5.34}$$

$$\ln \widetilde{p}_i^{-}(s_i) = \alpha_2 - s_i^2/2 + \ln \cosh(s_i) \tag{5.35}$$

where α_1 and α_2 are positive parameters that are fixed so as to make these two functions logarithms of probability densities. Actually, these constants can be ignored in the follow-

ing. The factor 2 in (5.34) is not important, but it is usually used here; also, the factor $1/2$ in (5.35) could be changed.

The motivation for these functions is that $\widetilde{p}_i^+$ is a supergaussian density, because the $\ln\cosh$ function is close to the absolute value that would give the Laplacian density. The density given by $\widetilde{p}_i^-$ is subgaussian, because it is like a gaussian log-density, $s_i^2/2$ plus a constant, that has been somewhat flattened by the $\ln\cosh$ function.

For $\widetilde{p}_i^+$, we have

$$\begin{aligned} g_i^+(s_i) &= \frac{\partial}{\partial s_i} \ln \widetilde{p}_i^+(s_i) \\ &= \frac{\partial}{\partial s_i} (\alpha_1 - 2\ln\cosh(s_i)) \\ &= -2\frac{1}{\cosh(s_i)} \frac{\partial}{\partial s_i} \cosh(s_i) \\ &= -2\frac{1}{\cosh(s_i)} \sinh(s_i) \\ &= -2\tanh(s_i), \end{aligned}$$

and

$$\frac{\partial g_i^+(s_i)}{\partial s_i} = -2\left(1 - \tanh^2(s_i)\right).$$

Thus the value of the nonpolynomial moment in (5.33) for $\widetilde{p}_i^+$ is

$$\begin{aligned} \mathrm{E}\left\{ s_i g_i^+(s_i) - \frac{\partial g_i^+(s_i)}{\partial s_i} \right\} &= \mathrm{E}\left\{ -2\tanh(s_i)\, s_i + 2\left(1 - \tanh^2(s_i)\right) \right\} \\ &= 2\mathrm{E}\left\{ 1 - \tanh(s_i)\, s_i - \tanh^2(s_i) \right\} \end{aligned} \tag{5.36}$$

Similarly, for $\widetilde{p_i^-}$, we have

$$\begin{aligned} g_i^- (s_i) &= \frac{\partial}{\partial s_i} \ln \widetilde{p_i^-} (s_i) \\ &= \frac{\partial}{\partial s_i} \left(\alpha_2 - s_i^2/2 + \ln \cosh (s_i)\right) \\ &= -s_i + \frac{1}{\cosh (s_i)} \frac{\partial}{\partial s_i} \cosh (s_i) \\ &= -s_i + \frac{1}{\cosh (s_i)} \sinh (s_i) \\ &= -s_i + \tanh (s_i) , \end{aligned}$$

and

$$\begin{aligned} \frac{\partial g_i^- (s_i)}{\partial s_i} &= -1 + \left(1 - \tanh^2 (s_i)\right) \\ &= -\tanh^2 (s_i) . \end{aligned}$$

Thus the value of the nonpolynomial moment in (5.33) for $\widetilde{p_i^-}$ is

$$\begin{aligned} \mathrm{E}\left\{ s_i g_i^- (s_i) - \frac{\partial g_i^- (s_i)}{\partial s_i} \right\} &= \mathrm{E}\left\{ \left(-s_i + \tanh (s_i)\right) s_i + \tanh^2 (s_i) \right\} \\ &= \mathrm{E}\left\{ -s_i^2 + \tanh (s_i) s_i + \tanh^2 (s_i) \right\} \\ &= -\mathrm{E}\left\{ s_i^2 \right\} + \mathrm{E}\left\{ \tanh (s_i) s_i + \tanh^2 (s_i) \right\} \\ &= -1 + \mathrm{E}\left\{ \tanh (s_i) s_i + \tanh^2 (s_i) \right\} \\ &= \mathrm{E}\left\{ -1 + \tanh (s_i) s_i + \tanh^2 (s_i) \right\} . \end{aligned} \tag{5.37}$$

We see that the signs of the expressions (5.36) and (5.37) are always opposite. Thus, for practically any distributions of the s_i, one of these functions fulfills the condition (5.33), i.e., has the desired sign, and estimation is possible. Of course, for some distribution of the s_i the nonpolynomial moment in the condition could be zero, which corresponds to the

case of zero kurtosis in cumulant-based estimation; such cases can be considered to be very rare.

Thus we can just compute the nonpolynomial moments for the two prior distributions in (5.34) and (5.35), and choose the one that fulfills the stability condition in (5.33). This can be done during the computation involving the maximization of the likelihood. This always provides a (locally) consistent estimator, and solves the problem of semiparametric estimation.

In fact, the nonpolynomial moment in question measures the shape of the density function in much the same way as kurtosis. For $g_i(s_i)$, we would actually obtain kurtosis. Thus, the choice of nonlinearity could be compared with the choice whether to minimize or maximize kurtosis.

5.4.2 Algorithms for Maximum Likelihood Estimation

To perform maximum likelihood estimation in practice, we need an algorithm to perform the numerical maximization of likelihood. In this section, we discuss different methods to this end. First, we show how to derive simple gradient algorithms, of which especially the natural gradient algorithm has been widely used. Then we show how to derive a fixed-point algorithm, a version of FastICA, that maximizes the likelihood faster and more reliably.

Gradient Algorithms

The Bell-Sejnowski Algorithm

The simplest algorithms for maximizing likelihood are obtained by gradient methods. We start with the log-likelihood equation (5.31):

$$\frac{1}{T}\ln \mathcal{L}\left(\mathbf{B}\right) = \ln\left|\det\left(\mathbf{B}\right)\right| + \mathrm{E}\left\{\sum_{i=1}^{n}\ln p_i\left(\mathbf{b}_i^T\mathbf{x}\right)\right\}. \tag{5.38}$$

where $\mathbf{b}_i^T$ represents the ith row of $\mathbf{B}$.

Our aim is to maximize the above function with respect to the matrix $\mathbf{B}$, thus we need to take the derivative of the log-likelihood function with respect to the elements of $\mathbf{B}$, b_{rs}. Let's start with the second term from the right side of (5.38):

$$\begin{aligned}
\frac{\partial}{\partial b_{rs}}\mathrm{E}\left\{\sum_{i=1}^{n}\ln p_i\left(\mathbf{b}_i^T\mathbf{x}\right)\right\} &= \mathrm{E}\left\{\frac{\partial}{\partial b_{rs}}\sum_{i=1}^{n}\ln p_i\left(\mathbf{b}_i^T\mathbf{x}\right)\right\} \\
&= \mathrm{E}\left\{\frac{\partial}{\partial b_{rs}}\sum_{i=1}^{n}\ln p_i\left(\sum_j b_{ij}x_j\right)\right\} \\
&= \mathrm{E}\left\{\frac{\partial}{\partial b_{rs}}\ln p_r\left(\sum_j b_{rj}x_j\right)\right\} \\
&= \mathrm{E}\left\{\frac{\partial}{\partial b_{rs}}\ln p_r\left(s_r\right)\right\} \\
&= \mathrm{E}\left\{\frac{\partial \ln p_r\left(s_r\right)}{\partial s_r}\frac{\partial s_r}{\partial b_{rs}}\right\} \\
&= \mathrm{E}\left\{g_r\left(s_r\right)\frac{\partial \mathbf{b}_r^T\mathbf{x}}{\partial b_{rs}}\right\} \\
&= \mathrm{E}\left\{g_r\left(\mathbf{b}_r^T\mathbf{x}\right)\frac{\partial \sum_j b_{rj}x_j}{\partial b_{rs}}\right\} \\
&= \mathrm{E}\left\{g_r\left(\mathbf{b}_r^T\mathbf{x}\right)x_s\right\}.
\end{aligned} \tag{5.39}$$

In matrix form, the matrix gradient (see Section 3.1) of the log-likelihood function with

respect to the matrix $\mathbf{B}$ is

$$\begin{aligned}\frac{\partial}{\partial \mathbf{B}} \mathrm{E}\left\{\sum_{i=1}^{n} \ln p_i\left(\mathbf{b}_i^T \mathbf{x}\right)\right\} &= \mathrm{E}\left\{\left\{\begin{array}{c} g_1\left(\mathbf{b}_1^T \mathbf{x}\right) \\ \vdots \\ g_n\left(\mathbf{b}_n^T \mathbf{x}\right) \end{array}\right\}\left[\begin{array}{ccc} x_1 & \ldots & x_n \end{array}\right]\right\} \\ &= \mathrm{E}\left\{\mathbf{g}\left(\mathbf{B}\mathbf{x}\right)\mathbf{x}^T\right\}. \end{aligned} \tag{5.40}$$

Here, $\mathbf{g}\left(\mathbf{y}\right)$ is a component-wise vector function that consists of the so-called score functions g_i of the distributions of s_i, defined in (5.32).

Now, we apply this identity as well as the result obtained for the matrix gradient of the logarithm of the determinant (3.13) to obtain the likelihood equation:

$$\frac{1}{T}\frac{\partial}{\partial \mathbf{B}}\mathcal{L} = \left(\mathbf{B}^T\right)^{-1} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{B}\mathbf{x}\right)\mathbf{x}^T\right\}. \tag{5.41}$$

This immediately gives the following gradient for the MLE

$$\Delta \mathbf{B} \propto \left(\mathbf{B}^T\right)^{-1} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{B}\mathbf{x}\right)\mathbf{x}^T\right\}. \tag{5.42}$$

for which the iterative algorithm to estimating $\mathbf{B}$ is

$$\mathbf{B} \leftarrow B + \mu\left[\left(\mathbf{B}^T\right)^{-1} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{B}\mathbf{x}\right)\mathbf{x}^T\right\}\right]. \tag{5.43}$$

This algorithm is often called the Bell-Sejnowski algorithm [70]. This algorithm converges very slowly, however, especially due to the inversion of the matrix $\mathbf{B}$ that is needed in every step. The convergence can be improved by decorrelating the data (see Section 5.2, and especially by using the natural gradient.

The natural (or relative) gradient method simplifies the maximization of the likelihood considerably, and makes it better conditioned. The principle of the natural gradient is

based on the geometrical structure of the parameter space, and is related to the principle of relative gradient, which uses a certain group structure of the ICA problem (see [82] for more details). The basic idea is to multiply the matrix gradient by some orthogonal matrix that would lead to a faster converging algorithm. In the case of basic ICA, both of these principles amount to multiplying the right-hand side by $\mathbf{B}\mathbf{B}^T$. Thus we obtain

$$\begin{aligned}
\Delta\mathbf{B} &\propto \left[\left(\mathbf{B}^T\right)^{-1} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{B}\mathbf{x}\right)\mathbf{x}^T\right\}\right]\mathbf{B}^T\mathbf{B} \\
&= \left(\mathbf{B}^T\right)^{-1}\mathbf{B}^T\mathbf{B} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{B}\mathbf{x}\right)\mathbf{x}^T\right\}\mathbf{B}^T\mathbf{B} \\
&= \mathbf{B} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{y}\right)\mathbf{x}^T\right\}\mathbf{B}^T\mathbf{B} \\
&= \left(\mathbf{I} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{y}\right)\mathbf{x}^T\right\}\mathbf{B}^T\right)\mathbf{B}. \qquad (5.44)
\end{aligned}$$

where $\mathbf{y} = \mathbf{B}\mathbf{x}$ is the estimate of the independent random vector.

The above natural gradient leads to the MLE iterative algorithm to estimating the ICA transformation matrix $\mathbf{B}$:

$$\begin{aligned}
\mathbf{B} &\leftarrow \mathbf{B} + \mu\left(\mathbf{I} + \mathrm{E}\left\{\mathbf{g}\left(\mathbf{y}\right)\mathbf{x}^T\right\}\mathbf{B}^T\right)\mathbf{B} \\
&= (1+\mu)\,\mathbf{B} + \mu\mathrm{E}\left\{\mathbf{g}\left(\mathbf{y}\right)\mathbf{x}^T\right\}\mathbf{B}^T\mathbf{B} \qquad (5.45)
\end{aligned}$$

Interestingly, this algorithm can be interpreted as nonlinear decorrelation. The idea is that the algorithm converges when $\mathrm{E}\left\{\mathbf{g}\left(\mathbf{y}\right)\mathbf{x}^T\right\}\mathbf{B}^T = \mathrm{E}\left\{\mathbf{g}\left(\mathbf{y}\right)\mathbf{y}^T\right\} = -\mathbf{I}$ which means that the y_i and $g_j\left(y_j\right)$ are uncorrelated for $i \neq j$. This is a nonlinear extension of the ordinary requirement of uncorrelatedness, and is thus labelled nonlinear decorrelation.

In practice, one can use, for example, the two densities described in Section 5.4.1. For supergaussian independent components, the pdf defined by (5.34) is usually used. This

means that the component-wise nonlinearity g is the tanh function:

$$g^{+}(y) = -2\tanh(y). \tag{5.46}$$

For subgaussian independent components, other functions must be used. For example, one could use the pdf in (5.35), which leads to

$$g^{-}(y) = \tanh(y) - y. \tag{5.47}$$

The choice between the two nonlinearities in (5.46) and (5.47) can be made by computing the nonpolynomial moment (5.36) using some estimates of the independent components. If this nonpolynomial moment is positive, the nonlinearity in (5.46) should be used, otherwise the nonlinearity in (5.47) should be used. This is because of the condition in Theorem 1.

The choice of nonlinearity can be made while running the iterative gradient algorithm, using the running estimates of the independent components to estimate the nature of the independent components (that is, the sign of the nonpolynomial moment). Note that the use of the polynomial moment requires that the estimates of the independent components are first scaled properly, constraining them to unit variance, as in the theorem. Such normalizations are often omitted in practice, which may in some cases lead to situations in which the wrong nonlinearity is chosen.

5.5 ICA by Tensorial Methods

Another approach for estimation of independent components of ICA consists of using higher-order cumulant tensor. Tensors can be considered as generalization of matrices,

or linear operators. Cumulant tensors are then generalizations of the covariance matrix that was used in POD (see Chapter 4). The covariance matrix is the second-order cumulant tensor, and the fourth order tensor is defined by the fourth-order cumulants $\mathrm{cum}\,(x_i, x_j, x_k, x_l)$ given by [82]

$$\mathrm{cum}\,(x_i, x_j, x_k, x_l) = \mathrm{E}\,\{x_i x_j x_k x_l\} - \mathrm{E}\,\{x_i x_j\}\,\mathrm{E}\,\{x_k x_l\} - \mathrm{E}\,\{x_i x_k\}\,\mathrm{E}\,\{x_j x_l\} - \mathrm{E}\,\{x_i x_l\}\,\mathrm{E}\,\{x_j x_k\} \tag{5.48}$$

As explained in Section 5.2, we can use the eigenvalue decomposition of the covariance matrix to decorrelate the components of data, which is what the POD method results in. This means that we transform the data so that second-order cross-correlations of the components of data are zero. As a generalization of this principle, we can use the fourth-order cumulant tensor of the data to transform the data such that the fourth-order cross-cumulants of the transformed data are zero, or at least as small as possible. This kind of (approximative) higher-order decorrelation gives one class of methods for ICA estimation.

5.5.1 Definition of Cumulant Tensor

We shall here consider only the fourth-order cumulant tensor, which we call for brevity the cumulant tensor. The cumulant tensor is a four-dimensional array whose entries are given by the fourth-order cross-cumulants of the data: $\mathrm{cum}\,(x_i, x_j, x_k, x_l)$, where the indices i, j, k, and l are from 1 to n. This can be considered as a four-dimensional matrix, since it has four different indices instead of the usual two.

In fact, all fourth-order cumulants of linear combinations of x_i can be obtained as linear combinations of the cumulants of x_i. This can be seen using the additive properties of the

cumulants [101]. The kurtosis of a linear combination $s = \sum_i b_i x_i$ is given by

$$\begin{aligned} \text{kurt}\,(s) &= \text{kurt}\left(\sum_i b_i x_i\right) \\ &= \text{cum}\left(\sum_i b_i x_i, \sum_j b_j x_j, \sum_k b_k x_k, \sum_l b_l x_l\right) \\ &= \sum_{ijkl} b_i^4 b_j^4 b_k^4 b_l^4 \text{cum}\,(x_i, x_j, x_k, x_l)\,. \end{aligned} \tag{5.49}$$

Thus the (fourth-order) cumulants contain all the fourth-order information of the data, just as the covariance matrix gives all the second-order information on the data. Note that if the s_i are independent, all the cumulants with at least two different indices are zero, and therefore we have

$$\text{kurt}\left(\sum_i q_i s_i\right) = \sum_i q_i^4 \text{kurt}\,(s_i)\,. \tag{5.50}$$

The cumulant tensor is an linear operator defined by the fourth-order cumulants $\text{cum}\,(x_i, x_j, x_k, x_l)$. This is analogous to the case of the covariance matrix with elements $\text{cov}\,(x_i, x_j)$, which defines a linear operator just as any matrix defines one. In the case of the tensor we have a linear transformation in the space of $n \times n$ matrices, instead of the space of n-dimensional vectors. The space of such matrices is a linear space of dimension $n \times n$, so there is nothing extraordinary in defining the linear transformation. The (i, j)th element of the matrix given by the transformation, say $\mathbf{F}_{ij}\,(\mathbf{M})$ is defined as

$$\mathbf{F}_{ij}\,(\mathbf{M}) = \sum_{kl} m_{kl} \text{cum}\,(x_i, x_j, x_k, x_l) \tag{5.51}$$

where m_{kl} are the elements of the matrix $\mathbf{M}$ that is to be transformed.

5.5.2 Tensor Eigenvalues Give Independent Components

As any symmetric linear operator, the cumulant tensor has an eigenvalue decomposition (EVD). An eigenmatrix of the tensor is by definition, a matrix $\mathbf{M}$ such that

$$\mathbf{F}(\mathbf{M}) = \lambda \mathbf{M} \tag{5.52}$$

i.e., $\mathbf{F}_{ij}(\mathbf{M}) = \lambda \mathbf{M}_{ij}$, where λ is a scalar eigenvalue.

The cumulant tensor is a symmetric linear operator. This is because in the expression $\mathrm{cum}(x_i, x_j, x_k, x_l)$, the order of the variables makes no difference.

Let us consider the case where the observed data follows the ICA model, with decorrelated data:

$$\mathbf{x} = \mathbf{VAs} = \mathbf{B}^{-1}s = \mathbf{B}^{T}s \tag{5.53}$$

where we denote the decorrelated mixing matrix by $\mathbf{B}^T$. This is due to the orthogonality of $\mathbf{B}$.

The cumulant tensor of $\mathbf{x}$ has a special structure which will be discussed next. In fact, every matrix of the form

$$\mathbf{M} = \mathbf{b}_m \mathbf{b}_m^T \tag{5.54}$$

for $m = 1, \ldots, n$ is an eigenmatrix, as stated in (5.52). The vector $\mathbf{b}_m$ here is one of the rows of the matrix $\mathbf{B}$, and thus one of the columns of the decorrelation mixing matrix $\mathbf{B}^T$.

To see this, we exploit the linearity properties of cumulants in conjunction with (5.51)

to obtain

$$\begin{aligned}\mathbf{F}_{ij}\left(\mathbf{M}=\mathbf{b}_m\mathbf{b}_m^T\right) &= \sum_{kl} b_{mk}b_{ml}\mathrm{cum}\left(x_i,x_j,x_k,x_l\right)\\ &= \sum_{kl} b_{mk}b_{ml}\mathrm{cum}\left(\sum_q b_{qi}s_q,\sum_r b_{rj}s_r,\sum_u b_{uk}s_u,\sum_v b_{vl}s_v\right)\\ &= \sum_{klqruv} b_{mk}b_{ml}b_{qi}b_{rj}b_{uk}b_{vl}\mathrm{cum}\left(s_q,s_r,s_u,s_v\right) \qquad (5.55)\end{aligned}$$

Now, due to the independence of the s_i, only those cumulants where $q=r=u=v$ are nonzero. Thus we have

$$\mathbf{F}_{ij}\left(\mathbf{b}_m\mathbf{b}_m^T\right)=\sum_{klq} b_{mk}b_{ml}b_{qi}b_{qj}b_{qk}b_{ql}\mathrm{kurt}\left(s_q\right). \qquad (5.56)$$

Due to the orthogonality of the rows of $\mathbf{B}$, we have $\sum_k b_{mk}b_{qk}=\delta_{mq}$, and similarly for index l. Sequentially, we can take the sum first with respect to k, and then with respect to l, which gives

$$\begin{aligned}\mathbf{F}_{ij}\left(\mathbf{b}_m\mathbf{b}_m^T\right) &= \sum_{lq} b_{ml}b_{qi}b_{qj}b_{ql}\mathrm{kurt}\left(s_q\right)\delta_{mq}\\ &= \sum_{q} b_{qi}b_{qj}\mathrm{kurt}\left(s_q\right)\delta_{mq}\delta_{mq}\\ &= b_{mi}b_{mj}\mathrm{kurt}\left(s_m\right). \qquad (5.57)\end{aligned}$$

Therefore we have

$$\mathbf{F}\left(\mathbf{b}_m\mathbf{b}_m^T\right)=\mathrm{kurt}\left(s_m\right)\left(\mathbf{b}_m\mathbf{b}_m^T\right), \qquad (5.58)$$

which satisfies (5.52).

This proves that matrices of the form in (5.54) are eigenmatrices of the tensor. The corresponding eigenvalues are given by the kurtosis of the independent components. Moreover, it can be proved that all other eigenvalues of the tensor are zero.

Thus we see that if we knew the eigenmatrices of the cumulant tensor, we could easily obtain the independent components. If the eigenvalues of the tensor, i.e., the kurtosis of the independent components, are distinct, every eigenmatrix corresponds to a nonzero eigenvalue of the form $\mathbf{b}_m\mathbf{b}_m^T$, giving one of the columns of the decorrelated mixing matrix.

5.5.3 Computing the Tensor Decomposition by a Power Method

In principle, using tensorial methods is simple. One could take any method for computing the EVD of a symmetric matrix, and apply it on the cumulant tensor.

To do this, we must first consider viewing the tensor from the space of $n \times n$ matrices. Let q be an index that goes through all the $n \times n$ couples (i, j). Then we can consider the elements of an $n \times n$ matrix $\mathbf{M}$ as a vector using Kronecker Algebra (see Section 3.3.4). This means that we are simply vectorizing the matrices. A tensor, therefore, can be considered to be a symmetric matrix $\mathbf{F}$ with elements $f_{qq'} = \mathrm{cum}(x_i, x_j, x_{i'}, x_{j'})$, where the indices (i, j) correspond to q, and similarly (i', j') to q'. It is on this matrix that we could apply ordinary EVD algorithms, for example the well-known QR methods. The special symmetry properties of the tensor could be used to reduce the complexity (See [102] for such algorithms).

The problem with such algorithm in this category, however, is the prohibitive memory requirement, which is of $O(n^4)$ units of memory. The computational requirement also grows substantially. In addition, the case of repeating eigenvalues is also problematic.

In what follows, we discuss a simple modification of the power method, that circumvents the computational problems with the tensor EVD. In general, the power method is a simple way of computing the eigenvector corresponding to the largest eigenvalue of a matrix [103,

104]. This algorithm consists of multiplying the matrix with the successive estimate of the eigenvector, and taking the product as the new value of the vector. The vector is then normalized to unit length, and the iteration is continued until convergence is achieved. The vector then gives the desired eigenvector.

We can apply the power method quite simply to the case of the cumulant tensor. Starting from a random matrix $\mathbf{M}$, we compute $\mathbf{F}(\mathbf{M})$ and take this as the new value of $\mathbf{M}$. Then we normalize $\mathbf{M}$ and go back to the iteration step. After convergence, $\mathbf{M}$ will be of the form $\mathbf{b}\mathbf{b}^T$, where $\mathbf{b}$ is one of the independent components. In practice, though, the eigenvectors will not be exactly of this form due to estimation errors. To find several independent components, we could simply project the matrix after every step on the space of matrices that are orthogonal to the previously found ones.

In fact, in the case of ICA, such an algorithm can be considerably simplified. Since we know that the matrices $\mathbf{M} = \mathbf{b}\mathbf{b}^T$ are eigenmatrices of the cumulant tensor, we can apply the power method inside that set of matrices $\mathbf{M} = \mathbf{b}\mathbf{b}^T$ only. After every computation of the product with the tensor, we must then project the obtained matrix back to the set of matrices of the form $\mathbf{b}\mathbf{b}^T$. A very simple way of doing this is to multiply the new matrix $\mathbf{M}^*$ by the old vector to obtain the new vector $\mathbf{b}^* = \mathbf{M}^*\mathbf{b}$ (which will be normalized as necessary). This can be interpreted as another power method, this time applied on the eigenmatrix to compute its eigenvectors. Since the best way of approximating the matrix $\mathbf{M}^*$ in the space of matrices of the form $\mathbf{b}\mathbf{b}^T$ is by using the dominant eigenvector, a single step of this ordinary power method (which is performed by multiplying $\mathbf{M}^*$ with $\mathbf{b}$) will at least take us closer to the dominant eigenvector, and thus to the optimal vector.

Thus we obtain an iteration of the form

$$\mathbf{b} \leftarrow \mathbf{b}^T \mathbf{F}\left(\mathbf{b}\mathbf{b}^T\right) \tag{5.59}$$

or

$$b_i \leftarrow \sum_j b_j \sum_{kl} b_k b_l \mathrm{cum}\left(x_i, x_j, x_k, x_l\right). \tag{5.60}$$

In fact, this can be manipulated algebraically to give much simpler forms. Stated equivalently, we have

$$b_i \leftarrow \mathrm{cum}\left(x_i, \sum_j b_j x_j, \sum_k b_k x_k, \sum_l b_l x_l\right) = \mathrm{cum}\left(x_i, y, y, y\right) \tag{5.61}$$

where we denote by $y = \sum_i b_i x_i$ the estimate of an independent component s. By definition of the cumulants, we have

$$\mathrm{cum}\left(x_i, y, y, y\right) = \mathrm{E}\left\{x_i y^3\right\} - 3\mathrm{E}\left\{x_i y\right\} \mathrm{E}\left\{y^2\right\}. \tag{5.62}$$

We can constrain y to have unit variance, or $\mathrm{E}\left\{y^2\right\} = 1$, as usual. Moreover, we have $\mathrm{E}\left\{x_i y\right\} = b_i$. Thus we have

$$b_i \leftarrow \mathrm{E}\left\{x_i y^3\right\} - 3b_i \tag{5.63}$$

which in vector form is equivalent to

$$\mathbf{b} \leftarrow \mathrm{E}\left\{\mathbf{x} y^3\right\} - 3\mathbf{b} \tag{5.64}$$

where $\mathbf{b}$ is normalized to unit norm after every iteration. To find several independent components, we can actually just constrain the $\mathbf{b}$ corresponding to different independent components to be orthogonal, as is usual for decorrelated data.

Somewhat surprisingly, (5.64) is very similar to the Maximum Likelihood algorithm given in (5.45).

In the special case where the nonlinearity $\mathbf{g}(y)$ in the Maximum Likelihood algorithm given in (5.45) is chosen to be $\mathbf{g}(y) = y^3$, the Maximum Likelihood algorithm has a form very similar to that of the tensorial method in (5.64).

5.6 System Model Reduction using ICA modes versus POD modes

We have so far described the the idea of ICA as well as different algorithms that compute the ICA components. We have described in Section 4.6 how POD is applied towards achieving a reduced-order model of a system described by a set of linear differential equations. We were able to use the POD transformation matrix Σ to obtain reducd-oder system matrices.

The idea for obtaining the reduced order ICA model is the same; we would use the ICA transformation matrix **B** obtained using the algorithms discussed in Section 5.4 and Section 5.5 just as we would use the transformation matrix Σ.

The difference between the POD and ICA reduced order models lies in the characteristics of the reduced order displacement vector $\mathbf{u}_m(t)$ from Equation (4.58). We have shown that $\mathbf{u}_m(t)$, under the transformation obtained using POD, is an uncorrelated vector. ICA achieves further decoupling of the elements of $\mathbf{u}_m(t)$ by making the elements of $\mathbf{u}_m(t)$ as independent as possible being achieved by non-linear decorrelation.

We have already mentioned that our proposed system matrix identification technique (see Section 6) can estimate, in theory, the system matrices of both the full-order system (Equation (4.53)) as well as the matrices of the reduced-order system (Equation (4.58)). As well, we have described the need to estimate a reduced order model instead of a full order

model (see Section 4.6). So how do we decide between POD and ICA for model reduction? The answer lies in the nature of the system's output (for example, the displacement vector).

We expect POD to achieve decoupling of the principle components of the response vector when the system output is gaussian, i.e., when the system output is described by a vector gaussian process. In this case, POD will decorrelate the system output vector, which is the same as obtaining an independent vector (decorrelation amounts to independence in the gaussian case). Since decorrelation is the first step in ICA, ICA will do no more decoupling as the vector would be already independent (see Section 5.3). Thus, when the system's output is expected to be gaussian, one should use POD to reduce the order of the system as ICA does not achieve anything more.

It is conjectured that when the system response is non-gaussian, ICA has a superior role in weakening the inter-dependence of the response components in contrast to POD. In the context of linear systems, the system response will be nongaussian when the system input (forcing) is nongaussian and/or when the system transfer function (TF) exhibits non-gaussian behavior. This case is all the more prevalent in dealing with nonlinear systems where the usefulness of ICA still remains to be investigated.

The system of differential equations can be represented in operator form as $A(\mathbf{u}) = \mathbf{f}$ where A is the operator describing the system behavior, $\mathbf{u}$ is the system output vector, and $\mathbf{f}$ is the input vector. In that case, the system output can be expressed as $\mathbf{u} = A^{-1}(\mathbf{f})$, where A^{-1} is the inverse operator of A. If $\mathbf{f}$, the system input, is nongaussian, then $\mathbf{u}$ will also be nongaussian. Furthermore, if $\mathbf{f}$ is gaussian and A is a nonlinear operator, then $\mathbf{u}$ will also be nongaussian.

In those two cases when the system output is nongaussian, ICA is believed to achieve

stronger decoupling as POD decorrelates the system output whereas ICA makes them as independent as possible.

Chapter 6

System Identification

In this chapter, we will present a system matrix identification method which capitalizes on the use of Kronecker Algebra. As will be shown, the use of the Kronecker Algebra and its properties permits a unique and simplified approach to estimating the system matrices from the knowledge of its output.

6.1 Proposed Methodology

We will consider the system of equations describing the forced vibration of a viscously damped linear discrete system with m degrees of freedom

$$\mathbf{M}_m\ddot{\mathbf{u}}(t) + \mathbf{C}_m\dot{\mathbf{u}}(t) + \mathbf{K}_m\mathbf{u}(t) = \mathbf{f}(t) \tag{6.1}$$

where $\mathbf{M}_m \in \mathbb{R}^{m\times m}$ is the mass matrix, $\mathbf{C}_m \in \mathbb{R}^{m\times m}$ is the damping matrix, $\mathbf{K}_m \in \mathbb{R}^{m\times m}$ is the stiffness matrix, $\mathbf{u}(t) \in \mathbb{R}^m$ is the displacement vector, and $\mathbf{f}(t) \in \mathbb{R}^m$ is the forcing vector at time t.

Transforming Equation (6.1) into the frequency domain, one obtains

$$\left[-\omega^2\mathbf{M}_m + \mathrm{i}\omega\mathbf{C}_m + \mathbf{K}_m\right]\mathbf{Q}_m(\omega) = \mathbf{F}_m(\omega) \tag{6.2}$$

where $\mathbf{Q}_m(\omega) \in \mathbb{C}^m$ and $\mathbf{F}_m(\omega) \in \mathbb{C}^m$ are the Fourier transforms of $\mathbf{q}_m(t)$ and $\mathbf{f}_m(t)$, respectively.

For a specific frequency ω_i, we can rewrite the left hand side of (6.2) as

$$\begin{aligned}\left[-\omega_i^2\mathbf{M}_m + \mathrm{i}\omega_i\mathbf{C}_m + \mathbf{K}_m\right]\mathbf{Q}_m(\omega_i) &= \left[-\omega_i^2\mathbf{I}_m\mathbf{M}_m + \mathrm{i}\omega_i\mathbf{I}_m\mathbf{C}_m + \mathbf{I}_m\mathbf{K}_m\right]\mathbf{Q}_m(\omega_i) \\ &= \begin{bmatrix} -\omega_i^2\mathbf{I}_m & \mathrm{i}\omega_i\mathbf{I}_m & \mathbf{I}_m \end{bmatrix}\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\mathbf{Q}_m(\omega_i),\end{aligned} \tag{6.3}$$

where $\mathbf{I}_m$ is the identity matrix of order m.

Thus for a specific frequency ω_i, the system of equations governing the system behavior can be written as

$$\begin{bmatrix} -\omega_i^2\mathbf{I}_m & \mathrm{i}\omega_i\mathbf{I}_m & \mathbf{I}_m \end{bmatrix}\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\mathbf{Q}_m(\omega_i) = \mathbf{F}_m(\omega_i). \tag{6.4}$$

Applying the vec operator (see Section 3.3.4) on both sides of (6.4) and using the identity $\mathrm{vec}\,(\mathbf{AYB}) = \left(\mathbf{B}^T \otimes \mathbf{A}\right)\mathrm{vec}\,(\mathbf{Y})$ (see Section 3.3.6), we obtain

$$\left(\mathbf{Q}_m(\omega_i)^T \otimes \begin{bmatrix} -\omega_i^2\mathbf{I}_m & \mathrm{i}\omega_i\mathbf{I}_m & \mathbf{I}_m \end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\right) = \mathrm{vec}\left(\mathbf{F}_m(\omega_i)\right) = \mathbf{F}_m(\omega_i), \tag{6.5}$$

where $\otimes$ denotes the Kronecker matrix product, or tensor product.

Equation (6.5) has three unknowns, namely the mass, damping, and stiffness matrices. Having measured the system response vector at J different frequencies in the frequency

band of interest, we can rewrite the generalized form of Equation (6.5) as follows

$$\begin{bmatrix} \mathbf{Q}_m(\omega_1)^T \otimes \begin{bmatrix} -\omega_1^2\mathbf{I}_m & \mathrm{i}\omega_1\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \\ \mathbf{Q}_m(\omega_2)^T \otimes \begin{bmatrix} -\omega_2^2\mathbf{I}_m & \mathrm{i}\omega_2\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \\ \vdots \\ \mathbf{Q}_m(\omega_J)^T \otimes \begin{bmatrix} -\omega_J^2\mathbf{I}_m & \mathrm{i}\omega_J\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \end{bmatrix} \mathrm{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix} \right) = \begin{Bmatrix} \mathbf{F}_m(\omega_1) \\ \mathbf{F}_m(\omega_2) \\ \vdots \\ \mathbf{F}_m(\omega_J) \end{Bmatrix}. \tag{6.6}$$

Equation (6.6) can be written as [105]

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{6.7}$$

where

$$\mathbf{x} = \mathrm{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix} \right) \in \mathbb{R}^{3m^2}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q}_m(\omega_1)^T \otimes \begin{bmatrix} -\omega_1^2\mathbf{I}_m & \mathrm{i}\omega_1\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \\ \mathbf{Q}_m(\omega_2)^T \otimes \begin{bmatrix} -\omega_2^2\mathbf{I}_m & \mathrm{i}\omega_2\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \\ \vdots \\ \mathbf{Q}_m(\omega_J)^T \otimes \begin{bmatrix} -\omega_J^2\mathbf{I}_m & \mathrm{i}\omega_J\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \end{bmatrix} \in \mathbb{C}^{mJ\times 3m^2} \tag{6.8}$$

$$\mathbf{y} = \begin{Bmatrix} \mathbf{F}_m(\omega_1) \\ \mathbf{F}_m(\omega_2) \\ \vdots \\ \mathbf{F}_m(\omega_J) \end{Bmatrix} \in \mathbb{C}^{mJ} \tag{6.9}$$

The above system of equations (6.6) is overdetermined in the case where $J > 3m$. The vectorized equivalent $\mathbf{x}$ containing the mass, damping, and stiffness matrices can be solved in the least-square sense using the least-square inverse of the matrix $\mathbf{A}$, as follows

$$\widehat{\mathbf{x}} = \left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T\mathbf{y}. \tag{6.10}$$

where $\widehat{\mathbf{x}}$ is the least-square estimate of $\mathbf{x}$ and $\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T$ is the least-square inverse of matrix $\mathbf{A}$, also known as the Moore-Penrose inverse of a matrix [106].

The least-square estimate has some strong statistical properties. Under certain conditions, the least-square estimate is the *best unbiased linear estimate* often denoted with the acronym BLUE [107].

6.2 Tikhonov Regularization

6.2.1 Definition

In the case of overdetermined or underdetermined system of equations (6.7), the estimated vector $\widehat{\mathbf{x}}$ is acceptable if the matrix-vector product $\mathbf{A}\widehat{\mathbf{x}}$ is close to $\mathbf{y}$. In the preceeding section, the estimate of $\mathbf{x}$ is obtained in the least-squares sense. One quantity for measuring the accuracy of the estimated $\widehat{\mathbf{x}}$ is the L_2-norm of the residual vector $\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}$ given by

$$C\left(\widehat{\mathbf{x}}\right) = \left\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}\right\| = \left(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}\right)^T\left(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}\right). \tag{6.11}$$

In the case of matrix $\mathbf{A}$ having rank less than $3m^2$, there exist one or more zero singular values of $\mathbf{A}$. The least-square solution vector $\widehat{\mathbf{x}}$ will have two components. One component lies in the subspace spanned by the singular vectors of $\mathbf{A}$ corresponding to nonzero singular values. The other non-zero component exists in the subspace spanned by the singular vectors with zero singular values. Only the first component can be reasonably estimated from the data set $\mathbf{y}$.

Therefore, there is a need to include additional information which permits the estimation of the component of $\mathbf{x}$ that lies in the null-space of $\mathbf{A}$. One approach to solving this problem

is to introduce another norm $D(\widehat{\mathbf{x}})$ that measures the error between $\widehat{\mathbf{x}}$ and some default solution $\mathbf{x}^{\infty}$ where $\mathbf{x}^{\infty}$ may be some prior information about $\mathbf{x}$. Thus, $D(\widehat{\mathbf{x}})$ has the form

$$D(\widehat{\mathbf{x}}) = \|\widehat{\mathbf{x}} - \mathbf{x}^{\infty}\|. \tag{6.12}$$

More generally, we try to estimate the result of a linear operator $\mathbf{L}$ in the form of a matrix acting on the difference $(\widehat{\mathbf{x}} - \mathbf{x}^{\infty})$. In that case, we have

$$D(\widehat{\mathbf{x}}) = \|\mathbf{L}(\widehat{\mathbf{x}} - \mathbf{x}^{\infty})\| = (\widehat{\mathbf{x}} - \mathbf{x}^{\infty})^{T}\mathbf{L}^{T}\mathbf{L}(\widehat{\mathbf{x}} - \mathbf{x}^{\infty}). \tag{6.13}$$

A well-known regularization technique is to form a weighted sum of $C(\mathbf{x})$ and $D(\mathbf{x})$ using a weighting factor λ^2. The estimate $\widehat{\mathbf{x}}$ is the value of $\mathbf{x}$ that minimizes this sum:

$$\widehat{\mathbf{x}} = \arg\min\left\{C(\mathbf{x}) + \lambda^2 D(\mathbf{x})\right\}. \tag{6.14}$$

The solution to (6.14) is obtained by setting the vector-gradient with respect to $\mathbf{x}$ equal

to zero:

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{x}} \left\{ C(\mathbf{x}) + \lambda^2 D(\mathbf{x}) \right\} \Big|_{\mathbf{x}=\widehat{\mathbf{x}}} \\
&= \frac{\partial}{\partial \mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{y}\| + \lambda^2 \|\mathbf{L}(\mathbf{x} - \mathbf{x}^{\infty})\| \right\} \Big|_{\mathbf{x}=\widehat{\mathbf{x}}} \\
&= \frac{\partial}{\partial \mathbf{x}} \left\{ (\mathbf{A}\mathbf{x} - \mathbf{y})^T (\mathbf{A}\mathbf{x} - \mathbf{y}) + \lambda^2 (\mathbf{x} - \mathbf{x}^{\infty})^T \mathbf{L}^T \mathbf{L} (\mathbf{x} - \mathbf{x}^{\infty}) \right\} \Big|_{\mathbf{x}=\widehat{\mathbf{x}}} \\
&= \frac{\partial}{\partial \mathbf{x}} \left\{ \left(\mathbf{x}^T \mathbf{A}^T - \mathbf{y}^T\right) (\mathbf{A}\mathbf{x} - \mathbf{y}) + \lambda^2 \left(\mathbf{x}^T - (\mathbf{x}^{\infty})^T\right) \mathbf{L}^T \mathbf{L} (\mathbf{x} - \mathbf{x}^{\infty}) \right\} \Big|_{\mathbf{x}=\widehat{\mathbf{x}}} \\
&= \frac{\partial}{\partial \mathbf{x}} \left\{ \begin{array}{c} \left[\mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x} - \mathbf{y}^T \mathbf{A}\mathbf{x} - \mathbf{x}^T \mathbf{A}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}\right] + \\ \lambda^2 \left[\mathbf{x}^T \mathbf{L}^T \mathbf{L}\mathbf{x} - (\mathbf{x}^{\infty})^T \mathbf{L}^T \mathbf{L}\mathbf{x} - \mathbf{x}^T \mathbf{L}^T \mathbf{L}\mathbf{x}^{\infty} + (\mathbf{x}^{\infty})^T \mathbf{x}^{\infty}\right] \end{array} \right\} \Big|_{\mathbf{x}=\widehat{\mathbf{x}}} \\
&= \left\{ \begin{array}{c} \left[2\mathbf{A}^T \mathbf{A}\mathbf{x} - \mathbf{A}^T \mathbf{y} - \mathbf{A}^T \mathbf{y} + \mathbf{0}_n\right] + \\ \lambda^2 \left[2\mathbf{L}^T \mathbf{L}\mathbf{x} - \mathbf{L}^T \mathbf{L}\mathbf{x}^{\infty} - \mathbf{L}^T \mathbf{L}\mathbf{x}^{\infty} + \mathbf{0}_n\right] \end{array} \right\} \Big|_{\mathbf{x}=\widehat{\mathbf{x}}} \\
&= 2\left(\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{L}^T \mathbf{L}\right) \widehat{\mathbf{x}} - 2\left(\mathbf{A}^T \mathbf{y} + \lambda^2 \mathbf{L}^T \mathbf{L}\mathbf{x}^{\infty}\right) \\
&= \mathbf{0}_n. \qquad (6.15)
\end{aligned}
$$

where $\mathbf{0}_n$ is the zero vector of order n.

Solving for $\widehat{\mathbf{x}}$ in (6.15), we have

$$\widehat{\mathbf{x}} = \left(\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{L}^T \mathbf{L}\right)^{-1} \left(\mathbf{A}^T \mathbf{y} + \lambda^2 \mathbf{L}^T \mathbf{L}\mathbf{x}^{\infty}\right). \qquad (6.16)$$

The above formulation leads to a family of solutions parametrized by the weighing factors λ, popularly known as the *regularization parameter* [2]. If the regularization parameter is very large, the constraint involving the observed data $\mathbf{y}$ weakly influences the solution $\widehat{\mathbf{x}}$ and the estimate of $\mathbf{x}$ is heavily influenced by the constraint $\mathbf{L}\mathbf{x} = \mathbf{L}\mathbf{x}^{\infty}$. On the other hand, if λ is chosen to be small, the solutions depends more heavily on the observed data. Of course, if λ is set to zero, the problem reverts back to solving Equation (6.7), posed as

an unconstrained optimization problem. Thus, the value for the regularization parameters is chosen depending on how strongly one would like to enforce the constraint $\mathbf{Lx} = \mathbf{Lx}^{\infty}$. This regularization method is generally known as Tikhonov Regularization [2].

6.2.2 Determining the Regularization Parameter Value

The difficulty in applying Tikhonov Regularization is to choose a good value of λ that regularizes the solution without losing too much information. There are three methods to determine the optimal value for λ (see [108, 109, 110, 111]):

1. Discrepancy Principle

   Choose λ such that the two terms in (6.14) contribute equally to the total error, i.e.

   $$C(\widehat{\mathbf{x}}) = \lambda^2 D(\widehat{\mathbf{x}}). \tag{6.17}$$

2. Generalized cross-validation

   Choose λ as to minimize the function

   $$G(\widehat{\mathbf{x}}) = \frac{\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}\|}{\mathrm{tr}\left(\mathbf{I} - \mathbf{A}\left(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{I}\right)^{-1}\mathbf{A}^T\right)}. \tag{6.18}$$

3. L-curve criterion

   Choose λ that corresponds to the "corner" of the curve $\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{y}\|$ versus $\|\mathbf{L}(\widehat{\mathbf{x}} - \mathbf{x}^{\infty})\|$, plotted in log-log scale.

6.2.3 Application to System Identification

In order to estimate the component of $\mathbf{x}$ that lies in the null-space of $\mathbf{A}$, we need to exploit some prior information concerning the system. For example, we can decide to choose

the solution that gives rise to an estimated mass, stiffness, and damping matrices that are symmetric, or as symmetric as possible, assuring that the underlying continuous operator is self-adjoint. Mathematically, this problem can be posed as a constrained optimization problem.

In order to satisfy symmetry, for instance, in the mass matrix $\mathbf{M}_m$, we need to have

$$\mathbf{M}_m = \mathbf{M}_m^T. \tag{6.19}$$

$\mathbf{M}_m$ can be rewritten as

$$\begin{aligned} \mathbf{M}_m &= \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix} \begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix} \begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix} \mathbf{I}_m. \end{aligned} \tag{6.20}$$

where $\mathbf{I}_m$ is the identity matrix of order m and $\mathbf{0}_{m\times m}$ is the square matrix of order $m \times m$ whose elements are zero.

Similarly, $\mathbf{M}_m^T$ can be rewritten as

$$\begin{aligned}
\mathbf{M}_m^T &= \begin{bmatrix} \mathbf{M}_m^T & \mathbf{C}_m^T & \mathbf{K}_m^T \end{bmatrix} \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}^T \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} \end{bmatrix} \\
&= \mathbf{I}_m \begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}^T \begin{bmatrix} \mathbf{I}_m \\ \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} \end{bmatrix}
\end{aligned} \tag{6.21}$$

The symmetry constraint (6.19) can be written as

$$\mathbf{M}_m - \mathbf{M}_m^T = \mathbf{0}_{m\times m}. \tag{6.22}$$

Applying the vec operator and Identities (6.20) and (6.21) and (3.95), we obtain as

$$\begin{aligned}
& \operatorname{vec}\left(\mathbf{M}_m - \mathbf{M}_m^T\right) = \operatorname{vec}\left(\mathbf{0}_{m\times m}\right) \qquad (6.23)\\
\Rightarrow\ & \operatorname{vec}\left(\mathbf{M}_m\right) - \operatorname{vec}\left(\mathbf{M}_m^T\right) = 0_{m^2}\\
\Rightarrow\ & \left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix}\right) \operatorname{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\right)\\
& - \left(\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix} \otimes \mathbf{I}_m\right) \operatorname{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}^T\right) = 0_{m^2}\\
\Rightarrow\ & \left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix}\right) \operatorname{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\right)\\
& - \left(\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix} \otimes \mathbf{I}_m\right) \mathbf{U}\operatorname{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\right) = 0_{m^2}\\
\Rightarrow\ & \left\{ \begin{array}{l} \left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix}\right) - \\ \left(\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix} \otimes \mathbf{I}_m\right) \mathbf{U} \end{array} \right\} \operatorname{vec}\left(\begin{bmatrix} \mathbf{M}_m \\ \mathbf{C}_m \\ \mathbf{K}_m \end{bmatrix}\right) = 0_{m^2}\\
\Rightarrow\ & \mathbf{L}_M\mathbf{x} = 0_{m^2},
\end{aligned}$$

where $\mathbf{U}$ is the vec-permutation matrix (3.101), 0_{m^2} is the zero vector of order m^2, and

$$\mathbf{L}_M = \left[\left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix}\right) - \left(\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} \end{bmatrix} \otimes \mathbf{I}_m\right) \mathbf{U}\right].$$

Therefore, symmetry condition in the mass matrix gives rise to the constraint equation:

$$\mathbf{L}_M\mathbf{x} = 0_{m^2} \qquad (6.24)$$

where the subscript in $\mathbf{L}_M$ indicates that the constraint is on the mass matrix.

A symmetry condition on the damping matrix $\mathbf{C}_m$ similarly leads to

$$\mathbf{L}_C\mathbf{x} = 0_{m^2} \tag{6.25}$$

where

$$\mathbf{L}_C = \left[\left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{I}_m & \mathbf{0}_{m\times m} \end{bmatrix}\right) - \left(\begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{I}_m & \mathbf{0}_{m\times m} \end{bmatrix} \otimes \mathbf{I}_m\right)\mathbf{U}\right].$$

Similarly, a symmetry property of the stiffness matrix can be achieved using the constraint equation

$$\mathbf{L}_K\mathbf{x} = 0_{m^2} \tag{6.26}$$

where

$$\mathbf{L}_K = \left[\left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{I}_m \end{bmatrix}\right) - \left(\begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \mathbf{I}_m \end{bmatrix} \otimes \mathbf{I}_m\right)\mathbf{U}\right].$$

Applying Tikhonov Regularization to estimate $\mathbf{x}$, we obtain the following solution

$$\widehat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A} + \lambda_M^2\mathbf{L}_M^T\mathbf{L}_M + \lambda_C^2\mathbf{L}_C^T\mathbf{L}_C + \lambda_K^2\mathbf{L}_K^T\mathbf{L}_K\right)^{-1}\left(\mathbf{A}^T\mathbf{y}\right). \tag{6.27}$$

The above solution depends on the values chosen for the regularization parameters λ_M, λ_C and λ_K. If the regularization parameters are very large, the constraint enforcing the symmetry condition predominates in the solution of $\mathbf{x}$. On the other hand, if λ values are chosen to be small, the symmetry constraint is less satisfied and the solutions depends more heavily on the observed data. The values for the regularization parameters are chosen using any of the methods discussed above (Section 6.2.2).

Chapter 7

Numerical Validation

In this chapter, we provide numerical a demonstration of the proposed theoretical formulation using a simple discrete dynamical system. The proposed system identification method identifies the reduced order model of the system. As stated previously, the identification performed on the reduced-order model significantly reduces the computational requisite in contrast to the case involving the comprehensive system. The model reduction is carried out using both POD and ICA methods. The efficiency and similarity of the POD and ICA approaches are demonstrated using numerical examples. The sensitivity of the identification step with respect to measurement noise is also investigated.

7.1 Homogeneous Discrete Linear System

In this section, a homogeneous proportionally damped linear array of mass-spring oscillators is considered to be the prototype system. Such discrete model normally arises from the Finite Element Method (FEM) discretization of a one-dimensional partial differential equa-

tion. Such simple model is well-suited to investigate the usefulness and feasibility of the proposed methodology without the undue computational complexity involved in dealing with a complex physical system.

7.1.1 Original System Model

The schematic diagram of the system is shown in Figure 7.1.

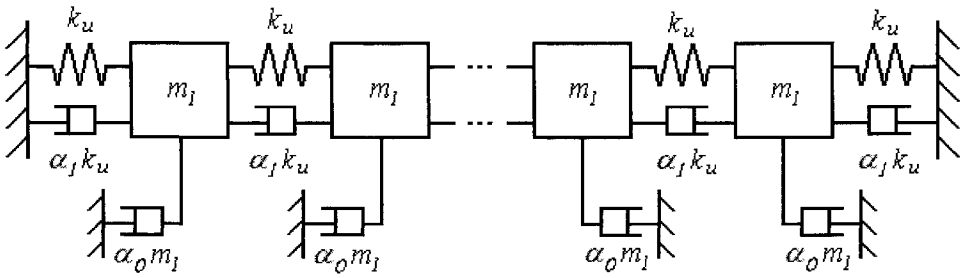

Figure 7.1: Linear array of mass-spring oscillators.

The mass matrix of the system has the form

$$\mathbf{M}_n = m_1 \mathbf{I}_n \tag{7.1}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. n and m_1 are chosen to be 100 DOFs and 1kg, respectively. The stiffness matrix of the system is given by

$$\mathbf{K}_n = k_u \begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & -1 & 2 & -1 & & \\ & & & \ddots & \ddots & -1 & \\ & & & & -1 & 2 & \end{bmatrix}, \tag{7.2}$$

with $k_u = 4 \times 10^5$ N/m.

Energy dissipation in the system is modelled by Rayleigh damping given by $\mathbf{C}_n = \alpha_0 \mathbf{M}_n + \alpha_1 \mathbf{K}_n$, where $\alpha_0 = 0.5$ and $\alpha_1 = 3 \times 10^{-5}$.

For a preliminary illustration, we consider the computer-simulated FRFs of this system as if they were experimentally measured. A typical FRF of the system is shown in Figure 7.2.

In the same figure, the frequency range considered for the construction of the POD as well as ICA is also shown.

7.1.2 Reduced Order Model: Forward Simulation

Under an input that is band-limited independent white noise with unit variance, the system response is calculated and the POD eigenvectors are extracted from the correlation matrix. Normalized eigenvalues of the correlation matrix, that is $\lambda/\lambda_{\max}$, are shown in Figure 7.3.

From Figure 7.3, note that only the first few eigenvalues are significantly large as is expected from the scree test (see Section 4.5.2). This justifies the approximation in Equation (4.45). A typical FRF of the POD-based reduced system is compared with the original FRF

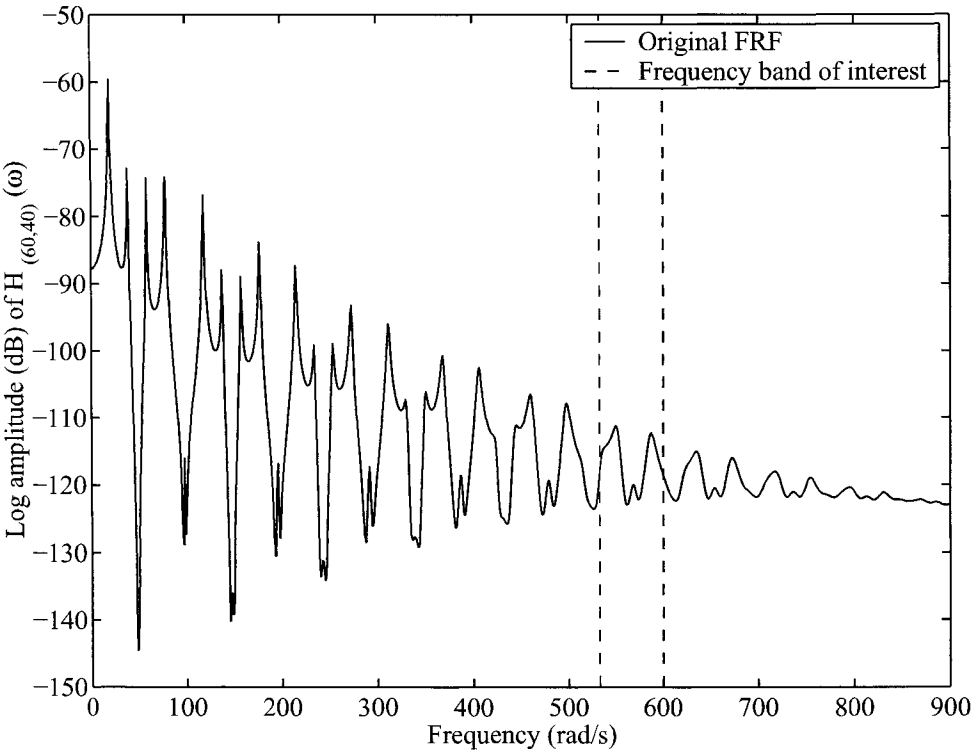

Figure 7.2: A typical FRF of the original homogeneous system and the frequency band of interest.

in Figure 7.4.

With only 16 POD modes, the FRF from the POD-based Reduced-Order Model (ROM) matches reasonably well with the original FRF in the frequency band of interest. The FRF of the POD-ROM does not match the FRF of the original system outside the frequency band of interest. This is expected; the POD was carried out within the frequency band, thus it fails to capture the system behavior outside this band. There are a number of factors that influence the accuracy of the reconstructed FRFs, for example (a) number of POD modes to retain, (b) level of damping, (c) the size and position of frequency window for the construction of POD.

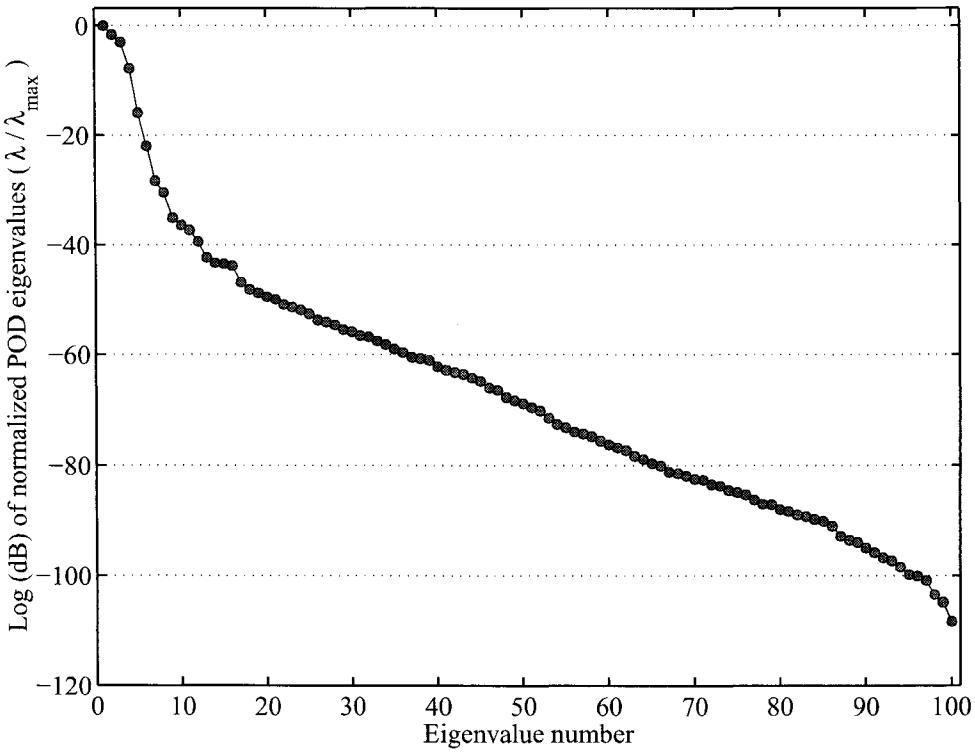

Figure 7.3: Normalized POD eigenvalues $(\lambda/\lambda_{\max})$ for the homogeneous system.

To compare POD with ICA, the FRF obtained from the ICA reduced model is compared with the original FRF as well as the POD reduced model FRF in Figure 7.5.

Again, with 16 ICA modes, the ICA reduced order model FRF is the same as that of the POD reduced order model. This might seem surprising at first, but as was mentioned in Section 5.3, in the case of gaussian components, all ICA can do is to decorrelate the data, which is already achieved by POD (POD being the first step in ICA). Seemingly gaussian data for the system response explains why POD and ICA produce the same results.

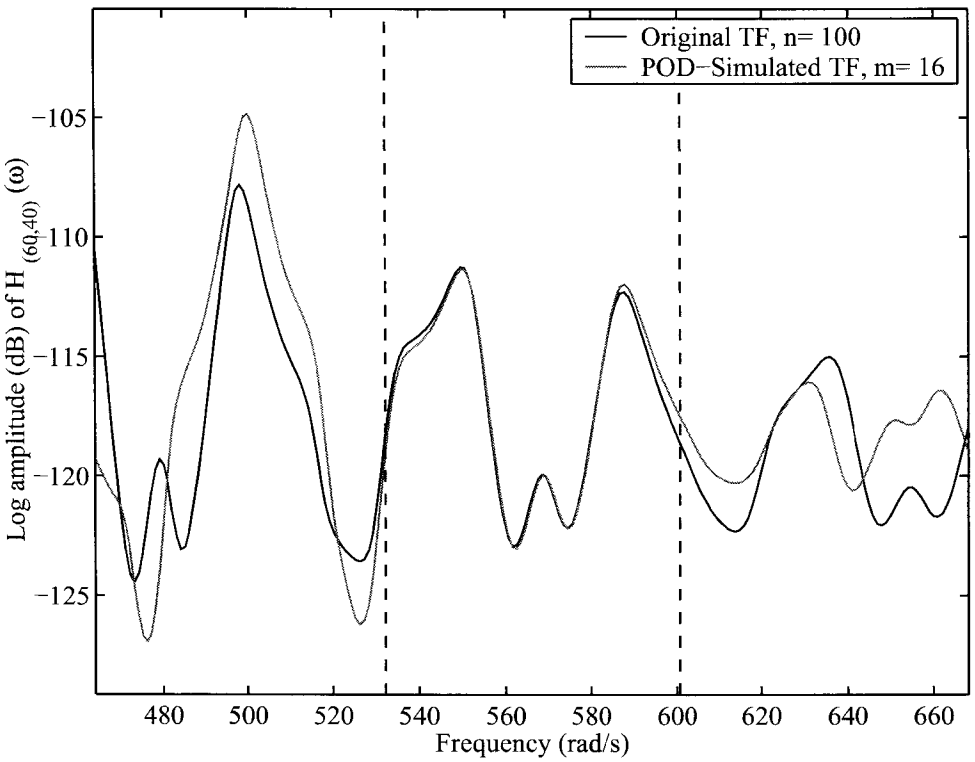

Figure 7.4: Original and POD reconstructed FRF for the homogeneous system.

7.1.3 Reduced Order Model Identification: Noiseless Case

Next, we apply the proposed system identification method discussed in Chapter 6. To simplify matters, we assume that the system response is obtained without noise. The case of noise-contaminated output will be examined subsequently. We first obtain the POD and ICA transformation matrix, choosing the number of modes to be 16 for each method. Consequently, we obtain the identified POD and ICA reduced order system matrices. The identified matrices are then used to obtain a typical FRF of the system. This FRF is compared to the original model FRF, as shown in Figure 7.6.

The symmetry constraint was applied in the identification process, with the value for

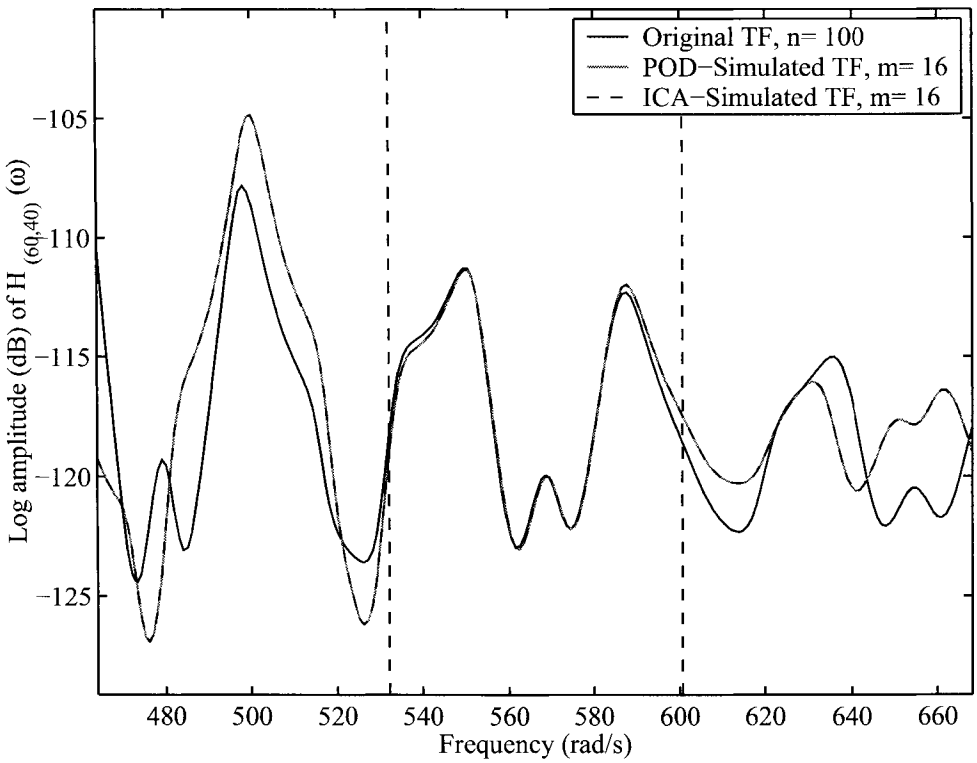

Figure 7.5: Original, ICA, and POD reconstructed FRF for the homogeneous system.

the regularization parameters, λ_M, λ_C and λ_K all being 100. We can see that the identified system matrices result in FRFs that match reasonably well the original FRF. Both ICA and POD result in the same FRF. This can be explained by the apparent gaussian nature of the system response.

Borrowing the idea of Signal-to Noise (SNR) ratio from telecommunication theory [112], a measure of the error between the identified and original FRF is

$$\text{error} = 10\log_{10}\frac{\|\|H_{orig}(\omega)\|\|}{\|\|H_{orig}(\omega)| - |H_{iden}(\omega)\|\|}\text{dB}. \tag{7.3}$$

where dB is the decibel unit, $\|\cdot\|$ denotes the norm, $|\cdot|$ the absolute value, $H_{orig}(\omega)$ the original model FRF, and $H_{iden}(\omega)$ the identified model FRF. Note that the higher the value

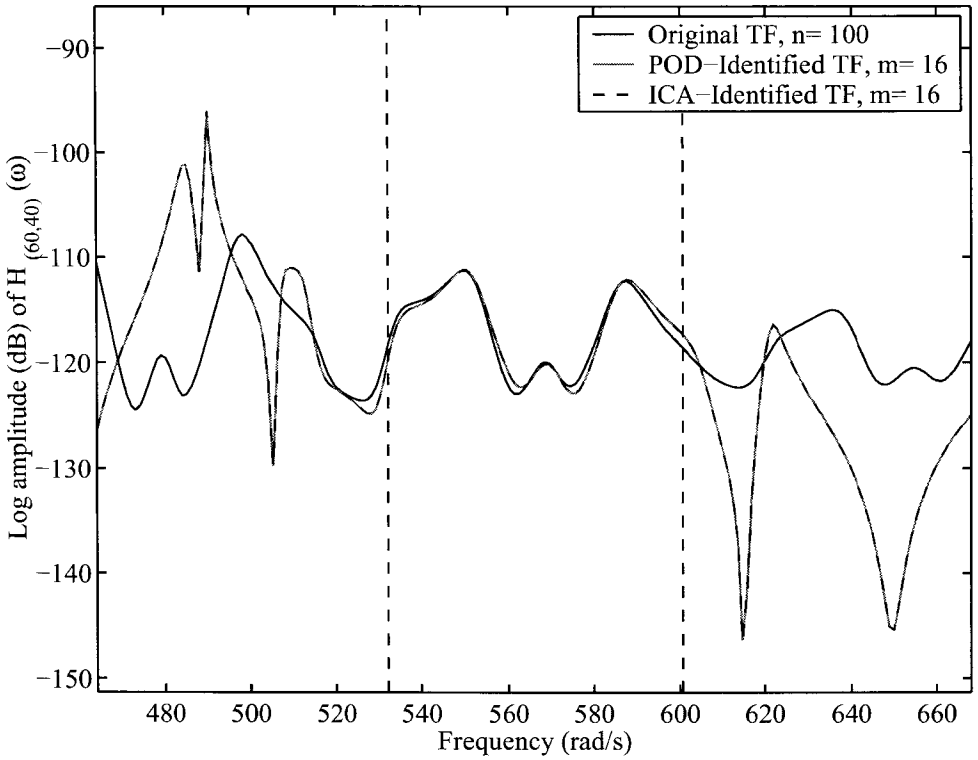

Figure 7.6: Original, ICA, and POD identified FRF for the homogeneous system.

of this measure, the less the error between the amplitudes of the original and identified FRF. For the identified FRF in Figure 7.6, the error is calculated to be 23.95 dB.

It is important to note that the POD and ICA modes are not exactly the same. Even though the system response is gaussian resulting in the same FRFs of the ICA and POD reduced order models, the ICA algorithm achieves minor rotations of the POD modes in the nth-dimension resulting in slightly different modes for ICA. The first mode obtained from ICA is compared to that obtained from POD in Figure 7.7.

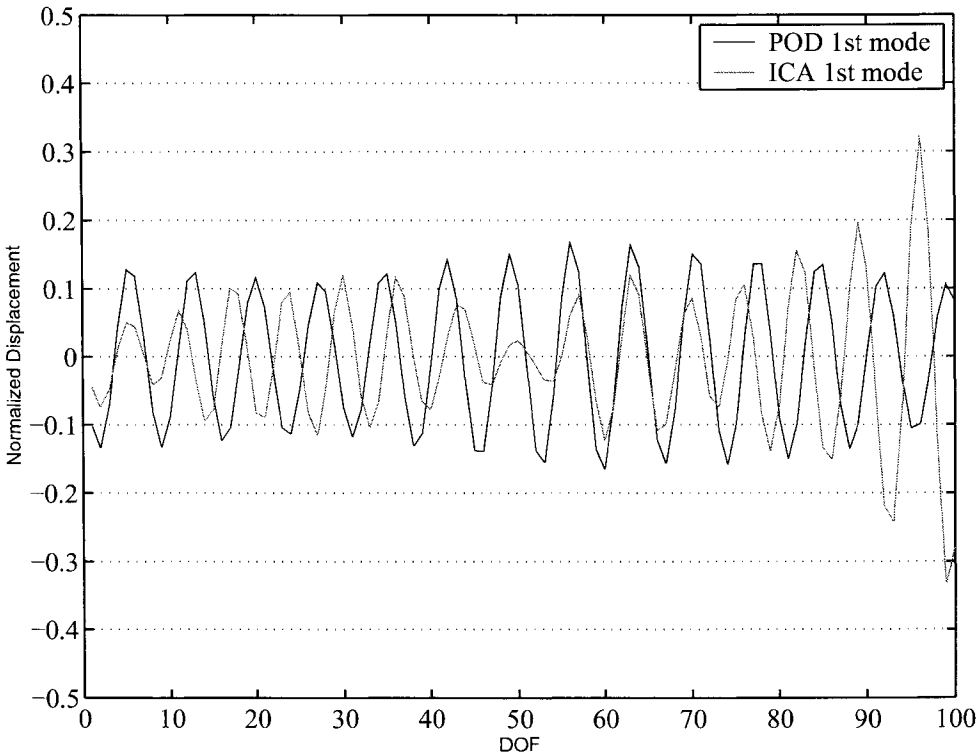

Figure 7.7: POD and ICA 1st mode for the homogeneous system.

7.1.4 Reduced Order Model Identification: Noisy Case

We have demonstrated the efficacy of the proposed method assuming that our data is not contaminated by noise. In the low frequency range, the signal-to-noise ratio (SNR) is sufficiently high that our assumption is reasonable. However, in the mid-frequency range, the SNR decreases significantly that the assumption no longer holds. The presence of noise obviously affects the confidence level of the estimated parameters.

We apply the same methodology described in the section above, with the added step of contaminating the system response vector with uncorrelated gaussian band-limited white noise to achieve a specific Signal-to-Noise (SNR) ratio prior to the identification process.

Let σ_n^2 and σ_s^2 denote the variance of the noise and the original signal, respectively. The SNR is calculated using

$$\text{SNR} = 10 \log_{10} \frac{\sigma_s^2}{\sigma_n^2} \text{dB}. \tag{7.4}$$

Thus, an SNR of 10 implies that the variance of the signal is 10 times higher than the variance of the contaminating noise. An SNR of 20 implies that the variance of the signal is 100 times the variance of the contaminating noise, and so on.

With an SNR of 30 dB, we obtain the identified FRF shown in Figure 7.8 with an error of 21.14 dB. Compared to the noiseless case in which the error in the identified FRF was 23.95 dB, a noise level of 30 dB SNR does not affect the identification process significantly.

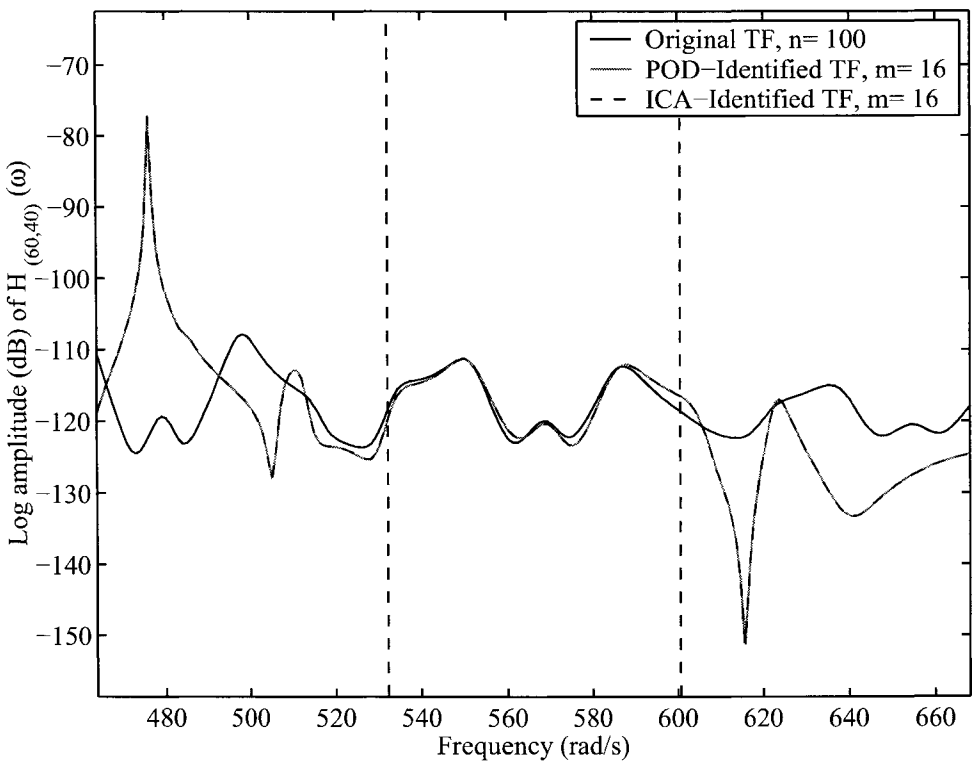

Figure 7.8: Original, ICA, and POD identified FRF for the homogeneous system under a noise level of 30 dB SNR.

With an SNR of 20 dB, we obtain the identified FRF shown in Figure 7.9 with an error of 7.21 dB. Compared to a noise level of 30 dB case in which the error in the identified FRF was 23.95 dB, a noise level of 20 dB SNR does affect the identification process, but the identified FRF matches the original FRF reasonably well in most of the bandwidth of interest.

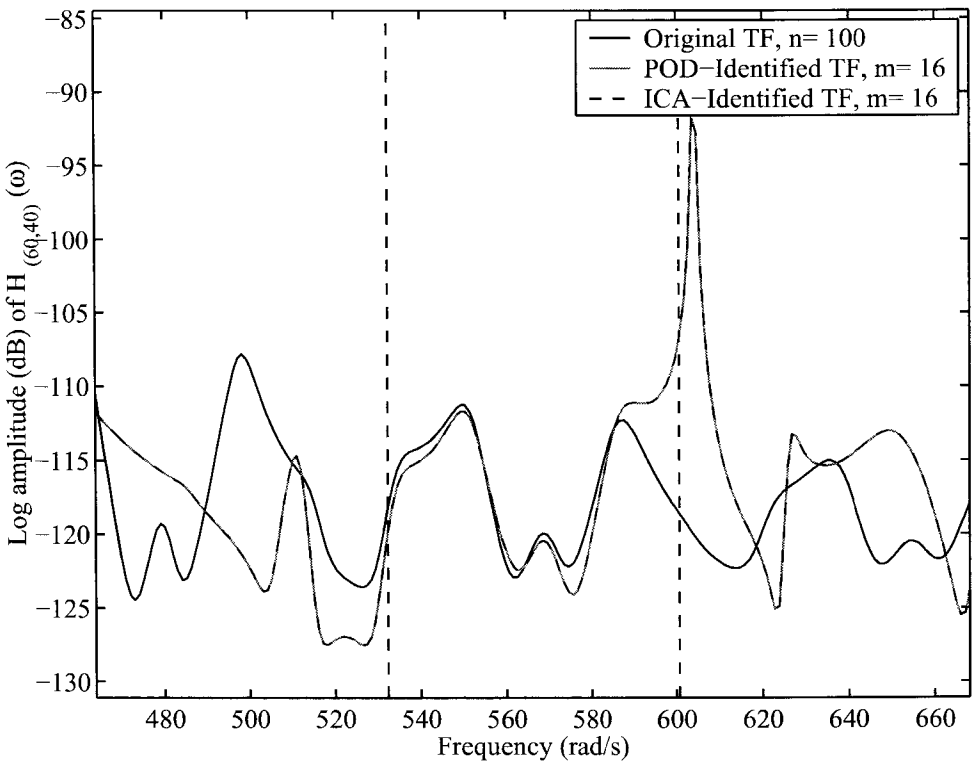

Figure 7.9: Original, ICA, and POD identified FRF for the homogeneous system under a noise level of 20 dB SNR.

Lastly, with an SNR of 10 dB, we obtain the identified FRF shown in Figure 7.10 with an error of 9.68 dB. Compared to the error of 7.21 dB for a noise level of 20 dB SNR, it seems strange that the identified FRF has a smaller error (higher decibel value), but by examining the FRF in Figure 7.9 and Figure 7.10, it is obvious that a stronger noise level

of 10 dB SNR results in a worse estimate of the FRF.

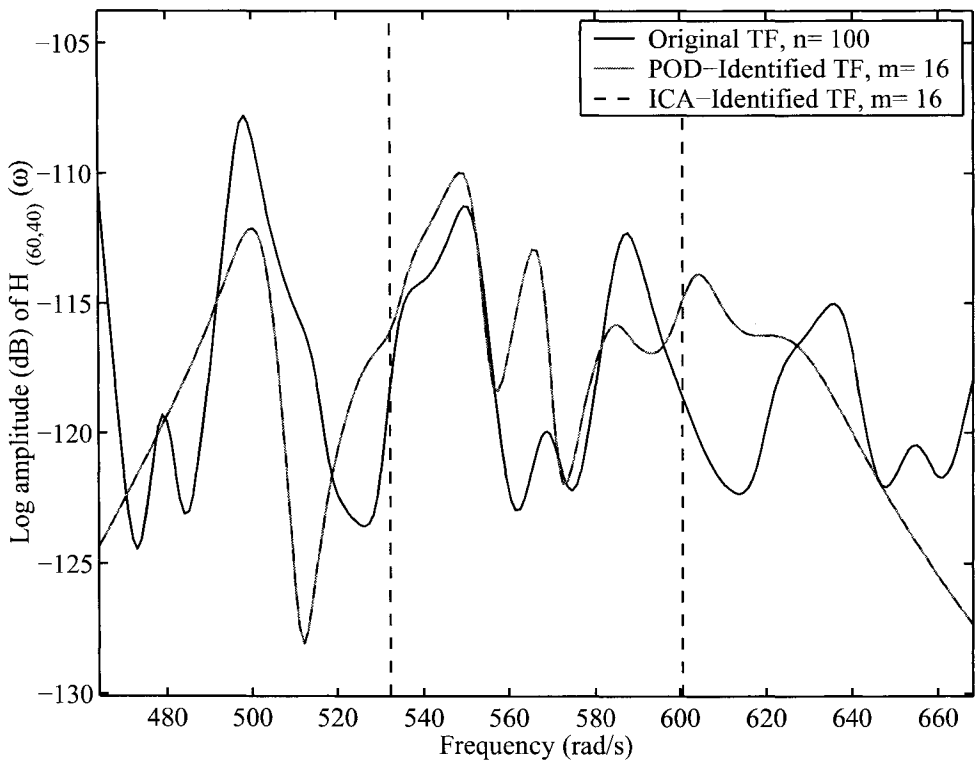

Figure 7.10: Original, ICA, and POD identified FRF for the homogeneous system under a noise level of 10 dB SNR.

7.2 Coupled Discrete Linear System

In this section, a coupled linear array of mass-spring oscillators is considered to be the original system. A lighter system is coupled with a heavier system. In other words, the lighter system posses higher modal densities compared to the heavier system.

7.2.1 Original System Model

The schematic of the system is shown in Figure 7.11.

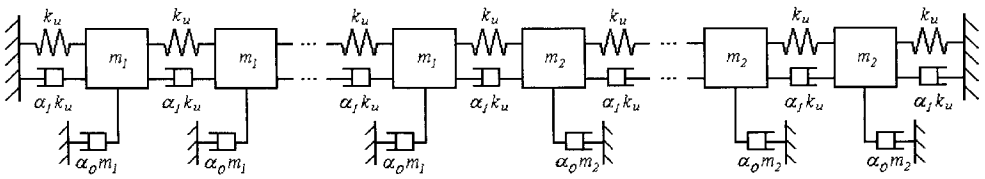

Figure 7.11: Linear array of mass-spring oscillators.

The mass matrix of the system has the form

$$\mathbf{M}_n = \begin{bmatrix} m_1\mathbf{I}_{n/2} & \mathbf{0}_{n/2} \\ \mathbf{0}_{n/2} & m_2\mathbf{I}_{n/2} \end{bmatrix} \tag{7.5}$$

where $\mathbf{I}_{n/2}$ is the $n/2 \times n/2$ identity matrix. n, m_1, and m_2 are chosen to be 100 DOFs, 0.1kg, and 1kg, respectively. The stiffness matrix of the system is given by

$$\mathbf{K}_n = k_u \begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & -1 & 2 & -1 & & \\ & & & \ddots & \ddots & -1 & \\ & & & & -1 & 2 & \end{bmatrix}, \tag{7.6}$$

with $k_u = 4 \times 10^5$ N/m.

The system is assumed to have Rayleigh damping by $\mathbf{C}_n = \alpha_0\mathbf{M}_n + \alpha_1\mathbf{K}_n$, where $\alpha_0 = 0.5$ and $\alpha_1 = 3 \times 10^{-5}$.

For a preliminary illustration, we again consider the computer-simulated FRFs of this system as if they were experimentally measured. A typical FRF of the system is shown in Figure 7.12.

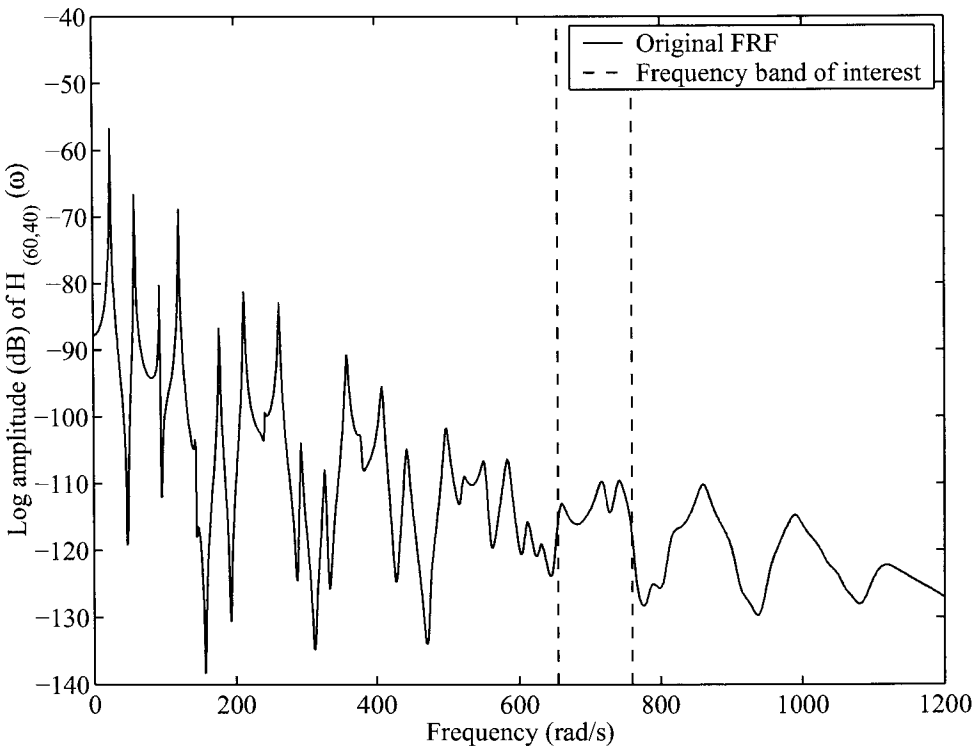

Figure 7.12: A typical FRF of the original coupled system and the frequency band of interest.

In the same figure, the frequency range considered for the construction of the POD and ICA is also plotted.

7.2.2 Reduced Order Model Forward Simulation

Using only the selected frequency range of the 'measured' FRFs, the correlation matrix is constructed and the POD eigensolutions are extracted. Normalized eigenvalues of the

correlation matrix, that is $\lambda/\lambda_{\max}$, are shown in Figure 7.13.

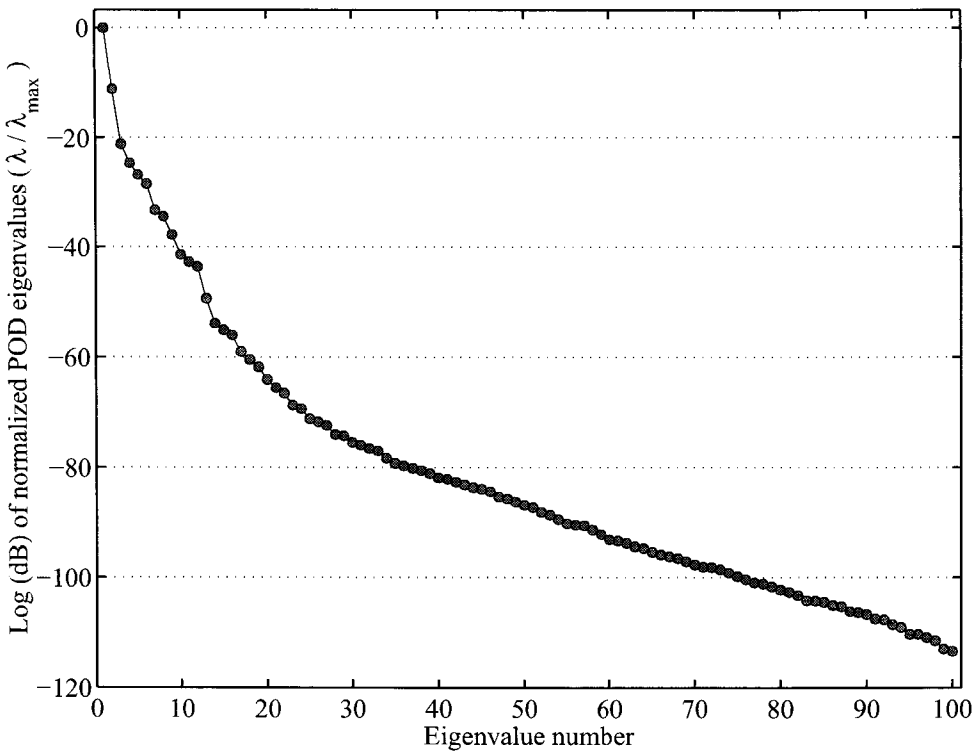

Figure 7.13: Normalized POD eigenvalues ($\lambda/\lambda_{\max}$) for the coupled system.

It is clear that the first few eigenvalues are significantly large compared to rest of the eigenvalues, as is expected from the scree test (see Section 4.5.2), as mentioned previously. This justifies the approximation in Equation (4.45). A typical FRF of the POD reduced system is compared with the original FRF in Figure 7.14.

With only 13 POD modes, the POD reduced order model FRF agrees reasonably well with the original FRF in the frequency band of interest.

To compare POD with ICA, the FRF obtained from the ICA reduced model is compared with the original FRF as well as the POD reduced model FRF in Figure 7.15.

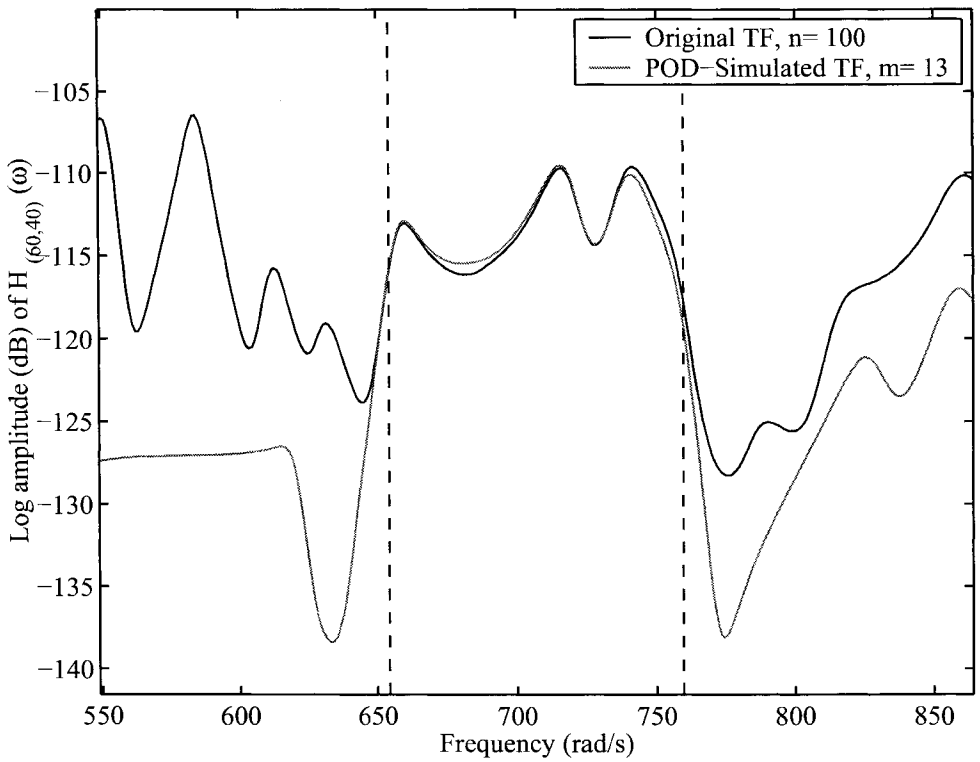

Figure 7.14: Original and POD reconstructed FRF for the coupled system.

Again, with 13 ICA modes, the ICA reduced order model FRF is the same as that of the POD reduced order model.

7.2.3 Reduced Order Model Identification: Noiseless Case

In the noise-free case , we first obtain the POD and ICA transformation matrix, choosing the number of modes to be 13 for each method. We then obtain the identified POD as well as ICA reduced order system matrices. The identified matrices are then used to obtain a typical FRF of the system which is compared to the original model FRF, as shown in Figure 7.16.

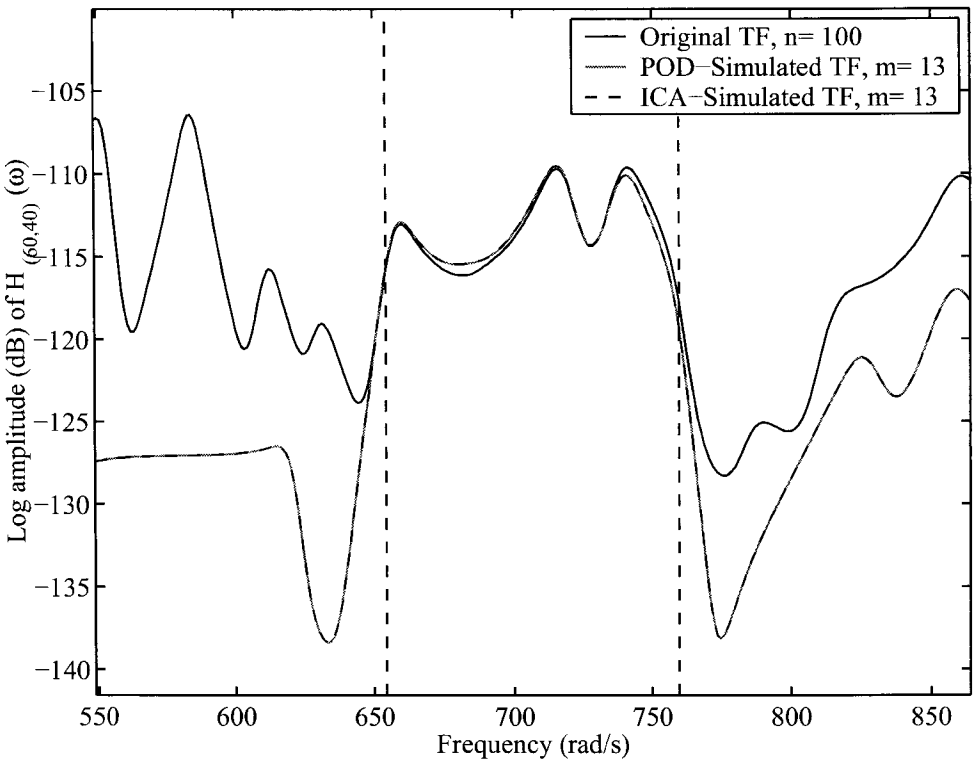

Figure 7.15: Original, ICA, and POD reconstructed FRF for the coupled system.

Again, the symmetry constraint was applied in the identification process, with the value for the regularization parameter, λ_M, λ_C, and λ_K all being 100. We can see that the identified system matrices result in FRFs that match reasonably well with the original system FRF.

For the identified FRF in Figure 7.16, the error is calculated to be 25.96 dB. The first mode obtained from ICA is compared to that obtained from POD in Figure 7.17.

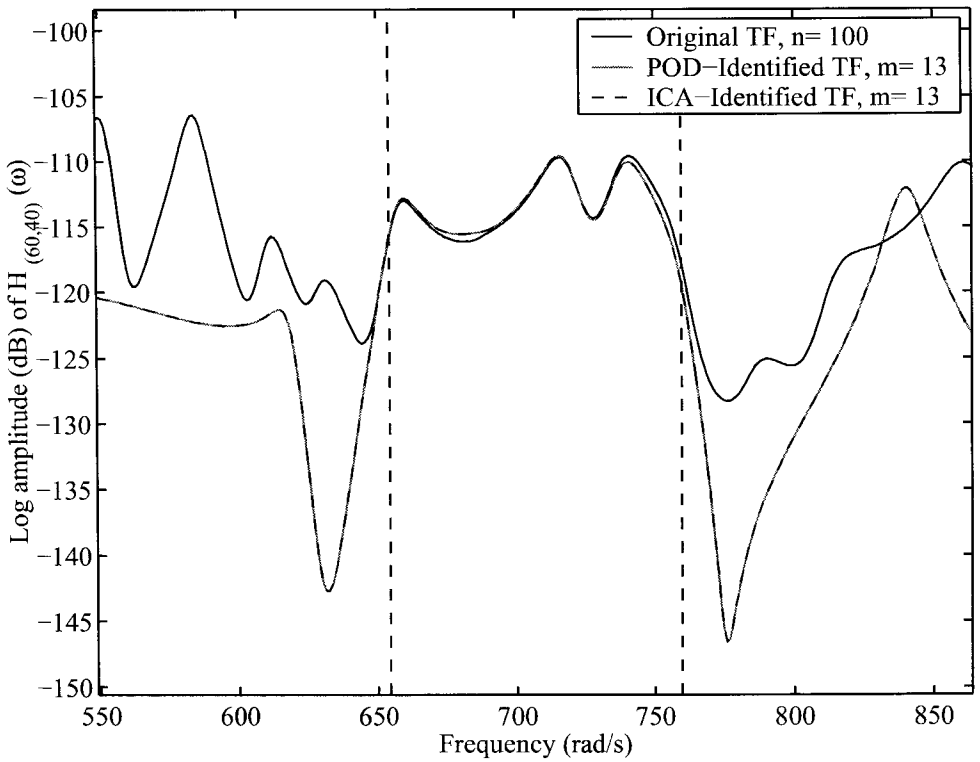

Figure 7.16: Original, ICA, and POD identified FRF for the coupled system.

7.2.4 Reduced Order Model Identification: Noisy Case

With an SNR of 30 dB, we obtain the identified FRF shown in Figure 7.18 with an error of 25.25 dB. Compared to the noiseless case in which the error in the identified FRF was 25.96 dB, a noise level of 30 dB SNR does not affect the identification process significantly.

With an SNR of 20 dB, we obtain the identified FRF shown in Figure 7.19 with an error of 22.43 dB. noiseless case in which the error in the identified FRF was 25.96 dB, a noise level of 20 dB SNR does not affect the identification process much neither.

Lastly, with an SNR of 10 dB, we obtain the identified FRF shown in Figure 7.20 with an error of 13.49 dB. Compared to the error of 22.43 dB for a noise level of 20 dB SNR, it

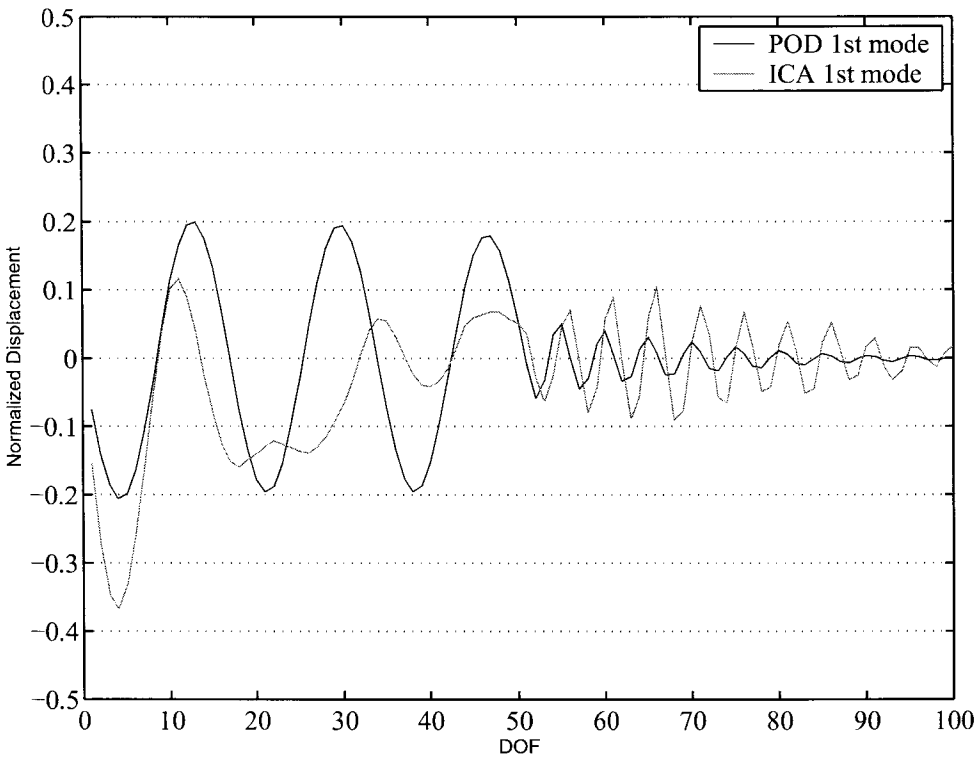

Figure 7.17: POD and ICA 1st mode for the coupled system.

is obvious that a stronger noise level of 10 dB SNR results in a worse estimate of the FRF.

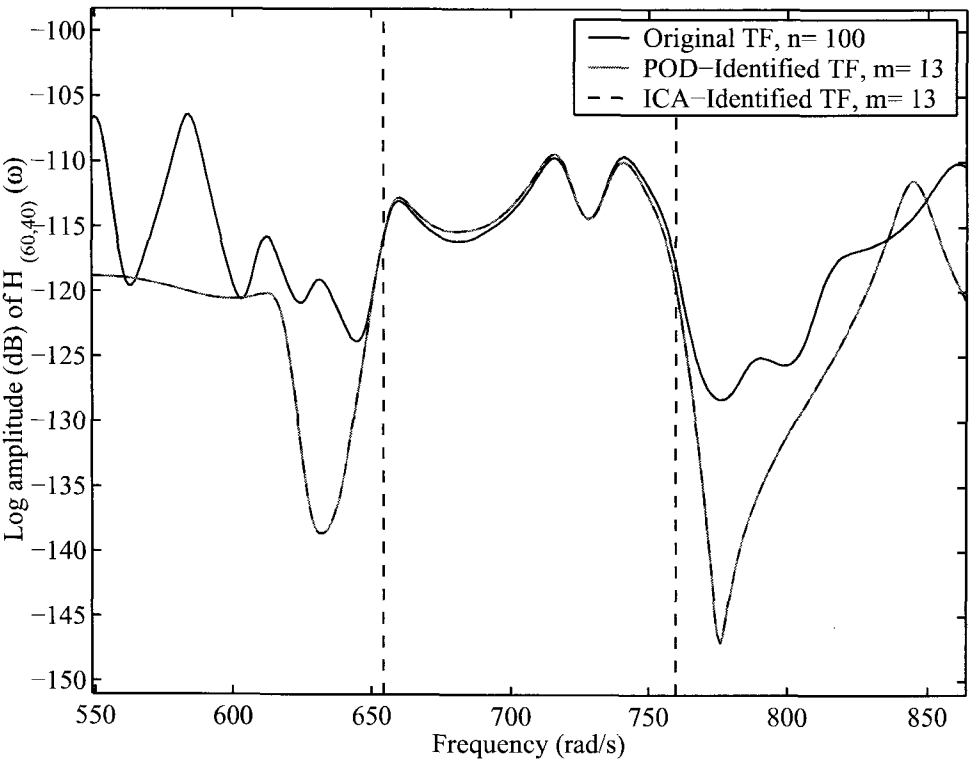

Figure 7.18: Original, ICA, and POD identified FRF for the coupled system under a noise level of 30 dB SNR.

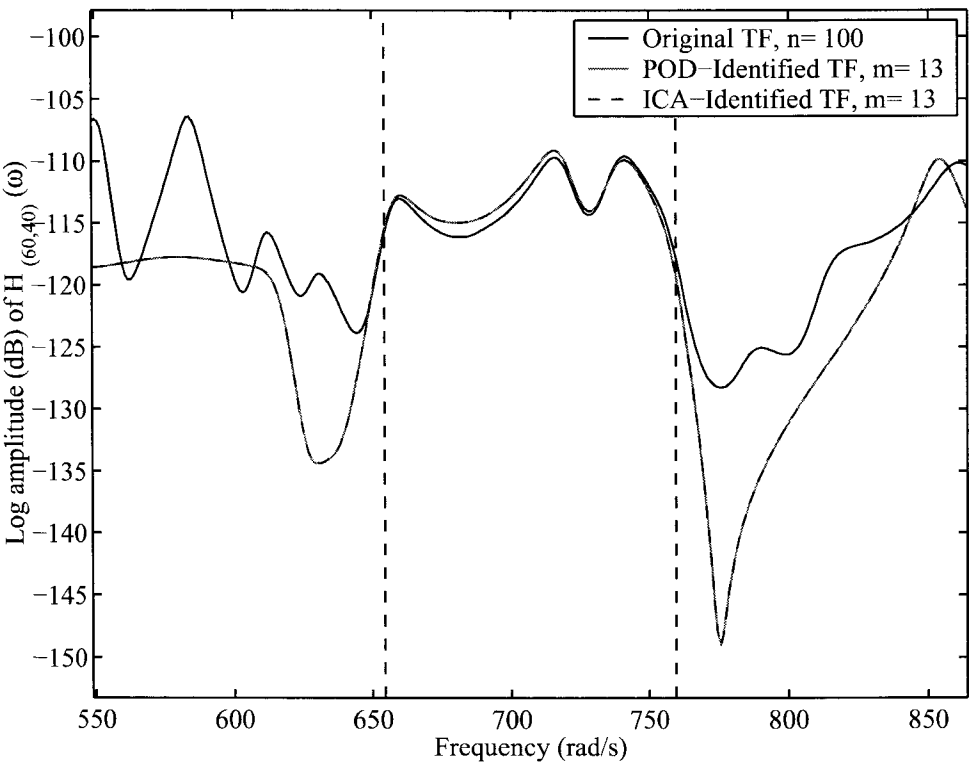

Figure 7.19: Original, ICA, and POD identified FRF for the coupled system under a noise level of 20 dB SNR.

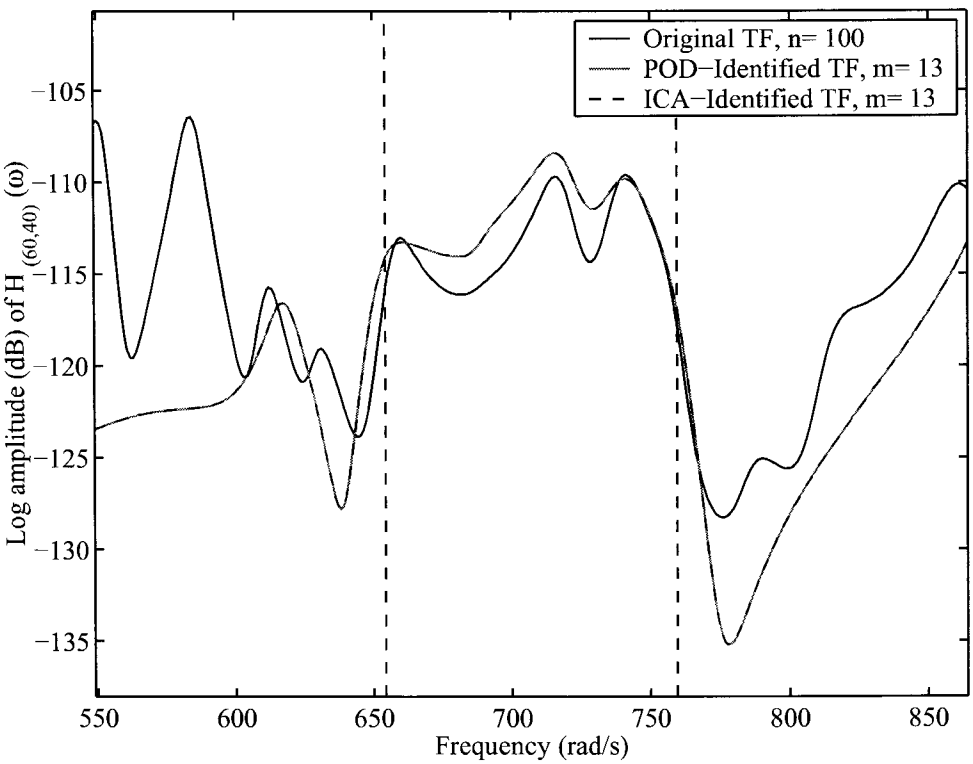

Figure 7.20: Original, ICA, and POD identified FRF for the coupled system under a noise level of 10 dB SNR.

Chapter 8

Conclusion

8.1 Summary of Findings

This thesis explored the feasibility of identifying a reduced order model of linear dynamical system in the mid-frequency regime. POD and ICA were used for model reduction strategy. Such a reduced-order model circumvented the limitations of traditional modal analysis developed for the low-frequency region.

The inverse problem relating to identification of the system matrices (namely mass, damping and stiffness matrices) was tackled in the framework of a linear least-square estimation. A mathematical framework of Kronecker Algebra elegantly handles the identification procedures involving system matrices. Additionally, the concept of Tikhonov Regularization was used to satisfy certain physical constraints involving the symmetric property of the identified matrices.

The salient features that emerged from the current investigation are:

1. At least for the discrete dynamical system investigated in this thesis, it is demonstrated that POD and ICA can be successfully applied for reduced-order modeling. The dimension of the reduced model may be an order of magnitude smaller than the corresponding comprehensive model. It also emerged that the ICA and POD reduced-order model may be indistinguishable in terms of the frequency response functions. However, the projection basis obtained by POD and ICA may exhibit discernable differences. On the other hand, ICA may be construed as a preconditioning step applied on the POD-based reduced order model.

2. Exploitation of Kronecker Algebra provided an elegant theoretical formulation involving identification of system matrices in the framework of linear least-square minimization techniques. To achieve the symmetry property of these matrices, the concept of Tikhonov Regularization was used.

3. It was also demonstrated that the predicted results obtained using the identified reduced-order model match reasonably well with the original system response. The robustness of the identification method was demonstrated by a noise-sensitivity study.

8.2 Future Research

Based on the current investigation, the writer believes that the following issues merit future investigation:

1. The reduced order model based on ICA and POD performs equally well, at least for the illustrative example considered in this investigation. Note that POD resolves

a spatio-temporal signal (i.e. vibration signature) into an optimal set of uncorrelated components. On the other hand, ICA achieves higher-order decorrelation of the signal components. The writer believes that the ICA-based model order reduction strategy introduced in this thesis may outperform the POD-based method in tackling strongly non-linear systems. This may be achieved by weakening the nonlinear coupling among the generalized coordinates of a discrete non-linear system (for example, refer to [42]).

2. The system identification scheme outlined in the thesis can be useful to tackle inverse problems arising in randomly heterogeneous dynamical systems. In this case, the system matrices are themselves random matrices, i.e. matrix-valued random variables (for example, refer to [113, 114]). Further exploration is necessary to address such problems.

Bibliography

[1] J. Hadamard. *Lectures on Cauchy's Problem in Linear Differential Equations*. Yale University Press, 1923.

[2] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston-Wiley, 1977.

[3] R. Ohayon and C. Soize. *Structural Acoustics and Vibration: Mechanical Models, Variational Formulations and Discretization*. Academic Press, 1998.

[4] R. H. Lyon. *Statistical Energy Analysis of Dynamical Systems : Theory and Applications*. MIT Press, 1975.

[5] U. Iemma, L. Moreno, and M. Diez. Digital holography and karhunenloève decomposition for the modal analysis of two-dimensional vibrating structures. *Journal of Sound and Vibration*, 291(1-2):107–131, 2006.

[6] R. Johansson. *System Modeling and Identification*. Prentice Hall, 1993.

[7] M. Geradin and D. Rixen. *Mechanical Vibrations: Theory and Applications to Structural Dynamics*. Wiley, 1997.

[8] R. W. Clough and J. Penzien. *Dynamics of Structures*. McGraw-Hill, 1975.

[9] J. Humar. *Dynamics of Structures*. Prentice Hall, 1990.

[10] D. J. Ewins. *Modal Testing: Theory and Practice*. Research Studies Press, second edition, 2000.

[11] N. M. M. Maia and J. M. M. Silva. *Theoretical and Experimental Modal Nalaysis*. Research Studies Press, 1997.

[12] P. Lancaster. Expression of damping matrices in linear vibration problems. *Journal of Aerospace Sciences*, 28:256, 1961.

[13] S. R. Ibrahim. Dynamic modeling of structures from measured complex modes. *AIAA Journal*, 21(6):898–901, 1983.

[14] S. Adhikari. Lancaster's method of damping identification revisited. *Journal of Vibration and Acoustics*, 124(4):617–627, 2002.

[15] M. J. Roemer and D. J. Mook. Mass, stiffness and damping matrix identification: an integrated approach. *Journal of Vibration and Acoustics*, 114:358–363, 1992.

[16] S. Y. Chen, M. S. Ju, and Y. G. Tsuei. Estimation of mass, stiffness and damping matrices from frequency response function. *Journal of Vibration and Acoustics*, 118:78–82, 1996.

[17] X. Zhao, Y. L. Xu, J. Li, and J. Chen. Hybrid identification method for multi-story buildings with unknown ground motion: theory. *Journal of Sound and Vibration*, 291(1-2):215–239, 2006.

[18] B. F. Yan, A. Miyamoto, and E. G. Bruhwiler. Wavelet transform-based modal parameter identification considering uncertainty. *Journal of Sound and Vibration*, 291(1-2):285–301, 2006.

[19] M. Baruch. Modal data are insufficient for identification of both mass and stiffness matrices. *AIAA Journal*, 35(11):1797–1798, 1997.

[20] M. Baruch. Correction of stiffness matrix using vibration tests. *AIAA Journal*, 20(3):441–442, 1982.

[21] M. Baruch. Optimal correction of mass and stiffness matrices using measured modes. *AIAA Journal*, 20(11):1623–1626, 1982.

[22] M. Baruch. Methods of reference basis for identification of linear dynamic structures. *AIAA Journal*, 22(4):561–564, 1984.

[23] R. Provasi and G. A. Zanetta. The extended kalman filter in the frequency domain for the identification of mechanical structures excited by sinusoidal multiple inputs. *Mechanical Systems and Signal Processing*, 14(3):327–341, 2000.

[24] A. Srikantha-Phani and J. Woodhouse. The challenge of reliable identification of complex modes. In *Proceedings of the IMAC-XX, 20th International Modal Analysis Conference*, Los Angeles, USA, 2002.

[25] S. Adhikari. Optimal complex modes and an index of damping non-proportionality. *Mechanical System and Signal Processing*, 18(1):1–27, 2004.

[26] L. Ljung. *System Identification*. Prentice Hall, 1999.

[27] A. P. Sage and J. L. Melsa. *System Identification*. Academic Press, 1971.

[28] R. Pintelon and J. Schoukens. *System Identification: A Frequency Domain Approach*. IEEE Press, 2001.

[29] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1965.

[30] C. P. Fritzen. Identification of mass, stiffness and damping matrices of mechanical systems. *Journal of Vibration, Acoustics, Stress, and Reliability in Design*, 108(1):9–16, 1986.

[31] J. E. Mottershead, A. W. Lees, and R. Stanway. A linear, frequency domain filter for parameter identification of vibrating structures. *Transactions of the ASME*, 109:262–269, 1987.

[32] G. A. Zanetta. Identification methods in the dynamics of turbogenerator rotors. In *Proceedings of the IMechE, International Conference on Viibrations in Rotating Machinery*, pages 173–181, Bath, UK, 1992.

[33] L. Ljung. *The System Identification Toolbox: The Manual*. The MathWorks Inc., fourth edition, 1995.

[34] N. Boivin and C. Pierre. Non-linear modal analysis of structural systems featuring internal resonances. *Journal of Sound and Vibration*, 182(2):336–341, 1995.

[35] M. Schetzen. *The Volterra and Wiener Theories on Nonlinear Systems*. Wiley, 1980.

[36] K. Worden and G. R. Tomlinson. *Nonlinearity in Structural Dynamics*. Taylor and Francis, 2001.

[37] S. J. Giffor and G. R. Tomlinson. Recent advances in the application of functional series to non-linear structures. *Journal of Sound and Vibration*, 135:289–317, 1989.

[38] K. Worden and G. R. Tomlinson. Random vibrations of a dung oscillator using the volterra series. *Journal of Sound and Vibration*, 217:781–789, 1998.

[39] A. A. Khan and N. S. Vyas. Non-linear parameter estimation using volterra and wiener theories. *Journal of Sound and Vibration*, 221:805–821, 1999.

[40] A. Sarkar and R. Ghanem. Mid-frequency structural dynamics with parameter uncertainty. *Computer Methods in Applied Mechanics and Engineering*, 191(47-48):5499–5513, November 2002.

[41] A. Sarkar and M. P. Paidoussis. A compact limit-cycle oscillation model of a cantilever conveying fluid. *Jounal of Fluids and Strcutures*, 17(4):525–539, March 2003.

[42] A. Sarkar and M. P. Paidoussis. A cantilever conveying fluid: coherent modes versus beam modes. *International Jounal of Non-Linear Mechanics*, 39(3):467–481, April 2004.

[43] V. Lenaerts, G. Kerschen, and J. C. Golinval. Physical interpretation of the proper orthogonal modes using the singular value decomposition. *Journal of Sound and Vibration*, 249(5):849–865, January 2002.

[44] B. F. Feeny. On proper orthogonal co-ordinates as indicators of modal activity. *Journal of Sound and Vibration*, 255(5):805–817, 2002.

[45] M. F. A. Azeez and A. F. Vakakis. Numerical and experimental analysis of the nonlinear dynamics due to impacts of a continuous overhung rotor. In *Proceedings of DETC97, ASME Design Engineering Technical Conferences*, Sacramento, USA, 1997.

[46] M. F. A. Azeez and A. F. Vakakis. Proper orthogonal decomposition of a class of vibroimpact oscillations. *Journal of Sound and Vibration*, 240(5):859–889, 2001.

[47] V. Lenaerts and J. C. Golinval. Non-linear generalization of principal component analysis: from a global to a local approach. *Journal of Sound and Vibration*, 254(5):867–876, 2002.

[48] G. Kerschen, B. F. Feeny, and J. C. Golinval. On the exploitation of chaos to build reduced-order models. *Computer Methods in Applied Mechanics and Engineering*, 192:1785–1795, 2003.

[49] G. Kerschen and J. C. Golinval. Generation of accurate finite element models of nonlinear systems application to an aeroplane-like structure. *Nonlinear Dynamics*, 39:129–142, 2003.

[50] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis*, 40(2):492–515, 2002.

[51] V. Lenaerts, G. Kerschen, and J. C. Golinval. Proper orthogonal decomposition for model updating of non-linear mechanical systems. *Mechanical Systems and Signal Processing*, 15(1):31–43, 2001.

[52] V. Lenaerts, G. Kerschen, and J.-C. Golinval. Identification of a continuous structure with a geometrical non-linearity. part ii: Proper orthogonal decomposition. *Journal of Sound and Vibration*, 262(4):907–919, May 2003.

[53] J. L. Lumley. *Stochastic Tools in Turbulence*. Academic Press, 1971.

[54] D. D. Kosambi. Statistics in function space. *Journal of the Indian Mathematical Society*, 7:76–88, 1943.

[55] M. Loéve. Functions aléatoire de second ordre. *Comptes Rendus de l'Academie des Sciences Paris*, 220, 1945.

[56] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae A1*, 34, 1946.

[57] V. S. Pougachev. General theory of the correlations of random functions. *Izvestiya Akademii Nauk SSSR Seriya Matematicheskaya*, 17:401–402, 1953.

[58] A. M. Obukhov. Statistical description of continuous fields. *Trudy Geofiz. In-ta Akademii Nauk SSSR*, 24:3–42, 1954.

[59] A. Rosenfeld and A. C. Kak. *Digital Picture Processing*. Academic Press, 1982.

[60] V. R. Algazi and D. J. Sakrison. On the optimality of the karhunen-loéve expansion. *IEEE Transactions on Information Theory*, 15:319–321, 1969.

[61] C. A. Andrews, J. M. Davies, and G. R. Schwartz. Adaptive data compression. In *Proceedings of the IEEE*, volume 55, pages 267–277, 1967.

[62] J. Hérault and B. Ans. Neural network with modifiable synapses - decoding of composite sensory messages under unsupervised and permanent learning. *Comptes Rendus de l'Academie des Sciences Series III - Sciences de la Vie*, 299(13):525–528, 1984.

[63] J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentisage non supervisé. In *Actes du Xème colloque GRETSI*, pages 1017–1022, Nice, France, 1985.

[64] B. Ans, J. Hérault, and C. Jutten. Adaptive neural architectures: detection of primitives. In *Proceedings of COGNITIVA'85*, pages 593–597, Paris, France, 1985.

[65] J.-F. Cardoso. Blind identification of independent components with higher-order statistics. In *Proceedings of the Workshop on Higher-Order Spectral Analysis*, pages 157–160, Vail, Colorado, 1989.

[66] P. Comon. Blind identification of independent signals. In *Proceedings of the Workshop on Higher-Order Spectral Analysis*, pages 174–179, Vail, Colorado, 1989.

[67] D. L. Donoho. On minimum entropy deconvolution. In *Applied Time Series Analysis II*, pages 565–608. Academic Press, 1981.

[68] O. Shalvi and E. Weinstein. New criteria for blind deconvlution of nonminimum phase systems. *IEEE Transactions on Information Theory*, 36(2):312–321, 1990.

[69] A. J. Bell and T. J. Sejnowski. A non-linear information maximization algorithm that performs blind separation. In *Advances in Neural Information Processing Systems 7*, pages 467–474. The MIT Press, 1995.

[70] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[71] S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. The MIT Press, 1996.

[72] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[73] A. Hyvärinen. A family of fixed-point algorithm for independent component analysis. In *Proceedings of the ICASSP'97, IEEE International Conference on Accoustics, Speech, and Signal Processing*, pages 3917–3920, Munich, Germany, 1997.

[74] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[75] M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1992.

[76] K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *Proceedings of the ICONIP'98, International Conference on Neural Information Processing*, volume 2, pages 895–898, Tokyo, Japan, 1998.

[77] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13, 2001.

[78] Z. N. Li, Y. Y. He, F. L. Chu, J. Han, and W. Hao. Fault recognition method for speed-up and speed-down process of rotating machinery based on independent component analysis and factorial hidden markov model. *Journal of Sound and Vibration*, 291(1-2):60–71, 2006.

[79] M. J. Zuo, J. Lin, and X. F. Fan. Feature separation using ica for a one-dimensional time series and its application in fault detection. *Journal of Sound and Vibration*, 287(3):614–624, 2005.

[80] S. I. Grossman. *Elementary Linear Algebra*. Wadsworth, third edition, 1987.

[81] J. M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Prentice Hall, second edition, 1995.

[82] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[83] A. W. F. Edwards. *Likelihood; an Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*. Cambridge University Press, 1972.

[84] A. Graham. *Kronecker Products and Matrix Calculus: With Applications*. Ellis Horwood, 1981.

[85] H. Neudecker. Some theorems on matrix differentiation with special reference to kronecker matrix products. *Journal of the American Statistical Association*, 64(327):953–963, September 1969.

[86] A. Graham. *Matrix Theory and Applications for Engineers and Mathematicians*. Ellis Horwood, 1979.

[87] H. V. Henderson and S. R. Searle. The vec-permutation matrix, the vec operator and kronecker products: A review. *Linear and Multilinear Algebra*, 9:271–288, 1981.

[88] D. S. G. Pollock. *The Algebra of Econometrics*. Wiley, 1979.

[89] P. Holmes, J. L. Lumley, and G Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, 1996.

[90] F. Riesz and B. Sz. Nagy. *Functional Analysis*. Ungar, 1955.

[91] J. C. Heinrich and C. A. Vionnet. On boundary-conditions for unbounded flows. *Communications in Numerical Methods in Engineering*, 11(2):179–185, 1995.

[92] R. Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. Springer-Verlag, 1988.

[93] A. Sarkar and R. Ghanem. A substructure approach for the midfrequency vibration of stochastic systems. *Journal of the Accoustical Society of America*, 113(4):1922–1934, April 2003.

[94] T. T. Soong and M. Grigoriu. *Random Vibration of Mechanical and Structural Systems*. Prentice Hall, 1993.

[95] R. L. Gorsuch. *Factor Analysis*. Lawrence Erlbaum Associates, second edition, 1983.

[96] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[97] T. Lee. *Independent Component Analysis: Theory and Applications*. Kluwer Academic Publishers, 1998.

[98] S. Roberts and R. Everson. *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[99] L. Tong, Y. Inouye, and R. Liu. Waveform preserving blind estimation of multiple independent sources. *IEEE Transactions on Signal Processing*, (41):2461–2470, 1999.

[100] P. Comon. Independent component analysis - a new concept? *Signal Processing*, (36):287–314, 1994.

[101] W. Feller. *Probability Theory and Its Applications*. Wiley, third edition, 1968.

[102] J.-F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *Proceedings of the ICASSP'90, IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2655–2658, Albuquerque, New Mexico, 1990.

[103] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Halsted Press, 1992.

[104] B. N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice Hall, 1980.

[105] M. Khalil, S. Adhikari, and A. Sarkar. Identification of damping using proper orthogonal decomposition. In *Proceedings of the Eighth International Conference on Computational Structures Technology*, Las Palmas de Gran Canaria, Spain, 2006. Accepted for publication.

[106] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51:406–413, 1955.

[107] H. Sorenson. *Parameter Estimation - Principles and Problems*. Marcel Dekker, 1980.

[108] R. Hunt. Three-dimensional flow in a general tube using a combination of finite and pseudospectral discretisations. *SIAM Journal on Scientific Computing*, 16(3):513–530, 1995.

[109] M. Hanke and P. C. Hansen. Regularization methods for alrge-scale problems. *Surveys on Mathematics for Industry*, 3:253–315, 1993.

[110] P. C. Hansen. Numerical tools for analysis and solution of fredholm integral equations of the first kind. *Inverse Problems*, 8:849–872, 1992.

[111] P. C. Hansen. A matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6:1–35, 1994.

[112] A. V. Oppenheim, A. S. Willsky, and I. T. Young. *Signals and Systems*. Prentice-Hall, 1983.

[113] S. Adhikari and M. I. Friswell. Random matrix eigenvalue problems in structural dynamics. *International Journal of Numerical Methods in Engineering*, 2006. Accepted for publication.

[114] T. Wagenknecht, K. Green, S. Adhikari, and W. Michiels. Structured pseudospectra and random eigenvalue problems in vibrating systems. *AIAA Journal*, 2006. Accepted for publication.