Chapter 6

Inverse problems and mapping inversion

6.1 Introduction

The concepts of "inverse problem" and "mapping inversion" are often used interchangeably in the machine learning literature, although they denote different things. The aim of this chapter is: to introduce the area of (Bayesian) inverse problem theory; to compare it with general Bayesian analysis and in particular with latent variable models; and to differentiate it from the problems of mapping inversion and mapping approximation.

Section 6.2 defines inverse problem theory, explains the reasons for non-uniqueness of the solution and reviews Bayesian inverse problem theory, including the topics of the choice of prior distributions, stability and regularisation. It also describes several examples of inverse problems in some detail to clarify the theory and its interpretation. Section 6.3 compares Bayesian inverse problem theory with general Bayesian analysis and in particular with latent variable models. Section 6.4 defines (statistical) mapping inversion, mapping approximation and universal mapping approximators.

6.2 Inverse problem theory

6.2.1 Introduction and definitions

Inverse calculations involve making inferences about models of physical systems from data[1]. The scientific procedure to study a physical system can be divided into three parts:

1. *Parameterisation of the system*: discovery of a minimal set of model parameters[2] whose values completely characterise the system.

2. *Forward modelling*: discovery of the physical laws allowing, for given values of the parameters, predictions of some observable or data parameters to be made.

3. *Inverse modelling*: use of actual measurements of the observed parameters to infer the values of the model parameters. This inference problem is termed the **inverse problem**.

The model parameters conform the model space $\mathcal{M}$, the observable parameters conform the data space $\mathcal{D}$ and the union of both parameter sets conforms the parameter space $\mathcal{X} = \mathcal{D} \times \mathcal{M}$. See section 6.2.3.1 for a further interpretation of the model parameters.

Usually the forward problem is a well-defined single-valued relationship (i.e., a function in the mathematical sense) so that given the values of the model parameters the values of the measured parameters are uniquely

[1]Tarantola (1987) claims that inverse problem theory in the wide sense has been developed by people working with geophysical data, because geophysicists try to understand the Earth's interior but can only use data collected at the Earth's surface. However, inverse problems appear in many other areas of physics and engineering, some of which are briefly reviewed in section 6.2.4.

[2]The term *parameters* is used in inverse problem theory to mean both the variables and the parameters of a model, as these terms are usually understood in machine learning. Throughout this chapter, we will keep the notation and naming convention which is standard in inverse problem theory. In section 6.3 we discuss the point of view of probabilistic models.

identified. This is often due to causality in the physical system. But often this forward mapping is many-to-one, so that the inverse problem is one-to-many: given values of the observed parameters, there is more than one model (possibly an infinite number) that corresponds to them.

Thus, if $\mathbf{g} : \mathcal{M} \to \mathcal{D}$ is the forward mapping, then $\mathbf{d} = \mathbf{g}(\mathbf{m})$ is unique given $\mathbf{m}$, but its inverse $\mathbf{g}^{-1}(\mathbf{d})$ can take several values for some observed $\mathbf{d} \in \mathcal{D}$.

The example in section 6.2.4.1 illustrates this abstract formulation.

6.2.1.1 Types of inverse problems

Inverse problems can be classified as:

Continuous Most inverse problems are of this type. The model to be estimated is a continuous function in several variables. For example, the mass density distribution inside the Earth as a function of the space coordinates.

Discrete There is a finite (actually numerable) number of model parameters to be estimated. Sometimes the problem itself is discrete in nature, e.g. the location of the epicentre in the example 6.2.4.1, which is parameterised by the epicentre coordinates X and Y. But most times, the problem was originally continuous and was discretised for computational reasons. For example, one can express the mass density distribution inside the Earth as a parameterised function in spherical coordinates (perhaps obtained as the truncation of an infinite parameterised series) or as a discrete grid (if the sampling length is small enough).

In this chapter we deal only with discrete inverse problems. Tarantola (1987) discusses both discrete and continuous inverse problems.

6.2.1.2 Why the nonuniqueness?

Nonuniqueness arises for several reasons:

- Intrinsic lack of data: for example, consider the problem of estimating the density distribution of matter inside the Earth from knowledge of the gravitational field at its surface. Gauss' theorem shows that there are infinitely many different distributions of matter density that give rise to identical exterior gravitational fields. In this case, it is necessary to have additional information (such as a priori assumptions on the density distribution) or additional data (such as seismic observations).
- Uncertainty of knowledge: the observed values always have experimental uncertainty and the physical theories of the forward problem are always approximations of the reality.
- Finiteness of observed data: continuous inverse problems have an infinite number of degrees of freedom. However, in a realistic experiment the amount of data is finite and therefore the problem is underdetermined.

6.2.1.3 Stability and ill-posedness of inverse problems

In the Hadamard sense, a well-posed problem must satisfy certain conditions of existence, uniqueness and continuity. Ill-posed problems can be numerically unstable, i.e., sensitive to small errors in the data (arbitrarily small changes in the data may lead to arbitrarily large changes in the solution).

Nonlinearity has been shown to be a source of ill-posedness (Snieder and Trampert, 1999), but linearised inverse problems can often be ill-posed too due to the fact that realistic data is finite. Therefore, inverse problems in general might not have a solution in the strict sense, or if there is a solution, it might not be unique or might not depend continuously on the data. To cope with this problem, stabilising procedures such as regularisation methods are often used (Tikhonov and Arsenin, 1977; Engl et al., 1996). Bayesian inversion is, in principle, always well-posed (see section 6.2.3.5). Mapping inversion (section 6.4) is also a numerically unstable problem but, again, probabilistic methods such as the one we develop in chapter 7 are well-posed.

6.2.2 Non-probabilistic inverse problem theory

For some inverse problems, such as the reconstruction of the mass density of a one-dimensional string from measurements of all eigenfrequencies of vibration of the string, an exact theory for inversion is available (Snieder and Trampert, 1999). Although these exact nonlinear inversion techniques are mathematically elegant, they are of limited applicability because:

- They are only applicable to idealistic situations which usually do not hold in practice. That is, the physical models for which an exact inversion method exists are only crude approximations of reality.
- They are numerically unstable.
- The discretisation of the problem caused by the fact that the data are only available in a finite amount makes the problem underdetermined.

Non-probabilistic inversion methods attempt to invert the mathematical equation of the forward mapping (for example, solving a linear system of equations by using the pseudoinverse). These methods cannot deal with data uncertainty and redundancy in a natural way, and we do not deal with such methods here. A more general formulation of inverse problems is obtained using probability theory.

6.2.3 Bayesian inverse problem theory

The standard reference for the Bayesian view of (geophysical) inversion is Tarantola (1987), whose notation we use in this section; the standard reference for the frequentist inverse theory is Parker (1994); other references are Scales and Smith (1998) and Snieder and Trampert (1999).

In the Bayesian approach to inverse problems, we use physical information about the problem, plus possibly uninformative prior distributions, to construct the following two models:

- A joint prior distribution $\rho(\mathbf{d}, \mathbf{m})$ in the parameter space $\mathcal{X} = \mathcal{D} \times \mathcal{M}$. This prior distribution is usually factorised as $\rho_{\mathcal{D}}(\mathbf{d})\rho_{\mathcal{M}}(\mathbf{m})$, because by definition the a priori information on the model parameters is independent of the observations. However, it may happen that part of this prior information was obtained from a preliminary analysis of the observations, in which case $\rho(\mathbf{d}, \mathbf{m})$ might not be factorisable. If no prior information is available, then an uninformative prior may be used (see section 6.2.3.2).
- Using information obtained from physical theories we solve the forward problem, deriving a deterministic forward mapping $\mathbf{d} = \mathbf{g}(\mathbf{m})$. If a noise model f (typically normal) is applied, a conditional distribution $\theta(\mathbf{d}|\mathbf{m}) = f(\mathbf{d} - \mathbf{g}(\mathbf{m}))$ may be derived. For greater generality, the information about the resolution of the forward problem is described by a joint density function $\theta(\mathbf{d}, \mathbf{m})$. However, usually $\theta(\mathbf{d}, \mathbf{m}) = \theta(\mathbf{d}|\mathbf{m})\mu_{\mathcal{M}}(\mathbf{m})$, where $\mu_{\mathcal{M}}(\mathbf{m})$ describes the state of null information on model parameters.

Tarantola (1987) postulates that the a posteriori state of information is given by the *conjunction* of the two states of information: the prior distribution on the $\mathcal{D} \times \mathcal{M}$ space and the information about the physical correlations between $\mathbf{d}$ and $\mathbf{m}$. The conjunction is defined as

$$\sigma(\mathbf{d}, \mathbf{m}) \stackrel{\text{def}}{=} \frac{\rho(\mathbf{d}, \mathbf{m})\theta(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} \tag{6.1}$$

where $\mu(\mathbf{d}, \mathbf{m})$ is a distribution representing the state of null information (section 6.2.3.2). Thus, all available information is assimilated into the posterior distribution of the model given the observed data, computed by marginalising the "joint posterior distribution" σ (assuming factorised priors $\rho(\mathbf{d}, \mathbf{m})$ and $\mu(\mathbf{d}, \mathbf{m})$):

$$\sigma_{\mathcal{M}}(\mathbf{m}) = \int_{\mathcal{D}} \sigma(\mathbf{d}, \mathbf{m})\, d\mathbf{d} = \rho_{\mathcal{M}}(\mathbf{m})L(\mathbf{m}) \tag{6.2}$$

where the likelihood function L, which measures the data fit, is defined as:

$$L(\mathbf{m}) \stackrel{\text{def}}{=} \int_{\mathcal{D}} \frac{\rho_{\mathcal{D}}(\mathbf{d})\theta(\mathbf{d}|\mathbf{m})}{\mu_{\mathcal{D}}(\mathbf{d})}\, d\mathbf{d}.$$

Thus, the "solution" of the Bayesian inversion method is the posterior distribution $\sigma_{\mathcal{M}}(\mathbf{m})$, which is unique (although it may be multimodal and even not normalisable, depending on the problem). Usually a maximum a posteriori (MAP) approach is adopted, so that we take the model maximising the posterior probability σ: $\mathbf{m}_{\text{MAP}} \stackrel{\text{def}}{=} \max_{\mathbf{m} \in \mathcal{M}} \sigma_{\mathcal{M}}(\mathbf{m})$.

Likewise, the posterior distribution in the data space is calculated as

$$\sigma_{\mathcal{D}}(\mathbf{d}) = \int_{\mathcal{M}} \sigma(\mathbf{d}, \mathbf{m})\, d\mathbf{m} = \frac{\rho_{\mathcal{D}}(\mathbf{d})}{\mu_{\mathcal{D}}(\mathbf{d})} \int_{\mathcal{M}} \theta(\mathbf{d}|\mathbf{m})\rho_{\mathcal{M}}(\mathbf{m})\, d\mathbf{m}$$

which allows to estimate posterior values of the data parameters (*recalculated data*).

In practice, all uncertainties are described by stationary Gaussian distributions:

- likelihood $L(\mathbf{m}) \sim \mathcal{N}(\mathbf{g}(\mathbf{m}), \mathbf{C}_{\mathcal{D}})$
- prior $\rho_{\mathcal{M}}(\mathbf{m}) \sim \mathcal{N}(\mathbf{m}_{\text{prior}}, \mathbf{C}_{\mathcal{M}})$.

A more straightforward approach that still encapsulates all the relevant features (uninformative priors and an uncertain forward problem) is simply to obtain the posterior distribution of the model given the observed data as

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})} \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m}). \tag{6.3}$$

This equation should be more familiar to statistical learning researchers.

6.2.3.1 Interpretation of the model parameters

Although the treatment of section 6.2.3 is perfectly general, it is convenient to classify the model parameters into one of two types:

- Parameters that describe the *configuration* or *state* of the physical system and completely characterise it. In this case, they could be called *state variables* as in dynamical system theory. We represent them as a vector $\mathbf{s}$. Examples of such parameters are the location of the epicentre in example 6.2.4.1 or the absorption coefficient distribution of a medium in CAT (example 6.2.4.3). These parameters are independent, in principle, of any measurements taken of the system (such as a particular projection in CAT or a measurement of the arrival time of the seismic wave in example 6.2.4.1).
- Parameters that describe the *experimental conditions* in which a particular measurement of the system was taken. Thus, for each measurement $\mathbf{d}_n$ we have a vector $\mathbf{c}_n$ indicating the conditions in which it was taken. For example, in 2D CAT (example 6.2.4.3) one measurement is obtained from a given X-ray source at plane coordinates x, y and at an angle θ; thus $\mathbf{c}_n = (x_n, y_n, \theta_n)$ and the measurement $\mathbf{d}_n$ is the transmittance. In example 6.2.4.1, one measurement is taken at the location (x_n, y_n) of station n; and so on. If there are N measurements, then the model parameters are $\{\mathbf{c}_n\}_{n=1}^N$ (in addition to the $\mathbf{s}$ model parameters) and one can postulate prior distributions for them to indicate uncertainties in their determination. However, usually one assumes that there is no uncertainty involved in the conditions of the measurement and takes these distributions as Dirac deltas. Of course, the measured value $\mathbf{d}_n$ can still have a proper distribution reflecting uncertainty in the actual measurement. In this way, the estimation problem is simplified, because all $\{\mathbf{c}_n\}_{n=1}^N$ model parameters are considered constant, and only the $\mathbf{s}$ model parameters are estimated.

From a Bayesian standpoint, there is no formal difference between both kinds of model parameters, state $\mathbf{s}$ and experimental conditions $\mathbf{c}$—or between model parameters $\mathbf{m}$ and data parameters $\mathbf{d}$, for that matter—because probability distributions are considered for all variables and parameters of the problem.

Thus, if there are N measurements, the forward mapping is $\mathbf{d} = \mathbf{g}(\mathbf{m})$ with $\mathbf{d} = (\mathbf{d}_1, \dots, \mathbf{d}_N)$ and $\mathbf{m}$ including both kinds of parameters (state variables and experimental conditions): $\mathbf{m} = (\mathbf{s}, \mathbf{c}_1, \dots, \mathbf{c}_N)$. Usually the measurements are taken independently, so that

$$\rho_{\mathcal{D}}(\mathbf{d}) = \prod_{n=1}^{N} \rho_{\mathcal{D},n}(\mathbf{d}_n) \qquad \mu_{\mathcal{D}}(\mathbf{d}) = \prod_{n=1}^{N} \mu_{\mathcal{D},n}(\mathbf{d}_n)$$

and the forward mapping decomposes into N equations $\mathbf{d}_n = \mathbf{g}_n(\mathbf{s}, \mathbf{c}_n)$. Thus $\theta(\mathbf{d}|\mathbf{m}) = \prod_{n=1}^N \theta_n(\mathbf{d}_n|\mathbf{s}, \mathbf{c}_n) = \prod_{n=1}^N f(\mathbf{d}_n - \mathbf{g}_n(\mathbf{s}, \mathbf{c}_n))$ and the likelihood function factorises as $L(\mathbf{m}) = \prod_{n=1}^N L_n(\mathbf{m})$ with

$$L_n(\mathbf{m}) \stackrel{\text{def}}{=} \int_{\mathcal{D}_n} \frac{\rho_{\mathcal{D},n}(\mathbf{d}_n)\theta_n(\mathbf{d}_n|\mathbf{m}, \mathbf{c}_n)}{\mu_{\mathcal{D},n}(\mathbf{d}_n)}\, d\mathbf{d}_n.$$

6.2.3.2 Choice of prior distributions

The controversial matter in the Bayesian approach is, of course, the construction of prior distributions. Usual ways to do this are (Jaynes, 1968; Kass and Wasserman, 1996):

- Define a measure of information, such as the entropy, and determine the distribution that optimises it (e.g. maximum entropy).

- Define properties that the noninformative prior should have, such as invariance to certain transformations (Jeffreys' prior). For example, for a given definition of the physical parameters $\mathbf{x}$, it is possible to find a unique density function $\mu(\mathbf{x})$ which is form invariant under the transformation groups which leave the fundamental equations of physics invariant.

- The previous choices, usually called "objective" or "noninformative," are constructed by some formal rule, but it is also possible to use priors based on subjective knowledge.

In any case, the null information distributions are obtained in each particular case, depending on the coordinate systems involved, etc. However, the choice of a noninformative distribution for a continuous, multidimensional space remains a delicate problem. Bernardo and Smith (1994, pp. 357–367) discuss this issue.

6.2.3.3 Bayesian linear inversion theory

Assuming that all uncertainties are Gaussian, if the forward operator is linear, then the posterior distribution σ will also be Gaussian. This is equivalent to factor analysis, which is a latent variable model where the prior distribution in latent space is Gaussian, the mapping from latent onto data space is linear and the noise model in data space is Gaussian.

Linear inversion theory is well developed and involves standard linear algebra techniques: pseudoinverse, singular value decomposition and (weighted) least squares problems. Tarantola (1987) gives a detailed exposition.

6.2.3.4 Bayesian nonlinear inversion theory

Almost all work in nonlinear inversion theory, particularly in geophysics, is based on linearising the problem using physical information. Usual linearisation techniques include the Born approximation (also called the single-scattering approximation), Fermat's principle and Rayleigh's principle (Snieder and Trampert, 1999).

6.2.3.5 Stability

In the Bayesian approach to inverse problems, it is not necessary in principle to invert any operators to construct the solution to the inverse problem, i.e., the posterior probability σ. Thus, from the Bayesian point of view, no inverse problem is ill-posed (Gouveia and Scales, 1998).

6.2.3.6 Confidence sets

Once a MAP model has been selected from the posterior distribution σ, confidence sets or other measures of resolution can be extracted from $\sigma(\mathbf{m}_{\mathrm{MAP}})$. Due to the mathematical complexity of this posterior distribution, only approximate techniques are possible, including the following ones:

- The forward operator $\mathbf{g}$ is linearised about the selected model $\mathbf{m}_{\mathrm{MAP}}$, so that the posterior becomes normal: $\sigma(\mathbf{m}) \sim \mathcal{N}(\mathbf{m}_{\mathrm{MAP}}, \mathbf{C}'_{\mathcal{M}})$. The posterior covariance matrix $\mathbf{C}'_{\mathcal{M}}$ is obtained as $\mathbf{C}'_{\mathcal{M}} = (\mathbf{G}^T \mathbf{C}_{\mathcal{D}}^{-1} \mathbf{G} + \mathbf{C}_{\mathcal{M}}^{-1})^{-1}$, where $\mathbf{G}$ is the derivative[3] of $\mathbf{g}$ with respect to the model parameters evaluated at $\mathbf{m}_{\mathrm{MAP}}$. This has the same form as the posterior covariance matrix in latent space of a factor analysis, as in eq. (2.59), where $\mathbf{G}$ would be the factor loadings matrix $\mathbf{\Lambda}$.

- Sampling the posterior distribution with Markov chain Monte Carlo methods (Mosegaard and Tarantola, 1995).

6.2.3.7 Occam's inversion

In Occam's inversion (Constable et al., 1987), the goal is to construct the smoothest model consistent with the data. This is not to say that one believes a priori that models are really smooth, but rather that a more conservative interpretation of the data should be made by eliminating features of the model that are not required to fit the data. To this effect, they define two measures:

[3]The linearised mapping $\mathbf{g}(\mathbf{m}_{\mathrm{MAP}}) + \mathbf{G}(\mathbf{m} - \mathbf{m}_{\mathrm{MAP}})$ is usually called Fréchet derivative in inverse problem theory, and is the linear mapping tangent to $\mathbf{g}$ at $\mathbf{m}_{\mathrm{MAP}}$.

- A measure of data fit (irrespective of any Bayesian interpretation of the models):

$$d(\mathbf{m}, \mathbf{d}) \stackrel{\text{def}}{=} (\mathbf{g}(\mathbf{m}) - \mathbf{d})^T \mathbf{C}_{\mathcal{D}}^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}),$$

  that is, the Mahalanobis distance between $\mathbf{g}(\mathbf{m})$ and $\mathbf{d}$ with matrix $\mathbf{C}_{\mathcal{D}}^{-1}$.

- A measure of model smoothness: $\|\mathbf{Rm}\|$, where $\mathbf{R}$ is a Tikhonov roughening operator (Tikhonov and Arsenin, 1977), e.g. a discrete second-difference operator, such as

$$\mathbf{R} = \begin{pmatrix} -2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 \end{pmatrix}. \tag{6.4}$$

Then, Occam's inversion finds a model being both smooth and fitting well the data by solving the optimisation problem:

$$\min_{\mathbf{m} \in \mathcal{M}} \|\mathbf{Rm}\| \text{ subject to } d(\mathbf{m}, \mathbf{d}) \leq \epsilon$$

for some tolerance ϵ. Practically, due to the distance d being a quadratic form, this can be conveniently implemented as a weighted least-squares problem with a Lagrange multiplier to control the tradeoff between model smoothness and data fit: for fixed λ, solve the weighted, regularised least-squares problem

$$\min_{\mathbf{m} \in \mathcal{M}} (\mathbf{g}(\mathbf{m}) - \mathbf{d})^T \mathbf{C}_{\mathcal{D}}^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}) + \lambda (\mathbf{m} - \mathbf{m}_{\text{prior}})^T \mathbf{R}^T \mathbf{R} (\mathbf{m} - \mathbf{m}_{\text{prior}}). \tag{6.5}$$

Then, increase λ until $(\mathbf{g}(\mathbf{m}) - \mathbf{d})^T \mathbf{C}_{\mathcal{D}}^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}) > \epsilon$.

Clearly, Occam's inversion is a particular case of Bayesian inversion, in which the uncertainty distributions are taken as Gaussians (with the appropriate covariance matrix) and the prior distribution over the models is used as a smoothness regularisation term by taking $\mathbf{C}_{\mathcal{M}}^{-1} = \lambda \mathbf{R}^T \mathbf{R}$. However, Bayesian inversion is more general than Occam's inversion in that the prior distributions allow to introduce physical knowledge into the problem. Gouveia and Scales (1997) compare Bayes' and Occam's inversion in a seismic data problem.

6.2.3.8 Locally independent inverse problems

In the general statement of inverse problems, the data parameters $\mathbf{d}$ depend on all the model parameters $\mathbf{m}$ which in turn are a continuous function of some independent variables, such as the spatial coordinates $\mathbf{x}$. Thus $\mathbf{m} = \mathbf{m}(\mathbf{x})$ and $\mathbf{d} = \mathbf{g}(\mathbf{m})$. A single datum parameter depends on the whole function $\mathbf{m}(\cdot)$, even if that datum was measured at point $\mathbf{x}$ only.

Sometimes we can assume locally independent problems, so that a datum measured at point $\mathbf{x}$ depends only on the value of $\mathbf{m}(\mathbf{x})$, not on the whole function $\mathbf{m}$ for all $\mathbf{x}$. If we have N measurements $\{\mathbf{d}_n\}_{n=1}^N$ at points $\{\mathbf{x}_n\}_{n=1}^N$ and we discretise the problem, so that we have one model parameter $\mathbf{m}_n$ at point $\mathbf{x}_n$, for $n = 1, \dots, N$, then:

$$\theta(\mathbf{d}|\mathbf{m}) = \prod_{n=1}^{N} \vartheta(\mathbf{d}_n|\mathbf{m}_n) \Longrightarrow \theta(\mathbf{d}, \mathbf{m}) \propto \rho_{\mathcal{M}}(\mathbf{m}) \prod_{n=1}^{N} \vartheta(\mathbf{d}_n|\mathbf{m}_n) \tag{6.6}$$

where the distribution ϑ is the same for all parameters because the function $\mathbf{g}$ is now the same for all values of $\mathbf{m}$. This is equivalent to inverting the mapping $\mathbf{x} \to \mathbf{m}(\mathbf{x}) \xrightarrow{\mathbf{g}} \mathbf{d}(\mathbf{m}(\mathbf{x}))$ at data values $\mathbf{d}_1, \dots, \mathbf{d}_N$ and obtaining values $\mathbf{m}_1 = \mathbf{g}^{-1}(\mathbf{d}_1), \dots, \mathbf{m}_N = \mathbf{g}^{-1}(\mathbf{d}_N)$. This approach is followed in the example of section 6.2.4.2. We deal with problems of this kind in section 6.4.1 and give examples. However, this simplification cannot be applied generally. For example, for the CAT problem (section 6.2.4.3) a measurement depends on all the points that they ray travels through.

If we also assume an independent prior distribution for the parameters, $\rho_{\mathcal{M}}(\mathbf{m}) = \prod_{n=1}^{N} \varrho_{\mathcal{M}}(\mathbf{m}_n)$, then the complete inverse problem factorises into N independent problems:

$$\theta(\mathbf{d}|\mathbf{m}) \propto \prod_{n=1}^{N} \varrho_{\mathcal{M}}(\mathbf{m}_n) \vartheta(\mathbf{d}_n|\mathbf{m}_n).$$

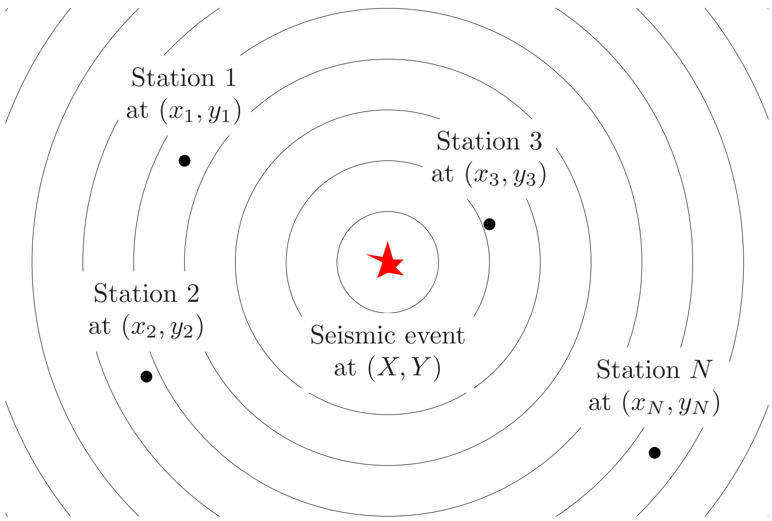

Figure 6.1: A seismic event takes place at time $\tau = 0$ at location (X, Y) and the seismic waves produced are recorded by several seismic stations of Cartesian coordinates $\{(x_n, y_n)\}_{n=1}^N$ at times $\{d_n\}_{n=1}^N$. Determining the location of the epicentre from the wave arrival times is an inverse problem.

6.2.4 Examples of inverse problems

To clarify the concepts exposed, we briefly review some examples of inverse problems, some of which have a rich literature.

6.2.4.1 Locating the epicentre of a seismic event

We consider the simplified problem[4] of estimating the epicentral coordinates of a seismic event (e.g. a nuclear explosion), depicted in fig. 6.1. The event takes place at time $\tau = 0$ at an unknown location (X, Y) on the surface of the Earth (considered flat). The seismic waves produced by the explosion are recorded in a network of N seismic stations of Cartesian coordinates $\{(x_n, y_n)\}_{n=1}^N$, so that $\mathbf{d}_n = d_n$ is the observed arrival time of the seismic wave at station n. The waves travel at a velocity v in all directions.

The model parameters to be determined from the data parameters $\{d_n\}_{n=1}^N$ are:

- State parameters: the coordinates of the epicentre (X, Y).
- Experimental condition parameters: the coordinates of each station (x_n, y_n), the time of the event τ and the wave velocity v.

Thus $\mathbf{m} = (\mathbf{s}, \mathbf{c}_1, \dots, \mathbf{c}_N) = (X, Y, \tau, v, x_1, y_1, \dots, x_N, y_N)$. Assuming that the station coordinates, the time of the event and the wave velocity are perfectly known, we can drop them and avoid defining prior distributions for them, so that $\mathbf{m} = (X, Y)$.

Given (X, Y), the arrival times of the seismic wave at the stations can be computed exactly as $\mathbf{d}_n = \mathbf{g}_n(X, Y) = \frac{1}{v}\sqrt{(x_n - X)^2 + (y_n - Y)^2}$ for $n = 1, \dots, N$, which solves the forward problem. Determining the epicentre coordinates (X, Y) from the arrival times at the different stations is the inverse problem, whose complete solution is given by Tarantola (1987).

6.2.4.2 Retrieval of scatterometer wind fields

A satellite can measure the amount of backscatter generated by small ripples on the ocean surface, in turn produced by oceanic wind fields. The satellite scatterometer consists of a line of cells, each capable to detect backscatter at the location to which it is pointing. At a given position in space of the satellite, each cell records a *local measurement*. A *field* is defined as a spatially connected set of local measurements obtained from a swathe, swept by the satellite along its orbit. For example, for the ESA satellite ERS-1, which follows a polar

[4] This example is adapted from problem 1.1 of Tarantola (1987, pp. 85–91).

orbit, the swathe contains 19 cells and is approximately 500 km wide. Each cell samples an area of around 50×50 km, with some overlap between samples.

The *backscatter* $\boldsymbol{\sigma}^0$ is a three-dimensional vector because each cell is sampled from three different directions by the fore, mid and aft beams, respectively. The near-surface *wind vector* $\mathbf{u}$ is a two-dimensional, quasicontinuous function of the oceanic spatial coordinates (although see comments about the wind continuity in section 7.9.6). Both $\boldsymbol{\sigma}^0$ and $\mathbf{u}$ contain noise, although the noise in $\boldsymbol{\sigma}^0$ is dominated by that of $\mathbf{u}$. A backscatter field is written as $\boldsymbol{\Sigma}^0 = (\boldsymbol{\sigma}_i^0)$ and a wind field as $\mathbf{U} = (\mathbf{u}_i)$.

The forward problem is to obtain $\boldsymbol{\sigma}^0$ from $\mathbf{u}$ and is single-valued and relatively easy to solve. The inverse problem, to obtain the wind field from the backscatter, is one-to-many and no realistic physically-based local inverse model is possible. The aim of the inversion is to produce a wind field $\mathbf{U}$ that can be used in data assimilation for numerical weather prediction (NWP) models. The most common method for inversion is to use lookup tables and interpolation. Following the standard Bayesian approach of inverse problem theory described in section 6.2.3, Cornford and colleagues[5] (Cornford et al., 1999a; Nabney et al., 2000; Evans et al., 2000) model the conditional distribution $p(\boldsymbol{\Sigma}^0|\mathbf{U})$ and the prior distribution of the wind fields $p(\mathbf{U})$. The prior is taken as a zero-mean normal, $p(\mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{U}})$. The conditional distribution of the backscatter field $\boldsymbol{\Sigma}^0$ given the wind field $\mathbf{U}$ can be factorised into the individual distributions at each point in the region (i.e., at each cell) as $p(\boldsymbol{\Sigma}^0|\mathbf{U}) = \prod_i p(\boldsymbol{\sigma}_i^0|\mathbf{u}_i)$ because theoretically there is a single-valued mapping $\mathbf{u} \rightarrow \boldsymbol{\sigma}^0$. However, rather than using a physical forward model to obtain a noise model $p(\boldsymbol{\sigma}^0|\mathbf{u})$, which is difficult, they use Bayes' theorem to obtain $p(\boldsymbol{\sigma}^0|\mathbf{u}) \propto p(\mathbf{u}|\boldsymbol{\sigma}^0)/p(\mathbf{u})$, the factor $p(\boldsymbol{\sigma}^0)$ being constant for a given data, and they model $p(\mathbf{u}|\boldsymbol{\sigma}^0)$ as a mixture density network (Bishop, 1994), which is basically a universal approximator for conditional densities (see section 7.11.3). Applying Bayes' theorem again, the posterior distribution is

$$p(\mathbf{U}|\boldsymbol{\Sigma}^0) \propto p(\mathbf{U}) \prod_i \frac{p(\mathbf{u}_i|\boldsymbol{\sigma}_i^0)}{p(\mathbf{u}_i)}.$$

Given a backscatter field $\boldsymbol{\Sigma}^0$, the corresponding wind field is determined by MAP: a mode of $p(\mathbf{U}|\boldsymbol{\Sigma}^0)$ is found using a conjugate gradients method.

The fact that this inverse problem can be factorised into independent mapping inversion problems (see section 6.4.1) and the quasicontinuous dependence of the wind on the space coordinates make this problem amenable to the technique described in chapter 7.

6.2.4.3 Computerised tomography

The aim of computerised tomography (Herman, 1980) is to reconstruct the spatially varying absorption coefficients within a medium (e.g. the human body) from measurements of intensity decays of X-rays sent through the medium. Typically, X-rays are sent between a point source and a point receiver which counts the number of photons not absorbed by the medium, thus giving an indication of the integrated attenuation coefficient along that particular ray path (fig. 6.2). Repeating the measurement for many different ray paths, conveniently sampling the medium, the spatial structure of the attenuation coefficient can be inferred and so an image of the medium can be obtained.

The transmittance ρ_n (the probability of a photon of being transmitted) along the nth ray is given by:

$$\rho_n \stackrel{\text{def}}{=} \exp\left(-\int_{R_n} \mathbf{m}(\mathbf{x}(s_n))\, ds_n\right)$$

where:

- $\mathbf{m}(\mathbf{x})$ is the linear attenuation coefficient at point $\mathbf{x}$ and corresponds to the probability per unit length of path of a photon arriving at $\mathbf{x}$ being absorbed.
- R_n is the ray path, identified by the coordinates of the X-ray source and its shooting angle (ray paths of X-rays through an animal body can be assimilated to straight lines with an excellent approximation).
- ds_n is the element of length along the ray path.
- $\mathbf{x}(s_n)$ is the current point considered in the line integral along the ray (in Cartesian or spherical coordinates).

[5] Papers and software can be found in the NEUROSAT project web page at `http://www.ncrg.aston.ac.uk/Projects/NEUROSAT`.

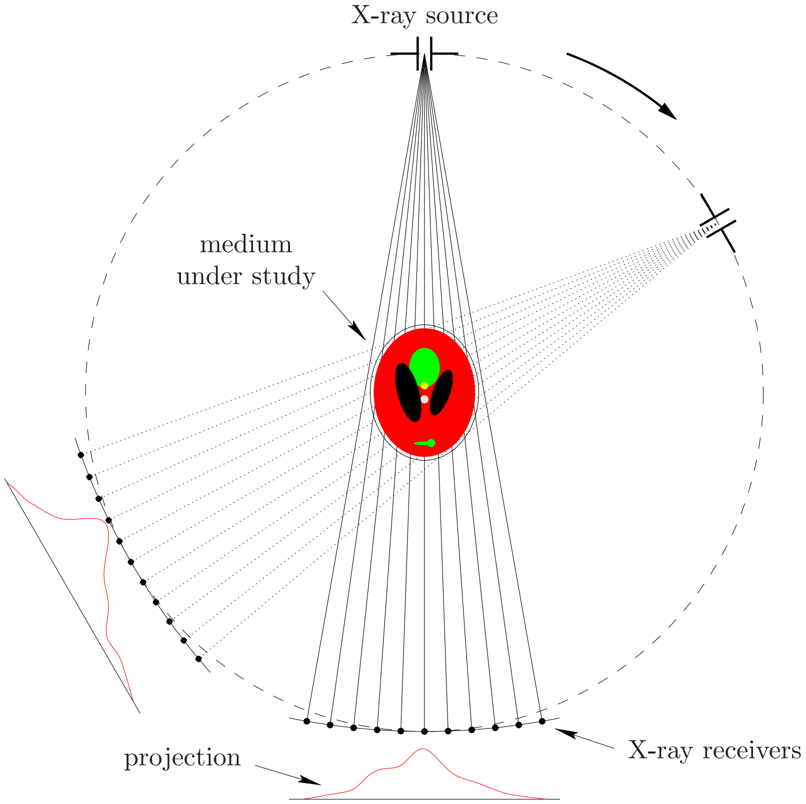

Figure 6.2: Setup for 2D X-ray tomography. A source sends a beam of X-rays through the object under study. Each individual X-ray is attenuated differently according to its path through the object. The X-rays are measured by an array of receivers, thus providing with a projection of the object. By rotating the source or having several sources surrounding the object we obtain several projections. Reconstructing the object density from these projections is the inverse problem of tomography.

Defining the data

$$d_n \overset{\text{def}}{=} -\ln \rho_n = \int_{R_n} \mathbf{m}(\mathbf{x}(s_n))\, ds_n \tag{6.7}$$

gives a linear relation between the data d_n and the unknown function $\mathbf{m}(\mathbf{x})$. Eq. (6.7) is the Radon transform of the function $\mathbf{m}(\mathbf{x})$, so that the tomography problem is the problem of inverting the Radon transform.

Thus, here the model parameters are the attenuation $\mathbf{m}(\mathbf{x})$ in the continuous case, or $\mathbf{m} = (m_{ijk})$ in a discretised version, and the observed parameters are the measured log-transmittance d_n. Given $\mathbf{m}(\mathbf{x})$, the forward problem is solved by the linear equation (6.7).

Similar problems appear in non-destructive testing and in geophysics. For example, in *geophysical acoustic tomography* the aim is to compute the acoustic structure of a region inside the Earth from seismic measurements. This allows, for example, to detect gas or oil deposits or to determine the radius of the Earth's metallic core. Acoustic waves are generated by sources at different positions inside a borehole and the travel times of the first wave front to receivers located in other boreholes around the region under study are recorded. The main difference with X-ray tomography is that the ray paths are not straight (they depend on the medium structure and are diffracted and reflected at boundaries), which makes the forward problem nonlinear. Acoustic tomography is an inverse scattering problem, in which one wants to determine the shape or the location of an obstacle from measurements of waves scattered by the obstacle.

Unlike example 6.2.4.2, this inverse problem does not factorise into independent mapping inversion problems and thus is not approachable by the technique of chapter 7.

6.3 Inverse problems vs Bayesian analysis in general

In Bayesian analysis in general, a parametric model is a function $p(\mathbf{x}; \mathbf{\Theta})$ where $\mathbf{x}$ are the variables of interest in the problem and $\mathbf{\Theta}$ the parameters, which identify the model. One is interested in inferences about both

Bayesian inverse problem theory	Latent variable models
Data space $\mathcal{D}$	Observed or data space $\mathcal{T}$
Model space $\mathcal{M}$	Latent space $\mathcal{X}$
Prior distribution of models $\rho_{\mathcal{M}}(\mathbf{m})$	Prior distribution in latent space $p(\mathbf{x})$
Forward mapping $\mathbf{g} : \mathcal{M} \to \mathcal{D}$	Mapping from latent space onto data space $\mathbf{f} : \mathcal{X} \to \mathcal{T}$
Uncertainty in the forward mapping $\theta(\mathbf{d}\|\mathbf{m}) = f(\mathbf{d} - \mathbf{g}(\mathbf{m}))$	Noise model $p(\mathbf{t}\|\mathbf{x}) = p(\mathbf{t}\|\mathbf{f}(\mathbf{x}))$

Table 6.1: Formal correspondence between continuous latent variable models and Bayesian inverse problem theory.

the parameters and the data, e.g. prediction via conditional distributions $p(x_2|x_1)$, etc. Bayesian inference is often approximated by fixing the parameters to a certain value given a sample $\{\mathbf{x}_n\}_{n=1}^N$ of the data, e.g. via maximum a posteriori (MAP) estimation:

$$\mathbf{\Theta}_{\mathrm{MAP}} \stackrel{\mathrm{def}}{=} \arg\max_{\mathbf{\Theta}} p(\mathbf{\Theta}|\mathbf{x}) = \arg\max_{\mathbf{\Theta}} p(\mathbf{x}|\mathbf{\Theta})p(\mathbf{\Theta}). \tag{6.8}$$

The model parameters $\mathbf{m}$ and the observed parameters $\mathbf{d}$ of inverse problem theory correspond to the parameters $\mathbf{\Theta}$ and the problem variables $\mathbf{x}$, respectively, of the Bayesian analysis in general. Bayesian inference about the model parameters $\mathbf{m}$, as shown in section 6.2.3, coincides with equation 6.8. But the emphasis is solely in inferences about the parameter estimates, i.e., how to find a single value of the model parameters that hopefully approximates well the physical reality. Thus, inverse problem theory is a *one-shot inversion problem*: use as much data as required to find a single value $\mathbf{m}_{\mathrm{MAP}}$. Even if a second inversion is performed, we would expect the new inverse value of the model parameters to be close to the previous one—assuming that the system has not changed, i.e., assuming it is stationary or considering it in a fixed moment of time.

6.3.1 Inverse problems vs latent variable models

As an interesting example of the differences in interpretation between inverse problem theory and general Bayesian inference, let us consider continuous latent variable models as defined in chapter 2. As mentioned before, estimating the parameters of any probabilistic model can be seen as an inverse problem. But even when these parameters have been estimated and fixed, there is a formal parallelism between latent variable models and Bayesian inverse problem theory (in fact, all the estimation formulas for factor analysis mirror those of linear inverse problem theory). In latent variable models, we observe the data variables $\mathbf{t} \in \mathcal{T}$ and postulate a low-dimensional space $\mathcal{X}$ with a prior distribution $p(\mathbf{x})$, a mapping from latent space onto data space $\mathbf{f} : \mathcal{X} \to \mathcal{T}$ and a noise model in data space $p(\mathbf{t}|\mathbf{x}) = p(\mathbf{t}|\mathbf{f}(\mathbf{x}))$. Thus, the *model* here means the whole combined choice of the prior distribution in latent space, $p(\mathbf{x})$, the noise model, $p(\mathbf{t}|\mathbf{x})$, and the mapping from latent onto data space, $\mathbf{f}$, as well as the dimensionality of the latent space, L. And all these elements are equipped with parameters (collectively written as $\mathbf{\Theta}$) that are estimated from a sample in data space, $\{\mathbf{t}_n\}_{n=1}^N$. Table 6.1 summarises the formal correspondence between continuous latent variable models and Bayesian inverse problem theory.

The choice of model parameters to describe a system in inverse problem theory is not unique in general, and a particular choice of model parameters is a *parameterisation* of the system, or a *coordinate system*. Two different parameterisations are equivalent if they are related by a bijection. Physical knowledge of the inverse problem helps to choose the right parameterisation and the right forward model. However, what really matters is the combination of both the prior over the models and the forward mapping, $\rho_{\mathcal{M}}(\mathbf{m})\theta(\mathbf{d}|\mathbf{m})$, because this gives the solution to the Bayesian inversion. The same happens in latent variable models: what matters is the density in observed space $p(\mathbf{t})$, which is the only observable of the problem, rather than the particular conceptualisation (latent space plus prior plus mapping) that we choose. Thus, we are reasonably free to choose a simple prior in latent space if we can have a universal approximator as mapping $\mathbf{f}$, so that a large class of $p(\mathbf{t})$ can be constructed. Of course, this does not preclude using specific functional forms of the mapping and distributions if the knowledge about the problem suggests so.

We said earlier that inverse problem theory is a one-shot problem in that given a data set one inverts the forward mapping once to obtain a unique model. In continuous latent variable models, the latent variables are interpreted as *state variables*, which can take any value in their domain and that determine the observed variables of the system (up to noise). That is, when the system is in the state $\mathbf{x}$, the observed data is $\mathbf{f}(\mathbf{x})$ (plus the noise). The model remains the same (same parameters $\mathbf{\Theta}$, same prior distribution, etc.) but the marginal distribution of the latent and observed variables $p(\mathbf{x}, \mathbf{t})$ can be used to make inferences about $\mathbf{t}$ given $\mathbf{x}$ (forward problem, related to the deterministic and now estimated mapping $\mathbf{f} : \mathcal{X} \to \mathcal{T}$) and about $\mathbf{x}$ given $\mathbf{t}$ (inverse problem, or dimensionality reduction), for different values of $\mathbf{x}$ and $\mathbf{t}$. Thus, once the latent variable model has been fixed (which requires estimating any parameters it may contain, given a sample in data space) it may be applied any number of times to different observed data and give completely different posterior distributions in latent space, $p(\mathbf{x}|\mathbf{t})$—unlike in inverse problem theory, where different data sets are expected to correspond to the same model.

The particular case of independent component analysis (ICA), discussed in section 2.6.3, cannot be considered as an inverse problem because we do not have a forward model to invert. Even though we are looking for the inverse of the mixing matrix $\mathbf{\Lambda}$ (so that we can obtain the sources $\mathbf{x}$ given the sensor outputs $\mathbf{t}$), $\mathbf{\Lambda}$ is unknown. ICA finds a particular linear transformation $\mathbf{A}$ and a nonlinear function f that make the sources independent, but we do not either invert a function (the linear transformation $\mathbf{\Lambda}$) or estimate an inverse from input-output data ($\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$).

6.4 Mapping inversion

Consider a function[6] $\mathbf{f}$ between sets $\mathcal{X}$ and $\mathcal{Y}$ (usually subsets of $\mathbb{R}^D$):

$$\begin{array}{rccc} \mathbf{f}: & \mathcal{X} & \longrightarrow & \mathcal{Y} \\ & \mathbf{x} & \mapsto & \mathbf{y} = \mathbf{f}(\mathbf{x}). \end{array}$$

Mapping inversion is the problem of computing the inverse $\mathbf{x}$ of any $\mathbf{y} \in \mathcal{Y}$:

$$\begin{array}{rccc} \mathbf{f}^{-1}: & \mathcal{Y} & \longrightarrow & \mathcal{X} \\ & \mathbf{y} & \mapsto & \mathbf{x} = \mathbf{f}^{-1}(\mathbf{y}). \end{array}$$

$\mathbf{f}^{-1}$ may not be a well-defined function: for some $\mathbf{y} \in \mathcal{Y}$, it may not exist or may not be unique. When $\mathbf{f}^{-1}$ is to be determined from a training set of pairs $\{(\mathbf{y}_n, \mathbf{x}_n)\}_{n=1}^N$, perhaps obtained by sampling $\mathcal{X}$ and applying a known function $\mathbf{f}$, the problem is indistinguishable from **mapping approximation** from data: *given a collection of input-output pairs, construct a mapping that best transforms the inputs into the outputs.*

A **universal mapping approximator** (UMA) for a given class of functions $\mathcal{F}$ from $\mathbb{R}^L$ to $\mathbb{R}^D$ is a set $\mathcal{U}$ of functions which contains functions arbitrarily close (in the squared Euclidean distance sense, for definiteness) to any function in $\mathcal{F}$, i.e., any function in $\mathcal{F}$ can be approximated as accurately as desired by a function in $\mathcal{U}$. For example, the class of multilayer perceptrons with one or more layers of hidden units with sigmoidal activation function or the class of Gaussian radial basis function networks are universal approximators for continuous functions in a compact set of $\mathbb{R}^D$ (see Scarselli and Tsoi, 1998 for a review). The functions in $\mathcal{U}$ will usually be parametric and the optimal parameter values can be found using a learning algorithm. There are important issues in statistical mapping approximation, like the existence of local minima of the error function, the reachability of the global minimum and the generalisation to unseen data. But for our purposes in this last part of the thesis what matters is that several kinds of UMAs exist, in particular the multilayer perceptron (MLP), for which practical training algorithms exist, like backpropagation.

A multivalued mapping assigns several images to the same domain point and is therefore not a function in the mathematical sense. Among other cases, multivalued mappings arise when computing the inverse of an injective mapping (i.e., a mapping that maps different domain points onto a same image point)—a very common situation. That is, if the direct or forward mapping verifies $\mathbf{f}(\mathbf{x}_1) = \mathbf{f}(\mathbf{x}_2) = \mathbf{y}$ then both $\mathbf{x}_1$ and $\mathbf{x}_2$ are inverse values of $\mathbf{y}$: $\mathbf{f}^{-1}(\mathbf{y}) \supseteq \{\mathbf{x}_1, \mathbf{x}_2\}$. UMAs work well with univalued mappings but not with multivalued mappings. In chapter 7 we give a method for reconstruction of missing data that applies as a particular case to multivalued mappings, and we compare it to UMAs as well as other approaches for mapping approximation like vector quantisation (section 7.11.4) and conditional modelling (section 7.11.3).

6.4.1 Inverse problems vs mapping inversion

Mapping inversion is a different problem from that of inverse problem theory: in inverse problem theory, one is interested in obtaining a unique inverse point $\mathbf{x}$ which represents a model of a physical system—of which

[6] We use the term *function* in its strict mathematical sense: a correspondence that, given an element $\mathbf{x} \in \mathcal{X}$, assigns to it one and only one element $\mathbf{y} \in \mathcal{Y}$. We use the term *mapping* for a correspondence which may be multivalued (one-to-many).

we observed the values $\mathbf{y} \in \mathcal{Y}$. In mapping inversion, we want to obtain an inverse mapping $\mathbf{f}^{-1}$ that we can use many times to invert different values $\mathbf{y} \in \mathcal{Y}$.

In section 6.2.3.8 we saw that the Bayesian inverse problem theory could be recast to solve such a mapping inversion problem. We rewrite eq. (6.6) with a simplification of the notation and decompose the model parameters $\mathbf{m}$ into parameters of *state* $\mathbf{s}$ and *experimental conditions* $\mathbf{c}_n$, $\mathbf{m} = (\mathbf{s}, \mathbf{c}_1, \dots, \mathbf{c}_N)$:

$$p(\mathbf{s}, \mathbf{c}_1, \dots, \mathbf{c}_N | \mathbf{d}_1, \dots, \mathbf{d}_N) \propto p(\mathbf{s}, \mathbf{c}_1, \dots, \mathbf{c}_N) \prod_{n=1}^{N} p(\mathbf{d}_n | \mathbf{s}, \mathbf{c}_n)$$

where $p(\mathbf{d}_n|\mathbf{s}, \mathbf{c}_n) = f(\mathbf{d}_n - \mathbf{g}(\mathbf{c}_n; \mathbf{s}))$ and $\mathbf{g}$ is the known forward mapping. This can be interpreted as a function $\mathbf{g}$ which maps $\mathbf{c}_n$ into $\mathbf{d}_n$ and has trainable parameters $\mathbf{s}$. However, $\{\mathbf{c}_n\}_{n=1}^N$ are *unknown parameters themselves, not data.* We are interested in constructing an inverse mapping $\mathbf{g}^{-1}$ given a data set consisting of pairs of values $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ so that $\mathbf{y}_n = \mathbf{g}(\mathbf{x}_n)$ for a mapping $\mathbf{g}$ (not necessarily known). Clearly, in these terms the distinction between inverse and forward mapping disappears and the problem, as before, becomes a problem of mapping approximation.

Practitioners of inverse problem theory may object that by considering the forward mapping unknown we are throwing away all the physical information. But the theorems about universal approximation of mappings and about universal approximation of probability density functions support the fact that, given enough data, we can obtain (ideally) a good approximation of the joint density of the observed data and thus capture the information about the forward mapping too. This has the added flexibility of making inferences about any group of variables given any other group of variables by constructing the appropriate conditional distribution from the joint density—which includes both the forward and inverse mappings.

Two well-known examples of mapping inversion problems with one-to-many inverse mappings (often referred to as inverse problems in the literature) are the **inverse kinematics problem of manipulators** (the robot arm problem) (Atkeson, 1989) and the **acoustic-to-articulatory mapping problem** of speech (Schroeter and Sondhi, 1994). We describe them and apply to them our own algorithm later in this thesis.

Bibliography

S. Amari. Natural gradient learning for over- and under-complete bases in ICA. *Neural Computation*, 11(8): 1875–1883, Nov. 1999.

S. Amari and A. Cichoki. Adaptive blind signal processing—neural network approaches. *Proc. IEEE*, 86(10): 2026–2048, Oct. 1998.

T. W. Anderson. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34 (1):122–148, Mar. 1963.

T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In J. Neyman, editor, *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, volume V, pages 111–150, Berkeley, 1956. University of California Press.

S. Arnfield. Artificial EPG palate image. The Reading EPG, 1995. Available online at `http://www.linguistics.reading.ac.uk/research/speechlab/epg/palate.jpg`, Feb. 1, 2000.

H. Asada and J.-J. E. Slotine. *Robot Analysis and Control*. John Wiley & Sons, New York, London, Sydney, 1986.

D. Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.*, 6:128–143, 1985.

B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoustic Soc. Amer.*, 63(5):1535–1555, May 1978.

C. G. Atkeson. Learning arm kinematics and dynamics. *Annu. Rev. Neurosci.*, 12:157–183, 1989.

H. Attias. EM algorithms for independent component analysis. In Niranjan (1998), pages 132–141.

H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, May 1999.

F. Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, Sept. 1991.

A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, B*, 61(3):579–602, 1999.

R. J. Baddeley. Searching for filters with "interesting" output distributions: An uninteresting direction to explore? *Network: Computation in Neural Systems*, 7(2):409–421, 1996.

R. Bakis. Coarticulation modeling with continuous-state HMMs. In *Proc. IEEE Workshop Automatic Speech Recognition*, pages 20–21, Arden House, New York, 1991. Harriman.

R. Bakis. An articulatory-like speech production model with controlled use of prior knowledge. Frontiers in Speech Processing: Robust Speech Analysis '93, Workshop CDROM, NIST Speech Disc 15 (also available from the Linguistic Data Consortium), Aug. 6 1993.

P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *J. Amer. Stat. Assoc.*, 87(417):7–16, Mar. 1992.

J. Barker, L. Josifovski, M. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. of the International Conference on Spoken Language Processing (ICSLP'00)*, Beijing, China, Oct. 16–20 2000.

J. P. Barker and F. Berthommier. Evidence of correlation between acoustic and visual features of speech. In Ohala et al. (1999), pages 199–202.

M. F. Barnsley. *Fractals Everywhere*. Academic Press, New York, 1988.

D. J. Bartholomew. The foundations of factor analysis. *Biometrika*, 71(2):221–232, Aug. 1984.

D. J. Bartholomew. Foundations of factor analysis: Some practical implications. *Brit. J. of Mathematical and Statistical Psychology*, 38:1–10 (discussion in pp. 127–140), 1985.

D. J. Bartholomew. *Latent Variable Models and Factor Analysis*. Charles Griffin & Company Ltd., London, 1987.

A. Basilevsky. *Statistical Factor Analysis and Related Methods*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1994.

H.-U. Bauer, M. Herrmann, and T. Villmann. Neural maps and topographic vector quantization. *Neural Networks*, 12(4–5):659–676, June 1999.

H.-U. Bauer and K. R. Pawelzik. Quantifying the neighbourhood preservation of self-organizing feature maps. *IEEE Trans. Neural Networks*, 3(4):570–579, July 1992.

J. Behboodian. On the modes of a mixture of two normal distributions. *Technometrics*, 12(1):131–139, Feb. 1970.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338, Dec. 1997.

R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, 1961.

Y. Bengio and F. Gingras. Recurrent neural networks for missing or asynchronous data. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 395–401. MIT Press, Cambridge, MA, 1996.

C. Benoît, M.-T. Lallouache, T. Mohamadi, and C. Abry. A set of French visemes for visual speech synthesis. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Designs*, pages 485–504. North Holland-Elsevier Science Publishers, Amsterdam, New York, Oxford, 1992.

P. M. Bentler and J. S. Tanaka. Problems with EM algorithms for ML factor analysis. *Psychometrika*, 48(2): 247–251, June 1983.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, Berlin, second edition, 1985.

M. Berkane, editor. *Latent Variable Modeling and Applications to Causality*. Number 120 in Springer Series in Statistics. Springer-Verlag, Berlin, 1997.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester, 1994.

N. Bernstein. *The Coordination and Regulation of Movements*. Pergamon, Oxford, 1967.

D. P. Bertsekas. *Dynamic Programming. Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, N.J., 1987.

J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, B*, 55(1):25–37, 1993.

J. C. Bezdek and N. R. Pal. An index of topological preservation for feature extraction. *Pattern Recognition*, 28(3):381–391, Mar. 1995.

E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1):32–48, Jan. 1982.

C. M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Neural Computing Research Group, Aston University, Feb. 1994. Available online at `http://www.ncrg.aston.ac.uk/Papers/postscript/NCRG_94_004.ps.Z`.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, Oxford, 1995.

C. M. Bishop. Bayesian PCA. In Kearns et al. (1999), pages 382–388.

C. M. Bishop, G. E. Hinton, and I. G. D. Strachan. GTM through time. In *IEE Fifth International Conference on Artificial Neural Networks*, pages 111–116, 1997a.

C. M. Bishop and I. T. Nabney. Modeling conditional probability distributions for periodic variables. *Neural Computation*, 8(5):1123–1133, July 1996.

C. M. Bishop, M. Svensén, and C. K. I. Williams. Magnification factors for the SOM and GTM algorithms. In *WSOM'97: Workshop on Self-Organizing Maps*, pages 333–338, Finland, June 4–6 1997b. Helsinki University of Technology.

C. M. Bishop, M. Svensén, and C. K. I. Williams. Developments of the generative topographic mapping. *Neurocomputing*, 21(1–3):203–224, Nov. 1998a.

C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, Jan. 1998b.

C. M. Bishop and M. E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 20(3):281–293, Mar. 1998.

A. Bjerhammar. *Theory of Errors and Generalized Matrix Inverses.* North Holland-Elsevier Science Publishers, Amsterdam, New York, Oxford, 1973.

C. S. Blackburn and S. Young. A self-learning predictive model of articulator movements during speech production. *J. Acoustic Soc. Amer.*, 107(3):1659–1670, Mar. 2000.

T. L. Boullion and P. L. Odell. *Generalized Inverse Matrices.* John Wiley & Sons, New York, London, Sydney, 1971.

H. Bourlard and Y. Kamp. Autoassociation by the multilayer perceptrons and singular value decomposition. *Biol. Cybern.*, 59(4–5):291–294, 1988.

H. Bourlard and N. Morgan. *Connectionist Speech Recognition. A Hybrid Approach.* Kluwer Academic Publishers Group, Dordrecht, The Netherlands, 1994.

M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11(5):1155–1182, July 1999.

C. Bregler and S. M. Omohundro. Surface learning with applications to lip-reading. In Cowan et al. (1994), pages 43–50.

C. Bregler and S. M. Omohundro. Nonlinear image interpolation using manifold learning. In Tesauro et al. (1995), pages 973–980.

L. J. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug. 1996.

S. P. Brooks. Markov chain Monte Carlo method and its application. *The Statistician*, 47(1):69–100, 1998.

C. P. Browman and L. M. Goldstein. Articulatory phonology: An overview. *Phonetica*, 49(3–4):155–180, 1992.

E. N. Brown, L. M. Frank, D. Tang, M. C. Quirk, and M. A. Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *J. Neurosci.*, 18(18):7411–7425, Sept. 1998.

G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8(4): 297–336, Oct. 1994.

D. Byrd, E. Flemming, C. A. Mueller, and C. C. Tan. Using regions and indices in EPG data reduction. *Journal of Speech and Hearing Research*, 38(4):821–827, Aug. 1995.

J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Letters on Signal Processing*, 4(4):112–114, Apr. 1997.

J.-F. Cardoso. Blind signal separation: Statistical principles. *Proc. IEEE*, 86(10):2009–2025, Oct. 1998.

M. Á. Carreira-Perpiñán. A review of dimension reduction techniques. Technical Report CS–96–09, Dept. of Computer Science, University of Sheffield, UK, Dec. 1996. Available online at `http://www.dcs.shef.ac.uk/~miguel/papers/cs-96-09.html`.

M. Á. Carreira-Perpiñán. Density networks for dimension reduction of continuous data: Analytical solutions. Technical Report CS–97–09, Dept. of Computer Science, University of Sheffield, UK, Apr. 1997. Available online at `http://www.dcs.shef.ac.uk/~miguel/papers/cs-97-09.html`.

M. Á. Carreira-Perpiñán. Mode-finding for mixtures of Gaussian distributions. Technical Report CS–99–03, Dept. of Computer Science, University of Sheffield, UK, Mar. 1999a. Revised August 4, 2000. Available online at `http://www.dcs.shef.ac.uk/~miguel/papers/cs-99-03.html`.

M. Á. Carreira-Perpiñán. One-to-many mappings, continuity constraints and latent variable models. In *Proc. of the IEE Colloquium on Applied Statistical Pattern Recognition*, Birmingham, UK, 1999b.

M. Á. Carreira-Perpiñán. Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 22(11):1318–1323, Nov. 2000a.

M. Á. Carreira-Perpiñán. Reconstruction of sequential data with probabilistic models and continuity constraints. In Solla et al. (2000), pages 414–420.

M. Á. Carreira-Perpiñán and S. Renals. Dimensionality reduction of electropalatographic data using latent variable models. *Speech Communication*, 26(4):259–282, Dec. 1998a.

M. Á. Carreira-Perpiñán and S. Renals. Experimental evaluation of latent variable models for dimensionality reduction. In Niranjan (1998), pages 165–173.

M. Á. Carreira-Perpiñán and S. Renals. A latent variable modelling approach to the acoustic-to-articulatory mapping problem. In Ohala et al. (1999), pages 2013–2016.

M. Á. Carreira-Perpiñán and S. Renals. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12(1):141–152, Jan. 2000.

J. Casti. Flight over Wall St. *New Scientist*, 154(2078):38–41, Apr. 19 1997.

T. Chen and R. R. Rao. Audio-visual integration in multimodal communication. *Proc. IEEE*, 86(5):837–852, May 1998.

H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *J. Amer. Stat. Assoc.*, 68(342):361–368, June 1973.

D. A. Cohn. Neural network exploration using optimal experiment design. *Neural Networks*, 9(6):1071–1083, Aug. 1996.

C. H. Coker. A model of articulatory dynamics and control. *Proc. IEEE*, 64(4):452–460, 1976.

P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36(3):287–314, Apr. 1994.

S. C. Constable, R. L. Parker, and C. G. Constable. Occam's inversion—a practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, 52(3):289–300, 1987.

D. Cook, A. Buja, and J. Cabrera. Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics*, 2(3):225–250, 1993.

M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 2000. To appear.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, June 2001.

D. Cornford, I. T. Nabney, and D. J. Evans. Bayesian retrieval of scatterometer wind fields. Technical Report NCRG/99/015, Neural Computing Research Group, Aston University, 1999a. Submitted to J. of Geophysical Research. Available online at `ftp://cs.aston.ac.uk/cornford/bayesret.ps.gz`.

D. Cornford, I. T. Nabney, and C. K. I. Williams. Modelling frontal discontinuities in wind fields. Technical Report NCRG/99/001, Neural Computing Research Group, Aston University, Jan. 1999b. Submitted to Nonparametric Statistics. Available online at `http://www.ncrg.aston.ac.uk/Papers/postscript/NCRG_99_001.ps.Z`.

R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1. Interscience Publishers, New York, 1953.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, London, Sydney, 1991.

J. D. Cowan, G. Tesauro, and J. Alspector, editors. *Advances in Neural Information Processing Systems*, volume 6, 1994. Morgan Kaufmann, San Mateo.

T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, New York, 1994.

J. J. Craig. *Introduction to Robotics. Mechanics and Control*. Series in Electrical and Computer Engineering: Control Engineering. Addison-Wesley, Reading, MA, USA, second edition, 1989.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

P. Dayan. Arbitrary elastic topologies and ocular dominance. *Neural Computation*, 5(3):392–401, 1993.

P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The Helmholtz machine. *Neural Computation*, 7(5): 889–904, Sept. 1995.

M. H. DeGroot. *Probability and Statistics*. Addison-Wesley, Reading, MA, USA, 1986.

D. DeMers and G. W. Cottrell. Non-linear dimensionality reduction. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 580–587. Morgan Kaufmann, San Mateo, 1993.

D. DeMers and K. Kreutz-Delgado. Learning global direct inverse kinematics. In J. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 589–595. Morgan Kaufmann, San Mateo, 1992.

D. DeMers and K. Kreutz-Delgado. Canonical parameterization of excess motor degrees of freedom with self-organizing maps. *IEEE Trans. Neural Networks*, 7(1):43–55, Jan. 1996.

D. DeMers and K. Kreutz-Delgado. Learning global properties of nonredundant kinematic mappings. *Int. J. of Robotics Research*, 17(5):547–560, May 1998.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

L. Deng. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, 24(4):299–323, July 1998.

L. Deng, G. Ramsay, and D. Sun. Production models as a structural basis for automatic speech recognition. *Speech Communication*, 22(2–3):93–111, Aug. 1997.

P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12(3):793–815, Sept. 1984.

K. I. Diamantaras and S.-Y. Kung. *Principal Component Neural Networks. Theory and Applications.* Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, New York, London, Sydney, 1996.

T. G. Dietterich. Machine learning research: Four current directions. *AI Magazine*, 18(4):97–136, winter 1997.

M. P. do Carmo. *Differential Geometry of Curves and Surfaces.* Prentice-Hall, Englewood Cliffs, N.J., 1976.

R. D. Dony and S. Haykin. Optimally adaptive transform coding. *IEEE Trans. on Image Processing*, 4(10): 1358–1370, Oct. 1995.

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, London, Sydney, 1973.

R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, editors. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

R. Durbin and G. Mitchison. A dimension reduction framework for understanding cortical maps. *Nature*, 343 (6259):644–647, Feb. 15 1990.

R. Durbin, R. Szeliski, and A. Yuille. An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation*, 1(3):348–358, Fall 1989.

R. Durbin and D. Willshaw. An analogue approach to the traveling salesman problem using an elastic net method. *Nature*, 326(6114):689–691, Apr. 16 1987.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer Academic Publishers Group, Dordrecht, The Netherlands, 1996.

K. Erler and G. H. Freeman. An HMM-based speech recognizer using overlapping articulatory features. *J. Acoustic Soc. Amer.*, 100(4):2500–2513, Oct. 1996.

G. Eslava and F. H. C. Marriott. Some criteria for projection pursuit. *Statistics and Computing*, 4:13–20, 1994.

C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan. Acoustic modeling of American English /r/. *J. Acoustic Soc. Amer.*, 108(1):343–356, July 2000.

J. Etezadi-Amoli and R. P. McDonald. A second generation nonlinear factor analysis. *Psychometrika*, 48(3): 315–342, Sept. 1983.

D. J. Evans, D. Cornford, and I. T. Nabney. Structured neural network modelling of multi-valued functions for wind vector retrieval from satellite scatterometer measurements. *Neurocomputing*, 30(1–4):23–30, Jan. 2000.

B. S. Everitt. *An Introduction to Latent Variable Models.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1984.

B. S. Everitt and D. J. Hand. *Finite Mixture Distributions.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1981.

K. J. Falconer. *Fractal Geometry: Mathematical Foundations and Applications.* John Wiley & Sons, Chichester, 1990.

K. Fan. On a theorem of Weyl concerning the eigenvalues of linear transformations II. *Proc. Natl. Acad. Sci. USA*, 36:31–35, 1950.

G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, Paris, second edition, 1970.

E. Farnetani, W. J. Hardcastle, and A. Marchal. Cross-language investigation of lingual coarticulatory processes using EPG. In J.-P. Tubach and J.-J. Mariani, editors, *Proc. EUROSPEECH'89*, volume 2, pages 429–432, Paris, France, Sept. 26–28 1989.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2 of *Wiley Series in Probability and Mathematical Statistics.* John Wiley & Sons, New York, London, Sydney, third edition, 1971.

D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, July 1994.

J. L. Flanagan. *Speech Analysis, Synthesis and Perception.* Number 3 in Kommunication und Kybernetik in Einzeldarstellungen. Springer-Verlag, Berlin, second edition, 1972.

M. K. Fleming and G. W. Cottrell. Categorization of faces using unsupervised feature extraction. In *Proc. Int. J. Conf. on Neural Networks (IJCNN90)*, volume II, pages 65–70, San Diego, CA, June 17–21 1990.

P. Földiák. Adaptive network for optimal linear feature extraction. In *Proc. Int. J. Conf. on Neural Networks (IJCNN89)*, volume I, pages 401–405, Washington, DC, June 18–22 1989.

D. Fotheringhame and R. Baddeley. Nonlinear principal components analysis of neuronal data. *Biol. Cybern.*, 77(4):283–288, 1997.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35 (2):109–135 (with comments: pp. 136–148), May 1993.

J. Frankel, K. Richmond, S. King, and P. Taylor. An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proc. of the International Conference on Spoken Language Processing (ICSLP'00)*, Beijing, China, Oct. 16–20 2000.

J. H. Friedman. Exploratory projection pursuit. *J. Amer. Stat. Assoc.*, 82(397):249–266, Mar. 1987.

J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67 (with comments, pp. 67–141), Mar. 1991.

J. H. Friedman and N. I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2): 123–143 (with discussion, pp. 143–162), Apr. 1999.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Stat. Assoc.*, 76(376):817–823, Dec. 1981.

J. H. Friedman, W. Stuetzle, and A. Schroeder. Projection pursuit density estimation. *J. Amer. Stat. Assoc.*, 79(387):599–608, Sept. 1984.

J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Computers*, C–23:881–889, 1974.

C. Fyfe and R. J. Baddeley. Finding compact and sparse distributed representations of visual images. *Network: Computation in Neural Systems*, 6(3):333–344, Aug. 1995.

J.-L. Gauvain and C.-H. Lee. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech and Audio Process.*, 2:1291–1298, 1994.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Texts in Statistical Science. Chapman & Hall, London, New York, 1995.

C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135 (with discussion, pp. 135–148), Feb. 1986.

Z. Ghahramani. Solving inverse problems using an EM approach to density estimation. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend, editors, *Proceedings of the 1993 Connectionist Models Summer School*, pages 316–323, 1994.

Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixture of factor analysers. In Solla et al. (2000), pages 449–455.

Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG–TR–96–1, University of Toronto, May 21 1996. Available online at `ftp://ftp.cs.toronto.edu/pub/zoubin/tr-96-1.ps.gz`.

Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In Cowan et al. (1994), pages 120–127.

W. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, New York, 1996.

M. Girolami, A. Cichoki, and S. Amari. A common neural network model for exploratory data analysis and independent component analysis. *IEEE Trans. Neural Networks*, 9(6):1495–1501, 1998.

M. Girolami and C. Fyfe. Stochastic ICA contrast maximization using Oja's nonlinear PCA algorithm. *Int. J. Neural Syst.*, 8(5–6):661–678, Oct./Dec. 1999.

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, Mar. 1995.

S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration: A Statistical Model-Based Approach*. Springer-Verlag, Berlin, 1998a.

S. J. Godsill and P. J. W. Rayner. Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. *IEEE Trans. Speech and Audio Process.*, 6(4):352–372, July 1998b.

B. Gold and N. Morgan. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, New York, London, Sydney, 2000.

D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins Press, Baltimore, third edition, 1996.

G. J. Goodhill and T. J. Sejnowski. A unifying objective function for topographic mappings. *Neural Computation*, 9(6):1291–1303, Aug. 1997.

R. A. Gopinath, B. Ramabhadran, and S. Dharanipragada. Factor analysis invariant to linear transformations of data. In *Proc. of the International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, Nov. 30 – Dec. 4 1998.

W. P. Gouveia and J. A. Scales. Resolution of seismic waveform inversion: Bayes versus Occam. *Inverse Problems*, 13(2):323–349, Apr. 1997.

W. P. Gouveia and J. A. Scales. Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis. *J. of Geophysical Research*, 130(B2):2759–2779, 1998.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, San Diego, fifth edition, 1994. Corrected and enlarged edition, edited by Alan Jeffrey.

R. M. Gray. Vector quantization. *IEEE ASSP Magazine*, pages 4–29, Apr. 1984.

R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inf. Theory*, 44(6):2325–2383, Oct. 1998.

M. Gyllenberg, T. Koski, E. Reilink, and M. Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. *J. Appl. Prob.*, 31:542–548, 1994.

P. Hall. On polynomial-based projection indices for exploratory projection pursuit. *Annals of Statistics*, 17 (2):589–605, June 1989.

W. J. Hardcastle, F. E. Gibbon, and W. Jones. Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders. *Brit. J. of Disorders of Communication*, 26:41–74, 1991a.

W. J. Hardcastle, F. E. Gibbon, and K. Nicolaidis. EPG data reduction methods and their implications for studies of lingual coarticulation. *J. of Phonetics*, 19:251–266, 1991b.

W. J. Hardcastle and N. Hewlett, editors. *Coarticulation: Theory, Data, and Techniques*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge, U.K., 1999.

W. J. Hardcastle, W. Jones, C. Knight, A. Trudgeon, and G. Calder. New developments in electropalatography: A state-of-the-art report. *J. Clinical Linguistics and Phonetics*, 3:1–38, 1989.

H. H. Harman. *Modern Factor Analysis*. University of Chicago Press, Chicago, second edition, 1967.

A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.

T. J. Hastie and W. Stuetzle. Principal curves. *J. Amer. Stat. Assoc.*, 84(406):502–516, June 1989.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Number 43 in Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1990.

G. T. Herman. *Image Reconstruction from Projections. The Fundamentals of Computer Tomography*. Academic Press, New York, 1980.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoustic Soc. Amer.*, 87(4):1738–1752, Apr. 1990.

H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. Speech and Audio Process.*, 2(4): 578–589, Oct. 1994.

J. A. Hertz, A. S. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Number 1 in Santa Fe Institute Studies in the Sciences of Complexity Lecture Notes. Addison-Wesley, Reading, MA, USA, 1991.

G. E. Hinton. Products of experts. In D. Wilshaw, editor, *Proc. of the Ninth Int. Conf. on Artificial Neural Networks (ICANN99)*, pages 1–6, Edinburgh, UK, Sept. 7–10 1999. The Institution of Electrical Engineers.

G. E. Hinton, P. Dayan, and M. Revow. Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural Networks*, 8(1):65–74, Jan. 1997.

T. Holst, P. Warren, and F. Nolan. Categorising [s], [ʃ] and intermediate electropalographic patterns: Neural networks and other approaches. *European Journal of Disorders of Communication*, 30(2):161–174, 1995.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. of Educational Psychology*, 24:417–441 and 498–520, 1933.

P. J. Huber. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1981.

P. J. Huber. Projection pursuit. *Annals of Statistics*, 13(2):435–475 (with comments, pp. 475–525), June 1985.

D. Husmeier. *Neural Networks for Conditional Probability Estimation*. Perspectives in Neural Computing. Springer-Verlag, Berlin, 1999.

J.-N. Hwang, S.-R. Lay, M. Maechler, R. D. Martin, and J. Schimert. Regression modeling in back-propagation and projection pursuit learning. *IEEE Trans. Neural Networks*, 5(3):342–353, May 1994.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In Jordan et al. (1998), pages 273–279.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, Oct. 1999a.

A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999b.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, New York, London, Sydney, 2001.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, June 2000.

N. Intrator and L. N. Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5(1):3–17, 1992.

E. Isaacson and H. B. Keller. *Analysis of Numerical Methods*. John Wiley & Sons, New York, London, Sydney, 1966.

M. Isard and A. Blake. CONDENSATION — conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

J. E. Jackson. *A User's Guide to Principal Components*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1991.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

M. Jamshidian and P. M. Bentler. A quasi-Newton method for minimum trace factor analysis. *J. of Statistical Computation and Simulation*, 62(1–2):73–89, 1998.

N. Japkowicz, S. J. Hanson, and M. A. Gluck. Nonlinear autoassociation is not equivalent to PCA. *Neural Computation*, 12(3):531–545, Mar. 2000.

E. T. Jaynes. Prior probabilities. *IEEE Trans. Systems, Science, and Cybernetics*, SSC-4(3):227–241, 1968.

F. V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, London, 1996.

I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, Berlin, 1986.

M. C. Jones. *The Projection Pursuit Algorithm for Exploratory Data Analysis*. PhD thesis, University of Bath, 1983.

M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society, A*, 150(1): 1–18 (with comments, pp. 19–36), 1987.

W. Jones and W. J. Hardcastle. New developments in EPG3 software. *European Journal of Disorders of Communication*, 30(2):183–192, 1995.

M. I. Jordan. Motor learning and the degrees of freedom problem. In M. Jeannerod, editor, *Attention and Performance XIII*, pages 796–836. Lawrence Erlbaum Associates, Hillsdale, New Jersey and London, 1990.

M. I. Jordan, editor. *Learning in Graphical Models*, Adaptive Computation and Machine Learning series, 1998. MIT Press. Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models, held in Erice, Italy, September 27 – October 7, 1996.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov. 1999.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, Mar. 1994.

M. I. Jordan, M. J. Kearns, and S. A. Solla, editors. *Advances in Neural Information Processing Systems*, volume 10, 1998. MIT Press, Cambridge, MA.

M. I. Jordan and D. E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16(3):307–354, July–Sept. 1992.

K. G. Jöreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482, Dec. 1967.

K. G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34 (2):183–202, June 1969.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, Sept. 1958.

N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516, Oct. 1997.

G. K. Kanji. 100 *Statistical Tests*. Sage Publications, London, 1993.

J. N. Kapur. *Maximum-Entropy Models in Science and Engineering*. John Wiley & Sons, New York, London, Sydney, 1989.

J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *J. Amer. Stat. Assoc.*, 91 (435):1343–1370, Sept. 1996.

M. S. Kearns, S. A. Solla, and D. A. Cohn, editors. *Advances in Neural Information Processing Systems*, volume 11, 1999. MIT Press, Cambridge, MA.

B. Kégl, A. Krzyzak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 22(3):281–297, Mar. 2000.

M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics Vol. 1: Distribution Theory.* Charles Griffin & Company Ltd., London, fourth edition, 1977.

W. M. Kier and K. K. Smith. Tongues, tentacles and trunks: The biomechanics of movement in muscular-hydrostats. *Zoological Journal of the Linnean Society*, 83:307–324, 1985.

S. King and A. Wrench. Dynamical system modelling of articulator movement. In Ohala et al. (1999), pages 2259–2262.

B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1–3):117–132, Aug. 1998.

T. K. Kohonen. *Self-Organizing Maps.* Number 30 in Springer Series in Information Sciences. Springer-Verlag, Berlin, 1995.

A. C. Kokaram, R. D. Morris, W. J. Fitzgerald, and P. J. W. Rayner. Interpolation of missing data in image sequences. *IEEE Trans. on Image Processing*, 4(11):1509–1519, Nov. 1995.

J. F. Kolen and J. B. Pollack. Back propagation is sensitive to initial conditions. *Complex Systems*, 4(3): 269–280, 1990.

A. C. Konstantellos. Unimodality conditions for Gaussian sums. *IEEE Trans. Automat. Contr.*, AC–25(4): 838–839, Aug. 1980.

M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *Journal of the American Institute of Chemical Engineers*, 37(2):233–243, Feb. 1991.

J. B. Kruskal and M. Wish. *Multidimensional Scaling.* Number 07–011 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Beverly Hills, 1978.

W. J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective.* Number 3 in Oxford Statistical Science Series. Oxford University Press, New York, Oxford, 1988.

S. Y. Kung, K. I. Diamantaras, and J. S. Taur. Adaptive principal component extraction (APEX) and applications. *IEEE Trans. Signal Processing*, 42(5):1202–1217, May 1994.

O. M. Kvalheim. The latent variable. *Chemometrics and Intelligent Laboratory Systems*, 14:1–3, 1992.

P. Ladefoged. Articulatory parameters. *Language and Speech*, 23(1):25–30, Jan.–Mar. 1980.

P. Ladefoged. *A Course in Phonetics.* Harcourt College Publishers, Fort Worth, fourth edition, 2000.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Stat. Assoc.*, 73 (364):805–811, Dec. 1978.

J. N. Larar, J. Schroeter, and M. M. Sondhi. Vector quantisation of the articulatory space. *IEEE Trans. Acoust., Speech, and Signal Process.*, ASSP-36(12):1812–1818, Dec. 1988.

F. Lavagetto. Time-delay neural networks for estimating lip movements from speech analysis: A useful tool in audio-video synchronization. *IEEE Trans. Circuits and Systems for video technology*, 7(5):786–800, Oct. 1997.

E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, editors. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley Series in Discrete Mathematics and Optimization. John Wiley & Sons, Chichester, England, 1985.

D. N. Lawley. A modified method of estimation in factor analysis and some large sample results. *Nord. Psykol. Monogr. Ser.*, 3:35–42, 1953.

P. F. Lazarsfeld and N. W. Henry. *Latent Structure Analysis*. Houghton-Mifflin, Boston, 1968.

M. LeBlanc and R. Tibshirani. Adaptive principal surfaces. *J. Amer. Stat. Assoc.*, 89(425):53–64, Mar. 1994.

D. D. Lee and H. Sompolinsky. Learning a continuous hidden variable model for binary data. In Kearns et al. (1999), pages 515–521.

T.-W. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, Feb. 1999.

C. J. Leggeter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, Apr. 1995.

S. E. Levinson and C. E. Schmidt. Adaptive computation of articulatory parameters from the speech signal. *J. Acoustic Soc. Amer.*, 74(4):1145–1154, Oct. 1983.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, Feb. 2000.

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological Review*, 74(6):431–461, 1967.

A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21:1–36, 1985.

B. G. Lindsay. The geometry of mixture likelihoods: A general theory. *Annals of Statistics*, 11(1):86–94, Mar. 1983.

R. Linsker. An application of the principle of maximum information preservation to linear systems. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1, pages 186–194. Morgan Kaufmann, San Mateo, 1989.

R. J. A. Little. Regression with missing X's: A review. *J. Amer. Stat. Assoc.*, 87(420):1227–1237, Dec. 1992.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1987.

S. P. Luttrell. A Bayesian analysis of self-organizing maps. *Neural Computation*, 6(5):767–794, Sept. 1994.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, May 1992a.

D. J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3): 448–472, May 1992b.

D. J. C. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research A*, 354(1):73–80, Jan. 1995a.

D. J. C. MacKay. Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995b.

D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Draft 3.7, Cavendish Laboratory, University of Cambridge, Dec. 19 1996. Available online at `http://wol.ra.phy.cam.ac.uk/mackay/abstracts/ica.html`.

D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, July 1999.

S. Maeda. A digital simulation method of the vocal tract system. *Speech Communication*, 1(3–4):199–229, 1982.

S. Makeig, T.-P. Jung, A. J. Bell, D. Ghahremani, and T. J. Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proc. Natl. Acad. Sci. USA*, 94:10979–10984, Sept. 1997.

E. C. Malthouse. Limitations of nonlinear PCA as performed with generic neural networks. *IEEE Trans. Neural Networks*, 9(1):165–173, Jan. 1998.

J. Mao and A. K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Networks*, 6(2):296–317, Mar. 1995.

A. Marchal and W. J. Hardcastle. ACCOR: Instrumentation and database for the cross-language study of coarticulation. *Language and Speech*, 36(2, 3):137–153, 1993.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Probability and Mathematical Statistics Series. Academic Press, New York, 1979.

A. D. Marrs and A. R. Webb. Exploratory data analysis using radial basis function latent variable models. In Kearns et al. (1999), pages 529–535.

T. M. Martinetz and K. J. Schulten. Topology representing networks. *Neural Networks*, 7(3):507–522, 1994.

G. P. McCabe. Principal variables. *Technometrics*, 26(2):137–144, 1984.

R. P. McDonald. *Factor Analysis and Related Methods.* Lawrence Erlbaum Associates, Hillsdale, New Jersey and London, 1985.

R. S. McGowan. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 14(1):19–48, Feb. 1994.

R. S. McGowan and A. Faber. Introduction to papers on speech recognition and perception from an articulatory point of view. *J. Acoustic Soc. Amer.*, 99(3):1680–1682, Mar. 1996.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1997.

G. J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 2000.

X.-L. Meng and D. van Dyk. The EM algorithm — an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, B*, 59(3):511–540 (with discussion, pp. 541–567), 1997.

P. Mermelstein. Determination of vocal-tract shape from measured formant frequencies. *J. Acoustic Soc. Amer.*, 41(5):1283–1294, 1967.

P. Mermelstein. Articulatory model for the study of speech production. *J. Acoustic Soc. Amer.*, 53(4):1070–1082, 1973.

L. Mirsky. *An Introduction to Linear Algebra.* Clarendon Press, Oxford, 1955. Reprinted in 1982 by Dover Publications.

B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 19(7):696–710, July 1997.

J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, Summer 1989.

D. F. Morrison. *Multivariate Statistical Methods.* McGraw-Hill, New York, third edition, 1990.

K. Mosegaard and A. Tarantola. Monte-Carlo sampling of solutions to inverse problems. *J. of Geophysical Research—Solid Earth*, 100(B7):12431–12447, 1995.

É. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP'97)*, volume 5, pages 3617–3620, Munich, Germany, Apr. 21–24 1997.

J. R. Movellan, P. Mineiro, and R. J. Williams. Modeling path distributions using partially observable diffusion networks: A Monte-Carlo approach. Technical Report 99.01, Department of Cognitive Science, University of California, San Diego, June 1999. Available online at `http://hci.ucsd.edu/cogsci/tech_reports/faculty_pubs/99_01.ps`.

F. Mulier and V. Cherkassky. Self-organization as an iterative kernel smoothing process. *Neural Computation*, 7(6):1165–1177, Nov. 1995.

I. T. Nabney, D. Cornford, and C. K. I. Williams. Bayesian inference for wind field retrieval. *Neurocomputing*, 30(1–4):3–11, Jan. 2000.

J.-P. Nadal and N. Parga. Non-linear neurons in the low-noise limit: A factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5(4):565–581, Nov. 1994.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG–TR–93–1, Dept. of Computer Science, University of Toronto, Sept. 1993. Available online at `ftp://ftp.cs.toronto.edu/pub/radford/review.ps.Z`.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer Series in Statistics. Springer-Verlag, Berlin, 1996.

R. M. Neal and P. Dayan. Factor analysis using delta-rule wake-sleep learning. *Neural Computation*, 9(8): 1781–1803, Nov. 1997.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan (1998), pages 355–368. Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models, held in Erice, Italy, September 27 – October 7, 1996.

W. L. Nelson. Physical principles for economies of skilled movements. *Biol. Cybern.*, 46(2):135–147, 1983.

N. Nguyen. EPG bidimensional data reduction. *European Journal of Disorders of Communication*, 30:175–182, 1995.

N. Nguyen, P. Hoole, and A. Marchal. Regenerating the spectral shape of [s] and [ʃ] from a limited set of articulatory parameters. *J. Acoustic Soc. Amer.*, 96(1):33–39, July 1994.

N. Nguyen, A. Marchal, and A. Content. Modeling tongue-palate contact patterns in the production of speech. *J. of Phonetics*, 24(1):77–97, Jan. 1996.

K. Nicolaidis and W. J. Hardcastle. Articulatory-acoustic analysis of selected English sentences from the EUR-ACCOR corpus. Technical report, SPHERE (Human capital and mobility program), 1994.

K. Nicolaidis, W. J. Hardcastle, A. Marchal, and N. Nguyen-Trong. Comparing phonetic, articulatory, acoustic and aerodynamic signal representations. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual Representations of Speech Signals*, pages 55–82. John Wiley & Sons, 1993.

M. A. L. Nicolelis. Actions from thoughts. *Nature*, 409(6818):403–407, Jan. 18 2001.

M. Niranjan, editor. *Proc. of the 1998 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing (NNSP98)*, Cambridge, UK, Aug. 31 – Sept. 2 1998.

D. A. Nix and J. E. Hogden. Maximum-likelihood continuity mapping (MALCOM): An alternative to HMMs. In Kearns et al. (1999), pages 744–750.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, editors. *Proc. of the 14th International Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, USA, Aug. 1–7 1999.

E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, Nov.–Dec. 1992.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 13 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.*, 37(23):3311–3325, Dec. 1997.

M. W. Oram, P. Földiák, D. I. Perret, and F. Sengpiel. The 'ideal homunculus': Decoding neural population signals. *Trends Neurosci.*, 21(6):259–265, June 1998.

D. Ormoneit and V. Tresp. Penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Trans. Neural Networks*, 9(4):639–650, July 1998.

M. Ostendorf, V. V. Digalakis, and O. A. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Process.*, 4(5):360–378, Sept. 1996.

G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoustic Soc. Amer.*, 92(2):688–700, Aug. 1992.

J. Park and I. W. Sandberg. Approximation and radial-basis-function networks. *Neural Computation*, 5(2): 305–316, Mar. 1993.

R. L. Parker. *Geophysical Inverse Theory.* Princeton University Press, Princeton, 1994.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, 1988.

B. A. Pearlmutter. Gradient calculation for dynamic recurrent neural networks: A survey. *IEEE Trans. Neural Networks*, 6(5):1212–1228, 1995.

B. A. Pearlmutter and L. C. Parra. A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing (ICONIP–96), Hong Kong*, pages 151–157, Sept. 1996.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6): 559–572, 1901.

D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, Oct. 2000.

H.-O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and Fractals: New Frontiers of Science.* Springer-Verlag, New York, 1992.

J. Picone, S. Pike, R. Reagan, T. Kamm, J. Bridle, L. Deng, J. Ma, H. Richards, and M. Schuster. Initial evaluation of hidden dynamic models on conversational speech. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP'99)*, volume 1, pages 109–112, Phoenix, Arizona, USA, May 15–19 1999.

A. Pisani. A nonparametric and scale-independent method for cluster-analysis. 1. The univariate case. *Monthly Notices of the Royal Astronomical Society*, 265(3):706–726, Dec. 1993.

C. Posse. An effective two-dimensional projection pursuit algorithm. *Communications in Statistics — Simulation and Computation*, 19(4):1143–1164, 1990.

C. Posse. Tools for two-dimensional exploratory projection pursuit. *Journal of Computational and Graphical Statistics*, 4:83–100, 1995.

S. Pratt, A. T. Heintzelman, and D. S. Ensrud. The efficacy of using the IBM Speech Viewer vowel accuracy module to treat young children with hearing impairment. *Journal of Speech and Hearing Research*, 29: 99–105, 1993.

F. P. Preparata and M. I. Shamos. *Computational Geometry: An Introduction.* Monographs in Computer Science. Springer-Verlag, New York, 1985.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, U.K., second edition, 1992.

L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Signal Processing Series. Prentice-Hall, Englewood Cliffs, N.J., 1993.

M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi. On the use of neural networks in articulatory speech synthesis. *J. Acoustic Soc. Amer.*, 93(2):1109–1121, Feb. 1993.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, Apr. 1984.

K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Speech Communication*, 27(1):19–42, Feb. 1999.

M. Revow, C. K. I. Williams, and G. Hinton. Using generative models for handwritten digit recognition. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 18(6):592–606, June 1996.

H. B. Richards and J. S. Bridle. The HDM: a segmental hidden dynamic model of coarticulation. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP'99)*, volume I, pages 357–360, Phoenix, Arizona, USA, May 15–19 1999.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, B*, 59(4):731–758, 1997.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, U.K., 1996.

S. J. Roberts. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2): 261–272, Feb. 1997.

S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 20(11):1133–1142, 1998.

W. J. J. Roberts and Y. Ephraim. Hidden Markov modeling of speech using Toeplitz covariance matrices. *Speech Communication*, 31(1):1–14, May 2000.

A. J. Robinson. An application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, 5(2):298–305, Mar. 1994.

T. Rögnvaldsson. On Langevin updating in multilayer perceptrons. *Neural Computation*, 6(5):916–926, Sept. 1994.

R. Rohwer and J. C. van der Rest. Minimum description length, regularization, and multimodal data. *Neural Computation*, 8(3):595–609, Apr. 1996.

E. T. Rolls and A. Treves. *Neural Networks and Brain Function*. Oxford University Press, 1998.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, Nov. 1998.

R. C. Rose, J. Schroeter, and M. M. Sondhi. The potential role of speech production models in automatic speech recognition. *J. Acoustic Soc. Amer.*, 99(3):1699–1709 (with comments, pp. 1710–1717), Mar. 1996.

E. Z. Rothkopf. A measure of stimulus similarity and errors in some paired-associate learning tasks. *J. of Experimental Psychology*, 53(2):94–101, 1957.

B. Rotman and G. T. Kneebone. *The Theory of Sets & Transfinite Numbers*. Oldbourne, London, 1966.

S. Roweis. EM algorithms for PCA and SPCA. In Jordan et al. (1998), pages 626–632.

S. Roweis. Constrained hidden Markov models. In Solla et al. (2000), pages 782–788.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, Dec. 22 2000.

A. E. Roy. *Orbital Motion*. Adam Hilger Ltd., Bristol, 1978.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1987.

D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, Mar. 1982.

D. B. Rubin and D. T. Thayer. More on EM for ML factor analysis. *Psychometrika*, 48(2):253–257, June 1983.

P. Rubin, T. Baer, and P. Mermelstein. An articulatory synthesizer for perceptual research. *J. Acoustic Soc. Amer.*, 70(2):321–328, Aug. 1981.

E. Saltzman and J. A. Kelso. Skilled actions: a task-dynamic approach. *Psychological Review*, 94(1):84–106, Jan. 1987.

J. W. Sammon, Jr. A nonlinear mapping for data structure analysis. *IEEE Trans. Computers*, C–18(5): 401–409, May 1969.

T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2:459–473, 1989.

T. D. Sanger. Probability density estimation for the interpretation of neural population codes. *J. Neurophysiol.*, 76(4):2790–2793, Oct. 1996.

L. K. Saul and M. G. Rahim. Markov processes on curves. *Machine Learning*, 41(3):345–363, Dec. 2000a.

L. K. Saul and M. G. Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Trans. Speech and Audio Process.*, 8(2):115–125, Mar. 2000b.

E. Saund. Dimensionality-reduction using connectionist networks. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 11(3):304–314, Mar. 1989.

J. A. Scales and M. L. Smith. *Introductory Geophysical Inverse Theory*. Samizdat Press, 1998. Freely available in draft form from `http://samizdat.mines.edu/inverse_theory/`.

F. Scarselli and A. C. Tsoi. Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, 11(1):15–37, Jan. 1998.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Number 72 in Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1997.

B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods. Support Vector Learning*. MIT Press, 1999a.

B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola. Input space vs. feature space in kernel-based methods. *IEEE Trans. Neural Networks*, 10(5):1000–1017, Sept. 1999b.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

M. R. Schroeder. Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoustic Soc. Amer.*, 41(4):1002–1010, 1967.

J. Schroeter and M. M. Sondhi. Dynamic programming search of articulatory codebooks. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP'89)*, volume 1, pages 588–591, Glasgow, UK, May 23–26 1989.

J. Schroeter and M. M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech and Audio Process.*, 2(1):133–150, Jan. 1994.

M. Schuster. *On Supervised Learning from Sequential Data with Applications for Speech Recognition*. PhD thesis, Graduate School of Information Science, Nara Institute of Science and Technology, 1999.

D. W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1992.

D. W. Scott and J. R. Thompson. Probability density estimation in higher dimensions. In J. E. Gentle, editor, *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, pages 173–179, Amsterdam, New York, Oxford, 1983. North Holland-Elsevier Science Publishers.

R. N. Shepard. Analysis of proximities as a technique for the study of information processing in man. *Human Factors*, 5:33–48, 1963.

K. Shirai and T. Kobayashi. Estimating articulatory motion from speech wave. *Speech Communication*, 5(2): 159–170, June 1986.

M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, editors. *Lévy Flights and Related Topics in Physics*. Number 450 in Lecture Notes in Physics. Springer-Verlag, Berlin, 1995. Proceedings of the International Workshop held at Nice, France, 27–30 June 1994.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Number 26 in Monographs on Statistics and Applied Probability. Chapman & Hall, London, New York, 1986.

L. Sirovich and M. Kirby. Low-dimensional procedure for the identification of human faces. *J. Opt. Soc. Amer. A*, 4(3):519–524, Mar. 1987.

D. S. Sivia. *Data Analysis. A Bayesian Tutorial*. Oxford University Press, New York, Oxford, 1996.

R. Snieder and J. Trampert. *Inverse Problems in Geophysics*. Samizdat Press, 1999. Freely available from `http://samizdat.mines.edu/snieder_trampert/`.

S. A. Solla, T. K. Leen, and K.-R. Müller, editors. *Advances in Neural Information Processing Systems*, volume 12, 2000. MIT Press, Cambridge, MA.

V. N. Sorokin. Determination of vocal-tract shape for vowels. *Speech Communication*, 11(1):71–85, Mar. 1992.

V. N. Sorokin, A. S. Leonov, and A. V. Trushkin. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30(1):55–74, Jan. 2000.

C. Spearman. General intelligence, objectively determined and measured. *Am. J. Psychol.*, 15:201–293, 1904.

D. F. Specht. A general regression neural network. *IEEE Trans. Neural Networks*, 2(6):568–576, Nov. 1991.

M. Spivak. *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Addison-Wesley, Reading, MA, USA, 1965.

M. Spivak. *Calculus*. Addison-Wesley, Reading, MA, USA, 1967.

M. Stone. Toward a model of three-dimensional tongue movement. *J. of Phonetics*, 19:309–320, 1991.

N. V. Swindale. The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems*, 7(2):161–247, May 1996.

A. Tarantola. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 1987.

J. B. Tenenbaum. Mapping a manifold of perceptual observations. In Jordan et al. (1998), pages 682–688.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, Dec. 22 2000.

G. Tesauro, D. S. Touretzky, and T. K. Leen, editors. *Advances in Neural Information Processing Systems*, volume 7, 1995. MIT Press, Cambridge, MA.

R. J. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Scripta Series in Mathematics. John Wiley & Sons, New York, London, Sydney, 1977. Translation editor: Fritz John.

M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, Feb. 1999a.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 61(3):611–622, 1999b.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1985.

L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang. The indeterminacy and identifiability of blind identification. *IEEE Trans. Circuits and Systems*, 38(5):499–509, May 1991.

V. Tresp, R. Neuneier, and S. Ahmad. Efficient methods for dealing with missing data in supervised learning. In Tesauro et al. (1995), pages 689–696.

A. Treves, S. Panzeri, E. T. Rolls, M. Booth, and E. A. Wakeman. Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Computation*, 11(3):601–632, Mar. 1999.

A. Treves and E. T. Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, Nov. 1991.

A. C. Tsoi. Recurrent neural network architectures — an overview. In C. L. Giles and M. Gori, editors, *Adaptive Processing of Temporal Information*, volume 1387 of *Lecture Notes in Artificial Intelligence*, pages 1–26. Springer-Verlag, New York, 1998.

UCLA. Artificial EPG palate image. The UCLA Phonetics Lab. Available online at `http://www.humnet.ucla.edu/humnet/linguistics/faciliti/facilities/physiology/EGP_picture.JPG`, Feb. 1, 2000.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128, Sept. 2000.

A. Utsugi. Hyperparameter selection for self-organizing maps. *Neural Computation*, 9(3):623–635, Apr. 1997a.

A. Utsugi. Topology selection for self-organizing maps. *Network: Computation in Neural Systems*, 7(4): 727–740, 1997b.

A. Utsugi. Bayesian sampling and ensemble learning in generative topographic mapping. *Neural Processing Letters*, 12(3):277–290, Dec. 2000.

A. Utsugi and T. Kumagai. Bayesian analysis of mixtures of factor analyzers. *Neural Computation*, 13(5): 993–1002, May 2001.

V. N. Vapnik and S. Mukherjee. Support vector method for multivariate density estimation. In Solla et al. (2000), pages 659–665.

S. V. Vaseghi. *Advanced Signal Processing and Digital Noise Reduction*. John Wiley & Sons, New York, London, Sydney, second edition, 2000.

S. V. Vaseghi and P. J. W. Rayner. Detection and suppression of impulsive noise in speech-communication systems. *IEE Proc. I (Communications, Speech and Vision)*, 137(1):38–46, Feb. 1990.

T. Villmann, R. Der, M. Hermann, and T. M. Martinetz. Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Trans. Neural Networks*, 8(2):256–266, Mar. 1997.

W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, Feb. 18 2000.

H. M. Wagner. *Principles of Operations Research with Applications to Managerial Decisions*. Prentice-Hall, Englewood Cliffs, N.J., second edition, 1975.

A. Webb. *Statistical Pattern Recognition*. Edward Arnold, 1999.

A. R. Webb. Multidimensional scaling by iterative majorization using radial basis functions. *Pattern Recognition*, 28(5):753–759, May 1995.

E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *J. Amer. Stat. Assoc.*, 85(411): 664–675, Sept. 1990.

J. R. Westbury. *X-Ray Microbeam Speech Production Database User's Handbook Version 1.0*. Waisman Center on Mental Retardation & Human Development, University of Wisconsin, Madison, WI, June 1994. With the assistance of Greg Turner & Jim Dembowski.

J. R. Westbury, M. Hashi, and M. J. Lindstrom. Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, 26(3):203–226, Nov. 1998.

J. Weston, A. Gammerman, M. O. Stitson, V. Vapnik, V. Vovk, and C. Watkins. Support vector density estimation. In Schölkopf et al. (1999a), chapter 18, pages 293–306.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, 1990.

P. Whittle. On principal components and least square methods of factor analysis. *Skand. Aktur. Tidskr.*, 36: 223–239, 1952.

J. Wiles, P. Bakker, A. Lynton, M. Norris, S. Parkinson, M. Staples, and A. Whiteside. Using bottlenecks in feedforward networks as a dimension reduction technique: An application to optimization tasks. *Neural Computation*, 8(6):1179–1183, Aug. 1996.

J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, New York, Oxford, 1965.

P. M. Williams. Using neural networks to model conditional multivariate densities. *Neural Computation*, 8 (4):843–854, May 1996.

B. Willmore, P. A. Watters, and D. J. Tolhurst. A comparison of natural-image-based models of simple-cell coding. *Perception*, 29(9):1017–1040, Sept. 2000.

R. Wilson and M. Spann. A new approach to clustering. *Pattern Recognition*, 23(12):1413–1425, 1990.

J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350, July 1970.

D. M. Wolpert and Z. Ghahramani. Computational principles of movement neuroscience. *Nat. Neurosci.*, 3 (Supp.):1212–1217, Nov. 2000.

D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7–8):1317–1329, Oct. 1998.

A. A. Wrench. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. In *Phonus*, volume 5, Saarbrücken, 2000. Institute of Phonetics, University of Saarland.

F. Xie and D. van Compernolle. Speech enhancement by spectral magnitude estimation —a unifying approach. *Speech Communication*, 19(2):89–104, Aug. 1996.

L. Xu, C. C. Cheung, and S. Amari. Learned parametric mixture based ICA algorithm. *Neurocomputing*, 22 (1–3):69–80, Nov. 1998.

E. Yamamoto, S. Nakamura, and K. Shikano. Lip movement synthesis from speech based on hidden Markov models. *Speech Communication*, 26(1–2):105–115, 1998.

H. H. Yang and S. Amari. Adaptive on-line learning algorithms for blind separation — maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, Oct. 1997.

H. Yehia and F. Itakura. A method to combine acoustic and morphological constraints in the speech production inverse problem. *Speech Communication*, 18(2):151–174, Apr. 1996.

H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. Using speech acoustics to drive facial motion. In Ohala et al. (1999), pages 631–634.

H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1–2):23–43, Oct. 1998.

G. Young. Maximum likelihood estimation and factor analysis. *Psychometrika*, 6:49–53, 1940.

S. J. Young. A review of large vocabulary continuous speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–57, Sept. 1996.

K. Zhang, I. Ginzburg, B. L. McNaughton, and T. J. Sejnowski. Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *J. Neurophysiol.*, 79(2): 1017–1044, Feb. 1998.

R. D. Zhang and J.-G. Postaire. Convexity dependent morphological transformations for mode detection in cluster-analysis. *Pattern Recognition*, 27(1):135–148, 1994.

Y. Zhao and C. G. Atkeson. Implementing projection pursuit learning. *IEEE Trans. Neural Networks*, 7(2): 362–373, Mar. 1996.

I. Zlokarnik. Adding articulatory features to acoustic features for automatic speech recognition. *J. Acoustic Soc. Amer.*, 97(5):3246, May 1995a.

I. Zlokarnik. Articulatory kinematics from the standpoint of automatic speech recognition. *J. Acoustic Soc. Amer.*, 98(5):2930–2931, Nov. 1995b.