Chapter 2

The continuous latent variable modelling formalism

This chapter gives the theoretical basis for continuous latent variable models. Section 2.1 defines intuitively the concept of latent variable models and gives a brief historical introduction to them. Section 2.2 uses a simple example, inspired by the mechanics of a mobile point, to justify and explain latent variables. Section 2.3 gives a more rigorous definition, which we will use throughout this thesis. Section 2.6 describes the most important specific continuous latent variable models and section 2.7 defines mixtures of continuous latent variable models. The chapter discusses other important topics, including parameter estimation, identifiability, interpretability and marginalisation in high dimensions. Section 2.9 on dimensionality reduction will be the basis for part II of the thesis. Section 2.10 very briefly mentions some applications of continuous latent variable models for dimensionality reduction. Section 2.11 shows a worked example of a simple continuous latent variable model. Section 2.12 give some complementary mathematical results, in particular the derivation of a diagonal noise GTM model and of its EM algorithm.

2.1 Introduction and historical overview of latent variable models

Latent variable models are probabilistic models that try to explain a (relatively) high-dimensional process in terms of a few degrees of freedom. Historically, the idea of latent variables arose primarily from psychometrics, beginning with the g factor of Spearman (1904) and continuing with other psychologists such as Thomson, Thurstone and Burt, who were investigating the mental ability of children as suggested by the correlation and covariance matrices from cognitive tests variables. This eventually led to the development of factor analysis. Principal component analysis, traditionally not considered a latent variable model, was first thought of by Pearson (1901) and developed as a multivariate technique by Hotelling (1933). Latent structure analysis, corresponding to models where the latent variables are categorical, originated with Lazarsfeld and Henry (1968) as a tool for sociological analysis and had a separate development from factor analysis until recently. Bartholomew (1987) gives more historical details and references.

The statistics literature classifies latent variable models according to the metric (continuous) or categorical (discrete) character of the variables involved, as table 2.1 shows. Also, in a broad sense many probabilistic models commonly used in machine learning can be considered as latent variable models inasmuch as they include probability distributions for variables which are not observed: mixture models (where the variable which indexes the components is a latent variable), hidden Markov models (the state sequence is unobserved),

This chapter is an extended version of references Carreira-Perpiñán (1996, 1997).

		Observed variables	
		Metrical	*Categorical*
Latent variables	*Metrical*	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis
		Analysis of mixtures	

Table 2.1: Classification of latent variable models (adapted from Bartholomew, 1987).

Helmholtz machines (Dayan et al., 1995) (the activations of each unit), elastic nets (Durbin et al., 1989) (the tour knots), etc. In this chapter we will concentrate exclusively on latent variable models where both the latent and the observed variables are continuous (although much of the general treatment applies to discrete variables too), going much further than the linear-normal model of factor analysis. In some points discrete variables will appear as a result of a Monte Carlo discretisation. The latent class model where the observed variables are binary and there is a single discrete latent variable corresponds to the mixture of multivariate Bernoulli distributions, which is discussed in chapter 3 and used in chapter 5.

In this work, we follow (with minor variations) the theoretical framework currently accepted in statistics for latent variable models, for which two good references are Everitt (1984) and Bartholomew (1987). However, these works do not include any of the more recent latent variable models such as GTM, ICA or IFA, and they do not discuss issues of importance in machine learning, such as the continuity of the dimensionality reduction mapping or the mixtures of continuous latent variable models. The work of MacKay (1995a) on density networks, which is the name he uses for nonlinear latent variable models, was pioneering in the introduction of the latent variable model framework to the machine learning community (in spite of an odd choice of journal).

The traditional treatment of latent variable models in the statistics literature is (Krzanowski, 1988):

1. formulate model (independence and distributional assumptions)
2. estimate parameters (by maximum likelihood)
3. test hypotheses about the parameters

which perhaps explains why most of the statistical research on latent variables has remained in the linear-normal model of factor analysis (with a few exceptions, e.g. the use of polynomial rather than linear functions; Etezadi-Amoli and McDonald, 1983; McDonald, 1985): rigorous analysis of nonlinear models is very difficult. Interesting developments concerning new models and learning algorithms have appeared in the literature of statistical pattern recognition in the 1990s, as the rest of this chapter will show.

Traditionally, most applications of latent variables have been in psychology and the social sciences, where hypothesised latent variables such as intelligence, verbal ability or social class are believed to exist; they have been less used in the natural sciences, where often most variables are measurable (Krzanowski, 1988, p. 502). This latter statement is not admissible under the generative point of view exposed in sections 2.2 and 2.3, whereby the processes of measurement and stochastic variation (noise) can make a physical system appear more high-dimensional than it really is. In fact, factor analysis has been applied to a number of "natural science" problems (e.g. in botany, biology, geology or engineering) as well as other kinds of latent variable models recently developed, such as the application of GTM to dimensionality reduction of electropalatographic data (Carreira-Perpiñán and Renals, 1998a) or of ICA to the removal of artifacts in electroencephalographic recordings (Makeig et al., 1997). Besides that, the most popular technique for dimensionality reduction, principal component analysis, has recently been recast in the form of a particular kind of factor analysis—thus as a latent variable model.

Whether the latent variables revealed can or cannot be interpreted (a delicate issue discussed in section 2.8.1), latent variable models are a powerful tool for data analysis when the intrinsic dimensionality of the problem is smaller than the apparent one and they are specially suited for dimensionality reduction and missing data reconstruction—as this thesis will try to demonstrate.

2.2 Physical interpretation of the latent variable model formalism

In this section we give a physical flavour to the generative view of latent variables. This generative view will be properly defined in section 2.3 and should be compared to the physical interpretation given below.

2.2.1 An example: eclipsing spectroscopic binary star

We will introduce the idea of latent variables through an example from astronomy, that of *eclipsing spectroscopic binary stars* (Roy, 1978). A binary system is a pair of stars that describe orbits about their common centre of mass, the two components being gravitationally bound together. An eclipsing spectroscopic binary is viewed as a single star because the components are so close that they cannot be resolved in a telescope, but its double nature can be revealed in several ways:

- If the orbit plane contains or is close to the line of sight, the components will totally or partially eclipse each other, which results in regular diminutions in the binary star's brightness.

- The Doppler effect[1] of the orbital velocities of the components produces shifts to red and blue on the spectral lines of the binary star.

A well-known example of this kind of star system is Algol (β Per), which is actually a ternary system.

We will consider a simplified model of eclipsing spectroscopic binary star, in which one star (the primary star) is much more massive than the other, thus remaining practically stationary in the centre-of-mass system. The secondary star follows an elliptical orbit around the primary one, which is located at one of the foci. Figures 2.1 and 2.2 show the trajectory of the secondary star and the basic form of the variation with the polar angle θ (true anomaly) of:

- The radial velocity v_r of the secondary star (i.e., the projection of the velocity vector along the line of sight from the Earth), to which the spectrum shift is proportional. We are assuming a stationary observer in Earth, so that the measured radial velocity must be corrected for the Earth's orbital motion about the Sun.

- The combined brightness B.

We would like to provide a scenario which, though idealised, is very close to an experimental situation that an astronomer could find. Thus, we are ignoring complicated effects, such as the facts that: the shape of both stars can be far from spherical, as they can fill their lobes (bounded by the Roche limit); the stars do not have uniform brightness across their discs, but decreasing towards the limb (limb darkening); the relativistic advance of the periastron with time; the perturbation due to third bodies; etc. Thus, the true theoretical curve for the brightness will not be piecewise linear and the actual curve will vary from one binary system to another, but it will always conserve the form indicated in fig. 2.2: a periodic curve with two falls corresponding to the eclipses. Similarly, the spectral shift will be a periodic oscillating function.

Now, an astronomer could collect a number of paired measurements (B, v_r) and construct with them brightness and radial velocity (or spectrum) curves such as those of fig. 2.2 (although perhaps replacing the true anomaly θ by a more natural variable in this case, such as the time t at which the measurement was collected). Detailed analysis of the brightness and spectrum curves under the laws of Kepler can provide knowledge of the eccentricity and inclination of the orbit, the major semiaxis, the radii, masses and luminosities of the stars, etc.

The knowledge of the astronomer that all measured variables (B and v_r) depend on a single variable, the true anomaly θ, is implicit. That is, given θ, the configuration of the binary system is completely defined and so are the values of all the variables we could measure: B, v_r, even r, etc.

Now let us suppose that we have such a collection of measurements but that we do not know anything about the underlying model that generated them, i.e., we do not know that the observed brightness and spectral shifts are (indirectly) governed by Kepler's law. Let us consider the following question: could it be possible that, although we are measuring two variables each time we observe the system, just one variable (not necessarily B or v_r) would be enough to determine the configuration of the system completely?[2] More concisely: could the number of degrees of freedom of the binary system be smaller than the number of variables that we measure from it?

As we discuss in section 4.4, this is a very difficult question to answer if the measurements are corrupted by noise or if each measurement consists of more than about 3 variables. But if we plotted the collection of pairs (B, v_r) in a plane coordinate system, with each variable being associated with one axis, we would observe the following fact (fig. 2.3 left): the points do not fall all over the plane, spanning an extensional (two-dimensional) area, but fall on a curve (dotted in the figure). If we accept that our instruments are imperfect, so that each point is slightly off the position where it should be, we would observe a situation like that of fig. 2.3 right: the points now occupy an extensional area (several oval patches), but it is apparent that the region over which the measurements fall is still a curve. This gives away the fact that the system has only one degree of freedom, however many different, apparently unrelated variables we want to measure from it. In this example, the intrinsic dimensionality of the binary system is one but we measure two variables.

Accepting that the intrinsic dimensionality of a system is smaller than the number of variables that we measure from it, the latent variable framework allows the construction of a generative model for the system.

[1] If a source emitting radiation has a velocity v relative to the observer, the received radiation that normally has a wavelength λ when the velocity relative to the observer is 0 will have a measured wavelength λ' with $\frac{\lambda'-\lambda}{\lambda} = \frac{v}{c}$, where c is the speed of light and the source is approaching for $v < 0$ and receding for $v > 0$. Thus, the Doppler shift is defined as $\frac{\Delta\lambda}{\lambda} = \frac{v}{c}$, being positive for red shift and negative for blue shift.

[2] Of course we could argue that two variables could not be enough to determine the system configuration completely, but we will suppose that this is not the case.

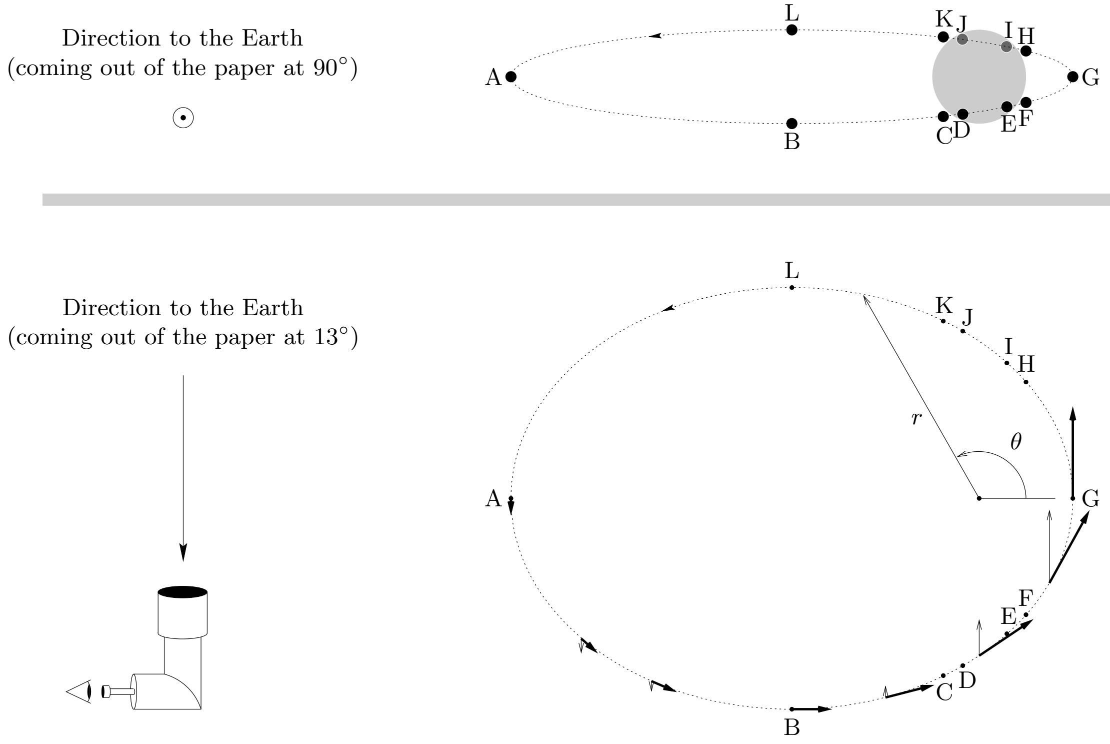


Figure 2.1: Schematic of the orbit for a binary system. The primary star is assumed stationary with the secondary following a Keplerian closed orbit around it: an ellipse of eccentricity $\epsilon = \frac{2}{3}$ with the primary star in a focus. The figure on the top shows the view of the binary system as seen from the Earth with a direction which has an inclination of 13° over the orbit plane and is orthogonal to the major axis of the orbit. The bottom figure shows the true orbit; the vectors of the velocity (thick arrows) and the radial velocity (in the direction of the Earth; solid arrows) are given for some points in the orbit.

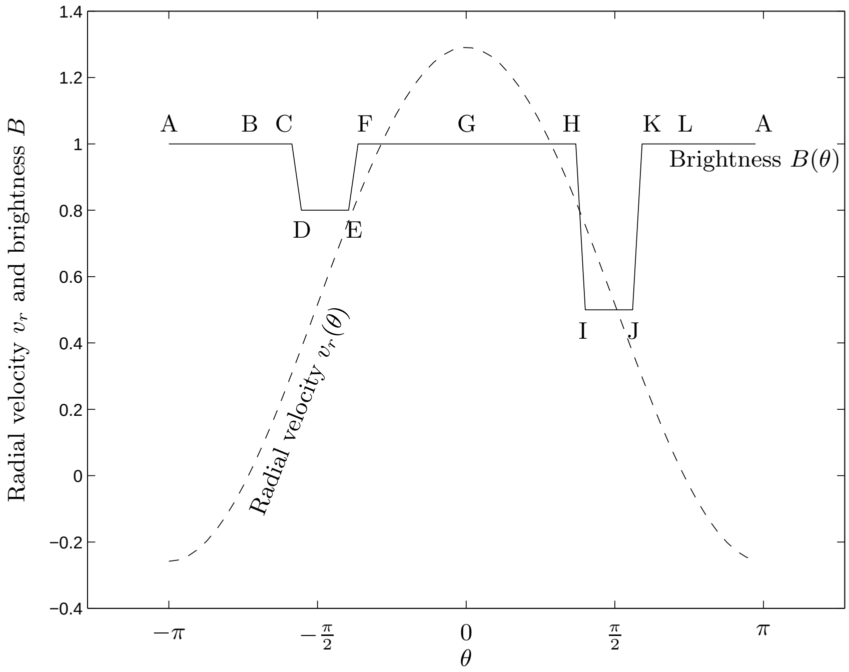


Figure 2.2: Variation of the radial velocity v_r and of the combined brightness B as a function of the true anomaly θ for the binary system of fig. 2.1. The units of the vertical axis of the graph have been normalised by a multiplying factor, so that only the shape of the curves is important. The brightness curve assumes that the secondary star has a very small radius compared to that of the primary one.

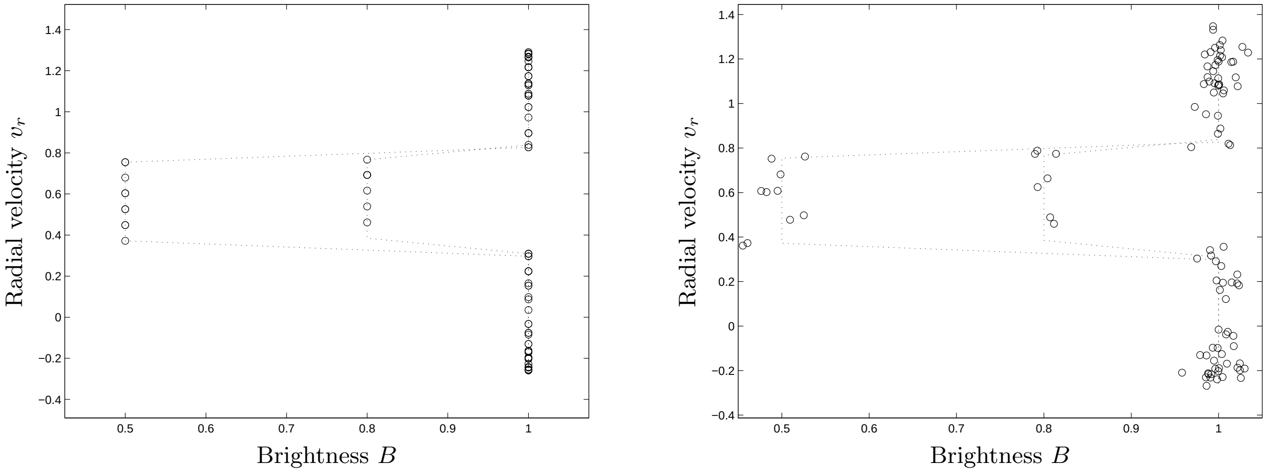


Figure 2.3: Plots of a number of measurements (B, v_r) in a coordinate system where the brightness goes in abscissas and the radial velocity in ordinates. The left graph corresponds to an ideal, noiseless measuring device: the measurements fall exactly on a curve. Due to the discontinuity of the brightness B, the curve in (B, v_r) space consists of several disconnected segments. The right graph corresponds to a real, noisy measuring device: the measurements still fall approximately on a curve, but now span a truly two-dimensional area. Again, due to the discontinuity of the brightness B, we observe several patches.

This generative model can, under certain circumstances, mimic the behaviour of the system inasmuch as it is a black box that generates data.

2.2.2 Physical interpretation of latent variables: measurement and noise

We can give a physical interpretation to the latent variable model formalism, to be defined more rigorously in section 2.3. Consider a (physical) system governed by L independent variables (or degrees of freedom) $x_1, \dots, x_L$, which we can represent as an L-dimensional vector $\mathbf{x} = (x_1, \dots, x_L)^T$ taking values in a certain subset $\mathcal{S}$ of $\mathbb{R}^L$, called *latent* or *state space* (we will consider only continuous variables). That is, every state of the system is given by a particular vector $\mathbf{x}$ in state space $\mathcal{S}$. The variables $x_1, \dots, x_L$ can be called *generalised coordinates* in classical mechanics or *quantum numbers* in quantum mechanics. In our example of the eclipsing spectroscopic binary star, the latent or state space is given by the variable θ varying continuously in the $[-\pi, \pi]$ interval[3].

The intrinsic dimensionality L of the system is unknown to the observer, who designs an experimental setup that allows to obtain one observation of the physical system by measuring a fixed number D of variables on it. Usually the number of observed variables D will be bigger (perhaps much bigger) than the true number of degrees of freedom L, because the observer needs to capture as much information from the system as possible. Any measured variable must be a function of the latent variables, so that we can represent the operation of obtaining one observation from the system by a *measurement* mapping $\mathbf{f} : \mathcal{S} \subset \mathbb{R}^L \longrightarrow \mathcal{M} \subset \mathbb{R}^D$. We will assume that the measurement mapping $\mathbf{f}$ is nonsingular[4], so that the observed variables will span an L-dimensional manifold $\mathcal{M} = \mathbf{f}(\mathcal{S})$ in $\mathbb{R}^D$; this manifold will be nonlinear in general. We refer to $\mathbb{R}^D$ as *data* or *observed space*. In our example of the eclipsing spectroscopic binary star, the observed variables would be the brightness B and the radial velocity v_r (or the spectral shift $\Delta\lambda$), which span a nonlinear one-dimensional manifold in $\mathbb{R}^2$, i.e., a plane curve. That is, if we plotted each observation in a two-dimensional coordinate system, with one axis being the brightness B and the other one the radial velocity v_r, all the points would fall on a curve instead of filling the plane or a two-dimensional region of it. This would make apparent the true dimensionality of the system (or, to be more strict, a lower bound of the true dimensionality).

Another set of observed variables, not measurable in practice for an eclipsing spectroscopic binary star, could be the x and y coordinates of the position vector of the secondary star drawn from the primary star (or

[3] One could argue that the state of the system is specified by the position variable θ and the velocity $(\dot{r}, \dot{\theta})$, but for our example both $\dot{r}$ and $\dot{\theta}$ are a function of θ.

[4] We say that a mapping $\mathbf{f}$ is *nonsingular* if the dimension of the image of $\mathbf{f}$ is equal to the dimension of its domain: $\dim \mathcal{S} = \dim \mathbf{f}(\mathcal{S})$. If $\mathbf{f}$ is linear, this means that the matrix associated with $\mathbf{f}$ is full-rank, or equivalently, that the image variables are linearly independent. If $\mathbf{f}$ is nonlinear, its Jacobian must be full-rank. See section A.7 for a discussion of L-manifolds in $\mathbb{R}^D$.

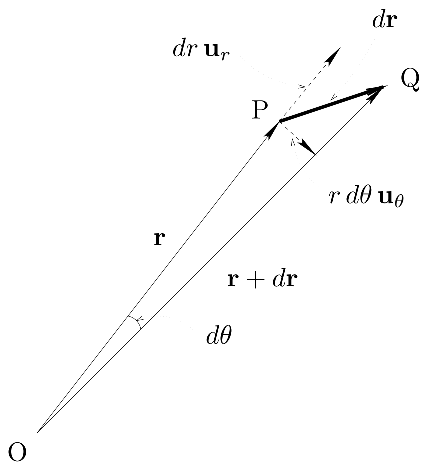


Figure 2.4: Differential displacement of a mobile point in plane polar coordinates (r, θ).

the polar coordinates (r, θ)!), which span precisely the elliptical orbit of figure 2.1. Yet another set of observed variables could be obtained by taking a picture of the binary system (again not possible practically for our example) and digitising it at $N \times N$ pixels; thus, $D = N^2$ could be made as large as desired, but the intrinsic dimension of the data, equal to the dimension of the generating process, would always be 1.

The mapping $\mathbf{f}$ describes an ideal **measurement process**, in which each point in latent space is exactly mapped onto a point in the L-dimensional manifold $\mathcal{M}$ in observed space. Thus, it would be possible to invert $\mathbf{f}$ and, for any point in $\mathcal{M}$, recover its corresponding point in latent space (again assuming that $L < D$ and that $\mathbf{f}$ is invertible): $\mathbf{f}^{-1} : \mathcal{M} \in \mathbb{R}^D \longrightarrow \mathcal{S} \in \mathbb{R}^L$. However, a real measurement process will introduce an error so that, for a latent point $\mathbf{x}$, the observer will measure a point $\mathbf{f}(\mathbf{x}) + \mathbf{e}$ in data space, where $\mathbf{e} \in \mathbb{R}^D$ is the error. The nature of this error is stochastic, i.e., repeated measurements of the system being in the same state would give different points in data space, even though the mapped point would be the same. The measurement error will hide to some extent the true, low-dimensional nature of the data, transforming the L-dimensional manifold $\mathcal{M}$ into a D-dimensional region in data space. The higher the error level, the more distortion $\mathcal{M}$ will suffer. It is customary to refer to this error as **noise**. We can assume that it follows a certain probability law, prescribed by a density function $p(\mathbf{t}|\mathbf{x})$, which gives the probability that latent point $\mathbf{x}$ will be observed as data point $\mathbf{t}$; we call this the *noise model*. The distribution of the noise will usually be a relatively generic, symmetric, zero-mean distribution, such as a normal $\mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Sigma})$, but any other distribution is acceptable if there is a reason for it (e.g. if one suspects that the error is systematic, a skewed distribution will be necessary).

2.2.3 Probabilities and trajectories

Here we analyse the probability distribution associated with planar movement[5]. This is an example intended to show how complex probability distributions of latent variables can arise even in a simple situation.

Assume we have a curve $\mathcal{C}$ parameterised by a certain variable θ; in this section, we will consider plane polar coordinates (r, θ) with $r(\theta) \geq 0$ and $-\pi \leq \theta \leq \pi$. We assume that any point of the curve $\mathcal{C}$ has a unique corresponding value of the variable θ, i.e., that $r(\theta)$ is invertible. We assume that a moving object is following the curve $\mathcal{C}$ with a certain instantaneous velocity of modulus $v(\theta) \geq 0$ at each point $(r(\theta), \theta)$. Given $r(\theta)$ and $v(\theta)$, we want to find a function $p(\theta)$ that gives at each point the probability density of finding the moving object at that point, i.e., the probability density of seeing the point under a given polar angle θ. That is, if we took a number of pictures of the object, we would find it more often in those parts of the curve where the velocity is small, or in those values of the variable θ for which the trajectory $r(\theta)$ varies quickly. Then, $p(\theta)\,d\theta$ is, by definition of probability density function, the probability that the object is in the interval $(\theta, \theta + d\theta)$. This probability will be proportional to the time dt that the object takes to move from $(r(\theta), \theta)$ to $(r(\theta + d\theta), \theta + d\theta)$: $p(\theta)\,d\theta \propto dt$. Assuming that the velocity is constant during the infinitesimal time interval dt, the arc length corresponding to the displacement $d\mathbf{r}$ will be $ds = v\,dt$, where $ds = |d\mathbf{r}|$ and $dr = d\,|\mathbf{r}|$ (see fig. 2.4). From the figure, we have $ds = \sqrt{(dr)^2 + (r\,d\theta)^2}$, where $dr = d\,|\mathbf{r}|$ is the variation in the radial

[5] Naturally, movement is not the only reason why probability distributions can appear in physics.

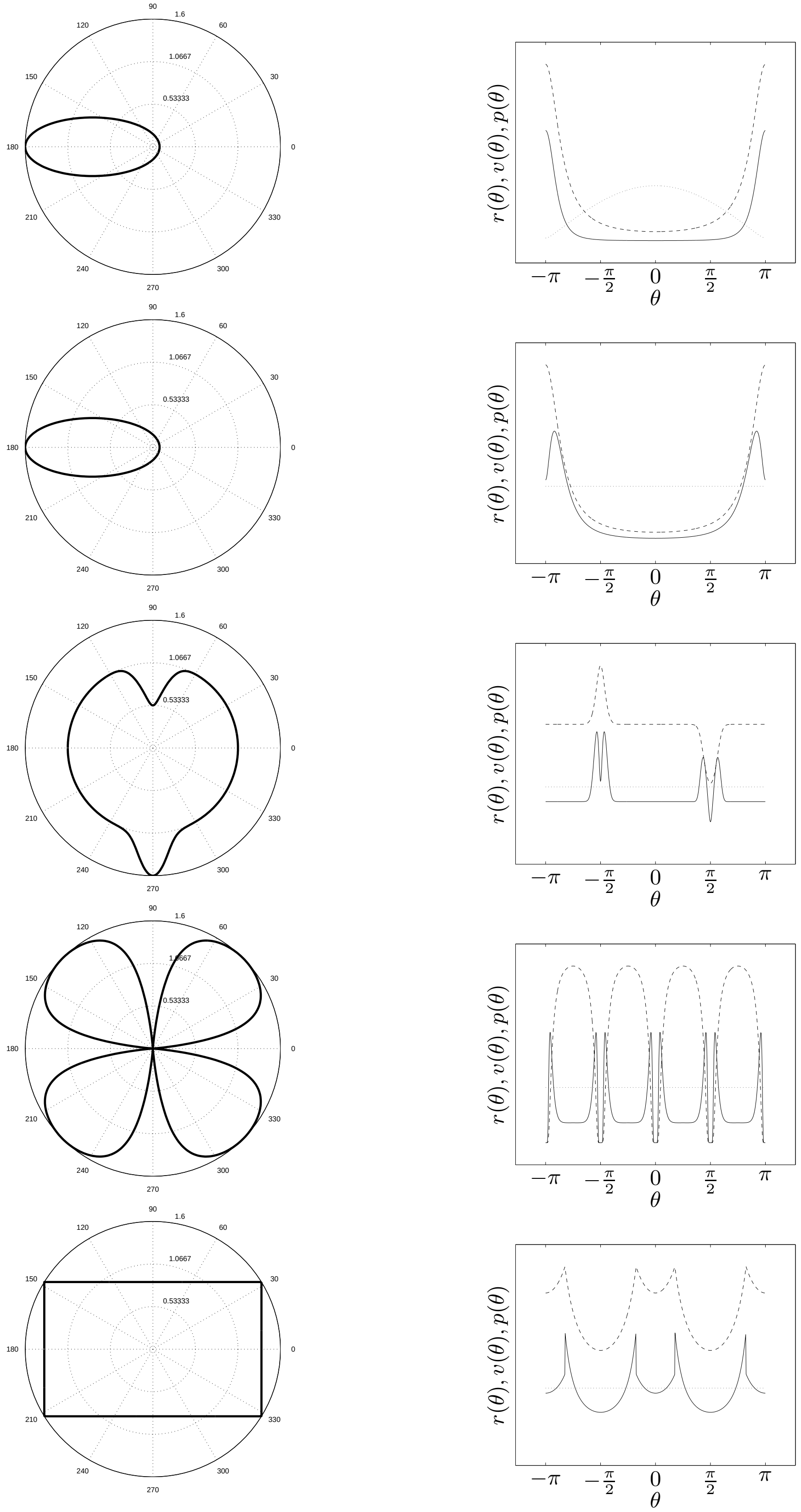


Figure 2.5: Examples of trajectories and the distributions generated by them. Each row represents a choice of the moduli of the radius vector $r(\theta)$ and the velocity $v(\theta)$ as functions of the polar angle θ, and each choice produces a certain probability distribution $p(\theta)$ of finding a mobile point at a certain polar angle θ. For each row, the figure on the left shows the trajectory $(r(\theta), \theta)$ in polar coordinates for $-\pi \leq \theta \leq \pi$ (thick line) and the graph on the right the form of the functions $r(\theta)$ (dashed line), $v(\theta)$ (dotted line) and $p(\theta)$ (solid line). All rows assume constant velocity modulus v except the first one, which assumes the Keplerian velocity (2.2).

direction and $r\,d\theta$ the variation in the tangential one. Then:

$$p(\theta) \propto \frac{dt}{d\theta} = \frac{1}{v}\frac{ds}{d\theta} = \frac{1}{v}\sqrt{\left(\frac{dr}{d\theta}\right)^2 + r^2} = \frac{1}{v}\sqrt{\dot{r}^2 + r^2}. \tag{2.1}$$

This function must be normalised dividing by $\int \frac{1}{v}\sqrt{\dot{r}^2 + r^2}\,d\theta$, the integral extended over the domain of θ. We can see from eq. (2.1) that the probability of finding the object in an interval $(\theta, \theta + d\theta)$ will be higher when:

- its velocity $v(\theta)$ is small there, or
- the distance to the origin $r(\theta)$ is high, or
- the trajectory varies quickly in the interval $(\theta, \theta + d\theta)$, i.e., the radial velocity $\dot{r}$ is large;

all in accordance with intuition. The trajectories for which the probability density is constant, when the point moves uniformly in θ-space, are given by the solutions $r(\theta)$, $v(\theta)$ of the differential equation $k^2v^2 = r^2 + \dot{r}^2$ for $k \in \mathbb{R}^+$. For example, for constant velocity this equation has two different solutions (where $R = kv$ is a positive constant):

- $r = R$, i.e., a circle centred at the origin;
- $r^2 + \dot{r}^2 = R^2 \Rightarrow r = R\,|\sin(\theta - \theta_0)|$, i.e., a circle passing through the origin.

For the familiar case of closed-orbit Keplerian movement of a two-body problem, the moduli of the radius vector and the velocity vary as a function of the polar angle θ as follows:

$$r(\theta) = \frac{a(1-\epsilon^2)}{1+\epsilon\cos\theta} \qquad v(\theta) = \sqrt{\mu\left(\frac{2}{r} - \frac{1}{a}\right)}. \tag{2.2}$$

That is, the trajectory is an ellipse of eccentricity $0 \le \epsilon < 1$ and major semiaxis a with a focus in the origin and $\mu = G(M+m)$, where G is the gravitational constant and M and m the masses of both bodies (Roy, 1978). According to the law of areas (Kepler's second law), the area swept by the radius vector per unit time is constant, i.e., $\left|\frac{\mathbf{r}\times d\mathbf{r}}{dt}\right|$ = constant (from fig. 2.4, $\left|\frac{1}{2}\mathbf{r} \times d\mathbf{r}\right|$ is the area of the triangle OPQ). Expanding $\mathbf{r} = r\mathbf{u}_r$ and $d\mathbf{r} = dr\,\mathbf{u}_r + r\,d\theta\,\mathbf{u}_\theta$ in the radial and tangential directions we obtain $r^2\frac{d\theta}{dt}$ = constant. From here and from the fact that $p(\theta) \propto \frac{dt}{d\theta}$ we obtain $p(\theta) \propto r^2$. The same result can be obtained by substituting the expressions for v and r from eq. (2.2) into eq. (2.1). Taking into account the following integral (Gradshteyn and Ryzhik, 1994):

$$\int_{-\pi}^{\pi} \frac{d\theta}{(1+\epsilon\cos\theta)^2} = \frac{2\pi}{(1-\epsilon^2)^{3/2}},$$

the exact expression for $p(\theta)$ turns out to be:

$$p(\theta) = \frac{(1-\epsilon^2)^{3/2}}{2\pi}\frac{1}{(1+\epsilon\cos\theta)^2} = \frac{r^2}{2\pi a^2\sqrt{1-\epsilon^2}}.$$

Figure 2.5 shows the trajectory and the form of the variation of r, v and p with θ for the example of Keplerian closed movement and for other trajectories for constant velocity. Observe how many different forms the distribution $p(\theta)$ can take, all physically possible, even for this simple example of planar movement. If we consider θ as a latent variable and the point coordinates (x, y) (for example) as observed variables, then $p(\theta)$ will be the prior distribution in latent space, as defined in section 2.3. Therefore, the distribution in latent space can be very complex, contrasting with the simple prior distributions of the latent variable models of section 2.6 (all of which are either normal of uniform, except for ICA). See sections 2.3.2 and 2.8 for a further discussion of this issue.

2.3 Generative modelling using continuous latent variables

In latent variable modelling the assumption is that the observed high-dimensional data is generated from an underlying low-dimensional process. The high dimensionality arises for several reasons, including stochastic variation and the measurement process. The objective is to learn the low dimensional generating process (defined by a small number of latent or hidden variables) along with a noise model, rather than directly

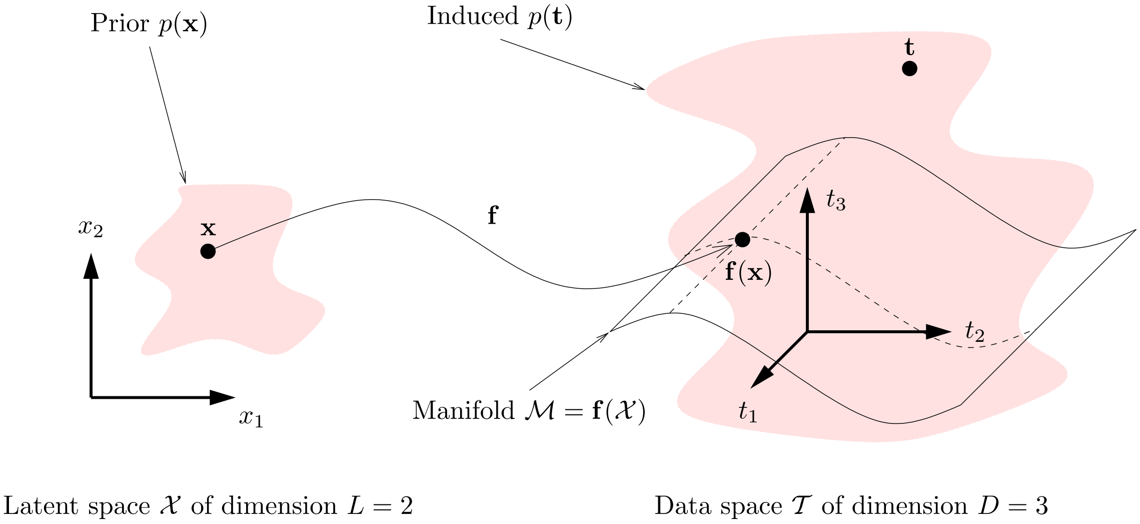

Prior $p(\mathbf{x})$
Induced $p(\mathbf{t})$
$\mathbf{t}$
x_2
$\mathbf{x}$
$\mathbf{f}$
t_3
$\mathbf{f}(\mathbf{x})$
t_2
x_1
Manifold $\mathcal{M} = \mathbf{f}(\mathcal{X})$
t_1
Latent space $\mathcal{X}$ of dimension $L = 2$
Data space $\mathcal{T}$ of dimension $D = 3$

Figure 2.6: Schematic of a continuous latent variable model with a 3-dimensional data space and a 2-dimensional latent space.

learning a dimensionality reducing mapping. We will consider that the latent variables are mapped by a fixed transformation into a higher-dimension observed space (measurement procedure) and noise is added there (stochastic variation). In contrast with this *generative*, bottom-up point of view, statisticians often consider latent variable models from an *explanatory*, top-down point of view (Bartholomew, 1987): given the empirical correlation between the observed variables, the mission of the latent variables is to explain those correlations via the axiom of local independence (explained in section 2.3.1); i.e., given an observed distribution, find a combination of latent distribution and noise model that approximates it well.

We will consider that both the observed and the latent variables are continuous. Call $\mathcal{T} \subseteq \mathbb{R}^D$ the D-dimensional **data** or **observed space**[6]. Consider an unknown distribution $p(\mathbf{t})$ in data space, for $\mathbf{t} \in \mathcal{T}$, of which we only see a sample $\{\mathbf{t}_n\}_{n=1}^N \subset \mathcal{T}$. In latent variable modelling we assume that the distribution in data space $\mathcal{T}$ is actually due to a small number $L < D$ of latent variables acting in combination. We refer to this L-dimensional space as the **latent space** $\mathcal{X} \subseteq \mathbb{R}^L$.

Thus, a point $\mathbf{x}$ in latent space $\mathcal{X}$ is generated according to a **prior distribution** $p(\mathbf{x})$ and it is mapped onto data space $\mathcal{T}$ by a smooth, nonsingular mapping $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{T}$. Because $\mathcal{M} = \mathbf{f}(\mathcal{X})$ is an L-dimensional manifold in $\mathcal{T}$, in order to extend it to the whole D-dimensional data space we define a distribution $p(\mathbf{t}|\mathbf{x}) = p(\mathbf{t}|\mathbf{f}(\mathbf{x}))$ on $\mathcal{T}$, called the **noise** or **error model**. Figure 2.6 illustrates the idea of latent variable models.

The joint probability density function in the product space $\mathcal{T} \times \mathcal{X}$ is $p(\mathbf{t}, \mathbf{x})$ and integrating over the latent space gives the marginal distribution in data space:

$$p(\mathbf{t}) = \int_{\mathcal{X}} p(\mathbf{t}, \mathbf{x}) \, d\mathbf{x} = \int_{\mathcal{X}} p(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) \, d\mathbf{x}. \tag{2.3}$$

This is called the **fundamental equation of latent variable models** by Bartholomew (1984). Thus, a model is essentially a specification of $p(\mathbf{x})$ and $p(\mathbf{t}|\mathbf{x})$—the specification of the mapping $\mathbf{f}$ can be absorbed into that of $p(\mathbf{t}|\mathbf{x})$. The only empirical evidence available concerns $p(\mathbf{t})$ through the sample $\{\mathbf{t}_n\}_{n=1}^N$ and so the only constraint on $p(\mathbf{x})$ and $p(\mathbf{t}|\mathbf{x})$, apart from the need to be nonnegative and integrate to 1, is given by eq. (2.3). In general, there are many combinations of $p(\mathbf{x})$ and $p(\mathbf{t}|\mathbf{x})$ that can satisfy (2.3) for a given $p(\mathbf{t})$.

Eq. (2.3) can also be seen as a continuous mixture model (Everitt and Hand, 1981), with the latent variable $\mathbf{x}$ "indexing" continuously the mixture component. In fact, any density function can be considered as a mixture density where extra variables have been integrated over.

Eq. (2.3) (or, for that matter, any marginalisation) can also be seen as a particular case of a *Fredholm integral equation of the first kind*. A Fredholm integral equation of the first kind (in one dimension) has the form:

$$f(t) = \int_{-\infty}^{\infty} K(t, x) g(x) \, dx$$

[6] In the statistical literature the variables in the data space are usually called *manifest variables*.

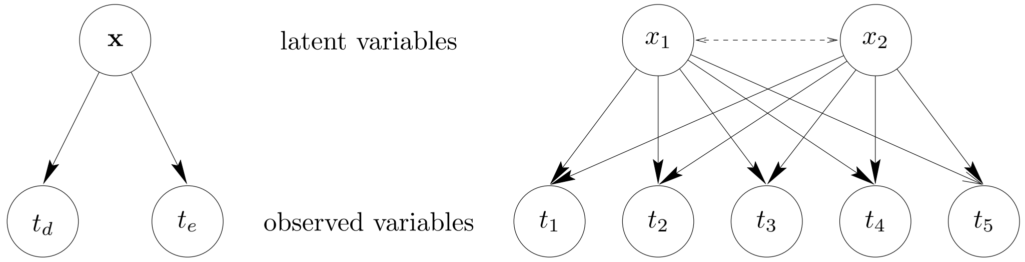


Figure 2.7: Graphical model representation of latent variable models. *Left*: the axiom of local independence. *Right*: a latent variable model with $D = 5$ observed variables and $L = 2$ latent variables. The dotted line indicates that the latent variables may or may not be independent in a particular model.

where the function $K(t, x)$ is called the kernel and both K and f are known while g is unknown. If all functions are probability densities and specifically $K(t, x) \equiv p(t|x)$ we obtain eq. (2.3). Note the similarity of the Fredholm integral equation with a matrix equation $\mathbf{f} = \mathbf{K}\mathbf{g}$, whose solution is $\mathbf{g} = \mathbf{K}^{-1}\mathbf{f}$. In fact, most inverse problems (to which chapter 6 is dedicated) are Fredholm integral equations of the first kind. Press et al. (1992, chapter 18) contains a brief discussion of numerical methods for Fredholm and Volterra integral equations and further references.

2.3.1 Noise model: the axiom of local independence

If the latent variables are to be efficient in representing faithfully the observed variables, we should expect that, given a value for the latent variables, the values of any group of observed variables are independent of the values of any other group of observed variables. Otherwise, the chosen latent variables would not completely explain the correlations between the observed variables and further latent variables would be necessary. Thus, for all $d, e \in \{1, \dots, D\}$,

$$p(t_d|t_e, \mathbf{x}) = p(t_d|\mathbf{x}) \Rightarrow p(t_d, t_e|\mathbf{x}) = p(t_d|t_e, \mathbf{x})p(t_e|\mathbf{x}) = p(t_d|\mathbf{x})p(t_e|\mathbf{x})$$

and we obtain that the distribution of the observed variables conditioned on the latent variables, or the noise model, is factorial (the subindex in p_d emphasises that the component distributions $p_d(t_d|\mathbf{x})$ need not be the same):

$$p(\mathbf{t}|\mathbf{x}) \overset{\text{def}}{=} \prod_{d=1}^{D} p_d(t_d|\mathbf{x}). \tag{2.4}$$

That is, for some $L \leq D$, *the observed variables are conditionally independent given the latent variables.* This is usually called the **axiom of local (or conditional) independence** (Bartholomew, 1984; Everitt, 1984). It will prove very convenient in section 7.12.3 because choosing factorial noise models simplifies the calculations of conditional distributions considerably.

It should be noted that, rather than an assumption, the axiom of local independence is a definition of what it means to have fully explained the joint distribution of the observed variables in terms of the latent ones. The aim is to find the smallest number of latent variables $L \leq D$ for which it holds (it does, trivially, for $L = D$ by taking $x_d \equiv t_d$, provided that (2.3) is satisfied). However, in practice one will need to try several values of L and select the best one.

Thus said, there have been some suggestions of models that violate the axiom of linear independence. One example has been proposed for the output distribution of a hidden Markov model: Gopinath et al. (1998) model the output distribution of state s as a factor analyser with nondiagonal covariance matrix $\mathbf{\Sigma}_T = \mathbf{U}\mathbf{\Psi}_s\mathbf{U}^T$ where $\mathbf{\Psi}_s$ is diagonal and $\mathbf{U}$ is an orthogonal matrix shared among states. It is not clear why this constrained covariance model should have any advantage over other parameter-tying methods (Young, 1996) or over using a mixture of factor analysers as output distribution (Saul and Rahim, 2000b).

The graphical model (Jensen, 1996; Jordan, 1998; Pearl, 1988; Whittaker, 1990) representing the local independence is shown in fig. 2.7 (left). Fig. 2.7 (right) shows the graphical model for a latent variable model with $D = 5$ and $L = 2$.

Regarding the actual choice of the functional form of the noise model $p(\mathbf{t}|\mathbf{f}(\mathbf{x}))$, it seems reasonable to use a density function with the following properties:

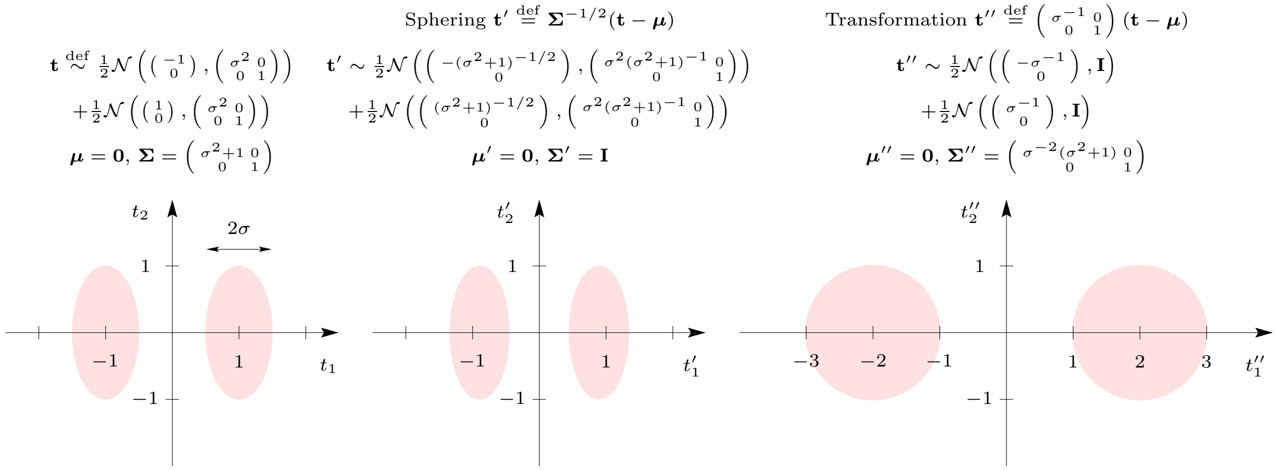


Figure 2.8: Sphering does not yield spherical local noise in general, as demonstrated here via a mixture of two normal distributions. *Left*: original distribution. *Centre*: sphered distribution. *Right*: linear transformation to local spherical noise (unknown, since the local noise is unknown in advance). Each normal distribution is represented as a unit standard deviation elliptical boundary in Mahalanobis distance. In each case, $\sigma = \frac{1}{2}$ and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix, respectively, of the mixture, computed as in section 8.7.

- It is centred at $\mathbf{f}(\mathbf{x})$, which would be the only possible point in the absence of noise:

  $$\forall \mathbf{x} \in \mathcal{X}: \ \mathrm{E}\left\{\mathbf{t}|\mathbf{x}\right\} = \mathbf{f}(\mathbf{x}). \tag{2.5}$$

  This is a relaxed form of the self-consistency condition (4.2) of principal curves.

- It decays gradually as the distance to $\mathbf{f}(\mathbf{x})$ increases, according to some parameter related to the noise covariance. However, it need not be symmetric around $\mathbf{f}(\mathbf{x})$.

- It assigns nonzero density to every point in the observed space, for two reasons:
  - No region of the observed space should have null probability unless knowledge about the problem at hand dictates otherwise.
  - Noise models which have a distribution with finite support (i.e., assigning nonnull probability only to points inside a finite neighbourhood of the centre, $\mathbf{f}(\mathbf{x})$) make very difficult to compute the data distribution, $p(\mathbf{t})$. This is so because the integration in latent space (2.3) to compute $p(\mathbf{t})$ for a given observed point $\mathbf{t}$ is restricted to those latent points $\mathbf{x}$ for which the neighbourhood of $\mathbf{f}(\mathbf{x})$ contains $\mathbf{t}$ (so that $p(\mathbf{t}|\mathbf{x}) > 0$).

- It should have a diagonal covariance matrix to account for different scales in the different observed variables $t_1, \dots, t_D$. This latter aspect cannot be overcome in general by sphering the data and using an isotropic noise model, because if the data has clusters, the intercluster distances affect the covariances; fig. 2.8 illustrates this with a simple example. The same happens if the data manifold is nonlinear. For cases where the dispersion attains orders of magnitude, a logarithmic transformation may be helpful.

In all the specific latent variable models discussed in section 2.6 (except ICA, which assumes no noise in its standard formulation) a normal noise model is assumed (thus symmetric), whose covariance matrix is either diagonal (factor analysis, IFA) or spherical (PCA, GTM). This is primarily due to the mathematical tractability of the multivariate normal distribution with respect to linear combinations, marginalisation and conditioning and its closure with respect to those operations—which simplifies the computation of integral (2.3) and the derivation of an EM algorithm for parameter estimation. It also adds flexibility to the model, since $p(\mathbf{t})$ ends up being a Gaussian mixture, which is a universal density approximator[7] (although the model itself may not be; see below). Yet another justification comes from the central limit theorem: if the noise is due to

[7] However, for mixtures the shape of each component is not crucial for density approximation, as is well known from kernel density estimation (Titterington et al., 1985; Scott, 1992), as long as there is a high enough number of components. Most localised kernels (decreasing from the central point) give rise to universal density approximators.

the combined additive action of a number of uncontrolled variables of finite variance, then its distribution will be asymptotically normal.

However, one disadvantage of the normal distribution is that its tails decay very rapidly, which reduces robustness against outliers: the presence of a small percentage of outliers in the training set can lead to significantly poor parameter estimates (Huber, 1981). Besides, many natural distributions have been proven to have long tails and would be better modelled by, say, a Student-t distribution or by infinite-variance members of the Lévy family[8], such as the Cauchy (or Lorentzian) distribution (Casti, 1997; Shlesinger et al., 1995). Another potential disadvantage of the normal distribution is its symmetry around the mean, when the noise is skewed. Skewed multivariate distributions may be obtained as normal mixtures or as multivariate extensions of univariate skewed distributions[9], but in both cases the analytical treatment may become very complicated—even if the axiom of local independence is followed, in which case a product of univariate skewed distributions may be used.

It should be noted that the noise may depend on the point in latent space (or on the point in data space). For example, in the binary system of section 2.2.1, the noise in the apastron area (point A in fig. 2.1) may be smaller than in the periastron area and its surroundings (points C to K) due to the interference with the central star. If, say, a normal noise model $\mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Psi})$ is assumed, then its covariance $\mathbf{\Psi}$ should be a function of $\mathbf{x}$ too: $\mathbf{\Psi} = \mathbf{\Psi}(\mathbf{x})$ (fig. 2.9). However, this would require implementing $\mathbf{\Psi}(\mathbf{x})$ via some function approximator (e.g. a multilayer perceptron) and the mathematical treatment becomes very complicated. None of the models described in section 2.6 implement a noise model dependent on $\mathbf{x}$, and therefore none of them would be able to represent a data distribution such as the one depicted in fig. 2.9. For example, GTM could probably capture the mapping $\mathbf{f}$ but would only find an average value for the noise covariance, thus missing the data density. In principle one could think that GTM could approximate any smooth density function, since the data density has the same form as a kernel estimator: eq. (2.8) or eq. (2.43) where the components are spherical Gaussian kernels of width, or smoothing parameter, σ. However, in the kernel estimator the kernel centres are free to move in data space while in GTM they are constrained by the mapping $\mathbf{f}$. Therefore, GTM is not a universal density approximator.

In latent variable models that use a constant noise model and a sampled latent space, like GTM, another disadvantage appears. The observed space distribution is a mixture in which all components have the same noise model, but are located at different places in observed space, as in eq. (2.8). Since the width of each component is the same, those areas of observed space that have low probability (few samples) will be assigned few, widely separated components compared to high-probability areas. As a result, the density estimate in those low-probability areas will not be smooth, presenting a characteristic ripple that gives rise to spurious modes, as discussed in section 7.9.1. This phenomenon is well-known in kernel density estimation with fixed-width kernels (Silverman, 1986, pp. 17–18) and is particularly noticeable in the tails of the distribution being approximated.

2.3.1.1 Latent variable models and principal curves

Hastie and Stuetzle (1989) define principal curves (reviewed in section 4.8) as smooth curves (or, in general, manifolds) that pass through the middle of a data set and satisfy the self-consistence condition (4.2), which we reproduce here for convenience:

$$\forall \mathbf{x} \in \mathcal{X}: \; \mathrm{E}\left\{\mathbf{t} | \mathbf{F}(\mathbf{t}) = \mathbf{x}\right\} = \mathbf{f}(\mathbf{x}) \qquad (4.2)$$

where the projection or dimensionality reduction mapping $\mathbf{F}: \mathbf{t} \in \mathcal{T} \longrightarrow \mathcal{X}$ is defined as the point in $\mathcal{X}$ whose image by $\mathbf{f}$ is closest to $\mathbf{t}$ in the Euclidean distance in $\mathcal{T}$. Ignoring boundary effects, $\mathbf{F}^{-1}(\mathbf{x})$ (the points in $\mathcal{T}$ projecting on $\mathbf{x}$) will be a subset of the manifold orthogonal to $\mathcal{M} = \mathbf{f}(\mathcal{X})$ at $\mathbf{f}(\mathbf{x})$, as fig. 2.10 shows.

[8] The Lévy, or stable, family of probability distributions (Feller, 1971, sec. VI.1–3) is defined as $\mathcal{F}\{P_N(x)\} \stackrel{\text{def}}{=} \mathcal{F}\{e^{-N|x|^\beta}\}$, where $\mathcal{F}$ is the Fourier transform of the probability $P_N(x)$ for N-step addition of random variables and $\beta \in (0, 2)$. Closed-forms for the pdf are only known in a few cases. All its members have an infinite variance and are scale invariant. The case $\beta = 1$ gives a Cauchy distribution and the limit case $\beta = 2$ gives the Gaussian distribution. The addition of N Lévy-distributed random variables follows a Lévy distribution. Lévy's theorem states that the addition of a number of random variables follows asymptotically a Lévy distribution (the Gaussian distribution if all the random variables have finite variance).

[9] Few useful multivariate extensions of nonnormal distributions exist. One such extension may be the multivariate skew-normal distribution $\mathcal{SN}(\mathbf{\Sigma}, \boldsymbol{\alpha})$ discussed by Azzalini and Capitanio (1999), with density:

$$p(\mathbf{t}) \stackrel{\text{def}}{=} 2\phi_D(\mathbf{t}; \mathbf{\Sigma})\Phi(\boldsymbol{\alpha}^T\mathbf{t}) \qquad \mathbf{t} \in \mathbb{R}^D$$

where ϕ_D is the D-variate pdf of the normal distribution with zero mean and covariance $\mathbf{\Sigma}$, Φ is the univariate cdf of the standard normal distribution $\mathcal{N}(0, 1)$ and the parameter $\boldsymbol{\alpha} \in \mathbb{R}^D$ partly regulates the skewness, with $\boldsymbol{\alpha} = \mathbf{0}$ giving the symmetric $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. This extension has some interesting properties, such as the fact that its marginal distributions are skew-normal too.

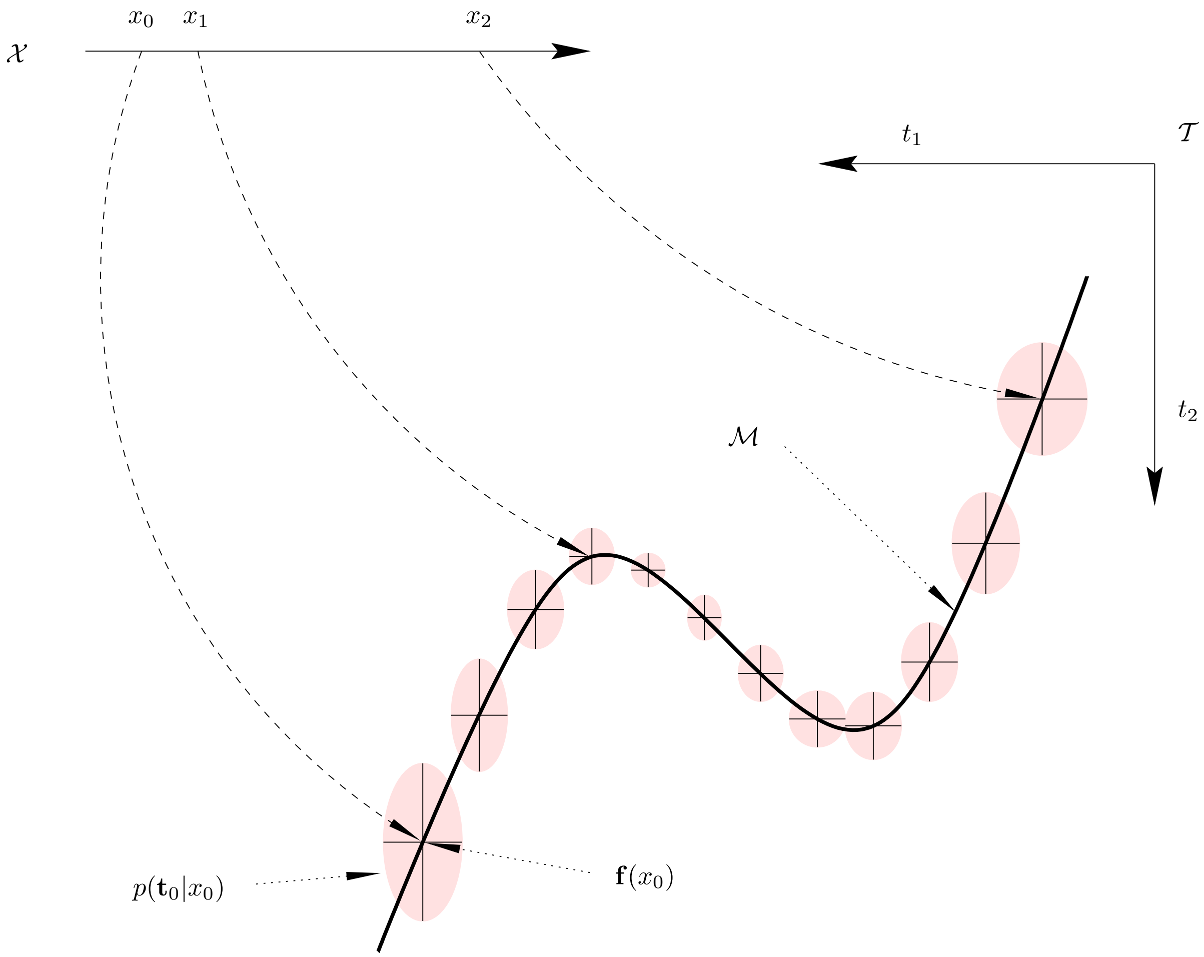


Figure 2.9: Variable noise in a latent variable model: the noise distribution $p(\mathbf{t}|\mathbf{x})$ depends on the latent point $\mathbf{x}$. The half bars on each ellipse measure one standard deviation of the corresponding noise distribution.

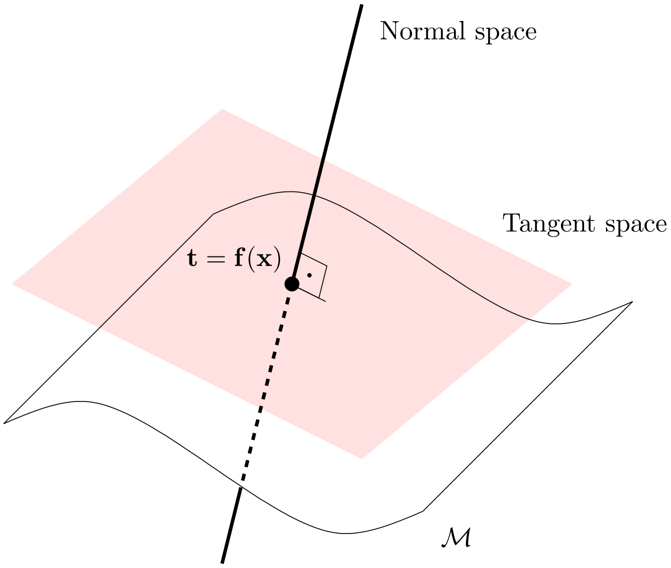


Figure 2.10: Self-consistency condition of principal curves. The graph shows a principal surface $\mathcal{M} = \mathbf{f}(\mathcal{X})$ (of dimension $L = 2$) and the normal space at $\mathbf{f}(\mathbf{x})$, in which $\mathbf{F}^{-1}(\mathbf{x})$, the set of data points projecting on a latent point $\mathbf{x}$, is contained. $\mathcal{M}$ is self-consistent if $\mathrm{E}\left\{\mathbf{t}|\mathbf{F}(\mathbf{t}) = \mathbf{x}\right\} = \mathbf{f}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$.

We can see that the principal curve self-consistency condition (4.2) is more restrictive than condition (2.5) (verified by unbiased latent variable models) in that the expectation is restricted to the set $\mathbf{F}^{-1}(\mathbf{x})$ rather than to the whole data space $\mathcal{T}$. In other words, the (unbiased) latent variable model condition (2.5) means that $p(\mathbf{t}|\mathbf{x})$ is centred at $\mathbf{f}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$, whereas the principal curves self-consistency condition (4.2) means that $p(\mathbf{t}|\mathbf{x})$ *restricted to the points in $\mathcal{T}$ projecting onto* $\mathbf{x}$ (which is a subset of the normal hyperplane to $\mathcal{M}$ at $\mathbf{f}(\mathbf{x})$) is centred at $\mathbf{f}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$.

Clearly then, redefining self-consistency as condition (2.5) and considering both $\mathbf{t}$ and $\mathbf{x}$ as random variables turns the principal curve into a latent variable model. This is what Tibshirani (1992) did, seeking to eliminate the bias intrinsic to the standard definition of principal curves. He then approached the estimation problem as nonparametric estimation of a continuous mixture (section 2.5.2) and, by assuming a diagonal normal distribution for $p(\mathbf{t}|\mathbf{f}(\mathbf{x}))$, reduced it to EM estimation of a Gaussian mixture. As mentioned in section 2.5.2, this has the disadvantage of defining the principal curve $\mathbf{f}$ only through a finite collection of points in $\mathcal{X}$ rather than as a smooth function defined for all points in $\mathcal{X}$.

Under a generative view, such as that adopted for latent variable models (section 2.3), the self-consistency condition (4.2) is unnatural, since it means that for a point that has been generated in latent space and mapped onto a data space point $\mathbf{t}$, the error added to $\mathbf{t}$ must be "clever" enough to perturb the point $\mathbf{t}$ in a direction orthogonal to the manifold $\mathcal{M}$ at $\mathbf{t}$ and do so with mean zero! This only seems possible in the trivial case where the tangent manifold is constant in direction, i.e., the principal curve has curvature zero. This is the case of the principal component subspaces and normal distributions, in which case latent variable models and principal curves coincide.

2.3.2 Prior distribution in latent space: interpretability

Given any prior distribution $p_{\mathbf{x}}(\mathbf{x})$ of the L latent variables $\mathbf{x}$, it is always possible to find an invertible transformation $\mathbf{g}$ to an alternative set of L latent variables $\mathbf{y} = (y_1, \dots, y_L) = \mathbf{g}(\mathbf{x})$ having another desired distribution $p_{\mathbf{y}}(\mathbf{y})$:

$$\begin{tikzcd}\mathcal{X} \arrow[r, "\mathbf{g}"] \arrow[rr, bend right, "\mathbf{f}"'] & \mathcal{Y} \arrow[r, "\mathbf{f}'"] & \mathcal{T}\end{tikzcd}$$

The mapping from the new latent space onto the data space becomes $\mathbf{f}' = \mathbf{f} \circ \mathbf{g}^{-1}$, i.e., $\mathbf{t} = \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{g}^{-1}(\mathbf{y})) = \mathbf{f}'(\mathbf{y})$ and the new prior distribution of the $\mathbf{y}$ variables becomes $p_{\mathbf{y}} = p_{\mathbf{x}} \left|\mathbf{J}_{\mathbf{g}}\right|^{-1}$, where $\mathbf{J}_{\mathbf{g}} \overset{\text{def}}{=} \left(\frac{\partial g_l}{\partial x_k}\right)$ is the Jacobian of the transformation $\mathbf{g}$. That is, given a space $\mathcal{X}$ with a distribution $p_{\mathbf{x}}$, to transform it into a space $\mathcal{Y}$ with a distribution $p_{\mathbf{y}}$ we apply[10] an invertible mapping $\mathbf{g} : \mathcal{X} \to \mathcal{Y}$ of Jacobian $\left|\mathbf{J}_{\mathbf{g}}\right| = \frac{p_{\mathbf{x}}}{p_{\mathbf{y}}}$. For example, if $\mathbf{x}$ are independently and normally distributed, transforming $y_l = e^{x_l}$ for $l = 1, \dots, L$ means that y_l follow a log-normal distribution.

Thus, fixing the functional form of the prior distribution in latent space is a convention rather than an assumption[11]. However, doing so requires being able to select $\mathbf{f}$ and $p_d(\mathbf{t}|\mathbf{x})$ from a broad class of functions so that eq. (2.3) still holds. GTM (section 2.6.5) is a good example: while it keeps the prior $p(\mathbf{x})$ simple (discrete uniform), its mapping $\mathbf{f}$ is a generalised linear model (which has universal approximation capabilities).

In particular, we can choose the latent variables to be independent[12] and identically distributed, $p_{\mathbf{x}}(\mathbf{x}) \overset{\text{def}}{=} \prod_{l=1}^{L} p(x_l)$. Unfortunately, while this is a particularly simple and symmetric choice for prior distribution in latent space, it does not necessarily simplify the calculation of the difficult L-dimensional integral (2.3), which is a major shortcoming of the latent variable modelling framework (see section 2.4).

We coincide with Bartholomew (1985) that the latent variables must be seen as constructs designed to simplify and summarise the observed variables, without having to look for an interpretation for them (which may exist in some cases anyway): all possible combinations of $p(\mathbf{x})$ and $p(\mathbf{t}|\mathbf{x})$ are equally valid as long as they satisfy eq. (2.3). Thus, we do not need to go into the issue of the interpretation of the latent variables, which has plagued the statistical literature for a very long time without reaching a general consensus.

Another issue is the topology of the true data manifold. For the one-dimensional example of figure 2.13, the data manifold is a closed curve (an ellipse). Thus, modelling it with a latent space that has the topology of an open curve (e.g. an interval of the Cartesian coordinate x) will lead to a discontinuity where both ends of

[10] Unfortunately, this is terribly complicated in practice, since solving the nonlinear system of partial differential equations $\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| = \frac{p_{\mathbf{x}}}{p_{\mathbf{y}}}$ to obtain $\mathbf{y} = \mathbf{g}(\mathbf{x})$ will be impossible except in very simple cases.

[11] The choice of a prior distribution in the latent space is completely different from, and much simpler than, the well-known problem of Bayesian analysis of choosing a noninformative prior distribution for the parameters of a model (mentioned in section 6.2.3.2)

[12] When $\mathbf{g}$ is linear and $y_1, \dots, y_L$ are independent, this is exactly the objective of independent component analysis (section 2.6.3).

the curve join: it is impossible to have a continuous mapping between spaces with different topological characteristics without having singularities, e.g. mapping a circle onto a line. A representation using a periodic latent variable is required[13] (e.g. the polar angle θ). Although some techniques exist for Gaussian mixture modelling of periodic variables (Bishop and Nabney, 1996), all the latent variable models considered in section 2.6 assume non-periodic latent variables.

A flexible and powerful representation of the prior distribution can be obtained with a mixture:

$$p(\mathbf{x}) \overset{\text{def}}{=} \sum_{m=1}^{M} p(m) p(\mathbf{x}|m)$$

which keeps the marginalisation in data space (2.3) analytically tractable (if the component marginals are):

$$p(\mathbf{t}) = \int_{\mathcal{X}} p(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) \, d\mathbf{x} = \sum_{m=1}^{M} p(m) \int_{\mathcal{X}} p(\mathbf{t}|\mathbf{x}) p(\mathbf{x}|m) \, d\mathbf{x}. \tag{2.6}$$

By making the latent space distribution more complex we can have a simpler mapping $\mathbf{f}$. The independent factor analysis model of Attias (1999), discussed in section 2.6.4, uses this idea, where $p(\mathbf{x})$ is a product of Gaussian mixtures, the mapping is linear and the noise model is normal.

2.3.3 Smooth mapping from latent onto data space: preservation of topographic structure

In section 2.3 we required the mapping $\mathbf{f} : \mathcal{X} \to \mathcal{T}$ to be smooth, that is, continuous and differentiable. There are two reasons for this:

- Continuity: this, by definition, will guarantee that points which lie close to each other in the latent space will be mapped onto points which will be close to each other in data space. In the context of Kohonen maps and similar algorithms for dimensionality reduction this is called topology or topography preservation (although quantifying this topography preservation is difficult; Bauer and Pawelzik, 1992; Martinetz and Schulten, 1994; Kohonen, 1995; Bezdek and Pal, 1995; Goodhill and Sejnowski, 1997; Villmann et al., 1997; Bauer et al., 1999). It expresses the essential requirement that a continuous trajectory followed by a point in latent space will generate a continuous trajectory in data space (without abrupt jumps). The question of whether the dimensionality reduction mapping from data space onto latent space is continuous will be dealt with in section 2.9.2.
- (Piecewise) differentiability: this is more of a practical requirement in order to be able to use derivative-based optimisation methods, such as gradient descent.

The more general the class of functions from which we can pick $\mathbf{f}$ is (expressed through its parameters), the more flexible the latent variable model is (in that more data space distributions $p(\mathbf{t})$ can be constructed for a given prior $p(\mathbf{x})$, eq. (2.3)), but also the more complex the mathematical treatment becomes and the more local optima appear. The quality of the local optima found for models which are very flexible can be dreadful, as that shown in fig. 2.13 for GTM. The problem of local optima is very serious since it affects all local optimisation methods, i.e., methods that start somewhere in parameter space and move towards a nearby optimum, such as the EM algorithm and gradient or Newton methods. The only way for such methods to find a good optimum is to start them from many different locations, but this really does not guarantee any good results.

In section 2.3 we also required the mapping $\mathbf{f}$ to be nonsingular (as defined in section 2.2.2). This is to ensure that the manifold $\mathcal{M} = \mathbf{f}(\mathcal{X})$ has the same dimension as the latent space $\mathcal{X}$; otherwise we would be wasting latent variables.

2.4 The problem of the marginalisation in high dimensions

The latent variable framework is very general, accomodating arbitrary mappings and probability distributions. However, this presents insurmountable mathematical and computational difficulties—particularly in the analytical evaluation of integral (2.3) but also when maximising the log-likelihood (2.9)—so that the actual choice

[13] It is not enough that a curve self-intersects, as it nearly happens in plot A of fig. 2.13: the almost coincident end points $\mathbf{f}(x_1)$ and $\mathbf{f}(x_K)$ of the model manifold correspond to the widely separated end latent grid points x_1 and x_K.

is limited. In fact, the only[14] tractable case in arbitrary dimensions seems to be when both the prior in latent space $p(\mathbf{x})$ and the noise model $p(\mathbf{t}|\mathbf{x})$ are Gaussian (or mixtures of Gaussians) and the mapping $\mathbf{f}$ linear; or when $p(\mathbf{x})$ is a mixture of Dirac deltas (as a result of Monte Carlo sampling) and the mapping is nonlinear; or when $p(\mathbf{t}|\mathbf{x})$ is a Dirac delta[15]. Combinations of these give the specific latent variable models of section 2.6.

Conditioning or marginalising a multivariate distribution, such as the joint distribution $p(\mathbf{t}, \mathbf{x})$ of eq. (2.3), requires the evaluation of an integral in several dimensions of the form:

$$I(\mathbf{u}) = \int_{\mathcal{V}} \mathbf{f}(\mathbf{u}, \mathbf{v}) p(\mathbf{v}) \, d\mathbf{v} \qquad \mathbf{u} \in \mathcal{U} \subseteq \mathbb{R}^U, \quad \mathbf{v} \in \mathcal{V} \subseteq \mathbb{R}^V. \tag{2.7}$$

This integral cannot be evaluated analytically for most forms of the function $\mathbf{f}$ and the distribution $p(\mathbf{v})$. In this case, a conceptually simple but computationally expensive workaround is to approximate it by Monte Carlo integration, as MacKay (1995a) suggested. For the case (2.7) this means sampling K times in the space $\mathcal{V}$ from $p(\mathbf{v})$ and approximating $I(\mathbf{u}) \approx \frac{1}{K} \sum_{k=1}^K \mathbf{f}(\mathbf{u}, \mathbf{v}_k)$, with an error of order $1/\sqrt{K}$ (Press et al., 1992, section 7.6). However, the sample size K in a space of V dimensions grows exponentially with V, as does the hypervolume of the V-dimensional region (this is basically the curse of the dimensionality, discussed in section 4.3). This severely limits the practical use of Monte Carlo methods. However, it should be noted that K is not a parameter of the model (it does not take part in the estimation) and so increasing it does not produce overfitting.

For eq. (2.3) Monte Carlo sampling yields

$$p(\mathbf{t}) \approx \frac{1}{K} \sum_{k=1}^K p(\mathbf{t}|\mathbf{x}_k) \tag{2.8}$$

with $\{\mathbf{x}_k\}_{k=1}^K$ drawn from $p(\mathbf{x})$. GTM (section 2.6.5) uses this method with a uniform $p(\mathbf{x})$. For the log-likelihood gradient $\nabla_{\mathbf{\Theta}} \mathcal{L}$ of eq. (2.10) Monte Carlo sampling yields

$$\nabla_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}) \approx \sum_{n=1}^N \frac{\sum_{k=1}^K \nabla_{\mathbf{\Theta}} p(\mathbf{t}_n|\mathbf{x}_k, \mathbf{\Theta})}{\sum_{k=1}^K p(\mathbf{t}_n|\mathbf{x}_k, \mathbf{\Theta})}.$$

The aim of integral (2.3) is not to obtain the numerical value of $p(\mathbf{t})$ at a given $\mathbf{t}$, but to obtain an analytical expression for $p(\mathbf{t})$ dependent on the parameters $\mathbf{\Theta}$ (section 2.5). This is so because to estimate the model parameters (e.g. by an EM algorithm) we generally need to be able to take the derivative of $p(\mathbf{t}|\mathbf{\Theta})$ with respect to $\mathbf{\Theta}$.

2.5 Parameter estimation

The prior in latent space $p(\mathbf{x})$, the smooth mapping $\mathbf{f}$ and the noise model $p(\mathbf{t}|\mathbf{x})$ are all equipped with parameters[16] which we collectively call $\mathbf{\Theta}$. Once their functional forms are fixed, we have freedom to set the parameters to those values that agree best with the data sample. These parameters are optimised, typically to maximise the likelihood of the observed data given the parameters, $p(\mathbf{t}_n|\mathbf{\Theta})$. This approach has the well-known problems of overfitting and model selection and could be overcome by a Bayesian treatment. In a Bayesian treatment, a prior distribution is placed on the parameters and all subsequent inferences are done by marginalising over the parameters:

$$p(\mathbf{t}, \mathbf{x}) = \int p(\mathbf{t}, \mathbf{x}|\mathbf{\Theta}) p(\mathbf{\Theta}) \, d\mathbf{\Theta}$$

where $p(\mathbf{\Theta})$ is the prior parameter distribution or the posterior parameter distribution after having seen some data $\{\mathbf{t}_n\}_{n=1}^N$. However, this adds an extra degree of intractability to that of eq. (2.3) and approximations are required. For example, there is current interest in approximate Bayesian inference using Markov chain

[14] One might think that using uniform distributions may simplify the mathematics. However, while they do simplify the expression of the integrand in (2.3), they complicate the integration region, preventing further treatment for any kind of mapping (Carreira-Perpiñán, 1997).

[15] This is the zero-noise case, which is not interesting since data space points not in $\mathbf{f}(\mathcal{X})$ receive zero density (although the standard formulation of ICA has zero noise).

[16] Strictly, the dimensionality L of the latent space is also a parameter of the latent variable model, but we will consider it fixed to some value, due to the practical difficulty of optimising it jointly with the other parameters.

Monte Carlo methods (Besag and Green, 1993; Brooks, 1998; Gilks et al., 1996; Neal, 1993) and variational methods (Jordan et al., 1999), and this has been applied to some latent variable models, such as principal component analysis (Bishop, 1999) and mixtures of factor analysers (Ghahramani and Beal, 2000; Utsugi and Kumagai, 2001). However, these are still preliminary results and the potential gain of using the Bayesian treatment (notably the autodetection of the optimal number of latent variables and mixture components) may not warrant the enormous complication of the computations, at least in high dimensions. In this thesis we will only consider maximum likelihood estimation unless indicated otherwise.

The log-likelihood of the parameters given the sample $\{\mathbf{t}_n\}_{n=1}^N$ is (assuming $\mathbf{t}_1, \dots, \mathbf{t}_N$ independent and identically distributed random variables):

$$\mathcal{L}(\mathbf{\Theta}) \overset{\text{def}}{=} \ln p(\mathbf{t}_1, \dots, \mathbf{t}_N | \mathbf{\Theta}) = \ln \prod_{n=1}^N p(\mathbf{t}_n|\mathbf{\Theta}) = \sum_{n=1}^N \ln p(\mathbf{t}_n|\mathbf{\Theta}) \tag{2.9}$$

which is to be maximised under the maximum likelihood criterion for parameter estimation. This will provide with a set of values for the parameters, $\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta})$, corresponding to a (local) maximum of the log-likelihood. The log-likelihood value $\mathcal{L}(\mathbf{\Theta}^*)$ allows the comparison of any two latent variable models (or, in general, any two probability models), however different these may be (although from a Bayesian point of view, in addition to their likelihood, one should take into account a prior distribution for the parameters and the evidence for the model; MacKay, 1995b).

One maximisation strategy is to find the stationary points of the log-likelihood (2.9):

$$\nabla_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}) = \sum_{n=1}^N \frac{1}{p(\mathbf{t}_n|\mathbf{\Theta})} \nabla_{\mathbf{\Theta}} p(\mathbf{t}_n|\mathbf{\Theta}) = \mathbf{0} \tag{2.10}$$

but maximum likelihood optimisation is often carried out using an **EM algorithm** (Dempster et al., 1977; McLachlan and Krishnan, 1997), which is usually simpler and is guaranteed to increase the log-likelihood monotonically. In the EM approach to latent variable models, the latent variables $\{\mathbf{x}_n\}_{n=1}^N$ (one per data point) are considered missing[17]. If their values were known, estimation of the parameters (e.g. the $\mathbf{\Lambda}$ matrix in eq. (2.14)) would be straightforward by least squares. However, for a given data point $\mathbf{t}_n$ we do not know the value of $\mathbf{x}_n$ that generated it. The EM algorithm operates in two steps which are repeated alternatively until convergence:

E step computes the expectation of the complete data log-likelihood with respect to the current posterior distribution $p(\mathbf{x}_n|\mathbf{t}_n, \mathbf{\Theta}^{(\tau)})$ (i.e., using the current parameter values), traditionally notated $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(\tau)})$:

$$Q(\mathbf{\Theta}|\mathbf{\Theta}^{(\tau)}) \overset{\text{def}}{=} \sum_{n=1}^N \mathrm{E}_{p(\mathbf{x}_n|\mathbf{t}_n,\mathbf{\Theta}^{(\tau)})} \left\{\mathcal{L}_{n,\text{complete}}(\mathbf{\Theta})\right\} \text{ where } \mathcal{L}_{n,\text{complete}}(\mathbf{\Theta}) \overset{\text{def}}{=} \ln p(\mathbf{t}_n, \mathbf{x}_n|\mathbf{\Theta}).$$

Thus, we average over the missing latent variables $\{\mathbf{x}_n\}_{n=1}^N$, effectively filling in their unknown values. Computing $\mathcal{L}_{n,\text{complete}}(\mathbf{\Theta})$ is possible because the joint distribution $p(\mathbf{t}, \mathbf{x}|\mathbf{\Theta})$ is known for the latent variable model in question.

M step determines new parameter values $\mathbf{\Theta}^{(\tau+1)}$ that maximise the expected complete-data log-likelihood:

$$\mathbf{\Theta}^{(\tau+1)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(\tau)}).$$

This increases the log-likelihood $\mathcal{L}(\mathbf{\Theta})$ unless it is already at a local maximum.

The standard EM algorithm has some disadvantages:

- It is a batch algorithm. However, by interpreting EM as an alternating maximisation of a negative free-energy-like function (Neal and Hinton, 1998), it is possible to derive online EM algorithms, suitable for online learning (e.g. in sequential tasks, where the data come one at a time).

- Its slow convergence after the first few steps, which are usually quite effective. Also, the greater the proportion of missing information, the slower the rate of convergence of EM (Dempster et al., 1977). However, methods for accelerating it are available; see, for example, Meng and van Dyk (1997), McLachlan and Krishnan (1997) and references therein.

[17] Depending on the model, additional missing variables may have to be introduced. For example, the component labels for mixture models.

Despite these shortcomings, EM usually remains the best choice for parameter estimation thanks to its reliability.

For a general choice of prior in latent space, mapping between latent and data space and noise model the log-likelihood surface can have many local maxima of varying height. In some cases, some or all of those maxima are equivalent, in the sense that the model produces the same distribution (and therefore the same log-likelihood value at all the maxima), i.e., the model is not identifiable (section 2.8). This is often due to symmetries of the parameter space, such as permutations (e.g. PCA) or general rotations of the parameters (e.g. factor analysis). In those cases, the procedure to follow is to find a first maximum likelihood estimate of the parameters (in general, by some suitable optimisation method, e.g. EM, although sometimes an analytical solution is available, as for PCA) and then possibly apply a transformation to them to take them to a canonical form satisfying a certain criterion (e.g. varimax rotation in factor analysis).

2.5.1 Relation of maximum likelihood with other estimation criteria

Least squares Using the least-squares reconstruction error as objective function for parameter estimation gives in general different estimates as the maximum likelihood criterion, although the latter usually results in a low reconstruction error. If we consider the unobserved values $\{\mathbf{x}_n\}_{n=1}^N$ as fixed parameters rather than random variables and assume that the noise model is normal with isotropic known variance, then a penalised maximum likelihood criterion results in

$$\sum_{n=1}^{N} \left\|\mathbf{t}_n - \mathbf{f}(\mathbf{x}_n)\right\|^2 + \text{ penalty term on } \mathbf{f}$$

which coincides with the spline-related definition of principal curves given by Hastie and Stuetzle (1989), mentioned in section 4.8.

Kullback-Leibler distance For $N \to \infty$, the normalised log-likelihood of $\mathbf{\Theta}$ converges in probability to its expectation by the law of large numbers:

$$\begin{aligned} \mathcal{L}_N(\mathbf{\Theta}) \overset{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} \ln p(\mathbf{t}_n|\mathbf{\Theta}) \xrightarrow{\mathcal{P}} \mathcal{L}_\infty(\mathbf{\Theta}) &\overset{\text{def}}{=} \mathrm{E}_{p_\mathbf{t}}\left\{\ln p(\mathbf{t}|\mathbf{\Theta})\right\} \\ &= \int_{\mathcal{T}} p_\mathbf{t}(\mathbf{t}) \ln p(\mathbf{t}|\mathbf{\Theta})\, d\mathbf{t} = -h(p_\mathbf{t}) - \mathrm{D}\left(p_\mathbf{t} \| p(\cdot|\mathbf{\Theta})\right) \end{aligned} \tag{2.11}$$

for any $\mathbf{\Theta}$. Since the entropy of the data distribution $h(p_\mathbf{t})$ does not depend on the parameters $\mathbf{\Theta}$, maximising the log-likelihood is asymptotically equivalent to minimising the Kullback-Leibler distance to the data density.

2.5.2 Relation with nonparametric estimation of continuous mixtures

As mentioned in section 2.3, the fundamental equation (2.3) can be interpreted as a continuous mixture model for $\mathbf{t}$, where $\mathbf{x}$ is the mixing variable. Assume that the functional form of $p(\mathbf{t}|\mathbf{x})$ is known and depends on $\mathbf{f}(\mathbf{x})$ and on parameters $\boldsymbol{\theta}(\mathbf{x})$. The log-likelihood of this model is (call $p_\mathbf{x}$ the density of the mixing variable $\mathbf{x}$):

$$\mathcal{L}(p_\mathbf{x}, \mathbf{f}, \boldsymbol{\theta}) \overset{\text{def}}{=} \sum_{n=1}^{N} \ln \int_{\mathcal{X}} p(\mathbf{t}_n|\mathbf{f}(\mathbf{x}), \boldsymbol{\theta}(\mathbf{x}))p(\mathbf{x})\, d\mathbf{x}.$$

Results from the theory of mixtures (Laird, 1978; Lindsay, 1983) dictate that for fixed $\mathbf{f}$ and $\boldsymbol{\theta}$ the nonparametric maximum likelihood estimate for $p(\mathbf{x})$ uniquely exists and is discrete with at most N support points (where N is the sample size). Denote these support points by $\{\mathbf{x}_m\}_{m=1}^M \subset \mathcal{X}$ where $M \le N$. This results then in

$$p(\mathbf{t}; \mathbf{\Theta}) = \sum_{m=1}^{M} p_m p(\mathbf{t}|\mathbf{f}_m, \boldsymbol{\theta}_m) \tag{2.12}$$

where $\mathbf{\Theta} \overset{\text{def}}{=} \{p_m, \mathbf{f}_m, \boldsymbol{\theta}_m\}_{m=1}^M$ contains the values of $\mathbf{f}$, $\boldsymbol{\theta}$ and $p_\mathbf{x}$ at the M unknown support points:

$$\mathbf{f}_m \overset{\text{def}}{=} \mathbf{f}(\mathbf{x}_m) \qquad \boldsymbol{\theta}_m \overset{\text{def}}{=} \boldsymbol{\theta}(\mathbf{x}_m) \qquad p_m \overset{\text{def}}{=} p(\mathbf{x}_m).$$

Model	*Prior in latent space* $p(\mathbf{x})$	*Mapping* $\mathbf{f}$ $\mathbf{x} \to \mathbf{t}$	*Noise model* $p(\mathbf{t}\|\mathbf{x})$	*Density in observed space* $p(\mathbf{t})$
Factor analysis (FA)	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	linear	diagonal normal	constrained Gaussian
Principal component analysis (PCA)	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	linear	spherical normal	constrained Gaussian
Independent component analysis (ICA)	unknown but factorised	linear	Dirac delta	depends
Independent factor analysis (IFA)	product of 1D Gaussian mixtures	linear	normal	constrained Gaussian mixture
Generative topographic mapping (GTM)	discrete uniform	generalised linear model	spherical normal	constrained Gaussian mixture

Table 2.2: Summary of specific continuous latent variable models.

So we end up with a parametric finite mixture model of M components and parameters $\mathbf{\Theta}$ which must be estimated from the sample $\{\mathbf{x}_n\}_{n=1}^N$: $\{p_m\}$ are the mixing proportions and, for each component m, $\mathbf{f}_m$ and $\boldsymbol{\theta}_m$ could be (for example) location and covariance parameters. Note that the log-likelihood is not a function of the support points $\{\mathbf{x}_m\}_{m=1}^M$, but only of $\mathbf{\Theta}$. An obvious choice of optimisation algorithm would be EM (Redner and Walker, 1984; McLachlan and Krishnan, 1997).

The nonparametric approach to parameter estimation in latent variable models has an important disadvantage: it only finds the values of $\mathbf{f}$ and $p_{\mathbf{x}}$ at the support points. To obtain their values at other, intermediate, points—which is necessary to perform dimensionality reduction and reconstruction (section 2.9)—they must be smoothed or interpolated using the known values, $\{\mathbf{f}_m\}_{m=1}^M$ for $\mathbf{f}$ and $\{p_m\}_{m=1}^M$ for $p_{\mathbf{x}}$. This is as hard a problem as the original one of estimating $\mathbf{f}$ parametrically. A less important disadvantage is that of the singularities of the log-likelihood due to the use of separate parameters for each component: if $\mathbf{f}_m$ becomes equal to some data point $\mathbf{t}_n$, then if its variance parameter $\boldsymbol{\theta}_m \to \mathbf{0}$ then $\mathcal{L} \to \infty$. This requires regularising the model.

Although the expression for $p(\mathbf{t})$ in eq. (2.12) looks similar to the one obtained by Monte Carlo sampling of the latent space, eq. (2.8), there is an important difference: the Monte Carlo method preserves a parametric form for the mapping $\mathbf{f}$, which the nonparametric method has lost.

Knowing that the number of support points is at most N is not useful in practice, since fitting a mixture of more than N components (each with separate parameters) to N data points would result in overfitting as well as being computationally expensive.

2.6 Specific latent variable models

A latent variable model is specified by the functional forms of:

- the prior in latent space $p(\mathbf{x})$
- the smooth mapping $\mathbf{f} : \mathcal{X} \to \mathcal{T}$ from latent space to data space
- the noise model in data space $p(\mathbf{t}|\mathbf{x})$

all of which are equipped with parameters (as before, we omit them for clarity). As mentioned in section 2.4, analytically tractable models can be obtained by clever combinations of (mixtures of) normal distributions with linear mappings or mixtures of Dirac deltas with nonlinear mappings.

We describe here several well-known specific latent variable models. Table 2.2 gives a summary of them. In all cases the parameters of the model may be estimated using the EM algorithm (in the case of PCA, an analytical solution is known as well). In all cases $\mathcal{T} \equiv \mathbb{R}^D$ and $\mathcal{X} \equiv \mathbb{R}^L$ except for GTM, which uses a discrete prior latent space.

Latent variable models can be classified as *linear* and *nonlinear* according to the corresponding character of the mapping $\mathbf{f}$. We call a latent variable model *normal* when both the prior in latent space and the noise model are normal. Thus, factor analysis and principal component analysis (defined below) are *linear-normal* latent variable models.

2.6.1 Factor analysis (FA)

Factor analysis[18] (Bartholomew, 1987; Everitt, 1984) uses a Gaussian distributed prior and noise model, and a linear mapping from data space to latent space. Specifically:

- The latent space prior $p(\mathbf{x})$ is unit normal:

$$\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.13}$$

  although there exist varieties of factor analysis where these factors are correlated. The latent variables $\mathbf{x}$ are often referred to as the *factors*.

- The mapping $\mathbf{f}$ is linear:

$$\mathbf{f}(\mathbf{x}) \overset{\text{def}}{=} \mathbf{\Lambda x} + \boldsymbol{\mu}. \tag{2.14}$$

  The columns of the $D \times L$ matrix $\mathbf{\Lambda}$ are referred to as the *factor loadings*. We assume $\operatorname{rank}(\mathbf{\Lambda}) = L$, i.e., linearly independent factors.

- The data space noise model is normal centred at $\mathbf{f}(\mathbf{x})$ with diagonal covariance matrix $\mathbf{\Psi}$:

$$\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Psi}). \tag{2.15}$$

  The D diagonal elements of $\mathbf{\Psi}$ are referred to as the *uniquenesses*.

The marginal distribution in data space can be computed analytically and it turns out to be normal with a constrained covariance matrix (theorems 2.12.1 and A.3.1(iv)):

$$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda\Lambda}^T + \mathbf{\Psi}). \tag{2.16}$$

The posterior in latent space is also normal:

$$\mathbf{x}|\mathbf{t} \sim \mathcal{N}\left(\mathbf{A}(\mathbf{t} - \boldsymbol{\mu}), (\mathbf{I} + \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\right) \tag{2.17}$$

$$\mathbf{A} = \mathbf{\Lambda}^T(\mathbf{\Lambda\Lambda}^T + \mathbf{\Psi})^{-1} = (\mathbf{I} + \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}. \tag{2.18}$$

The reduced-dimension representative (defined in section 2.9.1) is taken as the posterior mean (coinciding with the mode) and is usually referred to as the *Thomson scores*:

$$\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \mathbf{A}(\mathbf{t} - \boldsymbol{\mu}). \tag{2.19}$$

The dimensionality reduction mapping $\mathbf{F}$ is linear and therefore smooth. As discussed in section 2.9.1.1, factor analysis with Thomson scores does not satisfy the condition that $\mathbf{F} \circ \mathbf{f}$ be the identity, because $\mathbf{A\Lambda} \neq \mathbf{I}$, except in the zero-noise limit.

If we apply an invertible linear transformation $\mathbf{g}$ with matrix $\mathbf{R}$ to the factors $\mathbf{x}$ to obtain a new set of factors $\mathbf{y} = \mathbf{Rx}$, the prior distribution $p(\mathbf{y})$ is still normal (theorem A.3.1(i)), $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{RR}^T)$, and the new mapping becomes $\mathbf{t} = \mathbf{f}'(\mathbf{y}) = \mathbf{f}(\mathbf{g}^{-1}(\mathbf{y})) = \mathbf{\Lambda R}^{-1}\mathbf{y} + \boldsymbol{\mu}$ (section 2.3.2). That is, the new factor loadings become $\mathbf{\Lambda}' = \mathbf{\Lambda R}^{-1}$. If $\mathbf{R}$ is an orthogonal matrix, i.e., $\mathbf{R}^{-1} = \mathbf{R}^T$, the new factors $\mathbf{y}$ will still be independent and $\mathbf{\Psi}' = \mathbf{\Psi}$ diagonal; this is called an *orthogonal rotation* of the factors in the literature of factor analysis. If $\mathbf{R}$ is an arbitrary nonsingular matrix, the new factors $\mathbf{y}$ will not be independent anymore; this is called an *oblique rotation* of the factors. Thus, from all the factor loadings matrices $\mathbf{\Lambda}$, we are free to choose that which is easiest to interpret according to some criterion, e.g. by varimax rotation[19]. However, we insist that, provided that the model $p(\mathbf{t})$ remains the same, all transformations—orthogonal or oblique—are equally valid. Section 2.8.1 further discusses this issue.

The log-likelihood of the parameters[20] $\mathbf{\Theta} = \{\mathbf{\Lambda}, \mathbf{\Psi}, \boldsymbol{\mu}\}$ is obtained as the log-likelihood of a normal distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ of covariance $\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}^T + \mathbf{\Psi}$:

$$\mathcal{L}(\mathbf{\Lambda}, \mathbf{\Psi}) = -\frac{N}{2}\left(D \ln 2\pi + \ln|\mathbf{\Sigma}| + \operatorname{tr}\left(\mathbf{S\Sigma}^{-1}\right)\right) \tag{2.20}$$

[18] Our Matlab implementation of factor analysis (EM algorithm, Rao's algorithm, scores, χ^2-test, etc.) and varimax rotation is freely available in the Internet (see appendix C).

[19] Varimax rotation (Kaiser, 1958) finds an orthogonal rotation of the factors such that, for each new factor, the loadings are either very large or very small (in absolute value). The resulting rotated matrix $\mathbf{\Lambda}'$ has many values clamped to (almost) 0, that is, each factor involves only a few of the original variables. This simplifies factor interpretation.

[20] The maximum likelihood estimate of the location parameter $\boldsymbol{\mu}$ is the sample mean $\bar{\mathbf{t}}$. If the covariance matrix of $p(\mathbf{t})$ in eq. (2.16) was unconstrained, the problem would be that of fitting a normal distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ to the sample $\{\mathbf{t}_n\}_{n=1}^N$. In this case, the maximum likelihood estimates for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ would be the natural ones: the sample mean $\bar{\mathbf{t}}$ and the sample covariance matrix $\mathbf{S}$, respectively (Mardia et al., 1979).

where $\mathbf{S} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \overline{\mathbf{t}})(\mathbf{t}_n - \overline{\mathbf{t}})^T$ is the sample covariance matrix and $\overline{\mathbf{t}} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbf{t}_n$ the sample mean. The log-likelihood gradient is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{\Lambda}} = -N(\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{S}\mathbf{\Sigma}^{-1})\mathbf{\Lambda}) \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{\Psi}} = -\frac{N}{2}\operatorname{diag}\left(\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{S}\mathbf{\Sigma}^{-1})\right).$$

The log-likelihood has infinite equivalent maxima resulting from orthogonal rotation of the factors. Apart from these, it is not clear whether the log-likelihood has a unique global maximum or there exist suboptimal ones. In theory, different local maxima are possible and should be due to an underconstrained model (e.g. a small sample), but there does not seem to be any evidence in the factor analysis literature about the frequency of multiple local maxima with actual data. Rubin and Thayer (1982), Bentler and Tanaka (1983) and Rubin and Thayer (1983) give an interesting discussion about this[21].

The parameters of a factor analysis model may be estimated using an EM algorithm (Rubin and Thayer, 1982):

E step: This requires computing the moments:

$$\begin{aligned} \mathrm{E}\left\{\mathbf{x}|\mathbf{t}_n\right\} &= \mathbf{A}^{(\tau)}(\mathbf{t}_n - \boldsymbol{\mu}) \\ \mathrm{E}\left\{\mathbf{x}\mathbf{x}^T|\mathbf{t}_n\right\} &= \mathbf{I} - \mathbf{A}^{(\tau)}\mathbf{\Lambda}^{(\tau)} + \mathbf{A}^{(\tau)}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^T(\mathbf{A}^{(\tau)})^T \end{aligned}$$

for each data point $\mathbf{t}_n$ given the current parameter values $\mathbf{\Lambda}^{(\tau)}$ and $\mathbf{\Psi}^{(\tau)}$.

M step: This results in the following update equations for the factor loadings $\mathbf{\Lambda}$ and uniquenesses $\mathbf{\Psi}$:

$$\begin{aligned} \mathbf{\Lambda}^{(\tau+1)} &= \left(\sum_{n=1}^{N}\mathbf{t}_n\,\mathrm{E}\left\{\mathbf{x}|\mathbf{t}_n\right\}^T\right)\left(\sum_{n=1}^{N}\mathrm{E}\left\{\mathbf{x}\mathbf{x}^T|\mathbf{t}_n\right\}^T\right)^{-1} \\ \mathbf{\Psi}^{(\tau+1)} &= \frac{1}{N}\operatorname{diag}\left(\sum_{n=1}^{N}\mathbf{t}_n\mathbf{t}_n^T - \mathbf{\Lambda}^{(\tau+1)}\,\mathrm{E}\left\{\mathbf{x}|\mathbf{t}_n\right\}\mathbf{t}_n^T\right) \end{aligned}$$

where the updated moments are used and the "diag" operator sets all the off-diagonal elements of a matrix to zero.

The location parameter $\boldsymbol{\mu}$ is estimated by the sample mean, and does not need to take part in the EM algorithm.

Apart from EM, there are a number of other methods for maximum likelihood parameter estimation for factor analysis, such as the methods of Jöreskog (1967)[22] or Rao (Morrison, 1990, pp. 357–362). Also, in common with other probabilistic models, factor analysis may be implemented using an autoencoder network, with weights implementing a recognition and a generative model: a special case of the Helmholtz machine having two layers of linear units with Gaussian noise, and trained with the wake-sleep algorithm[23] (Neal and Dayan, 1997).

There are also estimation criteria for factor analysis other than maximum likelihood, e.g. principal factors (Harman, 1967) or minimum trace factor analysis (Jamshidian and Bentler, 1998).

2.6.2 Principal component analysis (PCA)

Principal component analysis[24] (PCA) can be seen as a maximum likelihood factor analysis in which the uniquenesses are constrained to be equal, that is, $\mathbf{\Psi} = \sigma^2\mathbf{I}$ is isotropic. This simple fact, already reported in the early factor analysis literature, seems to have gone unnoticed until Tipping and Bishop (1999b) and Roweis (1998) recently rediscovered it. Indeed, a number of textbooks and papers (e.g. Krzanowski, 1988,

[21] Rubin and Thayer (1982), reanalysing an example with 9 observed variables, 4 factors and around 36 free parameters in which Jöreskog had found a single maximum with the LISREL method, claimed to find several additional maxima of the log-likelihood using their EM algorithm. By carefully checking the gradient and the Hessian of the log-likelihood at those points, Bentler and Tanaka (1983) showed that none except Jöreskog's were really maxima. This is disquieting in view of the large number of parameters used in pattern recognition applications.

[22] The method of Jöreskog (1967) is a second-order Fletcher-Powell method. It is also applicable to *confirmatory factor analysis*, where some of the loadings have been set to fixed values (usually zero) according to the judgement of the user (Jöreskog, 1969).

[23] Strictly speaking, Neal and Dayan (1997) give no proof that wake-sleep learning works for factor analysis, only empirical support that it usually does in some simulations.

[24] Our Matlab implementation of principal component analysis is freely available in the Internet (see appendix C).

p. 502 or Hinton et al., 1997, beginning of section III) wrongly quote that "PCA does not propose a model for the data" as a disadvantage when compared with factor analysis.

The approach of considering an isotropic noise model in factor analysis had already been adopted in the Young-Whittle factor analysis model (Young, 1940; Whittle, 1952) and its maximum likelihood solution assuming σ known found analytically (Anderson, 1963; Basilevsky, 1994, pp. 361–363). Lawley (1953) and Anderson and Rubin (1956) showed that stationary points of the log-likelihood as a function of the loadings $\mathbf{\Lambda}$ and uniqueness σ^2 occur at the value of eq. (2.28), although they did not prove it to be global maximum, which Tipping and Bishop (1999b) did. In addition to this direct solution, Roweis (1998) and Tipping and Bishop (1999b) give an EM algorithm for estimating $\mathbf{\Lambda}$ and σ^2.

Consider then the factor analysis model of the previous section with an isotropic error model. There exists a unique (although possibly degenerate, if some eigenvalues are equal) maximum likelihood estimate closely related to the L principal components of the data[25]. If the sample covariance matrix is decomposed as $\mathbf{S} = \mathbf{U}\mathbf{V}\mathbf{U}^T$, with $\mathbf{V} = \operatorname{diag}(v_1, \dots, v_D)$ containing the eigenvalues (ordered decreasingly) and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_D)$ the associated eigenvectors, then $\mathbf{\Lambda} = \mathbf{U}_L(\mathbf{V}_L - \sigma^2\mathbf{I})^{1/2}$ with $\mathbf{U}_L = (\mathbf{u}_1, \dots, \mathbf{u}_L)$, $\mathbf{V}_L = \operatorname{diag}(v_1, \dots, v_L)$ and $\sigma^2 = \frac{1}{D-L}\sum_{j=L+1}^{D} v_j$. Therefore eqs. (2.13)–(2.19) become:

$$\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.21}$$

$$\mathbf{f}(\mathbf{x}) \overset{\text{def}}{=} \mathbf{\Lambda}\mathbf{x} + \boldsymbol{\mu} \tag{2.22}$$

$$\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{f}(\mathbf{x}), \sigma^2\mathbf{I}) \tag{2.23}$$

$$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda}\mathbf{\Lambda}^T + \sigma^2\mathbf{I}) \tag{2.24}$$

$$\mathbf{x}|\mathbf{t} \sim \mathcal{N}\left(\mathbf{A}(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{V}_L^{-1}\right) \tag{2.25}$$

$$\mathbf{A} = \mathbf{V}_L^{-1}(\mathbf{V}_L - \sigma^2\mathbf{I})^{1/2}\mathbf{U}_L^T \tag{2.26}$$

$$\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \mathbf{A}(\mathbf{t} - \boldsymbol{\mu}) \tag{2.27}$$

with maximum likelihood estimates:

$$\mathbf{\Lambda} = \mathbf{U}_L(\mathbf{V}_L - \sigma^2\mathbf{I})^{1/2} \qquad \sigma^2 = \frac{1}{D-L}\sum_{j=L+1}^{D} v_j. \tag{2.28}$$

The Thomson scores give $\mathbf{F}(\mathbf{t}) = \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \mathbf{A}(\mathbf{t} - \boldsymbol{\mu})$, which still does not satisfy the condition that $\mathbf{F} \circ \mathbf{f}$ be the identity, except in the zero-noise limit ($\sigma \to 0$), as discussed in section 2.9.1.1.

Thus, in the latent variable model scenario PCA does define a probability model. Still, PCA is not usually constructed as a probability model but as a reconstruction and dimensionality reduction technique, since its most attractive property is that it is the linear mapping that minimises the least squares reconstruction error of a sample. We describe this aspect of PCA in section 4.5.

Since the noise model variance σ^2 is the same for all observed variables, the directions of the columns of $\mathbf{\Lambda}$ will be more influenced by the variables that have higher noise—unlike with factor analysis, which should be able to separate the linear correlations from the noise. The PCA model is then more restricted than the factor analysis one. However, in many practical situations the directions $\mathbf{\Lambda}$ found by factor analysis and PCA do not differ much.

The log-likelihood of the estimate (2.28) is[26]

$$\mathcal{L}(\mathbf{\Lambda}, \sigma^2) = -\frac{N}{2}\left(D(1 + \ln 2\pi) + \ln|\mathbf{S}| + (D - L)\ln\frac{a}{g}\right) \tag{2.29}$$

where a and g are the arithmetic and geometric means of the $D - L$ smallest eigenvalues of the sample covariance matrix $\mathbf{S}$. Although there is an EM algorithm that finds the L principal components by maximising the log-likelihood (Tipping and Bishop, 1999b; Roweis, 1998), the fastest way to perform PCA is via numerical singular value decomposition (Golub and van Loan, 1996; Press et al., 1992).

Again, as in factor analysis, it is possible to orthogonally rotate the latent variables keeping the same distribution.

[25] To differentiate it from conventional PCA, where no probabilistic model is explicitly defined and only a linear orthogonal projection is considered (section 4.5), this latent variable model is called *probabilistic PCA* by Tipping and Bishop (1999b) and *sensible PCA* by Roweis (1998). We do not deem necessary to use and additional, different name for what is essentially the same thing, so in this thesis we will only use the name PCA and the context will make clear what aspect we mean—probability model or linear mapping.

[26] As with factor analysis, the maximum likelihood estimate of the location parameter $\boldsymbol{\mu}$ is the sample mean $\bar{\mathbf{t}}$.

For a fixed data set, PCA has the property of *additivity*, insofar that the principal components obtained using a latent space of dimension L are exactly the same as the ones obtained using a latent space of dimension $L-1$ plus a new, additional principal component. However, this additivity property does not necessarily hold for either PCA followed by (varimax) rotation or for factor analysis. That is, the factors found by a factor analysis of order L are, in general, all different from those found by a factor analysis of order $L-1$. That means that one can only talk about the joint collection of L factors (or the linear subspace spanned by them).

2.6.3 Independent component analysis (ICA)

Independent component analysis (ICA) or blind source separation consists of recovering independent sources given only sensor observations that are unknown linear mixtures of the sources (Comon, 1994; Cardoso, 1998; Hyvärinen, 1999b; Hyvärinen et al., 2001). Its basic formulation is as follows. Denote the time variable by τ (continuous or discrete). Call $\mathbf{x}(\tau) \in \mathbb{R}^L$ the L time-varying *source signals*[27], assumed independent and zero-mean. Call $\mathbf{t}(\tau) \in \mathbb{R}^D$ the D *data signals*, measured without noise and with instantaneous mixing[28] (that is, there is no time delay between source l mixing into channel d). Then $\mathbf{t}(\tau) \overset{\text{def}}{=} \mathbf{\Lambda x}(\tau)$, where $\mathbf{\Lambda}_{D\times L}$ is the *mixing matrix* and $D \geq L$ (although we will assume $D = L$ in most of this section[29]). The goal of ICA is, given a sample $\{\mathbf{t}_n\}_{n=1}^N$ of the sensor outputs, to find a linear transformation $\mathbf{A}_{L\times D}$ (*separating matrix*) of the sensor signals $\mathbf{t}$ that makes the outputs $\mathbf{u}(\tau) = \mathbf{At}(\tau) = \mathbf{A\Lambda x}(\tau)$ as independent as possible. If $\mathbf{\Lambda}$ was known, $\mathbf{A} = \mathbf{\Lambda}^+ = (\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T$ (assuming rank $(\mathbf{\Lambda}) = L$) would recover the sources exactly, but $\mathbf{\Lambda}$ is unknown; all we have is a finite set of realisations of the sensor outputs $\{\mathbf{t}_n\}_{n=1}^N$.

If we take the sources $\mathbf{x} = (x_1, \dots, x_L)^T$ as the latent variables and the sensor outputs $\mathbf{t} = (t_1, \dots, t_D)^T$ as the observed variables, ICA can be seen as a latent variable model with the following elements (the dependence on the time τ is omitted for clarity):

- The prior distribution in latent space $\mathcal{X} = \mathbb{R}^L$ is factorised but unknown: $p(\mathbf{x}) = \prod_{l=1}^L p_l(x_l)$.
- The mapping between latent and observed space is linear: $\mathbf{t} = \mathbf{\Lambda x}$.
- The noise model is a Dirac delta, i.e., the observed variables are generated without noise: $p(\mathbf{t}|\mathbf{x}) = \delta(\mathbf{t} - \mathbf{\Lambda x}) = \prod_{d=1}^D \delta(t_d - \sum_{l=1}^L a_{dl}x_l)$. However, we can imagine that noise is an independent source and segregate it from the other sources.

Therefore, the only free parameter in the latent variable model is the matrix $\mathbf{\Lambda}_{D\times L} = (\lambda_{dl})$ (or $\mathbf{A}_{L\times D} = \mathbf{\Lambda}^+ = (a_{ld})$), but it is also necessary to determine the unknown functions $\{p_l(x_l)\}_{l=1}^L$. How to deal with this unknown prior distribution is what makes ICA different from the other latent variable models discussed here. Ideally one would learn the functions $\{p_l\}_{l=1}^L$ nonparametrically, but this is difficult. For practical estimation, all current methods use a fixed functional form for the latent space prior distribution with or without parameters. A flexible parametric model of the prior distribution is a Gaussian mixture and this gives rise to the IFA model of section 2.6.4.

The very effective **infomax** learning rule of Bell and Sejnowski (1995) is the following online iterative algorithm:

$$\mathbf{A}^{(\tau+1)} = \mathbf{A}^{(\tau)} + \Delta\mathbf{A}^{(\tau+1)} \tag{2.30}$$

$$\Delta\mathbf{A}^{(\tau+1)} \propto ((\mathbf{A}^{(\tau)})^T)^{-1} + \mathbf{g}(\mathbf{u})\mathbf{t}^T \tag{2.31}$$

where $\mathbf{t}$ is a sensor output, $\mathbf{u} \overset{\text{def}}{=} \mathbf{A}^{(\tau)}\mathbf{t}$, $\mathbf{g}(\mathbf{u}) \overset{\text{def}}{=} (g(u_1), \dots, g(u_L))^T$ and $g(u) \overset{\text{def}}{=} \frac{\partial \ln |f'(u)|}{\partial u}$. The choice of the nonlinear function $f : \mathbb{R} \to \mathbb{R}$ is critical since it can be proven (see below and Bell and Sejnowski, 1995) that $p(x_l) \propto |f'(x_l)|$. That is, f is the one-dimensional cumulative distribution function in latent space and fixing its form is equivalent to assuming a prior distribution $p(x_l)$ for the latent variables (the sources). Taking $g(u) = -2u$ (or equivalently $f = \text{erf}$) means assuming normally distributed sources, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (although with a Dirac-delta noise model), which gives second-order decorrelation only—in which case using factor analysis would have been enough. The use of a nonlinear function, whose Taylor series expansion has terms of many orders, may be more sensitive to the right higher-order statistics of the inputs. In practice, a sigmoidal function is chosen, such as the logistic function. Several forms of the nonlinearity f are summarised in table 2.3. The computationally simpler nonlinearities, such as the logistic function or the hyperbolic tangent, assume

[27] Instead of time-varying sources $\mathbf{x}(\tau)$ one can consider space-varying sources $\mathbf{x}(x, y)$, such as images.

[28] If time delays are considered, the problem is called *blind deconvolution* rather than *blind separation*.

[29] The case $D < L$ (more sources than sensors) is called *overcomplete representation* and has been advocated because of its greater robustness to noise, sparsity and flexibility in matching structure in the data (Amari, 1999; Lewicki and Sejnowski, 2000).

Nonlinearity name	$f(u)$	$f'(u)$	$g(u) \stackrel{\text{def}}{=} \frac{\partial}{\partial u}\ln \lvert f'(u)\rvert = \frac{f''(u)}{f'(u)}$	Kurtosis k_4
Logistic	$\frac{1}{1+e^{-u}}$	$f(1-f)$	$1-2f$	$+$
Generalised logistic		$f^p(1-f)^r$	$\left(\frac{p}{f}-\frac{r}{1-f}\right) f^p(1-f)^r$	$\begin{cases} +, \ p,r>1 \\ -, \ p,r<1 \end{cases}$
Hyperbolic tangent	$\tanh u$	$1-f^2$	$-2f$	$+$
Generalised tanh		$1-\lvert f\rvert^r$	$-r\lvert f\rvert^{r-1}\operatorname{sgn}(f)$	$\begin{cases} +, \ r<2 \\ -, \ r>2 \end{cases}$
		$e^{-u^2/2}\cosh u$	$-u+\tanh u$	$-$
Error function	$\operatorname{erf} u$	$\frac{2}{\sqrt{\pi}}e^{-u^2}$	$-2u$	0
Generalised erf		$e^{-\lvert u\rvert^r}$	$-r\lvert u\rvert^{r-1}\operatorname{sgn}(u)$	$\begin{cases} +, \ r<2 \\ -, \ r>2 \end{cases}$
Arctangent	$\arctan u$	$\frac{1}{1+u^2}$	$-\frac{2u}{1+u^2}$	$+\infty$
cdf of Student's t		$\frac{\Gamma(\frac{1}{2}(r+1))}{\sqrt{\pi r}\Gamma(\frac{1}{2}r)}\left(1+\frac{u^2}{r}\right)^{-\frac{r+1}{2}}$	$-\frac{(r+1)u}{u^2+r}$	$\frac{6}{r-4}$

Table 2.3: Different nonlinearities f and their associated slopes f' and functions g for the infomax rule (2.31). The kurtosis is that of the associated p.d.f. in latent space, $p(u) \propto |f'(u)|$. If $f(u)$ is omitted, it is defined as $\int_{-\infty}^{u} f'(v)\,dv$. $f = \operatorname{erf}$ gives the normal distribution and $f = \arctan$ the Cauchy one. In all cases $p, r > 0$.

supergaussian (positive kurtosis) source distributions and thus are not suitable for subgaussian ones. A better nonlinearity in those cases is the one given by $g(u) = -u + \tanh u$. However, Bell and Sejnowski claim that most real-world analog signals are supergaussian. Subgaussian source separation is a topic of current research (e.g. Lee et al., 1999).

The rule (2.31) is nonlocal, requiring a matrix inversion, and thus does not seem biologically plausible. Being a standard gradient descent, it is also noncovariant: different variables have different units. However, rescaling the gradient by a metric matrix $\mathbf{A}^T\mathbf{A}$, we obtain a **natural gradient** algorithm (Yang and Amari, 1997; Amari and Cichoki, 1998):

$$\Delta\mathbf{A} \propto (\mathbf{I} + \mathbf{g}(\mathbf{u})\mathbf{u}^T)\mathbf{A} \tag{2.32}$$

which is simpler and converges faster. Batch algorithms are also much faster, such as FastICA (Hyvärinen, 1999a), which is based on a fixed-point iteration.

The rule (2.31), or approximate versions of it, can be justified from several points of view:

Information maximisation In the low noise case, the *infomax principle* states that if the mutual information between the inputs $\mathbf{t}$ and the outputs $\mathbf{y}$ of a processor is maximum, then the output distribution $p(\mathbf{y})$ is factorial (Linsker, 1989; Nadal and Parga, 1994). Consider a single-layer feedforward neural network with inputs $\mathbf{t} = (t_1, \dots, t_D)^T$, net inputs $\mathbf{u} \stackrel{\text{def}}{=} \mathbf{At}$ and outputs $y_l \stackrel{\text{def}}{=} f(u_l)$, $l = 1, \dots, L$, where $\mathbf{A}_{L\times D}$ is a matrix of adjustable parameters and f a fixed invertible (thus monotonic) and differentiable nonlinearity. Maximising the mutual information $I(\mathbf{y}, \mathbf{t})$ by varying the parameters $\mathbf{A}$ is equivalent to maximising the output entropy $h(\mathbf{y})$:

$$I(\mathbf{y}, \mathbf{t}) \stackrel{\text{def}}{=} h(\mathbf{y}) - h(\mathbf{y}|\mathbf{t}) \Rightarrow \frac{\partial}{\partial\mathbf{A}} I(\mathbf{y}, \mathbf{t}) = \frac{\partial}{\partial\mathbf{A}} h(\mathbf{y})$$

because $h(\mathbf{y}|\mathbf{t})$ does not depend on $\mathbf{A}$. Taking into account that $p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{t}}(\mathbf{t})\, |J(\mathbf{t}, \mathbf{y})|^{-1}$ (where $J(\mathbf{t}, \mathbf{y}) = \left(\frac{\partial y_l}{\partial t_d}\right)$ is the Jacobian of the transformation), $p(u_l) = p(y_l)\, |f'(u_l)|$ and $h(\mathbf{y}) \stackrel{\text{def}}{=} -\operatorname{E}_{p_{\mathbf{y}}}\{\ln p_{\mathbf{y}}\}$, we obtain the rule (2.31) with $g(u) \stackrel{\text{def}}{=} \frac{\partial \ln|f'(u)|}{\partial u}$.

The maximum entropy is obtained when each y_l is distributed uniformly (assuming $\{y_l\}_{l=1}^{L}$ are amplitude-bounded random variables), in which case $p(u_l) = p(y_l)\, |f'(u_l)| \propto |f'(u_l)|$. That is, f is the c.d.f. of u_l.

Finally, observe that the output $\mathbf{y}$ of our single-layer neural network is used only for minimising the mutual information. We are really interested in $\mathbf{u}$, which are the recovered sources.

Maximum likelihood estimation MacKay (1996) shows for the case $L = D$ that the infomax algorithm can be derived by maximising the log-likelihood of the ICA latent variable model by gradient ascent.

Let us compute the probability distribution induced in the observed space for the particular case $L = D$ by marginalising the joint distribution as in eq. (2.3):

$$p(\mathbf{t}) = \int_{\mathcal{X}} p(\mathbf{t}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} = \int_{\mathcal{X}} \delta(\mathbf{t} - \mathbf{\Lambda}\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} = |\mathbf{\Lambda}|^{-1}\, p(\mathbf{\Lambda}^{-1}\mathbf{t}) = |\mathbf{\Lambda}|^{-1} \prod_{l=1}^{L} p_l \left(\sum_{d=1}^{D} \lambda_{ld}^{-1} t_d \right).$$

So we get:

$$\begin{aligned} \text{as a function of } \mathbf{\Lambda}\text{: } \ln p(\mathbf{t}|\mathbf{\Lambda}) &= -\ln|\mathbf{\Lambda}| + \sum_{l=1}^{L} \ln p_l \left(\sum_{d=1}^{D} \lambda_{ld}^{-1} t_d \right) \\ \text{as a function of } \mathbf{A} = \mathbf{\Lambda}^{-1}\text{: } \ln p(\mathbf{t}|\mathbf{A}) &= \ln|\mathbf{A}| + \sum_{l=1}^{L} \ln p_l \left(\sum_{d=1}^{D} a_{ld} t_d \right). \end{aligned}$$

Calling $\mathbf{u} \overset{\text{def}}{=} \mathbf{A}\mathbf{t}$ and $\mathbf{g}(\mathbf{u}) \overset{\text{def}}{=} \left(\frac{d \ln p_1(u_1)}{du_1}, \dots, \frac{d \ln p_L(u_L)}{du_L} \right)^T$ and using the matrix differentiation identities (A.4) we obtain the gradient of the log-likelihood:

$$\begin{aligned} \frac{\partial}{\partial \lambda_{dl}} \ln p(\mathbf{t}|\mathbf{\Lambda}) &= -a_{ld} - u_l \sum_{l'=1}^{L} a_{l'd}\, g_{l'}(u_{l'}) \\ \frac{\partial}{\partial a_{ld}} \ln p(\mathbf{t}|\mathbf{\Lambda}) &= \lambda_{dl} + g_l(u_l) t_d \end{aligned}$$

which coincides with the infomax one, eq. (2.31).

An alternative derivation from a maximum likelihood estimation point of view has been given by Pearlmutter and Parra (1996) and Cardoso (1997). Consider a parametric model $p_{\mathbf{\Theta}}(\mathbf{t})$ for the true sensor distribution $p_{\mathbf{t}}(\mathbf{t})$. Eq. (2.11) shows that maximising the log-likelihood of $\mathbf{\Theta}$ is equivalent to minimising the Kullback-Leibler distance to the data density $p_{\mathbf{t}}(\mathbf{t})$. Assuming $\mathbf{A}$ and $\mathbf{\Lambda}$ invertible, and (a) since $H(p_{\mathbf{t}})$ is independent of $\mathbf{A}$ and (b) the Kullback-Leibler divergence is invariant under invertible transformations ($\mathbf{t} = \mathbf{\Lambda}\mathbf{x}$ and $\mathbf{u} = \mathbf{A}\mathbf{t}$), maximum likelihood estimation produces:

$$\Delta \mathbf{A} \propto \frac{\partial \mathcal{L}(\mathbf{\Theta})}{\partial \mathbf{A}} \overset{\text{(a)}}{=} -\frac{\partial}{\partial \mathbf{A}} \mathrm{D}\left(p_{\mathbf{t}}(\mathbf{t}) \| p_{\mathbf{\Theta}}(\mathbf{t}) \right) \overset{\text{(b)}}{=} -\frac{\partial}{\partial \mathbf{A}} \mathrm{D}\left(p(\mathbf{x}) \| p_{\mathbf{\Theta}}(\mathbf{u}) \right)$$

which coincides with the infomax rule of eq. (2.31).

Negentropy maximisation and projection pursuit From a projection pursuit point of view (section 4.6), ICA looks for linear projections where the data become independent. Girolami et al. (1998) show that negentropy maximisation projection pursuit performs ICA on sub- or supergaussian sources and leads exactly to the infomax rule (2.31). Negentropy of a distribution of density p is defined as the Kullback-Leibler divergence between p and the Gaussian distribution $p_{\mathcal{N}}$ with the same mean and covariance as p: $\mathrm{D}\left(p \| p_{\mathcal{N}}\right)$. It is zero if p is normal and positive otherwise.

Cumulant expansion If $\mathbf{u}$ and $\mathbf{x}$ are symmetrically distributed and approximately normal, their mutual information can be approximated using a Gram-Charlier or Edgeworth polynomial expansion (Kendall and Stuart, 1977) of fourth order in terms of the cumulants (Comon, 1994) and used as objective function for gradient ascent. However, this method requires more computations than the infomax algorithm and gives worse separation—an order higher than 4 is necessary to improve separation.

Nonlinear PCA Karhunen and Joutsensalo (1994) show that a neural network trained by least squares to compute the function $\mathbf{A}f(\mathbf{A}\mathbf{t})$ on prewhitened data can separate signals. Girolami and Fyfe (1999) show that its cost function is approximately equivalent to that of the cumulants method.

2.6.4 Independent factor analysis (IFA)

Independent factor analysis (Attias, 1998, 1999) uses a factorial Gaussian mixture prior, a linear mapping from data space to latent space and a normal noise model[30]. Specifically:

[30] Similar approaches have also been proposed that use mixture models for the source density: Pearlmutter and Parra (1996) (mixture of logistic densities, gradient descent) and Moulines et al. (1997) and Xu et al. (1998) (Gaussian mixture, EM algorithm).

- Each latent variable is modelled independently of the others as a mixture of M_l one-dimensional Gaussians:

$$p(\mathbf{x}) \overset{\text{def}}{=} \prod_{l=1}^{L} p_l(x_l) \qquad p_l(x_l) \overset{\text{def}}{=} \sum_{m_l=1}^{M_l} p(m_l)p(x_l|m_l) \qquad x_l|m_l \overset{\text{def}}{\sim} \mathcal{N}(\mu_{l,m_l}, v_{l,m_l}). \tag{2.33}$$

  Thus the latent space prior $p(\mathbf{x})$ is a product of one-dimensional Gaussian mixtures, which results in a mixture of L-dimensional diagonal Gaussians but with constrained means and covariance matrices (since a general mixture of diagonal Gaussians need not be factorised):

$$p(\mathbf{x}) = \sum_{\mathbf{m}} p(\mathbf{m})p(\mathbf{x}|\mathbf{m}) \begin{cases} \mathbf{m} \overset{\text{def}}{=} (m_1, \dots, m_L) \\ p(\mathbf{m}) = \prod_{l=1}^{L} p(m_l) \\ \mathbf{x}|\mathbf{m} \overset{\text{def}}{\sim} \mathcal{N}(\boldsymbol{\mu}_{\mathbf{m}}, \mathbf{V}_{\mathbf{m}}) \text{ with } \begin{cases} \boldsymbol{\mu}_{\mathbf{m}} \overset{\text{def}}{=} (\mu_{1,m_1}, \dots, \mu_{L,m_L})^T \\ \mathbf{V}_{\mathbf{m}} \overset{\text{def}}{=} \operatorname{diag}(v_{1,m_1}, \dots, v_{L,m_L}). \end{cases} \end{cases} \tag{2.34}$$

  The summation over $\mathbf{m}$ includes all combinations of tuples $\mathbf{m} = (m_1, \dots, m_L)$ where $m_l = 1, \dots, M_l$ for each $l = 1, \dots, L$.

- The mapping $\mathbf{f}$ is linear (the data space distribution is assumed zero-mean, which can be obtained by centring the data sample before fitting the model):

$$\mathbf{f}(\mathbf{x}) \overset{\text{def}}{=} \boldsymbol{\Lambda}\mathbf{x}. \tag{2.35}$$

  We assume $\operatorname{rank}(\boldsymbol{\Lambda}) = L$.

- The data space noise model is normal centred at $\mathbf{f}(\mathbf{x})$ with covariance matrix $\boldsymbol{\Phi}$:

$$\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{f}(\mathbf{x}), \boldsymbol{\Phi}). \tag{2.36}$$

Thus, factor analysis (and PCA) is a particular case of IFA where $\boldsymbol{\Phi}$ is diagonal and each $p_l(x_l)$ is $\mathcal{N}(0,1)$.

The marginal distribution in data space can be computed analytically and turns out to be a constrained mixture of Gaussians (eq. (2.6) and theorems 2.12.1 and A.3.1(iv)):

$$p(\mathbf{t}) = \sum_{\mathbf{m}} p(\mathbf{m})p(\mathbf{t}|\mathbf{m}) \qquad \mathbf{t}|\mathbf{m} \overset{\text{def}}{\sim} \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{\mu}_{\mathbf{m}}, \boldsymbol{\Lambda}\mathbf{V}_{\mathbf{m}}\boldsymbol{\Lambda}^T + \boldsymbol{\Phi}). \tag{2.37}$$

The posterior in latent space is also a mixture of Gaussians:

$$p(\mathbf{x}|\mathbf{t}) = \sum_{\mathbf{m}} p(\mathbf{m}|\mathbf{t})p(\mathbf{x}|\mathbf{m},\mathbf{t}) \begin{cases} p(\mathbf{m}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{m})p(\mathbf{m})}{\sum_{\mathbf{m}'} p(\mathbf{t}|\mathbf{m}')p(\mathbf{m}')} \\ (\mathbf{x}|\mathbf{m},\mathbf{t}) \overset{\text{def}}{\sim} \mathcal{N}(\nu_{\mathbf{m},\mathbf{t}}, \boldsymbol{\Sigma}_{\mathbf{m}}) \begin{cases} \boldsymbol{\Sigma}_{\mathbf{m}} \overset{\text{def}}{=} (\boldsymbol{\Lambda}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\Lambda} + \mathbf{V}_{\mathbf{m}}^{-1})^{-1} \\ \nu_{\mathbf{m},\mathbf{t}} \overset{\text{def}}{=} \boldsymbol{\Sigma}_{\mathbf{m}}(\boldsymbol{\Lambda}^T\boldsymbol{\Phi}^{-1}\mathbf{t} + \mathbf{V}_{\mathbf{m}}^{-1}\boldsymbol{\mu}_{\mathbf{m}}). \end{cases} \end{cases} \tag{2.38}$$

The reduced-dimension representative can be taken as the posterior mean, which is simple to compute (and it need be computed in each iteration of the EM algorithm):

$$\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \sum_{\mathbf{m}} p(\mathbf{m}|\mathbf{t})\nu_{\mathbf{m},\mathbf{t}}. \tag{2.39}$$

Since the posterior (2.38) can be asymmetric or multimodal, one could take its mode instead (which could be computed iteratively with the algorithms of chapter 8). In the limit of zero noise, the reduced-dimension representative can be taken as the projection via the pseudoinverse $\boldsymbol{\Lambda}^+ = (\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T$, as in ICA.

Attias (1999) gives an EM algorithm for IFA. Taking the limit $\boldsymbol{\Phi} \to \mathbf{0}$ in it results in an EM algorithm for PCA, as presented by Tipping and Bishop (1999b) and Roweis (1998). Attias (1999) also gives an EM and a generalised EM algorithm for noiseless IFA ($\boldsymbol{\Phi} = \mathbf{0}$). The latter results in a combination of a rule similar to the infomax rule (2.31) for updating the mixing matrix and update rules for the Gaussian mixture parameters similar to those of the standard EM algorithm for a Gaussian mixture. That is, it combines separating the sources $x_1, \dots, x_L$ with learning their densities.

Again, a major disadvantage of the IFA model is that the number of parameters grows exponentially with the dimensionality of the latent space: the prior distribution for the latent variables includes $\prod_{l=1}^{L}(2M_l - 1)$ parameters, which is $\mathcal{O}(e^L)$. Attias (1999) proposes a variational approximation of the EM algorithm that reduces the number of operations—but the number of parameters remains $\mathcal{O}(e^L)$.

2.6.5 The generative topographic mapping (GTM)

The generative topographic mapping (GTM) (Bishop, Svensén, and Williams, 1998b) is a nonlinear latent variable model which has been proposed as a principled alternative to self-organising feature maps (Kohonen, 1995). Specifically:

- The L-dimensional latent space $\mathcal{X}$ is discrete[31]. The prior in latent space, $p(\mathbf{x})$, is discrete uniform, assigning nonzero probability only to the points $\{\mathbf{x}_k\}_{k=1}^K \subset \mathbb{R}^L$, usually arranged in a regular grid (for visualisation purposes):

$$p(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{x}_k). \tag{2.40}$$

  This discrete prior can be seen as a fixed approximation of a continuous, uniform distribution in a hyperrectagle of $\mathbb{R}^L$ (see section 2.4 on Monte Carlo sampling).

- The mapping $\mathbf{f}$ is a generalised linear model:

$$\mathbf{f}(\mathbf{x}) \overset{\text{def}}{=} \mathbf{W}\boldsymbol{\phi}(\mathbf{x}), \tag{2.41}$$

  where $\mathbf{W}$ is a $D \times F$ matrix and $\boldsymbol{\phi}$ an $F \times 1$ vector of fixed basis functions.

- The noise model $p(\mathbf{t}|\mathbf{x})$ is an isotropic normal centred at $\mathbf{f}(\mathbf{x})$:

$$\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{f}(\mathbf{x}), \sigma^2 \mathbf{I}). \tag{2.42}$$

The marginal distribution in data space is a constrained mixture of Gaussians (in the sense that the Gaussian centres cannot move independently, but only by changing the mapping $\mathbf{f}$ through $\mathbf{W}$):

$$p(\mathbf{t}) = \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{t}|\mathbf{x}_k), \tag{2.43}$$

and the posterior in latent space is discrete:

$$p(\mathbf{x}_k|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}_k)}{\sum_{i=1}^{K} p(\mathbf{t}|\mathbf{x}_i)}. \tag{2.44}$$

The reduced-dimension representative can be taken as the posterior mean or the posterior mode, both of which can be easily computed since the posterior distribution is discrete. The mean and the mode can be quite different from each other if the posterior distribution is multimodal. The dimensionality reduction mapping $\mathbf{F}$ for the posterior mode is not continuous in general, although it will be approximately continuous if the posterior distribution (2.44) is unimodal and sharply peaked for most points in data space (section 2.9.2).

The log-likelihood of the parameters $\boldsymbol{\Theta} = \{\mathbf{W}, \sigma^2\}$ is

$$\mathcal{L}(\mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \ln \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{t}_n|\mathbf{x}_k; \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \ln \frac{1}{K(2\pi\sigma^2)^{\frac{D}{2}}} \sum_{k=1}^{K} e^{-\frac{1}{2\sigma^2}\left\|\mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}(\mathbf{x}_k)\right\|^2}$$

and is known to contain a number of suboptimal maxima (see fig. 2.13 for some examples).

The parameters of a GTM model may be estimated using the EM algorithm:

E step This requires computing the *responsibility* $R_{nk} = p(\mathbf{x}_k|\mathbf{t}_n)$ of each latent space point $\mathbf{x}_k$ having generated point $\mathbf{t}_n$ using eqs. (2.40)-(2.44) with the current parameter values $\mathbf{W}^{(\tau)}$ and $\sigma^{(\tau)}$.

M step This results in the following update equations for the parameters $\mathbf{W}$ and σ, respectively:

$$\boldsymbol{\Phi}^T \mathbf{G}^{(\tau)} \boldsymbol{\Phi} (\mathbf{W}^{(\tau+1)})^T = \boldsymbol{\Phi}^T (\mathbf{R}^{(\tau)})^T \mathbf{T} \tag{2.45a}$$

$$(\sigma^{(\tau+1)})^2 = \frac{1}{ND} \sum_{k=1}^{K} \sum_{n=1}^{N} R_{nk}^{(\tau)} \left\|\mathbf{t}_n - \mathbf{f}(\mathbf{x}_k)\right\|^2 \tag{2.45b}$$

where $\boldsymbol{\Phi} \overset{\text{def}}{=} (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K)^T$, $\mathbf{T} \overset{\text{def}}{=} (\mathbf{t}_1, \dots, \mathbf{t}_N)^T$, $\mathbf{R}$ is an $N \times K$ matrix with elements R_{nk} and $\mathbf{G}$ is a $K \times K$ diagonal matrix with elements $g_{kk} = \sum_{n=1}^{N} R_{nk}$. Solving for $\mathbf{W}$ requires (pseudo-)inverting the matrix $\boldsymbol{\Phi}^T \mathbf{G} \boldsymbol{\Phi}$ at each iteration.

[31] Strictly, the latent space is continuous and $p(\mathbf{x})$ is a density, but we will call it "discrete uniform" for short (see also sections 2.4 and 2.9.2).

A simple regularised version of GTM to control the mapping $\mathbf{f}$ is obtained by placing an isotropic Gaussian prior distribution of variance λ^{-1} on the mapping weights $\mathbf{W} = (w_{df})$, $\mathbf{W}|\lambda \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. This leads to a maximum a posteriori (MAP) estimator:

$$\mathcal{L}_{\text{MAP}}(\mathbf{W}, \sigma^2) = \mathcal{L}(\mathbf{W}, \sigma^2) - \frac{N\lambda}{2} \left\|\mathbf{W}\right\|^2 \tag{2.46}$$

where λ is the regularisation coefficient, $\left\|\mathbf{W}\right\|^2 = \sum_{d=1}^{D} \sum_{f=1}^{F} w_{df}^2$ and an additive term independent of $\mathbf{W}$ and σ^2 has been omitted. This results in a modified M step of the EM algorithm:

$$(\mathbf{\Phi}^T \mathbf{G}^{(\tau)} \mathbf{\Phi} + \lambda\sigma^2 \mathbf{I})(\mathbf{W}^{(\tau+1)})^T = \mathbf{\Phi}^T (\mathbf{R}^{(\tau)})^T \mathbf{T}.$$

An approximate Bayesian treatment of the hyperparameter λ is given by Bishop et al. (1998a) and Utsugi (2000).

The major shortcoming of GTM is that, being based on (fixed) Monte Carlo sampling of the latent space mentioned in section 2.4, both the number of latent grid points K and the number of basis functions F required for the generalised linear model (2.41) grow exponentially with the dimension of the latent space, L. This limits the practical applicability of GTM to about 2 latent variables. Another shortcoming, shared by most global methods (i.e., that find an estimate for the whole data space) that are very flexible parameterised models, is the fact that the EM estimate very often converges to very bad local maxima, as those shown in the lower half of fig. 2.13; for twisted manifolds, a good maximum can only be reached it starting EM from a small region of the parameter space.

The **density networks** of MacKay (1995a) are a model similar to GTM: the mapping $\mathbf{f}$ is implemented by a multilayer perceptron, all the distributions are assumed isotropic Gaussian and a Monte Carlo fixed sampling is necessary to obtain $p(\mathbf{t})$. The log-likelihood is maximised with a conjugate gradients method. Like GTM, this approach suffers from the exponential complexity of the Monte Carlo sampling. MacKay applies density networks to a discrete problem, modelling a protein family. The insight of (Bishop et al., 1998b) is to implement the mapping $\mathbf{f}$ with a generalised linear model, which results in a tractable M step of the EM algorithm, since only the weights of the linear layer have to be estimated.

GTM pursues a similar goal as Kohonen's self-organising maps (to adapt a topographical arrangement of knots to a data distribution) but with the important advantage that it defines a probabilistic model for the data and dimensionality reduction and reconstruction mappings. Table 4.4 compares succintly both models.

2.6.5.1 Extensions to GTM

Bishop et al. (1998a) propose several extensions to the GTM model:

- A manifold-aligned noise model, where the covariance matrix of the noise model (2.42) is not isotropic anymore but approximately aligned (depending on the value of η below) with the manifold $\mathcal{M} = \mathbf{f}(\mathcal{X})$ locally at each mapped point $\mathbf{f}(\mathbf{x})$:

  $$\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Sigma}(\mathbf{x})) \qquad \mathbf{\Sigma}(\mathbf{x}) \overset{\text{def}}{=} \sigma^2\mathbf{I} + \eta \sum_{l=1}^{L} \frac{\partial \mathbf{f}}{\partial x_l} \left(\frac{\partial \mathbf{f}}{\partial x_l}\right)^T$$

  where the new hyperparameter η must be adjusted manually and $\mathbf{\Sigma}(\mathbf{x})$ is computed only at the points $\{\mathbf{x}_k\}_{k=1}^{K}$. The goal is to ensure that the variance of the noise distribution in directions tangential to the manifold is never significantly less than the square of the typical distance between neighbouring points $\mathbf{f}(\mathbf{x}_k)$, so that there is a smooth distribution along the manifold even when the noise variance perpendicular to the manifold becomes small.

  It would seem that this model violates the axiom of local independence (section 2.3.1), since $\mathbf{\Sigma}(\mathbf{x})$ is not diagonal and so the noise model would not be factorised anymore. However, the model can be derived from taking the model with factorised axis-aligned noise, but approximating the uniform density by a grid of Gaussians (rather than deltas) and then linearising the manifold about the centre of each Gaussian (Chris Williams, pers. comm.).

  It remains to be seen whether this approach has any advantages over using a mixture of linear-normal latent variable models (since each GTM mixture component represents now a local linear subspace).

- Modelling of discrete observed variables by defining a Bernoulli or multinomial noise model depending on $\mathbf{f}(\mathbf{x})$ via a logistic or softmax function, respectively. However, estimation of the parameters via the EM algorithm is much more difficult, requiring nonlinear optimisation in the M step.

- A semilinear model where some of the latent variables are discretised and mapped with the generalised linear model (2.41) as in the standard GTM model and the rest are continuous and mapped linearly.

- An incremental EM algorithm based on the approach of Neal and Hinton (1998).

- Approximate Bayesian inference for the regularisation parameter λ of eq. (2.46), using the Laplace approximation (MacKay, 1992a), i.e., a local Gaussian approximation at the mode of the posterior distribution of the parameters $\mathbf{W}$. The Laplace approximation breaks down when the posterior parameter distribution is multimodal or skewed—a likely situation when training data is scarce (Richardson and Green, 1997). Utsugi (2000) uses a Gibbs sampler and an ensemble learning method to approximate Bayesian inference.

- A Gaussian process formulation.

Further extensions by other authors include:

- Marrs and Webb (1999) propose an average generalised unit-speed constraint on $\mathbf{f}$ to preserve geodetic distances in data space (outlined in section 2.8.3).

- We propose a diagonal noise GTM model (dGTM), where the noise model covariance is diagonal rather than spherical: $\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Psi})$ with $\mathbf{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_D)$. In section 2.12.4 we show that the EM algorithm remains the same except for eq. (2.45b), which becomes D equations:

$$\psi_d^{(\tau+1)} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} R_{nk}^{(\tau)} (t_{nd} - f_d(\mathbf{x}_k))^2 \quad d = 1, \dots, D \tag{2.45b'}$$

  and the responsibilities R_{nk} depend on $\mathbf{\Psi}$ rather than on σ^2.

  A diagonal noise model is necessary to account for different scales and noise levels in the different data variables $t_1, \dots, t_D$ (which in general cannot be overcome by presphering the data, as fig. 2.8 shows).

2.7 Finite mixtures of latent variable models

Finite mixtures (Everitt and Hand, 1981) of latent variable models can be constructed in the usual way as[32]

$$p(\mathbf{t}) \overset{\text{def}}{=} \sum_{m=1}^{M} p(m) p(\mathbf{t}|m) \tag{2.47}$$

where:

- $p(\mathbf{t}|m)$, $m = 1, \dots, M$ are latent variable models based on latent spaces $\mathcal{X}_m$ of dimension L_m (not necessarily equal), i.e.,

$$p(\mathbf{t}|m) = \int_{\mathcal{X}_m} p(\mathbf{t}, \mathbf{x}|m) \, d\mathbf{x} = \int_{\mathcal{X}_m} p(\mathbf{t}|\mathbf{x}, m) p(\mathbf{x}|m) \, d\mathbf{x}$$

  where $p(\mathbf{x}|m)$ is the prior distribution in the latent space of the mth component, $p(\mathbf{t}|\mathbf{x}, m)$ its noise model and $\mathbf{f}_m : \mathcal{X}_m \to \mathcal{T}$ its mapping from latent space into data space.

- $p(m)$ are the mixing proportions.

The joint density is $p(\mathbf{t}, \mathbf{x}, m) = p(\mathbf{t}|\mathbf{x}, m) p(\mathbf{x}|m) p(m)$ and the finite mixture distribution can be expressed as the marginalisation of $p(\mathbf{t}, \mathbf{x}, m)$ over $\mathbf{x}$ and m:

$$p(\mathbf{t}) = \sum_{m=1}^{M} \int_{\mathcal{X}_m} p(\mathbf{t}, \mathbf{x}, m) \, d\mathbf{x} = \sum_{m=1}^{M} p(m) \int_{\mathcal{X}_m} p(\mathbf{t}|\mathbf{x}, m) p(\mathbf{x}|m) \, d\mathbf{x} = \sum_{m=1}^{M} p(m) p(\mathbf{t}|m).$$

The advantage of finite mixtures of latent variable models is that they can place different latent variable models in different regions of data space, where each latent variable model models locally the data. This allows the use of simple local models (e.g. linear-normal, like factor analysis or principal component analysis) that build a complex global model (piecewise linear-normal). In other words, finite mixtures of latent variable models combine clustering with dimensionality reduction.

[32] The IFA model results in a data space distribution (2.37) with the same form as eq. (2.47). But in the IFA model the mixture takes place in the latent space while here it takes place in the data space.

2.7.1 Parameter estimation

Once the model is formulated and the functional forms of each latent variable model are fixed, estimation of the parameters can be done by maximum likelihood. As usual with mixture models, the mixing proportions are taken as parameters $p(m) = \pi_m$ and included in the estimation process; one parameter π_m is not free due to the constraint $\sum_{m=1}^{M} \pi_m = 1$.

Maximum likelihood estimation can be conveniently accomplished with an EM algorithm, where for each data point $\mathbf{t}_n$ the missing information is not only the values of the latent variables $\mathbf{x}_n$ (as in section 2.5), but also the index of the mixture component that generated $\mathbf{t}_n$:

E step This requires computation of the *responsibility* $R_{nm} = p(m|\mathbf{t}_n)$ of each component m having generated point $\mathbf{t}_n$ using Bayes' theorem:

$$R_{nm} = \frac{p(\mathbf{t}_n|m)p(m)}{p(\mathbf{t}_n)} = \frac{\pi_m p(\mathbf{t}_n|m)}{\sum_{m=1}^{M} \pi_m p(\mathbf{t}_n|m)}$$

where $p(\mathbf{t}|m)$ is given by the latent variable model with the parameter values of the current iteration.

M step This results in several update equations for the parameters. The update equations for the mixing proportions are independent of the type of latent variable model used:

$$\pi_m^{(\tau+1)} = \frac{1}{N} \sum_{n=1}^{N} R_{nm}^{(\tau)} \pi_m^{(\tau)}.$$

The equations for the rest of the parameters (from the individual latent variable models) depend on the specific functional form of $p(\mathbf{t}|\mathbf{x}, m)$ and $p(\mathbf{x}|m)$, but often they are averages of the usual statistics weighted by the responsibilities and computed in a specific order.

2.7.2 Examples

Ghahramani and Hinton (1996) construct a **mixture of factor analysers** where each factor analyser, of L_m factors, is characterised by two kinds of parameters: the mean vector $\boldsymbol{\mu}_m \in \mathbb{R}^{L_m}$ and the loadings matrix $\boldsymbol{\Lambda}_m$ (containing L_m loading vectors), in addition to the mixing proportion π_m. All analysers share a common noise model diagonal covariance matrix $\boldsymbol{\Psi}$ for simplicity (although this implies a loss of generality). They give an EM algorithm for estimating the parameters $\{\{\pi_m, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}_{m=1}^{M}, \boldsymbol{\Psi}\}$ by maximum likelihood from a sample. As with many other mixture models, the log-likelihood surface contains singularities (where the covariance matrix of a component model tends to zero and the lilkelihood tends to infinity) to which the EM algorithm can be attracted (see sections 3.2.2 and 5.4.4.1). Hinton et al. (1997) also used a mixture of factor analysers to model handwritten characters, where each factor analyser was implemented via an autoencoder. McLachlan and Peel (2000, chapter 8) give an AECM algorithm[33] for a mixture of factor analysers where each component has a different, diagonal covariance matrix $\boldsymbol{\Psi}_m$. Utsugi and Kumagai (2001) derive Bayesian estimation algorithms using a Gibbs sampler and its deterministic approximation for this same model. Ghahramani and Beal (2000) apply a variational approximation of Bayesian inference to the model of Ghahramani and Hinton (1996) to automatically determine the optimal number of components M and the local dimensionality of each component L_m.

Tipping and Bishop (1999a) define **mixtures of principal component analysers** and give for them an EM algorithm and a simpler and faster generalised EM algorithm. They show good results in image compression and handwritten digit recognition applications, although the reconstruction error attained is in general larger than that attained by a nonprobabilistic mixture of PCAs trained to minimise the reconstruction error (the VQPCA algorithm of Kambhatla and Leen, 1997, discussed in section 4.7). This is reasonable since the probabilistic mixture of PCAs is trained to maximise the log-likelihood rather than minimise the reconstruction error.

A mixture of diagonal Gaussians and a mixture of spherical Gaussians can be seen, as limit cases, as a mixture of factor analysers with zero factors per component model and a mixture of principal component analysers with zero principal components per component model, respectively. Thus, Gaussian mixtures explain the data by assuming that it its exclusively due to noise—without any underlying (linear) structure.

[33] The alternating expectation conditional-maximisation (AECM) algorithm (Meng and van Dyk, 1997) replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximisation steps and allows the specification of the complete data to be different on each such step.

Bishop and Tipping (1998) use hierarchical mixtures of latent variable models with a two-dimensional latent space to visualise the structure of a data set at different levels of detail; for example, a first, top level can show the general cluster structure of the data while a second level can show the internal structure of the clusters, and so on. The tree structure of the hierarchy can be built iteratively by the user in a top-down way.

Section 4.7 mentions other local models for dimensionality reduction not based on mixtures of latent variable models.

2.7.3 Comparison with Gaussian mixtures

In general, mixtures of latent variable models whose distribution in data space $p(\mathbf{t})$ results in a Gaussian mixture (such as mixtures of factor analysers or PCAs) have two advantages over usual mixtures of Gaussian distributions:

- Each component latent variable model locally models both the (linear) mapping and the noise, rather than just the covariance.
- They use fewer parameters per component, e.g. $D(L+1)$ for a factor analyser versus $\frac{D(D+1)}{2}$ for a Gaussian (of course, L should not be too small for the model to remain good).

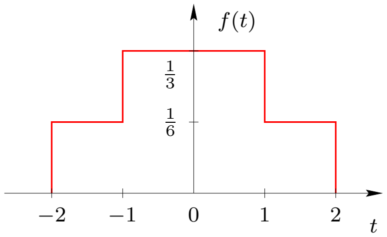
A small number of free parameters requires less computation and less training data (thus reducing the risk of overfitting when data is scarce). Using an unconstrained Gaussian mixture requires more parameters because each mixture component has its own covariance matrix. The total number of parameters can be reduced by one of the following approaches:

- Using diagonal or even spherical components (which still are universal density approximators, as full-covariance mixtures are; Titterington et al., 1985; Scott, 1992).
- Using *parameter tying* or *sharing* (Young, 1996). This method, very popular in the literature of speech recognition based on hidden Markov models, consists of allocating a pool of parameters, typically covariance matrices, to be shared by several or all the components of the mixture. Thus, several components may have the same covariance matrix but different means and mixture proportions. However, the training algorithm—often EM-based—becomes considerably complex and the task of deciding what parameters to tie is not straightforward.

Exactly what approach (full, diagonal or spherical covariance; and unconstrained, latent variable or parameter-tied mixture) is more efficient depends on each particular case; "efficient" here means to use as few parameters as possible to achieve a given performance (such as approximation error, classification error, word error rate, etc.). Saul and Rahim (2000b) give an EM algorithm for hidden Markov models whose output distribution is a mixture of factor analysers and show good performance with a reduced number of parameters when compared with other methods.

2.8 Identifiability, interpretability and visualisation

Identifiability of a class of probabilistic models refers to the existence of a unique characterisation for any model in that class, i.e., to the fact that no two different values of the model parameters give rise to the same distribution (Titterington et al., 1985; Everitt and Hand, 1981; McLachlan and Peel, 2000). For example, the class of finite mixtures of uniform distributions is not identifiable, as a simple counterexample shows. The following three mixtures produce the same distribution:



$$f_1 \overset{\text{def}}{\sim} \frac{1}{3}\mathcal{U}(-1,1) + \frac{2}{3}\mathcal{U}(-2,2) \tag{2.48a}$$

$$f_2 \overset{\text{def}}{\sim} \frac{1}{2}\mathcal{U}(-2,1) + \frac{1}{2}\mathcal{U}(-1,2) \tag{2.48b}$$

$$f_3 \overset{\text{def}}{\sim} \frac{1}{6}\mathcal{U}(-2,1) + \frac{2}{3}\mathcal{U}(-1,1) + \frac{1}{6}\mathcal{U}(1,2) \tag{2.48c}$$

since $f_1(t) = f_2(t) = f_3(t)$ for all $t \in (-\infty, \infty)$.

For a mixture distribution, the parameters include the number of mixture components too. Thus, identifiability is defined formally as follows for mixtures (and, as a particular case, for non-mixture models):

Definition 2.8.1. A class of finite mixtures parameterised by $\mathbf{\Theta}$ is said to be identifiable if for any two members

$$p(\mathbf{t};\mathbf{\Theta}) \stackrel{\text{def}}{=} \sum_{m=1}^{M} \pi_m p(\mathbf{t};\theta_m) \qquad p(\mathbf{t};\mathbf{\Theta}^*) \stackrel{\text{def}}{=} \sum_{m=1}^{M^*} \pi_m^* p(\mathbf{t};\theta_m^*)$$

then $p(t;\mathbf{\Theta}) = p(t;\mathbf{\Theta}^*)$ for all $\mathbf{t} \in \mathcal{T}$ if and only if $M = M^*$ and $\pi_m = \pi_m^*$ and $\theta_m = \theta_m^*$ for $m = 1,\dots,M$ (perhaps with a reordering of the indices). Trivial cases where $\pi_m = 0$ or $\theta_m = \theta_{m'}$ for some m, m' are disregarded, being exactly represented by a mixture with fewer components.

Consider the example of equations (2.48). Given a sample $\{t_n\}_{n=1}^N$ generated from f_1, we cannot tell from which of f_1, f_2 or f_3 it was generated (given the sample alone). Estimating the parameters of a uniform mixture with two components could end up (approximately) in any of f_1 or f_2 or not work at all, depending on the algorithm used, the starting point if it is iterative, etc. The sample could be interpreted as coming from two populations (in the case of f_1 and f_2) or from three (in the case of f_3). Thus, if we want (1) to be able to interpret in a unique way the parameters estimated from a sample, and (2) to avoid that the estimation procedure may break down in case of ill-posedness, then the model class being considered must be identifiable.

However, theoretical identifiability is not the whole story:

- The existence of local, suboptimal maxima of the log-likelihood (or other objective function) makes very difficult to obtain the global maximum—for example, when estimating a Gaussian mixture or a GTM model (fig. 2.13).

- Theoretical identifiability does not guarantee practical identifiability, as discussed in section 3.3.3.2.

Non-identifiability often arises with discrete distributions, for the following reason: if there are C categories we cannot set up more than $C-1$ independent equations (because $\sum p(\mathbf{c}) = 1$) and hence can only determine $C-1$ parameters at most. Fortunately, with continuous distributions—the case that concerns us—it is usually not a problem.

Identifiability is a mathematical property that depends on the choice of model and there are no general necessary and sufficient conditions for all models. The identifiability of the latent variable models of section 2.6, if known, is discussed below, while that of finite mixtures of multivariate Bernoulli distributions (which are used in chapter 5) is discussed separately in section 3.3. A further treatment of the identifiability of mixtures is given by Titterington et al. (1985).

2.8.1 Interpretability

The factor analysis literature, especially regarding social science applications, has debated the issue of the interpretability of factors for decades. The problem arises from the fact that factor analysis is non-identifiable with respect to orthogonal rotations of the factors.

As noticed in section 2.3.2, in an ideal latent variable model we could adjust freely both the latent space prior distribution $p_l(\mathbf{x})$ and the mapping $\mathbf{f}$ via an invertible transformation $\mathbf{g}$ of the latent space, so that we could have infinitely many equivalent combinations $(p_l, \mathbf{f})$, all giving rise to the same data distribution $p(\mathbf{t})$ in eq. (2.3). In other words, we could have infinitely many different coordinate systems, each one with its one interpretation, equally able to explain the data. In more practical latent variable models only some restricted kinds of indeterminacy will exist (such as orthogonal rotations in factor analysis) but the point remains: there is no **empirical** ground to prefer one model over another if both give rise to the same $p(\mathbf{t})$.

The key matter is that the only variables with a real existence are the observed variables and that the latent variables are sheer mathematical constructs subordinated to explaining the observed ones in a compact way (Bartholomew, 1987; Everitt, 1984). Therefore, one should not try to reify the latent variables[34] or look too hard for an interpretation of the estimated parameters if the model under consideration belongs to an non-identifiable class. Also, one should not attribute a causal nature to the latent variables: the generative view of section 2.3 is just a convenient conceptualisation of the probabilistic latent variable framework.

In contrast, let us mention the technique of **principal variables** (McCabe, 1984), related to principal component analysis, whose aim is to select a subset of variables that contain, in some sense, as much information as possible. The principal variables can be readily interpreted because they are observed variables—unlike the principal components, which are linear combinations of the observed variables.

[34] Although in some particular cases this may be possible. For example, Kvalheim (1992) claims that, in chemistry, latent variables are often interpretable and independently verifiable.

However, what can be interpretable is the manifold in data space spanned by the mapping from latent onto data space, $\mathcal{M} = \mathbf{f}(\mathcal{X})$. This manifold is invariant under a coordinate change, while the coordinates themselves (the latent variables) are not. As an example, consider PCA: what reason is there to choose a particular basis of the hyperplane spanned by the principal components? Besides, sample-based estimates of the principal components can vary considerably, in particular when the covariance matrix has several eigenvalues of approximately the same value, since the directions of the associated eigenvectors will be strongly influenced by the noise. But the subspace spanned by those principal components is still unique and thus identifiable and interpretable.

Another issue is that often the model assumptions are wrong for the problem under consideration, due to the nature of the relationship between variables being unknown or to the mathematical impossibility of using an appropriate but overly complex model. Consider, for example, fitting a principal component analysis to the data of figure 2.13. Interpretation of the estimated parameters in such a situation may be completely misleading. Yet another difficulty lies in whether the number of parameters in the model and the way it is estimated from a data set can lead to overfitting.

Fortunately, in machine learning applications the data is usually of a very high dimensionality and the relationships between variables is nonlinear, so that more often than not the user does not have strong preconceptions about what the latent variables should be. Besides, as a consequence of the philosophical approach of statistical machine learning, which pursues to simulate the behaviour of a system, the utility of a model is related to how well it can extract the structure of a high-dimensional data set—irrespective of whether it actually matches the true underlying system (always an idealisation anyway) or just mimicks it. Of course, overfitting remains a difficult problem.

To summarise:

- In the ideal case where we know the model from which the data comes (i.e., we know the number of latent variables and the functional forms for the prior distribution in latent space, the mapping and the noise model) and all is necessary is to fit its parameters, we need to check the issues of identifiability and overfitting.

- If we are guessing the model, apart from the identifiability and overfitting we need to assess the model's goodness, which is very difficult in high-dimensional cases when there are nonlinear relationships. For example, the noise can easily mask nonlinear relationships when using a linear model.

2.8.2 Identifiability of specific latent variable models

2.8.2.1 Factor analysis

From section 2.6.1 we know that:

- If $L > 1$ then an orthogonal rotation of the factors produces the same data distribution, so that factor analysis is non-identifiable with respect to orthogonal rotations. If $L = 1$ then the rotation becomes a sign reversal, which is irrelevant.

- If we consider a generalised view of factor analysis where the factors are distributed normally but not necessarily $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then factor analysis is non-identifiable with respect to any invertible linear transformation of the factors.

In confirmatory factor analysis, where some elements of the loading matrix $\mathbf{\Lambda}$ are fixed, the rotation indeterminacy may disappear. In explanatory factor analysis, where all elements of $\mathbf{\Lambda}$ are free, one usually applies a restriction to make it identifiable. The most typical one is to choose the factors so that the first factor makes a maximum contribution to the variance of the observed variables, the second makes a maximum contribution subject to being uncorrelated with the first one, and so on. This can be shown to be equivalent to choosing $\mathbf{\Lambda}$ such that $\mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda}$ is diagonal and has the effect of imposing $\frac{1}{2}L(L-1)$ constraints, so that the total number of free parameters becomes

$$\underbrace{D}_{\mathbf{\Psi}} + \underbrace{DL}_{\mathbf{\Lambda}} + \underbrace{\frac{1}{2}L(L-1)}_{\text{constraints}}$$

which for consistency must be smaller or equal than the number of elements in the covariance matrix, $\frac{1}{2}D(D+1)$, i.e., $\frac{1}{2}((D-L)^2 - (D+L))$ must be positive, which gives an upper bound for the number of factors to be extracted.

Since scale changes of the data ($\mathbf{t}' = \mathbf{D}\mathbf{t}$ with $\mathbf{D}$ diagonal) result in an irrelevant scale change of the factors and the uniquenesses, the same results are obtained whether using the covariance matrix or the correlation matrix.

2.8.2.2 PCA

PCA as defined in section 2.6.2 only differs from factor analysis in the noise model and is constrained to produce the principal components in decreasing order of their associated variance, i.e., $\mathbf{\Lambda}^T(\sigma^2\mathbf{I})^{-1}\mathbf{\Lambda}$ is diagonal from eq. (2.28). Thus, PCA is always identifiable except when several eigenvalues of the sample covariance matrix are equal, in which case the corresponding eigenvectors (columns of $\mathbf{U}_L$) can be rotated orthogonally inside their own subspace at will.

Unlike with factor analysis, with PCA different results are obtained whether using the covariance matrix or the correlation matrix because there is a single uniqueness parameter shared by all data variables. PCA does depend on the scale, so one has to decide whether to sphere the data or not.

2.8.2.3 ICA

Arbitrary scaling, reordering or in general invertible linear mapping of the latent variables (sources) gives rise to the same observed (sensor) distribution: $(\mathbf{\Lambda}, \mathbf{x})$, $(\mathbf{\Lambda}\mathbf{V}, \mathbf{V}^{-1}\mathbf{x})$, $(\mathbf{\Lambda}\mathbf{P}, \mathbf{P}^T\mathbf{x})$ and $(\mathbf{\Lambda}\mathbf{R}, \mathbf{R}^{-1}\mathbf{x})$ all produce the same distribution $p(\mathbf{t})$ if $\mathbf{V}$ is diagonal nonsingular, $\mathbf{P}$ is a permutation matrix and $\mathbf{R}$ is nonsingular. However, only scaling and reordering are acceptable indeterminacies, because they do not alter the "waveform of the signals" or their statistical properties (independence in particular). Thus, it can be proven that the identifiability of the ICA model (up to reordering and rescaling) is guaranteed if (Tong et al., 1991; Comon, 1994):

- At most one source is distributed normally (since the sum of two normals is itself normal, it would not possible to separate the individual components in a unique way).
- There are fewer sources than sensors: $L \leq D$.
- The mixing matrix is full-rank: $\operatorname{rank}(\mathbf{\Lambda}) = L$.

2.8.2.4 Other latent variable models

To our knowledge, no identifiability results are known for GTM, IFA or mixtures of factor analysers or principal component analysers—but suboptimal local maxima of the log-likelihood do exist for all of them.

2.8.3 Visualisation

A different matter from interpretability is visualisation of data. While all coordinate systems related by an invertible map (reparametrisation) are equivalent for dimensionality reduction, some may be more appropriate than others in that the structure in the data is shown more clearly. Rigid motion transformations (translation, rotation or reflection) have a trivial effect on visualisation, since our visual system can still recognise the structure in the data, but invertible transformations that alter the distances between points (distortions) can both make apparent but also completely mask clusters or other accidents in the data.

It is possible to include constraints in the latent variable model being used to ensure that the interpoint distances in data space are approximately preserved in the latent space. These constraints affect exclusively the mapping $\mathbf{f}$. However, if the prior distribution in latent space is not flexible enough to represent a large class of distributions (and it is not for factor analysis, PCA and GTM, since it is fixed) then by constraining $\mathbf{f}$ we are reducing the class of distributions that the model can approximate.

These distortion constraints are different from the usual smoothness constraints: the latter force the manifold $\mathcal{M} = \mathbf{f}(\mathcal{X})$ traced by the function $\mathbf{f}$ to be smooth but do not care about the coordinate system (the parametrisation); they usually bound (in a soft fashion) the second derivative of $\mathbf{f}$. In contrast, the distortion constraints do not care whether the manifold $\mathcal{M}$ is smooth or rough, but force the distances in data space to be preserved in latent space. How this is done depends on how the distance between two data space points is defined: as geodetic distance (i.e., following the shortest path between the two points along the data manifold, in which case it is appropriate to use a unit-speed parametrisation[35] that bounds the first derivative of $\mathbf{f}$) or as Euclidean distance (i.e., following the straight line segment joining both points freely in the data space, whether it goes out of the data manifold or not). This consideration enters the realm of distance-preserving methods and multidimensional scaling, briefly described in section 4.10.

Such distortion constraints can be incorporated as a regularisation term in the log-likelihood or, equivalently, viewed as a prior distribution on the parameters of $\mathbf{f}$. In any case they lead to the appearance of hyperparameters that control the relative importance of the fitting term (log-likelihood) and the regularisation term (distortion constraint). Determining good values for the hyperparameters is a well-known problem that we will meet several times in this thesis, but on which we shall not dwell.

For example, it has been observed that the estimates found by GTM often give a distorted view of the data due to the latent space stretching like a rubber sheet (Tenenbaum, 1998; Marrs and Webb, 1999). This metric distortion is revealed by high values of an appropriately defined magnification factor. A *magnification factor* $M(\mathbf{x})$ is a scalar function dependent on the latent space point that measures the local distortion at latent space point $\mathbf{x}$ induced by the mapping $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{T}$. Bishop et al. (1997b) defined it in the usual sense of differential geometry as the ratio of hypervolumes in the data space and the latent space, equal to the Jacobian of the mapping $\mathbf{f}$: $M(\mathbf{x};\mathbf{W}) \overset{\text{def}}{=} \frac{dV_{\mathcal{T}}}{dV_{\mathcal{X}}} = \mathbf{J_f} = \sqrt{|\mathbf{K}^T\mathbf{W}^T\mathbf{W}\mathbf{K}|}$ where $(K)_{fl} \overset{\text{def}}{=} \frac{\partial \phi_f}{\partial x_l}$. A value of one for this magnification factor corresponds to no distortion (although Marrs and Webb (1999) have pointed out that it can also be one for some distortions; their slightly different definition of the magnification factor is given below). Marrs and Webb (1999) implement an average generalised unit-speed constraint in GTM's function $\mathbf{f}$ to preserve geodetic distances:

$$\mathrm{E}_{p(\mathbf{t})}\left\{\mathbf{K}^T\mathbf{W}^T\mathbf{W}\mathbf{K}\right\} = \mathbf{I}_L$$

which for a one-dimensional latent space ($L = 1$) reduces to $\mathrm{E}\left\{\left\|\frac{d\mathbf{f}}{dx}\right\|^2\right\} = 1$, i.e., a parametrisation of average unit squared speed. This constraint results in a modified M step of the EM algorithm for GTM, eq. (2.45a). The results can be evaluated by checking that the magnification factor $M(\mathbf{x};\mathbf{W}) \overset{\text{def}}{=} \left\|\mathbf{K}^T\mathbf{W}^T\mathbf{W}\mathbf{K} - \mathbf{I}\right\|^2$ (a zero value of which corresponds to no distortion) is small in all points of the latent space.

2.9 Mapping from data onto latent space

2.9.1 Dimensionality reduction and vector reconstruction

In dimensionality reduction (reviewed in chapter 4) we want, given a data point $\mathbf{t}$, to obtain a representative of it in latent space, $\mathbf{x}^* = \mathbf{F}(\mathbf{t})$, for a certain mapping $\mathbf{F} : \mathcal{T} \longrightarrow \mathcal{X}$. The latent variable modelling framework allows the definition of a natural dimensionality reduction mapping. Once the parameters[36] $\boldsymbol{\Theta}$ are fixed, Bayes' theorem gives the posterior distribution in latent space given a data vector $\mathbf{t}$, i.e., the distribution of the probability that a point $\mathbf{x}$ in latent space was responsible for generating $\mathbf{t}$:

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})} = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{\int_{\mathcal{X}} p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}}. \tag{2.49}$$

The latent variable model gives us all this information about $\mathbf{x}$ for fixed $\mathbf{t}$. Summarising this distribution $\mathbf{x}|\mathbf{t}$ in a single latent space point $\mathbf{x}^*$ results in a **reduced-dimension representative**[37] of $\mathbf{t}$. This defines a corresponding mapping $\mathbf{F}$ from data space onto latent space, so that every data point $\mathbf{t}$ is assigned a representative in latent space, $\mathbf{x}^* = \mathbf{F}(\mathbf{t})$. Thus, it can be considered as an *inverse mapping* of $\mathbf{f}$.

[35] For a curve $\mathbf{f} : x \in \mathbb{R} \rightarrow \mathbb{R}^D$, a unit-speed parametrisation verifies $\left\|\frac{d\mathbf{f}}{dx}\right\|^2 = \sum_{d=1}^D \left(\frac{df_d}{dx}\right)^2 = 1$. Equivalently, the arc length between two points $\mathbf{f}(x_1)$ and $\mathbf{f}(x_2)$ on the curve is

$$\left|\int_{x_1}^{x_2} \sqrt{\sum_{d=1}^D \left(\frac{df_d}{dx}\right)^2}\,dx\right| = |x_2 - x_1|$$

so that this parametrisation is also called *arc-length parametrisation*. It means that a unit step in latent space produces a unit step along the curve in data space.

[36] In what follows, we omit the parameters from the formulae for clarity, i.e., $p(\mathbf{x}|\mathbf{t})$ really means $p(\mathbf{x}|\mathbf{t},\boldsymbol{\Theta})$, and so on.

[37] Admittedly, *reduced-dimension representative* of an observed point $\mathbf{t}$ is a pedantic denomination, but it points to the two things we are interested in: that it (1) represents the observed point (2) in latent space. *Latent space representative* is also acceptable, although less general. Other, more compact terms have been proposed but we find them unsatisfactory. For example, *cause* of an observed point $\mathbf{t}$ may attribute to the reduced-dimension representative more than it really is worth (and ring unwanted bells concerning its interpretation). And the term *scores*, very popular in statistics, sounds too vague.

When the posterior distribution $p(\mathbf{x}|\mathbf{t})$ is unimodal, defining $\mathbf{F}$ as the **posterior mean**:

$$\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \mathrm{E}_{p(\mathbf{x}|\mathbf{t})}\left\{\mathbf{x}\right\}$$

is the obvious choice since it is optimal in the least-squares sense (see section 7.3.3). But if $p(\mathbf{x}|\mathbf{t})$ may be multimodal for some points $\mathbf{t}$ (as happens with IFA and GTM), then the posterior mean (while still being the least-squares optimum and defining a continuous mapping) may be inappropriate, since the mean of a multimodal distribution can be a low-probability point. Defining $\mathbf{F}$ as one of the **posterior modes** (perhaps the global one):

$$\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \arg\max_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\mathbf{t})$$

ensures that the reduced-dimension representative $\mathbf{F}(\mathbf{t})$ is a high-probability point, but then which mode to choose is a problem, and besides $\mathbf{F}$ may become discontinuous. These issues are discussed at length in sections 2.9.2 and 7.3.

If $\mathbf{f}$ is injective (the manifold $\mathcal{M} = \mathbf{f}(\mathcal{X})$ does not self-intersect) then $\mathbf{f}$ is invertible[38] on $\mathcal{M}$; i.e., there exists a function $\mathbf{f}^{-1} : \mathcal{M} \to \mathcal{X}$ such that $\mathbf{f}^{-1}(\mathbf{t}) = \mathbf{x}$ if $\mathbf{t} = \mathbf{f}(\mathbf{x}) \in \mathcal{M}$ for $\mathbf{x} \in \mathcal{X}$. But we are interested in defining a dimensionality reduction mapping $\mathbf{F}$ not only for data points in $\mathcal{M}$, but in the whole data space $\mathcal{T}$, which we can attain thanks to the noise model, that assigns nonzero probability to a neighbourhood of every point in $\mathcal{M}$ (remarkably, this very essence of the probabilistic modelling implies that in general $\mathbf{F}\circ\mathbf{f} \neq$ identity, as discussed below). Since $\dim\mathcal{M} < \dim\mathcal{T}$ (for dimensionality reduction to make sense) this will mean that a whole manifold of dimension $\dim\mathcal{T} - \dim\mathcal{X} = D - L$ will be mapped onto the same latent space point $\mathbf{x}$: $\mathbf{F}^{-1}(\mathbf{x}) \overset{\text{def}}{=} \{\mathbf{t}\in\mathcal{T} : \mathbf{F}(\mathbf{t}) = \mathbf{x}\}$. For example, if $\mathbf{F}$ is linear with matrix $\mathbf{A}$ (assumed full-rank), $\mathbf{F}(\mathbf{t}) = \mathbf{A}(\mathbf{t}-\boldsymbol{\mu})$, then $\mathbf{F}^{-1}(\mathbf{x}) = \{\mathbf{t}\in\mathcal{T} : \mathbf{F}(\mathbf{t}) = \mathbf{A}(\mathbf{t}-\boldsymbol{\mu}) = \mathbf{x}\} = \{\boldsymbol{\mu} + \mathbf{B}\mathbf{x}\} + \ker\mathbf{A}$ where $\mathbf{B}$ is any matrix that satisfies $\mathbf{B}\mathbf{A} = \mathbf{I}$ (such as the pseudoinverse of $\mathbf{A}$, $\mathbf{A}^+ = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$, but there may be others, as section 2.9.1.1 discusses) and $\ker\mathbf{A} \overset{\text{def}}{=} \{\mathbf{t}\in\mathcal{T} : \mathbf{A}\mathbf{t} = \mathbf{0}\}$ is the kernel or null-space of the matrix $\mathbf{A}$. $\mathbf{F}^{-1}(\mathbf{x})$ has dimension $D - L$ since $\dim\ker\mathbf{F} + \dim\operatorname{im}\mathbf{F} = \dim\mathcal{T}$ and $\dim\operatorname{im}\mathbf{F} = \dim\mathcal{X}$.

Applying the mapping $\mathbf{f}$ to the reduced-dimension representative $\mathbf{x} = \mathbf{F}(\mathbf{t})$ we obtain the **reconstructed** data vector $\mathbf{t}^* = \mathbf{f}(\mathbf{x}^*)$. Then, the reconstruction error for that point $\mathbf{t}$ is defined as $d(\mathbf{t},\mathbf{t}^*)$ (or a function of it) for some suitable distance d in data space and the average reconstruction error for the sample is defined as $E_d = \frac{1}{N}\sum_{n=1}^N d(\mathbf{t}_n,\mathbf{t}_n^*)$. For example, taking the square of the Euclidean distance, $d(\mathbf{t},\mathbf{t}^*) = \|\mathbf{t}-\mathbf{t}^*\|_2^2$, results in the usual mean squared error criterion $E_2 = \frac{1}{N}\sum_{n=1}^N \|\mathbf{t}_n - \mathbf{t}_n^*\|_2^2$. It is unknown what the relationship between the maximum likelihood criterion and a distance criterion is. While in general they are different, leading to different estimates of the parameters, in practice maximum likelihood estimation often produces very good estimates in terms of reconstruction error.

A desirable feature of the dimension reduction mapping $\mathbf{F}$ would be to satisfy that $\mathbf{F}\circ\mathbf{f}$ be the identity, so that $\mathbf{F}(\mathbf{f}(\mathbf{x})) = \mathbf{x}$ for any latent space point. That is, that $\mathbf{F}$ be the inverse of $\mathbf{f}$ in the manifold $\mathcal{M} = \mathbf{f}(\mathcal{X}) = \operatorname{im}\mathbf{f}$ (the image, or range, space of $\mathbf{f}$). This implies perfect reconstruction of data points in $\mathcal{M}$, since if $\mathbf{t} = \mathbf{f}(\mathbf{x}) \in \mathcal{M}$, then the reconstructed point of $\mathbf{t}$ is $\mathbf{t}^* = \mathbf{f}(\mathbf{F}(\mathbf{t})) = \mathbf{f}(\mathbf{F}(\mathbf{f}(\mathbf{x}))) = \mathbf{f}(\mathbf{x}) = \mathbf{t}$. In general, this condition is not satisfied (as sections 2.9.1.1–2.9.1.3 show) except in the zero-noise limit. In the latter case, the data space points in $\mathcal{T}\setminus\mathbf{f}(\mathcal{X})$ are unreachable under the model (in the sense that $p(\mathbf{t}) = 0$ for such points) and the mapping $\mathbf{f} : \mathcal{X} \to \mathcal{M}$ is invertible on the image space of $\mathcal{X}$ by $\mathbf{F}$ (assuming that $\mathbf{f}$ is injective, i.e., $\mathcal{M} = \mathbf{f}(\mathcal{X})$ does not self-intersect).

The following sections analyse the dimensionality reduction mappings of each specific latent variable model.

2.9.1.1 Linear-normal models (factor analysis and PCA): scores matrix

Consider a general linear-normal model as in the statement of theorem 2.12.1, i.e., $\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}_L(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ in $\mathbb{R}^L$, $\mathbf{f}(\mathbf{x}) \overset{\text{def}}{=} \boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\mu}_T$ and $\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}_D(\mathbf{f}(\mathbf{x}), \boldsymbol{\Sigma}_T)$ in $\mathbb{R}^D$ (remember that for factor analysis $\boldsymbol{\mu}_X = \mathbf{0}$, $\boldsymbol{\Sigma}_X = \mathbf{I}$ and $\boldsymbol{\Sigma}_T = \boldsymbol{\Psi}$ is diagonal and for PCA $\boldsymbol{\mu}_X = \mathbf{0}$, $\boldsymbol{\Sigma}_X = \mathbf{I}$ and $\boldsymbol{\Sigma}_T = \sigma^2\mathbf{I}$ is isotropic). Consider the dimensionality reduction mapping defined by the posterior mean (identical to the posterior mode), $\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \mathbf{A}(\mathbf{t}-\boldsymbol{\mu}) + \boldsymbol{\mu}_X$. The matrix $\mathbf{A}$ of eq. (2.58) is called the **Thomson scores** in the factor analysis literature (eq. (2.18)). Theorem 2.12.3 shows that the posterior distribution is always narrower than the prior distribution, the narrower the smaller the noise is.

From eq. (2.57), the condition $\mathbf{F}\circ\mathbf{f} \equiv$ identity is equivalent to:

$$\mathbf{F}(\mathbf{f}(\mathbf{x})) = \mathbf{A}(\mathbf{f}(\mathbf{x}) - \boldsymbol{\mu}_T) + \hat{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X = \mathbf{A}\boldsymbol{\Lambda}\mathbf{x} + \hat{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X = \mathbf{x} \qquad \forall\mathbf{x}\in\mathbb{R}^L$$

[38] And even if $\mathbf{f}$ is not injective, local invertibility is assured by the inverse function theorem A.6.1, since $\mathbf{f}$ is smooth.

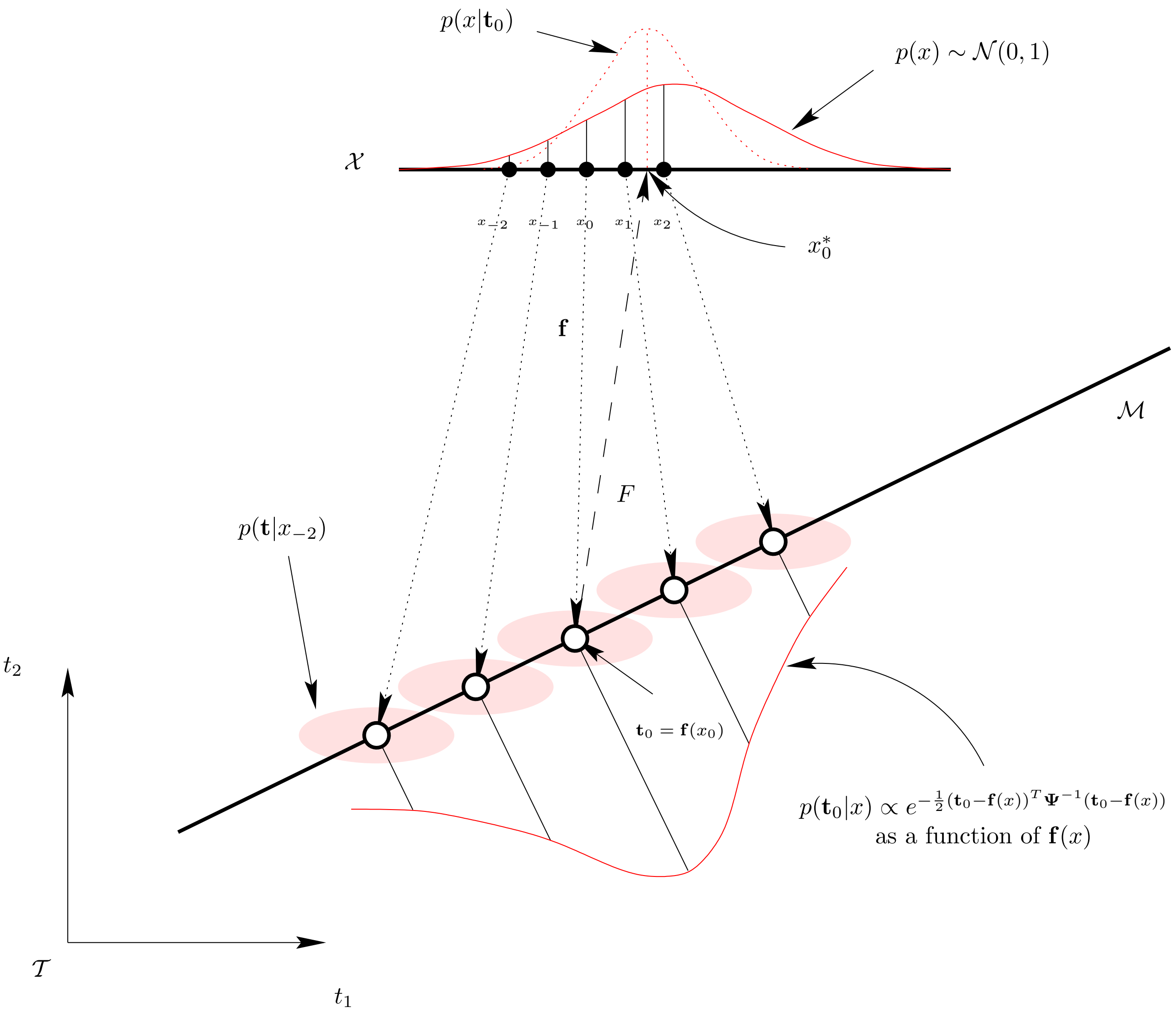

$p(x|\mathbf{t}_0)$
$p(x) \sim \mathcal{N}(0,1)$
$\mathcal{X}$
x_{-2}
x_{-1}
x_0
x_1
x_2
x_0^*
$\mathbf{f}$
$\mathcal{M}$
F
$p(\mathbf{t}|x_{-2})$
t_2
$\mathbf{t}_0 = \mathbf{f}(x_0)$
$p(\mathbf{t}_0|x) \propto e^{-\frac{1}{2}(\mathbf{t}_0-\mathbf{f}(x))^T\mathbf{\Psi}^{-1}(\mathbf{t}_0-\mathbf{f}(x))}$
as a function of $\mathbf{f}(x)$
$\mathcal{T}$
t_1

Figure 2.11: Offset of the reduced-dimension representative when using the posterior mean. This example demonstrates how mapping a latent space point onto data space and back using the posterior mean leads to a reduced-dimension representative which is different from the original latent space point. The latent variable model is a factor analysis with one-dimensional latent space and two-dimensional observed space. The latent space shows its prior distribution $p(x)$ (solid line) centred symmetrically around its mean. A point x_0 is mapped onto data space point $\mathbf{t}_0 = \mathbf{f}(x_0)$. Given the noise model $p(\mathbf{t}|x) \sim \mathcal{N}(\mathbf{f}(x), \mathbf{\Psi})$ alone, other nearby latent points x_{-1}, x_1, etc. could also have generated $\mathbf{t}_0$, but less likely than x_0 (and symmetrically so as we move away from $\mathbf{t}_0$, as marked, because of the symmetry of $p(\mathbf{t}_0|x)$). Each shaded ellipse represents $p(\mathbf{t}_0|x_i)$ for $i \in \{-2, \dots, 2\}$. But given the prior distribution in latent space, latent points to the nearby right side of x_0 are more likely a priori than latent points to the left: $p(x_{-2}) < p(x_{-1}) < p(x_0) < p(x_1) < p(x_2)$. Therefore, the posterior distribution in latent space $p(x|\mathbf{t}_0) \propto p(\mathbf{t}_0|x)p(x)$ (dotted line) is offset towards latent points which are more likely a priori (towards the prior mean in this example) and thus its centre does not match the original latent point x_0. This offset is inherent to the choice of prior distribution in latent space and noise model and is therefore unavoidable.

which implies two conditions:

$$\boldsymbol{\mu}_X = \mathbf{0} \tag{2.50a}$$
$$\mathbf{A}\boldsymbol{\Lambda} = \mathbf{I}. \tag{2.50b}$$

Condition (2.50a) is readily satisfied by both factor analysis and PCA, and we assume it holds from now on. As for condition (2.50b), using eqs. (2.58) and (2.59) it becomes:

$$\mathbf{I} = \mathbf{A}\boldsymbol{\Lambda} \overset{(2.58)}{=} \hat{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda} \overset{(2.59)}{=} (\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda} \Leftrightarrow \boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda} \tag{2.51}$$

which is impossible. Therefore *no linear-normal latent variable model exists that satisfies* $\mathrm{E}\left\{\mathbf{x}|\mathbf{f}(\mathbf{x}_0)\right\} = \mathbf{x}_0$, in particular factor analysis and PCA. Figure 2.11 explains intuitively this.

Let us analyse when condition (2.50b) holds approximately. This will happen when $\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda} \gg \boldsymbol{\Sigma}_X^{-1}$ or equivalently when $\mathbf{M}^T\boldsymbol{\Sigma}_T^{-1}\mathbf{M} \gg \mathbf{I}$ with $\mathbf{M} \overset{\text{def}}{=} \boldsymbol{\Lambda}\boldsymbol{\Sigma}_X^{1/2}$. Since $\mathbf{M}\mathbf{M}^T = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_T$ is the covariance associated with the latent variables and $\boldsymbol{\Sigma}_T$ the covariance associated to the noise, then $\mathbf{F} \circ \mathbf{f} \equiv$ identity will hold approximately when the covariance of the noise is much smaller than the covariance due to the latent variables. We consider two limit cases:

- The **low noise limit**, where

  $$\mathbf{A} = \hat{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1} = (\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1} \approx \hat{\mathbf{A}} \overset{\text{def}}{=} (\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}.$$

  The matrix $\hat{\mathbf{A}}$ is called the **Bartlett scores** in the factor analysis literature. It can also be derived from a distribution-free argument as follows: if the mapping from latent to data space is represented as $\mathbf{t} = \boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\mu}_T + \mathbf{e}$ where $\mathbf{e} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_T)$ and the mapping from latent to data space as $\mathbf{x}^* = \mathbf{B}(\mathbf{t} - \boldsymbol{\mu}) \overset{(2.55)}{=} \mathbf{B}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}_X) + \mathbf{B}\mathbf{e}$, then it can be proven (Bartholomew, 1987, pp. 66–69) that:

  $$\hat{\mathbf{A}} = \arg\min_{\mathbf{B}\boldsymbol{\Lambda}=\mathbf{I}} \mathbf{B}\boldsymbol{\Sigma}_T\mathbf{B}^T.$$

  That is, $\hat{\mathbf{A}}$ is the matrix $\mathbf{B}$ satisfying $\mathbf{B}\boldsymbol{\Lambda} = \mathbf{I}$ that minimises the residual variance of the reduced-dimension representative, $\mathrm{var}\left\{\mathbf{e}\right\} = \mathbf{B}\boldsymbol{\Sigma}_T\mathbf{B}^T$.

  In practice there tends to be little difference between Thomson and Bartlett scores, since $\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is often approximately diagonal and the difference is just a rescaling of the $\mathbf{x}$ variables.

- The **isotropic zero noise limit**, where from corollary 2.12.2 and assuming $\boldsymbol{\Lambda}$ full-rank:

  $$\boldsymbol{\Sigma}_T = k\mathbf{I} \text{ and } k \to 0^+ \Rightarrow \begin{cases} \mathbf{A} \to \boldsymbol{\Lambda}^+ = (\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T \\ \mathbf{x}|\mathbf{t} \to \delta(\boldsymbol{\Lambda}^+(\mathbf{t} - \boldsymbol{\mu}_T)). \end{cases}$$

  The matrix $\boldsymbol{\Lambda}^+$ is the **pseudoinverse** of $\boldsymbol{\Lambda}$. The pseudoinverse is the matrix $\boldsymbol{\Lambda}^*$ that minimises $\left\|\mathbf{I} - \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}\right\|_2^2$ (Boullion and Odell, 1971); thus, it is the matrix that achieves the least squares reconstruction error. The Thomson scores are also optimal in the least squares sense, being the mean of the posterior distribution in latent space. The difference is that the pseudoinverse gives equal importance to each point $\mathbf{x}$ in the latent space, while the Thomson scores weight each value $\mathbf{x}$ according to the normal distribution $\mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\}$. Thus, it "pulls" points towards the mean, since $p(\mathbf{x}|\mathbf{t})$, being normal, decreases when going away from its mean. The disagreement of both scores is then due to the Thomson scores using a joint probability model for $\mathbf{x}$ and $\mathbf{t}$ and the pseudoinverse using no model.

  If the noise tends to zero but not isotropically, then $\mathbf{A} \not\to \boldsymbol{\Lambda}^+$ necessarily; e.g. in fig. 2.12 if $\boldsymbol{\Sigma}_T = k\left(\begin{smallmatrix} S_1 & 0 \\ 0 & S_2 \end{smallmatrix}\right)$ with $S_1 \neq S_2$ and $k \to 0^+$ then $\mathbf{A}$, $\hat{\mathbf{A}} \to \mathbf{A}_0 \overset{\text{def}}{=} \frac{1}{S_2\lambda_1^2 + S_1\lambda_2^2}(S_2\lambda_1 \; S_1\lambda_2)$ which is different from $\boldsymbol{\Lambda}^+ = \frac{1}{\lambda_1^2 + \lambda_2^2}(\lambda_1 \; \lambda_2)$. However, $\mathbf{A}_0$ and $\boldsymbol{\Lambda}^+$ only differ in how they map points $\mathbf{t} \notin \mathcal{M}$ because for $\mathbf{t} \in \mathcal{M}$ they are equivalent (since $\hat{\mathbf{A}}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^+\boldsymbol{\Lambda} = \mathbf{I}$). Since the points $\mathbf{t} \notin \mathcal{M}$ are unreachable in the zero-limit noise, the difference is irrelevant.

$\mathbf{A}$ (Thomson scores) never coincides with $\boldsymbol{\Lambda}^+$ (pseudoinverse) since it does not satisfy any of the Penrose conditions (A.1). $\hat{\mathbf{A}}$ (Bartlett scores) satisfies all of the Penrose conditions except (A.1a) because $\boldsymbol{\Lambda}\hat{\mathbf{A}}$ is not symmetric in general and so it does not coincide with the pseudoinverse either (although it is a left weak generalised inverse of $\boldsymbol{\Lambda}$; Boullion and Odell, 1971); but $\hat{\mathbf{A}}$ does coincide with the pseudoinverse in

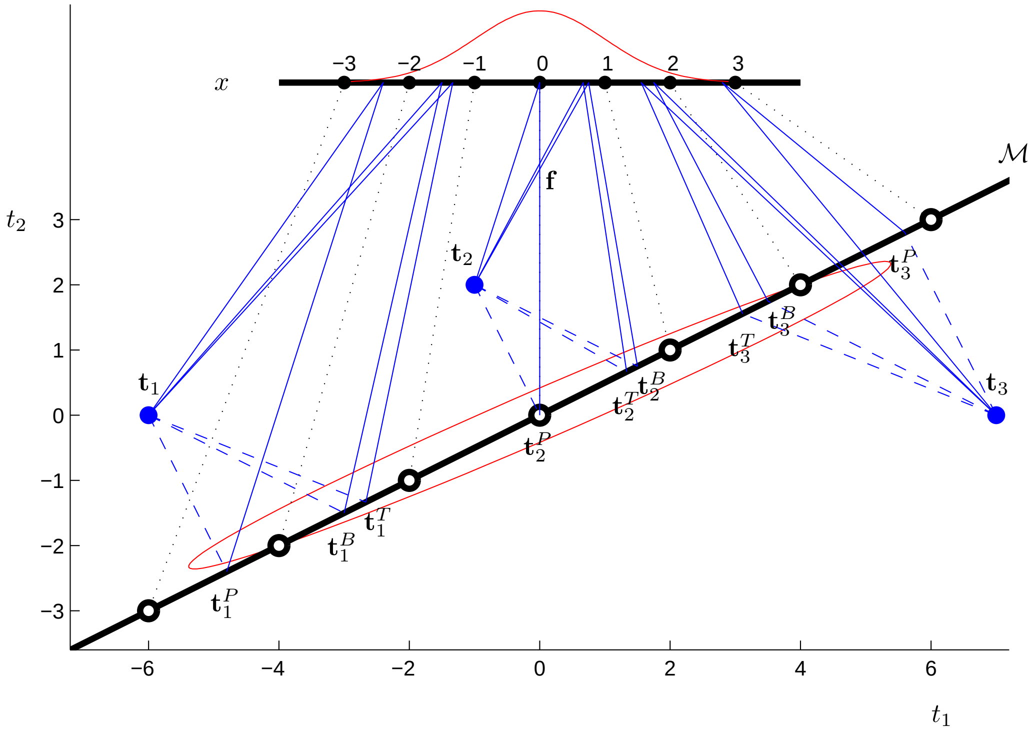


Figure 2.12: Different dimensionality reduction mappings for linear-normal models. Using the symbols of theorem 2.12.1, the numerical values are: $\mu_X = \mathbf{0}$, $\mathbf{\Sigma}_X = \mathbf{I}$, $\mathbf{\Lambda} = \binom{2}{1}$, $\boldsymbol{\mu}_T = \mathbf{0}$, $\mathbf{\Sigma}_T = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$. Therefore $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & \frac{5}{4} \end{pmatrix}$, $\mathbf{A} = \frac{1}{9}(2\ 4)$, $\hat{\mathbf{A}} = \frac{1}{8}(2\ 4)$ and $\mathbf{\Lambda}^+ = \frac{1}{5}(2\ 1)$. The graph shows schematically how the latent space $\mathcal{X} = \mathbb{R}$, with a normal prior distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_X, \mathbf{\Sigma}_X)$ is mapped onto $\mathcal{M}$, inducing a normal distribution $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ (indicated by its unit Mahalanobis ellipse) in $\mathcal{T} = \mathbb{R}^2$. $\mathbf{t}_n^T$, $\mathbf{t}_n^B$ and $\mathbf{t}_n^P$ represent the reconstructed points using $\mathbf{A}$ (Thomson scores), $\hat{\mathbf{A}}$ (Bartlett scores) and $\mathbf{\Lambda}^+$ (pseudoinverse), respectively, for data point $\mathbf{t}_n$ (where $n = 1, 2, 3$). The dimension-reduced representatives are marked on the latent space by the corner of the solid lines; the dashed lines join the original and reconstructed points. The dashed lines for the pseudoinverse are parallel with respect to each other and orthogonal to $\mathcal{M}$: $\mathbf{\Lambda}\mathbf{\Lambda}^+$ is an orthogonal projection. The dashed lines for the Bartlett scores are parallel with respect to each other but not orthogonal to $\mathcal{M}$: $\mathbf{\Lambda}\hat{\mathbf{A}}$ is an oblique projection. The dashed lines for the Thomson scores are not parallel with respect to each other: $\mathbf{\Lambda}\mathbf{A}$ is not a projection.

the particular case where $\mathbf{\Sigma}_T$ is isotropic (precisely the case of PCA). As mentioned before, in the isotropic zero-noise limit all three scores coincide.

Observe that $\mathbf{f} \circ \mathbf{F}$ maps an arbitrary point in data space onto the manifold $\mathcal{M}$. However, of $\mathbf{A}$, $\hat{\mathbf{A}}$ and $\mathbf{\Lambda}^+$ only $\hat{\mathbf{A}}$ and $\mathbf{\Lambda}^+$ give rise to projection matrices: oblique $\mathbf{\Lambda}\hat{\mathbf{A}}$ and orthogonal $\mathbf{\Lambda}\mathbf{\Lambda}^+$, respectively[39].

Table 2.4 summarises the three types of scores and figure 2.12 illustrates the typical situation when all of them differ.

2.9.1.2 Independent component analysis

Given the discussion of section 2.9.1.1, since in the standard ICA model there is no noise and $\mathbf{\Lambda}$ is full rank, the dimensionality reduction mapping to use is $\mathbf{F}(\mathbf{t}) \overset{\text{def}}{=} \mathbf{\Lambda}^+\mathbf{t} = (\mathbf{\Lambda}^T\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\mathbf{t}$ (i.e., the unmixing matrix $\mathbf{A} = \mathbf{\Lambda}^+$ of section 2.6.3), which satisfies $\mathbf{F} \circ \mathbf{f} \equiv$ identity.

The goal of ICA is not really dimensionality reduction but separation of linearly mixed sources, which is attained by the unmixing matrix. Besides, the literature of ICA has mostly been concerned with the case where the number of sensors is equal to the number of sources, although research of more general situations

[39] A matrix $\mathbf{P}$ which verifies symmetry ($\mathbf{P} = \mathbf{P}^T$) and idempotence ($\mathbf{P}^2 = \mathbf{P}$) is an **orthogonal projection** matrix. If it only verifies idempotence but not symmetry then it is an **oblique projection** matrix.

Scores name	Matrix expression	Justification
Thomson	$\mathbf{A} \overset{\text{def}}{=} (\mathbf{\Sigma}_X^{-1} + \mathbf{\Lambda}^T \mathbf{\Sigma}_T^{-1} \mathbf{\Lambda})^{-1} \mathbf{\Lambda}^T \mathbf{\Sigma}_T^{-1}$	Least-squares for linear normal model
Bartlett	$\hat{\mathbf{A}} \overset{\text{def}}{=} (\mathbf{\Lambda}^T \mathbf{\Sigma}_T^{-1} \mathbf{\Lambda})^{-1} \mathbf{\Lambda}^T \mathbf{\Sigma}_T^{-1}$	Low-noise limit of $\mathbf{A}$; also distribution-free
Pseudoinverse	$\mathbf{\Lambda}^{+} = (\mathbf{\Lambda}^T \mathbf{\Lambda})^{-1} \mathbf{\Lambda}^T$	Isotropic zero-noise limit of $\mathbf{A}$; also model-free least-squares

Table 2.4: Dimensionality reduction matrices for linear-normal models with the posterior mean.

is progressing. ICA has been successfully applied to a number of blind separation problems, including elimination of artifacts from electroencephalographic (EEG) data, separation of speech sources (the cocktail party problem), processing of arrays of radar and sonar signals and restoration of images (see references in Hyvärinen and Oja, 2000; Hyvärinen et al., 2001).

2.9.1.3 Independent factor analysis

The posterior distribution in latent space for IFA (2.38) is a convex combination of factor analysis-like posterior probabilities (2.54), so we expect $\mathbf{F} \circ \mathbf{f} \neq$ identity except in the zero-noise limit.

2.9.1.4 GTM

We consider two cases for the definition of $\mathbf{F}$:

- Posterior mean, $\mathbf{F}(\mathbf{t}) = \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\}$. Then:

  $$\mathbf{F}(\mathbf{t}) = \mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \sum_{k=1}^{K} \mathbf{x}_k p(\mathbf{x}_k|\mathbf{t}) = \sum_{k=1}^{K} \mathbf{x}_k \rho_k(\mathbf{t})$$

  where

  $$\rho_k(\mathbf{t}) = p(\mathbf{x}_k|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}_k)p(\mathbf{x}_k)}{p(\mathbf{t})} = \frac{p(\mathbf{t}|\mathbf{x}_k)}{\sum_{k=1}^{K} p(\mathbf{t}|\mathbf{x}_k)} = \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{t}-\mathbf{f}(\mathbf{x}_k)\|^2}}{\sum_{k=1}^{K} e^{-\frac{1}{2\sigma^2}\|\mathbf{t}-\mathbf{f}(\mathbf{x}_k)\|^2}} \in (0,1).$$

  Thus, $\mathbf{F}(\mathbf{t})$ is a point in the convex hull of $\{\mathbf{x}_k\}_{k=1}^{K}$. Observe that for any $k' \in \{1,\dots,K\}$, $\rho_k(\mathbf{f}(\mathbf{x}_{k'})) > 0$ for all $k = 1,\dots,K$. So $\mathbf{F}(\mathbf{f}(\mathbf{x}_{k'})) \neq \mathbf{x}_{k'}$. However, $\mathbf{F}(\mathbf{f}(\mathbf{x}_{k'}))$ will be very close to $\mathbf{x}_{k'}$ when the Gaussian components in data space, centred at $\{\mathbf{f}(\mathbf{x}_k)\}_{k=1}^{K}$, are widely separated with respect to the noise standard deviation σ, since the tails of the Gaussian fall rapidly and then $\rho_k(\mathbf{f}(\mathbf{x}_{k'})) \approx \delta_{k'k}$ (again the zero-noise limit).

- Posterior mode, $\mathbf{F}(\mathbf{t}) = \arg\max_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\mathbf{t})$: it clearly satisfies $\mathbf{F}(\mathbf{f}(\mathbf{x})) = \mathbf{x}$ for $\{\mathbf{x}_k\}_{k=1}^{K}$ but not for any other $\mathbf{x} \in \mathbb{R}^L \setminus \{\mathbf{x}_k\}_{k=1}^{K}$ (which get mapped onto some $\mathbf{x}_k$).

  Also, since $p(\mathbf{t})$ is independent of $\mathbf{x}$ and $p(\mathbf{x}_k) = \frac{1}{K}$, then $\arg\max_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\mathbf{t}) = \arg\max_{k=1,\dots,K} p(\mathbf{t}|\mathbf{x}_k) = \arg\min_{k=1,\dots,K} \|\mathbf{t} - \mathbf{f}(\mathbf{x}_k)\|$, i.e., the latent grid point whose image is closest to $\mathbf{t}$. Thus, the dimensionality reduction mapping behaves like vector quantisation in the observed space based on the Euclidean distance using $\{\mathbf{f}(\mathbf{x}_k)\}_{k=1}^{K}$ as codebook.

2.9.2 Continuity of the dimensionality reduction mapping and regularisation

If the latent variable model has captured the structure of the data appropriately, one would not expect a multimodal distribution $p(\mathbf{x}|\mathbf{t})$, except perhaps if the data manifold is so twisted that it intersects itself or nearly so. Unfortunately, in practice the manifold $\mathcal{M}$ induced in data space can be badly twisted even when the data manifold is relatively smooth, as fig. 2.13 shows, due to a suboptimal local maximum of the log-likelihood. In this case, the induced manifold $\mathcal{M}$ can retrace itself and give rise to multimodal posterior distributions in latent space, since the same area of the data manifold is covered by different areas of the latent space. Regularising the mapping $\mathbf{f}$ (from latent to data space) so that highly twisted mappings are penalised is a possible solution. From a Bayesian point of view this requires introducing a prior distribution on the mapping parameters controlled by a hyperparameter (as in GTM), but determining good hyperparameter values is a hard problem (MacKay, 1999).

The continuity of the dimensionality reduction mapping $\mathbf{x} = \mathbf{F}(\mathbf{t})$ depends on several factors. If we define $\mathbf{F}$ as the mean of the posterior distribution $p(\mathbf{x}|\mathbf{t})$ of eq. (2.49) then $\mathbf{F}$ will be continuous by the first fundamental theorem of calculus (Spivak, 1967, p. 240, Th. 1 and also p. 230, Th. 8), because the mean is defined as an integral; further, $\mathbf{F}$ will be differentiable at every point where $p(\mathbf{x}|\mathbf{t})$ is continuous itself (as a function of $\mathbf{t}$). For example, the dimensionality reduction mappings for factor analysis and PCA are continuous, as is obvious from the linearity of the function (2.19). But if we define $\mathbf{F}$ as the (global) mode of $p(\mathbf{x}|\mathbf{t})$, then it may not be continuous if $\mathbf{F}$ can be multimodal, as fig. 2.14 shows. Section 9.2 shows more examples of the discontinuity of the $\arg\max(\cdot)$ function.

For GTM, $\mathbf{F}(\mathbf{t})$ defined as the posterior mode can be discontinuous, as would happen in the case of figure 2.13 (right). However, its continuity is likely in practical situations where the posterior distribution is unimodal and sharply peaked for the majority of the training set data points.

For latent variable models (like GTM) that sample the latent space, the concept of continuity in the latent space—which is now discretised— becomes blurred if the posterior mode is used (the posterior mean remains continuous, although it will produce points not in the latent space grid). In this case, discontinuities of the dimensionality reduction mapping $\mathbf{F}$ will manifest themselves as abrupt jumps in the grid (larger than the distance between two neighbouring latent grid points) when the point in data space changes slightly.

Finally, let us consider the case of a dimensionality reduction mapping defined as the point in the manifold $\mathcal{M}$ which is closest to the data point $\mathbf{t}$; that is, orthogonal projection ot $\mathbf{t}$ onto $\mathcal{M}$. This is the approach followed by the principal curves method. As mentioned in section 4.8, this definition leads to a discontinuous dimensionality reduction mapping if the manifold $\mathcal{M}$ is nonlinear. But, as discussed earlier, the presence of a noise model and a prior distribution in latent space preclude the use of an orthogonal projection.

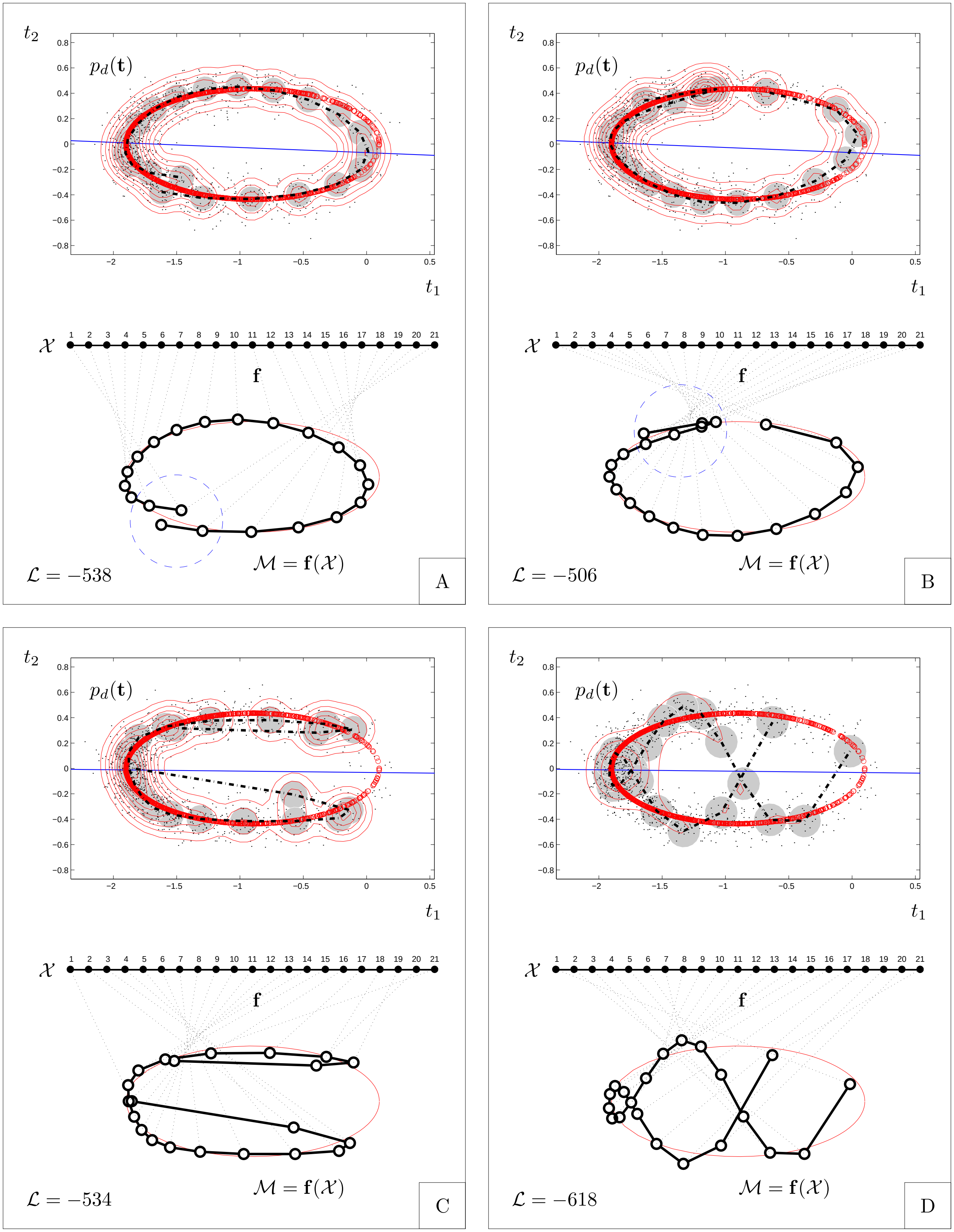
2.9.3 Finite mixtures of latent variable models

We turn now to the subject of dimensionality reduction in finite mixtures of latent variable models. If the dimension of all the latent spaces was the same, one could think of using $\mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\}$ as the latent space representative of data point $\mathbf{t}$:

$$p(\mathbf{x}|\mathbf{t}) = \sum_{m=1}^{M} p(\mathbf{x}, m|\mathbf{t}) = \sum_{m=1}^{M} p(m|\mathbf{t}) p(\mathbf{x}|m, \mathbf{t}) \Longrightarrow$$

$$\mathrm{E}\left\{\mathbf{x}|\mathbf{t}\right\} = \int \mathbf{x} p(\mathbf{x}|\mathbf{t})\, d\mathbf{x} = \int \mathbf{x} \left(\sum_{m=1}^{M} p(m|\mathbf{t}) p(\mathbf{x}|m, \mathbf{t}) \right) d\mathbf{x} = \sum_{m=1}^{M} p(m|\mathbf{t})\, \mathrm{E}\left\{\mathbf{x}|m, \mathbf{t}\right\}$$

Figure 2.13 *(following page)*: Twisted manifolds and continuity of the dimensionality reduction mapping. Each of the plots A to D shows an estimate for a training set of $N = 1\,000$ points (t_1, t_2) generated (with additive normal noise of variance 0.01) from the Keplerian movement distribution of fig. 2.5 (first row), discussed in section 2.2.3: a one-dimensional manifold embedded in the plane. In each plot, an unregularised GTM model was fitted that used $K = 21$ latent grid points and $F = 5$ radial basis functions with a standard deviation $s = 1$ times their separation. Two different training sets were used, one for plots A, B and another one for plots C, D. The EM starting point was random for plots A, C and the line of the first principal component for plots B, D. For each plot, the upper graph shows: the GTM manifold (thick dashed line), the true, elliptical data manifold (small circles, whose density mimicks the distribution along the data manifold), the K Gaussian components associated to the latent grid points (eqs. (2.42) and (2.43), with a radius σ; the big, grey circles), the first principal component of the training set (solid straight line, almost coinciding with the $t_2 = 0$ axis), a contour plot of the GTM data space distribution $p_d(\mathbf{t})$ and the training set (scattered dots). The lower graph shows explicitly the way the one-dimensional latent space is twisted in two dimensions: the latent space grid is the horizontal top line, which is mapped onto the GTM data space manifold $\mathcal{M} = \mathbf{f}(\mathcal{X})$ (thick solid line), to be compared with the true data manifold (thin ellipse). All models have approximately the same log-likelihood $\mathcal{L}$ (except D, which is much worse), but the way the GTM manifold twists itself in data space affects differently the continuity of the dimensionality reduction mapping $\mathbf{F}$ when the posterior mode is used. Model B has discontinuities in the area marked by the dashed circle in the lower graph due to the latent grid points x_{20}–x_{21} retracing over x_{16}–x_{19}. Retracing is worse in model C (points x_{20}–x_{21} retracing over x_{15}–x_{19} and points x_1–x_2 over x_3–x_{12}) and much worse in model D (with multiple branch-switching). A further discontinuity appears in all four models (marked by the dashed circle in the lower graph of plot A) due to the mismatch between the latent space topology (open curve) and the data manifold topology (closed curve).


t_2
$p_d(\mathbf{t})$
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−2
−1.5
−1
−0.5
0
0.5
t_1
$\mathcal{X}$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
$\mathbf{f}$
$\mathcal{M} = \mathbf{f}(\mathcal{X})$
$\mathcal{L} = -538$
A
$\mathcal{L} = -506$
B
$\mathcal{L} = -534$
C
$\mathcal{L} = -618$
D

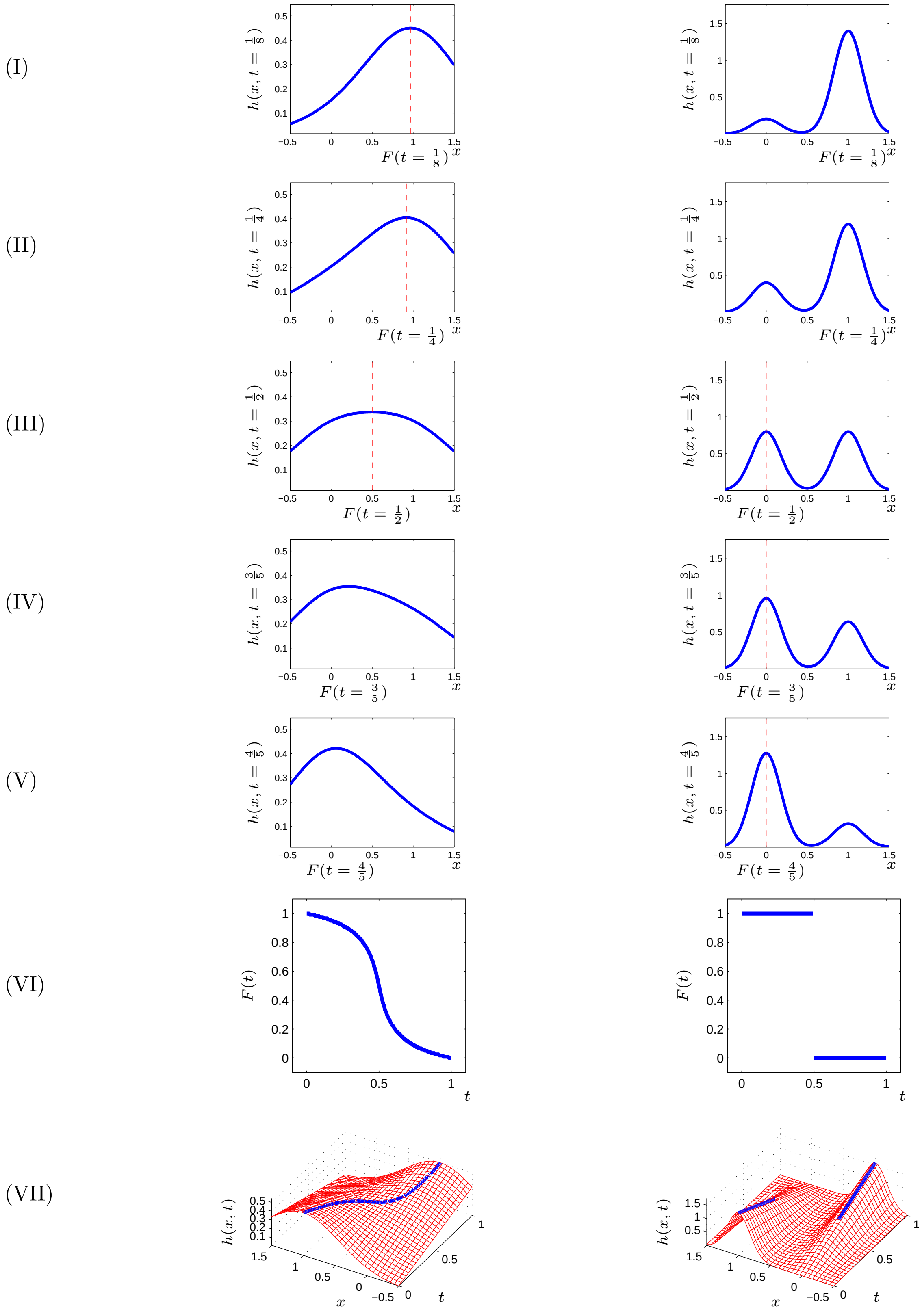


Figure 2.14: Discontinuity of the $\arg\max(\cdot)$ or mode function. The figure shows plots of the function $F(t) = \arg\max_x h(x,t)$ where $h(x,t) \overset{\text{def}}{=} (1-t)\mathcal{N}(x;\mu_1,\sigma^2)+t\mathcal{N}(x;\mu_2,\sigma^2)$ for $t \in [0,1]$ and $x \in (-\infty,\infty)$, with $\mu_1 = 0$ and $\mu_2 = 1$. The left column corresponds to $\sigma = 0.8$, so that the two normal functions overlap heavily and $h(x;t)$ is unimodal for any given $t \in [0,1]$: $F(t)$ is continuous. The right column corresponds to $\sigma = 0.25$, so that the two normal functions are widely separated and $h(x;t)$ is bimodal for (almost) any given $t \in [0,1]$: $F(t)$ is discontinuous. Rows (I)–(V) show, for several values of t, the function $h(x;t)$ and the location of $F(t)$. Row (VI) shows $F(t)$. Row (VII) shows a 3D view of $h(x,t)$ and, superimposed as a thick line, $F(t)$.

where $\mathrm{E}\{\mathbf{x}|m,\mathbf{t}\}$ is the representative in mth latent space proposed by component m, with responsibility $R_m = p(m|\mathbf{t})$ for having generated $\mathbf{t}$.

However, even if all the latent spaces had the same dimension, they will not necessarily correspond to the same coordinate systems (e.g. consider the effect of rotating differently several factor spaces of dimension L). Therefore, when dealing with mixtures of latent variable models, it makes no sense to talk about a single representative, but about M of them, one for each of the M latent spaces, and averaging these representatives is meaningless. Alternatively, a reduced-dimension representative can be obtained as the reduced-dimension representative of the mixture component with the highest responsibility: $\mathbf{x}^* = \mathbf{F}(\mathbf{t}) = \mathbf{x}^*_{m^*}$ such that $m^* = \arg\max_m p(m|\mathbf{t})$.

2.9.3.1 Reconstruction

While averaging the reduced-dimension representatives is not possible, it intuitively makes sense to average the reconstructed vectors (with respect to the responsibilities), because the data space coordinates are the same for any component:

$$\mathbf{t}^* = \sum_{m=1}^{M} p(m|\mathbf{t})\mathbf{f}_m(\mathbf{F}_m(\mathbf{t})) = \sum_{m=1}^{M} p(m|\mathbf{t})\mathbf{t}^*_m \tag{2.52}$$

where $\mathbf{t}^*_m = \mathbf{f}_m(\mathbf{F}_m(\mathbf{t}))$ is the reconstructed vector by the latent variable model of component m. This can be seen as a nonlinear projection of the original vector $\mathbf{t}$ onto the convex hull of $\{\mathbf{t}^*_m\}_{m=1}^M$, which is a subset of the linear subspace spanned by $\{\mathbf{t}^*_m\}_{m=1}^M$. The projection is nonlinear because even if $\mathbf{f}_m \circ \mathbf{F}_m$ is linear for all m, the $p(m|\mathbf{t})$ are not. The vectors $\{\mathbf{t}^*_m\}_{m=1}^M$ are not fixed, but depend on the particular input vector $\mathbf{t}$.

2.9.3.2 Classification

For classification purposes, one could assign the data vector $\mathbf{t}$ to the most responsible component $m^* = \arg\max_m p(m|\mathbf{t})$ and reconstruct the vector according to that component alone (cf. eq. (2.52)):

$$\mathbf{t}^* = \mathbf{t}^*_{m^*}. \tag{2.53}$$

As with latent variable models, the representatives could be computed as some suitable summary value of the posterior distribution $p(\mathbf{x}|m^*,\mathbf{t})$, e.g. the mean, $\mathrm{E}\{\mathbf{x}|m^*,\mathbf{t}\}$, or the mode, $\arg\max_m p(\mathbf{x}|m^*,\mathbf{t})$.

As with general mixture models, if for a given data vector no posterior probability is big enough, no component of the mixture will explain it properly. This may be due to the mixture model being insufficient, to the data vector lying on the boundary of two (or more) components, or to the data vector itself being an outlier (or a novel point). In the latter case the point could be rejected if $p(\mathbf{t}) < \theta$ for a suitable threshold $\theta > 0$ (which needs to be determined heuristically).

2.9.3.3 Reconstruction in general finite mixtures

We consider again equation (2.47), but where now the components $p(\mathbf{t}|m)$ are not latent variable models but certain arbitrary densities or probability mass functions. For the purposes of reconstruction, each component must provide with a reconstructed vector as a function of the original data vector $\mathbf{t}$. For a general probability distribution (without an underlying latent variable model) the best we can do is to choose a value that summarises the distribution, such as the mean, median or mode. We will call this the **prototype** of the distribution, $\mathbf{t}^*_m$. Then, the vector reconstructed by the finite mixture distribution will be $\mathbf{t}^* = \sum_{m=1}^M \pi_m \mathbf{t}^*_m$, as in the finite mixture of latent variable models. Since the $\{\mathbf{t}^*_m\}_{m=1}^M$ are fixed (unlike in reconstruction in mixtures of latent variable models), this amounts to a sort of weighted vector quantisation using the $\{\mathbf{t}^*_m\}_{m=1}^M$ as codebook vectors (although M will usually be small), which gives poor results. Effectively, this means that the finite mixture is performing cluster analysis rather than reconstruction or dimensionality reduction.

For example, for a normal distribution $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, the prototype is the parameter $\boldsymbol{\mu}$, which is the mean, median and mode. For a multivariate Bernoulli distribution $\mathcal{B}(\mathbf{p})$, with $p(\mathbf{t}) = \prod_{d=1}^D p_d^{t_d}(1-p_d)^{1-t_d}$, the prototype can be given by the mean $\mathbf{p}$ (which will not be a binary vector in general) or the mode $\lfloor \mathbf{p} + \frac{1}{2}\mathbf{1} \rfloor$ (which is binary). For binary data, a finite mixture of multivariate Bernoulli distributions (with M components) is called *latent class analysis* (with M classes), as mentioned in section 2.1.

For fixed $\{\mathbf{t}^*_m\}_{m=1}^M$, the set

$$\left\{ \sum_{m=1}^M \pi_m \mathbf{t}^*_m : \sum_{m=1}^M \pi_m = 1, \quad 0 \le \pi_m \le 1, \quad m = 1,\dots,M \right\}$$

is the convex hull of $\{\mathbf{t}^*_m\}_{m=1}^M$, which is a subset of the linear subspace spanned by $\{\mathbf{t}^*_m\}_{m=1}^M$. Hence, this method of reconstruction is more limited than unrestricted linear reconstruction—and therefore more restricted than PCA using M components, whatever the individual component distributions are. But even if the original data vector $\mathbf{t}$ lies in the convex hull of $\{\mathbf{t}^*_m\}_{m=1}^M$, the reconstructed vector $\mathbf{t}^*$ will not necessarily be equal to $\mathbf{t}$.

In a *responsibility plot*, the responsibility vector $\mathbf{R}(\mathbf{t}) = (p(1|\mathbf{t}), \dots, p(M|\mathbf{t}))^T$ is plotted in the $[0,1]^M$ hypercube, with each axis corresponding to one component. For the mixture to model the data well, the points should accumulate near the axes, indicating that for each point, there is always one component clearly responsible for having generated it. In other words, the distribution $p(m|\mathbf{t})$ should be almost degenerate, concentrating most of the mass in one value of m. Because $\sum_{m=1}^M p(m|\mathbf{t}) = 1$ is the equation of a hyperplane in $\mathbb{R}^M$, we can plot the points in the intersection between the hypercube $[0,1]^M$ and that hyperplane, which meets the coordinate axes at distance $+1$ from the origin. This region is a line for $M = 2$ and an equilateral triangle for $M = 3$. The mixture will perform well if the projected points fall near the vertices of that region.

2.10 Applications for dimensionality reduction

The traditional types of continuous latent variable models (factor analysis and PCA) have been extensively used for dimensionality reduction and related problems such as feature extraction or covariance structure analysis. Examples of such applications can be found in textbooks, e.g. Bartholomew (1987) or Everitt (1984) for factor analysis and Jolliffe (1986) or Diamantaras and Kung (1996) for PCA. Application of other types of continuous latent variable models has only started recently. ICA has been widely applied to signal separation problems, but in general not with the goal of dimensionality reduction, as noted in section 2.9.1.2. A handful of applications of GTM exist, often as a replacement of a Kohonen self-organising map; we study in detail our own application of GTM to a dimensionality reduction problem of speech in chapter 5. Also, we are currently working on the application of continuous latent variable models to a computational neuroscience problem, cortical map modelling (Swindale, 1996), from the point of view of dimensionality reduction. No applications of independent factor analysis exist yet.

2.11 A worked example

To concrete some of the abstract concepts discussed in this chapter, we conclude with a textbook-style example. The example also demonstrates several facts about latent variable models:

- That the marginalisation of the joint probability distribution $p(\mathbf{t}, \mathbf{x})$ can be analytically very difficult.
- The effect of varying noise levels in the induced probability distribution in data space $p(\mathbf{t})$, which is always between two limits:
  - No noise: $p(\mathbf{t}|\mathbf{x}) \to \delta(\mathbf{f}(\mathbf{x})) \Rightarrow p(\mathbf{t}) = \int_{\mathcal{X}} p(\mathbf{t}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \to \begin{cases} 0 & \mathbf{t} \notin \mathcal{M} = \mathbf{f}(\mathcal{X}) \\ p(\mathbf{x}) & \mathbf{t} \in \mathcal{M} = \mathbf{f}(\mathcal{X}), \mathbf{t} = \mathbf{f}(\mathbf{x}) \end{cases}$.
  - Large noise: $p(\mathbf{t}|\mathbf{x}) \approx$ constant $\forall \mathbf{x} \Rightarrow p(\mathbf{t}) \approx p(\mathbf{t}|\mathbf{x})$.

  In both limits the utility of the latent space is lost.

We consider a latent variable model with the following characteristics[40]:

- Latent space of dimension $L = 1$ and piecewise linear prior distribution $p_l(x) = \frac{2}{a^2+b^2} |x|$ on the interval $[-a, b]$ for $a, b \geq 0$.
- Linear mapping from latent onto data space $\mathbf{f}(x) = \mathbf{u}x + \mathbf{v}$.
- D-dimensional data space with Gaussian noise model $p_n(\mathbf{t}|x) = \mathcal{N}(\mathbf{f}(x), \mathbf{\Sigma})$.

Thus, the induced distribution in data space $p_d(\mathbf{t})$ can be obtained as follows:

$$p_d(\mathbf{t}) = \int_{-a}^{b} p_n(\mathbf{t}|x) p_l(x)\, dx = \int_{-a}^{b} \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{t}-(\mathbf{u}x+\mathbf{v}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}-(\mathbf{u}x+\mathbf{v}))} \frac{2}{a^2+b^2} |x|\, dx.$$

[40] We notate the prior, noise and data density functions as p_l, p_n and p_d, respectively.

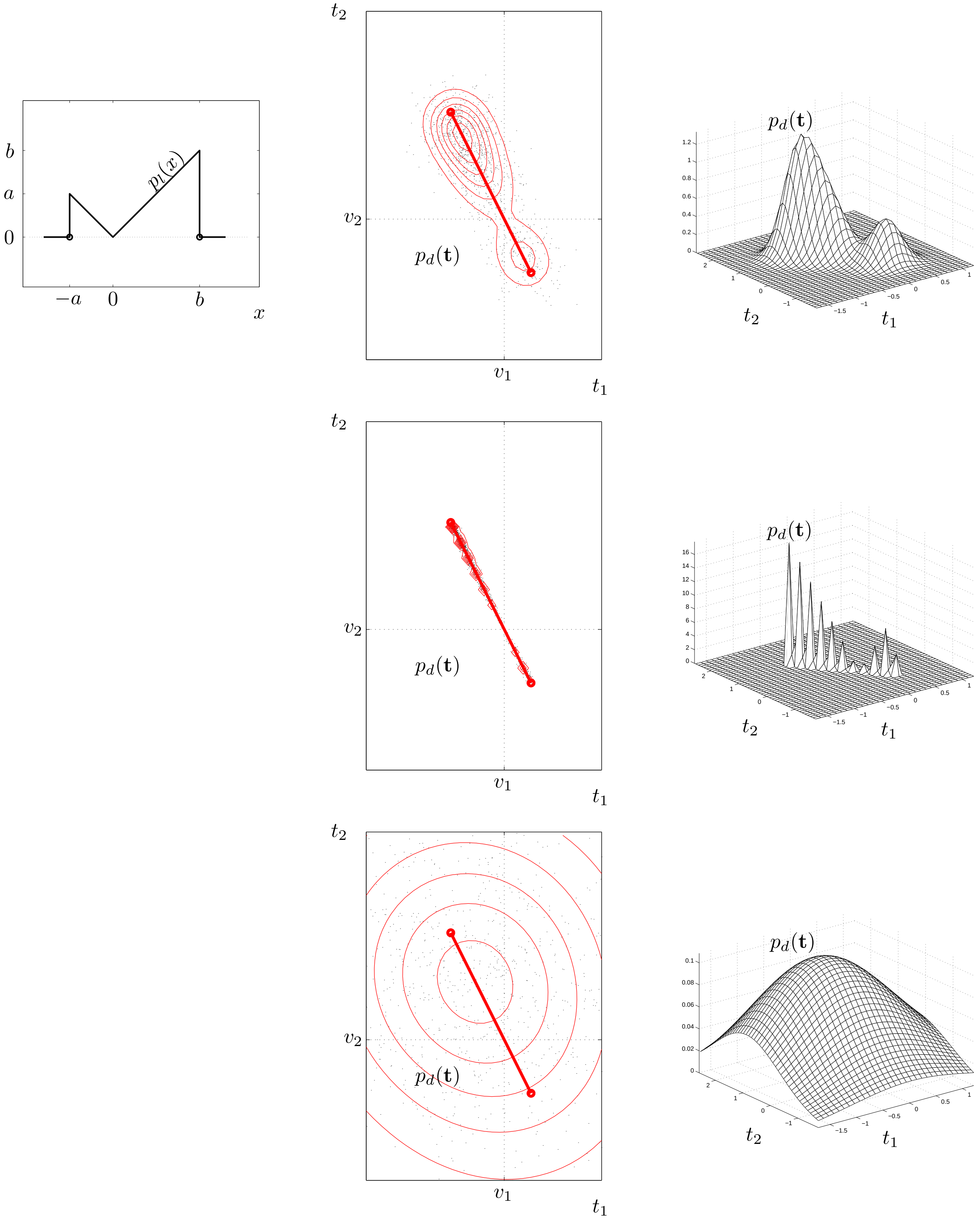


Figure 2.15: Distributions for the example of section 2.11 and effect of varying noise. The top row shows the prior $p_l(x)$ in latent space (left), a sample of 1000 points and a contour plot of the induced distribution $p_d(\mathbf{t})$ in data space (centre) and a surface plot of $p_d(\mathbf{t})$ (right). The manifold $\mathcal{M} = \mathbf{f}([-a, b])$ is linear because the mapping $\mathbf{f}$ is linear, and is represented by the thick solid segment in the centre graphs. Observe how the maxima of the induced density are close, but do not coincide, with the ends of the segment (represented by the circular points), due to the noise added. For this example, the actual parameters had the values: $a = \frac{1}{3}$, $b = \frac{2}{3}$, $\mathbf{u} = (-1, 2)^T$, $\mathbf{v} = (0, 0)^T$ and $\mathbf{\Sigma} = \frac{1}{20}\,\mathbf{I}$ (where $\mathbf{I}$ is the identity matrix). The second and third rows show the effect of very low noise ($\mathbf{\Sigma} = \frac{1}{2000}\,\mathbf{I}$) and very high noise ($\mathbf{\Sigma} = \frac{5}{4}\,\mathbf{I}$), respectively. The "pyramids" in the second row are visual artifacts produced by the mesh being too coarse.

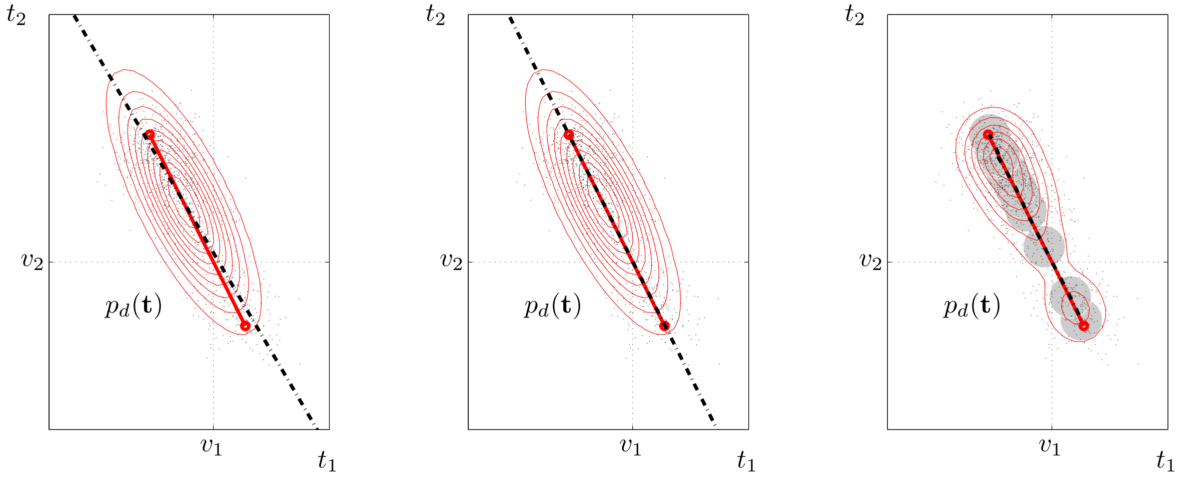


Figure 2.16: Models for the example of section 2.11 (from left to right): factor analysis ($L = 1$ factor), principal component analysis ($L = 1$ principal component) and GTM ($K = 11$ latent grid points, $F = 5$ radial basis functions with a standard deviation $s = 1$ times their separation, EM starting point: the line of the first principal component). For each model, the thick dashed line corresponds to the manifold $\mathcal{M} = \mathbf{f}([-a, b])$ (which in GTM's case overlaps the true one, represented by the thick solid segment). The grey circles in GTM's graph represent the K Gaussian components associated to the latent grid points (eqs. (2.42) and (2.43), with a radius σ). The training set contained $N = 10\,000$ points.

Changing variables to $A = \mathbf{u}^T\boldsymbol{\Sigma}^{-1}\mathbf{u}$, $B = -2(\mathbf{t}-\mathbf{v})^T\boldsymbol{\Sigma}^{-1}\mathbf{u}$, $C = (\mathbf{t}-\mathbf{v})^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\mathbf{v})$, $m = \frac{-B}{2A}$ and $\sigma = A^{-1/2}$ we obtain:

$$p_d(\mathbf{t}) = \frac{2}{a^2+b^2}\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\int_{-a}^{b}|x|\,e^{-\frac{1}{2}(Ax^2+Bx+C)}\,dx = \frac{2}{a^2+b^2}\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\int_{-a}^{b}|x|\,e^{-\frac{1}{2}\left[\left(\frac{x-m}{\sigma}\right)^2+\left(\frac{m}{\sigma}\right)^2-C\right]}\,dx$$

and, changing again to $\alpha = \frac{a+m}{\sigma\sqrt{2}}$, $\beta = \frac{b-m}{\sigma\sqrt{2}}$ and $\gamma = \frac{m}{\sigma\sqrt{2}}$, we finally obtain:

$$p_d(\mathbf{t}) = \frac{2}{a^2+b^2}\frac{\sigma^2}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}e^{\gamma^2-\frac{C}{2}}\left\{-e^{-\alpha^2}-e^{-\beta^2}+2e^{-\gamma^2}+\gamma\sqrt{\pi}\,(\operatorname{erf}(\beta)-\operatorname{erf}(\alpha)+2\operatorname{erf}(\gamma))\right\},$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$ is the error function, which does not have a closed-form expression. α, β and γ are functions of $\mathbf{t}$ and σ depends on the noise and mapping but not on $\mathbf{t}$.

Figure 2.15 shows the prior in latent space $p_l(x)$ and the induced distribution in data space $p_d(\mathbf{t})$ for the true distribution, as well as the effects of low and high noise. Figure 2.16 shows the data space distribution for factor analysis, principal component analysis and GTM. Figure 2.17 comparatively shows the distribution in data space along the manifold segment.

Even though for this problem the true manifold in data space is linear, the distribution $p_d(\mathbf{t})$ in data space is not normal because the prior distribution $p_l(x)$ in latent space is not normal. Thus, the linear-normal models (factor analysis and principal component analysis) produce a bad model while GTM models the true distribution very well. It is interesting to observe that both factor analysis and PCA produce exactly the same normal distribution $p_d(\mathbf{t})$, but while the PCA manifold $\mathcal{M}_{\mathrm{PCA}}$ is aligned with the Gaussian's principal axis, the factor analysis manifold $\mathcal{M}_{\mathrm{FA}}$ is not. The reasons are:

- In this example, where $D = 2$ and $L = 1$, the number of free parameters for factor analysis is $D(L+1) = 4$ and for principal component analysis $DL+1 = 3$. Since the covariance matrix of a bidimensional normal distribution is determined by only $\frac{D(D+1)}{2} = 3$ parameters, both models will account exactly for (and coincide with) the sample covariance matrix—which is the best they can do. Furthermore, since the factor analysis model has more parameters than needed—rotation constraints are not possible because the latent space is one-dimensional—it is undertermined (i.e., non-identifiable): there is an infinite number of equivalent combinations for the values of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ (one of them being the PCA solution).

- For factor analysis, the uniqueness matrix $\boldsymbol{\Psi}$ is not isotropic in general (except when it coincides with the PCA solution), as happens for the maximum likelihood estimate found in fig. 2.16, where $\psi_1 < \psi_2$. Thus

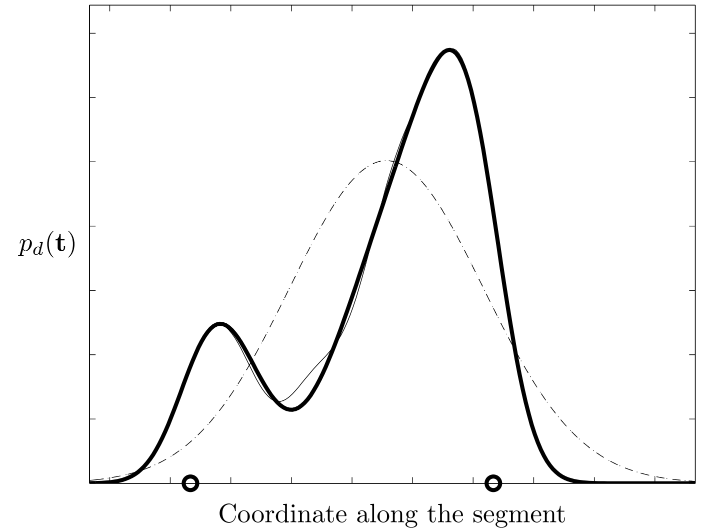


Figure 2.17: Distribution in data space along the manifold segment for the true distribution (thick, solid line) and for each model: factor analysis (dashed line), principal component analysis (dotted line, overlapping with the factor analysis one) and GTM (solid line, almost overlapping with the true one). The circles correspond to the location of the segment ends.

the principal axis of the Gaussian is not parallel to the $\mathbf{\Lambda}$ vector of eq. (2.14). For principal component analysis, the uniqueness matrix $\sigma^2\mathbf{I}$ is isotropic and therefore the principal axis of the Gaussian and the $\mathbf{U}_L$ vector of section 2.6.2 are parallel.

In high-dimensional situations, where factor analysis is identifiable (up to orthogonal rotations), factor analysis will be a better model than PCA, as has been discussed by various researchers (e.g. Hinton et al. 1997; Neal and Dayan 1997). Factor analysis attempts to model the underlying mapping and separately for each variable the noise, while PCA forces the mapping to be aligned with the sample covariance—but the principal component is not directed along the mapping if the noise level is high in another direction.

2.12 Mathematical appendix

2.12.1 Linear-normal models

Theorem 2.12.1. *Consider random variables* $\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}_L(\boldsymbol{\mu}_X, \mathbf{\Sigma}_X)$ *in* $\mathbb{R}^L$ *and* $\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}_D(\mathbf{\Lambda x} + \boldsymbol{\mu}_T, \mathbf{\Sigma}_T)$ *in* $\mathbb{R}^D$, *for symmetric positive definite matrices* $\mathbf{\Sigma}_X$, $\mathbf{\Sigma}_T$, *a* $D \times L$ *matrix* $\mathbf{\Lambda}$ *and vectors* $\boldsymbol{\mu}_X \in \mathbb{R}^L$, $\boldsymbol{\mu}_T \in \mathbb{R}^D$. *Then:*

$$\begin{pmatrix}\mathbf{t}\\ \mathbf{x}\end{pmatrix} \sim \mathcal{N}_{D+L}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}_X\end{pmatrix}, \begin{pmatrix}\mathbf{\Sigma} & \mathbf{\Lambda\Sigma}_X\\ \mathbf{\Sigma}_X\mathbf{\Lambda}^T & \mathbf{\Sigma}_X\end{pmatrix}\right) \qquad \mathbf{t} \sim \mathcal{N}_D(\boldsymbol{\mu}, \mathbf{\Sigma}) \qquad \mathbf{x}|\mathbf{t} \sim \mathcal{N}_L(\hat{\boldsymbol{\mu}}_X, \hat{\mathbf{\Sigma}}_X^{-1}) \tag{2.54}$$

where

$$\boldsymbol{\mu} = \mathbf{\Lambda}\boldsymbol{\mu}_X + \boldsymbol{\mu}_T \tag{2.55}$$

$$\mathbf{\Sigma} = \mathbf{\Sigma}_T + \mathbf{\Lambda\Sigma}_X\mathbf{\Lambda}^T = \mathbf{\Sigma}_T(\mathbf{\Sigma}_T - \mathbf{\Lambda}\hat{\mathbf{\Sigma}}_X^{-1}\mathbf{\Lambda}^T)^{-1}\mathbf{\Sigma}_T \tag{2.56}$$

$$\hat{\boldsymbol{\mu}}_X = \mathbf{A}(\mathbf{t} - \boldsymbol{\mu}) + \boldsymbol{\mu}_X = \mathbf{A}(\mathbf{t} - \boldsymbol{\mu}_T) + \hat{\mathbf{\Sigma}}_X^{-1}\mathbf{\Sigma}_X^{-1}\boldsymbol{\mu}_X \tag{2.57}$$

$$\mathbf{A} = \mathbf{\Sigma}_X\mathbf{\Lambda}^T\mathbf{\Sigma}^{-1} = \hat{\mathbf{\Sigma}}_X^{-1}\mathbf{\Lambda}^T\mathbf{\Sigma}_T^{-1} \tag{2.58}$$

$$\hat{\mathbf{\Sigma}}_X = \mathbf{\Sigma}_X^{-1} + \mathbf{\Lambda}^T\mathbf{\Sigma}_T^{-1}\mathbf{\Lambda} = \mathbf{\Sigma}_X^{-1}(\mathbf{I}_L - \mathbf{\Sigma}_X\mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{\Lambda})^{-1}. \tag{2.59}$$

Proof. From $p(\mathbf{t}, \mathbf{x}) = p(\mathbf{t}|\mathbf{x})p(\mathbf{x})$ and

$$p(\mathbf{t}|\mathbf{x}) = |2\pi\mathbf{\Sigma}_T|^{-1/2}\, e^{-\frac{1}{2}(\mathbf{t}-\mathbf{\Lambda x}-\boldsymbol{\mu}_T)^T\mathbf{\Sigma}_T^{-1}(\mathbf{t}-\mathbf{\Lambda x}-\boldsymbol{\mu}_T)}$$

$$p(\mathbf{x}) = |2\pi\mathbf{\Sigma}_X|^{-1/2}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_X)^T\mathbf{\Sigma}_X^{-1}(\mathbf{x}-\boldsymbol{\mu}_X)}$$

algebraic manipulation gives:

$$(\mathbf{t}-\boldsymbol{\Lambda}\mathbf{x}-\boldsymbol{\mu}_T)^T\boldsymbol{\Sigma}_T^{-1}(\mathbf{t}-\boldsymbol{\Lambda}\mathbf{x}-\boldsymbol{\mu}_T)+(\mathbf{x}-\boldsymbol{\mu}_X)^T\boldsymbol{\Sigma}_X^{-1}(\mathbf{x}-\boldsymbol{\mu}_X)= \\ \begin{pmatrix}\mathbf{t}-\boldsymbol{\mu}\\ \boldsymbol{\mu}_X\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Sigma}_T^{-1} & -\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\\ -\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1} & \boldsymbol{\Sigma}_X^{-1}+\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\end{pmatrix}\begin{pmatrix}\mathbf{t}-\boldsymbol{\mu}\\ \boldsymbol{\mu}_X\end{pmatrix}$$

with $\boldsymbol{\mu}=\boldsymbol{\Lambda}\boldsymbol{\mu}_X+\boldsymbol{\mu}_T$. Theorem A.1.1(ii) proves

$$\begin{pmatrix}\boldsymbol{\Sigma}_T^{-1} & -\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\\ -\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1} & \boldsymbol{\Sigma}_X^{-1}+\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\end{pmatrix}^{-1}=\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Lambda}\boldsymbol{\Sigma}_X\\ \boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T & \boldsymbol{\Sigma}_X\end{pmatrix}$$

with $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_T(\boldsymbol{\Sigma}_T-\boldsymbol{\Lambda}\hat{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{\Lambda}^T)^{-1}\boldsymbol{\Sigma}_T$, $\hat{\boldsymbol{\Sigma}}_X=\boldsymbol{\Sigma}_X^{-1}+\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}=\hat{\boldsymbol{\Sigma}}_X^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}$. The Sherman-Morrison-Woodbury formula (A.2) proves first that $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_T+\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T$ and then that $\hat{\boldsymbol{\Sigma}}_X=\boldsymbol{\Sigma}_X^{-1}(\mathbf{I}_L-\boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}$. Theorem A.1.1(i) proves that

$$\begin{vmatrix}\boldsymbol{\Sigma}_T^{-1} & -\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\\ -\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1} & \boldsymbol{\Sigma}_X^{-1}+\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\end{vmatrix}=\left|\boldsymbol{\Sigma}_X\right|^{-1}\left|\boldsymbol{\Sigma}_T\right|^{-1}\ \text{ and }\ \begin{vmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Lambda}\boldsymbol{\Sigma}_X\\ \boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T & \boldsymbol{\Sigma}_X\end{vmatrix}=\left|\boldsymbol{\Sigma}\right|\left|\hat{\boldsymbol{\Sigma}}_X\right|^{-1}=\left|\boldsymbol{\Sigma}_X\right|\left|\boldsymbol{\Sigma}_T\right|.$$

Finally, theorem A.3.1(iv) and the previous results prove that $\mathbf{x}|\mathbf{t}\sim\mathcal{N}(\hat{\boldsymbol{\mu}}_X,\hat{\boldsymbol{\Sigma}}_X^{-1})$ with $\hat{\boldsymbol{\mu}}_X$ as defined above. □

Corollary 2.12.2 (Isotropic zero-noise limit of theorem 2.12.1). *In the same conditions of theorem 2.12.1, when $\boldsymbol{\Sigma}_T=k\mathbf{I}$ and $k\to 0^+$ then:*

$$\begin{pmatrix}\mathbf{t}\\ \mathbf{x}\end{pmatrix}\to\mathcal{N}_{D+L}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}_X\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Lambda}\boldsymbol{\Sigma}_X\\ \boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T & \boldsymbol{\Sigma}_X\end{pmatrix}\right)\qquad \mathbf{t}\to\mathcal{N}_D(\boldsymbol{\mu},\boldsymbol{\Sigma})\qquad \mathbf{x}|\mathbf{t}\to\delta_L(\hat{\boldsymbol{\mu}}_X)$$

where

$$\boldsymbol{\mu}=\boldsymbol{\Lambda}\boldsymbol{\mu}_X+\boldsymbol{\mu}_T \tag{2.60}$$
$$\boldsymbol{\Sigma}=\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X\boldsymbol{\Lambda}^T \tag{2.61}$$
$$\hat{\boldsymbol{\mu}}_X=\mathbf{A}(\mathbf{t}-\boldsymbol{\mu}_T) \tag{2.62}$$
$$\mathbf{A}=\mathbf{A}^+=(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda} \tag{2.63}$$
$$\hat{\boldsymbol{\Sigma}}_X\to\infty. \tag{2.64}$$

The normal distributions above are degenerate in the space of $\mathbf{t}$*: only those data points* $\mathbf{t}\in\operatorname{span}\{\boldsymbol{\Lambda}\}=\operatorname{im}\mathbf{f}$ *have nonzero density.*

The following theorem proves that, for linear-normal models, the posterior distribution in latent space $\mathbf{x}|\mathbf{t}\sim\mathcal{N}_L(\hat{\boldsymbol{\mu}}_X,\hat{\boldsymbol{\Sigma}}_X^{-1})$ is always narrower than the prior distribution $\mathbf{x}\sim\mathcal{N}_L(\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X)$, the narrower the smaller the noise is: $\boldsymbol{\Sigma}_T\to\mathbf{0}\Rightarrow\hat{\boldsymbol{\Sigma}}_X^{-1}\to\mathbf{0}$.

Theorem 2.12.3. *In the same conditions of theorem 2.12.1, $\left|\hat{\boldsymbol{\Sigma}}_X^{-1}\right|<\left|\boldsymbol{\Sigma}_X\right|$.*

Proof. We have that:

- $\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}$ is positive definite because, for any $\mathbf{x}\neq\mathbf{0}$, $\mathbf{x}^T\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\mathbf{x}=\mathbf{y}^T\boldsymbol{\Sigma}_T^{-1}\mathbf{y}>0$, since $\boldsymbol{\Sigma}_T$ is positive definite and $\mathbf{y}\overset{\text{def}}{=}\boldsymbol{\Lambda}\mathbf{x}\neq\mathbf{0}$ (since $\boldsymbol{\Lambda}$ is full-rank).
- $\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X$ is positive definite because the product of positive definite matrices is positive definite.
- If $\mathbf{B}$ is positive definite then $\left|\mathbf{I}+\mathbf{B}\right|>1$, since decomposing spectrally $\mathbf{B}=\mathbf{U}\mathbf{V}\mathbf{U}^T$ with $\mathbf{V}>0$ diagonal and $\mathbf{U}$ orthogonal: $\left|\mathbf{I}+\mathbf{B}\right|=\left|\mathbf{U}\mathbf{U}^T+\mathbf{U}\mathbf{V}\mathbf{U}^T\right|=\left|\mathbf{U}\right|\left|\mathbf{I}+\mathbf{V}\right|\left|\mathbf{U}^T\right|=\left|\mathbf{I}+\mathbf{V}\right|=\prod_{l=1}^L(1+v_l)>1$.

Hence $\left|\hat{\boldsymbol{\Sigma}}_X\right|\left|\boldsymbol{\Sigma}_X\right|=\left|\hat{\boldsymbol{\Sigma}}_X\boldsymbol{\Sigma}_X\right|=\left|\mathbf{I}+\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X\right|>1\Rightarrow\left|\hat{\boldsymbol{\Sigma}}_X^{-1}\right|<\left|\boldsymbol{\Sigma}_X\right|$. □

2.12.2 Independence relations

Theorem 2.12.4 (Pairwise local independence). $p(t_i t_j|\mathbf{x}) = p(t_i|\mathbf{x})p(t_j|\mathbf{x})\ \forall i,j \in \{1,\dots,D\}$.

Proof. Call $\mathcal{I} = \{1,\dots,D\} \setminus \{i,j\}$. Then:

$$p(t_i t_j|\mathbf{x}) = \int p(\mathbf{t}|x)\, d\mathbf{t}_{\mathcal{I}} = \int \prod_{d\in\{1,\dots,D\}} p(t_d|\mathbf{x})\, d\mathbf{t}_{\mathcal{I}} = p(t_i|\mathbf{x})p(t_j|\mathbf{x}) \int p(\mathbf{t}_{\mathcal{I}}|\mathbf{x})\, d\mathbf{t}_{\mathcal{I}} = p(t_i|\mathbf{x})p(t_j|\mathbf{x})$$

where we have used the axiom of local independence (2.4). □

Definition 2.12.1 and theorems 2.12.5 and 2.12.6 are from Cover and Thomas (1991).

Definition 2.12.1. The random variables X, Y and Z form a Markov chain $X \to Y \to Z$ in that order if the conditional distribution of Z depends only on Y and is conditionally independent of X, i.e., $p(x,y,z) = p(x)p(y|x)p(z|y) = p(x,y)p(z|y)$.

Theorem 2.12.5. *$X \to Y \to Z$ if and only if X and Z are conditionally independent given Y.*

Proof. $p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y)$. □

Theorem 2.12.6. *If $X \to Y \to Z$ then:*

(i) $X \to Y \to Z \Rightarrow Z \to Y \to X$.

(ii) $I(X;Z) \le I(X;Y)$*: no clever manipulation of the data Y (deterministic or random) can improve the inferences that can be made from the data, i.e., the information that Y contains about X (data processing inequality).*

(iii) $I(X;Y|Z) \le I(X;Y)$*: the dependency of X and Y is decreased, or remains unchanged, by observation of a downstream random variable Z. Note, though, that $I(X;Y|Z) > I(X;Y)$ may happen if X, Y and Z do not form a Markov chain.*

Theorem 2.12.7 (Pairwise Markov chains). *$t_i \to \mathbf{x} \to t_j\ \forall i,j \in \{1,\dots,D\}$, i.e., there is a Markov chain between any two observed variables via the latent ones.*

Proof. From theorems 2.12.4 and 2.12.5. □

2.12.3 Latent variable models and entropy

The following results are easily proven using theorem 2.12.4 and the information theory results from section A.4.

Theorem 2.12.8.

- $I(t_i;t_j|\mathbf{x}) = 0$.
- $h(t_i|t_j,\mathbf{x}) = h(t_i|\mathbf{x})\ \forall i,j \in \{1,\dots,D\}$.
- $h(t_i|\mathbf{x}) + h(t_j|\mathbf{x}) = h(t_i,t_j|\mathbf{x})\ \forall i,j \in \{1,\dots,D\}$.
- $h(\mathbf{t}|\mathbf{x}) = \sum_{d=1}^{D} h(t_d|\mathbf{x})$.
- $h(\mathbf{t},\mathbf{x}) = h(\mathbf{t}) + h(\mathbf{x}|\mathbf{t}) = h(\mathbf{x}) + h(\mathbf{t}|\mathbf{x}) = h(\mathbf{x}) + \sum_{d=1}^{D} h(t_d|\mathbf{x})$. *If the latent variables are mutually independent,* $h(\mathbf{t},\mathbf{x}) = \sum_{l=1}^{L} h(x_l) + \sum_{d=1}^{D} h(t_d|\mathbf{x})$.
- $0 \le I(t_i;\mathbf{x}|t_j) \le I(t_i;t_j) \le \min\left(I(t_i;\mathbf{x}), I(t_j;\mathbf{x})\right) \le \min\left(h(\mathbf{x}), h(t_i), h(t_j)\right)\ \forall i,j \in \{1,\dots,D\}$.

For models where the prior distribution in latent space has been sampled, $p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}_k)\delta(\mathbf{x}-\mathbf{x}_k)$, the distributions $p(\mathbf{t})$, $p(\mathbf{x}|\mathbf{t})$ and $p(\mathbf{t}_{\mathcal{I}}|\mathbf{t}_{\mathcal{J}})$ are finite mixtures, whose entropy cannot be computed analytically. However, we can give bounds for it.

Theorem 2.12.9. *Whether the prior distribution in latent space is continuous or has been sampled, if $\mathbf{\Psi}$ is independent of $\mathbf{x}$ and $\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Psi})$, then $h(\mathbf{t}|\mathbf{x}) = \frac{1}{2}\ln|2\pi e\mathbf{\Psi}|$.*

Proof.

- Continuous $p(\mathbf{x})$: $h(\mathbf{t}|\mathbf{x}) \overset{\text{def}}{=} -\int_{\mathbb{R}^L} p(\mathbf{x}) \int_{\mathbb{R}^D} p(\mathbf{t}|\mathbf{x}) \ln p(\mathbf{t}|\mathbf{x})\, d\mathbf{t}\, d\mathbf{x} = \int_{\mathbb{R}^L} p(\mathbf{x}) \frac{1}{2} \ln |2\pi e \mathbf{\Psi}|\, d\mathbf{x} = \frac{1}{2} \ln |2\pi e \mathbf{\Psi}|$.
- Sampled $p(\mathbf{x})$: $h(\mathbf{t}|\mathbf{x}) \overset{\text{def}}{=} -\sum_{k=1}^{K} p(\mathbf{x}_k) \int_{\mathbb{R}^D} p(\mathbf{t}|\mathbf{x}) \ln p(\mathbf{t}|\mathbf{x})\, d\mathbf{t} = \sum_{k=1}^{K} p(\mathbf{x}_k) \frac{1}{2} \ln |2\pi e \mathbf{\Psi}| = \frac{1}{2} \ln |2\pi e \mathbf{\Psi}|$.

□

Remark. Although the axiom of local independence prescribes factorised noise models, theorem 2.12.9 holds even if $\mathbf{\Psi}$ is not diagonal.

Theorem 2.12.10 (Entropy for linear-normal models). *Consider random variables* $\mathbf{x} \sim \mathcal{N}_L(\boldsymbol{\mu}_X, \mathbf{\Sigma}_X)$ *in* $\mathbb{R}^L$ *and* $\mathbf{t}|\mathbf{x} \sim \mathcal{N}_D(\mathbf{\Lambda}\mathbf{x} + \boldsymbol{\mu}_T, \mathbf{\Sigma}_T)$ *in* $\mathbb{R}^D$, *for positive definite matrices* $\mathbf{\Sigma}_X$, $\mathbf{\Sigma}_T$, *a* $D \times L$ *matrix* $\mathbf{\Lambda}$ *and vectors* $\boldsymbol{\mu}_X \in \mathbb{R}^L$, $\boldsymbol{\mu}_T \in \mathbb{R}^D$. *Then, with* $\mathbf{\Sigma}$ *and* $\hat{\mathbf{\Sigma}}_X$ *given as in theorem 2.12.1:*

- $h(\mathbf{x}) = \frac{1}{2} \ln |2\pi e \mathbf{\Sigma}_X|$
- $h(\mathbf{t}|\mathbf{x}) = \frac{1}{2} \ln |2\pi e \mathbf{\Sigma}_T|$
- $h(\mathbf{t}, \mathbf{x}) = \frac{1}{2} \ln (|2\pi e \mathbf{\Sigma}_X|\, |2\pi e \mathbf{\Sigma}_T|)$
- $h(\mathbf{t}) = \frac{1}{2} \ln |2\pi e \mathbf{\Sigma}| = \frac{1}{2} \ln \left| 2\pi e (\mathbf{\Sigma}_T + \mathbf{\Lambda} \mathbf{\Sigma}_X \mathbf{\Lambda}^T) \right|$
- $h(\mathbf{x}|\mathbf{t}) = \frac{1}{2} \ln \left| 2\pi e \hat{\mathbf{\Sigma}}_X^{-1} \right| = \frac{1}{2} \ln \left(|2\pi e \mathbf{\Sigma}_X|\, |2\pi e \mathbf{\Sigma}_T|\, |2\pi e \mathbf{\Sigma}|^{-1} \right)$.

Theorem 2.12.11 (Entropy for factor analysis). *If* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$ *and* $\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{\Lambda}\mathbf{x} + \boldsymbol{\mu}, \mathbf{\Psi})$ *with* $\mathbf{\Psi}$ *diagonal and independent of* $\mathbf{x}$, *then:*

- $h(\mathbf{x}) = \frac{L}{2} \ln (2\pi e)$
- $h(\mathbf{t}|\mathbf{x}) = \frac{1}{2} \ln |2\pi e \mathbf{\Psi}|$
- $h(\mathbf{t}, \mathbf{x}) = \frac{1}{2} \ln \left((2\pi e)^{D+L} |\mathbf{\Psi}| \right)$
- $h(\mathbf{t}) = \frac{1}{2} \ln \left((2\pi e)^D \left| \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \right| \right)$
- $h(\mathbf{x}|\mathbf{t}) = \frac{1}{2} \ln \left((2\pi e)^L \left| \mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda} + \mathbf{I}_L \right|^{-1} \right)$.

Theorem 2.12.12 (Entropy for principal component analysis). *If* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$ *and* $\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{\Lambda}\mathbf{x} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_D)$ *with* σ *independent of* $\mathbf{x}$, *then:*

- $h(\mathbf{x}) = \frac{L}{2} \ln (2\pi e)$
- $h(\mathbf{t}|\mathbf{x}) = \frac{D}{2} \ln (2\pi e \sigma^2)$
- $h(\mathbf{t}, \mathbf{x}) = \frac{1}{2} \ln \left((2\pi e)^{D+L} \sigma^{2D} \right)$
- $h(\mathbf{t}) = \frac{1}{2} \ln \left((2\pi e)^D \left| \mathbf{\Lambda}\mathbf{\Lambda}^T + \sigma^2 \mathbf{I}_D \right| \right)$
- $h(\mathbf{x}|\mathbf{t}) = \frac{1}{2} \ln \left((2\pi e \sigma^2)^L \left| \mathbf{\Lambda}^T \mathbf{\Lambda} + \sigma^2 \mathbf{I}_L \right|^{-1} \right)$.

Theorem 2.12.13 (Entropy for GTM). *If* $p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{x}_k)$ *and* $\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{f}(\mathbf{x}), \sigma^2 \mathbf{I}_D)$ *with* σ *independent of* $\mathbf{x}$ *and* $\mathbf{f}$ *a generalised linear model, then:*

- $h(\mathbf{x}) = \ln K$
- $h(\mathbf{t}|\mathbf{x}) = \frac{D}{2} \ln (2\pi e \sigma^2)$
- $h(\mathbf{t}, \mathbf{x}) = \frac{1}{2} \ln \left(K^2 (2\pi e \sigma^2)^D \right)$.

Theorems 2.12.9–2.12.13 show that, for normal noise models with fixed covariance $\mathbf{\Psi}$, the entropy of the observed variables, $h(\mathbf{t})$, is bounded below by the entropy of the normal distribution of covariance $\mathbf{\Psi}$. This means that we cannot have a distribution $p(\mathbf{t})$ in observed space which is, loosely speaking, more peaked than a $\mathcal{N}(\cdot, \mathbf{\Psi})$. Since it is to be expected that real distributions will have variable "peakiness" depending on the region considered, the optimal $\mathbf{\Psi}$ will be a compromise between the covariance in areas of large noise and in areas of small noise. In GTM this could be overcome by having many points in latent space (high K) and a small covariance $\mathbf{\Psi} = \sigma^2 \mathbf{I}_D$. However, allowing the covariance $\mathbf{\Psi}$ to depend on $\mathbf{x}$ (like the mean of the noise model does, $\mathbf{f}(\mathbf{x})$) would be the obvious workaround.

2.12.4 Diagonal GTM (dGTM)

We give here the details for dGTM, the diagonal noise model for GTM that we proposed in section 2.6.5.1. First let us fully generalise GTM so that it has diagonal noise dependent on each latent grid point and both the values of the noise covariance matrix and the values of the prior distribution in latent space are trainable. Thus

$$\mathbf{t}|\mathbf{x} \overset{\text{def}}{\sim} \mathcal{N}(\mathbf{f}(\mathbf{x}), \mathbf{\Psi}(\mathbf{x})) \tag{2.42'}$$

with $\mathbf{\Psi} : \mathcal{X} \to (\mathbb{R}^+)^D$ and

$$p(\mathbf{t}) = \sum_{k=1}^{K} \pi_k p(\mathbf{t}|\mathbf{x}_k)$$

with $\pi_k \overset{\text{def}}{=} p(\mathbf{x}_k)$ and $\mathbf{\Psi}_k \overset{\text{def}}{=} \mathbf{\Psi}(\mathbf{x}_k) = \operatorname{diag}(\psi_{k1}, \dots, \psi_{kD})$. The dependence of $p(\mathbf{t}|\mathbf{x}_k)$ on all the relevant parameters is not explicitly written for clarity of notation. The EM equations are derived by minimising

$$\sum_{n=1}^{N} \ln p(\mathbf{t}_n) - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

where λ is a Lagrange multiplier that ensures that the values of the prior distribution in latent space add to one. This gives

$$\pi_k^{(\tau+1)} = \frac{1}{N} \sum_{n=1}^{N} p(\mathbf{x}_k|\mathbf{t}_n) = \frac{1}{N} \sum_{n=1}^{N} R_{nk}^{(\tau)}$$

for the prior distribution in latent space, using the responsibilities $R_{nk} = p(\mathbf{x}_k|\mathbf{t}_n)$, and for the rest of the parameters we obtain

$$\sum_{n=1}^{N} \sum_{k=1}^{K} R_{nk} \frac{\partial}{\partial \theta} \ln p(\mathbf{t}_n|\mathbf{x}_k) = 0 \tag{2.65}$$

where θ represents a parameter w_{df} or ψ_{kd}. Recalling that $f_d(\mathbf{x}_k) = \sum_{f=1}^{F} w_{df} \phi_{kf}$ where $\phi_{kf} = \phi_f(\mathbf{x}_k)$ and plugging

$$\frac{\partial}{\partial w_{df}} \ln p(\mathbf{t}_n|\mathbf{x}_k) = \frac{\phi_{kf}}{\psi_{kd}} \left(t_{nd} - \sum_{f'=1}^{F} w_{df'} \phi_{kf'} \right) \quad d = 1, \dots, D \quad f = 1, \dots, F$$

$$\frac{\partial}{\partial \psi_{kd}} \ln p(\mathbf{t}_n|\mathbf{x}_k) = -\frac{1}{2} \left(\frac{1}{\psi_{kd}} - \frac{(t_{nd} - f_d(\mathbf{x}_k))^2}{\psi_{kd}^2} \right) \quad k = 1, \dots, K \quad d = 1, \dots, D$$

$$\frac{\partial}{\partial \psi_{kd}} \ln p(\mathbf{t}_n|\mathbf{x}_{k'}) = 0 \text{ if } k \neq k'$$

into (2.65), we obtain a system of $D(F+K)$ equations that, together with the equations for $\{\pi_k\}_{k=1}^{K}$, define the M step:

$$\sum_{k=1}^{K} \frac{\phi_{kf}}{\psi_{kd}} g_{kk} \sum_{f'=1}^{F} \phi_{kf'} w_{df'} = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\phi_{kf}}{\psi_{kd}} R_{nk} t_{nd} \quad d = 1, \dots, D \quad f = 1, \dots, F$$

$$\psi_{kd} = \frac{1}{g_{kk}} \sum_{n=1}^{N} R_{nk} (t_{nd} - f_d(\mathbf{x}_k))^2 \quad k = 1, \dots, K \quad d = 1, \dots, D$$

where $g_{kk} \overset{\text{def}}{=} \sum_{n=1}^{N} R_{nk}$ as in section 2.6.5. Solving for $\{w_{df}\}_{d,f=1}^{D,F}$ is cumbersome, since the first group of equations cannot be put into a nice matrix equation form. The second group of equations shows that $\mathbf{\Psi}_k$ is a weighted average of the squared componentwise deviations of the data points from the "reference point" $\mathbf{f}(\mathbf{x}_k)$, where the weights are given by the responsibilities (normalised over all reference points).

Compared to the standard GTM model, this extended version has $K(D+1)-2$ more parameters: $\{\pi_k\}_{k=1}^{K-1}$ and $\{\mathbf{\Psi}_k\}_{k=1}^{K}$. Thus, we lose the attractive property that using a large number of latent points K (exponentially dependent on the latent space dimension L) does not increase the number of parameters. Even if the EM algorithm was straightforward, such a model would require a large training set for stable statistical estimation[41]

[41] Note, however, that the standard GTM model does depend on F, the number of radial basis functions needed for the mapping $\mathbf{f}$, which also depends exponentially on L (but can be kept quite smaller than K without losing approximation power).

and would be prone to singularity problems ($\mathbf{\Psi}_k \to \mathbf{0}$ for some k), as mentioned in section 2.5.2, among other places. Defining $\mathbf{\Psi}(\mathbf{x})$ as a generalised linear model (as $\mathbf{f}$ is, eq. (2.41), and reusing the $\boldsymbol{\phi}$ function) eliminates the direct dependence on K and ensures a smooth variability of the noise variance over the data manifold, but the resulting M step is unsolvable, as can be readily checked. Also, since $\mathbf{f}$ is a universal approximator, it is not really necessary to make the prior distribution in latent space trainable as well.

Taking into account these considerations, for all $k = 1, \dots, K$ we keep $p(\mathbf{x}_k) = \frac{1}{K}$ constant as in the standard GTM model and the noise model covariance matrix $\mathbf{\Psi}_k = \mathbf{\Psi}$ diagonal constant too. As can easily be seen by recalculating $\frac{\partial}{\partial \psi_d} \ln p(\mathbf{t}_n|\mathbf{x}_k)$, the M step becomes exactly solvable and leads to the following update equations for the parameters $\mathbf{W}$ and $\mathbf{\Psi}$, respectively:

$$\mathbf{\Phi}^T \mathbf{G}^{(\tau)} \mathbf{\Phi} (\mathbf{W}^{(\tau+1)})^T = \mathbf{\Phi}^T (\mathbf{R}^{(\tau)})^T \mathbf{T} \tag{2.45a}$$

$$\psi_d^{(\tau+1)} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} R_{nk}^{(\tau)} (t_{nd} - f_d(\mathbf{x}_k))^2 \quad d = 1, \dots, D \tag{2.45b'}$$

the first of which is the same one as for the standard GTM model and the second of which has the same interpretation as before as weighted average of the squared componentwise deviations of the data points from $\mathbf{f}(\mathbf{x}_k)$. Further modifications can still be done, such as a Gaussian regularisation term on $\mathbf{W}$, eq. (2.46), or a unit-speed constraint on $\mathbf{f}$, section 2.8.3.

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.