Density Networks for Dimension Reduction of Continuous Data: Analytical Solutions*

Miguel Á. Carreira-Perpiñán

Technical Report CS–97–09
Dept. of Computer Science
University of Sheffield
M.Carreira@dcs.shef.ac.uk

April 23, 1997

Abstract

MacKay's density networks (1995) provide a framework for generative modelling which can be adapted to specific problems by conveniently selecting a prior in latent space, a noise model in data space and a smooth, parametric mapping from latent to data space. The particular model thus obtained expresses the structure of a distribution in a high-dimensional data space in terms of a small number of variables in latent space.

Here, we consider the problem of dimension reduction of continuous, high-dimensional data by density networks. We explore the possibilities of obtaining analytical expressions of the optimal parameters for general classes of distributions and mappings under maximum likelihood of the data. Only the case of a linear mapping and Gaussian prior and noise model seems to admit a straightforward solution, equivalent to factor analysis (and principal factor analysis). Nonlinear mappings and other families of distributions are not analytically tractable; approximate methods (e.g. Monte Carlo) are necessary in such situations, but then the complexity of the procedure becomes exponential on the number of latent variables due to the curse of the dimensionality.

1 Introduction

1.1 Definition of the problem

In *dimension reduction of continuous data* [4], we want to model a certain distribution $p(\mathbf{t})$ in data space $\mathbb{R}^D$, given a sample $\{\mathbf{t}_n\}_{n=1}^N$ drawn independently from it, in terms of a small number L of *latent variables* [1, 5]. As a result, we obtain a projection from the manifold spanned by the data onto the latent space that preserves the topography of the manifold. We call this projection a *coordinate change*.

For convenience, we will represent the data sample in matrix form as $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_N)$.

1.2 The density networks framework

Density networks (MacKay [8]) are a form of Bayesian learning that provides a generative model of the data in terms of latent variables. The basic elements of a density network are:

- A prior distribution on L-dimensional latent space: $p(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^L$.
- A smooth mapping $\mathbf{y}$ from the latent space onto an L-dimensional[1] manifold $\mathcal{Y}$ in data space, with parameters θ:

$$\begin{array}{rccc} \mathbf{y}: & \mathbb{R}^L & \longrightarrow & \mathcal{Y} \subset \mathbb{R}^D \\ & \mathbf{x} & \longmapsto & \mathbf{y}(\mathbf{x};\theta) \end{array}$$

This jump from L to D dimensions is the key for the dimension reduction.

*This work has been partially funded by the Spanish Ministry of Education.

[1] We will only consider one-to-one mappings in which the topological dimension of the spaces involved is preserved.

- The error function or noise model on D-dimensional data space: $p(\mathbf{t}|\mathbf{x},\theta) = p(\mathbf{t}|\mathbf{y})$, $\mathbf{t} \in \mathbb{R}^D$.

$p(\mathbf{x})$ and $p(\mathbf{t}|\mathbf{x},\theta)$ can also have additional parameters, which we include in θ for convenience. Using Bayes' rule, we can compute the posterior in latent space:

$$p(\mathbf{x}|\mathbf{t},\theta) = \frac{p(\mathbf{t}|\mathbf{x},\theta)p(\mathbf{x})}{p(\mathbf{t}|\theta)} \tag{1}$$

with normalisation constant

$$p(\mathbf{t}|\theta) = \int p(\mathbf{t}|\mathbf{x},\theta)p(\mathbf{x})\,d\mathbf{x}. \tag{2}$$

A further application of Bayes' rule would give the posterior in parameter space:

$$p(\theta|\{\mathbf{t}_n\}_{n=1}^N) = \frac{p(\{\mathbf{t}_n\}_{n=1}^N|\theta)p(\theta)}{p(\{\mathbf{t}_n\}_{n=1}^N)}$$

where $p(\{\mathbf{t}_n\}_{n=1}^N|\theta) = \prod_{n=1}^N p(\mathbf{t}_n|\theta) = L(\theta)$ is the likelihood of the parameters and $p(\theta)$ is some parameter prior. Alternatively, the parameters θ can be learned by maximum likelihood[2]:

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

by finding the parameter sets that make zero the gradient of the log-likelihood $l(\theta) = \sum_{n=1}^N \ln p(\mathbf{t}_n|\theta)$:

$$\hat{\theta} : \nabla_\theta l(\hat{\theta}) = \sum_{n=1}^N \frac{1}{p(\mathbf{t}_n|\hat{\theta})} \left.\frac{\partial p(\mathbf{t}_n|\theta)}{\partial\theta}\right|_{\hat{\theta}} = \mathbf{0}. \tag{3}$$

1.3 Selection of mapping, prior distribution and noise model

Assuming that the parameters are found by maximum likelihood, the analytical treatment of a density network becomes difficult in two points: when marginalising over the latent space, integral (2), and when finding the roots of the gradient of the log-likelihood, eq. (3). In fact, the only tractable case seems to be when both distributions (prior in latent space and noise model in data space) are normal and the mapping is linear; we study this case in the next section. The use of uniform distributions simplifies the expressions of the distribution by complicating the integration regions, preventing further treatment for any kind of mapping; see appendix B for details. Similar difficulties can be expected from other, more complex distributions and from nonlinear mappings.

If the prior in latent space is a discrete distribution, integral (2) becomes a discrete sum. We can consider this situation, at least from a practical point of view, as a Monte Carlo sampling of a continuous distribution (see section 3).

2 Linear Gaussian density networks

In this section we consider:

- A linear mapping characterised by a $D \times L$ matrix $\mathbf{W}$ and a D-dimensional vector $\mathbf{b}$:

$$\mathbf{y}(\mathbf{x};\theta) = \mathbf{W}\mathbf{x} + \mathbf{b}.$$

- A Gaussian prior in latent space $\mathcal{N}_L(\mathbf{0}, \mathbf{\Sigma}_X)$, of fixed covariance $\mathbf{\Sigma}_X$:

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

  The normal distribution has, for fixed variance, the highest entropy, i.e. it is the least informative distribution about the data. This justifies choosing it as the prior in latent space, apart from the fact that it simplifies the calculations. Obviously the location of the mean is irrelevant and we set it to the origin for convenience.

- A Gaussian noise model in data space $\mathcal{N}_D(\mathbf{y}(\mathbf{x};\theta), \mathbf{\Sigma}_Y)$, centred in $\mathbf{y}(\mathbf{x};\theta)$ and with adjustable covariance $\mathbf{\Sigma}_Y$:

$$p(\mathbf{t}|\mathbf{x};\theta) = (2\pi)^{-D/2}|\mathbf{\Sigma}_Y|^{-1/2}e^{-\frac{1}{2}(\mathbf{t}-\mathbf{W}\mathbf{x}-\mathbf{b})^T\mathbf{\Sigma}_Y^{-1}(\mathbf{t}-\mathbf{W}\mathbf{x}-\mathbf{b})}.$$

The parameters to be determined by maximum likelihood are $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{\Sigma}_Y\}$.

[2] Which is also simpler, because it is not necessary to marginalise over the parameters in order to obtain $p(\{\mathbf{t}_n\}_{n=1}^N) = \int L(\theta)p(\theta)\,d\theta$.

2.1 Marginalisation over the latent space

Integral (2) can be computed exactly, as the convolution of two Gaussians is a Gaussian as well; the details are shown in appendix C. The data distribution for given values of the parameters is $p(\mathbf{t}|\theta) \sim \mathcal{N}_D(\mathbf{b}, \mathbf{\Sigma})$ and the posterior in latent space (1) is $p(\mathbf{x}|\mathbf{t}, \theta) \sim \mathcal{N}_L(\boldsymbol{\mu}_X, \mathbf{A}^{-1})$, where:

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Sigma}_Y + \mathbf{W}\mathbf{\Sigma}_X\mathbf{W}^T \\ \mathbf{A} &= \mathbf{\Sigma}_X^{-1} + \mathbf{W}^T\mathbf{\Sigma}_Y^{-1}\mathbf{W} \\ \boldsymbol{\mu}_X &= \mathbf{A}^{-1}\mathbf{W}^T\mathbf{\Sigma}_Y^{-1}(\mathbf{t} - \mathbf{b}). \end{aligned} \tag{4}$$

Observe that the decomposition of the covariance as $\mathbf{\Sigma} = \mathbf{\Sigma}_Y + \mathbf{W}\mathbf{\Sigma}_X\mathbf{W}^T$ is exactly the one that is obtained in **factor analysis** [9], where the latent variables $\mathbf{x}$ are the common factors (with covariance[3] $\mathbf{\Sigma}_X$), $\mathbf{W}$ contains the factor loadings and $\mathbf{\Sigma}_Y$ contains the specific variances, a result reported by Bishop *et al.* [2] and also known much before out of the context of density networks [1].

We take now $\mathbf{\Sigma}_X = \mathbf{I}_L$, which amounts to absorbing $\mathbf{\Sigma}_X^{1/2}$ into $\mathbf{W}$. Then $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$ and the previous formulae are slightly simplified[4]:

$$\left.\begin{aligned} p(\mathbf{x}) &\sim \mathcal{N}_L(\mathbf{0}, \mathbf{I}_L) \\ p(\mathbf{t}|\mathbf{x}, \theta) &\sim \mathcal{N}_D(\mathbf{W}\mathbf{x} + \mathbf{b}, \mathbf{\Sigma}_Y) \end{aligned}\right\} \Longrightarrow \begin{aligned} p(\mathbf{t}|\theta) &\sim \mathcal{N}_D(\mathbf{b}, \mathbf{\Sigma}) \\ p(\mathbf{x}|\mathbf{t}, \theta) &\sim \mathcal{N}_L(\boldsymbol{\mu}_X, \mathbf{A}^{-1}) \end{aligned}$$

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Sigma}_Y + \mathbf{W}\mathbf{W}^T \\ \mathbf{A} &= \mathbf{I}_L + \mathbf{W}^T\mathbf{\Sigma}_Y^{-1}\mathbf{W} \\ \boldsymbol{\mu}_X &= \mathbf{A}^{-1}\mathbf{W}^T\mathbf{\Sigma}_Y^{-1}(\mathbf{t} - \mathbf{b}). \end{aligned} \tag{5}$$

The decomposition of the covariance as $\mathbf{\Sigma} = \mathbf{\Sigma}_Y + \mathbf{W}\mathbf{W}^T$ corresponds to **factor analysis** with standardised[5] factors.

2.2 Maximum likelihood estimation of the parameters

We can now estimate the parameters by maximum likelihood[6]. Notice that if $\mathbf{\Sigma}$ was unconstrained the problem would be that of fitting a normal distribution $\mathcal{N}(\mathbf{b}, \mathbf{\Sigma})$ to a sample $\{\mathbf{t}_n\}_{n=1}^N$. In this case, the maximum likelihood estimators are the natural ones: the sample mean $\mathbf{b} = \bar{\mathbf{t}} = \frac{1}{N}\sum_{n=1}^N \mathbf{t}_n$ and the sample covariance $\mathbf{\Sigma} = \mathbf{S} = \frac{1}{N}\sum_{n=1}^N \mathbf{t}_n\mathbf{t}_n^T = \frac{1}{N}\mathbf{T}\mathbf{T}^T$ (see [9] for a proof). It is easy to see (appendix C) that if both $\mathbf{W}$ and $\mathbf{\Sigma}_Y$ are unconstrained, there are infinitely many solutions $\{\hat{\mathbf{W}}, \hat{\mathbf{\Sigma}}_Y\}$ that exactly reconstruct $\mathbf{S}$: the model is not *identified* (a problem well known in factor analysis). To obtain meaningful estimates of $\mathbf{W}$ and $\mathbf{\Sigma}_Y$ it is necessary to impose some constraints on them[7] so that there are more parameters than equations. Here we will consider a full-rank[8] matrix $\mathbf{W}$, but otherwise unrestricted, and several constrained forms for $\mathbf{\Sigma}_Y$, but always definite or semidefinite positive. We refer the reader to the specialised literature on factor analysis for a more comprehensive treatment of the problems of identification and estimation [1, 6, 9]. We also take $\mathbf{b} = \mathbf{0}$ and assume centred data, which is equivalent to estimating $\mathbf{b}$ by the data mean $\bar{\mathbf{t}}$.

2.3 Diagonal covariance: Gaussian with uncorrelated variables

From a practical point of view, in order to compute the gradient of the log-likelihood we need to invert $\mathbf{A}$ analytically. One way to do this is to select $\mathbf{W}$ and $\mathbf{\Sigma}_Y$ such that $\mathbf{W}^T\mathbf{\Sigma}_Y^{-1}\mathbf{W} = \operatorname{diag}\left(R_i^2\right)$. Taking $\mathbf{\Sigma}_Y = \operatorname{diag}\left(\sigma_i^2\right)$ this condition reduces to:

$$(\mathbf{W}^T\mathbf{\Sigma}_Y^{-1}\mathbf{W})_{ij} = \sum_{k=1}^D \frac{w_{ki}w_{kj}}{\sigma_k^2} = R_i^2\delta_{ij} \Rightarrow \begin{cases} \sum_{k=1}^D \frac{w_{ki}w_{kj}}{\sigma_k^2} = 0 & i \neq j \\ \sum_{k=1}^D \left(\frac{w_{ki}}{\sigma_k}\right)^2 = R_i^2 & i = j \end{cases}$$

[3] We have assumed that this covariance is fixed, as part of the prior in latent space, but it could be included in the model parameters θ and take part in the optimisation process as well.

[4] They can be further simplified without loss of generality by centring the data and disregarding the $\mathbf{b}$ parameters, because the maximum likelihood estimator for $\mathbf{b}$ is the data mean (see section 2.2).

[5] Factors which are uncorrelated with each other and have unit variance.

[6] In factor analysis, other criteria can be used for the estimation, such as unweighted least squares or generalised least squares.

[7] This can also ease the estimation procedure.

[8] Defective-rank matrices are due to correlated latent variables, which are redundant and without practical interest.

which means that $\{\mathbf{w}_i\}_{i=1}^L$, the column vectors of $\mathbf{W}$, have to lie on a hyperellipsoid centred in the origin with semiaxes $\{R_i\sigma_k\}_{k=1}^D$ parallel to the coordinate directions and be mutually orthogonal with respect to the metric of $\mathbf{\Sigma}_Y^{-1}$, i.e. $\{\mathbf{w}_i\}_{i=1}^L$ must be *conjugate axes* of that hyperellipsoid. In that case, the model reduces to:

$$\mathbf{\Sigma}_Y = \operatorname{diag}\left(\sigma_i^2\right) \qquad \mathbf{A} = \operatorname{diag}\left(1 + R_i^2\right) \tag{6}$$

$$p(\mathbf{t}|\mathbf{W}, \sigma_1^2, \ldots \sigma_D^2) = (2\pi)^{-D/2} \left(\prod_{i=1}^D \sigma_i^2\right)^{-\frac{1}{2}} \left(\prod_{i=1}^L (1 + R_i^2)\right)^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{t}^T\left(\mathbf{\Sigma}_Y^{-1} - \mathbf{\Sigma}_Y^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{W}^T\mathbf{\Sigma}_Y^{-1}\right)\mathbf{t}}.$$

In appendix D it is shown that, if the R_i are constant, maximum likelihood leads to the following system of coupled equations:

$$\mathbf{T}\mathbf{T}^T\mathbf{\Sigma}_Y^{-1}\mathbf{W} = \mathbf{0} \qquad \operatorname{diag}\left(-\frac{N}{2}\mathbf{\Sigma}_Y + \frac{1}{2}\mathbf{T}\mathbf{T}^T - \mathbf{W}\mathbf{A}^{-1}\mathbf{W}^T\mathbf{\Sigma}_Y^{-1}\mathbf{T}\mathbf{T}^T\right) = \mathbf{0} \tag{7}$$

together with the constraints $\mathbf{A} = \operatorname{diag}\left(1 + R_i^2\right)$ and $\mathbf{\Sigma}_Y = \operatorname{diag}\left(\sigma_i^2\right)$, which can be solved iteratively. Geometrically, the equation $\mathbf{T}\mathbf{T}^T\mathbf{\Sigma}_Y^{-1}\mathbf{W} = \mathbf{0}$ means that $L \leq D - \operatorname{rank}(\mathbf{T})$ and $\{\mathbf{w}'_i\}_{i=1}^L \subset \operatorname{span}\{\mathbf{T}\}^\perp$, where $w'_{ij} = w_{ij}/\sigma_j^2$.

2.4 Covariance proportional to the identity matrix: spherical Gaussian

If we further simplify equations (6) by taking $\mathbf{\Sigma}_Y = \sigma^2\mathbf{I}_D$ (i.e. the noise model is a spherical Gaussian) and $R_i = R = 1/\sigma$, $i = 1, \ldots, L$ (i.e. $\mathbf{W}^T\mathbf{W} = \mathbf{I}_L$: $\{\mathbf{w}_i\}_{i=1}^L$ are orthonormal), the model becomes:

$$\mathbf{\Sigma}_Y = \sigma^2\mathbf{I}_D \qquad \mathbf{A} = \left(\frac{\sigma^2 + 1}{\sigma^2}\right)\mathbf{I}_L \tag{8}$$

$$p(\mathbf{t}|\mathbf{W}, \sigma^2) = (2\pi)^{-D/2}\sigma^{L-D}(\sigma^2 + 1)^{-L/2}e^{-\frac{1}{2\sigma^2}\mathbf{t}^T\left(\mathbf{I}_D - \frac{\mathbf{W}\mathbf{W}^T}{\sigma^2+1}\right)\mathbf{t}}. \tag{9}$$

In this case (see appendix E for details), maximum likelihood estimation has an analytic solution[9]: $L \leq D - \operatorname{rank}(\mathbf{T})$, $\{\mathbf{w}_i\}_{i=1}^L$ orthonormal $\subset \operatorname{span}\{\mathbf{T}\}^\perp$ and

$$\sigma^2 = \frac{-\left(1 - \frac{L}{D} - d\right) + \sqrt{\left(1 - \frac{L}{D} - d\right)^2 + 4d}}{2}$$

where $d = \frac{1}{ND}\sum_{n=1}^N \mathbf{t}_n^T\mathbf{t}_n = \frac{1}{D}\operatorname{tr}(\mathbf{S})$ is the "average square norm of the data per dimension." In the particular case $L = 0$ (no latent variables!) $\sigma^2 = d$.

The posterior distribution in latent space $p(\mathbf{x}|\mathbf{t}, \mathbf{W}, \sigma^2)$ becomes $\mathcal{N}(\boldsymbol{\mu}_X, \mathbf{A}^{-1})$ with $\boldsymbol{\mu}_X = \frac{1}{\sigma^2+1}\mathbf{W}^T\mathbf{t}$ and $\mathbf{A} = \left(\frac{\sigma^2+1}{\sigma^2}\right)\mathbf{I}_L$. We notice the following two limit cases:

- Small variance, $\sigma \to 0$: $p(\mathbf{x}|\mathbf{t}, \mathbf{W}, \sigma^2) \to \delta_L(\mathbf{x} - \mathbf{W}^T\mathbf{t})$, i.e. $\mathbf{W}^T\mathbf{t}$, the components of the projection of $\mathbf{t}$ on $\{\mathbf{w}_i\}_{i=1}^L$, is the only point in latent space "responsible" for $\mathbf{t}$.

- Large variance, $\sigma \to \infty$: $p(\mathbf{x}|\mathbf{t}, \mathbf{W}, \sigma^2) \to \mathcal{N}(\mathbf{0}, \mathbf{I}_L) \equiv p(\mathbf{x})$, i.e. the model does not improve what was known (an undesirable situation).

Observe that, for any point $\mathbf{t}_n$ of the data sample, its posterior distribution in latent space is $\mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$, because $\mathbf{W}^T\mathbf{t}_n = \mathbf{0}$, $n = 1, \ldots, N$: the "point in latent space most responsible for the sample" is the origin.

The interpretation of the optimal solution is striking: the subspace spanned by the mapping must be included in the subspace orthogonal to the one spanned by the data sample. As a consequence, the larger the linear dimension of the sample subspace, the fewer latent variables are available to "explain" it!

[9]Thus not suffering from the curse of the dimensionality at all!

2.5 Null covariance: delta distribution

Assume centred data for simplicity, so that $\mathbf{b} = \mathbf{0}$. If now $\mathbf{\Sigma}_Y = \mathbf{0}$, the noise model becomes a Dirac delta centred in the mapped point, $p(\mathbf{t}|\mathbf{x}, \theta) = p(\mathbf{t}|\mathbf{x}, \mathbf{W}) \sim \delta_D(\mathbf{t} - \mathbf{W}\mathbf{x})$, and therefore

$$p(\mathbf{t}|\theta) = p(\mathbf{t}|\mathbf{W}) = \int p(\mathbf{x})\delta_D(\mathbf{t} - \mathbf{W}\mathbf{x})\, d\mathbf{x} = \begin{cases} 0 & \mathbf{t} \neq \mathbf{W}\mathbf{x} \quad \forall \mathbf{x} \in \mathbb{R}^L \\ p((\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{t}) & \text{otherwise} \end{cases}$$

because if there is a point $\mathbf{x}$ in latent space that maps onto $\mathbf{t}$, then: $\mathbf{W}\mathbf{x} = \mathbf{t} \Rightarrow \mathbf{W}^T\mathbf{W}\mathbf{x} = \mathbf{W}^T\mathbf{t} \Rightarrow \mathbf{x} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{t}$ (remember that $\mathbf{W}^T\mathbf{W}$ is nonsingular if $\mathbf{W}$ is not rank defective). Notice that $p(\cdot)$ here is the prior distribution in latent space.

Now if span $\{\mathbf{T}\} \nsubseteq$ span $\{\mathbf{W}\}$, at least one of the data points will have $p(\mathbf{t}_n|\mathbf{W}) = 0$ and therefore $L(\mathbf{W}) = 0$; this is the case $L <$ rank $(\mathbf{T})$ for any $\mathbf{W}$. Thus, we are left with the trivial case $L \geq$ rank $(\mathbf{T})$, in which the maximum likelihood estimator for $\mathbf{W}$ is obviously the one that recovers the sample covariance matrix $\mathbf{S}$ exactly (see section 2.2), since now $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^T = \mathbf{S}$. This estimator is $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{G}^T$, where $\mathbf{G}_{L\times L}$ is any orthogonal matrix and $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the spectral decomposition of $\mathbf{S}$, i.e. $\mathbf{U} = (\mathbf{u}_1 \dots \mathbf{u}_L)$ with $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_L) > 0$ contain the eigenvectors and eigenvalues of $\mathbf{S}$, respectively. This is the **principal factors** solution.

If instead of the maximum likelihood criterion we use the least squares one ($\min_{\mathbf{W}} \|\mathbf{S} - \mathbf{W}\mathbf{W}^T\|^2$), we obtain $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{G}^T$ again, but now L can take any value and $\mathbf{\Lambda}$ and $\mathbf{U}$ are correspondingly truncated to the largest L eigenvalues and associated eigenvectors, respectively.

The posterior distribution in latent space becomes $p(\mathbf{x}|\mathbf{t}, \mathbf{W}) \sim \delta_L(\mathbf{t} - \mathbf{W}\mathbf{x})$, which means that if $\mathbf{t} \in$ span $\{\mathbf{W}\}$ then there is a unique point in latent space $\mathbf{x} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{t}$ "responsible" for $\mathbf{t}$ (the *inverse* of $\mathbf{t}$ by $\mathbf{W}$); but if $\mathbf{t} \notin$ span $\{\mathbf{W}\}$, no point in latent space can explain $\mathbf{t}$.

3 Conclusions

MacKay's density networks offer a framework that can give rise to a number of methods for dimensionality reduction by selecting different functional classes for the three basic pieces that make them up: the prior distribution in latent space, the mapping between the latent space and the data space and the noise model in data space. Unfortunately, the possibilities of obtaining general, exact solutions of the equations involved are very limited, due to the complexity of the marginalisation over the latent variables and to the computation of the log-likelihood gradient. We have developed the method for, perhaps, the only combination for which analytic treatment is feasible: Gaussian distributions and a linear mapping. In this case, the density network optimal solution corresponds to factor analysis by maximum likelihood, but constraints over the solution are necessary to identify the problem and obtain meaningful results. We have given expressions for the optimal parameters (sometimes in implicit form) assuming several convenient constraints. These expressions should not be taken as a replacement for other general methods for maximum likelihood factor analysis (e.g. the Jöreskog method [7]).

When the noise model tends to a delta function located at the mapped point, the density network is equivalent to principal factor analysis, where the optimal mapping is the projection over the first L principal components of the data.

In order to use nonlinear mappings, integral (2) must be approximated. An obvious way to do this, proposed by MacKay [8], are Monte Carlo methods, by which we sample the prior distribution in latent space $p(\mathbf{x})$ at points $\{\mathbf{x}_r\}_{r=1}^R$ and obtain

$$l(\theta) = \sum_{n=1}^N \ln \int p(\mathbf{t}_n|\mathbf{x}, \theta)p(\mathbf{x})\, d\mathbf{x} \approx \sum_{n=1}^N \ln \frac{1}{R} \sum_{r=1}^R p(\mathbf{t}_n|\mathbf{x}_r, \theta)$$

$$\nabla_\theta l(\theta) \approx \sum_{n=1}^N \frac{\sum_{r=1}^R \nabla_\theta p(\mathbf{t}_n|\mathbf{x}_r, \theta)}{\sum_{r=1}^R p(\mathbf{t}_n|\mathbf{x}_r, \theta)}$$

for integral (2) and the log-likelihood gradient. Because the functional form of $p(\mathbf{t}|\mathbf{x}, \theta)$ is known, so is that of the gradient, and it is possible —at least in theory— to find its stationary points. The disadvantage of this procedure is that the number of points R grows exponentially with the number of latent variables L. An application of these ideas is Bishop *et al.*'s generative topographic mapping (GTM) [3], in which the prior in latent space is a discrete grid of points, the noise model a spherical Gaussian and a nonlinear mapping is implemented via a generalised linear model. Both the number of points in the grid and the number of radial basis functions in the generalised linear model grow exponentially with L, which in practice limits the model to not more than 3 latent variables.

A Notation

We represent scalars in italics, vectors by boldface lowercase letters and matrices by boldface uppercase letters. $\mathbf{I}_L$ is the identity matrix in L dimensions and $\mathbf{0}$ represents the vector or matrix (depending on the context) with null elements. $|\mathbf{A}|$ is the determinant of a square matrix $\mathbf{A}$ and $\|\mathbf{v}\|_\infty = \max_i |v_i|$. $\operatorname{span}\{\mathbf{A}\}$ is the subspace spanned by the column vectors of $\mathbf{A}$ and $\operatorname{span}\{\mathbf{A}\}^\perp$ the orthogonal subspace to it. $\delta_L(\mathbf{x})$ represents the delta distribution in L dimensions and $\mathcal{N}_L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the normal distribution in L dimensions with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

B Uniform distributions

Consider a D-dimensional uniform distribution in a hypercube of side $\sqrt{2v}$ centred on $\boldsymbol{\mu}$; then its density is $(12v)^{-D/2}$ for points inside the hypercube and 0 otherwise, and its covariance matrix is $v\mathbf{I}_D$. We represent it as $U_D(\boldsymbol{\mu}, v)$.

Assume the prior distribution in latent space is uniform in the hypercube $[0,1]^L$ and consider the following two particular cases for the noise model:

- Uniform noise model $U_D(\mathbf{y}(\mathbf{x};\theta'), v)$ for parameters $\theta = \{\theta', v\}$; then:

$$p(\mathbf{t}|\theta) = \int p(\mathbf{t}|\mathbf{x},\theta)p(\mathbf{x})\, d\mathbf{x} = (12v)^{-D/2}\operatorname{vol}(\mathcal{R})$$

  where $\mathcal{R}(\theta) = \{\mathbf{x} \in [0,1]^L : \|\mathbf{t} - \mathbf{y}(\mathbf{x};\theta')\|_\infty \leq \sqrt{3v}\}$ is a region in latent space and $\operatorname{vol}(\mathcal{R}) = \int_{\mathcal{R}} d\mathbf{x}$ its volume. It is difficult to simplify $p(\mathbf{t}|\theta)$ for any class of mappings and since its functional dependence on θ is unknown we can't compute its gradient. Monte Carlo methods will not help here because they don't give the functional form of $p(\mathbf{t}|\theta)$ either.

- Normal noise model $\mathcal{N}_D(\mathbf{W}\mathbf{x}, v\mathbf{I}_D)$ for parameters $\theta = \{\mathbf{W}, v\}$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}_L$; then:

$$p(\mathbf{t}|\theta) = \int_{[0,1]^L} (2\pi)^{-D/2} v^{-D/2} e^{-\frac{1}{2v}(\mathbf{t}-\mathbf{W}\mathbf{x})^T(\mathbf{t}-\mathbf{W}\mathbf{x})}\, d\mathbf{x} =$$
$$= (2\pi)^{-D/2} v^{-D/2} e^{-\mathbf{t}^T\mathbf{t}/2v} \prod_{i=1}^{L} \int_0^1 e^{-(x_i^2 - 2(\mathbf{t}^T\mathbf{W})_i x_i)/2v}\, dx_i$$

  which can be expressed in terms of the function $\operatorname{erf}(x) = \int_0^x e^{-u^2/2}\, du$, but further analytical treatment becomes impossible.

We see that the use of uniform distributions simplifies the integrand in (2) at the cost of complicating the integration region. As a result, it is not possible to obtain a closed-form expression for $L(\theta)$.

C Likelihood of the parameters for Gaussian distributions

If the prior on latent space is $\mathcal{N}_L(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$:

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

and the noise model is $\mathcal{N}_D(\mathbf{W}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_Y)$:

$$p(\mathbf{t}|\mathbf{x};\theta) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}_Y|^{-1/2} e^{-\frac{1}{2}(\mathbf{t}-\mathbf{W}\mathbf{x}-\mathbf{b})^T\boldsymbol{\Sigma}_Y^{-1}(\mathbf{t}-\mathbf{W}\mathbf{x}-\mathbf{b})}$$

then, using the following formula for the Gaussian integral (where $\mathbf{w}$ is a W-dimensional vector)

$$\int e^{-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}+\mathbf{h}^T\mathbf{w}}\, d\mathbf{w} = (2\pi)^{-W/2}|\mathbf{A}|^{-1/2} e^{\frac{1}{2}\mathbf{h}^T\mathbf{A}^{-1}\mathbf{h}}$$

we obtain:

$$p(\mathbf{t}|\theta) = \int p(\mathbf{t}|\mathbf{x},\theta)p(\mathbf{x})\,d\mathbf{x} =$$
$$(2\pi)^{-(D+L)/2}|\boldsymbol{\Sigma}_X|^{-1/2}|\boldsymbol{\Sigma}_Y|^{-1/2}e^{-\frac{1}{2}(\boldsymbol{\mu}_X^T\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X+(\mathbf{t}-\mathbf{b})^T\boldsymbol{\Sigma}_Y^{-1}(\mathbf{t}-\mathbf{b}))}\times$$
$$\times\int e^{-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Sigma}_X^{-1}+\mathbf{W}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{W})\mathbf{x}+(\boldsymbol{\mu}_X^T\boldsymbol{\Sigma}_X^{-1}+(\mathbf{t}-\mathbf{b})^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{W})\mathbf{x}}\,d\mathbf{x} =$$
$$(2\pi)^{-D/2}|\boldsymbol{\Sigma}_X|^{-1/2}|\boldsymbol{\Sigma}_Y|^{-1/2}|\mathbf{A}|^{-1/2}e^{-\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})}e^{-\frac{1}{2}(\boldsymbol{\mu}_X^T(\boldsymbol{\Sigma}_X^{-1}-\boldsymbol{\Sigma}_X^{-1}\mathbf{A}^{-1}\boldsymbol{\Sigma}_X^{-1})\boldsymbol{\mu}_X-(\boldsymbol{\mu}-\mathbf{b})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{b}))}$$

where $\mathbf{A} = \boldsymbol{\Sigma}_X^{-1}+\mathbf{W}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{W}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_Y^{-1} - \boldsymbol{\Sigma}_Y^{-1}\mathbf{W}\mathbf{A}^{-1}\mathbf{W}^T\boldsymbol{\Sigma}_Y^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_Y^{-1}\mathbf{W}\mathbf{A}^{-1}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X + \mathbf{b}$. Using the Sherman-Morrison-Woodbury formula for the inverse of a sum:

$$(\mathbf{P}+\mathbf{QRS})^{-1} = \mathbf{P}^{-1} - \mathbf{P}^{-1}\mathbf{Q}(\mathbf{R}^{-1}+\mathbf{SP}^{-1}\mathbf{Q})^{-1}\mathbf{SP}^{-1}$$

and the following formula for the determinant of a sum:

$$|\mathbf{P}+\mathbf{QS}| = |\mathbf{P}||\mathbf{I}_p+\mathbf{P}^{-1}\mathbf{QS}| = |\mathbf{P}||\mathbf{I}_q+\mathbf{SP}^{-1}\mathbf{Q}|$$

where the matrices are $\mathbf{P}_{p\times p}$, $\mathbf{Q}_{p\times q}$, $\mathbf{R}_{q\times q}$ and $\mathbf{S}_{q\times p}$, it is now easy (but tedious) to prove that the second exponential term reduces to 1 and that $|\boldsymbol{\Sigma}_X|^{-1/2}|\boldsymbol{\Sigma}_Y|^{-1/2}|\mathbf{A}|^{-1/2} = |\boldsymbol{\Sigma}|^{-1/2}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_Y + \mathbf{W}\boldsymbol{\Sigma}_X\mathbf{W}^T = \boldsymbol{\Sigma}_Y(\boldsymbol{\Sigma}_Y - \mathbf{W}\mathbf{A}^{-1}\mathbf{W}^T)^{-1}\boldsymbol{\Sigma}_Y$. Therefore $p(\mathbf{t}|\theta) \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. In a similar way we can prove that the posterior in latent space is $p(\mathbf{x}|\mathbf{t},\theta) \sim \mathcal{N}(\boldsymbol{\mu}'_X, \mathbf{A}^{-1})$, where $\boldsymbol{\mu}'_X = \mathbf{A}^{-1}\mathbf{W}^T\boldsymbol{\Sigma}_Y^{-1}(\mathbf{t}-\mathbf{b}) + \mathbf{A}^{-1}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X$. Equations (4) follow immediately for $\boldsymbol{\mu}_X = \mathbf{0}$ (with a slight change of notation).

Now we show that any symmetric, positive definite matrix can be decomposed as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_Y + \mathbf{W}\boldsymbol{\Sigma}_X\mathbf{W}^T$ in an infinite number of ways satisfying $\boldsymbol{\Sigma}_Y > 0$ and rank $(\mathbf{W}) = L$ for fixed $\boldsymbol{\Sigma}_X > 0$. For $\epsilon > 0$ we take $\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma} - \epsilon\begin{pmatrix}\mathbf{I}_L & 0\\ 0 & 0\end{pmatrix}$, which can always be made positive definite by choosing ϵ small enough:

$$\|\mathbf{v}\| = 1 : 0 < \mathbf{v}^T\boldsymbol{\Sigma}_Y\mathbf{v} = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \epsilon\sum_{i=1}^{L} v_i^2 \le \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - \epsilon \Longrightarrow \text{ take } \epsilon < \min_{\|\mathbf{u}\|=1} \mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}.$$

Then $\mathbf{W}\boldsymbol{\Sigma}_X\mathbf{W}^T = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_Y = \epsilon\begin{pmatrix}\mathbf{I}_L & 0\\ 0 & 0\end{pmatrix}$ and we can take $\mathbf{W} = \sqrt{\epsilon}\begin{pmatrix}\mathbf{I}_L\\ 0\end{pmatrix}\boldsymbol{\Sigma}_X^{-1/2}$, with rank L.

D Diagonal covariance: uncorrelated Gaussian

Taking partial derivatives with respect to the parameters of the log-likelihood

$$l(\mathbf{W},\sigma_1^2,\dots,\sigma_D^2) = \sum_{n=1}^{N}\ln p(\mathbf{t}|\mathbf{W},\sigma_1^2,\dots\sigma_D^2)$$

and taking into account equations (6) and (7) we obtain:

$$\frac{\partial l}{\partial w_{ij}} = 0 \Longrightarrow \sum_{n=1}^{N}\frac{\partial}{\partial w_{ij}}\left\{\sum_{k=1}^{D}\frac{t_{kn}^2}{\sigma_k^2} - \sum_{k=1}^{D}\sum_{l=1}^{D}\sum_{m=1}^{L}\frac{t_{kn}t_{ln}w_{km}w_{lm}}{\sigma_k^2\sigma_l^2(R_m^2+1)}\right\} =$$
$$-2\sum_{n=1}^{N}\sum_{k=1}^{D}\frac{t_{in}t_{kn}w_{kj}}{\sigma_i^2\sigma_k^2(R_j^2+1)} = 0 \Longrightarrow \mathbf{TT}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{W} = \mathbf{0}$$

and

$$\frac{\partial l}{\partial v_i} = 0 \Longrightarrow \sum_{n=1}^{N}\left\{-\frac{1}{2v_i} + \frac{t_{in}^2}{2v_i^2} - \sum_{k=1}^{D}\sum_{m=1}^{L}\frac{t_{kn}t_{in}w_{km}w_{im}}{v_k v_i^2(R_m^2+1)}\right\} = 0 \Longrightarrow$$
$$-\frac{N}{2}v_i + \frac{1}{2}\sum_{n=1}^{N}t_i^2 n - \sum_{k=1}^{D}\sum_{m=1}^{L}\frac{t_{kn}t_{in}w_{km}w_{im}}{v_k(R_m^2+1)} = 0 \Longrightarrow$$
$$\operatorname{diag}\left(-\frac{N}{2}\boldsymbol{\Sigma}_Y + \frac{1}{2}\mathbf{TT}^T - \mathbf{W}\mathbf{A}^{-1}\mathbf{W}^T\boldsymbol{\Sigma}_Y^{-1}\mathbf{TT}^T\right) = \mathbf{0}$$

where $v_i = \sigma_i^2$ and assuming R_i, $i = 1,\dots,L$ constant.

E Covariance proportional to identity: spherical Gaussian

Taking partial derivatives with respect to the parameters of the log-likelihood

$$l(\mathbf{W},\sigma^2)=\sum_{n=1}^{N}\ln p(\mathbf{t}|\mathbf{W},\sigma^2)$$

and taking into account equations (8) and (9) we obtain:

$$\frac{\partial l}{\partial w_{ij}}=0\Longrightarrow\sum_{n=1}^{N}\frac{\partial}{\partial w_{ij}}\left\{\frac{1}{\sigma^2}\sum_{k=1}^{D}t_{kn}^2-\frac{1}{\sigma^2(\sigma^2+1)}\sum_{k=1}^{D}\sum_{l=1}^{D}\sum_{m=1}^{L}t_{kn}t_{ln}w_{km}w_{lm}\right\}=$$
$$-2\frac{1}{\sigma^2(\sigma^2+1)}\sum_{n=1}^{N}\sum_{k=1}^{D}t_{in}t_{kn}w_{kj}=0\Longrightarrow\mathbf{T}\mathbf{T}^T\mathbf{W}=\mathbf{0}$$

and

$$\frac{\partial l}{\partial v}=0\Longrightarrow\sum_{n=1}^{N}\left\{\frac{\frac{\partial}{\partial v}\left(v^{(L-D)/2}(v+1)^{-L/2}\right)}{v^{(L-D)/2}(v+1)^{-L/2}}+\frac{\partial}{\partial v}\left(-\frac{1}{2v}\mathbf{t}_n^T\left(\mathbf{I}_D-\frac{\mathbf{W}\mathbf{W}^T}{v+1}\right)\mathbf{t}_n\right)\right\}=$$
$$\sum_{n=1}^{N}\left\{\frac{L-D}{2v}-\frac{L}{2(v+1)}+\frac{1}{2v^2}\mathbf{t}_n^T\mathbf{t}_n\right\}=0\Longrightarrow v^2+\left(1-\frac{L}{D}-d\right)v-d=0$$

where we have applied the fact that $\mathbf{T}\mathbf{T}^T\mathbf{W}=\mathbf{0}\Longrightarrow\mathbf{W}^T\mathbf{t}_n=\mathbf{0}$, $n=1,\dots,N$ and $d=\frac{1}{ND}\sum_{n=1}^{N}\mathbf{t}_n^T\mathbf{t}_n$. Thus, we arrive to a second-degree equation on v that has as solutions:

$$v=\frac{-\left(1-\frac{L}{D}-d\right)\pm\sqrt{\left(1-\frac{L}{D}-d\right)^2+4d}}{2}$$

of which one is always positive (for nontrivial data) and the other always negative. Only the positive one is meaningful, because $v=\sigma^2$ must not be negative.

Observe that this result cannot be derived from the result of appendix D, where we supposed R_i constant; here $R\sigma=1$.

References

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

[2] C. M. Bishop, M. Svensén, and C. K. I. Williams, *EM optimization of latent-variable density models*, in Advances in Neural Information Processing Systems, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., vol. 8, MIT Press, Cambridge, MA, 1996, pp. 465–471.

[3] ——, *GTM: A principled alternative to the self-organising map*, in Advances in Neural Information Processing Systems, M. C. Mozer, M. I. Jordan, and T. Petsche, eds., vol. 9, MIT Press, Cambridge, MA, 1997.

[4] M. Á. Carreira Perpiñán, *A review of dimension reduction techniques*, Tech. Rep. CS–96–09, Dept. of Computer Science, University of Sheffield, Dec. 1996.

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

[6] H. H. Harman, *Modern Factor Analysis*, University of Chicago Press, Chicago, second ed., 1967.

[7] K. G. Jöreskog, *A general approach to confirmatory maximum likelihood factor analysis*, Psychometrika, 34 (1969), pp. 183–202.

[8] D. J. C. MacKay, *Bayesian neural networks and density networks*, Nuclear Instruments and Methods in Physics Research A, 354 (1995), pp. 73–80.

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