1 Curved Gaussian Models

Juan K Lin & Peter Dayan

In principal component analysis, data are modeled with arbitrary linear transformations of a spherical Gaussian. In this paper, we present a simple extension to generalised linear transformations, which allow for an efficient local description of curvature. In a predetermined coordinate system, direct analytic computation of the curvature parameters is possible; otherwise, the coordinate system can also be inferred.

1.1 Introduction

Gaussian models for data, as found in standard density models, principal components analysis, factor analysis, are key *building blocks* for more complicated density estimation models. They have numerous advantages, notably including the simplicity with which they can be fit to data. There is a plethora of more complicated density estimation edifices built using these blocks, including simple mixtures of either radial Gaussians or Gaussians with arbitary covariance matrices (*eg* Nowlan, 1991); and, in more recent work, mixtures of principal component or factor analysers (Bregler & Omohundro, 1995; Kambhatala & Leen, 1994; Hinton, Dayan & Revow, 1997; Roweis, 1998; Tipping & Bishop, 1997).

Figure 1.1 illustrates one rationale underlying the use of mixture models. Figure 1.1A shows 1769 two dimensional data points from a particular financial time series. One reasonable way to characterise these points is as lying near to a one-dimensional curved manifold. In this representation, two dimensional Gaussian models are ellipses (assuming that the full covariance matrices are modeled). Figures 1.1B;C show two alternative fits of four two-dimensional Gaussians to the data (using standard expectation maximisation algorithm, Dempster, Laird & Rubin, 1977, to fit mixtures of Gaussian). The ellipses show $\sqrt{2}\sigma$ contours for the Gaussians; the widths of the outlines are proportional to the prior probabilities (mixing proportions) of the Gaussians. These Gaussian construct the manifold by 'bolting' pieces together – each piece is strongly elliptical, matching the one-dimensional structure of the data.

These Gaussian models capture part of the essence of the data, and indeed, had we used more components, would have fit them even better. However, they miss the simple observation that the data seem to lie on a curved manifold. In this paper, we consider ways of changing the Gaussian building block so that it can capture curvatures of this form, whilst retaining as much of its computational attractiveness as possible.

Section 1.2 discusses the basic curved Gaussian model; section 1.3 considers the

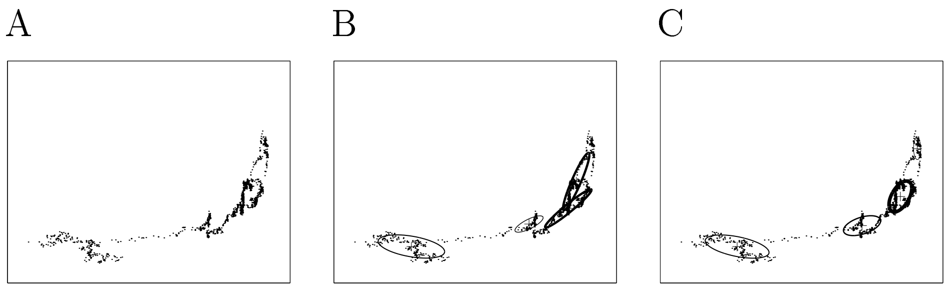

Figure 1.1
Mixture of Gaussians models for curved data. A) Daily values of the Japanese Yen *versus* the Korean Won measured in US Dollar units from 23rd October 1993 to 27th August 1998 normalised relative to their values at the end. B;C) Two different mixtures of Gaussian fits to these data using four full covariance two-dimensional Gaussians. The centers of the Gaussians are marked by a cross, $\sqrt{2}\sigma$ values by the ellipses, and the mixing proportions by the widths of the outlines of the ellipses.

problem of inferring appropriate coordinate systems in which to describe the curvature. Relationships with methods such as principal curves (Hastie & Tibshirani, 19??) are described in section 1.4.

1.2 THE BASIC BUILDLING BLOCK

1.2.1 Two-dimensions

One way to write a Gaussian distribution $\mathcal{N}[0, \Sigma]$ over two variables $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$ is as:

$$p(x_1, x_2) \propto \sqrt{|\Omega|} e^{-\mathbf{x^T \Omega x/2}} = \sqrt{w_1} e^{-w_1 \phi_1^2/2} \sqrt{w_2} e^{-w_2 \varphi_2^2/2}, \qquad \text{where}$$

$\Omega = \Sigma^{-1}$, and

$$\varphi_1(\mathbf{x}) = x_1 \ , \tag{1.1}$$

$$\varphi_2(\mathbf{x}) = x_2 + k_{21} x_1 \ , \tag{1.2}$$

$$w_1 = \Omega_{11} - \Omega_{12}^2/\Omega_{22} \quad , \quad w_2 = \Omega_{22} \quad , \quad k_{21} = \Omega_{12}/\Omega_{22} \ . \tag{1.3}$$

The key facet of the change of coordinates from x_1, x_2 to φ_1, φ_2 is the linear dependence of the mean of φ_2 on x_1 (and hence φ_1). Figure 1.2A shows an example of the contours of such a distribution (the ellipses), and the underlying linear

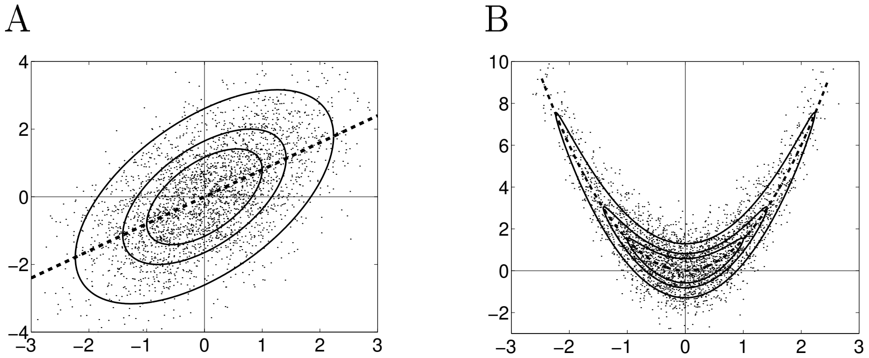

Figure 1.2
x_1, x_2 plots of conventional and curved Gaussians. A) Conventional two-dimensional Gaussian (in x_1, x_2 space) with covariance matrix $\Sigma = \begin{pmatrix} 1 & .8 \\ .8 & 2 \end{pmatrix}$. The ellipses are $\sigma, \sqrt{2}\sigma$ and $\sqrt{5}\sigma$ contours; the dashed line is $x_2 = 0.8x_1$ as in equation 1.2; 3000 points were drawn from the distribution. B) Curved two-dimensional Gaussian with $w_1 = 1$, $w_2 = 3$ and $k_{21} = 1.5$. The ellipses are $\sigma, \sqrt{2}\sigma$ and $\sqrt{5}\sigma$ contours; the dashed line is the skeleton $x_2 = k_{21}x_1^2$; 3000 points were drawn from the distribution.

dependence (the dashed line). Note that the principal axis of the covariance matrix is not necessarily aligned with this linear relation.

This linear change in coordinates also suggests a natural generalisation:

$$\varphi_2 = x_2 - \sum_j k_{2j}\psi_{2j}(x_1) \tag{1.4}$$

where $\psi_{2j}(x)$ are any suitably well-behaved (everywhere finite and differentiable) and *fixed* basis functions of x_1, and k_{2j} are parameters. This is a *generalised linear model* (Hastie & Tibshirani, 19??) for x_2, since the mean depends linearly on the underlying variable parameters (k_{2j}).

One of the simplest possible generalised linear models is the quadratic:

$$\varphi_2 = x_2 k_{21} x_1^2 \tag{1.5}$$

Consider therefore the full Gaussian model:

$$p(\mathbf{x}; \mathbf{w}; \mathbf{k_{21}}) \propto (\mathbf{w_1 w_2})^{\mathbf{1/2}} \mathbf{e}^{-\mathbf{w_1}\varphi_\mathbf{1}^\mathbf{2} - \mathbf{w_2}\varphi_\mathbf{2}^\mathbf{2}}$$

Since the Jacobian of the transformation from x_1, x_2 to φ_1, φ_2 using equation 1.5 in place of equation 1.2 is lower diagonal everywhere, with the diagonal elements consisting of ones, the determinant of the Jacobian is always one. Therefore, this is the probability density function in the new curved coordinate system. More generally, the multivariate statistical models being considered are subsets of volume preserving non–linear transformations of multivariate Gaussian models. Figure 1.2B shows a contour plot of this Gaussian—the 'scaffolding' for x_2 provided by the value of x_1 (the dashed line) is now quadratic rather than linear.

1.2.2 n dimensions

More generally, in the case that the input data $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$ are n-dimensional, consider the following non-linear change of variables:

$$\begin{aligned}
\varphi_1 &= x_1 + k_{10} \\
\varphi_2 &= x_2 + k_{20} + \sum_j k_{2j}\psi_{2j}(x_1) \\
\varphi_3 &= x_3 + k_{30} + \sum_j k_{3j}\psi_{3j}(x_1, x_2) \\
&\vdots \\
\varphi_n &= x_n + k_{n0} + \sum_j k_{3j}\psi_{nj}(x_1, x_2, \ldots, x_{n-1}) \ ,
\end{aligned}$$

together with a likelihood model

$$p(\mathbf{x}; \mathbf{w}, \mathbf{k_1}, \mathbf{k_2}, ..., \mathbf{k_n}) \propto (\mathbf{w_1}...\mathbf{w_n})^{\mathbf{1/2}} \exp[-\mathbf{w_1}\varphi_{\mathbf{1}}^{\mathbf{2}}... - \mathbf{w_{n-1}}\varphi_{\mathbf{n-1}}^{\mathbf{2}} - \mathbf{w_n}\varphi^{\mathbf{2}}], \tag{1.6}$$

where ϕ_{ij} are all differentiable functions of their arguments and k_{ij} are parameters. Again, the determinant of the Jacobian of the transformation is 1, and so the model is well defined. Natural choices for ϕ_{ij} are polynomials of non-negative orders in their parameters, although other choices are possible.

1.2.3 Parameter estimation

Consider fitting a model of the form of section 1.2.2 to data $\mathcal{D} = \{\mathbf{x^m}\}$. Maximising the log likelihood with respect to the model parameters gives optimising equations

$$k_{ij} = \left[\left(\Phi^i\right)^{-1} \mathbf{c^i}\right]_j \tag{1.7}$$

where $\Phi^i_{ab} = \langle\phi_{ia}\phi_{ib}\rangle$, $c^i_a = \langle\phi_{ia}x_i\rangle$, where $\langle\cdot\rangle$ indicates taking an average over $\mathcal{D}$, and, for convenience, we define $\phi_{l0} = 1$ for all l. Equation 1.7 is only valid if matrix Φ^i is invertible; otherwise, the maximum likelihood solution is degenerate

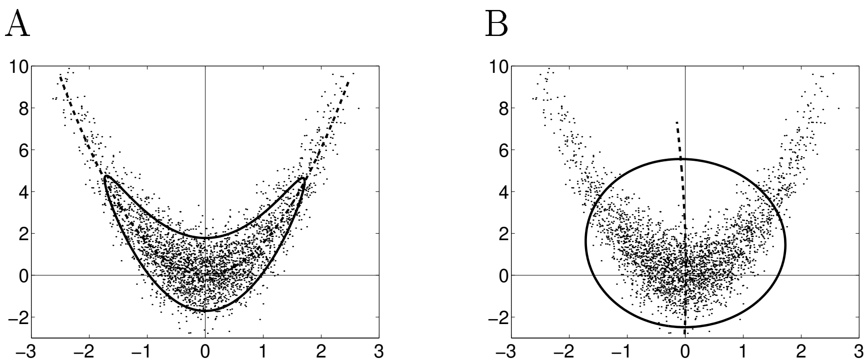

Figure 1.3
Curved Gaussian fits using equations 1.7 and 1.9 to the data in figure 1.2B with $\phi_{21} = x_1^2$. The dashed line shows the quadratic skeleton; the solid line the $\sqrt{1.5}\sigma$ contour. A) Fit using the correct ordering of the variables x_1, x_2. B) Fit using the opposite ordering x_2, x_1.

with whole manifolds of solutions. Equating to zero the partial derivatives of the log likelihood with respect to the inverse variance parameters, we find

$$\frac{\partial L}{\partial w_j} = \frac{1}{2w_j} - \langle \varphi_j^2 \rangle. \tag{1.8}$$

Thus the maximum likelihood solution is

$$w_j = \frac{1}{2\langle \varphi_j^2 \rangle}. \tag{1.9}$$

using the optimal values of k_{ij} given in equation 1.7.

Figure 1.3A shows the maximum likelihood fit of a curved Gaussian to the data in figure 1.2B. The curved model captures the structure of the data points well, as should be expected, given the way that they were generated.

We have therefore suggested a natural curved generalisation of a conventional Gaussian model. Given the coordinate systems (*ie* the *ordering*) of the variables, computing the maximum likelihood values of the parameters of the generalised linear model requires nothing more than matrix inversion—in particular, it is *not* an iterative process. One could use even more general parameterised versions of the functions ϕ_{ij}, and determine the values of the parameters by gradient ascent in the log likelihood.

1.3 COORDINATE TRANSFORMATIONS

Figure 1.3B shows an extreme example of the consequence of using a different coordinate system to fit the curved model – in this case, one that has the direction of the non-linear dependence exactly backwards. In this case, the model clearly misfits the data – doing just as badly, or perhaps even worse, as a non-curved Gaussian model. For fitting general datasets, such as that in figure 1.1, it is therefore necessary to be able to change coordinate systems.

Drawing inspiration from Independent Component Analysis, one way to do this is to consider a curved, source datapoint $\mathbf{x}$ being used to generate an observed datapoint $\mathbf{y}$ through a linear transformation:

$$\mathbf{y} = \mathbf{W^{-1}x} + \mathbf{b} \tag{1.10}$$

where $\mathbf{b}$ is an offset, and $\mathbf{W^{-1}}$ is an invertible 'mixing' transformation. Given a dataset $\mathcal{D} = \{\mathbf{y^m}\}$, the task is to find the unmixing transformations $\mathbf{b}$ and $\mathbf{W}$, together with the parameters $\mathbf{w}$ and k_{ij} of the curved source model so as to fit the data as tightly as possible, by maximising the log likelihood, the contribution to which from $\mathbf{y^m}$ is:

$$\log|\mathbf{W^{-1}}| + \log \mathbf{p}(\mathbf{W^{-1}}(\mathbf{y^m} - \mathbf{b}), \mathbf{k_1}, \mathbf{k_2}, ..., \mathbf{k_n})$$

using the expression in equation 1.6. It does not seem possible, in general, to derive closed-form optima for the likelihood as a function of all the paramters. Instead, we consider a form of coordinate-wise ascent of the likelihood, finding the global optimum in $\mathbf{w}$ and k_{ij} for values of $\mathbf{W}$ and $\mathbf{b}$, and then changing the latter variables by gradient ascent.

The derivatives of the log likelihood with respect to the mixing parameters are straightforward to calculate, albeit rather messy. Note that there is substantial redundancy in the model – $\mathbf{W}$ can be restricted to only $n(n-1)/2$ parameters corresponding to an arbitrary orthogonal matrix with unit determinant and there are also interdependencies between $\mathbf{b}$ and k_{i0} for i. However, the learning rule is considerably simpler with the model parameterization given above. Just as for ICA, we can use the natural gradient ascent algorithm (Amari & Cichocki & Yang 1995, Cardoso & Laheld 1996) to update the mixing transformation along the direction of steepest ascent by right multiplying the partial derivatives of the log likelihood with respect to $\mathbf{W}$ by $\mathbf{W^TW}$. One way to choose initial conditions for the mixing transformation is to apply standard principal components analysis – although this can be problematical depending on the scaling of the variables.

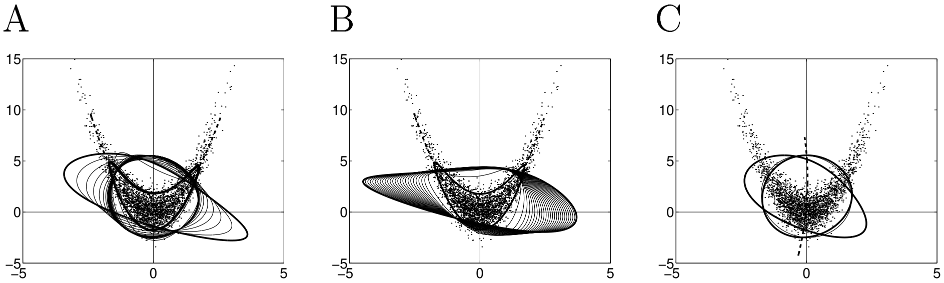

Figure 1.4
Inference of the mixing matrix. All graphs show the starting and ending $\sqrt{1.5}\sigma$ contours, the ending skeleton (dashed line), and intermediate contours every 10 iterations. A;B) Examples of inference to the global optimal coordinate transformation. C) Example of inference to the poor local optimum.

Figures 1.4A-C show examples of the way the algorithm fits the curved Gaussian data. In each case, the initial values of the mixing transformation are arbitrary, and the shape of the $\sqrt{1.5}\sigma$ contour is plotted every 10 iterations (using a rather arbitrary step-size). Figures 1.4A;B show examples of successful inference of the transformation and the curved Gaussian; figure 1.4C shows an example of failure, in which only the local maximum of the likelihood (shown in figure 1.3B) is attained. The absolute values of the likelihood can be used to select between alternative solutions.

Finally, we can apply the full algorithm to the currency data from figure 1.1. Figure **??** shows an example of a full curved Gaussian model for the data. It clearly captures the structure of the data more competently than the mixture of non-curved Gaussians shown in figures 1.1B;C.

1.4 DISCUSSION

We have presented a simple extension to a Gaussian density model that captures weakly curved manifolds. We used a form of generalised linear model for the mean of one latent variable given the values of the other latent variables that precede it in some particular ordering of the latent variables. The form of the non-linear variable transformation chosen is convenient since the Jacobian is guaranteed to be 1. The advantage of such generalised linear models is that there is a simple procedure for fitting them, although more general models also possible. We also showed how to use ideas from independent components analysis to infer the appropriate coordinate

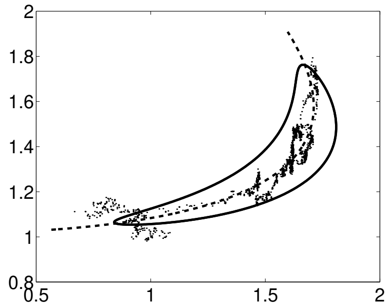

Figure 1.5
Full curved model fit to the currency data from figure 1.1A. The dashed line is the quadratic skeleton; the solid line the $\sqrt{2}\sigma$ contour for the model.

system in which to fit the curved model, although this leads to an algorithm that has to be fit iteratively.

There are various natural extensions of the work. Particularly important is the idea of using *mixtures* of the curved Gaussian distributions, in the same way that one uses mixtures of standard Gaussian distributions. This presents no conceptual or computational hurdle – and can be done simply using the expectation-maximisation algorithm.

Another extension is that the local curvature information could be used for things other than fitting a Gaussian model – for instance, it could be used to enhance local kernel methods in something like a curved form of tangent distance (Simard, Le Cun & Denker, 1993). Also, more general curved models could be used in place of the simple quadratic form we used to fit the data of figure 1.1A. As ever, there are trade-offs of bias for variance in assuming more flexible parameterisations, and more sophisticated cross-validation or Bayesian methods could be considered to choose the orders of the polynomials or the forms of the curvature.

A further possibility is to consider reduced-order, curved, Gaussian models of data (which is closely related to the idea of performing ICA when there are fewer sources than inputs, or in the presence of assumed noise in the transformation from the sources to the inputs, Attias, 19??). Inference of the coordinate system is more difficult in this case, since the equivalent of matrix $\mathbf{W}$ is singular, but standard methods can be applied.