M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 514

Automatic choice of dimensionality for PCA

Thomas P. Minka
MIT Media Laboratory, Vision and Modeling Group
20 Ames Street; Cambridge, MA 02139
tpminka@media.mit.edu
December 29, 2000

Abstract

A central issue in principal component analysis (PCA) is choosing the number of principal components to be retained. By interpreting PCA as density estimation, this paper shows how to use Bayesian model selection to determine the true dimensionality of the data. The resulting estimate is simple to compute yet guaranteed to pick the correct dimensionality, given enough data. The estimate involves an integral over the Steifel manifold of k-frames, which is difficult to compute exactly. But after choosing an appropriate parameterization and applying Laplace's method, an accurate and practical estimator is obtained. In simulations, it is more accurate than cross-validation and other proposed algorithms, plus it runs much faster.

1 Introduction

Principal component analysis (PCA) decomposes high-dimensional data into a low-dimensional subspace component and a noise component. This decomposition is useful for data compression as well as de-noising, making it a common first step for many data processing tasks. Tipping & Bishop (1997b) have shown that PCA can be interpreted as maximum-likelihood density estimation. This paper extends their work by applying Bayesian model selection to the probabilistic PCA model, providing a simple and fast criterion for choosing the dimensionality of the subspace.

2 Probabilistic PCA

This section reviews the results of Tipping & Bishop (1997b). The model is that a high-dimensional random vector $\mathbf{x}$ can be expressed as a linear combination of basis vectors plus noise:

$$\begin{aligned} \mathbf{x} &= \sum_{j=1}^{k} \mathbf{h}_j w_j + \mathbf{m} + \mathbf{e} && (1) \\ &= \mathbf{H}\mathbf{w} + \mathbf{m} + \mathbf{e} && (2) \\ p(\mathbf{e}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}) && (3) \end{aligned}$$

where $\mathbf{x}$ has length d and $\mathbf{w}$ has smaller length k. The vector $\mathbf{m}$ defines the mean of $\mathbf{x}$, while $\mathbf{H}$ and $\mathbf{V}$ define its variance. For PCA, the noise variance $\mathbf{V}$ is spherical:

$$\mathbf{V} = v\mathbf{I}_d \tag{4}$$

And the density of $\mathbf{w}$ is spherical Gaussian:

$$p(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k) \tag{5}$$

This model for PCA was also discussed by Moghaddam & Pentland (1995) and Roweis (1997). It is directly related to factor analysis: the only difference is that, in factor analysis, the noise variance $\mathbf{V}$ is a general diagonal matrix.

The goal of PCA is to estimate the basis vectors $\mathbf{H}$ and the noise variance v from a data set $D = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$. Under the model, the probability of observing a vector $\mathbf{x}$ is

$$\begin{aligned} p(\mathbf{x}|\mathbf{w}, \mathbf{H}, \mathbf{m}, v) &\sim \mathcal{N}(\mathbf{H}\mathbf{w} + \mathbf{m}, v\mathbf{I}) && (6) \\ p(\mathbf{x}|\mathbf{H}, \mathbf{m}, v) &= \int_{\mathbf{w}} p(\mathbf{x}|\mathbf{w}, \mathbf{H}, \mathbf{m}, v)p(\mathbf{w}) && (7) \\ &\sim \mathcal{N}(\mathbf{m}, \mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I}) && (8) \end{aligned}$$

The probability of the data set is therefore

$$\begin{aligned} p(D|\mathbf{H}, \mathbf{m}, v) &= \prod_i p(\mathbf{x}_i|\mathbf{H}, \mathbf{m}, v) && (9) \\ &= (2\pi)^{-Nd/2} \left|\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I}\right|^{-N/2} \exp(-\frac{1}{2}\mathrm{tr}((\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1}\mathbf{S})) && (10) \\ \mathbf{S} &= \sum_i (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^{\mathrm{T}} && (11) \end{aligned}$$

Regardless of $\mathbf{H}$ and $\mathbf{V}$, the maximum-likelihood value of $\mathbf{m}$ is obviously the sample mean:

$$\hat{\mathbf{m}} = \frac{1}{N}\sum_i \mathbf{x}_i \tag{12}$$

As shown by Tipping & Bishop (1997b), the maximum of (10) over $\mathbf{H}$ occurs at the eigenvectors of the covariance matrix $\mathbf{S}/N$, weighted by the eigenvalues and subject to an arbitrary rotation within the subspace. Specifically,

$$\hat{\mathbf{H}} = \mathbf{U}(\Lambda_k - v\mathbf{I}_k)^{1/2}\mathbf{R} \tag{13}$$

where orthogonal matrix $\mathbf{U}$ contains the top k eigenvectors of $\mathbf{S}/N$, diagonal matrix Λ_k contains the corresponding eigenvalues, and $\mathbf{R}$ is an arbitrary orthogonal matrix. The square root operation is safe because $\lambda_j - v$ will turn out to be positive when we estimate v. For this choice of $\mathbf{H}$, the likelihood reduces to

$$p(D|\mathbf{H} = \hat{\mathbf{H}}, \mathbf{m}, v) = (2\pi)^{-Nd/2} \left(\prod_{j=1}^{k} \lambda_j\right)^{-N/2} v^{-N(d-k)/2} \exp(-\frac{N}{2v}\sum_{j=k+1}^{d} \lambda_j) \exp(-\frac{Nk}{2}) \tag{14}$$

where λ_j is the jth eigenvalue of $\mathbf{S}/N$. From this formula the maximum-likelihood noise variance is seen to be the average of the left-out eigenvalues:

$$\hat{v} = \frac{\sum_{j=k+1}^{d} \lambda_j}{d-k} \tag{15}$$

so the maximized likelihood is simply

$$p(D|\mathbf{H} = \hat{\mathbf{H}}, \mathbf{m}, v = \hat{v}) = (2\pi)^{-Nd/2} \left(\prod_{j=1}^{k} \lambda_j\right)^{-N/2} \hat{v}^{-N(d-k)/2} \exp(-\frac{Nd}{2}) \tag{16}$$

At these parameter values, the covariance matrix of $\mathbf{x}$ is $\mathbf{U}_d\hat{\Lambda}\mathbf{U}_d^{\mathrm{T}}$ where $\mathbf{U}_d$ contains all the eigenvectors of $\mathbf{S}/N$ and

$$\hat{\Lambda} = \begin{bmatrix} \Lambda_k & 0 \\ 0 & v\mathbf{I}_{d-k} \end{bmatrix} \tag{17}$$

In other words, it is the maximum likelihood estimate of covariance, but with the smallest $d-k$ eigenvalues set to their average. The PCA model is equivalent to an equality constraint among the $d-k$ smallest eigenvalues.

3 Bayesian model selection

Bayesian model selection uses the rules of probability theory to select among different hypotheses. It is completely analogous to Bayesian classification. It automatically encodes a preference for simpler, more constrained models, as illustrated in figure 1. Simple models, e.g. linear regression, only fit a small fraction of data sets. But they assign correspondingly higher probability to those data sets. Flexible models spread themselves out more thinly.

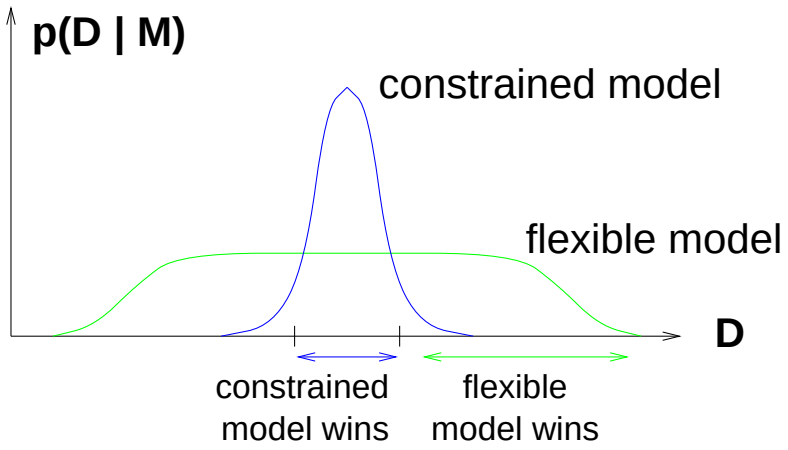

Figure 1: Why Bayesian model selection prefers simpler models

The probability of the data given the model is computed by integrating over the unknown parameter values in that model:

$$p(D|M) = \int_\theta p(D|\theta)p(\theta|M)d\theta \tag{18}$$

This quantity is called the evidence for model M. MacKay (1995) and Kass & Raftery (1993) discuss Bayesian model selection in detail. A useful property of Bayesian model selection is that it is guaranteed to select the true model, if it is among the candidates, as the size of the dataset grows to infinity.

3.1 The evidence for probabilistic PCA

For the PCA model, we want to select the subspace dimensionality k. To do this, we compute the probability of the data for each possible dimensionality. For a given dimensionality, this requires integrating over all PCA parameters $(\mathbf{m}, \mathbf{H}, v)$. First we need to define a prior density for these parameters. Assuming there is no information other than the data D, the prior should be as noninformative as possible. A noninformative prior for $\mathbf{m}$ is uniform:

$$p(\mathbf{m}) = (\text{constant}) \tag{19}$$

The constant depends on the prior range we choose for $\mathbf{m}$. But since this constant has no influence on model selection, we can let $\mathbf{m}$ range over the entire space and assume the constant is 1. With this prior, $\mathbf{m}$ can be integrated out analytically, leaving

$$p(D|\mathbf{H}, v) = N^{-d/2}(2\pi)^{-(N-1)d/2} \left|\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I}\right|^{-(N-1)/2} \exp(-\frac{1}{2}\mathrm{tr}((\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1}\mathbf{S})) \tag{20}$$

$$\text{where } \mathbf{S} = \sum_i (\mathbf{x}_i - \hat{\mathbf{m}})(\mathbf{x}_i - \hat{\mathbf{m}})^{\mathrm{T}} \tag{21}$$

Unlike $\mathbf{m}$, $\mathbf{H}$ must have a proper prior since it varies in dimension for different models. Let $\mathbf{H}$ be decomposed just as in (13):

$$\begin{aligned} \mathbf{H} &= \mathbf{U}(\mathbf{L} - v\mathbf{I}_k)^{1/2}\mathbf{R} && (22)\\ \mathbf{U}^{\mathrm{T}}\mathbf{U} &= \mathbf{I}_k && (23)\\ \mathbf{R}^{\mathrm{T}}\mathbf{R} &= \mathbf{I}_k && (24) \end{aligned}$$

where $\mathbf{L}$ is diagonal with diagonal elements l_i. The orthogonal matrix $\mathbf{U}$ is the basis, $\mathbf{L}$ is the scaling (corrected for noise), and $\mathbf{R}$ is a rotation within the subspace (which will turn out to be irrelevant).

A conjugate prior for $(\mathbf{U}, \mathbf{L}, \mathbf{R}, v)$, parameterized by α, is

$$\begin{aligned} p(\mathbf{U}, \mathbf{L}, \mathbf{R}, v) &\propto \left|\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I}\right|^{-(\alpha+2)/2} \exp(-\frac{\alpha}{2}\mathrm{tr}((\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1})) && (25)\\ &\propto |\mathbf{L}|^{-(\alpha+2)/2}\, v^{-(\alpha+2)(d-k)/2} \exp(-\frac{\alpha}{2}\mathrm{tr}(\mathbf{L}^{-1})) \exp(-\frac{\alpha(d-k)}{2v}) && (26) \end{aligned}$$

This distribution factors into separate terms for $(\mathbf{U}, \mathbf{L}, \mathbf{R}, v)$, which means they are a-priori independent:

$$\begin{aligned} p(\mathbf{U}, \mathbf{L}, \mathbf{R}, v) &= p(v)p(\mathbf{U})p(\mathbf{R}) \prod_{i=1}^{k} p(l_i) && (27)\\ p(v) &\sim \chi^{-2}(\alpha(d-k), (\alpha+2)(d-k) - 2) && (28)\\ &= \frac{1}{\Gamma((\alpha+2)(d-k)/2 - 1)v} \left(\frac{\alpha(d-k)}{2v}\right)^{(\alpha+2)(d-k)/2-1} \exp(-\frac{\alpha(d-k)}{2v}) && (29)\\ p(\mathbf{U})p(\mathbf{R}) &= (\text{constant—defined in (50)}) && (30)\\ p(l_i) &\sim \chi^{-2}(\alpha, \alpha) && (31)\\ &= \frac{1}{\Gamma(\alpha/2)l_i} \left(\frac{\alpha}{2l_i}\right)^{\alpha/2} \exp(-\frac{\alpha}{2l_i}) && (32) \end{aligned}$$

The hyperparameter α controls the sharpness of the prior. For a noninformative prior, α should be small, making the prior diffuse. The prior (27) does not enforce $l_i > v$, but the likelihood will rule out such situations.

Combining the likelihood with the prior gives

$$\begin{aligned} p(D|k) &= c_k \int_{\mathbf{U},\mathbf{L},v} \left|\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I}\right|^{-n/2} \exp(-\frac{1}{2}\mathrm{tr}((\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1}(\mathbf{S} + \alpha\mathbf{I}))) \, d\mathbf{U}d\mathbf{L}dv && (33)\\ n &= N + 1 + \alpha && (34)\\ c_k &= \frac{N^{-d/2}(2\pi)^{-(N-1)d/2}p(\mathbf{U})}{\Gamma((\alpha+2)(d-k)/2 - 1)} \left(\frac{\alpha(d-k)}{2}\right)^{(\alpha+2)(d-k)/2-1} \frac{1}{\Gamma(\alpha/2)^k} \left(\frac{\alpha}{2}\right)^{\alpha k/2} && (35) \end{aligned}$$

In this formula $\mathbf{R}$ has already been integrated out; the likelihood does not involve $\mathbf{R}$ so we just get a multiplicative factor of $\int_{\mathbf{R}} p(\mathbf{R})\, d\mathbf{R} = 1$.

3.2 Laplace approximation

It is possible to integrate (33) over $\mathbf{L}$ and v analytically. However, this leads to a complicated integral for $\mathbf{U}$. A simpler approach is to approximate the whole integral using Laplace's method (see Kass & Raftery (1993) for a description of Laplace's method):

$$\begin{aligned} \int f(\theta)d\theta &\approx f(\hat{\theta})(2\pi)^{\mathrm{rows}(A)/2} \left|\mathbf{A}\right|^{-1/2} && (36)\\ \hat{\theta} = \operatorname*{argmax}_{\theta} f(\theta) \qquad & \mathbf{A} = -\left[\frac{d^2 \log f(\theta)}{d\theta_i d\theta_j}\right]_{\theta=\hat{\theta}} && (37) \end{aligned}$$

For (33), $\theta = (\mathbf{U}, \mathbf{L}, v)$ and

$$\log f(\theta) = -\frac{n}{2}\log|\mathbf{L}| - \frac{n(d-k)}{2}\log(v) - \frac{1}{2}\mathrm{tr}((\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1}(\mathbf{S} + \alpha\mathbf{I})) \tag{38}$$

This expression can be simplified using the identity

$$(\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1} - v^{-1}\mathbf{I} = -(\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1}\mathbf{H}\mathbf{H}^{\mathrm{T}}v^{-1} = \mathbf{U}(\mathbf{L}^{-1} - v^{-1}\mathbf{I})\mathbf{U}^{\mathrm{T}} \tag{39}$$

which gives

$$-\frac{1}{2}\mathrm{tr}((\mathbf{H}\mathbf{H}^{\mathrm{T}} + v\mathbf{I})^{-1}(\mathbf{S} + \alpha\mathbf{I})) = -\frac{1}{2v}\mathrm{tr}(\mathbf{S} + \alpha\mathbf{I}) - \frac{1}{2}\mathrm{tr}((\mathbf{L}^{-1} - v^{-1}\mathbf{I})\mathbf{U}^{\mathrm{T}}(\mathbf{S} + \alpha\mathbf{I})\mathbf{U}) \tag{40}$$

$$= -\frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{2v} - \frac{1}{2}\mathrm{tr}(\mathbf{L}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U}) - \frac{\alpha}{2}\mathrm{tr}(\mathbf{L}^{-1}) \tag{41}$$

The key to getting a good approximation is choosing a good parameterization for $\mathbf{U}$, $\mathbf{L}$, and v. Since l_i and v are positive scale parameters, it is best to use $l_i' = \log(l_i)$ and $v' = \log(v)$. This transformation has Jacobian $J_{l' \to l} = l_i$. The derivatives with respect to l_i' at the maximum-likelihood value of $\mathbf{U}$ are

$$\frac{d\log f(\theta)}{dl_i'} = -\frac{n}{2} + \frac{N\lambda_i + \alpha}{2l_i} + 1 \tag{42}$$

$$\frac{d^2\log f(\theta)}{(dl_i')^2} = -\frac{N\lambda_i + \alpha}{2l_i} \tag{43}$$

which determine

$$\hat{l}_i = (N\lambda_i + \alpha)/(N - 1 + \alpha) \tag{44}$$

$$\left.\frac{d^2\log f(\theta)}{(dl_i')^2}\right|_{\theta=\hat{\theta}} = -\frac{N - 1 + \alpha}{2} \tag{45}$$

The derivatives with respect to $v' = \log(v)$ are (using (41))

$$\frac{d\log f(\theta)}{dv'} = -\frac{n(d-k)}{2} + \frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{2v} + 1 \tag{46}$$

$$\frac{d^2\log f(\theta)}{(dv')^2} = -\frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{2v} \tag{47}$$

which determine

$$\hat{v} = \frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{n(d-k) - 2} = \frac{N\sum_{j=k+1}^{d}\lambda_j}{n(d-k) - 2} \tag{48}$$

$$\left.\frac{d^2\log f(\theta)}{(dv')^2}\right|_{\theta=\hat{\theta}} = -\frac{n(d-k) - 2}{2} \tag{49}$$

The matrix $\mathbf{U}$ is an orthogonal k-frame and therefore lives on the Stiefel manifold (James, 1954), which is defined by condition (23). The dimension of the manifold is $m = dk - k(k+1)/2$, since we are imposing $k(k+1)/2$ constraints on a $d \times k$ matrix. The prior density for $\mathbf{U}$ is therefore the reciprocal of the area of the manifold (James, 1954):

$$p(\mathbf{U}) = 2^{-k}\prod_{i=1}^{k}\Gamma((d - i + 1)/2)\pi^{-(d-i+1)/2} \tag{50}$$

The manifold can be parameterized by Euler vector coordinates:

$$\mathbf{U} = \mathbf{U}_d \exp(\mathbf{Z}) \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \tag{51}$$

where $\mathbf{U}_d$ is a fixed orthogonal matrix and $\mathbf{Z}$ is a skew-symmetric matrix of parameters, e.g.

$$\mathbf{Z} = \begin{bmatrix} 0 & z_{12} & z_{13} \\ -z_{12} & 0 & z_{23} \\ -z_{13} & -z_{23} & 0 \end{bmatrix} \tag{52}$$

The free parameters in this matrix are the top k rows of the upper triangle, i.e. the entries z_{ij} with $i < j$ and $i \leq k$; the others are constant. This gives $d(d-1)/2 - (d-k)(d-k-1)/2 = m$ parameters, as desired. For example, in the case $(d = 3, k = 1)$ the free parameters are z_{12} and z_{13}, which define a coordinate system for the sphere.

Using (39) we find that as a function of $\mathbf{U}$, the integrand is simply

$$p(\mathbf{U}|D, \mathbf{L}, v) \propto \exp(-\frac{1}{2}\mathrm{tr}((\mathbf{L}^{-1} - v^{-1}\mathbf{I})\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})) \tag{53}$$

This distribution was studied by Bingham (1974) for the case $(d = 3, k = 1)$, where it is a distribution over the sphere. Figure 2 plots a typical instance of this distribution. The generalization to the Stiefel manifold was mentioned by Khatri & Mardia (1977) and is known as the *matrix Bingham distribution*. The density is maximized when $\mathbf{U}$ contains the top k eigenvectors of $\mathbf{S}$. However, the density is unchanged if we negate any column of $\mathbf{U}$. This means that there are actually 2^k different maxima, and we need to apply Laplace's method to each. Fortunately, these maxima are identical so can simply multiply (36) by 2^k to get the integral over the whole manifold. If we set $\mathbf{U}_d$ to the eigenvectors of $\mathbf{S}$:

$$\mathbf{U}_d^{\mathrm{T}}\mathbf{S}\mathbf{U}_d = N\Lambda \tag{54}$$

then we just need to apply Laplace's method at $\mathbf{Z} = \mathbf{0}$.

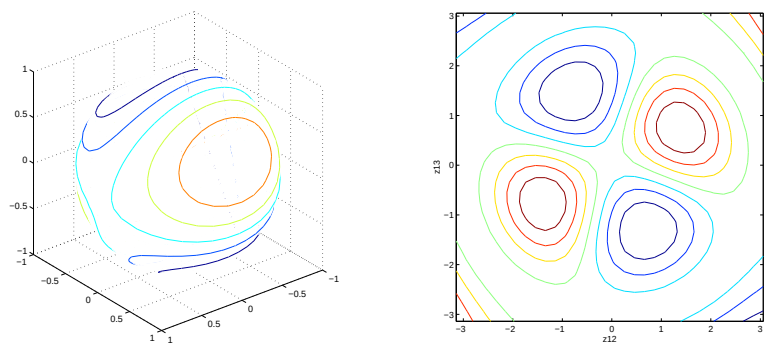

Figure 2: The posterior distribution for the first principal component in three dimensions, contour-plotted on the sphere and in Euler coordinates. It is equivalent to constraining a full-covariance Gaussian density to the sphere. Euler coordinates unwrap the sphere so that both modes, on opposite sides of the sphere, are visible.

Since

$$\exp(\mathbf{Z}) = \mathbf{I} + \mathbf{Z} + \frac{1}{2}\mathbf{Z}^2 + \frac{1}{6}\mathbf{Z}^3 + \cdots \tag{55}$$

the differential of $\mathbf{U}$ in Euler coordinates is

$$d\mathbf{U} = \mathbf{U}_d(d\mathbf{Z} + \frac{1}{2}(\mathbf{Z}d\mathbf{Z} + d\mathbf{Z}\mathbf{Z}) + \cdots) \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \tag{56}$$

$$d\mathbf{U}|_{\mathbf{Z}=\mathbf{0}} = \mathbf{U}_d d\mathbf{Z} \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \tag{57}$$

The second differential of $\mathbf{U}$ is

$$d^2\mathbf{U} = \mathbf{U}_d(d\mathbf{Z}^2 + \cdots)\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix} \tag{58}$$

$$d^2\mathbf{U}\big|_{\mathbf{Z}=\mathbf{0}} = \mathbf{U}_d d\mathbf{Z}^2\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix} \tag{59}$$

Therefore the differential of $\log f$ is

$$d\log f(\theta) = -\mathrm{tr}((\mathbf{L}^{-1} - v^{-1}\mathbf{I})\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathrm{d}\mathbf{U}) \tag{60}$$

and the second differential is

$$d^2\log f(\theta) = -\mathrm{tr}((\mathbf{L}^{-1} - v^{-1}\mathbf{I})d\mathbf{U}^{\mathrm{T}}\mathbf{S}d\mathbf{U}) - \mathrm{tr}((\mathbf{L}^{-1} - v^{-1}\mathbf{I})\mathbf{U}^{\mathrm{T}}\mathbf{S}d^2\mathbf{U}) \tag{61}$$

$$d^2\log f(\theta)\big|_{\mathbf{Z}=\mathbf{0}} = -N\mathrm{tr}(\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}(\mathbf{L}^{-1} - v^{-1}\mathbf{I})\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}^{\mathrm{T}} d\mathbf{Z}^{\mathrm{T}}\Lambda d\mathbf{Z})$$

$$-N\mathrm{tr}(d\mathbf{Z}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}(\mathbf{L}^{-1} - v^{-1}\mathbf{I})\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}^{\mathrm{T}}\Lambda d\mathbf{Z}) \tag{62}$$

$$= -N\mathrm{tr}((\mathbf{B}d\mathbf{Z}^{\mathrm{T}} + (d\mathbf{Z})\mathbf{B})\Lambda d\mathbf{Z}) \tag{63}$$

$$= -N\mathrm{tr}(\mathbf{T}\Lambda d\mathbf{Z}) \tag{64}$$

$$\text{where } \mathbf{B} = \begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}(\mathbf{L}^{-1} - v^{-1}\mathbf{I})\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}^{\mathrm{T}} \quad \text{(a diagonal matrix)} \tag{65}$$

$$\mathbf{T} = \mathbf{B}d\mathbf{Z}^{\mathrm{T}} + (d\mathbf{Z})\mathbf{B} = (d\mathbf{Z})\mathbf{B} - \mathbf{B}d\mathbf{Z} \tag{66}$$

$$t_{ij} = b_j dz_{ij} - b_i dz_{ij} \tag{67}$$

If we define the estimated eigenvalue matrix (analogous to (17))

$$\hat{\Lambda} = \begin{bmatrix}\hat{\mathbf{L}} & 0\\ 0 & \hat{v}\mathbf{I}_{d-k}\end{bmatrix} \tag{68}$$

then the (i,j) element of $\mathbf{T}$ is

$$t_{ij} = (\hat{\lambda}_j^{-1} - \hat{\lambda}_i^{-1})dz_{ij} \tag{69}$$

Now exploit the fact that $dz_{ji} = -dz_{ij}$ to get

$$d^2\log f(\theta)\big|_{\mathbf{Z}=\mathbf{0}} = -\sum_{i=1}^{k}\sum_{j=i+1}^{d}(\hat{\lambda}_j^{-1} - \hat{\lambda}_i^{-1})(\lambda_i - \lambda_j)Ndz_{ij}^2 \tag{70}$$

Note that there are no cross derivatives; the Hessian matrix $\mathbf{A}_Z$ is diagonal. So its determinant is the product of these second derivatives:

$$|\mathbf{A}_Z| = \prod_{i=1}^{k}\prod_{j=i+1}^{d}(\hat{\lambda}_j^{-1} - \hat{\lambda}_i^{-1})(\lambda_i - \lambda_j)N \tag{71}$$

Laplace's method requires this to be nonsingular, so we must have $k < N$.

The cross-derivatives between the parameters are all zero:

$$\left.\frac{d^2\log f(\theta)}{dl_i d\mathbf{Z}}\right|_{\theta=\hat{\theta}} = \left.\frac{d^2\log f(\theta)}{dv d\mathbf{Z}}\right|_{\theta=\hat{\theta}} = \left.\frac{d^2\log f(\theta)}{dl_i dv}\right|_{\theta=\hat{\theta}} = 0 \tag{72}$$

so $\mathbf{A}$ is block diagonal:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_Z & & \\ & \mathbf{A}_L & \\ & & \mathbf{A}_v \end{bmatrix} \tag{73}$$

$$|\mathbf{A}| = |\mathbf{A}_Z|\,|\mathbf{A}_L|\,|\mathbf{A}_v| \tag{74}$$

We know $\mathbf{A}_L$ from (45), $\mathbf{A}_v$ from (49), and $\mathbf{A}_Z$ from (71). We now have all of the terms needed in (36), and so the evidence approximation is

$$p(D|k) \approx 2^k c_k \left|\hat{\mathbf{L}}\right|^{-n/2} \hat{v}^{-n(d-k)/2} \exp(-\frac{nd}{2})(2\pi)^{(m+k+1)/2} \,|\mathbf{A}_Z|^{-1/2} \,|\mathbf{A}_L|^{-1/2} \,|\mathbf{A}_v|^{-1/2} \tag{75}$$

For model selection, the only terms that matter are those that strongly depend on k, and since α is small and N reasonably large we can simplify this to

$$p(D|k) \approx p(\mathbf{U}) \left(\prod_{j=1}^{k} \lambda_j\right)^{-N/2} \hat{v}^{-N(d-k)/2} (2\pi)^{(m+k)/2} \,|\mathbf{A}_Z|^{-1/2} \,N^{-k/2} \tag{76}$$

$$\hat{l}_i = \lambda_i \qquad \hat{v} = \frac{\sum_{j=k+1}^{d} \lambda_j}{d-k} \tag{77}$$

which is the recommended formula. Given the eigenvalues, the cost of computing $p(D|k)$ is $O(\min(d,N)k)$, which is less than one loop over the data matrix.

A simplification of Laplace's method is the BIC approximation (Kass & Raftery, 1993). This approximation drops all terms which do not grow with N, which in this case leaves only

$$p(D|k) \approx \left(\prod_{j=1}^{k} \lambda_j\right)^{-N/2} \hat{v}^{-N(d-k)/2} N^{-(m+k)/2} \tag{78}$$

This approximation is compared to Laplace in section 5.

4 Other approaches

Rajan & Rayner (1997) perform model selection on a slightly different probabilistic PCA model. In fact they have two different models—one with a Gaussian density in the subspace and one with a uniform density:

$$\mathbf{x} = \mathbf{U}\mathbf{w} + \mathbf{m} + \mathbf{e} \tag{79}$$

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_k \tag{80}$$

$$p(\mathbf{w}|\alpha) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\alpha) \tag{81}$$

$$\text{or } p(w_i|\beta) \sim \mathcal{U}(-\beta, \beta) \qquad \text{for all } i \tag{82}$$

They also included an assumption that $\mathbf{U}$ is smooth, which we omit. Under this model, the covariance of $\mathbf{x}$ is

$$E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathrm{T}}] = \begin{bmatrix} (\alpha^{-1}+v)\mathbf{I}_k & 0 \\ 0 & v\mathbf{I}_{d-k} \end{bmatrix} \tag{83}$$

$$\text{or} = \begin{bmatrix} (\beta^2/6+v)\mathbf{I}_k & 0 \\ 0 & v\mathbf{I}_{d-k} \end{bmatrix} \tag{84}$$

This is different since it implies all subspace components have the same variance, i.e. the true eigenvalues are constant over $\lambda_1, ..., \lambda_k$ as well as constant over $\lambda_{k+1}, ..., \lambda_d$. For the first model, the probability of a data set is

$$p(D|\mathbf{U}, \mathbf{m}, v, \alpha) = (2\pi)^{-Nd/2} \left|\alpha^{-1}\mathbf{U}\mathbf{U}^{\mathrm{T}} + v\mathbf{I}\right|^{-N/2} \exp(-\frac{1}{2}\mathrm{tr}((\alpha^{-1}\mathbf{U}\mathbf{U}^{\mathrm{T}} + v\mathbf{I})^{-1}\mathbf{S})) \quad (85)$$

$$= (2\pi)^{-Nd/2}(\alpha^{-1} + v)^{-Nk/2} v^{-N(d-k)/2} \exp(-\frac{\mathrm{tr}(\mathbf{S})}{2v} + \frac{\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{2v(1+\alpha v)}) \quad (86)$$

(because $(\alpha^{-1}\mathbf{U}\mathbf{U}^{\mathrm{T}} + v\mathbf{I})^{-1} = v^{-1}\mathbf{I} - \mathbf{U}v^{-1}(\alpha + v^{-1})^{-1}v^{-1}\mathbf{U}^{\mathrm{T}}$). Rather than integrate over the parameters $(\mathbf{U}, \mathbf{m}, v, \alpha)$ to get the evidence, Rajan and Rayner suggest simply using the maximum of this likelihood for model selection. The maximum-likelihood value of $\mathbf{U}$ and $\mathbf{m}$ are the same as before. Rajan and Rayner give an approximate formula for $\hat{\alpha}$ and $\hat{v}$; the exact maximum-likelihood values are

$$\hat{v} = \frac{\sum_{j=k+1}^{d} \lambda_j}{d-k} \quad (87)$$

$$\hat{\alpha}^{-1} = \frac{\sum_{j=1}^{k} \lambda_j}{k} - \hat{v} \quad (88)$$

which gives the maximized likelihood (cf (16))

$$p(D|\hat{\mathbf{U}}, \hat{\mathbf{m}}, \hat{v}, \hat{\alpha}) = (2\pi)^{-Nd/2} \left(\frac{\sum_{j=1}^{k} \lambda_j}{k}\right)^{-Nk/2} \hat{v}^{-N(d-k)/2} \exp(-\frac{Nd}{2}) \quad (89)$$

We will call this the RR-N algorithm, with the caveat that it is not identical to what Rajan and Rayner proposed. For the second model, the probability of a data set is

$$p(D|\mathbf{U}, \mathbf{m}, v, \beta) = (2\pi v)^{-Nd/2}(2\beta)^{-Nk} \prod_{i=1}^{N} \int_{\mathbf{w}} \exp(-\frac{1}{2v}(\mathbf{x}_i - \mathbf{m} - \mathbf{U}\mathbf{w})^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m} - \mathbf{U}\mathbf{w}))d\mathbf{w} \quad (90)$$

$$= (2\pi v)^{-Nd/2}(2\beta)^{-Nk} \exp(-\frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{2v}) \prod_{i=1}^{N} \int_{\mathbf{w}} \exp(-\frac{1}{2v}(\mathbf{w} - \mathbf{U}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m}))^{\mathrm{T}}(\mathbf{w} - \mathbf{U}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m})))d\mathbf{w} \quad (91)$$

$$= (2\pi v)^{-Nd/2}(2\beta)^{-Nk} \exp(-\frac{\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U})}{2v}) \prod_{i=1}^{N} \prod_{j=1}^{k} \sqrt{\pi v/2} \left(\mathrm{erf}\left(\frac{\beta - \mathbf{u}_j^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m})}{\sqrt{2v}}\right) - \mathrm{erf}\left(\frac{-\beta - \mathbf{u}_j^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m})}{\sqrt{2v}}\right)\right) \quad (92)$$

(In this formula, an error of Rajan and Rayner has been corrected.) Rajan and Rayner estimate v and β with

$$\hat{v} = \frac{\sum_{j=k+1}^{d} \lambda_j}{d-k} \quad (93)$$

$$\hat{\beta} = \max_{j,i} |\mathbf{u}_j^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m})| \quad (94)$$

We will call this the RR-U algorithm.

Everson & Roberts (2000) also perform Bayesian model selection on a slightly different probabilistic PCA model. They use an approximate generative model for the observed eigenvalues, which decouples as a function of the true eigenvalues:

$$p(\Lambda|\mathbf{L}, v) = \prod_{j=1}^{d} \prod_{i=1}^{k} f_{ij}((\lambda_j - l_i)/v) \prod_{i=k+1}^{d} f_{ij}((\lambda_j - 0)/v) \quad (95)$$

The f_{ij} are d^2 different functions relating each observed eigenvalue to each true eigenvalue. The evidence integral simplifies into k univariate integrals over l_i which are evaluated numerically. The noise variance v is not integrated out but chosen to maximize the evidence for each dimensionality k; a choice which must be done numerically. This technique will be called the ER algorithm.

Bishop's (1998) algorithm is different from the others in that it does not score each dimensionality but only reports the best dimensionality. It is an iterative estimation algorithm for $\mathbf{H}$ which sets columns to zero unless they are supported by the data. The number of nonzero columns at convergence is the estimate of dimensionality. The algorithm is based on MacKay's (1995) automatic relevance determination framework and so here it is called the ARD algorithm.

5 Results

To test the performance of these various algorithms for model selection, we can sample data from a known model and see how often the correct dimensionality is recovered. The seven estimators implemented and tested in this study are Laplace's method (76), BIC (78), Rajan and Rayner's RR-N (89), RR-U (92), Everson and Roberts' ER algorithm, Bishop's ARD algorithm, and 5-fold cross-validation. In the latter method, the data set is divided into 5 equal parts, and in turn we use one part to test the PCA model fitted to the remaining parts. The score for each division is the log-probability assigned to the held-out data. The score for a given dimensionality is the average score across the five divisions.

Most of these estimators work exclusively from the eigenvalues of the sample covariance matrix. The exceptions are RR-U, cross-validation, and ARD; the latter two require diagonalizing a series of different matrices constructed from the data. In our implementation, the algorithms are ordered from fastest to slowest as RR-N, BIC, Laplace, cross-validation, RR-U, ARD, and ER (ER is slowest because of the numerical integrations required). All of the estimators are guaranteed to recover the true dimensionality for a large enough data set, except for RR-N and RR-U because they use a restrictive model for the subspace.

The first experiment tests the data-rich case where $N >> d$. The data is generated from a 10-dimensional Gaussian distribution with variance in 5 directions given by [10 8 6 4 2] and variance 1 in the remaining 5 directions. Figure 3 plots the eigenvalues of the true covariance matrix and the observed covariance matrix for one particular realization of 100 samples. For each choice of dimensionality, figure 4 plots the maximized likelihood and the scores given by the various estimators. Most of them, including ARD, report $k = 5$ for this set of data. RR-N picks $k = 4$ and RR-U picks $k = 1$. The results over 60 replications are reported in figure 5. The differences between ER, Laplace, and CV are not statistically significant. Results below the dashed line are worse than Laplace with a significance level of 95%.

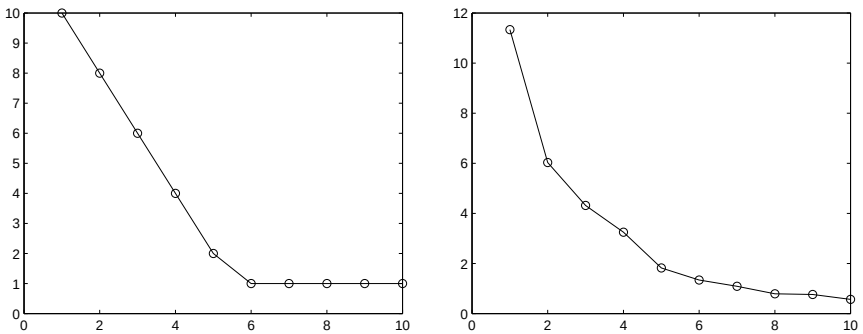

Figure 3: True vs. observed covariance matrix eigenvalues for 100 points in 10 dimensions. The latent dimensionality is 5.

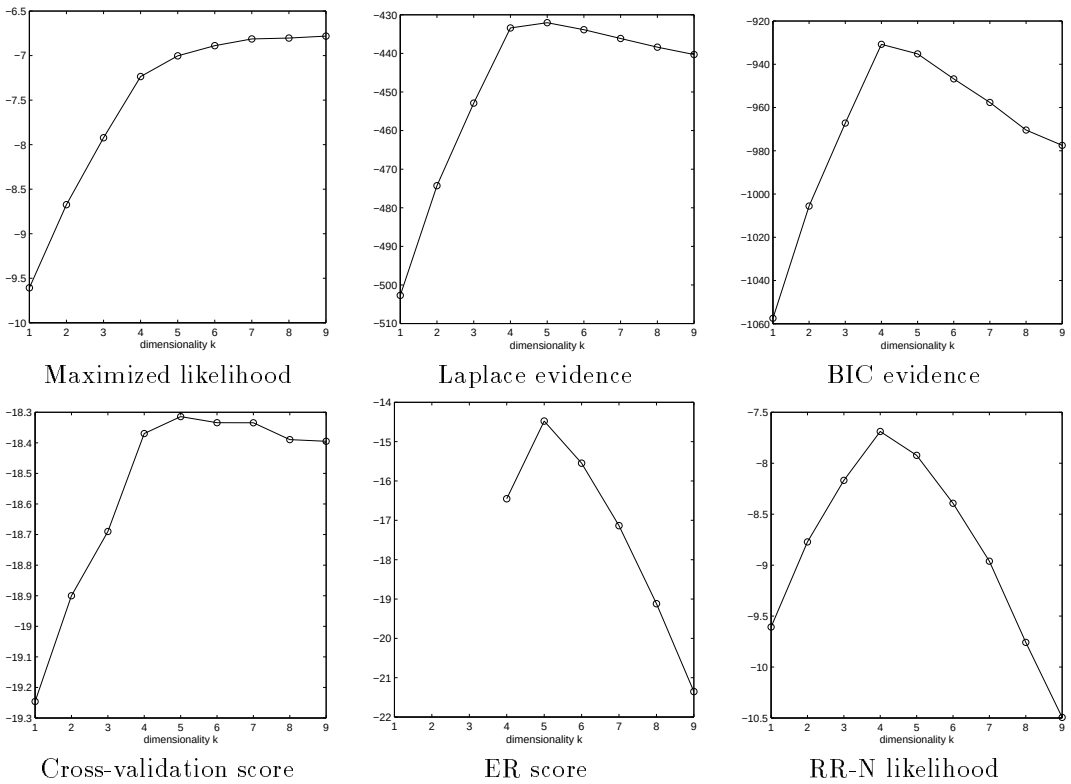

Figure 4: The score for each dimensionality, evaluated in six different ways. The true value is $k = 5$.

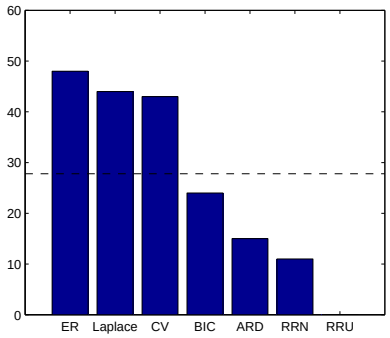

Figure 5: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 10, k = 5, N = 100$)

The second experiment tests the case of sparse data and low noise. The dimensionality is $d = 15$; the variance in the first 5 directions is the same but now the variance is 0.1 in the remaining 10 directions. There are only 10 data points. The results over 60 replications are reported in figure 6.

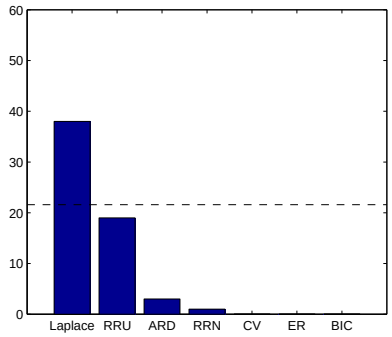

Figure 6: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 15, k = 5, N = 10$)

The third experiment tests the case of high noise dimensionality. The data is generated from a 100-dimensional Gaussian distribution with variance in 5 directions given by [10 8 6 4 2] and variance 1/4 in the remaining 95 directions. Figure 7 plots the eigenvalues of the true covariance matrix and the observed covariance matrix for one particular realization of 60 samples. For each choice of dimensionality, figure 8 plots the maximized likelihood and the scores given by the various estimators. Notice that BIC, which was derived as a large N approximation, is unreliable when the dimensionality is comparable to N. Fortunately, we can reject such solutions out of hand if there is a clear peak elsewhere. The results over 60 replications are reported in figure 9. The ER algorithm was not run in this case because of its excessive computation time for large d.

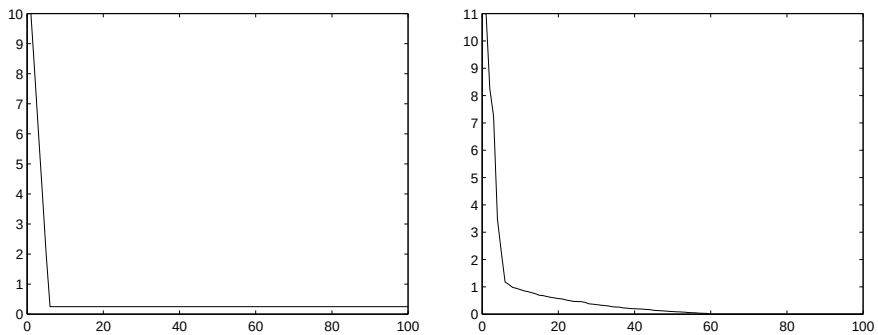

Figure 7: True (left) and observed (right) covariance matrix eigenvalues for 60 points in 100 dimensions.

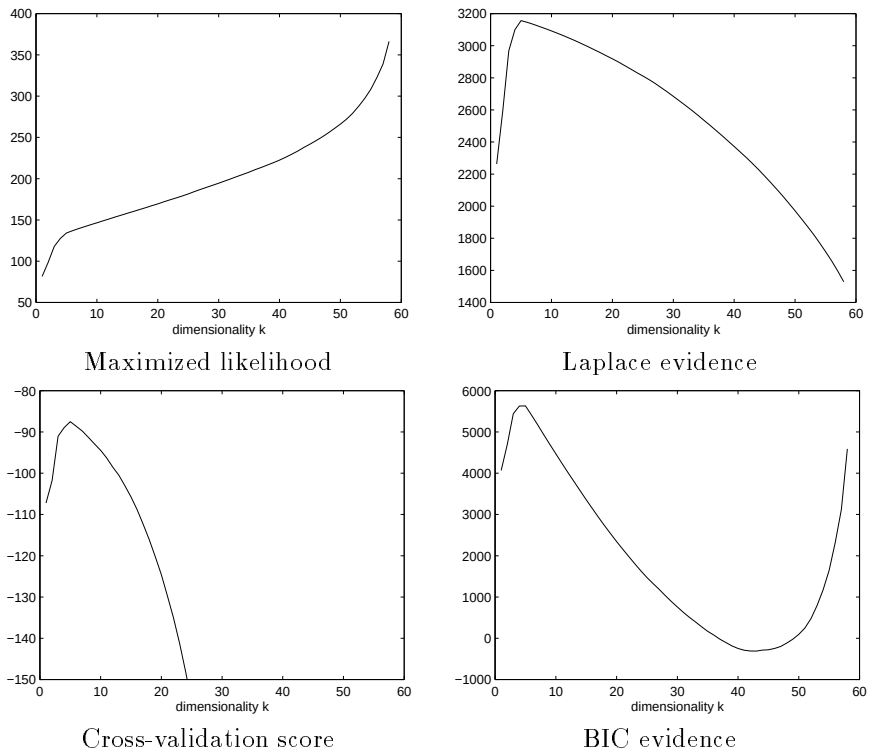

Figure 8: The score for each dimensionality, evaluated in four different ways. The cross-validation curve drops off quickly after $k = 15$. All except the likelihood peak at the true value in this case.

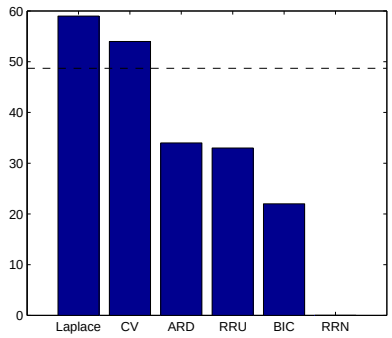

Figure 9: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 100, k = 5, N = 60$)

Since the data in these experiments really does follow the generative model, we should expect Bayesian model selection to be optimal. The Laplace approximation turns out to be excellent; it is a consistent top performer. Cross-validation is also a good performer, but it is expensive to compute. The algorithms RR-N, RR-U, and ER are effective only under certain conditions. The ARD algorithm does not give performance commensurate with its expense. Furthermore, the algorithms based on scoring can employ a smart search algorithm for the best k, but ARD cannot be accelerated in a simple way.

The next experiment tests the robustness to having a non-Gaussian data distribution within the subspace. We start with four sound fragments of $N = 100$ samples each. To make things especially non-Gaussian, the values in third fragment are squared and the values in the fourth fragment are cubed. All fragments are standardized to zero mean and unit variance. Figure 10 plots a kernel estimate of the distribution of values in each fragment. They are clearly non-Gaussian.

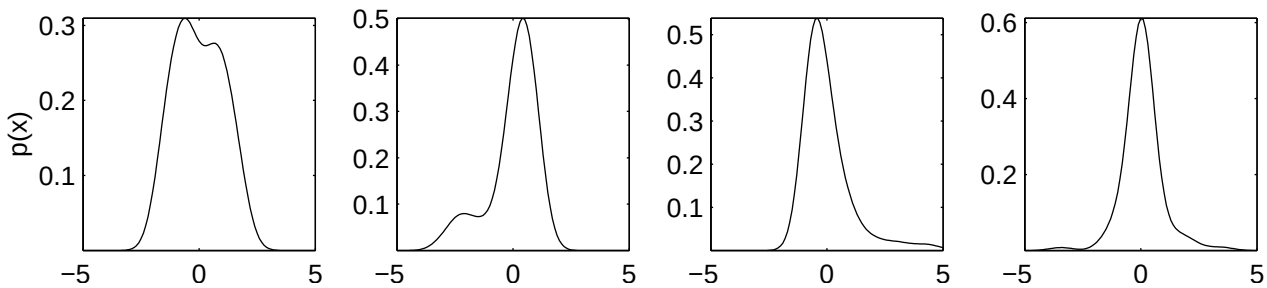

Figure 10: The distribution of samples in the sound fragments used in the second experiment. They are clearly non-Gaussian.

To this 4-dimensional data is added $d = 20$ dimensional Gaussian noise with variance $v = 1/2$ in all directions. Figure 11 plots the eigenvalues of the true covariance matrix and the observed covariance matrix for one particular realization of the noise. Figure 12 reports the results over 60 replications of the noise (the signals were constant).

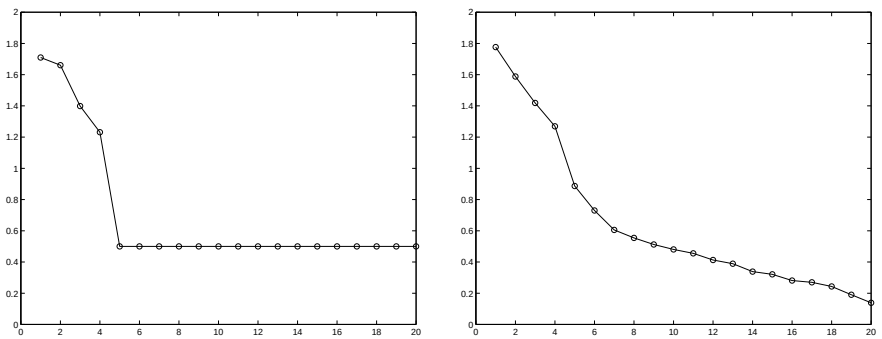

Figure 11: True (left) and observed (right) covariance matrix eigenvalues for 4 sounds embedded in 20-dimensional noise.

In an experiment where the true eigenvalues do not level off, but continue downward, all of the estimators pick the largest possible dimensionality, given a large enough dataset (except RR-N and RR-U because of their restrictive model). This underscores the fact that these estimators are for density estimation, i.e. accurate representation of the data, and are not necessarily appropriate for other purposes like reducing computation or extracting salient features. For example, on a database of 301 face images the Laplace evidence picked 120 dimensions, which is far more than one would use for feature extraction. (This result also suggests that probabilistic PCA is not a good generative model for face images.) A more appropriate use of these estimators

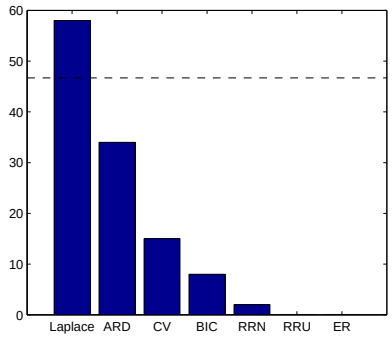

Figure 12: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 20, k = 4, N = 100$)

is fitting different PCA models to different classes, for use in Bayesian classification (Moghaddam & Pentland, 1997; Moghaddam et al., 1998).

6 Future directions

Bayesian model selection has been shown to provide excellent performance when the assumed model is correct or partially correct. The evaluation criterion was the number of times the correct dimensionality was chosen. It would also be useful to evaluate the trained model with respect to its performance on new data. It is conceivable that a method like ARD, which encompasses a soft blend between different dimensionalities, might perform better by this criterion than selecting one dimensionality.

The probabilistic PCA model can be incorporated into a larger probabilistic model, such as a mixture model (Tipping & Bishop, 1997a). Indeed, the ARD algorithm was designed for this purpose. A brute force approach to Bayesian model selection would be impractical, since we would need to try every combination of mixture component models. A more reasonable approach is to optimize each component model in turn, holding the others fixed. For a given mixture component, the Laplace formula (76) can be applied to the eigenvalues of the local responsibility-weighted covariance matrix (defined by Tipping & Bishop (1997a)).

Acknowledgment

This work was supported by the MIT Media Lab Digital Life Consortium.

References

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *Annals of Statistics*, *2*, 1201–1225.

Bishop, C. (1998). Bayesian PCA. *Neural Information Processing Systems 11* (pp. 382–388).

Everson, R., & Roberts, S. (2000). Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Trans Signal Processing*, *48*, 2083–2091. `http://www.robots.ox.ac.uk/~sjrob/Pubs/spectrum.ps.gz`.

James, A. (1954). Normal multivariate analysis and the orthogonal group. *Annals of Mathematical Statistics*, *25*, 40–75.

Kass, R. E., & Raftery, A. E. (1993). *Bayes factors and model uncertainty* (Technical Report 254). University of Washington. `http://www.stat.washington.edu/tech.reports/tr254.ps`.

Khatri, C. G., & Mardia, K. V. (1977). The von Mises-Fisher matrix distribution in orientation statistics. *J Royal Statistical Society B, 39*, 95–106.

MacKay, D. J. C. (1995). Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems, 6*, 469–505. `http://wol.ra.phy.cam.ac.uk/mackay/abstracts/network.html`.

Moghaddam, B., Jebara, T., & Pentland, A. (1998). Bayesian modeling of facial similarity. *Neural Information Processing Systems 11* (pp. 910–916).

Moghaddam, B., & Pentland, A. (1995). Probabilistic visual learning for object detection. *Int Conf on Comp Vision* (pp. 786–793). `ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-326.ps.Z`.

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Trans Pattern Analysis and Machine Intelligence, 19*, 696–710.

Rajan, J. J., & Rayner, P. J. W. (1997). Model order selection for the singular value decomposition and the discrete Karhunen-Loéve transform using a Bayesian approach. *IEE Vision, Image and Signal Processing, 144*, 166–123.

Roweis, S. (1997). EM algorithms for PCA and SPCA. *Neural Information Processing Systems 10* (pp. 626–632).

Tipping, M. E., & Bishop, C. M. (1997a). *Mixtures of probabilistic principal component analysers* (Technical Report NCRG/97/003). Neural Computing Research Group, Aston University. `http://neural-server.aston.ac.uk/Papers/postscript/NCRG_97_003.ps.Z`.

Tipping, M. E., & Bishop, C. M. (1997b). *Probabilistic principal component analysis* (Technical Report NCRG/97/010). Neural Computing Research Group, Aston University. `http://neural-server.aston.ac.uk/Papers/postscript/NCRG_97_010.ps.Z`.