Datalogisk Institut
KØBENHAVNS UNIVERSITET

Bayes PCA Revisited

Jon Sporring

Technical Report no. 08-09
ISSN: 0107-8283

Dept. of Computer Science
University of Copenhagen • Universitetsparken 1
DK-2100 Copenhagen • Denmark

Bayes PCA Revisited

Jon Sporring*

Department of Computer Science, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen, Denmark

June 26, 2008

Abstract

Principle Component Analysis is a simple tool to obtain linear models for stochastic data and is used both for a data reduction or equivalently noise elimination and for data analysis. Principle Component Analysis fits a multivariate Gaussian distribution to the data, and the typical method is by using the log-likelihood estimator. However for small sets of high dimensional data, the log-likelihood estimator is often far from convergence, and therefore reliable models must be obtained by use of prior information. In this paper, we will examine an earlier work on reconstructing missing data using statistical knowledge and regularization, we will show the circumstances for which this is equivalent to a Bayes estimation, we will give an expository presentation of Bayes Principle Component Analysis for a range of exponential type priors, and we will develop algorithms for their estimate.

1 Elasticity in Linear Shape-Variation on Tooth Landmarks

In [1], shape variation of landmarks on teeth was studied using an extension of Principle Component Analysis [2, 3], where it was suggested to estimate of the covariance matrix as,

$$C = \frac{1}{M} X X^T + K \tag{1}$$

where $X = [x_1, x_2, \ldots, x_M]$ is a matrix of concatenated N dimensional points in Euclidean space and $\cdot^T$ is the transpose operator. It was observed that the generalization error was reduced in terms of the leave-one-out error, when matrix K had the following form,

$$K_{ij} = f(d_{ij}) \left(\frac{n_i \cdot n_j}{2} + \frac{1}{2} \right) \tag{2}$$

where f is an exponentially decreasing function such that $f(0) = 1$, d_{ij} is the geodesic distance between two points on the surface of a tooth x_i and x_j, and n_i and n_j are the corresponding outward normals.

We will in the following show, that (1) may be considered a Bayes estimate of the covariance matrix, where the prior is given as an inverted Wishart distribution

$$P(C) = k \exp\left(-\frac{M}{2} \mathrm{tr}\left(K C^{-1} \right) \right), \tag{3}$$

*Danish Technical University, Informatics and Mathematical Modelling, Richard Petersens Plads DTU - Building 321, DK-2800 Lyngby, Denmark is greatfully acknowledge for making research facilities available during part of this work.

for a suitable normalization constant k, arbitrary constant matrix $M > 0$, and when K is positive definite. We will further show that K is positive definite, when $f(d_{ij})$ is diagonally dominant.

2 Log-Likelihood Estimation of Mean and Covariance

In the following we will give an expository derivation of the classical log-likelihood estimates of the mean vector and covariance matrix for random data sets using the method of Matrix Differential Calculus [4], and this material will be used as a reference point for deriving expressions for Bayes estimates.

Consider an N dimensional Euclidean space and M sample points in this space, $x_m \in \Re^N$. We will assume that the sample points are identically, independently, and normally distributed in $\Re^N$ according to,

$$P(x_m|\mu, C) = (2\pi)^{\left(\frac{N}{2}\right)} |C|^{\left(\frac{1}{2}\right)} \exp\left(-\frac{1}{2}(x_m - \mu)^T C^{-1}(x_m - \mu)\right), \tag{4}$$

for unknown covariance matrix $C \in \Re^{N\times N}$ and mean $\mu \in \Re^N$. Hence, the joint distribution is given as,

$$P(x_1, ..., x_m|\mu, C) = \left((2\pi)^{\left(\frac{N}{2}\right)} |C|^{\left(\frac{1}{2}\right)}\right)^{-M} \exp\left(-\frac{1}{2}\sum_{m=1}^{M}(x_m - \mu)^T C^{-1}(x_m - \mu)\right), \tag{5}$$

To estimate C and μ from a set of samples, we seek the maximum point of $P(x_1, ..., x_m|\mu, C)$ as,

$$\frac{\partial P(x_1, ..., x_m|\mu, C)}{\partial C_{ij}} = 0, \tag{6a}$$

$$\frac{\partial P(x_1, ..., x_m|\mu, C)}{\partial \mu_m} = 0, \tag{6b}$$

or equivalently

$$\frac{\partial \log P(x_1, ..., x_m|\mu, C)}{\partial C_{ij}} = 0, \tag{7a}$$

$$\frac{\partial \log P(x_1, ..., x_m|\mu, C)}{\partial \mu_m} = 0, \tag{7b}$$

due to the strict monotonicity of the logarithm function. For practical reasons we also rewrite the sum under the exponential function as,

$$\sum_{m=1}^{M}(x_m - \mu)^T C^{-1}(x_m - \mu) = \mathrm{tr}\left(C^{-1}XX^T\right) \tag{8}$$

where

$$X = [(x_1 - \mu), \ldots, (x_m - \mu)] \tag{9}$$

and rewrite the logarithm of the Gaussian distribution as,

$$L(x_1, ..., x_m|\mu, C) = -\frac{MN}{2}\log(2\pi) - \frac{M}{2}\log|C| - \frac{1}{2}\mathrm{tr}\left(C^{-1}XX^T\right), \tag{10}$$

This is the same as minus the optimal code length [5] of the total set of data points from our assumed Gaussian source, and where C and μ is fixed and known to both the sender and receiver.

The differential of L only considering C and μ as variables is found to be,

$$dL = -\mathrm{tr}\left(\frac{M}{2}C^{-1}dC + \frac{1}{2}\left(dC^{-1}\right)XX^T + \frac{1}{2}C^{-1}d\left(XX^T\right)\right). \tag{11}$$

To identify the partial derivatives of dL we isolate terms containing dC and $d\mu$. For C we need only consider the first two terms in (11), and together with

$$dC^{-1} = -C^{-1}\left(dC\right)C^{-1}, \tag{12}$$

we find that

$$\begin{aligned} 0 &= dL_C && \text{(13a)}\\ &= -\mathrm{tr}\left(\frac{M}{2}C^{-1}dC - \frac{1}{2}C^{-1}\left(dC\right)C^{-1}XX^T\right) && \text{(13b)}\\ &= -\frac{1}{2}\mathrm{tr}\left(C^{-1}\left(dC\right)\left(MI - C^{-1}XX^T\right)\right). && \text{(13c)} \end{aligned}$$

A non-trivial solution is,

$$\begin{aligned} 0 &= MI - C^{-1}XX^T && \text{(14a)}\\ &\Downarrow \\ C &= \frac{1}{M}XX^T, && \text{(14b)} \end{aligned}$$

which may be recognized as the log-likelihood estimate of the covariance matrix [6, Theorem 3.2.1]. For $d\mu$ we need only consider the third term in (11), and through similar calculations we find the non-trivial solution to be,

$$\mu = \frac{1}{M}\sum_{m=1}^{M} x_m, \tag{15}$$

Again this is verified to be the standard (biased) log-likelihood estimate.

3 Bayes Estimation Covariance

In the following sections, we only treat estimators for the covariance matrix. Applying Bayes theorem on Mean and Covariances we find that

$$P(\mu, C|x_1, ..., x_m) = \frac{P(x_1, ..., x_m|\mu, C)P(\mu, C)}{P(x_1, ..., x_m)}, \tag{16}$$

where we denote $P(\mu, C|x_1, ..., x_m)$ the posterior, $P(x_1, ..., x_m|\mu, C)$ the likelihood, $P(\mu, C)$ the prior, and $P(x_1, ..., x_m)$ the evidence. The point of Maximum Posterior, also known as Maximum A Posterior (MAP), is found as the maximum as the point of zero partial derivatives by the differential of the log-Posterior w.r.t. μ and C,

$$\begin{aligned} d\log P(\mu, C|x_1, ..., x_m) &= d\log P(x_1, ..., x_m|\mu, C) + d\log P(\mu, C) - d\log P(x_1, ..., x_m) && \text{(17a)}\\ &= d\log P(x_1, ..., x_m|\mu, C) + d\log P(\mu, C). && \text{(17b)} \end{aligned}$$

Again this has an information theoretical equivalent as the differential of minus the two-parts Minimum Description Length (MDL) [7]. In the following we will discuss estimates based on various priors of increasing complexity.

3.1 Gaussian Prior on Covariance Matrices

As the simplest prior we will in the following assume identical and independent Gaussians the elements of μ and C. Independence among other this imply,

$$P(\mu, C) = P(\mu)P(C), \tag{18}$$

which by the logarithm becomes additive terms, and w.r.t. the differential we can henceforth focus on the Bayes estimate of C,

$$P(C) = \left(2\pi s^2\right)^{-\frac{N^2}{2}} \exp\left(-\frac{1}{2s^2}||C-B||^2\right) \tag{19a}$$

$$= \left(2\pi s^2\right)^{-\frac{N^2}{2}} \exp\left(-\frac{1}{2s^2}\mathrm{tr}\left(\left(C-B\right)^T\left(C-B\right)\right)\right), \tag{19b}$$

$$\tag{19c}$$

where B and s are the mean and variances of the covariance matrix. The differential of the log-prior of the covariance matrix is,

$$d\log P(C) = -\frac{1}{2s^2} d\mathrm{tr}\left(\left(C-B\right)^T\left(C-B\right)\right) \tag{20a}$$

$$= -\frac{1}{2s^2}\mathrm{tr}\left(\left(dC^T\right)\left(C-B\right) + \left(C-B\right)^T dC\right) \tag{20b}$$

Combining the above results with (11) and only considering terms involving dC, we now have,

$$0 = d\log P(\mu, C|x_1, ..., x_m) \tag{21a}$$

$$= -\frac{1}{2}\mathrm{tr}\left(MC^{-1}dC - C^{-1}\left(dC\right)C^{-1}XX^T + \frac{1}{s^2}\left(dC^T\right)\left(C-B\right) + \frac{1}{s^2}\left(C-B\right)^T dC\right) \tag{21b}$$

$$= -\frac{1}{2}\mathrm{tr}\left(MC^{-1}dC - C^{-1}XX^TC^{-1}dC + \frac{2}{s^2}\left(C-B\right)^T dC\right) \tag{21c}$$

$$= -\frac{1}{2}\mathrm{tr}\left(\left(MC^{-1} - C^{-1}XX^TC^{-1} + \frac{2}{s^2}\left(C-B\right)^T\right)dC\right), \tag{21d}$$

where we have used that tr is linear, $\mathrm{tr}AB = \mathrm{tr}BA$, and $\mathrm{tr}A = \mathrm{tr}A^T$. A non-trivial solution is found to be,

$$0 = MC^{-1} - C^{-1}XX^TC^{-1} + \frac{2}{s^2}\left(C-B\right)^T. \tag{22}$$

This is a non-linear equation in C, and we may simplify it by pre- and post multiplication with C to yield,

$$0 = MC - XX^T + \frac{2}{s^2}C\left(C-B\right)^T C \tag{23a}$$

$$\Downarrow$$

$$C = \frac{1}{M}\left(XX^T - \frac{2}{s^2}C\left(C-B\right)^T C\right). \tag{23b}$$

The above is seen to be a third degree matrix polynomial in C, where the log-likelihood estimate act as zero order term. A solution may be found by a gradient descent, where

```
function  Lest = bayespca(y,B,sigma,stepsize);
% BAYESPCA A estimation of covariance matrices using Gaussian prior.


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Estimate covariance matrix by bayes method with Gaussian on norm of
% covariance matrix.
%
M = size(y,2);
normdnew = inf;
while normdnew > 0.0001
  %Cest = y*y'/M;
  Cest = B;
  Cest = (ones(size(Cest))+1*randn(size(Cest))).*Cest;
  dnew = Cest - 1/M*(y*y' - 2/sigma^2 * (Cest*Cest'*Cest-Cest*B'*Cest));
  dold = abs(dnew)+1;
  i = 0;
  while (i<1000)
    i=i+1;
    dold = dnew;
    % The following should be,
    % dnew = Cest - 1/M*(y*y' - 2/sigma^2 * (Cest*Cest'*Cest-Cest*B'*Cest));
    % However, this apparently has convergence problems.  Instead we use
    % the symmetry of Cest and B,
    dnew = Cest - 1/M*(y*y' - 2/sigma^2 * (Cest^3-Cest*B*Cest));
    normdnew = max(abs(dnew(:)));
    Cest = Cest - stepsize*dnew;
  end
end
[Xest,Lest,V] = svd(Cest);
```

Figure 1: A Matlab program implementing Bayes estimation for Gaussian Priors.

the gradient already is calculated as the right-hand side of (22), and the solution is sought starting in the log-likelihood estimate. An example of a Matlab program solving for C is given in Figures 1- 2. Outputs are shown in Figure 3 for randomly generated data from a known source with covariance C^*, optimal but in real life unavailable mean $B = C^*$, and various values of s. It is seen, that the estimate is increasingly close to the log-likelihood estimate as s increases, as should be expected.

Convergence of the algorithm may be improved by applying a continuation method, where a sequence of values $s_{\max}, \ldots, s_0$ is chosen, and solutions are found for each s_i using the previous point s_{i-1} as the starting point. The maximal value $s_{\max}$ is typically set to ∞ in which case the prior becomes uniform, and the solution to C is seen to be identical to the log-likelihood estimate.

3.2 Reverse Engineering a Prior

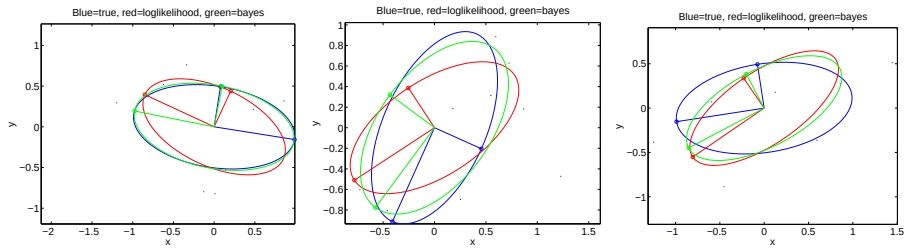
Returning to the study on shape variation of teeth as described in Section 1, we will now interpret the addition of the standard log-likelihood estimate (14) of the covariance matrix by a constant matrix K,

$$C = \frac{1}{M} X X^T + K. \tag{24}$$

as Bayes estimate of the covariance matrix, where K is part the essential part of a prior.

```
function testbayespca
% TESTBAYESPCA A test program for bayes estimation of covariance matrices.

% Global parameters
N = 2; % Dimension of space
M = 10; % Number of samples
Sdiag = [1,2]; % The standard deviation of the datagenerating density
stepsize = 0.01; % Stepsize in the Eulerian solver
sigma = 0.3; % The standard deviation for the Gaussian probability on covariance matrices

% Coordinate system
x = randn(N,N);
for i = 1:N
  for j = i-1:-1:1
    x(:,i) = x(:,i) - (x(:,i)'*x(:,j))*x(:,j);
  end
  x(:,i) = x(:,i)/sqrt(sum(x(:,i).^2));
end
B = x*diag(1./Sdiag.^2)*x'; % Prior mean matrix
a = inv(diag(Sdiag))*randn(N,M); % Random points according to a non isotropic Gaussian distribution
y = x*a;

blue = [0,0,1];
green = [0,1,0];
red = [1,0,0];
plot(y(1,1:min(100,M)),y(2,1:min(100,M)),'.','color',blue);
axis equal
hold on; drawUnitCircle(x,Sdiag,blue); hold off; drawnow;

% Estimate covariance matrix by maximum likelihood.
Cest = y*y'/M;
[Xest,Lest,V] = svd(Cest);
hold on; drawUnitCircle(Xest,sqrt(diag(inv(Lest))),red); hold off; drawnow;

% Estimate covariance matrix by maximum priori.
Lest = bayespca(y,B,sigma,stepsize);
hold on; drawUnitCircle(Xest,sqrt(diag(inv(Lest))),green); hold off; drawnow;
title('Blue=true, red=loglikelihood, green=bayes','fontsize',18);
print 'bayesCovariance.ps' -depsc

function drawUnitCircle(x,Sdiag,color)
% DRAWUNITCIRCLE draws a circle and axes in the x*Sdiag.^2*x' coordinate system
%
  plot([0,x(1,1)]/Sdiag(1),[0,x(2,1)]/Sdiag(1),'-','color',color)
  hold on;
  plot(x(1,1)/Sdiag(1),x(2,1)/Sdiag(1),'o','color',color)
  plot([0,x(1,2)]/Sdiag(2),[0,x(2,2)]/Sdiag(2),'-','color',color)
  plot(x(1,2)/Sdiag(2),x(2,2)/Sdiag(2),'o','color',color)
  t = linspace(0,2*pi,100);
  t = [t,0];
  p = x(1:2,1:2)*inv(diag(Sdiag(1:2)))*[cos(t);sin(t)];
  plot(p(1,:),p(2,:),'-','color',color);
  hold off
  set(gca,'fontsize',18);
  xlabel('x');
  ylabel('y');
```

Figure 2: Example of using the Bayes estimation program in Figure 1.



(a) (b) (c)

Figure 3: Comparing Log-likelihood to Bayes Estimate on 10 randomly sampled points for (a) $s = 0.1$, (b) $s = 0.2$, and (c) $s = 0.3$.

Targeting a Gaussian likelihood we pre multiply (24) with $\frac{1}{2}MC^{-1}(dC)C^{-1}$, rearranging terms, and taking the trace, which gives

$$0 = -\frac{1}{2}\mathrm{tr}\left(MC^{-1}dC - C^{-1}(dC)C^{-1}XX^T - MC^{-1}(dC)C^{-1}K\right) \tag{25a}$$

$$= -\frac{1}{2}\mathrm{tr}\left(MC^{-1}dC - C^{-1}XX^TC^{-1}dC - MC^{-1}(dC)C^{-1}K\right) \tag{25b}$$

Comparing with (11) we realize that the first two terms may be attributed the Gaussian likelihood, and we thus choose to attribute the last term to the prior. This yet unknown prior must have the differential,

$$d\log P(C) = -\frac{1}{2}\mathrm{tr}\left(-MC^{-1}(dC)C^{-1}K\right) \tag{26a}$$

$$= -\frac{1}{2}\mathrm{tr}\left(-MKC^{-1}(dC)C^{-1}\right) \tag{26b}$$

$$= -\frac{1}{2}\mathrm{tr}\left(MKdC^{-1}\right) \tag{26c}$$

Thus, the prior must be a linear function in C^{-1}, and we conclude that

$$P(C) = \exp\left(-\frac{1}{2}\mathrm{tr}\left(MKC^{-1} + D\right)\right) \tag{27a}$$

$$= k\exp\left(-\frac{M}{2}\mathrm{tr}\left(KC^{-1}\right)\right), \tag{27b}$$

for a suitable, constant matrix D normalizing the integral of P by $k = \exp\left(-\frac{1}{2}\mathrm{tr}(D)\right)$. When K is positive semi-definite, this is identified to be an inverted Wishart distribution [6, Chapter 7.7]. A Wishart distribution $W(\Sigma, n)$, is the distribution of matrices X^TX, where the p columns of X are random n-dimensional vectors identically but independently drawn from a normal distribution with mean value μ and covariance Σ, $N(\mu, \Sigma)$. If C is distributed as $W(\Sigma, n)$, then C^{-1} is distributed as (27), where $K = \Sigma^{-1}$.

The matrix K i positive definite, when $f(d_{ij})$ is is diagonally dominant, i.e. for all $i = 1 \ldots M$

$$|f(d_{ii})| \geq \sum_{j=1,\, j\neq i}^{M} |f(d_{ij})|. \tag{28}$$

This is proven as follows: Using the Hadamard product, $A \odot B = \{a_{ij}b_{ij}\}$, we may rewrite (2) as,

$$K = \frac{1}{2}F \odot (N^TN + 1), \tag{29}$$

where $N = [n_1|n_2|\dots|n_M]$. The matrix F has a positive diagonal, since $f(d_{ii}) = 1$. Further, assuming that F is diagonally dominant, we find that $F \odot N^T N$ is diagonally dominant, since the element of $|(N^T N)_{ij}| \leq 1$ and has 1 along the diagonal. Finally, as a consequence of (28), the sum of two diagonally dominant matrices is also diagonally dominant, and we therefore conclude that K is diagonally dominant when F is. A consequence of Gershgorin's theorem [8, As presented in [9]] is that all diagonally dominant matrices with positive diagonals have positive eigenvalues, and therefore we conclude that K is positive definite, when F is diagonally dominant.

4 A Catalog of Priors and Estimators

In the following we will discuss a number of common priors and algorithms for estimating the corresponding covariance matrix.

4.1 Scalar Priors on Matrices

We will in the following consider scalar functions of square matrices, $g : \Re^{N\times N}$, and in particular the following functions $\mathrm{tr}(C)$, $\det(C)$, $||C||$, the later being the two or Frobenius norm. These have differentials given by,

$$d\mathrm{tr}(C) = \mathrm{tr}(dC), \tag{30a}$$

$$d\det(C) = \det(C)\mathrm{tr}(C^{-1}dC), \tag{30b}$$

$$d||C|| = d\sqrt{\mathrm{tr}(C^T C)} = \frac{1}{||C||}\mathrm{tr}(C^T dC) \tag{30c}$$

and which may easily be extended with differentiable functions of these $f : \Re \to \Re$, such that

$$df(g(C)) = f'(g(C))dg(C). \tag{31}$$

This allows us to concentrate on probability distributions of one variable, $P(f(g(C)))$, covering a wide range of distributions on matrices. In the general case

$$d\log P(f(g(C))) = \frac{f'(g(C))}{P(f(g(C)))}dg(C). \tag{32}$$

Distributions based on the exponential function on scalars are popular and particular simple in terms of the differential of their logarithm. E.g. for the following distributions

$$P_{\mathrm{Gauss}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{33a}$$

$$P_{\mathrm{Exp}}(x) = \frac{1}{\mu}\exp\left(\frac{-x}{\mu^2}\right), \quad x \geq 0 \tag{33b}$$

$$P_{\mathrm{LogNormal}}(x) = \frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right), \quad x \geq 0 \tag{33c}$$

the differential of their logarithms are,

$$d\log P_{\mathrm{Gauss}}(x) = \frac{-x}{\sigma^2}dx \tag{34a}$$

$$d\log P_{\mathrm{Exp}}(x) = \frac{-1}{\mu^2}dx, \quad x \geq 0 \tag{34b}$$

$$d\log P_{\mathrm{LogNormal}}(x) = \frac{-\left(\sigma^2 - \mu + \log(x)\right)}{\sigma^2 x}dx, \quad x \geq 0 \tag{34c}$$

For both tr and $||\cdot||$ these distributions further act as independent distributions on the involved entries of C, and since the covariance matrix is known to be positive, semi-definite, and thus $\text{tr}(C) \geq 0$, and $\det(C) \geq 0$.

$$0 = MC^{-1} - C^{-1}XX^TC^{-1} +$$

	P_{Gauss}	P_{Exp}	$P_{\text{LogNormal}}$
$\text{tr}(C)$	$\frac{-\text{tr}(C)}{\sigma^2}$	$\frac{-1}{\mu^2}$	$\frac{-\left(\sigma^2 - \mu + \log(\text{tr}(C))\right)}{\sigma^2 \text{tr}(C)}$
$\det(C)$	$\frac{-\det(C)^2}{\sigma^2}C^{-1}$	$\frac{-1}{\mu^2}\det(C)C^{-1}$	$\frac{-\left(\sigma^2 - \mu + \log(\det(C))\right)}{\sigma^2}C^{-1}$
$\|\|C\|\|$	$\frac{-1}{\sigma^2}C^T$	$\frac{-1}{\mu^2\|\|C\|\|}C^T$	$\frac{-\left(\sigma^2 - \mu + \log(\|\|C\|\|)\right)}{\sigma^2\|\|C\|\|^2}C^T$

(35)

Pre and post multiplication with C and isolating the MC term gives,

$$MC = XX^T +$$

	P_{Gauss}	P_{Exp}	$P_{\text{LogNormal}}$
$\text{tr}(C)$	$\frac{\text{tr}(C)}{\sigma^2}C^2$	$\frac{1}{\mu^2}C^2$	$\frac{\sigma^2 - \mu + \log(\text{tr}(C))}{\sigma^2 \text{tr}(C)}C^2$
$\det(C)$	$\frac{\det(C)^2}{\sigma^2}C$	$\frac{1}{\mu^2}\det(C)C$	$\frac{\sigma^2 - \mu + \log(\det(C))}{\sigma^2}C$
$\|\|C\|\|$	$\frac{1}{\sigma^2}CC^TC$	$\frac{1}{\mu^2\|\|C\|\|}CC^TC$	$\frac{\sigma^2 - \mu + \log(\|\|C\|\|)}{\sigma^2\|\|C\|\|^2}CC^TC$

(36)

These are all non-linear equations, which must be solved iteratively.

4.2 Probability Densities on Logarithm of Matrices

A more complicated model of covariance matrices is to assume that the sorted eigenvalues have a polynomial or exponential relation. That is, covariance matrices are symmetric and semi-definite implying that

$$C = V\Lambda V^T \tag{37}$$

for some orthonormal matrix V of C's eigenvectors and for a diagonal matrix Λ of positive eigenvalues both ordered according to eigenvalue. We thus assume that $\text{diag}(\Lambda)$ has a simple expression,

$$\text{diag}(\Lambda) = h(\Lambda_{ii}), \tag{38}$$

where $h(x)$ is a polynomial or exponential in x. In that case we may easily calculate the matrix logarithm as,

$$C = V\log(\Lambda)V^T, \tag{39}$$

where $\log(\Lambda)$ is the diagonal matrix of $\log(\Lambda_{ii})$. The implication is for that the det and tr operations of C are simply given as,

$$\det C = \prod \text{diag}(\log(\Lambda)), \tag{40a}$$

$$\text{tr}C = \sum \text{diag}(\log(\Lambda)), \tag{40b}$$

and their differentials are

$$d \det C = \sum_i \frac{1}{\Lambda_{ii}} \prod_{j \neq i} \log \Lambda_{jj} d\Lambda_{ii}, \tag{41a}$$

$$d\mathrm{tr}C = \sum_i \frac{1}{\Lambda_{ii}} d\Lambda_{ii}. \tag{41b}$$

These may be used as input to the differential of the exponential type distributions above in the standard way.

For a Gaussian of the norm on the matrix logarithm of the covariance matrix results in a prior term as follows,

$$d \log P_{\mathrm{Gauss}}(|| \log(C)||) = \frac{-1}{2} \mathrm{tr} \left(V (\log(\Lambda))^2 dV^T + \Lambda^{-1} d\Lambda \right) \tag{42}$$

This cannot be refactored into terms only depending on dC, but using

$$dC = (dV)\Lambda V^T + V(d\Lambda)V^T + V\Lambda dV^T, \tag{43}$$

we may refactor the log-likelihood terms in dV and $d\Lambda$.

5 Summary

The maximum posterior method has been developed for many common distributions applied to many common matrix operators, and a program has been written to demonstrate the consequence of applying a prior to covariance estimates. Further, an existing covariance estimation scheme has been reformulated to reveal its corresponding prior.

References

[1] Katrine Hommelhoff Jensen and Jon Sporring. Reconstructing teeth with bite information. In Bjarne Ersbøll and Kim Steenstrup Pedersen, editors, *Proceedings of the Scandinavian Conference on Image Analysis (SCIA '07)*, volume 4522 of *Lecture Notes in Computer Science*, pages 102–111. Springer Verlag, June 2007.

[2] T. F. Cootes and C. J. Taylor. A mixture model for representing shape variation. *Image and Vision Computing*, 17:567–573, 1999.

[3] Fred L. Bookstein. Shape and the information in medical images: A decade of morphometric synthesis. *Computer Vision and Image Understanding*, 66(2):97–118, 1997.

[4] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, 1988.

[5] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, 1949.

[6] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, 3rd edition, 2003.

[7] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[8] S. Gerschgorin. Über die abgrenzung der eignewerte einer matrix. *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk 7*, pages 749–754, 1931.

[9] Gershgorin circle theorem. `http://en.wikipedia.org/wiki/Gershgorin_circle_theorem`, September8 2007.

[10] C.M. Bishop. Bayesian pca. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 382–388, 1999.

[11] P.T. Fletcher, Lu Congliln, S.M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004.