University of Leeds

**SCHOOL OF COMPUTER STUDIES**

**RESEARCH REPORT SERIES**

Report 97.40

# Scaling Additional Contributions to Principal Components Analysis

by

**Roger D Boyle**

August 1997

## Abstract

Principal Components Analysis (PCA) is of great use in representation of multi-dimensional data sets, often providing a useful compression mechanism. Sometimes, input data sets are drawn from disparate domains, such that components of the input are heterogeneous, making them difficult to compare in scale. When this occurs, it is possible for one component to dominate another in the PCA at the expense of the information content of the original data.

We present an approach to balancing the contributions of different components that is constructive; it generalises to the case of the addition of several variables. Conjectures about improved approaches and more complex data sets are presented.

The approach is demonstrated on two current research applications.

# 1 Introduction

The technique of Principal Components Analysis (PCA) has been in use for a considerable time and is well understood. As a method for representing multidimensional data sets it proves useful in pinpointing the nature of variation, and can therefore be used to great effect in data compression by neglecting components of lower importance. Recently, PCA has been used with great success in the computer vision domain in the construction of Point Distribution Models (PDMs) [2, 3], a statistical model of shape which permits non-rigid bodies (for example, faces [6], bones [4, 5] or silhouettes of pedestrians [1]) to be characterised and therefore located in 'difficult' scenes.

PCA accepts a data set $X$ of $n$-dimensional vectors, $X = \{\mathbf{x}_\alpha\}$ , $\alpha = 1, 2, \ldots, N$ and derives a linear transform $P$ to give $\mathbf{x} = \bar{\mathbf{x}} + P\mathbf{b}$ (details are provided below); it is customary to expect the vectors $\mathbf{x}$ to have internal coherence in the sense that the components have corresponding meaning, and therefore scale. An illustration of this would be the derivation of a PDM in which each vector is a concatenation of co-ordinates of boundary points of a shape instance – thus the scale of the components (or, more particularly, the scale of the *variation* in the components) is the same. Various applications exist, however, in which this need not be (or is not) the case – we examine two of these in Section 3, in which the units of measurement of the components of the data vector differ, giving immediate problems in scale comparison.

This paper presents an approach to augmenting data vectors with new components in which the appropriate scale is not known. An argument is given that a particular scale is optimal; the argument is constructive and yields the appropriate factor. Two example applications in which the technique is relevant are briefly described.

# 2 Formulation

PCA is well understood, and a brief summary only is given here. Commencing with a population of $N$ $n$-dimensional vectors,

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

$$\mathbf{x}_\alpha = (x^1_\alpha, x^2_\alpha, \ldots, x^n_\alpha)$$

we compute the covariance matrix C,

$$C_{ij} = Covar(x^i, x^j) = E((x^i - \bar{x}^i)(x^j - \bar{x}^j)) \tag{1}$$

$C$ is a real symmetric matrix and therefore has $n$ real, non-negative eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$; the associated eigen-vectors (appropriately normalised) $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ form an orthonormal basis permitting the vectors $\mathbf{x}$ to be rewritten

$$\mathbf{x} = \bar{\mathbf{x}} + P\mathbf{b} \tag{2}$$

where $P$ is a matrix whose columns are the eigen-vectors of $C$.

Without loss of generality we may assume the eigen-vectors of $C$ are ordered, so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. The eigen-analysis of $C$ is particularly useful; the eigen-vector corresponding to $\lambda_1$ is the direction of maximum variation (the principal component) of the original data $X$; the next most significant eigen-vector (corresponding to $\lambda_2$) is the next most significant, and so on. It is thus possible to use Equation 2 as the basis of a compression scheme; defining $P(k)$ to be the first $k$ columns of $P$, and $\mathbf{b}(k)$ to be the first $k$ components of $\mathbf{b}$, we have

$$\mathbf{x} \approx \bar{\mathbf{x}} + P(k)\mathbf{b}(k)$$

Since the higher indexed elements of $\mathbf{b}$ are minimal contributors to the data, for suitable $k$ this truncation can be an efficient representation; indeed, it may have the effect of suppressing noise in the data rather than losing 'true' information.

Consider now the effect of augmenting the original data by a new component $y$ to provide the population $\hat{X}$, so that

$$\hat{\mathbf{x}}_\alpha = (x^1_\alpha, x^2_\alpha, \ldots, x^n_\alpha, y_\alpha)$$

When calculating the covariance matrix of this augmented population, Equation 1 removes the influence of the value of the mean of individual components, but the scale of variation may be unknown. Trivially, if the $x^j$ are measured in miles, whether the $y$ are measured in miles or inches will influence the eigen-analysis. Therefore we write

$$\hat{\mathbf{x}}_\alpha(s) = (x^1_\alpha, x^2_\alpha, \ldots, x^n_\alpha, sy_\alpha) \tag{3}$$

for a scalar $s$, and ask how the choice of $s$ affects the analysis.

The covariance matrix $\hat{C}$ is easily calculated, being $C$ augmented by a column and row

$$\begin{aligned} \hat{C}_{i(n+1)} = \hat{C}_{(n+1)i} &= s\, Covar(x^i, y) \\ &= s\psi_i \qquad i = 1, 2, \ldots, n \\ \hat{C}_{(n+1)(n+1)} &= s^2 Var(y) \\ &= s^2\sigma_y^2 \end{aligned} \tag{4}$$

Well established results [9] inform us that the eigenvalues $\mu_j$ of $\hat{C}$ will interleave those of $C$,

$$\mu_1 \geq \lambda_1 \geq \mu_2 \geq \lambda_2 \geq \ldots \lambda_n \geq \mu_{n+1} \geq 0 \tag{5}$$

Supposing we know the eigen-vectors of $C$ to be $\lambda_1, \lambda_2, \ldots, \lambda_n$, the characteristic polynomial of $\hat{C}$ is straightforward to calculate as

$$P(\mu) = (\mu - s^2\sigma_y^2)\prod_{i=1}^{n}(\mu - \lambda_i) - \sum_{i=1}^{n} s^2\psi_i^2 \prod_{j=1, j\neq i}^{n}(\mu - \lambda_j) \tag{6}$$

Now behaviour in the limiting cases is clear; when $s = 0$, we have $\lambda_i = \mu_i$, $i = 1, 2, \ldots, n$, and $\mu_{n+1} = 0$, while as $s$ grows, so does $\mu_1$, while all the other $\mu_j$ are bounded by the interval in which they lie (Equation 5).

## 2.1 Eigen-variance

The distribution of eigenvalues - or *eigen-spectrum* – of $\hat{C}$ is a measure of the correlation between the variables. For compression purposes, we seek a non-uniform distribution; conversely, when introducing a new variable we are interested in it communicating as much as possible to the model.

Clearly when $s = 0$ there is no contribution from $y$, while as $s$ becomes large the influence of the $y$ component dominates the analysis, ultimately suppressing any contribution from the $x_i$ – between these two extremes we might seek a position in which the eigen-spectrum of $\hat{C}$ provides maximal information, where we are maximising the descriptive power of the new variable $y$. One way of doing this is to ask for the spectrum to exhibit minimum variance, thereby requiring that each component, in a relative sense, contributes as much as it can to the representation.

Specifically, we take the distribution of normalised eigenvalues

$$\begin{aligned} M &= \sum_{j=1}^{n+1} \mu_j \\ \tilde{\mu_i} &= \frac{\mu_i}{M} \end{aligned} \tag{7}$$

and consider the variance of these quantities

$$V = V(s) = \frac{1}{n+1}\sum_{j=1}^{n+1} \tilde{\mu}_j^2 \quad - \quad \frac{1}{(n+1)^2} \tag{8}$$

We call $V$ the *eigen-variance*. As $s \to \infty$ (at which $\mu_1$ dominates all other eigenvalues), $V(s) \to 1 - \frac{1}{n+1}$. The minimum value $V(s) = 0$ is assumed when all eigenvalues are equal (and there is no correlation at all between the components of the original data).

**Theorem 1:** For $s \geq 0$, $V(s)$ is unimodal, reaching a minimum at

$$s = \sqrt{\frac{\sigma_y^2 \sum_{i=1}^n \lambda_i^2 - \sum_{i=1}^n \psi_i^2 \sum_{i=1}^n \lambda_i}{\sigma_y^2(\sigma_y^2 \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \psi_i^2)}}$$

**Proof:** If the characteristic polynomial of $\hat{C}$ is written

$$P(\mu) = a_0 + a_1\mu + a_2\mu^2 + \ldots + a_n\mu^n + \mu^{n+1}$$

then

$$\begin{aligned} -a_n &= \sum_{j=1}^{n+1} \mu_j = M \\ a_{n-1} &= \frac{1}{2}\sum_{j=1}^{n+1}\sum_{k=1,k\neq j}^{n+1} \mu_j\mu_k \end{aligned}$$

Then

$$\begin{aligned} V(s) &= \frac{1}{n+1}\sum_{j=1}^{n+1}\tilde{\mu}_j^2 \quad - \quad \frac{1}{(n+1)^2} \\ &= \frac{1}{M^2(n+1)}\sum_{j=1}^{n+1}\mu_j^2 \quad - \quad \frac{1}{(n+1)^2} \\ &= \frac{1}{M^2(n+1)}((\sum_{j=1}^{n+1}\mu_j)^2 - \sum_{j=1}^{n+1}\sum_{k=1,k\neq j}^{n+1}\mu_j\mu_k) \quad - \quad \frac{1}{(n+1)^2} \\ &= \frac{1}{a_n^2(n+1)}(a_n^2 - 2a_{n-1}) \quad - \quad \frac{1}{(n+1)^2} \\ &= \frac{n}{(n+1)^2} \quad - \quad \frac{2a_{n-1}}{a_n^2(n+1)} \end{aligned} \tag{9}$$

From Equation 6

$$\begin{aligned} a_n &= -(s^2\sigma_y^2 + \sum_{i=1}^{n}\lambda_i) \\ a_{n-1} &= \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}\lambda_j\lambda_k \; + \; s^2\sigma_y^2\sum_{i=1}^{n}\lambda_i \; - \; s^2\sum_{i=1}^{n}\psi_i^2 \end{aligned} \tag{10}$$

Writing

$$\begin{aligned} \Lambda = \textstyle\sum_{i=1}^{n}\lambda_i \qquad & \Lambda_x = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}\lambda_j\lambda_k \\ \Lambda_2 = \textstyle\sum_{i=1}^{n}\lambda_i^2 \qquad & \Psi = \sum_{i=1}^{n}\psi_i^2 \end{aligned} \tag{11}$$

Equations 9 and 10 give

$$V(s) = \frac{n}{(n+1)^2} \; - \; \frac{2(\Lambda_x + s^2\sigma_y^2\Lambda - s^2\Psi)}{(s^2\sigma_y^2 + \Lambda)^2(n+1)}$$

and hence

$$\begin{aligned}
\frac{dV}{ds} &= \frac{8s\sigma_y^2(\Lambda_x + s^2\sigma_y^2\Lambda - s^2\Psi)}{(s^2\sigma_y^2+\Lambda)^3(n+1)} - \frac{4s(\sigma_y^2\Lambda - \Psi)}{(s^2\sigma_y^2+\Lambda)^2(n+1)} \\
&= \frac{4s(2\sigma_y^2(\Lambda_x + s^2\sigma_y^2\Lambda - s^2\Psi) - (s^2\sigma_y^2+\Lambda)(\sigma_y^2\Lambda - \Psi))}{(s^2\sigma_y^2+\Lambda)^3(n+1)} \\
&= \frac{4s(s^2(\sigma_y^4\Lambda - \sigma_y^2\Psi) - (\sigma_y^2\Lambda^2 - 2\sigma_y^2\Lambda_x - \Psi\Lambda))}{(s^2\sigma_y^2+\Lambda)^3(n+1)} \\
&= \frac{4s(s^2\sigma_y^2(\sigma_y^2\Lambda - \Psi) - (\sigma_y^2\Lambda_2 - \Psi\Lambda))}{(s^2\sigma_y^2+\Lambda)^3(n+1)} \qquad (12)
\end{aligned}$$

Equating $\frac{dV}{ds}$ to 0 gives

$$\begin{aligned}
s &= 0 \\
s &= \sqrt{\frac{\sigma_y^2\Lambda_2 - \Psi\Lambda}{\sigma_y^2(\sigma_y^2\Lambda - \Psi)}} \\
s &= -\sqrt{\frac{\sigma_y^2\Lambda_2 - \Psi\Lambda}{\sigma_y^2(\sigma_y^2\Lambda - \Psi)}} \qquad (13)
\end{aligned}$$

The quantity under the square root is positive, unless $y_\alpha = k, \alpha = 1, 2, \ldots, N$ for some constant $k$, in which case it is 0 and $V(s)$ is constant. To see this, we write

$$\begin{aligned}
Z = Z(\mathbf{y}) &= \sigma_y^2\Lambda_2 - \Psi\Lambda \\
&= \frac{\Lambda_2}{N}\sum_{\alpha=1}^{N}(y_\alpha - \bar{y})^2 - \frac{\Lambda}{N^2}\sum_{i=1}^{n}(\sum_{\alpha=1}^{N}(x_\alpha^i - \bar{x}^i)(y_\alpha - \bar{y}))^2
\end{aligned}$$

Then

$$\begin{aligned}
\frac{dZ}{dy_\alpha} &= \frac{\Lambda_2}{N}(2(y_\alpha - \bar{y})(1 - \frac{1}{N})) - \frac{\Lambda}{N^2}\sum_{i=1}^{n}(2(x_\alpha^i - \bar{x}^i)(y_\alpha - \bar{y})(1 - \frac{1}{N})) \\
&= \frac{\Lambda_2}{N}(1 - \frac{1}{N})(y_\alpha - \bar{y})
\end{aligned}$$

Thus these derivatives are simultaneously 0 when all the $y_\alpha$ are equal; clearly this provides a minimum at which $Z = 0$. Note also that

$$\begin{aligned}
\sigma_y^2\Lambda - \Psi &= \sigma_y^2\Lambda^2 - \Lambda\Psi \\
&\geq \sigma_y^2\Lambda_2 - \Lambda\Psi \\
&= Z
\end{aligned}$$

and thus the denominator is non-negative too.

The limiting case of $y_\alpha = k$ is of no interest since it provides nothing toward the description of the population $\hat{X}$.

■

Thus we have a unique positive value of $s$ that minimises $V(s)$; Figure 1 illustrates $V$ for a particular data set.

## 2.2 Extension to several variables

It is natural to consider extending this approach to more than one additional variable; thus, instead of equation 3, we may have

$$\hat{\mathbf{x}}_\alpha(s) = (x^1_\alpha, x^2_\alpha, \ldots, x^n_\alpha, sy^1_\alpha, sy^2_\alpha, \ldots, sy^k_\alpha) \tag{14}$$

Following the same approach as in Section 2.1, we seek the terms in $\mu^{n+k-1}$ and $\mu^{n+k-2}$ of the characteristic polynomial of the covariance matrix $\hat{C}$ derived from these data. Define

$$\psi_{ij} = Covar(x^i, y^j)$$

If we suppose that the eigenvalues of the augmented data $\mathbf{y}_\alpha = (y^1_\alpha, y^2_\alpha, \ldots, y^k_\alpha)$ are $\omega_1, \omega_2, \ldots, \omega_k$, it is possible to find a rotation of the augmented data whose covariance matrix is

$$\begin{aligned} \hat{C}_{ii} &= \lambda_i, & i &= 1, 2, \ldots, n \\ \hat{C}_{(n+i)(n+i)} &= s^2\omega_i, & i &= 1, 2, \ldots, k \\ \hat{C}_{i(n+j)} = \hat{C}_{(n+j)i} &= s\psi_{ij}, & i &= 1, 2, \ldots, n; j = 1, 2, \ldots, k \end{aligned} \tag{15}$$

with all other entries 0.

For convenience, we will write

$$\begin{aligned} \Omega &= \textstyle\sum_{i=1}^k \omega_i & \Omega_x &= \frac{1}{2}\displaystyle\sum_{j=1}^{k}\sum_{k=1,k\neq j}^{k} \omega_j\omega_k \\ \Omega_2 &= \textstyle\sum_{i=1}^k \omega_i^2 & \Psi &= \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{k} \psi_{ij}^2 \end{aligned} \tag{16}$$

with $\Lambda, \Lambda_x, \Lambda_2$ as before. Analogously to Equation 10, it is then straightforward to see that the relevant polynomial coefficients are

$$\begin{aligned} a_n &= -(L + s^2\Omega) \\ a_{n-1} &= \Lambda_x + s^4\Omega_x + s^2\Lambda\Omega - s^2\Psi \end{aligned} \tag{17}$$

Hence

$$V(s) = \frac{n}{(n+1)^2} - \frac{2(\Lambda_x + s^4\Omega_x + s^2\Lambda\Omega - s^2\Psi)}{(s^2\Omega + \Lambda)^2(n+1)}$$

and

$$\frac{dV}{ds} = \frac{4s(s^2(\Lambda\Omega_2 - \Psi\Omega) - (\Lambda_2\Omega - \Psi\Lambda))}{(s^2\Omega + \Lambda)^3(n+1)}$$

whence, for $s > 0$, equating $\frac{dV}{ds}$ to 0 gives

$$s = \sqrt{\frac{\Lambda_2\Omega - \Psi\Lambda}{\Lambda\Omega_2 - \Psi\Omega}} \tag{18}$$

A similar argument to before guarantees that this quantity is real.

More generally, it might not be reasonable to assume that the extra variables are 'tied' to each other, in which case each would demand a different scale;

$$\hat{\mathbf{x}}_\alpha(s) = (x_\alpha^1, x_\alpha^2, \ldots, x_\alpha^n, s_1 y_\alpha^1, s_2 y_\alpha^2, \ldots, s_k y_\alpha^k) \tag{19}$$

and we would seek 'optimal' values of $\mathbf{s} = (s_1, s_2, \ldots, s_k)$. The hypersurface $V(\mathbf{s})$ has a unique minimum: We can deduce this by noting that if $V(\mathbf{s})$ had two minima, at $\mathbf{s_1}$ and $\mathbf{s_2}$, say, it would be possible to construct a rotation of the data such that one axis was parallel to the line going $\mathbf{s_1}$ and $\mathbf{s_2}$, and thence to construct a counterexample to Theorem 1.

The location of the minimum is not obvious, although a strategy for the selection of a suboptimal $\mathbf{s}$ exists by adding the variables $y^j$ singly, and selecting each value of $s_j$ via equation 13.

### 2.3 Eigen-entropy

A more common measure of information content than variance is entropy. Corresponding to Equation 8 we might define the *eigen-entropy* as

$$\begin{aligned} E = E(s) &= -\sum_{i=1}^{n+1} \tilde{\mu}_i \log(\tilde{\mu}_i) \\ &= \log(M) - \frac{1}{M} \sum_{i=1}^{n+1} \mu_i \log(\mu_i) \end{aligned} \tag{20}$$

and then ask where maxima of this quantity lie.

In all the data sets we have observed – both single and multi-variable – the quantity $E(s)$ has been unimodal, exhibiting a maximum at which the information content of the eigenvalues is at a peak, Further, this peak is achieved by a value of $s$ (or $\mathbf{s}$) very close to the value that provides the minimum of $V(s)$. It remains a conjecture that this quantity is indeed unimodal for suitable $s$ ($\mathbf{s}$), and unknown as to what that mode may be.

Figure 1 illustrates $E$ for a particular data set augmented by one variable $y$.

## 3 Applications

This work has been prompted by two independent applications.

### 3.1 The 'Robotic Sheepdog' project

The 'robotic sheepdog' project is constructing an autonomous robot that has the intelligence to herd living creatures (for the purposes of demonstration, ducks). The project and its motivation are described elsewhere [7, 8].

The robot receives its information about its own and the ducks' parameters (position and velocity) via a camera monitoring the scene. The ducks behave in a *flocking* manner, with

the coarse shape of the flock being a good indicator of their behaviour - it is thus natural to model the flock using the Point Distribution Models which have proved very successful in modelling non-rigid shapes in other domains. This approach has been seen to work well.

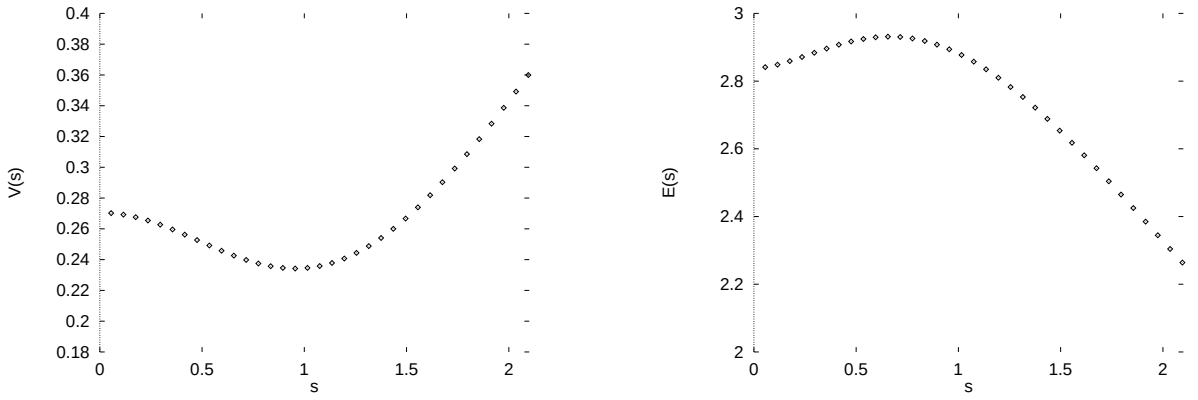


Figure 1: Data describing duck flock shape (concatenation of co-ordinate pairs) augmented by robot distance: (a) $V(s)$, (b) $E(s)$.

The PDM so constructed is interesting but solely descriptive, while a predictive model is sought. In order to proceed it is necessary to add to the model information about the velocity of the flock, and the (relative) position and velocity of the robot. Figure 2 illustrates these essential parameters.

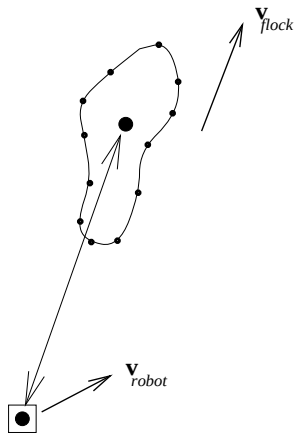


Figure 2: Parameters of the duck tracking system: Robot position and velocity, duck flock shape, position and velocity; separating distance and direction.

We thus have parameters as follows

- Flock shape: $N$ Cartesian co-ordinate pairs, used to define a spline ($N = 20$ here).
- Flock centroid velocity: 2 polar co-ordinates
- Robot distance and direction from flock centroid: 2 polar co-ordinates
- Robot velocity: 2 polar co-ordinates

We thus need to add six dimensions to the data used to describe shape alone; this we have done using the approach outlined in Section 2.2. The orthogonal eigen-system so derived contains significant redundancy (correlation between the components) and, after truncation of the less significant eigen-vectors from the representation, is being used as the basis of a predictor of duck behaviour derived from actual observation.

As an illustration, Figures 1 and 3 consider the addition of solely the robot distance parameter, and illustrate the eigen-spectrum in the cases

- $s = 0$: This distribution is exactly that of the unaugmented data set.
- $s = 0.955$: This is the optimum derived from Equation 13, so this distribution of values carries maximal information according to the definition of minimum variance. Figure 1 shows that a lower value would give maximal information according to an entropy definition.
- $s = 1.8$: Here it is clear that the most significant eigenvalue is already dominating significantly.

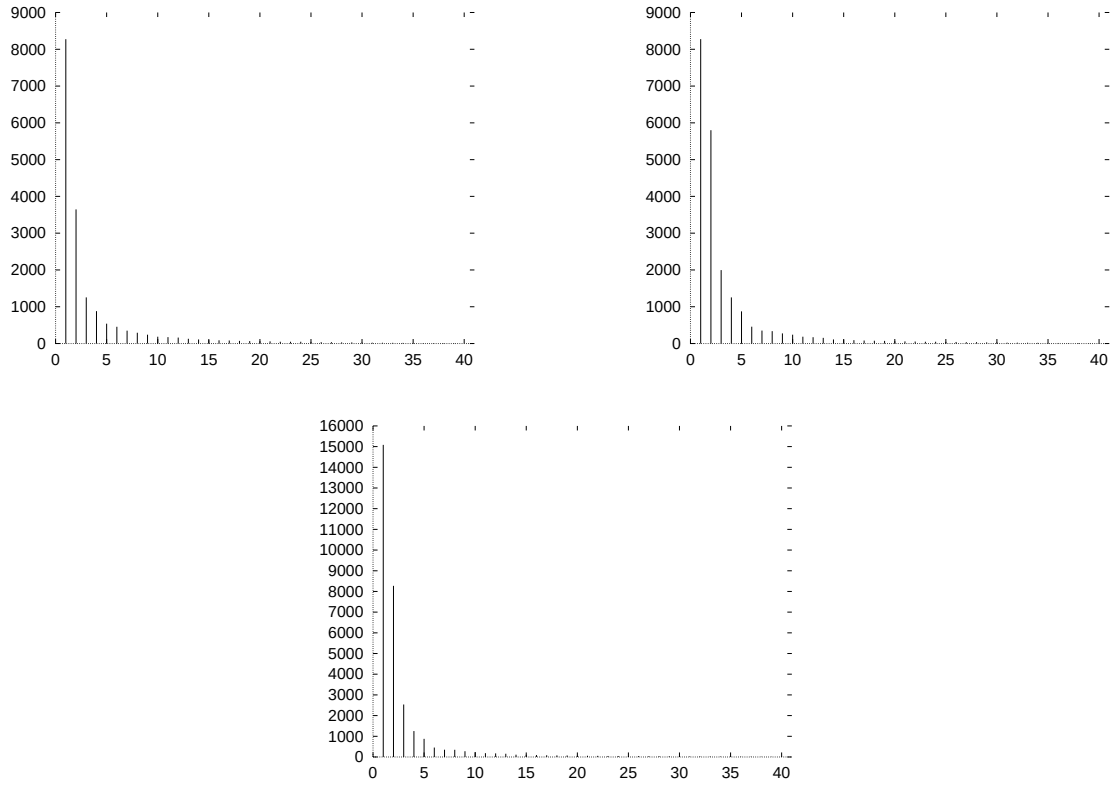


Figure 3: The distribution of eigenvalues for factors of the robot distance: (a) $s = 0$, $V(s) = 0.271$; (b) $s = 0.955$, $V(s) = 0.234$; (c) $s = 1.8$, $V(s) = 0.309$.

In fact, for the purposes of the demonstration we need also to add in a parameter describing the ducks' proximity to the boundary of their arena, since this also influences their behaviour. It is also possible that we should consider the acceleration of both the flock centroid and the robot. In due course we shall also try to observe the behaviour of individual birds. All of these extra parameters may be included in the descriptive model using the approach we have described.

## 3.2 Analysis of motorway flow data

The M25 motorway is an orbital highway around the city of London of circumference in excess of 300 kilometres; for the majority of its length it carries four lanes of traffic in both directions. The usually heavy flow is monitored at minute intervals by detectors placed approximately 500 metres apart; at each detector station four quantities are measured in each lane – speed (kilometres per hour), flow (vehicles per hour), occupancy (percentage of time the detector is 'occupied') and headway (the time interval between the same point of successive vehicles). This apparatus generates very significant quantities of uninterrupted data that describe a highly complex and dynamic system (although it may be immediately apparent that the data will contain significant redundancy).

It is common to consider sub-networks for monitoring and diagnostic purposes – for example, the detectors in the neighbourhood of an interchange, or a small number of consecutive detectors on a junction-free stretch; at simplest, we might consider two detectors that would describe the behaviour of approximately one kilometre of road by delivering (in each direction) a 32-dimensional vector every minute.

The behaviour of traffic makes these raw data very noisy, although patterns are clearly visible within it; for example, rush hour behaviour is clearly different to steady state mid-day flow or the light flow of the night. It is also apparent when acute congestion occurs, and that the 'recovery' of flow from an incident is different again. It is useful therefore to consider approaches to extracting these patterns from the mass of data. One approach to doing this is to perform a PCA and to consider only the components that dominate the variations.

| | $\mu$ | $\sigma^2$ |
|---|---|---|
| Speed | 98.1 | 12.2 |
| Flow | 1291.2 | 48204.2 |
| Occupancy | 10.9 | 5.2 |
| Headway | 2.8 | 0.3 |

Table 1: Data derived from clockwise lane 2 M25 flow on 8th October 1996, 0700 - 0900 from detector 4747b (near junction 10, Wisley).

Table 1 illustrates a small snapshot of behaviour; it is clear that the scale of the measurements differs widely, and in constructing a PCA suitable choices of multiplying factors need to be made.

This preliminary experiment combined observations from two consecutive detectors near Wisley; arbitrarily, we perform an analysis of the speed (eight measurements) which we may assume come from a homogeneous sub-population, and then extend the model by combining in the eight measurements for flow using Equation 18. The factor delivered in this instance is 0.01335, which gives an optimal model according to the derivation of Section 2.2. Combining further with occupancy and then headway extends the model further, but suboptimally; this can be seen by comparing results from different data combinations. Table 2 illustrates the eigen-variances and compactness, measured as the percentage variation described by the first three and five principal components, of a selection of models, built from these data.

| Components | Dimension | Eigen-variance | % $\lambda_1$ - $\lambda_3$ | % $\lambda_1$ - $\lambda_5$ |
|---|---|---|---|---|
| S | 8 | 0.252 | 63.6 | 83.7 |
| SF | 16 | 0.187 | 57.0 | 69.5 |
| SFO | 24 | 0.176 | 54.7 | 68.8 |
| SFOH | 32 | 0.161 | 52.4 | 66.2 |
| SH | 16 | 0.174 | 54.1 | 69.6 |
| SHF | 24 | 0.164 | 53.2 | 66.2 |
| SHFO | 32 | 0.157 | 51.8 | 64.8 |

Table 2: Results from building composite models from four components: S=speed, F=flow, O=Occupancy, H=Headway. Note that the order of combination affects the result.

This work is ongoing and is being used to provide an alarm mechanism to indicate 'incidents' during normal flow, which itself varies recognizably and which may be identified from this model.

# 4 Conclusion

We have presented an approach to extending Principal Components Analysis when the input variables are from different domains, and their relative scale is unknown. The approach considers distributions of eigenvalues of covariance matrices and is constructive; in simple cases it is optimal according to an elementary definition.

Two examples of the algorithm in use have been mentioned, both of which depend upon extracting as much as possible from disparate data sources. Earlier uses of these data had taken inefficient approaches (for example, scaling all inputs to be zero mean, unit variance), which had the effect of reducing the descriptive power of some data components. It seems probable that many other applications exist.

It has been noted that the definition of 'information' used to derive the algorithm is not the usual (or best) one, and it remains to demonstrate that using entropy maximisation rather than variance minimisation can deliver a constructive solution, although empirically we discover it can, and such a solution is remarkably close to that given. It also remains to derive an optimal solution in the case of data being augmented by more than one new data subset.

# 5 Acknowledgements

Acknowledgements are due to Neil Sumpter for data on the Robotic Sheepdog, and Haibo Chen for data from the M25. Jenny Gregory is to be thanked for conducting preliminary experiments that demonstrated the need for this analysis.

# References


[1] A M Baumberg and D C Hogg. Learning flexible models from image sequences. In *Proceedings of the European Conference on Computer Vision*, pages 299–308, May 1994.

[2] T F Cootes and C J Taylor. Active shape models – 'smart snakes'. In D C Hogg and R D Boyle, editors, *Proceedings of the British Machine Vision Conference, Leeds, UK*, pages 266–275. Springer Verlag, 1992.

[3] T F Cootes, C J Taylor, D H Cooper, and J Graham. Training models of shape from sets of examples. In D C Hogg and R D Boyle, editors, *Proceedings of the British Machine Vision Conference, Leeds, UK*, pages 9–18. Springer Verlag, 1992.

[4] N D Efford. Knowledge-based segmentation and feature analysis of hand and wrist radiographs. *Proc. SPIE*, 1905:596–608, 1993.

[5] N D Efford. Knowledge generation techniques for the model-driven segmentation of hand and wrist radiographs. In *Proceedings of the Eighth Scandinavian Conference on Image Analysis*, pages 251–256, Tromso, Norway, May 1993.

[6] A Lanitis, C J Taylor, and T F Cootes. An automatic face identification systems using flexible appearance models. In E Hancock, editor, *Proceedings of the British Machine Vision Conference, York, UK*, volume 1, pages 65–74. BMVA Press, 1994.

[7] N Sumpter, R D Boyle, and R Tillet. Modelling collective animal behaviour using extended PDMs. In A Clark, editor, *Proceedings of the British Machine Vision Conference, 1997*, Colchester, UK, 1997. To appear.

[8] R Vaughn, J Henderson, and N Sumpter. Introducing the robot sheepdog project. In *Proceedings of the International Workshop on Robotics and Automated Machinery for Bio-productions*, Valencia, Spain, September 1997. To appear.

[9] J H Wilkinson. *The algebraic eigenvalue problem.* Clarendon Press, 1965.