NEURAL COMPUTING
NCRG
RESEARCH GROUP

Neural Computing Research Group
Dept of Computer Science & Applied Mathematics
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Probabilistic Principal Component Analysis

Michael E. Tipping
M.E.Tipping@aston.ac.uk

Christopher M. Bishop
C.M.Bishop@aston.ac.uk

Technical Report NCRG/97/010 September 4, 1997

Submitted for publication.

Abstract

Principal component analysis (PCA) is a ubiquitous technique for data analysis and processing, but one which is not based upon a probability model. In this paper we demonstrate how the principal axes of a set of observed data vectors may be determined through maximum–likelihood estimation of parameters in a latent variable model closely related to factor analysis. We consider the properties of the associated likelihood function, giving an EM algorithm for estimating the principal subspace iteratively, and discuss the advantages conveyed by the definition of a probability density function for PCA.

1 Introduction

Principal component analysis (PCA) (Jolliffe 1986) is a well–established technique for dimension reduction, and a chapter on the subject may be found in practically every text on multivariate analysis. Examples of its many applications include data compression, image processing, visualization, exploratory data analysis, pattern recognition and time series prediction.

The most common derivation of PCA is in terms of a standardised linear projection which maximises the variance in the projected space (Hotelling 1933). For a set of observed d-dimensional data vectors $\{\mathbf{t}_n\}$, $n \in \{1 \dots N\}$, the q *principal axes* $\mathbf{w}_j$, $j \in \{1 \dots q\}$, are those orthonormal axes onto which the retained variance under projection is maximal. It can be shown that the vectors $\mathbf{w}_j$ are given by the q dominant eigenvectors (i.e. those with the largest associated eigenvalues λ_j) of the sample covariance matrix $\mathbf{S} = E[(\mathbf{t} - \boldsymbol{\mu})(\mathbf{t} - \boldsymbol{\mu})^{\mathrm{T}}]$ such that $\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j$. The q principal components of the observed vector $\mathbf{t}_n$ are given by the vector $\mathbf{x}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$, where $\mathbf{W}^{\mathrm{T}} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_q)^{\mathrm{T}}$. The variables x_j are then decorellated such that the covariance matrix $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ is diagonal with elements λ_j.

A complementary property of PCA, and that most closely related to the original discussions of Pearson (1901) is that, of all orthogonal linear projections $\mathbf{x}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$, the principal component projection minimises the squared reconstruction error $\sum_n \| \mathbf{t}_n - \hat{\mathbf{t}}_n \|^2$, where the optimal linear reconstruction of $\mathbf{t}_n$ is given by $\hat{\mathbf{t}}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}$.

One limiting disadvantage of both these definitions of PCA is the absence of a probability density model and associated likelihood measure. Deriving PCA from the perspective of density estimation would offer a number of important advantages including:

- The definition of a likelihood measure permits comparison with other density–estimation techniques and facilitates statistical testing.
- Bayesian inference methods may be applied (e.g. for model comparison) by combining the likelihood with a prior.
- If PCA is used to model the class–conditional densities in a classification problem, the posterior probabilities of class membership may be computed.
- The probability density function gives a measure of the novelty of a new data point.
- The single PCA model may be extended to a mixture of such models.

The key result of this paper is to show that principal component analysis may indeed be obtained from a probability model. This follows from incorporating $\mathbf{W}$ within a particular form of latent variable density model which is closely related to statistical factor analysis. Under this formulation, the maximum–likelihood estimator of $\mathbf{W}$ is the matrix of (scaled and rotated) principal axes of the data. Estimation of $\mathbf{W}$ in this way, using an iterative EM algorithm for example, is generally more computationally expensive than the standard eigen-decomposition approach. However, using the given derivation $\mathbf{W}$ may be computed in the standard fashion and subsequently incorporated in the model in order to realise the advantages listed above.

In the next section we briefly introduce the concept of latent variable models, and outline factor analysis in particular. Section 3 then shows how principal component analysis emerges from a particular model parameterisation, and we conclude with a discussion in Section 4. Proofs of key results are left to the appendix.

2 Latent Variable Models

A latent variable model seeks to relate the set of d–dimensional observed data vectors $\{\mathbf{t}_n\}$ to a corresponding set of q–dimensional latent variables $\{\mathbf{x}_n\}$:

$$\mathbf{t} = \mathbf{y}(\mathbf{x}; \boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}(\mathbf{x}, \boldsymbol{\theta})$ is a function of the latent variable $\mathbf{x}$ with parameters $\boldsymbol{\theta}$, and $\boldsymbol{\epsilon}$ is an $\mathbf{x}$-independent noise process. Generally, $q < d$ such that the latent variables offer a more parsimonious description of the data. By defining a prior distribution over $\mathbf{x}$, equation (1) induces a corresponding distribution in the data space and the model parameters may be determined by maximum–likelihood.

In standard factor analysis (Bartholomew 1987) the mapping $\mathbf{y}(\mathbf{x}; \boldsymbol{\theta})$ is linear:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{2}$$

where the latent variables $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ have a unit isotropic Gaussian distribution. The error, or noise, model is Gaussian such that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi}$ diagonal, the $(d \times q)$ parameter matrix $\mathbf{W}$ contains the *factor loadings*, and $\boldsymbol{\mu}$ is a constant whose maximum–likelihood estimator is the mean of the data. Given this formulation, the model for $\mathbf{t}$ is also normal $N(\boldsymbol{\mu}, \mathbf{C})$, where the covariance $\mathbf{C} = \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}$. The motivation, and indeed key assumption, for this model is that, because of the diagonality of $\boldsymbol{\Psi}$, the observed variables $\mathbf{t}$ are *conditionally independent* given the values of the latent variables, $\mathbf{x}$. Thus the reduced–dimensional distribution $\mathbf{x}$ is intended to model the dependencies between the observed variables while $\boldsymbol{\epsilon}$ represents the independent noise. This is in contrast to PCA which treats the inter–variable dependencies and the independent noise identically. In factor analysis the columns of $\mathbf{W}$ will generally *not* correspond to the principal subspace of the data. Furthermore, unlike PCA, there is no analytic solution for $\mathbf{W}$ and $\boldsymbol{\Psi}$, and so their values must be determined by iterative procedures. Note also that because of the $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ term, the covariance $\mathbf{C}$, and thus likelihood, is invariant with respect to orthogonal post–multiplication of $\mathbf{W}$. That is, $\mathbf{W}\mathbf{R}$, where $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal matrix, gives an equivalent $\mathbf{C}$.

3 A Probability Model for PCA

Because of the diagonal noise model $\boldsymbol{\Psi}$, the factor loadings $\mathbf{W}$ will, in general, differ from the principal axes (even when taking the arbitrary rotation into account). As considered by Anderson (1963), principal components emerge when the data is assumed to comprise a systematic component, plus an independent error term for each variable with common variance σ^2. This implies that the diagonal elements of the error matrix $\boldsymbol{\Psi}$ in factor analysis above should be identical. Indeed, the similarity between the factor loadings and the principal axes has often been observed in situations in which the elements of $\boldsymbol{\Psi}$ are approximately equal (Rao 1955). Basilevsky (1994) further notes that when the model $\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$ is exact, and therefore equal to $\mathbf{S}$, the factor loadings are identifiable and can be determined analytically through eigen–decomposition of $\mathbf{S}$, without resort to iteration.

As well as assuming the accuracy of the model, such observations do not consider the maximum–likelihood context. By considering the model given by (2) with an isotropic noise structure, such that $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$, we show in this paper that even when the covariance model is approximate, the maximum–likelihood estimator $\mathbf{W}_{\mathrm{ML}}$ is that matrix whose columns are the scaled and rotated principal eigenvectors of the sample covariance matrix $\mathbf{S}$. An important consequence of this derivation is that PCA may be expressed in terms of a density model, the definition of which now follows.

3.1 The Probability Model

For the isotropic, noise model $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, equation (2) implies a probability distribution over $\mathbf{t}$–space for a given $\mathbf{x}$ given by

$$p(\mathbf{t}|\mathbf{x}) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2} \parallel \mathbf{t} - \mathbf{W}\mathbf{x} - \boldsymbol{\mu} \parallel^2\right\}. \tag{3}$$

With a Gaussian prior over the latent variables defined by

$$p(\mathbf{x}) = (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right\}, \tag{4}$$

we obtain the marginal distribution of $\mathbf{t}$ in the form

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{5}$$

$$= (2\pi)^{-d/2}|\mathbf{C}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{t}-\boldsymbol{\mu})\right\}, \tag{6}$$

where the model covariance is

$$\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}. \tag{7}$$

Using Bayes' rule, the *posterior* distribution of the latent variables $\mathbf{x}$ given the observed $\mathbf{t}$ may be calculated:

$$\begin{aligned} p(\mathbf{x}|\mathbf{t}) \quad = \quad & (2\pi)^{-q/2}|\sigma^{-2}\mathbf{M}|^{1/2} \times \\ & \exp\left[-\frac{1}{2}\left\{\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}-\boldsymbol{\mu})\right\}^{\mathrm{T}}(\sigma^{-2}\mathbf{M})\left\{\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}-\boldsymbol{\mu})\right\}\right], \end{aligned} \tag{8}$$

where the posterior covariance matrix is given by

$$\sigma^2\mathbf{M}^{-1} = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}. \tag{9}$$

Note that $\mathbf{M}$ is $q \times q$ while $\mathbf{C}$ is $d \times d$.

The log–likelihood of observing the data under this model is:

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^{N} \ln\{p(\mathbf{t}_n)\}, \\ &= -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{N}{2}\mathrm{tr}\left[\mathbf{C}^{-1}\mathbf{S}\right], \end{aligned} \tag{10}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}, \tag{11}$$

the sample covariance matrix of the observed $\{\mathbf{t}_n\}$. The parameters for this model can thus be estimated by maximising the log–likelihood $\mathcal{L}$, and an EM algorithm to achieve this is given in Appendix B.

3.2 Properties of the Maximum–Likelihood Estimators

The log–likelihood (10) is maximised when the columns of $\mathbf{W}$ span the principal subspace of the data. To show this we consider the derivative of (10) with respect to $\mathbf{W}$:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}), \tag{12}$$

which may be obtained from standard matrix differentiation results [see Krzanowski and Marriott 1994, pp 133]. In Appendix A it is shown, with $\mathbf{C}$ given by (7), that the only non–zero stationary points of (12) occur for:

$$\mathbf{W} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{13}$$

where the q column vectors in $\mathbf{U}_q$ are eigenvectors of $\mathbf{S}$, with corresponding eigenvalues in the diagonal matrix $\mathbf{\Lambda}_q$, and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix. Furthermore, it is also shown that the stationary point corresponding to the *global maximum* of the likelihood occurs when $\mathbf{U}_q$ comprises the *principal* eigenvectors of $\mathbf{S}$, and that all other combinations of eigenvectors represent saddle–points of the likelihood surface. Thus, from (13), the columns of the maximum–likelihood estimator $\mathbf{W}_{\mathrm{ML}}$ contain the principal eigenvectors of $\mathbf{S}$, with a scaling determined by the corresponding eigenvalue and the parameter σ^2, and with arbitrary rotation.

It may also be shown that for $\mathbf{W} = \mathbf{W}_{\mathrm{ML}}$, the maximum–likelihood estimator for σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j, \tag{14}$$

which has a clear interpretation as the variance 'lost' in the projection, averaged over the lost dimensions.

It should be noted that the columns of $\mathbf{W}_{\mathrm{ML}}$ are not orthogonal since

$$(\mathbf{W}_{\mathrm{ML}})^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}} = \mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{R}, \tag{15}$$

which is not diagonal for $\mathbf{R} \neq \mathbf{I}$. In common with factor analysis, and indeed many other iterative PCA algorithms, there exists an element of rotational ambiguity. An orthonormal basis for the principal subspace may easily be extracted using standard techniques if required. Furthermore, the actual principal axes may also be determined by noting that equation (15) represents an eigenvector decomposition of $(\mathbf{W}_{\mathrm{ML}})^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}$, where the transposed rotation matrix $\mathbf{R}^{\mathrm{T}}$ is simply the matrix whose columns are the eigenvectors of the $q \times q$ matrix $(\mathbf{W}_{\mathrm{ML}})^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}$.

However, with reference to the optimal reconstruction property of PCA, further processing of the parameters is not necessary. From (8) it may be seen that the *posterior mean* projection of $\mathbf{t}_n$ is given by $\langle\mathbf{x}_n\rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$. When $\sigma^2 \to 0$, $\mathbf{M}^{-1} \to (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}$ and $\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}$ then becomes an orthogonal projection, and so PCA is recovered. However, the density model then becomes singular, and thus undefined, while for $\sigma^2 > 0$, the projection onto the manifold becomes skewed towards the origin as a result of the prior over $\mathbf{x}$. Because of this, $\mathbf{W}\langle\mathbf{x}_n\rangle$ is *not* an orthogonal projection of $\mathbf{t}_n$. However, each data point may still be optimally reconstructed from the latent variable by taking this skewing into account. With $\mathbf{W} = \mathbf{W}_{\mathrm{ML}}$ the required reconstruction is given by

$$\hat{\mathbf{t}}_n = \mathbf{W}_{\mathrm{ML}}\{(\mathbf{W}_{\mathrm{ML}})^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\}^{-1}\mathbf{M}\langle\mathbf{x}_n\rangle, \tag{16}$$

and is derived in Appendix C. Thus the latent variables convey the necessary information to reconstruct the original data vector optimally, even in the case of $\sigma^2 > 0$.

4 Discussion

In this paper we have shown how principal component analysis may be viewed as a maximum–likelihood procedure based on a probability density model of the observed data.

In addition, we have given an EM algorithm for determining the necessary model parameters, and although we are not necessarily advocating that standard principal components should be estimated in this way, the EM algorithm plays a crucial rôle when, for example, extending the approach to mixture models. (Even for standard PCA, there may be an advantage in an iterative

approach for large d since the algorithm derived in this paper requires at most the inversion of a $q \times q$ matrix, in contrast to a full eigen–decomposition of the $d \times d$ covariance matrix. However, in such instances there are other iterative algorithms available.)

Rather than consider the algorithmic perspective of determining principal components, we would emphasise the advantages, outlined in the introduction, of associating a probability model with PCA. In many applications, these advantages may be realised by computing $\mathbf{U}_q$ and $\mathbf{\Lambda}_q$ by standard eigen–decomposition of the covariance matrix, and subsequently incorporating those parameters within the probability model using equations (13) and (14), thereby avoiding the use of the EM algorithm.

In practice, the choice of the isotropic noise covariance $\sigma^2\mathbf{I}$ within the model conveys an advantage over the diagonal covariance $\mathbf{\Psi}$ used in standard factor analysis. In the latter method, considerable care must be taken in the choice of the number of factors q. An inappropriate choice can easily give misleading results, and some practitioners have been quite emphatic in their warnings (notably Chatfield and Collins 1980, chapter 5). A major problem is that if the observations can be explained sufficiently by, say, two factors, a model which attempts to identify only a single factor may often fail to find either of the sufficient two, but may instead find a third alternative. This is ultimately a result of mis–specification of q being compensated for in the factor loadings $\mathbf{W}$, an effect which does not occur in the case of the proposed model for PCA. In this latter case, the use of the isotropic noise model implies that the first two principal axes will clearly include the first alone.

Formulating PCA as a probability model can offer considerable practical benefits. For example, we are currently incorporating individual PCA models in a mixture model framework (Tipping and Bishop 1997a; Tipping and Bishop 1997b). An EM algorithm, based on that given in Appendix B, can be derived for estimating all the model parameters. Such a mixture model has been employed both for image compression, where the optimal linear reconstruction property of PCA can be effectively exploited, and for visualization, where the implicitly defined PCA projections may be utilised.

A further important implication of such an approach to density modelling (either with individual or mixture models) is the capacity to control the model complexity through choice of q, by limiting the number of parameters used to define the covariance structure. This enables density models to be constructed in high–dimensional spaces where fully–parameterised covariance matrices would be hopelessly under–constrained, and at the same time avoiding an inappropriate diagonal or spherical constraint. Classification through the modelling of class–conditional densities can thus become a realistic option even when d is large.

In addition to placing traditional PCA on a more general statistical footing, the probabilistic formalism opens the door to a richer class of density estimation techniques with much scope for practical application. The illustrative examples from the previous paragraph serve to emphasise that the proposed model has considerable potential.

Acknowledgements: This work was supported by EPSRC contract GR/K51808: *Neural Networks for Visualization of High Dimensional Data.*

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics 34*, 122–148.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. London: Charles Griffin & Co. Ltd.

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*. New York: Wiley.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Chatfield, C. and A. J. Collins (1980). *Introduction to Multivariate Analysis*. London: Chapman & Hall.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B 39*(1), 1–38.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 417–441.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Krzanowski, W. J. and F. H. C. Marriott (1994). *Multivariate Analysis Part I: Distributions, Ordination and Inference*. London: Edward Arnold.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series 2*, 559–572.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika 20*, 93–111.

Rubin, D. B. and D. T. Thayer (1982). EM algorithms for ML factor analysis. *Psychometrika 47*(1), 69–76.

Tipping, M. E. and C. M. Bishop (1997a). Hierarchical models for data visualization. In *Proceedings of the IEE Fifth International Conference on Artificial Neural Networks*, Cambridge, pp. 70–75. London: IEE.

Tipping, M. E. and C. M. Bishop (1997b). Mixtures of principal component analysers. In *Proceedings of the IEE Fifth International Conference on Aritificial Neural Networks*, Cambridge, pp. 13–18. London: IEE.

A Maximum-Likelihood PCA

A.1 The Stationary Points of the Log-Likelihood

The expression for the gradient of the log-likelihood (10) with respect to the weight matrix $\mathbf{W}$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}). \tag{17}$$

At the stationary points:

$$\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{C}^{-1}\mathbf{W}, \tag{18}$$

and hence

$$\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}, \tag{19}$$

assuming that $\sigma^2 > 0$, and thus that $\mathbf{C}^{-1}$ exists. This is a necessary and sufficient condition for the density model to remain nonsingular, and we will restrict ourselves to such cases. It will be seen shortly that $\sigma^2 > 0$ if $q < \text{rank}(\mathbf{S})$, so this assumption implies no loss of practicality.

There are three possible classes of solutions to equation (19):

1. $\mathbf{W} = \mathbf{0}$. This will be seen to be a minimum of the log-likelihood.
2. $\mathbf{C} = \mathbf{S}$. This is the case where the covariance model is exact, such as is discussed by Basilevsky (1994). In the context of standard PCA, such a result is only attainable if $q \geq \text{rank}(\mathbf{S})$. For probabilistic PCA it is necessary to consider the case in which the $d - q$ smallest eigenvalues of $\mathbf{S}$ are identical (or trivially, $q = d - 1$), because $\mathbf{C} = \mathbf{S}$ is attainable with $\sigma^2 = \lambda_{\min}$, the smallest eigenvalue of $\mathbf{S}$. As discussed in Section 3, $\mathbf{W}$ is then identifiable since
$$\begin{aligned} \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}} &= \mathbf{S}, \\ \Rightarrow \quad \mathbf{W}\mathbf{W}^{\mathrm{T}} &= \mathbf{S} - \sigma^2\mathbf{I}, \end{aligned} \tag{20}$$
which has a known solution at $\mathbf{W} = \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, where $\mathbf{U}$ is a square matrix whose columns are the eigenvectors of $\mathbf{S}$, with $\mathbf{\Lambda}$ the corresponding diagonal matrix of eigenvalues, and $\mathbf{R}$ is an arbitrary orthogonal (rotation) matrix.
3. $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$, with $\mathbf{W} \neq \mathbf{0}$ and $\mathbf{C} \neq \mathbf{S}$.

We are interested in case 3 where $\mathbf{C} \neq \mathbf{S}$ and the model is approximate. First, we express the weight matrix $\mathbf{W}$ in terms of its singular value decomposition:

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \tag{21}$$

where $\mathbf{U}$ is a $d \times q$ matrix of orthonormal column vectors, $\mathbf{L} = \text{diag}(l_1, l_2, \ldots, l_q)$ is the $q \times q$ diagonal matrix of singular values, and $\mathbf{V}$ is a $q \times q$ orthogonal matrix. Now,

$$\begin{aligned} \mathbf{C}^{-1}\mathbf{W} &= (\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}, \\ &= \mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}, \\ &= \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}(\sigma^2\mathbf{I} + \mathbf{V}\mathbf{L}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}})^{-1}, \\ &= \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}\mathbf{V}(\sigma^2\mathbf{I} + \mathbf{L}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{L})^{-1}\mathbf{V}^{\mathrm{T}}, \\ &= \mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^{\mathrm{T}}. \end{aligned} \tag{22}$$

Then at the stationary points,

$$\begin{aligned} & \mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}, \\ \Rightarrow\ & \mathbf{S}\mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \\ \Rightarrow\ & \mathbf{S}\mathbf{U}\mathbf{L} = \mathbf{U}(\sigma^2\mathbf{I} + \mathbf{L}^2)\mathbf{L}. \end{aligned} \tag{23}$$

For $l_j \neq 0$, equation (23) implies that if $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q]$, then each column vector $\mathbf{u}_j$ must be an eigenvector of $\mathbf{S}$, with corresponding eigenvalue λ_j such that $\sigma^2 + l_j^2 = \lambda_j$, and so

$$l_j = (\lambda_j - \sigma^2)^{1/2}. \tag{24}$$

For $l_j = 0$, $\mathbf{u}_j$ is arbitrary. All potential solutions for $\mathbf{W}$ may thus be written as

$$\mathbf{W} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{25}$$

where $\mathbf{U}_q$ is a $d \times q$ matrix comprising q column eigenvectors of $\mathbf{S}$, and $\mathbf{K}_q$ is a $q \times q$ diagonal matrix with elements:

$$k_j = \begin{cases} \lambda_j, & \text{the corresponding eigenvalue to } \mathbf{u}_j, \text{ or,} \\ \sigma^2, & \end{cases} \tag{26}$$

where the latter case may be seen to be equivalent to $l_j = 0$. Again, $\mathbf{R}$ is an arbitrary orthogonal (rotation) matrix.

A.2 The Global Maximum of the Likelihood

The matrix $\mathbf{U}_q$ may contain any of the eigenvectors of $\mathbf{S}$, so to identify those which maximise the likelihood, the expression for $\mathbf{W}$ in (25) is substituted into the log-likelihood function (10) to give

$$\mathcal{L} = -\frac{N}{2}\left\{ d\ln(2\pi) + \sum_{j=1}^{q'} \ln(\lambda_j) + \frac{1}{\sigma^2}\sum_{j=q'+1}^{d} \lambda_j + (d-q')\ln\sigma^2 + q' \right\}, \tag{27}$$

where q' is the number of non-zero l_j. Differentiating the log-likelihood (10) with respect to σ^2 and substituting for $\mathbf{W}$ from (25) gives

$$\sigma^2 = \frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j, \tag{28}$$

and so

$$\mathcal{L} = -\frac{N}{2}\left\{ \sum_{j=1}^{q'} \ln(\lambda_j) + (d-q')\ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right) + d\ln(2\pi) + d \right\}. \tag{29}$$

Note that (28) implies that $\sigma^2 > 0$ if $\text{rank}(\mathbf{S}) > q$ as stated earlier. We wish to find the maximum of the log-likelihood (29) with respect to the choice of vectors $\mathbf{u}_j$ to incorporate in $\mathbf{W}$. The corresponding 'retained' eigenvalues λ_j, $j \in \{1, \ldots, q'\}$, appear in the first term in (29), while those 'discarded' (and which determine σ^2) are found in the second term. Equation (29) is thus maximised over all possible choices of λ_j when the expression:

$$\sum_{j=1}^{q'} \ln(\lambda_j) + (d-q')\ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right), \tag{30}$$

is minimised. Noting that the right-hand term in (30) is the logarithm of an average, Jensen's inequality can be applied to re-write (30) as

$$\sum_{j=1}^{q'} \ln(\lambda_j) + \sum_{j=q'+1}^{d} \ln(\lambda_j) + A, \tag{31}$$

where $A \geq 0$ represents $(d-q')$ times the difference between the mean of the log-eigenvalues and the log of the mean eigenvalue, and is given by

$$A = (d-q')\ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right) - \sum_{j=q'+1}^{d} \ln(\lambda_j). \tag{32}$$

Since the sum of the first two terms in (31) is constant regardless of the choice of retained or discarded eigenvalues, maximisation of the likelihood is thus equivalent to minimisation of A. We examine this by first assuming that $d - q'$ discarded eigenvalues have been chosen arbitrarily, and, by differentiation, consider how a single such value λ_k affects A:

$$\frac{\partial A}{\partial \lambda_k} = \frac{(d-q')\lambda_k - \sum_{j=q'+1}^{d} \lambda_j}{\lambda_k \left(\sum_{j=q'+1}^{d} \lambda_j \right)}. \tag{33}$$

From (33), it can be seen that A has a single minimum when λ_k is equal to the mean of the remaining discarded eigenvalues λ_j. The eigenvalue λ_k can only take discrete values, but it is evident that if a retained eigenvalue λ_j, $j \in \{1 \dots q'\}$, lies between λ_k and the mean, then exchanging the two eigenvalues will result in a decrease in A and an increase in the likelihood. If we consider that the eigenvalues of $\mathbf{S}$ are ordered, for any combination of discarded eigenvalues which includes a 'gap' occupied by a retained eigenvalue, there will always be a sequence of contiguous eigenvalues with a lower value of A. It follows then that at the minimum of A with respect to all possible λ_k, the discarded eigenvalues λ_j, $j \in \{q'+1 \dots d\}$ must be contiguous within the spectrum of all eigenvalues of $\mathbf{S}$.

Without any additional constraint, no further analytic progress may be made with respect to which continuous block of eigenvalues minimises A. However, equation (24) indicates that not all combinations of retained and discarded eigenvalues are stationary points, and that only those where all retained λ_j are greater than σ^2 can exist. By reference to equation (28), we can deduce from this that the *smallest* eigenvalue must be discarded and included in the right-hand term of (30). Given the requirement that the discarded eigenvalues must be contiguous, A must then be minimised when the smallest $d - q'$ eigenvalues are present in the right-hand term of (30) and so $\mathcal{L}$ is maximised when λ_j, $j \in \{1, \dots, q\}$, are the largest eigenvalues of $\mathbf{S}$.

It should also be noted that A is minimised, with respect to q', when there are fewest terms in the sum in (32) which occurs when $q' = q$ and therefore no l_j is zero. Furthermore, $\mathcal{L}$ is *minimised* when $\mathbf{W} = \mathbf{0}$, which may be seen to be equivalent to the case of $q' = 0$.

A.3 The Nature of Other Stationary Points

If stationary points represented by minor eigenvector solutions are stable maxima, then local maximisation (via an EM algorithm for example) is not guaranteed to converge on the optimal solution comprising the principal eigenvectors. We may show, however, that minor eigenvector solutions are in fact saddle points on the likelihood surface.

Consider a stationary point of the gradient equation (17) at $\widehat{\mathbf{W}} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, where $\mathbf{U}_q$ may contain q arbitrary eigenvectors of $\mathbf{S}$ and $\mathbf{K}_q$ contains either the corresponding eigenvalue or σ^2. Then consider a perturbation to this solution of the form $\mathbf{W} = \widehat{\mathbf{W}} + \epsilon\mathbf{V}\mathbf{R}$, where ϵ is an arbitrarily small constant and the $d \times q$ matrix $\mathbf{V}$ is given by:

$$\mathbf{V} = \begin{bmatrix} \mathbf{u}_i & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}. \tag{34}$$

It will be sufficient to only consider those $\mathbf{u}_i$ that are not in $\mathbf{U}_q$. (A solution with a repeated eigenvector implies one l_j becoming zero and thus a decrease in the likelihood.) Arbitrary permutations of the columns of $\mathbf{V}$ with all valid $\mathbf{u}_i$ thus implies that the resulting vectors $\mathrm{vec}[\mathbf{V}\mathbf{R}]$ are a complete orthogonal basis for the directions of interest on the likelihood surface[1]. The solutions $\widehat{\mathbf{W}}$ will be stable if $\mathrm{vec}[\mathbf{V}\mathbf{R}]^{\mathrm{T}}\mathrm{vec}[\mathbf{G}]$ is negative for all such directions, where $\mathbf{G} = (\partial\mathcal{L}/\partial\mathbf{W})/N$ evaluated at $\mathbf{W} = \widehat{\mathbf{W}} + \epsilon\mathbf{V}\mathbf{R}$. Now, from (17),

$$\begin{aligned} \mathbf{C}\mathbf{G} &= \mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{W}, \\ &= \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} - \mathbf{W}, \\ &= \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}} + \epsilon^2\mathbf{R}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{R})^{-1} - \mathbf{W}, \end{aligned} \tag{35}$$

[1] The 'vec[·]' operator converts a matrix into a vector by 'stacking' its columns one above the other. It thus has the property that $\mathrm{vec}[\mathbf{A}]^{\mathrm{T}}\mathrm{vec}[\mathbf{B}] = \mathrm{tr}\,[\mathbf{A}^{\mathrm{T}}\mathbf{B}]$.

since $\mathbf{V}^{\mathrm{T}}\mathbf{W} = \mathbf{0}$. Ignoring the term in ϵ^2 then gives:

$$\begin{aligned}\mathbf{CG} &= \mathbf{S}(\widehat{\mathbf{W}} + \epsilon\mathbf{VR})(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - (\widehat{\mathbf{W}} + \epsilon\mathbf{VR}),\\ &= \mathbf{S}\widehat{\mathbf{W}}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - \widehat{\mathbf{W}} + \mathbf{S}\epsilon\mathbf{VR}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - \epsilon\mathbf{VR},\\ &= \epsilon\mathbf{SVR}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - \epsilon\mathbf{VR}, \qquad (36)\end{aligned}$$

since $\mathbf{S}\widehat{\mathbf{W}}(\sigma^2\mathbf{I}+\widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}}) = \widehat{\mathbf{W}}$ at the stationary point. Then substituting for $\widehat{\mathbf{W}}$ gives $\sigma^2\mathbf{I}+\widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}} = \mathbf{R}^{\mathrm{T}}\mathbf{K}_q\mathbf{R}$, such that

$$\begin{aligned}\mathbf{CG} &= \epsilon\mathbf{SVR}(\mathbf{R}^{\mathrm{T}}\mathbf{K}_q^{-1}\mathbf{R}) - \epsilon\mathbf{VR}, \text{ so}\\ \mathbf{G} &= \epsilon\mathbf{C}^{-1}\mathbf{V}(\mathbf{\Lambda}_i\mathbf{K}_q^{-1} - \mathbf{I})\mathbf{R}, \qquad (37)\end{aligned}$$

where

$$\mathbf{\Lambda}_i = \begin{bmatrix}\lambda_i & 0 & \dots\\ 0 & 0 & \dots\\ \dots & \dots & \dots\end{bmatrix}, \qquad (38)$$

with λ_i in the corresponding position to $\mathbf{u}_i$ in $\mathbf{V}$. Then

$$\begin{aligned}\mathrm{vec}[\mathbf{VR}]^{\mathrm{T}}\mathrm{vec}[\mathbf{G}] &= \mathrm{tr}\left[\mathbf{G}^{\mathrm{T}}\mathbf{VR}\right],\\ &= \epsilon\mathrm{tr}\left[\mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}_i\mathbf{K}_q^{-1} - \mathbf{I})\mathbf{V}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{VR}\right],\\ &= \epsilon(\lambda_i/k_i - 1)\mathbf{u}_i^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{u}_i, \qquad (39)\end{aligned}$$

where k_i is the value in $\mathbf{K}_q$ in the corresponding position to λ_i. Since $\mathbf{C}^{-1}$ is positive definite, clearly $\mathbf{u}_i^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{u}_i$ is always positive. When $k_i = \lambda_j$, the expression given by (39) is negative (and the maximum a stable one) for $\lambda_i < \lambda_j$. For $\lambda_i > \lambda_j$ the critical point must be a saddle point. If $k_i = \sigma^2$, the stationary point can never be stable since, from (28), σ^2 is the average of $d - q'$ eigenvalues, and so $\lambda_i > \sigma^2$ for at least one of those eigenvalues, *except* when all those eigenvalues are identical. Such a case is considered in the next section.

From this, by considering all possible perturbations $\mathbf{V}$, it can be seen that the only stable maximum occurs when $\mathbf{W}$ comprises the q principal eigenvectors, for which $\lambda_i < \lambda_j$, $\forall i \neq j$.

A.4 Equality of Eigenvalues

Equality of any of the q principal eigenvalues does not affect the presented analysis. However, consideration must be given to the instance when all the $d - q$ minor (discarded) eigenvalue(s) are equal and identical to the smallest principal (retained) eigenvalue(s). (In practice, particularly in the case of sampled covariance matrices, this is unlikely.)

Consider the example of extracting two components from data with a covariance matrix possessing eigenvalues 2, 1 and 1. In this case, the second principal axis is not uniquely defined within the minor subspace. The spherical noise distribution defined by σ^2, in addition to explaining the residual variance, can also optimally explain the second principal component. Because $\lambda_2 = \sigma^2$, the variable l_2 from equation (24) is zero, and $\mathbf{W}$ effectively only comprises a single vector (its two columns will be linearly dependent). The combination of this single vector and the noise distribution represents the maximum of the likelihood.

B An EM Algorithm for PCA

We now derive an EM algorithm for maximising the likelihood (10), following Rubin and Thayer (1982).

In the EM approach, we consider the latent variables $\{\mathbf{x}_n\}$ to be 'missing' data. If their values were known, estimation of $\mathbf{W}$ would be straightforward by maximising the likelihood for the model given by equation (3), which is equivalent to the standard least–squares solution to equation (2). However, for a given $\mathbf{t}_n$, we are ignorant of the value of $\mathbf{x}_n$ which generated it, although we do know the joint distribution of the observed and latent variables, $p(\mathbf{t}, \mathbf{x})$. In the E–step we use this quantity to calculate the *expectation* of the corresponding *complete–data* log–likelihood with respect to the posterior distribution of $\mathbf{x}_n$ given the observed $\mathbf{t}_n$ and the current parameter values. In the M–step, new parameter values $\widetilde{\mathbf{W}}$ and $\widetilde{\sigma}^2$ are determined which maximise the expected complete–data log–likelihood and this is guaranteed to increase the likelihood of interest, $\prod_n p(\mathbf{t}_n)$, unless it is already at a local maximum (Dempster, Laird, and Rubin 1977; Bishop 1995).

The complete–data log–likelihood is given by:

$$\mathcal{L}_C = \sum_{n=1}^{N} \ln\{p(\mathbf{t}_n, \mathbf{x}_n)\}, \tag{40}$$

where, from equations (3) and (4)

$$p(\mathbf{t}_n, \mathbf{x}_n) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{\| \mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu} \|^2}{2\sigma^2}\right\} (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n\right\}. \tag{41}$$

In the E-step, we take the expectation with respect to the distribution $p(\mathbf{x}_n|\mathbf{t}_n, \mathbf{W}, \sigma^2)$:

$$\begin{aligned}\langle \mathcal{L}_C \rangle = \text{constant terms} - \frac{d}{2}\ln\sigma^2 - \sum_{n=1}^{N}\Bigg\{ &-\frac{1}{2}\mathrm{tr}\left[\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle\right] \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left[(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}} - 2(\mathbf{t}_n - \boldsymbol{\mu})\langle \mathbf{x}_n\rangle^{\mathrm{T}}\mathbf{W}^{\mathrm{T}} + \mathbf{W}\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle\mathbf{W}^{\mathrm{T}}\right]\Bigg\}\end{aligned} \tag{42}$$

with

$$\langle \mathbf{x}_n \rangle = (\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}), \tag{43}$$

$$\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} \rangle = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} + \langle \mathbf{x}_n\rangle\langle \mathbf{x}_n\rangle^{\mathrm{T}}. \tag{44}$$

Note that these statistics are computed using the current (fixed) values of the parameters, and that (43) is simply the posterior mean from equation (8), where we exploit the identity that $\mathbf{W}^{\mathrm{T}}(\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1} = (\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}$. This latter form is preferred as we only need invert a $q \times q$ matrix, rather than the $d \times d$ matrix $\mathbf{C}$. Together with (9), this leads to equation (44).

In the M-step, $\langle \mathcal{L}_C \rangle$ is maximised with respect to $\mathbf{W}$ and σ^2 by differentiating equation (42) and setting the derivatives to zero. Calculating these derivatives, substituting for $\langle \mathbf{x}_n \rangle$ and $\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} \rangle$ and some further manipulation leads to the parameter updates:

$$\widetilde{\mathbf{W}} = \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}, \text{ and} \tag{45}$$

$$\widetilde{\sigma}^2 = \frac{1}{d}\mathrm{tr}\left[\mathbf{S} - \mathbf{S}\mathbf{W}\mathbf{M}^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}}\right], \tag{46}$$

where $\mathbf{S}$ and $\mathbf{M}$ are again given by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}, \tag{47}$$

$$\mathbf{M} = \sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W}. \tag{48}$$

Note that the first instance of $\mathbf{W}$ in equation (46) above is the *old* value of the weights, while the second instance $\widetilde{\mathbf{W}}$ is the *new* value calculated from equation (45).

To maximise the likelihood then, in the E–step the necessary statistics from (43) and (44) are calculated implicitly using (47) and (48), and the new parameters calculated in the M–step using (45) and (46). This procedure is repeated until the algorithm is judged to have converged.

C Optimal Least-Squares Reconstruction

One of the motivations for adopting PCA in many applications, notably in data compression, is the property of optimal least-squares linear reconstruction. That is for all orthogonal projections $\mathbf{x} = \mathbf{W}^{\mathrm{T}}\mathbf{t}$ of the data, the least-squares reconstruction error

$$E^2_{\mathrm{rec}} = \frac{1}{N}\sum_{n=1}^{N} \| \mathbf{t}_n - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n \|^2 \tag{49}$$

is minimised when the columns of $\mathbf{W}$ span the principal subspace of the data covariance matrix. (For simplification, and without loss of generality, we assume here that the data has zero mean.)

We may still obtain this property from our probabilistic formalism, without the need to determine the exact orthogonal projection $\mathbf{W}$, by finding the optimal reconstruction of the posterior mean vectors $\langle \mathbf{x}_n \rangle$. To do this we simply minimise

$$E^2_{\mathrm{rec}} = \frac{1}{N}\sum_{n=1}^{N} \| \mathbf{t}_n - \mathbf{B}\langle \mathbf{x}_n \rangle \|^2, \tag{50}$$

over the reconstruction matrix $\mathbf{B}$, which is equivalent to a linear regression problem giving

$$\mathbf{B} = \mathbf{T}^{\mathrm{T}}\langle \mathbf{X} \rangle \left(\langle \mathbf{X} \rangle^{\mathrm{T}} \langle \mathbf{X} \rangle \right)^{-1}, \tag{51}$$

where $\mathbf{T}$ is the $N \times d$ matrix whose rows are $\mathbf{t}_n$ and $\mathbf{X}$ the $N \times q$ matrix with corresponding rows $\langle \mathbf{x}_n \rangle$.

Since, from (43), $\langle \mathbf{X} \rangle = \mathbf{T}\mathbf{W}\mathbf{M}^{-1}$, (51) gives

$$\mathbf{B} = \mathbf{S}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}\mathbf{M}. \tag{52}$$

where $\mathbf{S} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$ and $\mathbf{M} = \sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W}$.

The reconstruction $\hat{\mathbf{t}}_n$ of $\mathbf{t}_n$ is then given by:

$$\begin{aligned} \hat{\mathbf{t}}_n &= \mathbf{B}\langle \mathbf{x}_n \rangle, \\ &= \mathbf{B}\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n, \\ &= \mathbf{S}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n. \end{aligned} \tag{53}$$

Note that in general this projection of $\mathbf{t}_n$ is not orthogonal. However, at the converged solution, with the substitution $\mathbf{W} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, (53) becomes:

$$\hat{\mathbf{t}}_n = \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{M}\langle \mathbf{x}_n \rangle, \tag{54}$$

$$= \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n, \tag{55}$$

which is the expected orthogonal projection. The implication is thus that in the data compression context, at the maximum likelihood solution, the variables $\langle \mathbf{x}_n \rangle$ can be transmitted down the channel and the original data vectors optimally reconstructed using equation (54) given the parameters $\mathbf{W}$ and σ^2. Substituting for $\mathbf{B}$ in equation (50) gives $E^2_{\mathrm{rec}} = (d-q)\sigma^2$ and the noise term σ^2 thus represents the expected squared reconstruction error per 'lost' dimension.