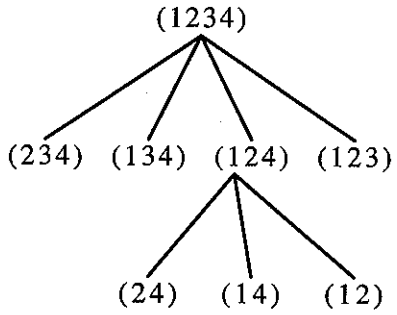


Figure 8.8. Sequential backward elimination of variables, again illustrated for the case of four features. Starting with the complete set, features are eliminated one at a time, such that at each stage the feature chosen for elimination is the one corresponding to the smallest reduction in the value of the selection criterion.

8.6 Principal component analysis

We have already discussed the problems which can arise in attempts to perform pattern recognition in high-dimensional spaces, and the potential improvements which can be achieved by first mapping the data into a space of lower dimensionality. In general, a reduction in the dimensionality of the input space will be accompanied by a loss of some of the information which discriminates between different classes (or, more generally, which determines the target values). The goal in dimensionality reduction is therefore to preserve as much of the relevant information as possible. We have already discussed one approach to dimensionality reduction based on the selection of a subset of a given set of features or inputs. Here we consider techniques for combining inputs together to make a (generally smaller) set of features. The procedures we shall discuss in this section rely entirely on the input data itself without reference to the corresponding target data, and can be regarded as a form of *unsupervised* learning. While they are of great practical significance, the neglect of the target data information implies they can also be significantly sub-optimal, as we discuss in Section 8.6.3.

We begin our discussion of unsupervised techniques for dimensionality reduction by restricting our attention to linear transformations. Our goal is to map vectors $\mathbf{x}^n$ in a d-dimensional space $(x_1, \ldots, x_d)$ onto vectors $\mathbf{z}^n$ in an M-dimensional space $(z_1, \ldots, z_M)$, where $M < d$. We first note that the vector $\mathbf{x}$ can be represented, without loss of generality, as a linear combination of a set of d orthonormal vectors $\mathbf{u}_i$

$$\mathbf{x} = \sum_{i=1}^{d} z_i \mathbf{u}_i \tag{8.12}$$

where the vectors $\mathbf{u}_i$ satisfy the orthonormality relation

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = \delta_{ij} \tag{8.13}$$

in which δ_{ij} is the Kronecker delta symbol defined on page xiii. Explicit expressions for the coefficients z_i in (8.12) can be found by using (8.13) to give

$$z_i = \mathbf{u}_i^{\mathrm{T}} \mathbf{x} \tag{8.14}$$

which can be regarded as a simple rotation of the coordinate system from the original x's to a new set of coordinates given by the z's (Appendix A). Now suppose that we retain only a subset $M < d$ of the basis vectors $\mathbf{u}_i$, so that we use only M coefficients z_i. The remaining coefficients will be replaced by constants b_i so that each vector $\mathbf{x}$ is approximated by an expression of the form

$$\widetilde{\mathbf{x}} = \sum_{i=1}^{M} z_i \mathbf{u}_i + \sum_{i=M+1}^{d} b_i \mathbf{u}_i. \tag{8.15}$$

This represents a form of dimensionality reduction since the original vector $\mathbf{x}$ which contained d degrees of freedom must now be approximated by a new vector $\mathbf{z}$ which has $M < d$ degrees of freedom. Now consider a whole data set of N vectors $\mathbf{x}^n$ where $n = 1, \ldots, N$. We wish to choose the basis vectors $\mathbf{u}_i$ and the coefficients b_i such that the approximation given by (8.15), with the values of z_i determined by (8.14), gives the best approximation to the original vector $\mathbf{x}$ on average for the whole data set. The error in the vector $\mathbf{x}^n$ introduced by the dimensionality reduction is given by

$$\mathbf{x}^n - \widetilde{\mathbf{x}}^n = \sum_{i=M+1}^{d} (z_i^n - b_i)\mathbf{u}_i. \tag{8.16}$$

We can then define the best approximation to be that which minimizes the sum of the squares of the errors over the whole data set. Thus, we minimize

$$E_M = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{x}^n - \widetilde{\mathbf{x}}^n\|^2 = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=M+1}^{d} (z_i^n - b_i)^2 \tag{8.17}$$

where we have used the orthonormality relation (8.13). If we set the derivative of E_M with respect to b_i to zero we find

$$b_i = \frac{1}{N} \sum_{n=1}^{N} z_i^n = \mathbf{u}_i^{\mathrm{T}} \overline{\mathbf{x}} \tag{8.18}$$

where we have defined the mean vector $\overline{\mathbf{x}}$ to be

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^n. \tag{8.19}$$

Using (8.14) and (8.18) we can write the sum-of-squares error (8.17) as

$$E_M = \frac{1}{2}\sum_{i=M+1}^{d}\sum_{n=1}^{N}\left\{\mathbf{u}_i^{\mathrm{T}}(\mathbf{x}^n - \overline{\mathbf{x}})\right\}^2$$

$$= \frac{1}{2}\sum_{i=M+1}^{d}\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_i \tag{8.20}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the set of vectors $\{\mathbf{x}^n\}$ and is given by

$$\boldsymbol{\Sigma} = \sum_n (\mathbf{x}^n - \overline{\mathbf{x}})(\mathbf{x}^n - \overline{\mathbf{x}})^{\mathrm{T}}. \tag{8.21}$$

There now remains the task of minimizing E_M with respect to the choice of basis vectors $\mathbf{u}_i$. It is shown in Appendix E that the minimum occurs when the basis vectors satisfy

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{8.22}$$

so that they are the eigenvectors of the covariance matrix. Note that, since the covariance matrix is real and symmetric, its eigenvectors can indeed be chosen to be orthonormal as assumed. Substituting (8.22) into (8.20), and making use of the orthonormality relation (8.13), we obtain the value of the error criterion at the minimum in the form

$$E_M = \frac{1}{2}\sum_{i=M+1}^{d}\lambda_i. \tag{8.23}$$

Thus, the minimum error is obtained by choosing the $d-M$ smallest eigenvalues, and their corresponding eigenvectors, as the ones to discard.

The linear dimensionality reduction procedure derived above is called the *Karhunen–Loéve transformation* or *principal component analysis* and is discussed at length in Jollife (1986). Each of the eigenvectors $\mathbf{u}_i$ is called a *principal component*. The technique is illustrated schematically in Figure 8.9 for the case of data points in two dimensions.

In practice, the algorithm proceeds by first computing the mean of the vectors $\mathbf{x}^n$ and then subtracting off this mean. Then the covariance matrix is calculated

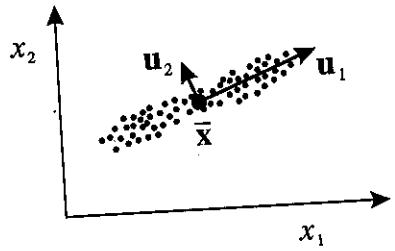


Figure 8.9. Schematic illustration of principal component analysis applied to data in two dimensions. In a linear projection down to one dimension, the optimum choice of projection, in the sense of minimizing the sum-of-squares error, is obtained by first subtracting off the mean $\overline{\mathbf{x}}$ of the data set, and then projecting onto the first eigenvector $\mathbf{u}_1$ of the covariance matrix.

and its eigenvectors and eigenvalues are found. The eigenvectors corresponding to the M *largest* eigenvalues are retained and the input vectors $\mathbf{x}^n$ are projected onto the eigenvectors to give the components of the transformed vectors $\mathbf{z}^n$ in the M-dimensional space. Thus, in Figure 8.9, each two-dimensional data point is transformed to a single variable z_1 representing the projection of the data point onto the eigenvector $\mathbf{u}_1$.

The error introduced by a dimensionality reduction using principal component analysis can be evaluated using (8.23). In some applications the original data has a very high dimensionality and we wish only to retain the first few principal components. In such cases use can be made of efficient algorithms which allow only the required eigenvectors, corresponding to the largest few eigenvalues, to be evaluated (Press *et al.*, 1992).

We have considered linear dimensionality reduction based on the sum-of-squares error criterion. It is possible to consider other criteria including data covariance measures and population entropy. These give rise to the same result for the optimal dimensionality reduction in terms of projections onto the eigenvectors of $\boldsymbol{\Sigma}$ corresponding to the largest eigenvalues (Fukunaga, 1990).

8.6.1 *Intrinsic dimensionality*

Suppose we are given a set of data vectors in a d-dimensional space, and we apply principal component analysis and discover that the first d' eigenvalues have significantly larger values than the remaining $d-d'$ eigenvalues. This tells us that the data can be represented to a relatively high accuracy by projection onto the first d' eigenvectors. We therefore discover that the effective dimensionality of the data is less than the apparent dimensionality d, as a result of correlations within the data. However, principal component analysis is limited by virtue of being a linear technique. It may therefore be unable to capture more complex non-linear correlations, and may therefore overestimate the true dimensionality

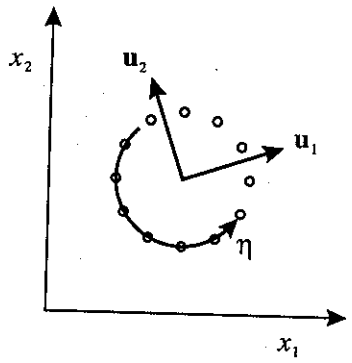


Figure 8.10. Example of a data set in two dimensions which has an intrinsic dimensionality $d' = 1$. The data can be specified not only in terms of the two variables x_1 and x_2, but also in terms of the single parameter η. However, a linear dimensionality reduction technique, such as principal component analysis, is unable to detect the lower dimensionality.

of the data. This is illustrated schematically in Figure 8.10, for data points which lie around the perimeter of a circle. Principal component analysis would give two eigenvectors with equal eigenvalues (as a result of the symmetry of the data). In fact, however, the data could be described equally well by a single parameter η as shown. More generally, a data set in d dimensions is said to have an *intrinsic dimensionality* equal to d' if the data lies entirely within a d'-dimensional subspace (Fukunaga, 1982).

Note that if the data is slightly noisy, then the intrinsic dimensionality may be increased. Figure 8.11 shows some data in two dimensions which is corrupted by a small level of noise. Strictly the data now lives in a two-dimensional space, but can nevertheless by represented to high accuracy by a single parameter.

8.6.2 *Neural networks for dimensionality reduction*

Multi-layer neural networks can themselves be used to perform non-linear dimensionality reduction, thereby overcoming some of the limitations of linear principal component analysis. Consider first a multi-layer perceptron of the form shown in Figure 8.12, having d inputs, d output units and M hidden units, with $M < d$ (Rumelhart *et al.*, 1986). The targets used to train the network are simply the input vectors themselves, so that the network is attempting to map each input vector onto itself. Due to the reduced number of units in the first layer, a perfect reconstruction of all input vectors is not in general possible. The network can be trained by minimizing a sum-of-squares error of the form

$$E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{d}\{y_k(\mathbf{x}^n) - x_k^n\}^2. \tag{8.24}$$

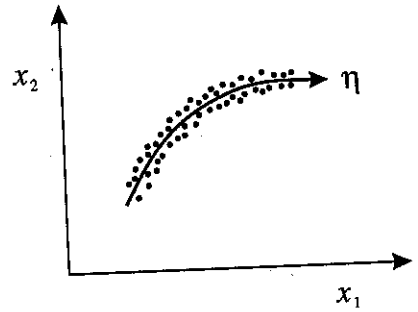


Figure 8.11. Addition of a small level of noise to data in two dimensions having an intrinsic dimensionality of 1 can increase its intrinsic dimensionality to 2. Nevertheless, the data can be represented to a good approximation by a single variable η and for practical purposes can be regarded as having an intrinsic dimensionality of 1.

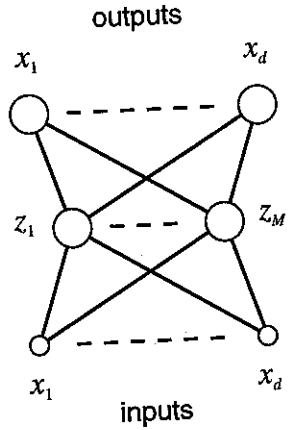


Figure 8.12. An auto-associative multi-layer perceptron having two layers of weights. Such a network is trained to map input vectors onto themselves by minimization of a sum-of-squares error. Even with non-linear units in the hidden layer, such a network is equivalent to linear principal component analysis. Biases have been omitted for clarity.

APPENDIX E

PRINCIPAL COMPONENTS

n Section 8.6, we showed that the optimal linear dimensionality reduction pro-cedure (in the sense of least squares) was determined by minimization of the ollowing function:

$$E_M = \frac{1}{2}\sum_{i=M+1}^{d}\sum_{n}\left\{\mathbf{u}_i^{\mathrm{T}}(\mathbf{x}^n - \overline{\mathbf{x}})\right\}^2$$

$$= \frac{1}{2}\sum_{i=M+1}^{d}\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_i \tag{E.1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix defined by (8.21). We now show that the solu-ion to this problem can be expressed in terms of the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$.

It is clear that (E.1) has a non-trivial minimum with respect to the $\mathbf{u}_i$ only if we impose some constraint. A suitable constraint is obtained by requiring the $\mathbf{u}_i$ o be orthonormal, and can be taken into account by the use of a set of Lagrange nultipliers μ_{ij} (Appendix C). We therefore minimize the function

$$\widehat{E}_M = \frac{1}{2}\sum_{i=M+1}^{d}\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_i - \frac{1}{2}\sum_{i=M+1}^{d}\sum_{j=M+1}^{d}\mu_{ij}(\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j - \delta_{ij}). \tag{E.2}$$

This is conveniently written in matrix notation in the form

$$\widehat{E}_M = \frac{1}{2}\mathrm{Tr}\left\{\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}\right\} - \frac{1}{2}\mathrm{Tr}\left\{\mathbf{M}(\mathbf{U}^{\mathrm{T}}\mathbf{U} - \mathbf{I}\right\} \tag{E.3}$$

where $\mathbf{M}$ is a matrix with elements μ_{ij}, $\mathbf{U}$ is a matrix whose columns consist of he eigenvectors $\mathbf{u}_i$, and $\mathbf{I}$ is the unit matrix. If we minimize (E.3) with respect o $\mathbf{U}$ we obtain

$$0 = (\boldsymbol{\Sigma} + \boldsymbol{\Sigma}^{\mathrm{T}})\mathbf{U} - \mathbf{U}(\mathbf{M} + \mathbf{M}^{\mathrm{T}}). \tag{E.4}$$

By definition, the matrix $\boldsymbol{\Sigma}$ is symmetric. Also, the matrix $\mathbf{M}$ can be taken to be symmetric without loss of generality, since the matrix $\mathbf{U}\mathbf{U}^{\mathrm{T}}$ is symmetric as is the unit matrix $\mathbf{I}$, and hence any anti-symmetric component in $\mathbf{M}$ would vanish in (E.3). Thus, we can write (E.4) in the form

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{M}. \tag{E.5}$$

Since, by construction, $\mathbf{U}$ has orthonormal columns, it is an orthogonal matrix satisfying $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$. Thus we can write (E.5) in the equivalent form

$$\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U} = \mathbf{M}. \tag{E.6}$$

Clearly one solution of this equation is to choose $\mathbf{M}$ to be diagonal so that the columns of $\mathbf{U}$ are the eigenvectors of $\boldsymbol{\Sigma}$ and the elements of $\mathbf{M}$ are its eigenvalues. However, this is not the only possible solution. Consider an arbitrary solution of (E.5). The eigenvector equation for $\mathbf{M}$ can be written

$$\mathbf{M}\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Lambda} \tag{E.7}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. Since $\mathbf{M}$ is symmetric, the eigen-vector matrix $\boldsymbol{\Psi}$ can be chosen to have orthonormal columns. Thus $\boldsymbol{\Psi}$ is an orthogonal matrix satisfying $\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \mathbf{I}$. From (E.7) we then have

$$\boldsymbol{\Lambda} = \boldsymbol{\Psi}^{\mathrm{T}}\mathbf{M}\boldsymbol{\Psi}. \tag{E.8}$$

Substituting (E.6) into (E.8) we obtain

$$\boldsymbol{\Lambda} = \boldsymbol{\Psi}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}\boldsymbol{\Psi}$$

$$= (\mathbf{U}\boldsymbol{\Psi})^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{U}\boldsymbol{\Psi})$$

$$= \widetilde{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{U}} \tag{E.9}$$

where we have defined

$$\widetilde{\mathbf{U}} = \mathbf{U}\boldsymbol{\Psi}. \tag{E.10}$$

Using $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} = \mathbf{I}$ we can write

$$\mathbf{U} = \widetilde{\mathbf{U}}\boldsymbol{\Psi}^{\mathrm{T}}. \tag{E.11}$$

Thus, an arbitrary solution to (E.6) can be obtained from the particular solution $\widetilde{\mathbf{U}}$ by application of an orthogonal transformation given by $\boldsymbol{\Psi}$. We now note that the value of the criterion E_M is invariant under this transformation since

$$E_M = \frac{1}{2}\mathrm{Tr}\left\{\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}\right\}$$

$$= \frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{\Psi}\widetilde{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{U}}\boldsymbol{\Psi}^{\mathrm{T}}\right\}$$

$$= \frac{1}{2}\mathrm{Tr}\left\{\widetilde{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{U}}\right\} \tag{E.12}$$

where we have used the fact that the trace is invariant to cyclic permutations of its argument, together with $\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \mathbf{I}$. Since all of the possible solutions give the same value for the residual error E_M, we can choose whichever is most convenient. We therefore choose the solution given by $\widetilde{\mathbf{U}}$ since, from (E.9), this has columns which are the eigenvectors of $\boldsymbol{\Sigma}$.

REFERENCES

Abu-Mostafa, Y. S. (1989). The Vapnik-Chervonenkis dimension: information versus complexity in learning. *Neural Computation* **1** (3), 312–317.

Ahmad, S. and V. Tresp (1993). Some solutions to the missing feature problem in vision. In S. J. Hanson, J. D. Cowan, and C. L. Giles (Eds.), *Advances in Neural Information Processing Systems*, Volume 5, pp. 393–400. San Mateo, CA: Morgan Kaufmann.

Aizerman, M. A., E. M. Braverman, and L. I. Rozonoer (1964). The probability problem of pattern recognition learning and the method of potential functions. *Automation and Remote Control* **25**, 1175–1190.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki (Eds.), *2nd International Symposium on Information Theory*, pp. 267–281. Tsahkadsov, Armenia, USSR.

Albertini, F. and E. D. Sontag (1993). For neural networks, function determines form. *Neural Networks* **6** (7), 975–990.

Anderson, J. A. (1982). Logistic discrimination. In P. R. Krishnaiah and L. N. Kanal (Eds.), *Classification, Pattern Recognition and Reduction of Dimensionality*, Volume 2 of *Handbook of Statistics*, pp. 169–191. Amsterdam: North Holland.

Anderson, J. A. and E. Rosenfeld (Eds.) (1988). *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley.

Arbib, M. A. (1987). *Brains, Machines, and Mathematics* (Second ed.). New York: Springer-Verlag.

Arnold, V. I. (1957). On functions of three variables. *Doklady Akademiia Nauk SSSR* **114** (4), 679–681.

Baldi, P. and K. Hornik (1989). Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks* **2** (1), 53–58.

Barnard, E. (1992). Optimization for training neural nets. *IEEE Transactions on Neural Networks* **3** (2), 232–240.

Barnard, E. and D. Casasent (1991). Invariance and neural nets. *IEEE Transactions on Neural Networks* **2** (5), 498–508.

Barron, A. R. (1984). Predicted squared error: a criterion for automatic model selection. In S. J. Farlow (Ed.), *Self-Organizing Methods in Modelling*, Vol-

Chris Bishop provides the
sive treatment of feed-
ks from the perspective of
n recognition. After intro-
concepts of pattern
describes techniques for
bility density functions,
e properties and relative
lti-layer perceptron and
ion network models. He
e use of various forms of
nd reviews the principal
ror function minimization.
d discussion of learning
n in neural networks, and
pics of data processing,
n, and prior knowledge
He concludes with an
ent of Bayesian techniques
tions to neural networks.

s
ercises
d introduction to
rn recognition
scussion of Bayesian

rofessor of Neural
on University.

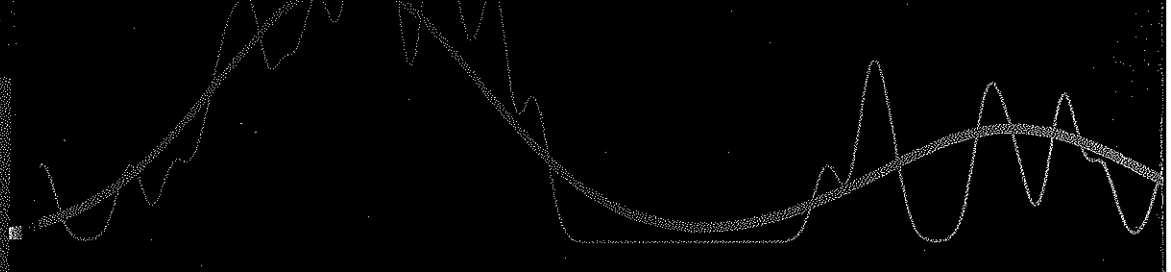
RSITY PRESS

ISBN 0-19-853864-2

Bishop

Neural Networks for Pattern Recognition

OXFORD



Neural Networks for Pattern Recognition

Christopher M. Bishop