

One Hidden Layer Linear Networks and Canonical Correlations

Zoubin Ghahramani

DRAFT — February 29, 1996

Abstract

This is a short note that relates learning one hidden layer neural networks and the statistical method known as canonical correlations analysis (CCA). It is assumed that the reader is familiar with CCA. [Abstract added: Nov 28, 2003]

Consider approximating a function from a p -dimensional input $\mathbf{x}$ to a q -dimensional target $\mathbf{y}$, using a linear network $\hat{\mathbf{y}} = VW'\mathbf{x}$, where the hidden units $\mathbf{z} = W'\mathbf{x}$ compute an m -dimensional projection of the inputs. We are interested in the class of linear networks where $m < \min(p, q)$, such that the hidden layer constitutes a bottle-neck. It is known that when $\mathbf{y} = \mathbf{x}$, the hidden unit representation spans the space of m Principal Components of $\mathbf{x}$ (eigenvectors with largest eigenvalue of the $\mathbf{x}$ covariance matrix, Σ_{xx})—we provide a generalization of this result.

Let the cost C be the expected squared error measured using a Mahalanobis metric based on the output covariance matrix, Σ_{yy} , and assume that both inputs and outputs have zero mean,

$$C = \frac{1}{2} \langle (\mathbf{y} - \hat{\mathbf{y}})' \Sigma_{yy}^{-1} (\mathbf{y} - \hat{\mathbf{y}}) \rangle \quad (1)$$

$$= \frac{1}{2} \langle (\mathbf{y} - VW'\mathbf{x})' \Sigma_{yy}^{-1} (\mathbf{y} - VW'\mathbf{x}) \rangle \quad (2)$$

$$= \frac{1}{2} \text{tr}(I_q) - \text{tr}(VW'\Sigma_{xy}\Sigma_{yy}^{-1}) + \frac{1}{2} \text{tr}(WV'\Sigma_{yy}^{-1}VW'\Sigma_{xx}) \quad (3)$$

$$= \frac{q}{2} - \text{tr}(W'\Sigma_{xy}\Sigma_{yy}^{-1}V) + \frac{1}{2} \text{tr}(V'\Sigma_{yy}^{-1}VW'\Sigma_{xx}W) \quad (4)$$

Define $U = \Sigma_{yy}^{-1/2}V$ and $S = \Sigma_{xx}^{1/2}W$. Then,

$$C = \frac{q}{2} - \text{tr}(S'\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}U) + \frac{1}{2} \text{tr}(U'US'S) \quad (5)$$

Scalar Case

For the $m = 1$ case, the arguments of the traces in (5) are all scalar. We examine separately the magnitudes and directions of the vectors U and S . Denote the magnitude of U by γ_u and the unit vector in the direction of U by $\mathbf{e}_u$, similarly for S . Then,

$$C = \frac{q}{2} - \gamma_u \gamma_s \mathbf{e}_s' \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \mathbf{e}_u + \frac{1}{2} \gamma_u^2 \gamma_s^2. \quad (6)$$

For fixed magnitudes, the minimum of C can be found by performing a Singular Value Decomposition (SVD) on $R = \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$. The SVD of R gives $R = Q_1\Lambda Q_2'$ where Q_1 is a $p \times q$ orthogonal matrix, Λ is a $q \times q$ diagonal matrix of eigenvalues, and Q_2 is a $q \times q$ orthogonal matrix. We therefore get that C is minimized when $\mathbf{e}_s$ is the vector in Q_1 associated with the largest eigenvalue of Λ , $\lambda_{\max}$, and $\mathbf{e}_u$ is the corresponding vector in Q_2 associated with $\lambda_{\max}$.

Having found the optimal directions, we have

$$C = \frac{q}{2} - \lambda_{\max}\gamma_u\gamma_s + \frac{1}{2}\gamma_u^2\gamma_s^2. \quad (7)$$

The magnitudes that minimize C are given by the constraint $\gamma_u\gamma_s = \lambda_{\max}$. Clearly, we cannot determine the individual magnitudes, as the cost is invariant to inversely proportional scalings of U and S .

Recall that in Canonical Correlation analysis, the first canonical direction for $\mathbf{x}$, $\mathbf{a}_x$ is given by

$$\mathbf{a}_x = \Sigma_{xx}^{-1/2}\mathbf{e}_x, \quad (8)$$

where $\mathbf{e}_x$ is the eigenvector with maximal eigenvalue obtained from the SVD of $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$. The vector W is therefore proportional to the first canonical direction for $\mathbf{x}$. The vector V is proportional to Σ_{yy} times the first canonical direction for $\mathbf{y}$.