Neural Networks, Vol. 2, pp. 53–58, 1989
Printed in the USA. All rights reserved.

0893-6080/89 $3.00 + .00

ORIGINAL CONTRIBUTION

Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima

PIERRE BALDI AND KURT HORNIK*

University of California, San Diego

(*Received* 18 *May* 1988; *revised and accepted* 16 *August* 1988)

Abstract—*We consider the problem of learning from examples in layered linear feed-forward neural networks using optimization methods, such as back propagation, with respect to the usual quadratic error function* E *of the connection weights. Our main result is a complete description of the landscape attached to* E *in terms of principal component analysis. We show that* E *has a unique minimum corresponding to the projection onto the subspace generated by the first principal vectors of a covariance matrix associated with the training patterns. All the additional critical points of* E *are saddle points (corresponding to projections onto subspaces generated by higher order vectors). The auto-associative case is examined in detail. Extensions and implications for the learning algorithms are discussed.*

Keywords—Neural networks, Principal component analysis, Learning, Back propagation.

1. INTRODUCTION

Neural networks can be viewed as circuits of highly interconnected units with modifiable interconnection weights. They can be classified, for instance, according to their architecture, algorithm for adjusting the weights, and the type of units used in the circuit. We shall assume that the reader is familiar with the basic concepts of the field; general reviews, complements, and references can be found in Rumelhart, McClelland, and the PDP Research Group (1986a), Lippman (1987), and Grossberg (1988).

The network architecture considered here is of the type often described in Rumelhart Hinton, and Williams (1986b), namely layered feed-forward networks with one layer of input units, one layer of output units, and one or several layers of hidden units. We assume that there are T input patterns x_t $(1 \leq t \leq T)$ and T corresponding target output patterns y_t which are used to train the network. For this purpose, a quadratic error function is defined as usual to be: $E = \Sigma_t \|y_t - F(x_t)\|^2$ where F is the current function implemented by the network. During the training phase, the weights (and hence F) are successively modified, according to one of several possible algorithms, in order to reduce E. Back propagation, the best known of such algorithms, is just a way of implementing a gradient descent method for E. The main thrust of this paper is not the study of a specific algorithm but rather a precise description of the salient features of the surface attached to E when the units are linear.

Linear units are the simplest one can use in these circuits. They are often considered as uninteresting for: (a) only linear functions can be computed in linear networks (and most "interesting" functions are nonlinear); and (b) a network with several layers of linear units can always be collapsed into a linear network without any hidden layer by multiplying the weights in the proper fashion.

As a result, nonlinear units are most commonly used: linear threshold gates or, when continuity or differentiability is required, units with a sigmoid input–output function. In this setting, the results of numerous simulations have led several people to believe that descent methods, such as back propagation, applied to the error function E are not seriously plagued by the problem of local minima (either because global minima are found, either because the local minima encountered are "good enough" for practical purposes) and that, for instance, the solu-

*Permanent address: Institut für Statistik and Wahrscheinlichkeitstheorie, Technische Universität Wien, Wiedner Haupstr. 8-10/107, A-1040 Wien, Austria.

The final stages of this work were supported by NSF grant DMS-8800323 to P. B.

Requests for reprints should be sent to Pierre Baldi, JPL 198-330, California Institute of Technology, Pasadena, CA 91109.

tions obtained have remarkable generalization properties. The complete absence, to this date, of any analytical result supporting these claims would alone by itself justify a careful investigation of the simpler linear case.

In addition, recent work of Linsker (1986a, 1986b, 1986c) and Cottrell, Munro, and Zipser (in press) seems to indicate that, for some tasks, linear units can still be of interest, not as much for the global map they implement but for the internal representation of the input data and the transformations that occur in the different layers during the learning period.

Linsker, for instance, has shown that in a layered feed-forward network of linear units with random inputs and a Hebb type of algorithm for adjusting the synaptic weights, spatial opponent and orientation selective units spontaneously emerge in successive hidden layers, in a way which does not contradict what is observed in the early visual system of higher animals. Cottrell *et al.* (in press) have used linear units together with the technique of auto-association to realize image compression. Auto-association, which is also called auto-encoding or identity mapping (see Ackley, Hinton, & Sejnowski; 1985; Ellman & Zipser, 1988) is a simple trick intended to avoid the need for having a teacher, that is, for knowing the target values y_t, by setting $x_t = y_t$. In this mode, the network will tend to learn the identity map which in itself is not too exciting. However, if this is done using one narrow layer of hidden units, one expects the network to find efficient ways of compressing the information contained in the input patterns. An analysis of linear auto-association has been provided by Bourlard and Kamp (1988) based on singular value decomposition of matrices. However, their results for the linear case, which are comprised by ours, do not give a description of the landscape of E.

Our notation will be as follows. All vectors are column vectors and prime superscripts denote transposition. To begin with, we shall assume that both x_t and y_t are n-dimensional vectors and that the network consists of one input layer with n inputs, one hidden layer with $p (p \leq n)$ units, and one output layer with n units (see Figure 1). The weights connecting the inputs to the hidden layer are described by a $p \times n$ real matrix B and those from the hidden layer to the output by an $n \times p$ real matrix A. With these assumptions, the error function can be written:

$$E(A, B) = \sum_{1 \leq t \leq T} \|y_t - ABx_t\|^2. \tag{1}$$

We define the usual sample covariance matrices $\Sigma_{XX} = \Sigma_t x_t x_t'$, $\Sigma_{XY} = \Sigma_t x_t y_t'$, $\Sigma_{YY} = \Sigma_t y_t y_t'$, and $\Sigma_{YX} = \Sigma_t y_t s_t'$. We consider the problem of finding the matrices A and B so as to minimize E. In Section 2, we use spectral analysis to describe the properties of the landscape attached to E in the general situation. The auto-associative case and its relations to principal component analysis follow immediately as a special case. In Section 3, we briefly examine some consequences for the optimization algorithms. All mathematical proofs are deferred to the Appendix.

It is important to notice from the onset that if C is any $p \times p$ invertible matrix, then $AB = ACC^{-1}B = (AC)(C^{-1}B)$. Therefore the matrices A and B are never unique since they can always be multiplied by appropriate invertible matrices. Whenever uniqueness occurs it is in terms of the global map $W = AB$ (equivalently, one could partition the matrices into equivalence classes). Notice also that W has rank at most p and recall that if Σ_{XX} is invertible the solution to the problem of minimizing $E(L) = \Sigma_t \|y_t - Lx_t\|^2$, where L is an $n \times n$ matrix without any rank restrictions, is unique and given by $L = \Sigma_{YX}\Sigma_{XX}^{-1}$ which is the usual slope matrix for the ordinary least squares regression of Y on X. Finally, if M is an $n \times p (p \leq n)$ matrix we shall denote by P_M the matrix of the orthogonal projection onto the subspace spanned by the columns of M. It is well known that $P_M^2 = P_M$ and $P_M' = P_M$. If in addition M is of full rank p, then $P_M = M(M'M)^{-1}M'$.

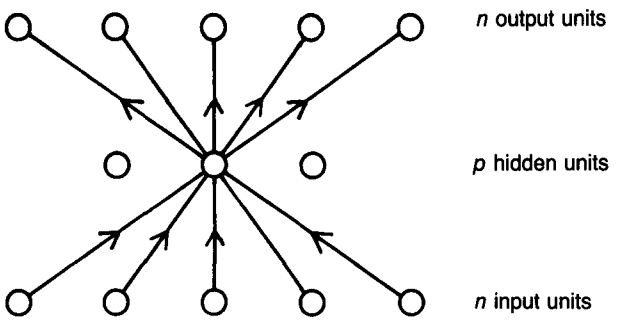

FIGURE 1. The network.

2. MAIN RESULTS: THE LANDSCAPE OF E

Our main result is that:
E has, up to equivalence, a unique local and global minimum corresponding to an orthogonal projection onto the subspace spanned by the first principal eigenvectors of a covariance matrix associated with the training patterns. All other critical points of E are saddle points.

More precisely, one has the following four facts.

Fact 1: For any fixed $n \times p$ matrix A the function $E(A, B)$ is convex in the coefficients of B and attains its minimum for any B satisfying the equation

$$A'AB\Sigma_{XX} = A'\Sigma_{YX}. \tag{1}$$

If Σ_{XX} is invertible and A is full rank p, then E is strictly convex and has a unique minimum reached when

$$B = \hat{B}(A) = (A'A)^{-1}A'\Sigma_{YX}\Sigma_{XX}^{-1}. \tag{3}$$

In the auto-associative case, (3) becomes

$$B = \hat{B}(A) = (A'A)^{-1}A'. \tag{3'}$$

Fact 2: For any fixed $p \times n$ matrix B the function $E(A, B)$ is convex in the coefficients of A and attains its minimum for any A satisfying the equation

$$AB\Sigma_{XX}B' = \Sigma_{YX}B'. \tag{4}$$

If Σ_{XX} is invertible and B is of full rank p, then E is strictly convex and has a unique minimum reached when

$$A = \hat{A}(B) = \Sigma_{YX}B'(B\Sigma_{XX}B')^{-1}. \tag{5}$$

In the auto-associative case, (5) becomes

$$A = \hat{A}(B) = \Sigma_{XX}B'(B\Sigma_{XX}B')^{-1}. \tag{5'}$$

Fact 3: Assume that Σ_{XX} is invertible. If two matrices A and B define a critical point of E (i.e., a point where $\partial E/\partial a_{ij} = \partial E/\partial b_{ij} = 0$) then the global map $W = AB$ is of the form

$$W = P_A\Sigma_{YX}\Sigma_{XX}^{-1} \tag{6}$$

with A satisfying

$$P_A\Sigma = P_A\Sigma P_A = \Sigma P_A \tag{7}$$

where $\Sigma = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$. In the auto-associative case, $\Sigma = \Sigma_{XX}$ and (6) and (7) become

$$W = AB = P_A \tag{6'}$$

$$P_A\Sigma_{XX} = P_A\Sigma_{XX}P_A = \Sigma_{XX}P_A. \tag{7'}$$

If A is of full rank p, then A and B define a critical point of E if and only if A satisfies (7) and $B = \hat{B}(A)$, or equivalently if and only if A and W satisfy (6) and (7).

Notice that in (4), the matrix $\Sigma_{YX}\Sigma_{XX}^{-1}$ is the slope matrix for the ordinary least squares regression of Y on X. It is easily seen that Σ is the sample covariance matrix of the best unconstrained linear approximation $\hat{y}_t = \Sigma_{YX}\Sigma_{XX}^{-1}x_t$ of Y based on X.

Fact 4: Assume that Σ is full rank with n distinct eigenvalues $\lambda_1 > \ldots > \lambda_n$. If $\mathcal{J} = \{i_1, \ldots, i_p\}$ $(1 \leq i_1 < \ldots < i_p \leq n)$ is any ordered p-index set, let $U_{\mathcal{J}} = [u_{i_1}, \ldots, u_{i_p}]$ denote the matrix formed by the orthonormal eigenvectors of Σ associated with the eigenvalues $\lambda_{i_1}, \ldots, \lambda_{i_p}$. Then two full rank matrices A and B define a critical point of E if and only if there exist an ordered p-index set $\mathcal{J}$ and an invertible $p \times p$ C matrix such that

$$A = U_{\mathcal{J}}C \tag{8}$$

$$B = C^{-1}U'_{\mathcal{J}}\Sigma_{YX}\Sigma_{XX}^{-1}. \tag{9}$$

For such a critical point we have

$$W = P_{U_{\mathcal{J}}}\Sigma_{YX}\Sigma_{XX}^{-1} \tag{10}$$

$$E(A, B) = tr\Sigma_{YY} - \sum_{i \in \mathcal{J}} \lambda_i. \tag{11}$$

Therefore a critical W of rank p is always the product of the ordinary least squares regression matrix followed by an orthogonal projection onto the subspace spanned by p eigenvectors of Σ. The critical map W associated with the index set $\{1, 2, \ldots, p\}$ is the unique local and global minimum of E. The remaining $\binom{n}{p} - 1$ p-index sets correspond to saddle points. All additional critical points defined by matrices A and B which are not of full rank are also saddle points and can be characterized in terms of orthogonal projections onto subspaces spanned by q eigenvectors, with $q < p$ (see Figure 2). In the auto-associative case, (8) (9) and (10) become

$$A = U_{\mathcal{J}}C \tag{8'}$$

$$B = C^{-1}U'_{\mathcal{J}} \tag{9'}$$

$$W = P_{U_{\mathcal{J}}} \tag{10'}$$

and therefore the unique locally and globally optimal map W is the orthogonal projection onto the space spanned by the first p eigenvectors of Σ_{XX}.

Remark: At the global minimum, if C is the identity I_p then the activities of the units in the hidden layer are given by $u_1'\hat{y}_t, \ldots, u_p'\hat{y}_t$ the so-called *principal components* of the $\hat{y}_t$'s (see for instance Kshirsagar, 1972). In the auto-associative case, these activities are given by $u_1'x_t, \ldots, u_p'x_t$. They are the coordinates of the vector x_t along the first p eigenvectors of Σ_{XX}.

The assumptions on the rank or eigenvalues of the matrices appearing in the statements of the facts are by no means restrictive. They are satisfied in most practical situations and also in the case of random matrices with probability one. For instance a non-invertible Σ_{XX} corresponds to a poor choice of the training patterns with linear dependencies and a rank deficient matrix A (or B) to a very poor utilization of the units in the network. For back propagation, the initial weights are usually set at random which yields, with probability one, matrices A and B of full rank. Σ is a covariance matrix and therefore its eigenvalues are always non-negative. To assume that they are all strictly positive is equivalent to assuming

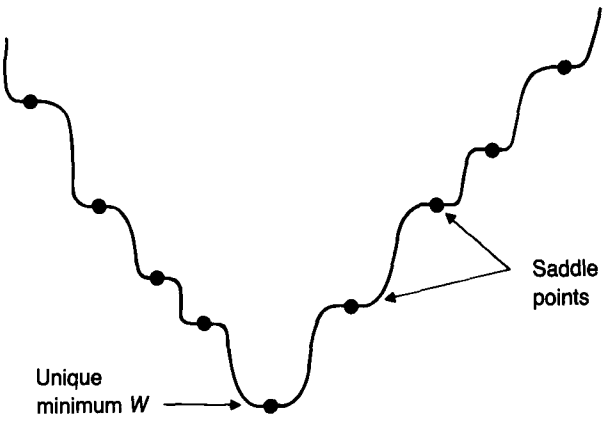

FIGURE 2. The landscape of *E*.

that both Σ_{XX} and Σ_{YX} are of full rank. Full rank matrices are dense and in a realistic environment with noise and finite precision, we can always slightly perturb the conditions so as to make Σ invertible and with distinct eigenvalues. Furthermore, in the proofs in Appendix B, we describe the structure of the critical points with deficient rank and what happens in the case where some of the eigenvalues of Σ are equal.

We have also restricted our analysis to the case of linear units without bias and to networks containing a single hidden layer. The generalization of our result to the affine case is straightforward either by presubtracting the mean from the input and target data, or by adding a unit which is kept at a fixed value. A rigorous extension to the nonlinear sigmoid case or the case involving linear threshold units seems more difficult. However, our results, and in particular the main features of the landscape of E, hold true in the case of linear networks with several hidden layers.

One of the central issues in learning from examples is the problem of generalization, that is, how does the network perform when exposed to a pattern never seen previously? In our setting, a precise quantitative answer can be given to this question. For instance, in the auto-associative case, the distortion on a new pattern is exactly given by its distance to the subspace generated by the first p eigenvectors of Σ_{XX}.

It is reasonable to think that for most solutions found by running a gradient descent algorithm on the function E, the final matrix C will not be the identity I_p. In fact, we even expect C to be rather "random" looking. This is the main reason why the relation of auto-association to principal component analysis was not apparent in earlier simulations described in the literature and why, in the solutions found by back propagation, the work load seems to be evenly distributed among the units of the hidden layer. If in (9′) we take $C = I_p$, then $B = U_p'$. Therefore the synaptic vector corresponding to the "first" hidden unit is exactly equal to the dominant eigenvector of the input correlation matrix. This is in fact exactly the same result as the one obtained by Oja (1982) in a different setting, using differential equations to approximate a constrained form of Hebbian learning on a single linear unit with n stochastic inputs. In other words, up to equivalence, the solution sought by a back propagation type of algorithm in the auto-associative case and by Hebbian learning are identical on one single linear "neuron." It remains to be checked whether simultaneous Hebbian learning on p units, probably with some appropriate form of laterial inhibition, leads to the same results as those encountered here for the auto-association.

3. CONCLUDING REMARKS ON THE ALGORITHMS

One of the nice features of the landscape of E is the existence, up to equivalence, of a unique local and global minimum which, in addition, can be described in terms of principal component analysis and least squares regression. Consequently, this optimum could also be obtained from several well-known algorithms for computing the eigenvalues and eigenvectors of symmetric positive definite matrices (see for instance Atkinson, 1978). By numerical analysis standards, these algorithms are superior to gradient methods for the class of problems considered here. However, though efficiency considerations are of importance, one should not disregard back propagation on this sole basis, for its introduction in the design of neural networks was guided by several other considerations. In particular, in addition to its simplicity, error back-propagation can be applied to nonlinear networks and to a variety of problems without having any detailed a priori knowledge of their structure or of the mathematical properties of the optimal solutions.

A second nice feature of the landscape of E is that if we fix A (resp. B) with full rank, then E is a strictly convex quadratic form and there exists a unique minimum reached for $B = \hat{B}(A)$ (resp. $A = \hat{A}(B)$). In this case, gradient descent with appropriate step width (or "learning rate") leads to a convergence with a residual error decaying exponentially fast. Of course, $\hat{B}(A)$ (resp. $\hat{A}(B)$) can also be obtained directly by solving the linear system in (2). This also suggests another optimization strategy which consists of successively computing, starting for instance from a random A, $\hat{B}(A)$, $\hat{A}(\hat{B}(A))$, . . . and so forth, which in fact is a Newton's type of method. In any case, from a theoretical standpoint, one should notice that, although E has no local minima, both gradient descent and Newton's type of methods could get stuck in a saddle point. However, as exemplified by simulations (Cottrell *et al.*, in press), this seems unlikely to happen, especially with the way error back-propagaton is usually implemented, with a descent direction computed by differentiating E after presentation of one or just a few training patterns. Such a direction is clearly distinct from a true gradient.

REFERENCES

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**, 147–169.

Atkinson, K. E. (1978). *An introduction to numerical analysis*. New York: John Wiley & Sons.

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294.

Cottrell, G. W., Munro, P. W., & Zipser, D. (in press). Image

compression by back propagation: A demonstration of extensional programming. In N. E. Sharkey (Ed.), *Advances in cognitive science* (Vol. 2). Norwood, NJ: Abbex.

Ellman, J. L., & Zipser, D. (1987). *Learning the hidden structure of speech*. (Tech. Rep. No. 8701). San Diego: Institute for Cognitive Science, University of California.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms and architectures. *Neural Networks*, **1**, 17–61.

Kshirsagar, A. N. (1972). *Multivariate analysis*. New York: Marcel Dekker, Inc.

Linsker, R. (1986a). From basic network principles to neural architecture: Emergence of spatial opponent cells. *Proceedings of the National Academy of Sciences USA*, **83**, 7508–7512.

Linsker, R. (1986b). From basic network principles to neural architecture: Emergence of orientation selective cells. *Proceedings of the National Academy of Sciences USA*, **83**, 8390–8394.

Linsker, R. (1986c). From basic network principle to neural architecture: Emergence of orientation columns. *Proceedings of the National Academy of Sciences USA*, **83**, 8779–8783.

Lippman, R. P. (1987). An introduction to computing with neural nets. *IEEE Transactions on Acoustics, Speech, and Signal Processing Magazine*, 4–22.

Magnus, J. R., & Neudecker, H. (1986). Symmetry, 0-1 matrices and Jacobians. *Econometric Theory*, **2**, 157–190.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, **15**, 267–273.

Pollock, D. S. G. (1979). The algebra of econometrics. New York: John Wiley & Sons.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986a). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vols. 1 & 2). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b). Learning internal representation by error propagaton. In Parallel distributed processing. *Explorations in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press.

APPENDIX A: MATHEMATICAL PROOFS

We have tried to write proofs which are self-contained up to very basic results of linear algebra. Slightly less elementary results which are often used in the proofs (sometimes without explicit mentioning) are listed below as a reminder for the reader. For any matrices P, Q, R we have $tr(PQR) = tr(RPQ) = tr(QRP)$, provided that these quantities are defined. Thus in particular if P is idempotent, that is, $P^2 = P$, then

$$tr(PQP) = tr(P^2Q) = tr(PQ). \quad \text{(a)}$$

If U is orthogonal, that is $U'U = I$, then

$$tr(UQU') = tr(U'UQ) = tr(Q). \quad \text{(b)}$$

The Kronecker product $P \otimes Q$ of any two matrices P and Q is the matrix obtained from the matrix P by replacing each entry p_{ij} of P with the matrix $p_{ij}Q$. If P is any $m \times n$ matrix and p_j its jth column, then vec P is the $mn \times 1$ vector vec $P = [p_1', \ldots, p_n']'$. Thus the vec operation transforms a matrix into a column vector by stacking the columns of the matrix one underneath the other. We then have (see for instance Magnus & Neudecker, 1986) for any matrices P, Q, R

$$tr(PQ') = (\text{vec } P)' \text{ vec } Q \quad \text{(c)}$$

$$\text{vec}(PQR') = (R \otimes P) \text{ vec } Q \quad \text{(d)}$$

$$(P \otimes Q)(R \otimes S) = PR \otimes QS \quad \text{(e)}$$

$$(P \otimes Q)^{-1} = P^{-1} \otimes Q^{-1} \quad \text{(f)}$$

$$(P \otimes Q)' = P' \otimes Q' \quad \text{(g)}$$

whenever these quantities are defined. Also:
if P and Q are symmetric and positive semidefinite (resp. positive definite) then $P \otimes Q$ is symmetric and positive semidefinite (resp. positive definite). (h)

Finally, let us introduce the input data matrix $X = [x_1, \ldots, x_T]$ and the output data matrix $Y = [y_1, \ldots, y_T]$. It is easily seen that $XX' = \Sigma_{XX}$, $XY' = \Sigma_{XY}$, $YY' = \Sigma_{YY}$, $YX' = \Sigma_{YX}$ and $E(A, B) = \|\text{vec}(Y - ABX)\|^2$. In the proofs of facts 1 and 2, we shall use the following well known lemma.

Lemma: The quadratic function

$$F(z) = \|c - Mz\|^2 = c'c - 2c'Mz + z'M'Mz$$

is convex. A point z corresponds to a global minimum of F if and only if it satisfies the equation $\nabla F = 0$, or equivalently $M'Mz = M'c$. If in addition $M'M$ is positive definite, then F is strictly convex and the unique minimum of F is attained for $z = (M'M)^{-1}M'c$.

Proof of fact 1: For fixed A, use (d) to write $\text{vec}(Y - ABX) = \text{vec } Y - \text{vec}(ABX) = \text{vec } Y - (X' \otimes A) \text{ vec } B$ and thus $E(A, B) = \|\text{vec } Y - (X' \otimes A) \text{ vec } B\|^2$. By the above lemma, E is convex in the coefficients of B and B corresponds to a global minimum if and only if $(X' \otimes A)'(X' \otimes A) \text{ vec } B = (X' \otimes A)' \text{ vec } Y$. Now on one hand $(X' \otimes A)'(X' \otimes A) \text{ vec } B = (X' \otimes A) \text{ vec } B = (XX' \otimes A'A) \text{ vec } B = (\Sigma_{XX} \otimes A'A) \text{ vec } B = \text{vec}(A'AB\Sigma_{XX})$. On the other hand $(X' \otimes A)' \text{ vec } Y = (X \otimes A') \text{ vec } Y = \text{vec}(A'YX') = \text{vec}(A'\Sigma_{YX})$. Therefore

$$A'AB\Sigma_{XX} = A'\Sigma_{YX},$$

which is (2). If A is full rank, $A'A$ is symmetric and positive definite. As a covariance matrix, Σ_{XX} is symmetric and positive semidefinite; if, in addition, Σ_{XX} is invertible, then Σ_{XX} is also positive definite. Because of (h), $(X' \otimes A)'(X' \otimes A) = \Sigma_{XX} \otimes A'A$ is also symmetric and positive definite. Applying the above lemma, we conclude that if Σ_{XX} is invertible and A is a fixed full rank matrix, then E is strictly convex in the coefficients of B and attains its unique minimum at the unique solution $B = \hat{B}(A) = (A'A)^{-1}A'\Sigma_{YX}\Sigma_{XX}^{-1}$ of (2), which is (3). In the auto-associative case, $x_t = y_t$. Therefore $\Sigma_{XX} = \Sigma_{YY} = \Sigma_{XY} = \Sigma_{YX}$ and the above expression simplifies to (3').

Proof of Fact 2: For fixed B, use (d) to write $\text{vec}(Y - ABX) = \text{vec } Y - \text{vec}(ABX) = \text{vec } Y - (X'B' \otimes I) \text{ vec } A$ and so $E(A, B) = \|\text{vec } Y - (X'B' \otimes I) \text{ vec } A\|^2$. By the above lemma, E is convex in the coefficients of A and A corresponds to a global minimum if and only if $(X'B' \otimes I)'(X'B' \otimes I) \text{ vec } A = (X'B' \otimes I)' \text{ vec } Y$. Since $(X'B' \otimes I)'(X'B' \otimes I) \text{ vec } A = (BXX'B \text{ pr} \otimes I) \text{ vec } A = (B\Sigma_{XX}B' \otimes I) \text{ vec } A = \text{vec}(AB\Sigma_{XX}B')$ and $(X'B' \otimes I)' \text{ vec } Y = (BX \otimes I) \text{ vec } Y = \text{vec}(YX'B') = \text{vec}(\Sigma_{YX}B')$ we have

$$AB\Sigma_{XX}B' = \Sigma_{YX}B',$$

which is (4). If B and Σ_{XX} are full rank, then the symmetric and positive semi-definite matrix $B\Sigma_{XX}B'$ becomes full rank and therefore positive definite. Because of (h), $(X'B' \otimes I)'(X'B' \otimes I) = (B\Sigma_{XX}B' \otimes I)$ is also positive definite and (5) and (5') are easily derived as in the end of the proof of fact 1.

Notice that from facts 1 and 2, two full rank matrices A and B define a critical point for E if and only if (2) and (4) are simultaneously satisfied. In all cases of practical interest where Σ_{YX} is full rank both $\hat{A}(B)$ and $\hat{B}(A)$ are full rank. In what follows, we shall always assume that A is of full rank p. The case *rank* $(A) < p$ is, although intuitively of no practical interest, slightly more technical and its treatment will be postponed to Appendix B.

Proof of Fact 3: Assume first that A and B define a critical point of E, with A full rank. Then from fact 1 we get $B = \hat{B}(A)$ and thus

$$W = AB = A(A'A)^{-1}A'\Sigma_{YX}\Sigma_{XX}^{-1} = P_A\Sigma_{YX}\Sigma_{XX}^{-1}$$

which is (6). Multiplication of (4) by A' on the right yields

$$W\Sigma_{XX}W' = AB\Sigma_{XX}B'A' = \Sigma_{YX}B'A' = \Sigma_{YX}W'$$

or

$$P_A\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY}P_A = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}P_A$$

or equivalently $P_A\Sigma P_A = \Sigma P_A$. Since both Σ and P_A are symmetric, $P_A\Sigma P_A = \Sigma P_A$ is also symmetric and therefore $\Sigma P_A = (\Sigma P_A)' = P_A'\Sigma' = P_A\Sigma$. So $P_A\Sigma = P_A\Sigma P_A = \Sigma P_A$, which is (7). Hence if A and B correspond to a critical point and A is full rank then (6) and (7) must hold and $B = \hat{B}(A)$.

Conversely, assume that A and W satisfy (6) and (7), with A full rank. Multiplying (6) by $(A'A)^{-1}A'$ on the left yields $B = (A'A)^{-1}A'\Sigma_{YX}\Sigma_{XX}^{-1} = \hat{B}(A)$ and (2) is satisfied. From $P_A\Sigma P_A = \Sigma P_A$ and using (6) we immediately get $AB\Sigma_{XX}B'A' = \Sigma_{YX}B'A'$ and multiplication of both sides by $A(A'A)^{-1}$ on the right yields $AB\Sigma_{XX}B' = \Sigma_{YX}B'$, which is (4). Thus A and B satisfy (2) and (4) and therefore they define a critical point of E.

Proof of Fact 4: First notice that since Σ is a real symmetric covariance matrix, it can always be written as $\Sigma = U\Lambda U'$ where U is an orthogonal column matrix of eigenvectors of Σ and Λ is the diagonal matrix with non-increasing eigenvalues on its diagonal. Also if Σ is full rank, then Σ_{XX}, Σ_{YX} and Σ_{XY} are full rank too.

Now clearly if A and B satisfy (8) and (9) for some C and some $\mathcal{I}$ then A and B are full rank p and satisfy (3) and (5). Therefore they define a critical point of E.

For the converse, we have

$$P_{U'A} = U'A(A'UU'A)^{-1}A'U = U'A(A'A)^{-1}A'U = U'P_AU$$

or, equivalently, $P_A = UP_{U'A}U'$. Hence (7) yields

$$UP_{U'A}U'U\Lambda U' = P_A\Sigma = \Sigma P_A = U\Lambda U'UP_{U'A}U'$$

and so $P_{U'A}\Lambda = \Lambda P_{U'A}$. Since $\lambda_1 > \ldots > \lambda_n > 0$, it is readily seen that $P_{U'A}$ is diagonal. $P_{U'A}$ is an orthogonal projector of rank p and its eigenvalues are 1 (p times) and 0 ($n - p$ times). Therefore there exists a unique index set $\mathcal{I} = \{i_1, \ldots, i_p\}$ with $1 \le i_1 < \ldots < i_p \le n$ such that $P_{U'A} = I_{\mathcal{I}}$ where $I_{\mathcal{I}}$ is the diagonal matrix with entry $i = 1$ if $i \in \mathcal{I}$ and 0 otherwise. It follows that

$$P_A = UP_{U'A}U' = UI_{\mathcal{I}}U' = U_{\mathcal{I}}U_{\mathcal{I}}',$$

where $U_{\mathcal{I}} = [u_{i_1}, \ldots, u_{i_p}]$. Thus P_A is the orthogonal projection onto the subspace spanned by the columns of $U_{\mathcal{I}}$. Since the column space of A coincides with the column space of $U_{\mathcal{I}}$, there exists an invertible $p \times p$ matrix C such that $A = U_{\mathcal{I}}C$. Moreover, $B = \hat{B}(A) = C^{-1}U_{\mathcal{I}}'\Sigma_{YX}\Sigma_{XX}^{-1}$ and (8) and (9) are satisfied. There are $\binom{n}{p}$ possible choices for $\mathcal{I}$ and therefore, up to equivalence, $\binom{n}{p}$ critical points with full rank.

From (8) and (9), (10) results immediately.

Remark: In the most general case with n-dimensional inputs x_t and m-dimensional outputs y_t, Σ has $r(r \le m)$ distinct eigenvalues $\lambda_1 \ge \ldots \ge \lambda_r \ge 0$ with multiplicities $m_1, \ldots, m_r$. Using the above arguments, it is easily seen that $P_{U'A}$ will now be block-diagonal $[P_1, \ldots, P_r]$ where $P_1, \ldots, P_r$ are orthogonal projectors of dimension $m_1, \ldots, m_r$ and thus A is of the form $A = (UV)_{\mathcal{I}}C$ where V is block-diagonal $[V_1, \ldots, V_r]$, $V_1, \ldots, V_r$ being orthogonal matrices of dimension $m_1, \ldots, m_r$. For all such choices of V, UV is a matrix of normalized eigenvectors of Σ corresponding to ordered eigenvalues of Σ. The geometric situation, as expected, does not really change but the parameterization becomes more involved as U is no longer unique.

To prove (11), use (c) to write $E(A, B) = (\mathrm{vec}(Y - ABX))'\,\mathrm{vec}(Y - ABX) = (\mathrm{vec}\,Y)'\,\mathrm{vec}\,Y - 2(\mathrm{vec}(ABX))'\,\mathrm{vec}\,Y + (\mathrm{vec}\,ABX)'\,\mathrm{vec}\,ABX = tr YY' - 2trABXY' + trABXX'B'A' = tr\Sigma_{YY} - 2trW\Sigma_{XY} + trW\Sigma_{XX}W'$. If A is full rank and $B = \hat{B}(A)$, then $W = A\hat{B}(A) = P_A\Sigma_{YX}\Sigma_{XX}^{-1}$ and therefore $tr(W\Sigma_{XX}W') = tr(P_A\Sigma P_A) = tr(P_A\Sigma) = tr(UP_{U'A}U'U\Lambda U') = tr(P_{U'A}U'U\Lambda) = tr(P_{U'A}\Lambda)$ and $tr(W\Sigma_{YX}) = tr(P_A\Sigma) = tr(P_{U'A}\Lambda)$. So for an arbitrary A of rank p,

$$E(A, \hat{B}(A)) = tr\Sigma_{YY} - trP_{U'A}\Lambda.$$

If A is of the form $U_{\mathcal{I}}C$, then $P_{U'A} = I_{\mathcal{I}}$. Therefore

$$E(A, \hat{B}(A)) = tr\Sigma_{YY} - trI_{\mathcal{I}}\Lambda = tr\Sigma_{YY} - \sum_{i \in \mathcal{I}} \lambda_i$$

which is (11).

We shall now establish that whenever A and B satisfy (8) and (9) with $\mathcal{I} \neq \{1, 2, \ldots, p\}$ there exist matrices $\tilde{A}$, $\tilde{B}$ arbitrarily close to A, B such that $E(\tilde{A}, \tilde{B}) < E(A, B)$. For this purpose it is enough to slightly perturb the column space of A in the direction of an eigenvector associated with one of the first p eigenvalues of Σ which is not contained in $\{\lambda_i, i \in \mathcal{I}\}$. More precisely, fix two indices j and k with $j \in \mathcal{I}$, $k \notin \mathcal{I}$. For any ϵ, put $\tilde{u}_j = (1 + \epsilon^2)^{-1/2}(u_j + \epsilon u_k)$ and construct $\tilde{U}_{\mathcal{I}}$ from $U_{\mathcal{I}}$ by replacing u_j with $\tilde{u}_j$. Since $k \notin \mathcal{I}$, we still have $\tilde{U}_{\mathcal{I}}'\tilde{U}_{\mathcal{I}} = I_p$. Now let $\tilde{A} = \tilde{U}_{\mathcal{I}}C$ and $\tilde{B} = \hat{B}(\tilde{A}) = C^{-1}\tilde{U}_{\mathcal{I}}'\Sigma_{YX}\Sigma_{XX}^{-1}$. A simple calculation shows that the diagonal elements $\tilde{\delta}_i$ of $P_{U'\tilde{A}}$ are

$$\tilde{\delta}_i = \begin{cases} 0 & \text{if } i \notin \mathcal{I} \cup \{k\} \\ 1 & \text{if } i \in \mathcal{I} \text{ and } i \neq j \text{ and } i \neq k \\ 1/(1 + \epsilon^2) & \text{if } i = j \\ \epsilon^2/(1 + \epsilon^2) & \text{if } i = k. \end{cases}$$

Therefore $E(\tilde{A}, \tilde{B}) = tr\Sigma_{YY} - trP_{U'\tilde{A}}\Lambda = tr\Sigma_{YY} - [\sum_{i \in \mathcal{I} - \{j\}}\lambda_i + \lambda_j/(1 + \epsilon^2) + \epsilon^2\lambda_k/(1 + \epsilon^2)] = tr\Sigma_{YY} - \sum_{i \in \mathcal{I}}\lambda_i - \epsilon^2(\lambda_k - \lambda_j)/(1 + \epsilon^2) = E(A, B) - \epsilon^2(\lambda_k - \lambda_j)/(1 + \epsilon^2)$. By taking values of ϵ arbitrarily small, we see that any neighborhood of A, B contains points of the form $\tilde{A}$, $\tilde{B}$ with a strictly smaller error function. Thus if $\mathcal{I} \neq \{1, 2, \ldots, p\}$, then (8) and (9) define a saddle point and not a local minimum. Notice that, in any case, it could not be a local maximum because of the strict convexity of E, with fixed full rank A, in fact 1.

APPENDIX B: THE RANK DEFICIENT CASE

We now complete the proof of fact 3 (equations (6) and (7)) and fact 4, in the case where A is not of full rank. Using the Moore-Penrose inverse A^+ of the matrix A (see for instance Pollock, 1979), the general solution to equation (2) can be written as

$$B = A^+\Sigma_{YX}\Sigma_{XX}^{-1} + (I - A^+A)L,$$

where L is an arbitrary $p \times n$ matrix. We have $P_A = AA^+$ and $AA^+A = A$ and so $W = AB = AA^+\Sigma_{YX}\Sigma_{XX}^{-1} + A(I - A^+A)L = P_A\Sigma_{YX}\Sigma_{XX}^{-1} + (A - AA^+A)L = P_A\Sigma_{YX}\Sigma_{XX}^{-1}$, which is (6). Multiplication of (4) by A' on the right yields $W\Sigma_{XX}W' = \Sigma_{YX}W'$ and (7) follows as usual. Observe that in order for A and B to determine a critical point of E, L must in general be constrained by (4); $L = 0$ is always a solution.

In any case, as in the proof of fact 4 for full rank A, if *rank* $A = r$ we conclude that $P_{U'A}$ is an orthogonal projector of rank r commuting with Λ, so that $P_{U'A} = I_{\mathcal{I}}$ for an index set $\mathcal{I} = \{i_1, \ldots, i_r\}$ with $1 \le i_1 < \ldots < i_r \le n$ and $P_A = UP_{U'A}U' = U_{\mathcal{I}}U_{\mathcal{I}}'$. Again as the column space of A is identical to the column space of $U_{\mathcal{I}}$, we can write A in the form

$$A = [U_{\mathcal{I}}, O]C,$$

where O denotes a matrix of dimension $n \times (p - r)$ with all entries 0. At any critical point A, B of E, A will be of the above form and, from (2), B will be of the form

$$B = A^+\Sigma_{YX}\Sigma_{XX}^{-1} + (I - A^+A)L,$$

where L is constrained by (4). No matter what L actually is, using

$$A^+ = C^{-1}\begin{bmatrix} U_{\mathcal{I}}' \\ O \end{bmatrix}$$

we obtain that

$$\begin{aligned} B &= C^{-1}\begin{bmatrix} U_{\mathcal{I}}' \\ O \end{bmatrix}\Sigma_{YX}\Sigma_{XX}^{-1} + \left(I - C^{-1}\begin{bmatrix} U_{\mathcal{I}}' \\ O \end{bmatrix}[U, O]C\right)L \\ &= C^{-1}\begin{bmatrix} U_{\mathcal{I}}'\Sigma_{YX}\Sigma_{XX}^{-1} \\ O \end{bmatrix} + C^{-1}\left(I_p - \begin{bmatrix} I_r \\ O \end{bmatrix}\right)CL \\ &= C^{-1}\begin{bmatrix} U_{\mathcal{I}}'\Sigma_{YX}\Sigma_{XX}^{-1} \\ P \end{bmatrix} + C^{-1}\begin{bmatrix} O \\ I_{p-r} \end{bmatrix}CL \\ &= C^{-1}\begin{bmatrix} U_{\mathcal{I}}'\Sigma_{YX}\Sigma_{XX}^{-1} \\ \text{last } p - r \text{ rows of } CL \end{bmatrix}. \end{aligned}$$

Now, by assumption, Σ has full rank n, and so $U_{\mathcal{I}}'\Sigma_{YX}\Sigma_{XX}^{-1}$ has full rank r. Upon slightly perturbing the last $p - r$ rows of CB (which are also the last $p - r$ rows of CL, we can always obtain $\tilde{B}$ arbitrarily close to B such that $\tilde{B}$ has maximal rank and $W = A\tilde{B} = AB$ and thus $E(A, B) = E(A, \tilde{B})$. Now $\tilde{B}$ has full rank and so E is strictly convex in the elements of A. Putting $\tilde{A} = (1 - \epsilon)A + \epsilon\hat{A}(\tilde{B})$ with $0 < \epsilon < 1$, we have $E(\tilde{A}, \tilde{B}) < E(A, \tilde{B}) = E(A, B)$. If $\epsilon \to 0$, $\tilde{A} \to A$ and therefore (A, B) is a saddle point for E.