Extremely Ill-posed Learning

L.K. Hansen[a)], B. Lautrup[b)], I. Law[c)], N. Mørch[a)], and J. Thomsen[a)]

[a)] CONNECT, Electronics Institute, build. 349
Technical University of Denmark, DK-2800 Lyngby, Denmark

[b)] CONNECT, The Niels Bohr Institute,
Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

[c)] Dept. of Neurology, The University Hospital of Copenhagen,
Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark

August 29, 1994

Abstract

Extremely ill-posed learning problems are common in image and spectral analysis. They are characterized by a vast number of highly correlated inputs, *e.g.* pixel or pin values, and a modest number of patterns, *e.g.* images or spectra. We show that it is possible to train neural networks to learn such patterns without using an excessive number of weights, and we devise a test to decide if new patterns should be included in the training set or whether they fall within the subspace already explored. The method is applied to the analysis of PET-images.

1 Introduction

The aim of learning is to match a model to data in such a way that generalization ability ensues. If the model is overly restrictive it cannot "capture the rule", hence, fails to implement the training set. On the other hand if we train a model with too high capacity for a given data set, it is unlikely that the model will generalize. The reason is that there will be many different ways to implement the training set in the model, *i.e.* to generalize from it. Training will pick up one rule, usually at random, and it is unlikely that this particular rule will generalize in any desirable way to new examples. We shall not go into the question of what constitutes a desirable generalization, but only note that this concept is often related to simplicity: The most economical model — in terms of free parameters — seems often to be the best.

The solution to the learning problem is therefore not unique, but constitutes an *ill-posed* problem (see [Poggio90] for a review). Many ingenious schemes have been devised in order to design models with various types of simplicity. Regularization by weight decay and by pruning are two prominent schemes for fine tuning network capacity (see for example [LeCun90, Poggio90, Svarer93]).

These schemes are, however, aimed at what could be called *marginally ill-posed learning*, where the number of parameters in the model is comparable to or smaller than the number of training examples. In neural net applications, one often faces a much more singular learning problem, where an example consists of a very large input vector (for example an image or a spectrum), but where it is nevertheless the aim to learn and generalize from a relatively small number of examples. This situation is what we will refer to as *extremely ill-posed learning*.

In this article we show how it is possible to simplify the extremely ill-posed learning problem by straightforward linear algebra *without loss of generality*. The basic idea is similar to the trick in Singular Value Decomposition [Press&al92], and works by transposing the problem from the high-dimensional input space to a low-dimensional "signal space". The success of this transformation depends on an assumption of strong correlations between the components of the input vector. We shall present an *a posteriori* test for the validity of the method.

In particular we show how this works for two paradigms: supervised learning based on feed-forward nets and unsupervised learning based on Sangers network. We use a specific example for illustration and this example has been investigated using Principal Component Analysis (PCA) in several variants including the fast one discussed here [Moeller87]. The example concerns feature extraction and interpretation of Positron Emission Tomography (PET) [Moeller87].

2 Learning in input or signal space

Let us consider a learning problem with a training set of p inputs: $\{\mathbf{x}_\alpha \,|\, \alpha = 1, ..., p\}$. For supervised learning we would also be provided a corresponding set of outputs $\{y_\alpha\}$. Let the dimension of the input space be denoted N. The extremely ill-posed problem occurs for $p \ll N$. In this case it is convenient to consider the linear subspace of input space spanned by the actual inputs of the training set $S = \mathrm{span}\{\mathbf{x}_\alpha\} = \{\mathbf{x} \,|\, \mathbf{x} = \sum_\alpha c_\alpha \mathbf{x}_\alpha\}$. For reference we call S the *signal space*, although this may not be entirely consistent since it is very likely that future (test) inputs will be found outside this subspace. We shall discuss this problem in section 3.

The learning problems that we shall consider are based on a number of adaptive linear forms of the type

$$h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} \tag{1}$$

where $\mathbf{w}$ is an N-dimensional weight vector. Since we only have have $p \ll N$ examples the problem of determining $\mathbf{w}$ from a set of values of this linear form is indeed extremely ill-posed.

One solution is obtained by restricting the weight vector to fall within the signal space, writing

$$\mathbf{w} \equiv \mathbf{w}_{||} = \sum_{\alpha=1}^{p} \gamma_\alpha \mathbf{x}_\alpha \tag{2}$$

with suitable coefficients γ_α. Assuming that the inputs are linearly independent, the coefficients will be uniquely determined. Let us define the *metric* in input space

$$g_{\alpha\beta} = \mathbf{x}_\alpha \cdot \mathbf{x}_\beta \tag{3}$$

which will be non-singular for linearly independent inputs. It is useful to define the conjugate vectors

$$\hat{\mathbf{x}}_\alpha = \sum_\beta (g^{-1})_{\alpha\beta} \mathbf{x}_\beta \tag{4}$$

satisfying[1]

[1] When working in rectilinear skew coordinate systems it is customary to define the conjugate metric $g^{\alpha\beta} = (g^{-1})_{\alpha\beta}$ and using this for raising and lowering indices: $\mathbf{x}^\alpha = g^{\alpha\beta}\mathbf{x}_\beta$, etc. We do not think it worthwhile to introduce this complication here.

$$\mathbf{x}_\alpha \cdot \hat{\mathbf{x}}_\beta = \delta_{\alpha\beta}$$

Then the coefficients are given by the scalar products

$$\gamma_\alpha = \mathbf{w} \cdot \hat{\mathbf{x}}_\alpha$$

It is clear from (1) that the component $\mathbf{w}_\perp$ orthogonal to the signal space (2) plays no role at all, and we shall see below that the network dynamics preserves this property as long as the dynamics only involves scalar products of the above form.

2.1 Supervised learning with feed-forward nets

Let us model an input-output relation by a standard feed-forward network

$$y(\mathbf{x}) = f\left(\sum_{j=1}^{n_\mathrm{H}} W_j g(\mathbf{w}_j \cdot \mathbf{x})\right),$$

where f, g respectively are the output and hidden squashing functions. The network has n_H hidden units, and is in fact a non-linear function of the linear forms, $\mathbf{w}_j \cdot \mathbf{x}$, $j = 1, ..., n_\mathrm{H}$ of the type discussed above. Clearly this network is extremely overparametrized, since we need to adapt more than Nn_H weight-parameters with only $p \ll Nn_\mathrm{H}$ examples.

A training scheme like *Backpropagation* [Rumelhart&al86] is based on a cost function, for example the mean square error function

$$E = \sum_{\alpha=1}^{p} (y_\alpha - y(\mathbf{x}_\alpha))^2 .$$

The cost function is, as noted above, independent of the orthogonal components for each input-to-hidden weight vector. Consequently, any derivative of E with respect to such a component is zero. We find in fact that the gradient is a linear combination of the inputs

$$\frac{\partial E}{\partial \mathbf{w}_j} = \sum_{\alpha=1}^{p} c_{j,\alpha} \mathbf{x}_\alpha$$

with

$$c_{j,\alpha} = 2\left(y\left(\mathbf{x}_{\alpha}\right) - y_{\alpha}\right) f'\left(\sum_{k=1}^{n_{\mathrm{H}}} W_k g\left(\mathbf{w}_k \cdot \mathbf{x}_{\alpha}\right)\right) W_j g'\left(\mathbf{w}_j \cdot \mathbf{x}_{\alpha}\right)$$

This implies that the training dynamics

$$\delta \mathbf{w}_j = -\eta \frac{\partial E}{\partial \mathbf{w}_j}$$

preserves signal space. If we initialize the weight-vectors within the signal space, the dynamics of back-propagation will leave them there.

Expanding the weight vectors in signal space

$$\mathbf{w}_j = \sum_{\alpha=1}^{p} \gamma_{j,\alpha} \mathbf{x}_{\alpha} \tag{5}$$

we note that the natural parameters to optimize are now the expansion coefficients $\gamma_{j,\alpha}$. This explicitly reduces the dimensionality of the optimization problem from $n_{\mathrm{H}} N$ to $n_H p$. We find explicitly the gradient with respect to the expansion coefficents

$$\frac{\partial E}{\partial \gamma_{j,\beta}} = \sum_{\alpha=1}^{p} c_{j,\alpha} g_{\alpha\beta}$$

where the coefficients $c_{j,\alpha}$ are functions of $\gamma_{j,\beta}$ and $g_{\alpha\beta}$ is given by (3). The gradient descent dynamics for the input-to-hidden weights

$$\delta \gamma_{j,\alpha} = -\eta \frac{\partial E}{\partial \gamma_{j,\alpha}}$$

may thus be formulated entirely in terms of the expansion coefficients, when it is used that $\mathbf{w}_j \cdot \mathbf{x}_{\alpha} = \sum_{\beta} \gamma_{j,\beta} g_{\beta\alpha}$. Consequently, one only has to calculate the metric once in order to find the optimal weights by means of gradient descent.

What we have achieved here is a *weight-sharing* construction [LeCun90a] in which the immense weight vectors $\mathbf{w}_j$ are controlled by the much smaller set of parameters $\gamma_{j,\beta}$. In section 3 we will discuss how to set up a "smoke alarm" that goes off whenever a test input has a significant orthogonal component, in which case it should either be rejected or included in the training set.

2.2 Ill-posed unsupervised learning: Sanger's rule

Principal component analysis is a very popular tool in exploratory statistics [Jackson91]. The principal components are defined to be the eigenvectors to the $(N \times N)$ covariance matrix of the (zero mean centered) inputs,

$$C_{ij} = \frac{1}{p} \sum_{\alpha=1}^{p} x_i^\alpha x_j^\alpha \tag{6}$$

The first principal component is the eigenvector corresponding to the maximal eigenvalue of C *et cetera*. By projecting the inputs onto a selected subset of the principal components significant data reduction can be obtained while keeping most of the *variance* in the data set.

Several network constructions have been proposed for estimation of principal components [Oja89, Sanger89]. Sangers network is convenient since it directly provides a given number M of principal components. The network consists of M linear neurons with output

$$y_{j,\alpha} = \mathbf{w}_j \cdot \mathbf{x}_\alpha, \qquad j = 1, \ldots, M$$

and is updated according to the rule [Sanger89]

$$\delta \mathbf{w}_j = \eta \sum_\alpha y_{j,\alpha} \left(\mathbf{x}_\alpha - \sum_{k=1}^{j} y_{k,\alpha} \mathbf{w}_k \right),$$

which guarantees that the principal directions fall out ordered according to size of eigenvalue, such that weight-vector $\mathbf{w}_j$ will contain the $j'th$ principal direction and y_j will be the $j'th$ principal component of the input vector.

Just as for the feed-forward network we would rather work in the p-dimensional signal space. Expanding the weight-vectors as in (5) we find the following update rule for the coefficients

$$\delta \gamma_{j,\alpha} = \eta \left(y_{j,\alpha} - \sum_\beta y_{j,\beta} \sum_{k=1}^{j} y_{k,\beta} \gamma_{k,\alpha} \right)$$

Note that this only depends on the output values

$$y_{j,\alpha} = \sum_\beta g_{\alpha\beta} \gamma_{j,\beta}$$

so that as before one only needs to calculate the metric once and for all in order to implement the network dynamics in signal space.

Working entirely in signal space it is also possible to calculate the principal components of the correlation matrix (6). Let us assume that $\mathbf{u}$ is an eigen-vector of the correlation matrix, *i.e.* that $\mathbf{Cu} = \lambda\mathbf{u}$. Multiplying from the left by $\mathbf{x}_\alpha$ we get

$$\frac{1}{p}\sum_\beta g_{\alpha\beta}\mathbf{x}_\beta \cdot \mathbf{u} = \lambda \mathbf{x}_\alpha \cdot \mathbf{u}$$

which shows that the signal space vector $\mathbf{x}_\alpha \cdot \mathbf{u}$ either vanishes or is an eigen-vector of the matrix $\frac{1}{p}g_{\alpha\beta}$ with eigen-value λ. Since all the eigen-values of the metric are non-vanishing by the assumption of linear independence, it follows that the two matrices C_{ij} and $\frac{1}{p}g_{\alpha\beta}$ have exactly the same non-vanishing eigen-values. The eigen-vectors of the non-vanishing eigen-values are related by

$$\mathbf{u} = \sum_\alpha u_\alpha \hat{\mathbf{x}}_\alpha, \qquad u_\alpha = \mathbf{x}_\alpha \cdot \mathbf{u}$$

The problem of finding the principal components of a small set of large input vectors has now been reduced to diagonalizing a matrix in the low-dimensional signal space. This diagonalization may conveniently be carried out using the Sanger network in signal space.

3 Generalization and rejection

In the preceeding section we have projected an unmanageably large set of inputs onto the much smaller signal space S. We must now address the question of what happens when new input vectors are included in the analysis, either for test or for further training.

A new input will most probably fall outside the already established signal space for any realistic system with noise. We therefore need to test whether the new input has a *significant* component orthogonal to the signal space, in which case we should reject the input or take actions to include the example in the training set, *i.e.* augment the signal space with the new example. If the orthogonal component is insignificant, on the other hand, we can hopefully trust the output of the network for this example.

The magnitude of the orthogonal component of an arbitrary vector $\mathbf{x}$ is easily found to be

$$\mathbf{x}_\perp^2 = \mathbf{x}^2 - \sum_{\alpha\beta}(g^{-1})_{\alpha\beta}(\mathbf{x}\cdot\mathbf{x}_\alpha)(\mathbf{x}\cdot\mathbf{x}_\beta)$$

expressed in quantities that refer to the signal space. A *leave-one-out* cross-validation scheme may now be used to obtain a scale for the expected magnitude of the orthogonal components [Jackson91].

To do so, we form p subsets of the training set, each containing $p-1$ training examples and one test example. Based on each subset we obtain as described above the magnitude of the orthogonal component of the left-out example. Since inversion of a $p \times p$ matrix effectively involves the inversion of all submatrices, it is not surprising that no further calculation has to be done beyond the inversion of the original $p \times p$ metric. From the definition (4) we find easily the relation

$$\mathbf{x}_\alpha = -\sum_{\beta \neq \alpha} \frac{(g^{-1})_{\alpha\beta}}{(g^{-1})_{\alpha\alpha}} \mathbf{x}_\beta + \frac{\hat{\mathbf{x}}_\alpha}{(g^{-1})_{\alpha\alpha}} \equiv \mathbf{x}_\alpha^{\parallel} + \mathbf{x}_\alpha^{\perp}$$

which resolves the example $\mathbf{x}_\alpha$ into a component parallel to the subspace spanned by all the other $p-1$ examples, and a component orthogonal to this subspace. The magnitude of the orthogonal component is now found to be

$$(\mathbf{x}_\alpha^{\perp})^2 = \frac{1}{(g^{-1})_{\alpha\alpha}}$$

The size of the orthogonal component relative to the size of the vector is

$$\frac{(\mathbf{x}_\alpha^{\perp})^2}{\mathbf{x}_\alpha^2} = \sin^2 \phi_\alpha = \frac{1}{g_{\alpha\alpha}(g^{-1})_{\alpha\alpha}}$$

defining the elevation ϕ_α between the vector and the subspace.

The statistics of the leave-one-out sample can be used to get a test for significance of the orthogonal components of future inputs. There are several options. We could test for significance under a hypothesis on their distribution. Alternatively, a pragmatic approach would be to let the alarm go off whenever an orthogonal component has an elevation larger than any of the ones seen in the training set. *Asymptotic* expressions for the statistics of the leave-one-out sample are given in [Jackson91].

4 Application to PET scans

Positron-Emission-Tomography (PET) is an important tool for providing high resolution 3-D images of metabolic and physiological processes and is a widely used clinical and experimental method for study of the human brain. When correlated with information about the physical stimuli and physiological state (cognitive functions,

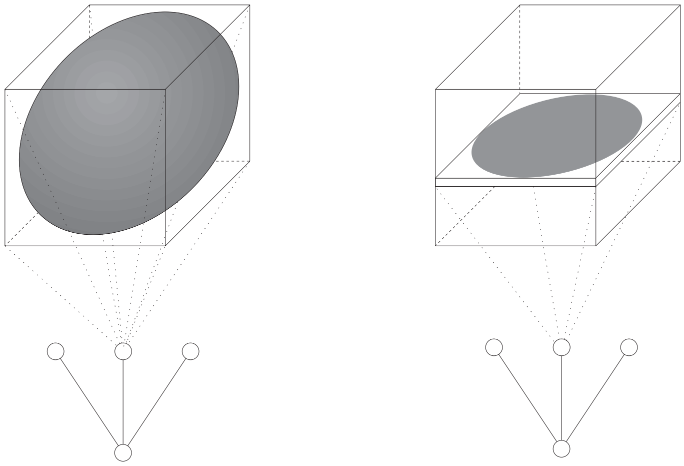

Figure 1: Network architecture (with 147030 input units, 3 hidden units and 1 output unit) for predicting the frequency of an activation paradigm from a PET volume scan of four subjects. The network is trained from 28 examples, hence, gravely overparametrized, *i.e.* an extremely ill-posed learning problem. By projecting the learning problem onto signal space, however, the computational burden is dramatically reduced. By further use of weight decay the problem is converted to a well-posed learning problem with about 10 effective parameters learned. In the right panel is indicated how the weights connecting to particular slice can be visualized as an image.

motion, etc.) such scans provide clues to the underlying functional connectivity between essential nodes of the brain at a given behavior.

Most previous studies on correlation of activity patterns and brain function are based on a combination of PCA and linear analysis. However, in a recent study, neural networks were used to discriminate PET images of a control group from that of patients with Alzheimer's desease [Kippenham92]. Singular Value Decomposition (SVD) techniques, as descibed in this paper, have been used on PET scans to facilitate (linear) PCA analysis. In particular, it is an integral part of the socalled *Scaled Subprofile Model* [Moeller87].

In a collaborative effort, involving several hospitals and other research institutions in the US, Japan, and Europe, we currently investigate the possibility of invoking artificial neural nets for analysis of functional connectivity in the human brain. In this report we use preliminary PET-based results to illustrate the role of signal space projections for *non-linear* ill-posed learning using neural nets. The PET images of this example were recorded at the Department of Neurology at The University Hospital of Copenhagen, more details regarding the experiment and the particular activation paradigms used may be found in [Law94].

Subjects were scanned under two conditions, *rest*, and a condition with a particular visio-motor activity (in PET-slang such conditions are referred to as *activation paradigms*). The activation paradigm is repeated at seven different frequencies (including rest, counted as zero frequency), resulting in a total of 7 scans per subject. In this experiment the objective is to predict the frequency from the filtered volume data from a PET scan, the training database containing data from four subjects (*i.e.* a total of 28 examples). Input to the network is created by a standardized normalization procedure aimed at eliminating relative displacements and rotations of subjects, and furthermore, the input volume is *centered* which means that the average activity pattern of the volume has been calculated and subtracted.

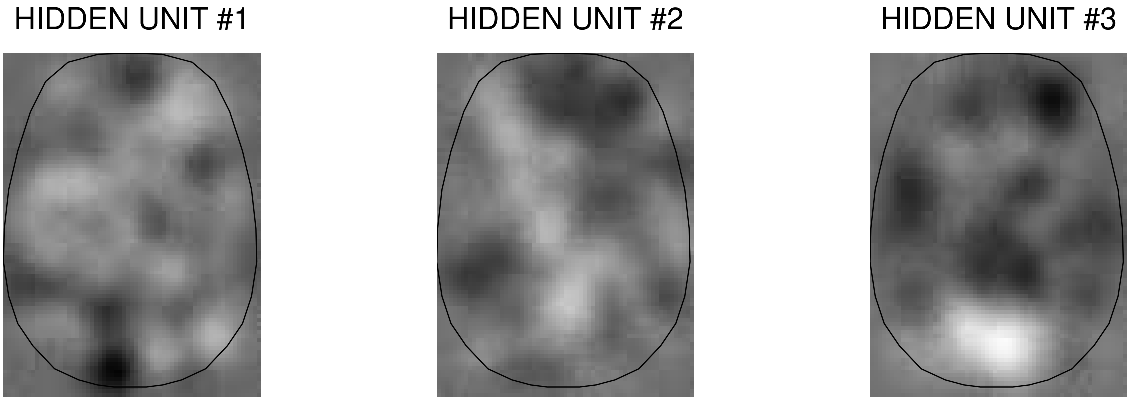

Figure 2: Image visualization of the weights connecting from three hidden units to a slice of the PET volume scan. The weights are shown in a linear gray scale with positive weights bright, and negative weights dark. Note that the hidden units pick up signals from different regions of the activated brain (outlined in black).

Since the volume scan contains 26 slices each holding 65×87 pixels, *i.e.* 147,030 voxels, the initial network, having 3 hidden units, and a single output, is gravely overparameterized (with more than 400,000 weights and only 28 examples), and the learning problem is indeed extremely ill-posed. By projecting the input volumes onto signal space the dimensionality of input space is brought from 147,030 down to 28. While this projection, on its own, does not hinder overfitting it does reduce the computational burden dramatically. To minimize overtraining weight decay has been applied. The magnitude of the weight decay parameter has been determined so that the *effective* number of fitted parameters[2] is $N_{\text{eff}} = 10$. The network architecture is visualized in figure 1. In the right panel of the figure we have marked a particular slice of the voxelated volume, – the weights of each of the three hidden units connecting to this slice are pictured in figure 2, note that the hidden units pick up different, and rather well defined, regions of the activated brain. Current research is aimed at interpretation of such weight images (weight volumes). For further illustration of the ability of the particular network we show in figure 3 the training set frequency predictions.

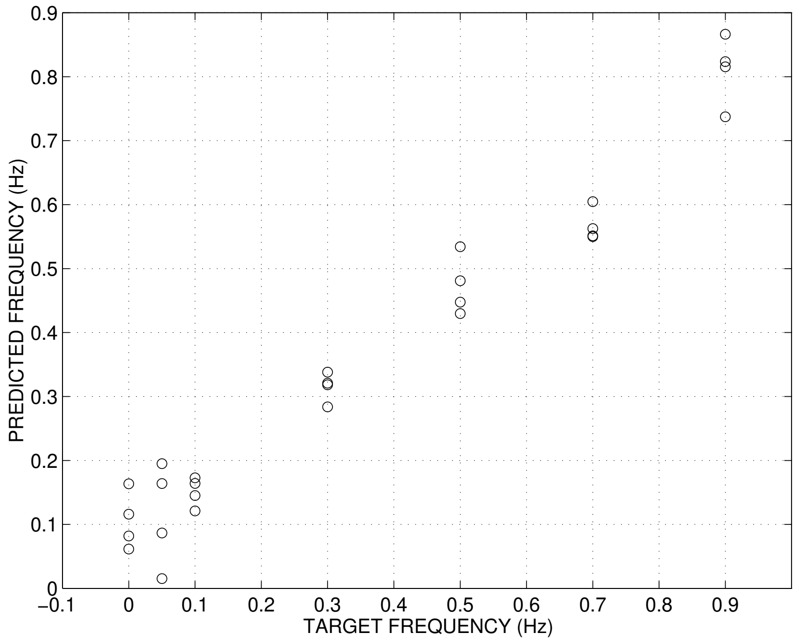

Figure 3: Prediction accuracy of the trained network on the 28 data points in the training set. The network was trained using a weight decay of $\alpha = 0.06$, this leaves about 10 effective degrees of freedom for the fit.

5 Conclusion

We have provided a general recipe for handling extremely ill-posed learning problems. Whenever a learning system based on adaptive linear forms on a huge input space is to be trained on a small training set, it is advantageous to reexpress the linear forms in terms of the training set input vectors without loss of information. The

[2]The effective number of parameters has been calculated as $N_{\text{eff}} = \text{Tr}\left[\mathbf{H}\mathbf{J}^{-1}\mathbf{H}\mathbf{J}^{-1}\right]$, where $\mathbf{H}$ is the second derivative matrix of the training set error, the *Hessian*, and, $\mathbf{J} = \mathbf{H} + \alpha\mathbf{1}$, is the Hessian of the cost function augmented by weight decay (see e.g. [Moody92, Svarer93])

mechanism can be viewed as a particular construction for obtaining massive weight sharing. We have shown how the mechanism works for supervised learning based on the conventional feed-forward net as well as for unsupervised learning based on the Sanger network. In addition to a dramatic reduction of computational effort, the scheme provides a natural mechanism for outlier rejection. In our example we have shown how a network with more than 400,000 weights may be adapted for analysis of PET images.

Acknowledgments

This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center, CONNECT. We thank Steve Strother for instructive discussions on PCA and PET images, Claus Svarer for assistance with the PET images, and Adriana Dumitras for useful comments on the manuscript.

References

[Jackson91] J. Edward Jackson:
A User's Guide to Principal Components,
Wiley Series on Probability and Statistics, John Wiley and Sons (1991).

[Kippenham92] J.S. Kippenham, W.W. Barker, S. Pascal, J. Nagel, and R. Duara:
Evaluation of a neural-network classifier for PET scans of normal and Alzheimers Disease Subjects,
J.Nucl.Med. **33** 1459-1467, (1992).

[Law94] Ian Law et al.:
Saccade Inhibition, Reflection, Selection, and Imagination: A PET-study,
In preparation (1994).

[LeCun90] Y. Le Cun, J.S. Denker, and S.A. Solla:
Optimal Brain Damage,
In *Advances in Neural Information Processing Systems 2*, 598-605, Morgan Kaufman, (1990).

[LeCun90a] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jakel:
Handwritten Digit Recognition with a Back-Propagation Network,
In *Advances in Neural Information Processing Systems 2*, 396-404. Morgan Kaufman (1990).

[Moeller87] J.R. Moeller, S.C. Strother, J.J. Sidtis, and D.A. Rottenberg:
Scaled Subprofile Model: A Statistical Approach to the Analysis of Functional Patterns in Positron Emission Tomographic Data,
J. Cereb. Blood Flow Metab. **7**, 649-658, (1987).

[Moody92] J.E. Moody:
The effective number of parameters: An analysis of generalization and regularization in non-linear learning systems,
in Neural Information Processing Systems 4, Eds. J. Moody et al.; Morgan Kaufmann, San Mateo CA, pp. 847-854, (1992).

[Oja89] E. Oja:
Neural Networks, Principal Components, and Subspaces,
International Journal of Neural Systems **1**, 61–68 (1989).

[Poggio90] T. Poggio and F. Girosi:
Networks for Approximation and Learning,
IEEE Proceedings **78** 1481-1497 (1990).

[Press&al92] W. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling:
Numerical Recipes in C, The Art of Scientific Computing,
Cambridge University Press, Cambridge (1992).

[Rumelhart&al86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams:
Learning Representations by Back-propagating Errors,
Nature **323**, 533–536 (1986).

[Sanger89] T. D. Sanger:
Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network,
Neural Networks **2**, 459–473 (1989).

[Svarer93] C. Svarer, L.K. Hansen, and J. Larsen:
On Design and Evaluation of Tapped-Delay Neural Networks,
Proc. of the IEEE Int. Conf. on Neural Networks 1993. Eds. H.R. Berenji et al., pp 45-51, (1993).