MASSIVE WEIGHT SHARING:
A CURE FOR EXTREMELY ILL-POSED PROBLEMS [*]

BENNY LAUTRUP
CONNECT, The Niels Bohr Institute
DK-2100 Copenhagen Ø, Denmark
e-mail: lautrup@connect.nbi.dk

LARS KAI HANSEN
CONNECT, Electronics Institute, Technical University of Denmark
DK-2800 Lyngby, Denmark

IAN LAW, NIELS MØRCH, CLAUS SVARER
Dept. of Neurology, University Hospital of Copenhagen
DK-2100 Copenhagen Ø, Denmark

and

STEPHEN C. STROTHER
PET Imaging Service, VA Medical Center
Minneapolis, Minnesota, USA

In most learning problems, adaptation to given examples is well-posed because the number of examples far exceeds the number of internal parameters in the learning machine. Extremely ill-posed learning problems are, however, 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. In this paper we show, for the case of a set of PET images differing only in the values of one stimulus parameter, that it is possible to train a neural network to learn the underlying rule without using an excessive number of network weights or large amounts of computer time. The method is based upon the observation that the standard learning rules conserve the subspace spanned by the input images.

1. Introduction

The aim of learning is to match a model to data in such a way that generalization ability ensues. Whether this is possible depends intricately on the training procedure and on the architecture of the learning machine. 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

[*] Talk presented at the workshop on Supercomputing in Brain Research: From Tomography to Neural Networks, HLRZ, Jülich, Germany, Nov. 21–23, 1994

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.

A first guide to the learning problem can be obtained from a comparison of the number of examples, p, used for training, and the number of parameters, n, used to implement the model. In order to allow a model to have sufficient initial capacity, the number of parameters may conveniently be chosen to be of the order of the number of examples $n \approx p$. Thus, the solution to the learning problem is in general not unique, but constitutes an *ill-posed* problem (see[10] for a review). Many ingenious schemes have been devised in order to reduce the complexity of the final model[6,10], such that in the end one has effectively $n \ll p$. Regularization by weight decay and by pruning are two prominent schemes for fine tuning of the network capacity (see for example[2]).

These schemes are, however, aimed at what could be called *marginally ill-posed learning*, where the number of parameters in the model initially is comparable to or smaller than the number of training examples, $n \leq p$. 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, where initially $n \gg p$, was recently analysed in[1] and is what we refer to as *extremely ill-posed learning.*

However, we showed how it is possible to cure the extremely ill-posed learning problem by straightforward linear algebra *without loss of generality.* The basic idea is similar to the trick that enters Singular Value Decomposition[11], which works by transposing the problem from the high-dimensional input space to a low-dimensional "signal space". The effect of the procedure is to introduce massive weight sharing by constraining the network weights to a low-dimensional subspace. By this transformation the extremely ill-posed problem is converted to a *marginally* ill-posed problem which can then be handled by, regularization, *e.g.*, using weight decay as in[1]. In this presentation we review the scheme for handling extremely ill-posed problems and we present an alternative approach for regularization of the transformed problem based on *pruning* of superflous network weights.

There is of course no such thing as a free lunch, so the success of the transformation depends on an assumption of strong correlations between the components of the input vector. We shall in turn present an *a posteriori* test for the validity of this assumption.

In particular we show how the cure works for backpropagation learning in a feed-forward network (it also works for unsupervised learning[1]). As learning problem we consider a set of images obtained from Positron Emission Tomography (PET)[5]. Neural networks have been used for diagnostic purposes by Kippenham et al[4]. In our case, a single parameter, the frequency of induced saccadic eye motion, is varied, and the aim is to learn to predict the frequency from the image. We show that even if the network *a priori* contains of the order of 420,000 parameters, it is nevertheless possible to capture the one-parametric rule, using only 48 images! Furthermore, we present an analysis of the pruned network, showing that the network model discards input channels that are related to "inter-subject" variation. Although these inputs carry large parts of the variance in the data set they are largely irrelevant from a modeling point of view. This finding extends recent results obtained from linear models (using the so-called Scaled Subprofile Model[13]), to the realm of non-linear modeling.

2. Supervised learning in signal space

Let us consider a supervised learning problem with a training set consisting of p inputs: $\{\mathbf{x}_\alpha \,|\, \alpha = 1, ..., p\}$ and a corresponding set of outputs $\{y_\alpha\}$. Let the dimension of the input space be denoted N. Let the network be of the usual feed-forward type with one layer of L hidden neurons and a single linear output neuron. Then the number of parameters — weights and thresholds — is $n = (N+2)L + 1$.

The input-output relation of this network is

$$y(\mathbf{x}) = \sum_{j=1}^{L} W_j f(\mathbf{w}_j \cdot \mathbf{x} - \theta_j) - \Theta, \tag{1}$$

where f is the hidden unit squashing functions and the thresholds are denoted θ_j and Θ for the hidden and output neurons, respectively. The input-to-hidden weights are vectors in the N-dimensional input space and denoted $\mathbf{w}_j$ for those weights connected with the j-th hidden neuron. The hidden-to-output weights are denoted W_j and may be thought of as importance coefficients for a committee made up by the neurons of the hidden layer.

A training scheme like *Backpropagation*[12] is based on a cost function, for example the mean square error

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

The training dynamics for the non-linear neurons in the hidden layer is of the gradient descent type[12]

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

whereas the hidden-to-output weights may be determined by global minimization of the cost function which is quadratic in these parameters.

It is evident from the form of the cost function and the network equation (1) that the gradient is a linear combination of the input vectors

$$-\frac{\partial E}{\partial \mathbf{w}_j} = \sum_{\alpha=1}^{p} c_j^\alpha \mathbf{x}_\alpha$$

where the coefficients are given by

$$c_j^\alpha = (y_\alpha - y(\mathbf{x}_\alpha)) W_j f'\left(\mathbf{w}_j \cdot \mathbf{x}_\alpha - \theta_j\right) \tag{2}$$

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

In the normal case where the number of examples is greater than the number of input values, this is of no interest. In the extremely ill-posed problem which occurs for $p \ll N$, it is highly useful to work entirely within the much smaller signal space spanned by the actual inputs of the training set $S = \mathrm{span}\{\mathbf{x}_\alpha\}$.

Expanding the weight vectors in signal space

$$\mathbf{w}_j = \sum_{\alpha=1}^{p} \gamma_j^\alpha \mathbf{x}_\alpha$$

we note that the natural parameters to optimize are now the expansion coefficients γ_j^α. This explicitly reduces the dimensionality of the optimization problem from LN to $Lp \ll LN$. The gradient descent dynamics for the γ-coefficients becomes

$$\delta\gamma_j^\alpha = -\eta \frac{\partial E}{\partial \gamma_j^\alpha}$$

and the gradient is explicitly found to be

$$-\frac{\partial E}{\partial \gamma_j^\beta} = \sum_{\alpha=1}^{p} c_j^\alpha g_{\alpha\beta}$$

where

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

is the *metric* tensor of the signal subspace of the full input space. The c-coefficients (2) may also be expressed in terms of the γ's

$$c_j^\alpha = (y_\alpha - y(\mathbf{x}_\alpha)) W_j f'\left(\sum_\beta g_{\alpha\beta}\gamma_j^\beta - \theta_j\right)$$

and this is also true for the network equation (1).

What we have achieved here is a *weight-sharing* construction[7] in which the immense weight vectors $\mathbf{w}_j$ are controlled by the much smaller set of parameters γ_j^α. A similar dimensional reduction may also be carried out for all extremely ill-posed learning problems that are based on adaptive linear forms of the type $h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$, for example Sanger's network for principal value decomposition (see ref[1]).

3. Using Principal components

In the generic case the metric $g_{\alpha\beta}$ will be non-singular, *i.e.* have non-zero eigenvalues, λ_α. Denoting the orthonormal eigenvectors of $g_{\alpha\beta}$ by $e_{\alpha\beta}$ (satisfying $\sum_\gamma g_{\beta\gamma} e_{\alpha\gamma} = \lambda_\alpha e_{\alpha\beta}$), we may define a set of eigen-images $\mathbf{e}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \sum_\beta e_{\alpha\beta} \mathbf{x}_\beta$, which are easily seen to be orthonormal themselves. These p images constitute the principal components of the input space and diagonalize (with the same eigenvalues) the input correlation matrix $\sum_\alpha \mathbf{x}_\alpha \mathbf{x}_\alpha$. The remaining $N - p$ eigen-images of this matrix all have vanishing eigen-values and are orthogonal to the signal space spanned by the inputs and thus of no importance for the training process. Notice that in order to calculate the non-trivial eigen-images, it is only necessary to diagonalize the generally much smaller matrix, $g_{\alpha\beta}$ (in accordance with SVD[11]).

It is sometimes convenient to formulate the training algorithm entirely in the principal images. Writing $\mathbf{w}_j \cdot \mathbf{x}_\alpha = \sum_\beta (\mathbf{w}_j \cdot \mathbf{e}_\beta)(\mathbf{e}_\beta \cdot \mathbf{x}_\alpha)$ we see that we may simply train the eigen-coordinates of the weights $(\mathbf{w}_j \cdot \mathbf{e}_\beta)$ using the eigen-coordinates of the images $(\mathbf{e}_\beta \cdot \mathbf{x}_\alpha)$ as inputs. This has been done in the example presented below.

4. 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[1]

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

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[3].

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. The magnitude of the orthogonal component is 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

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

where we have introduced 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.

5. Application to PET scans

Positron Emission Tomography (PET) is an important tool for mapping metabolic processes in the human brain. PET scans provide reasonably fine grained, three-dimensional information on patterns of metabolic activity. When correlated with information about physical and mental conditions (cognitive functions, motion, etc.) such scans provide clues to brain function and functional connectivity.

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 scans of a control group from those of patients with Alzheimer's disease[4]. Singular Value Decomposition (SVD) techniques have been used on PET scans to facilitate (linear) PCA analysis. In particular, it has become an integral part of the socalled *Scaled Subprofile Model*[8,13].

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[5].

Subjects were scanned under two conditions, one during rest, and one with a particular visual activity, in this case a voluntary saccadic eye movement consisting in tracking a moving light. The activation paradigm is repeated at six different frequencies and two rest states, counted as zero frequency, resulting in a total of 8 scans per subject. Altogether 8 subjects were examined resulting in a total of 64 brain volume images. Of these 48 were randomly chosen as training examples and 16 used for validation and testing.

In this analysis the objective is to predict the frequency from the filtered volume data from a PET scan. 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.

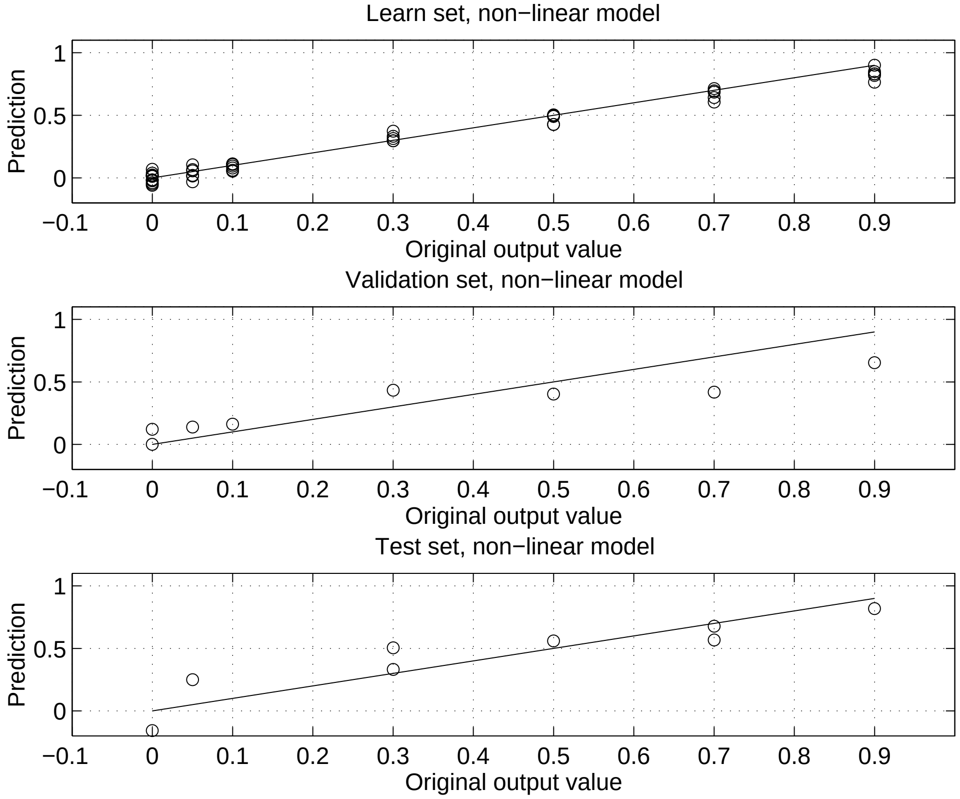

Fig. 1. Training and test error for an extremely ill-posed problem. In the first panel of the figure the output of the pruned network is compared to the control frequency. In the second and third, the performance of the pruned network on the validation and test sets is displayed.

Since the volume scan after coregistration and stereotactic normalization contains 25 slices each holding 65 × 87 pixels, *i.e.* 141,375 voxels, the initial network, having 3 hidden units, and a single output, is gravely overparameterized (with about 420,000 weights and only 48 examples), The learning problem is indeed extremely ill-posed. By projecting the input volumes onto signal space the dimensionality of input space is brought from 141,375 down to 47 (the average image has been subtracted).

While this projection, on its own, does not hinder overfitting it does reduce the computational burden dramatically. To further optimize the capacity of the trained network, we here investigate pruning by *Optimal Brain Damage*[6]. In this scheme second order properties of the error function (the Hessian) are used to select the candidates for connections to be severed. As pruning proceeds, the training set error increases because of the loss of degrees freedom available for the fit. To estimate the more interesting *generalization ability* we adopt a cross-validation procedure. The test set consisting of 16 randomly selected cases is divided into a test set and a validation set. The validation set is used to identify the optimal network among the nested family of pruned nets. The remaining test cases are used to provide an unbiased estimate of the generalization error of the selected network.

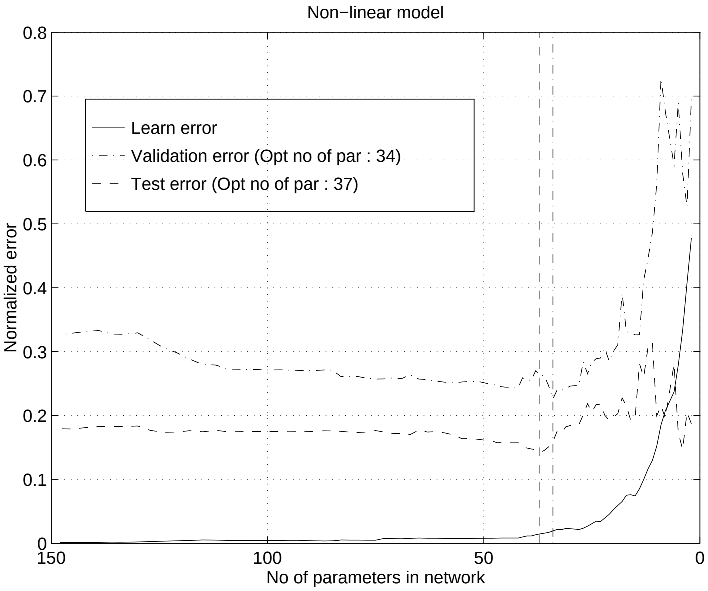

Fig. 2. Pruning run, in which the number of parameters (weights) are progressively removed from the network. The test set has been split into two sets each holding 8 examples. The two dashed lines represent the errors on the two test subsets. Selecting the “optimal” net on either of the two test sets produce the two vertical dashed lines. The optimal nets have 34 and 37 parameters respectively.

6. Results

The main result of training the network is shown in fig. 1. In this case 48 of the 64 examples have been used for training and the remaining for testing. Whereas the training is almost perfect with a residual error of 0.0103, the error on the test set is 0.224. In spite of this large generalisation error, it is evident from the figure that some generalisation is in fact obtained.

The optimal network selection process is depicted in fig. 2. In this figure the number of parameters is reduced from the originally fully connected state with 148 parameters to increasingly more sparsely connected networks. To select a net from the nested family of networks, we use the mentioned crosvalidation scheme in which a set of 8 examples were randomly chosen as validation set.

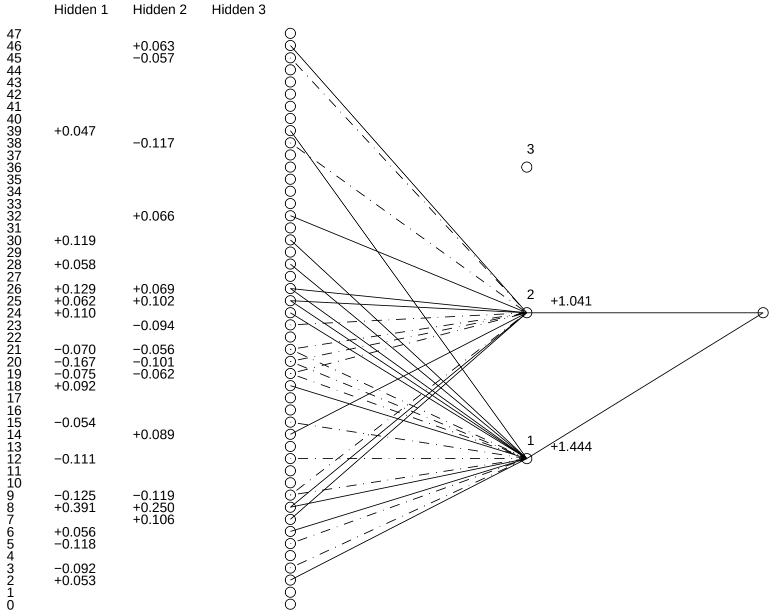

Fig. 3. Pruned net corresponding to minimal validation error. Excitatory connections are marked with full, inhibitory with broken lines. Note that several input channels (principal components) have been eliminated from the model, and that one hidden unit has been discarded. It is interesting that the network emphasizes the eighth principal component and puts less emphasis on the first seven components that carry much more variance. The finding is in line with recent linear model based studies: it turns out that the variance carried by the seven first component is in fact associated with intersubject variation, hence, to be considered as “noise” when it comes to predict the frequency parameter of the activation paradigm.

The optimal net obtained after pruning to minimum test error contains 34 parameters and is shown in fig. 3. In order to decorrelate the input images, they have been resolved into principal components[3], which corresponds to diagonalization of the metric tensor. This preprocessing is a reversible linear operation which does not change the basic problem. It is interesting, however, to note that the nonlinear model emphasizes (by large weights) the eighth principal component. This is in line with recent findings in[13] based on linear modeling. It was found that the first principal components correspond to intersubject variation.

Like in[13] the variation of the PC's are partitioned into intersubject, frequency and residual components. This partitioning may be carried out for each PC,

$$\Sigma^2_{\text{total}} = \Sigma^2_{\text{intersubject}} + \Sigma^2_{\text{frequency}} + \Sigma^2_{\text{residual}} \tag{3}$$

where $\Sigma^2_{\text{total}} = \sum_{s=1}^{N_s} \sum_{f=1}^{N_f} (y_{s,f} - y_{*,*})^2$ quantifies the total variation in the data set, $\Sigma^2_{\text{intersubject}} = \sum_{s=1}^{N_s} (y_{s,*} - y_{*,*})^2$ the intersubject variation, and $\Sigma^2_{\text{frequency}} = \sum_{f=1}^{N_f} (y_{*,f} - y_{*,*})^2$ the variation induced by the frequency stimulus parameter. In these expressions we have denoted averaged quantities with an asterisk.

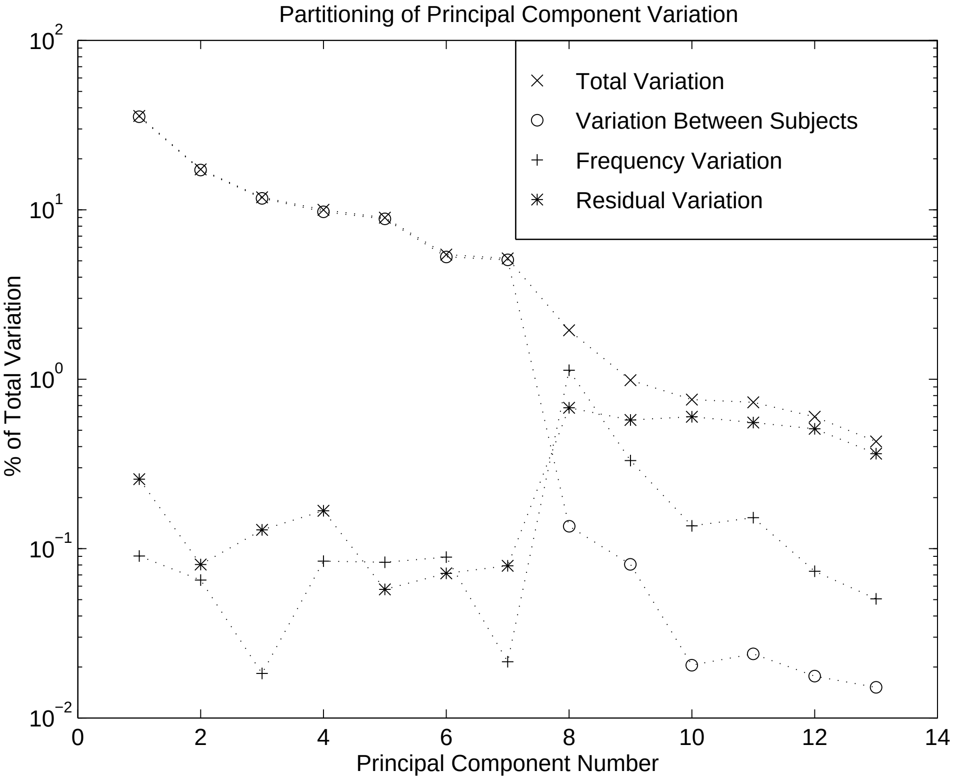

Fig. 4. Partitioning of the variation for the first 13 principal components. The total variation (×) is partitioned into intersubject variation (∘), frequency variation (+), and residual variation ($*$). For the first seven PC's the variation is solely carried by intersubject variation, while for the eight'th PC a sizeable fraction of the variation stems from the frequency stimulus.

As seen in figure 4 the variation associated with the first seven PC's is due to intersubject variation, i.e., variation that stems from the difference in activation patterns among different subjects. However, for the 8'th PC, which is important to the neural network, about 60% of the variation is accounted for by the frequency stimulus.

To analyse the consistency of the signal space projection approach we have computed the cross validation distribution of elevation angles as shown in the upper panel of fig. 5. Most examples have very small angles, reflecting the high correlation between voxel values. One training example seems to be an outlier and should possibly be excluded from the training set. The elevation angles of the test set (lower panel), however, are accepted as samples from the distribution obtained by crossvalidation within the training set.

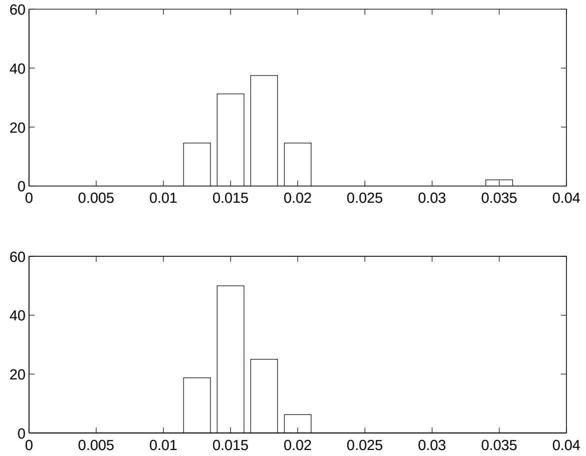

Fig. 5. Elevation distribution for full set of images, computed by leave one out cross validation. The ordinate is percent of maximum, the abscissa the angle in radians. On top the distribution of elevations for the training set, at bottom the angles between the training set and test images. The outlier in the training set corresponds (not unsurprisingly) to zero frequency scans. All patterns in the test set are accepted as samples from the empirical distribution derived by cross validation with the training set.

7. 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 mechanism can be viewed as a particular construction for obtaining massive weight sharing. 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 of the order of 420,000 weights may be adapted for analysis of PET images. By further pruning of the weight shared network we rediscovered the fact that most of the variance in the data set is intersubject variation hence irrelevant for modeling of the activation paradigm.

Acknowledgements

This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center, CONNECT.

References

1. L.K. Hansen, B. Lautrup, I. Law, N. Mørch, S.C. Strother, and J. Thomsen: *Extremely Ill-posed Learning,* CONNECT preprint, August 1994.
2. J. Hertz, A. Krogh and R.G. Palmer: *Introduction to the Theory of Neural Computation,* Addison Wesley, New York (1991).
3. J. Edward Jackson: *A User's Guide to Principal Components*, Wiley Series on Probability and Statistics, John Wiley and Sons (1991).
4. 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).
5. I. Law et al: *Saccade inhibition, reflection, selection and imagination: A PET-study,* In preparation, 1994.
6. 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).
7. 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).
8. 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).
9. J.E. Moody: *The effective number of parameters: An analysis of generalization and 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).
10. T. Poggio and F. Girosi: *Networks for Approximation and Learning,* IEEE Proceedings **78** 1481-1497 (1990).
11. 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).
12. D.E. Rumelhart, G.E. Hinton, and R.J. Williams: *Learning Representations by Backpropagating Errors*, Nature **323**, 533–536 (1986).
13. S.C. Strother, J.R. Anderson, K.A. Schaper, J.J. Sidtis, J.S. Liow, R.P. Woods, and D.A. Rottenberg: *Principal Component Analysis and the Subprofile Scaling Model Compared to Intersubject Averaging and Statistical Parametric Mapping: I. "Functional Connectivity of the Human Motor System Studied with* $[^{15}O]$*water PET",* Preprint, VA Medical Center, Minneapolis, Minn., USA.
14. 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).