Nonlinear versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover

Niels Mørch[1,2], Lars K. Hansen[2], Stephen C. Strother[3], Claus Svarer[1]
David A. Rottenberg[3], Benny Lautrup[4], Robert Savoy[5], Olaf B. Paulson[1]

[1] Neurobiology Research Unit
Copenhagen University Hospital, Rigshospitalet
DK-2100 Copenhagen Ø, Denmark

[2] Department for Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

[3] Radiology and Neurology Departments
University of Minnesota
and
PET Imaging Service
Minneapolis VA Medical Center
Minnesota, 55417, USA

[4] Niels Bohr Institute
University of Copenhagen
DK-2100 Copenhagen Ø

[5] Massachusetts General Hospital
Boston, Massachusetts, USA

Abstract. We introduce the concept of generalization for models of functional neuroactivation, and show how it is affected by the number, N, of neuroimaging scans available. By plotting generalization as a function of N (i.e. a "learning curve") we demonstrate that while simple, linear models may generalize better for small N's, more flexible, low-biased nonlinear models, based on artificial neural networks (ANN's), generalize better for larger N's. We demonstrate that for sets of scans of two simple motor tasks—one set acquired with $[O^{15}]$water using PET, and the other using fMRI—practical N's exist for which "generalization crossover" occurs. This observation supports the application of highly flexible, ANN models to sufficiently large functional activation datasets.
Keywords: Multivariate brain modeling, ill-posed learning, generalization, learning curves.

1 Introduction

Datasets that result from functional activation studies of the living, human brain typically consist of two corresponding sets of observables, the *microscopic* and the *macroscopic* [26]. The brains haemodynamic response, reflecting the microscopic neuronal firing pattern, is measured by modern three-dimensional (3D) imaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) by integrating in space and time [21]. Along with the resulting set of 3D image volumes (scans) a corresponding set of macroscopic descriptors governs the overall conditions of the experiment. This set can include experimentally controlled factors, such as paradigm labels and variables, and physiological and demographic measures, such as age and heart-rate. The micro- and macroscopic observables are generally both sets of multivariate, stochastic variables. Arranging the microscopic variables (the 3D image volumes) in vectors $\mathbf{x}$ and the macroscopic variables in vectors $\mathbf{g}$ a functional activation dataset $\mathcal{D}$ consisting of N observations can be written as

$$\mathcal{D} = \{(\mathbf{x}_j, \mathbf{g}_j) \mid j = 1, \ldots, N\} \quad . \tag{1}$$

Generally, we will assume the observations to be random, independent samples of an underlying stationary process with distribution $\mathrm{P}(\mathbf{x}, \mathbf{g})$. As we shall see this distribution plays a central role in the analysis of functional activation datasets [18].

In the following we discuss the so-called "curse of dimensionality" that results from the extremely ill-posed nature of typical functional activation datasets [6,23]. The problem is discussed in terms of probability density estimation and we briefly mention ways to remedy the inevitable over-parameterization that otherwise occurs in modeling procedures based on such datasets [12]. The main point we hope to convey is how model generalization—as studied intensively in other fields dealing with probability density estimation and multivariate modeling [8,13,17,20]—applies to functional neuroimaging [18], and specifically how it is affected by the number, N, of available observations.

2 Models of Functional Activation Datasets

In terms of $\mathbf{x}$ and $\mathbf{g}$ the analysis of functional activation datasets can be phrased as the estimation (of properties) of $\mathrm{P}(\mathbf{x}, \mathbf{g})$. For instance, we can estimate the conditional mean, $\mathrm{E}\{\mathbf{x}|\mathbf{g}\}$, using multivariate linear models as in [7], thus effectively modeling the expected scan from a set of macroscopic variables. Or, we can estimate the alternative conditional mean $\mathrm{E}\{\mathbf{g}|\mathbf{x}\}$, using multivariate linear models as in [18], effectively modeling the expected value of a set of macroscopic variables from the scan[1].

In general, we employ parameterized models of the properties we wish to estimate. In this work we focus on models that estimate $\mathrm{E}\{\mathbf{g}|\mathbf{x}\}$. Being a function of $\mathbf{x}$ we denote these models $f_\theta(\mathbf{x})$, explicitly indicating the dependency on the set of parameters θ. Parameter values are estimated using some or all of the available data. We call such a set of data used for parameter estimation the *training set*,

$$\mathcal{D}_{train} = \{(\mathbf{x}_j, \mathbf{g}_j) \mid j = 1, \ldots, N_{train}\} \quad . \tag{2}$$

[1] In fact, it can be shown that the two linear models are analogous and simple relations between the parameters exist.

For a given set of parameters model performance is quantified using the *cost function*, $\mathrm{c}(\mathbf{x}, \mathbf{g}, \theta)$, which is often derived from maximum likelihood (ML) arguments [4,10,14]. Parameter values are estimated by optimizing the cost function based on the observations in the training set (we say that the model is trained, hence the name). Averaged over the training set this evaluates to

$$\mathrm{C}(\mathcal{D}_{train}, \theta) = \iint \mathrm{c}(\mathbf{x}, \mathbf{g}, \theta) \mathrm{P}_{train}(\mathbf{x}, \mathbf{g}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{g} \ . \tag{3}$$

By using the empirical density estimate $\mathrm{P}_{train}(\mathbf{x}, \mathbf{g}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \delta(\mathbf{x} - \mathbf{x}_j, \mathbf{g} - \mathbf{g}_j)$ we get the so-called *training error*

$$\mathrm{C}(\mathcal{D}_{train}, \theta) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \mathrm{c}(\mathbf{x}_j, \mathbf{g}_j, \theta), \quad (\mathbf{x}_j, \mathbf{g}_j) \in \mathcal{D}_{train} \ . \tag{4}$$

The choice of cost function will depend on the noise model and potential constraints we impose on the model outputs (e.g. to make them interpretable as probabilities). For more details on these issues see [3,10,14].

Equipped with a training set, a model, and a cost function we are ready to gain knowledge about $\mathrm{P}(\mathbf{x}, \mathbf{g})$ and, hopefully, underlying information processing relationships in the human brain. However, several important additional issues must be considered before attempting to build practical models. Rather than using (4) to model $\mathrm{E}\{\mathbf{g}|\mathbf{x}\}$ from the observations directly we can reduce the computational burden dramatically by taking the extremely ill-posed nature of typical functional activation datasets into account.

2.1 Ill-posed Datasets

While we often include only a few descriptors in the macroscopic variables $\mathbf{g}$ making them low-dimensional, the microscopic variables $\mathbf{x}$ that represent the scans are often high-dimensional. Despite preprocessing that, among other things, mask out voxels outside the brain more than 40000 voxels often remain. Using $\mathcal{I}$ to denote the space in which all possible observations fall (i.e., the *input space*) we have $\dim(\mathcal{I}) \sim 10^4$. The space spanned by the actual observations in the dataset is called *signal space* and denoted $\mathcal{S}$. Often no more than a few hundred observations are available, so $\dim(\mathcal{S}) \sim 10^2$.

Typically $\dim(\mathcal{S}) \ll \dim(\mathcal{I})$, making $\mathcal{S}$ a small subspace of $\mathcal{I}$. This is exactly what characterizes extremely ill-posed datasets. In Fig. 1 an ill-posed situation is illustrated. Input space is 3D Euclidean space indicated by the dashed vectors. With only two observations in the dataset represented by the solid vectors, signal space is a 2D subspace, i.e. a plane. The dataset does not contain information about the parts of $\mathcal{I}$ that are orthogonal to $\mathcal{S}$.

Because the dimension of $\mathcal{S}$ is low we have a correspondingly low number of degrees of freedom available in any subsequent modeling, and naive estimation based directly on the observation pairs $(\mathbf{x}, \mathbf{g})$ will result in strong linear relations between the estimated parameters; the original basis in which observations in

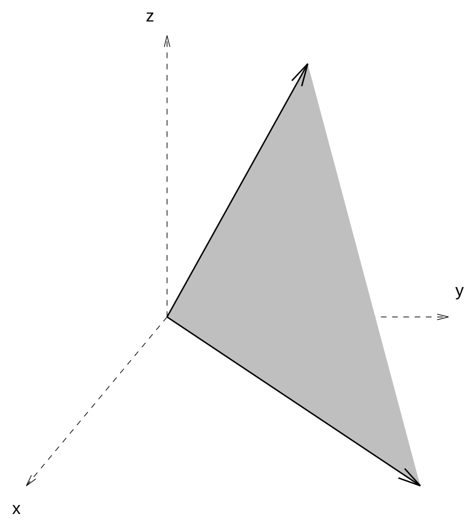

Fig. 1. Illustration of an ill-posed dataset. With input space, $\mathcal{I}$, being three-dimensional (represented by the dashed vectors) the signal space, $\mathcal{S}$, which is the space spanned by the two observations in the dataset (represented by the solid vectors), is the plane indicated in gray. The dataset contains no information about the parts of input space that are orthogonal to signal space because $\dim(\mathcal{S}) < \dim(\mathcal{I})$.

input space are represented is a poor choice when it comes to representing observations efficiently in signal space. We can easily construct other, more efficient bases, however, that reduce the dimensionality of the *representation* without loss of information [12,19]. The only requirement is that the basis chosen spans signal space. One particularly choice of basis is to use the observations in the dataset themselves. Even-though efficient in reducing an extremely ill-posed problem to an only marginally ill-posed one bases that reveal more about the signal structure are available. In particular, a singular value decomposition (SVD) basis [11,15,16] has been shown to reveal an interesting subspace structure [12,22,23]. In the following $\mathbf{v}$ will denote the projection of a scan $\mathbf{x}$ onto an efficient basis that spans signal space; for more details see [18].

2.2 Model Flexibility and Bias

Having reduced the extremely ill-posed dataset to a marginally ill-posed one where the dimension of each observation, $\mathbf{v}$, equals the number of observations, it is now part of the modeling task to impose further constraints in order to avoid over-fitting. Different model families approach this in various ways, by limiting model flexibility and thus the effective dimensionality of the parameterization to match the available degrees of freedom.

In the following we focus on models for classification. Assuming the macroscopic variables to be univariate labels we seek to build models that optimally classify the microscopic variables[2], $\mathbf{x}$, into the correct classes. In other words, we seek a *decision boundary* in signal space that allows the observations to be correctly classified according to their macroscopic labels. More specifically we will apply two model families that differ in model flexibility:

- **Fishers Linear Discriminant**
  Fishers Linear Discriminant (FLD) is a family of linear classifier that are based on a cost function that measures the difference between class means relative to the within class variance [4,14]. The term linear refers to the fact that the models are linear in the parameters which makes parameter estimation straight forward. However, this relatively high *bias* limits the flexibility of the relationships (decision boundaries) that the models can implement.

- **Artificial neural network (ANN) classifiers**
  Artificial neural networks is a family of parameter efficient models that deal with the curse of dimensionality by employing nonlinearities [2,9]. The models are nonlinear in the parameters in contrast to FLD. This complicates parameter estimation but makes the models less biased and allow them to implement a much more flexible and wider range of relationships (decision boundaries) [10,24].

3 Generalization

Although cost functions allow us to quantify model performance the training error in (3) is the average over the *specific* and *limited* training set only. If the distribution of observations in this set, $\mathrm{P}_{train}(\mathbf{x},\mathbf{g})$, does not match the true distribution, $\mathrm{P}(\mathbf{x},\mathbf{g})$, sufficiently well the cost function value will not reflect model performance correctly. Rather, as training sets are often small we should use *generalization error*,

$$\mathrm{G}(\theta_{train}) = \iint \mathrm{c}(\mathbf{x},\mathbf{g},\theta_{train})\mathrm{P}(\mathbf{x},\mathbf{g})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{g} \quad . \tag{5}$$

as our measure of model quality. Unfortunately this requires complete knowledge of $\mathrm{P}(\mathbf{x},\mathbf{g})$ which, of course, we do not have. Instead we can estimate generalization either analytically [1,20] or empirically [24]. The latter is often called *test error*

$$\widehat{\mathrm{G}}(\theta_{train}) = \mathrm{C}(\mathcal{D}_{test},\theta_{train}) \tag{6}$$

$$= \frac{1}{N_{test}}\sum_{j=1}^{N_{test}} \mathrm{c}(\mathbf{x}_j,\mathbf{g}_j,\theta_{train}), \quad (\mathbf{x}_j,\mathbf{g}_j) \in \mathcal{D}_{test} \tag{7}$$

and evaluated using an independent set of observations organized in a *test set*

$$\mathcal{D}_{test} = \{(\mathbf{x}_j,\mathbf{g}_j) \mid j = 1,\ldots,N_{test}\} \quad . \tag{8}$$

[2] In practice we use $\mathbf{v}$ of course, thus efficiently representing the scans using a basis that spans signal space.

In (5) we have indicated how generalization error depends on the training set via the estimated parameters θ_{train}. To eliminate this dependency we average over training sets of size N_{train} to yield the *expected generalization error*

$$\mathrm{E}_{N_{train}}(\mathrm{G}) = \int \mathrm{G}(\theta_{train})\mathrm{P}(\mathcal{D}_{N_{train}})\, \mathrm{d}\mathcal{D}_{N_{train}} \quad , \tag{9}$$

which can be estimated empirically by using the test error in (7) to estimate $\mathrm{G}(\theta_{train})$. Clearly, using a set of the available observations to independently estimate generalization reduces the number of observations left for training. The optimal split of the available data into training- and test sets constitutes a non-trivial problem that has been studied in the context of ANN's and statistical re-sampling techniques [5]. In the remainder of this paper we will fix the size of the test set as well as the observations therein to allow measures of model performance that are unbiased—or at least comparable between different model families.

3.1 Learning Curves and Generalization Crossover

Using generalization we are now ready to investigate how the number of observations in the training set, N_{train}, affects model performance. We hypothesize that, as N_{train} increases, generalization error will decrease. This downwards slope of the so-called *learning curve* is caused by the improved estimates of $\mathrm{P}(\mathbf{x}, \mathbf{g})$ (on which the models are based) that increasingly larger training sets provide.

For a given model family the learning curve will eventually flatten out as additional observations no longer improve model performance due to limitations in the models themselves. This naturally leads to the further hypothesis that learning curves look different for different model families. Models that are very flexible typically need many examples to obtain stable parameter estimates. These models will in return generalize very well. In contrast, the implicit constraints in highly biased models enable them to obtain stable parameter estimates from fewer observations. However, they may not generalize as well as their more flexible counterparts. Thus, while generalization error is highest for very flexible models for small training sets, it decreases to a lower level than for highly biased, less flexible models as N_{train} increases. This means that a *generalization crossover* occurs at which point the data support the use of the more flexible models. The situation is illustrated in Fig. 2.

4 Methods

To estimate learning curves data from two functional activation studies, both involving simple motor tasks, was used.

4.1 [O^{15}]Water PET Scanning

A set of 30 subjects were each scanned 8 times using a Siemens-ECAT 953B PET scanner while alternately resting and performing a simple finger opposition

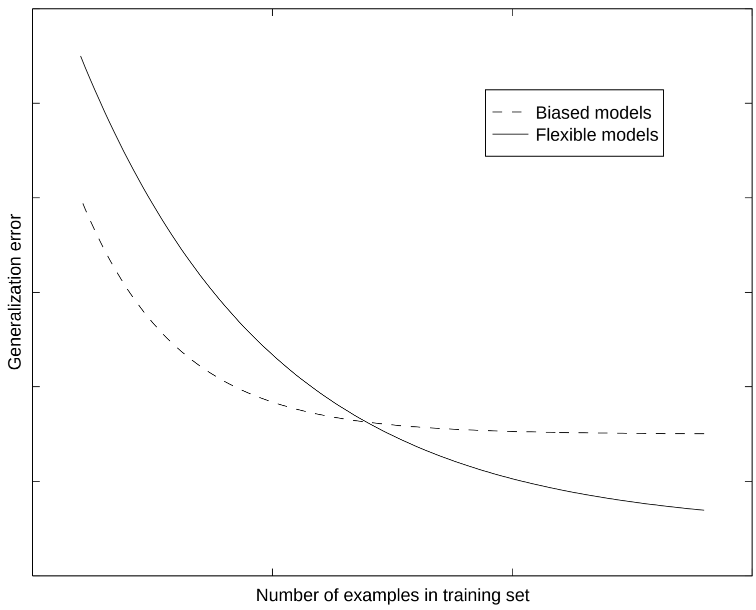

Fig. 2. Model generalization as a function of number of observations, N_{train}, used to estimate model parameters. Generalization error decreases with increasing N_{train} for both highly flexible and more biased models. The decrease is more rapid for the latter, whereas the former reaches a lower level for large values of N_{train}. At the point of generalization crossover enough data is available to support the use of more flexible, low-biased models.

task with their left hand [22]. For each subject four scans were acquired in each of the two states yielding a total of 240 scans.

Before scanning $[O^{15}]$water was automatically injected in the subjects right arm leaving the left arm free to perform the task. With the eyes covered by a patch an auditory timing signal was delivered by insert earphones.

For baseline (rest) scans, subjects were instructed to lie still and remain awake; they received no stimulation. For motor activation scans, the subjects left arm was positioned perpendicular to the scanning couch. At the start of the injection, the timing signal was initiated and the finger-thumb opposition task continued for 60 s. The finger-thumb opposition task consisted of sequential opposition of the thumb and successive digits, and back again $(2, 3, 4, 5, 4, 3, 2, 3, 4, \ldots)$ at a rate of 1 Hz.

PET scanning commenced when the radioactive material reached the brain, typically 10–20 s after injection, and data acquisition continued for 90 s. Each scanning session consisted of eight 90 s PET scans separated by 10 min rest periods to allow for O^{15} decay, for a total experimental time of approximately 90 min. The first, third, fifth, and seventh scans were acquired in the baseline state, and the second, fourth, sixth, and eighth scans were acquired in the activated state. Scans corrected for randoms, dead-time, and attenuation, but not for scatter, were reconstructed using 3D filtered back-projection.

4.2 fMRI Scanning

A single subject performing a left-handed finger-to-thumb opposition task was scanned during eight 180 s runs. In each run 24 baseline, 24 activation, and 24 baseline whole brain echo planar scans were acquired (2.5s/scan) with an interslice distance of 8 mm and an in plane voxel resolution of 3.1×3.1 mm^2. This yielded a total of 576 scans. During activation the task was timed with an auditory signal at a rate of 1 Hz.

4.3 Scan Alignment and Preprocessing

The PET and fMRI scans were intra-subject aligned using AIR (Automated Image Registration) [27] and only the PET scans were then stereo-tactically normalized to a simulated PET reference volume in Talairach space [25] using the 12 parameter linear transformation described in [28] (see [22] for more details). This yielded scans with 48 slices, inter-slice distance of 3.4 mm and in plane voxel resolution of 3.1×3.1 mm^2. After masking out voxels outside the brain an SVD basis was computed based on the entire[3] set of scans.

4.4 Modeling

After normalizing the singular vectors, $\mathbf{v}$, to zero mean and a standard deviation of one, a fixed test set was randomly selected (100 for the PET data and 200 for the fMRI data). The remaining observations were utilized to yield training sets of increasing size. A number of training sets of each size (25 for the PET data and 20 for the fMRI data) were randomly sampled with replacement[4] from the singular vectors. For each of the resulting training sets a linear (FLD) and a nonlinear (ANN) classifier were estimated. Model performance was then assessed using the fixed test set. The linear and nonlinear classifiers are based on different cost functions, so to allow a quantitative comparison generalization was measured as the mean misclassification on the independent test set.

5 Results

Figure 3 depicts the learning curves for the linear and nonlinear classifiers on the PET data. The two curves are slightly offset horizontally to better show the

[3] Basing models on an SVD of the entire set of observation limits results from generalization measures to the specific set of subjects in the PET case, and the specific subject in the fMRI case. Thus, generalization error does not implicate the extent to which models generalize to subjects other than those included in the datasets.

[4] Estimators based on sampling with replacement (also known as bootstrapping), where the same observation may appear more than once in the same sample, are asymptotically central [5]—however counter-intuitive this may seem.

error-bars that indicate one standard deviation of the mean for each training set size. As hypothesized both learning curves decrease. The nonlinear classifier

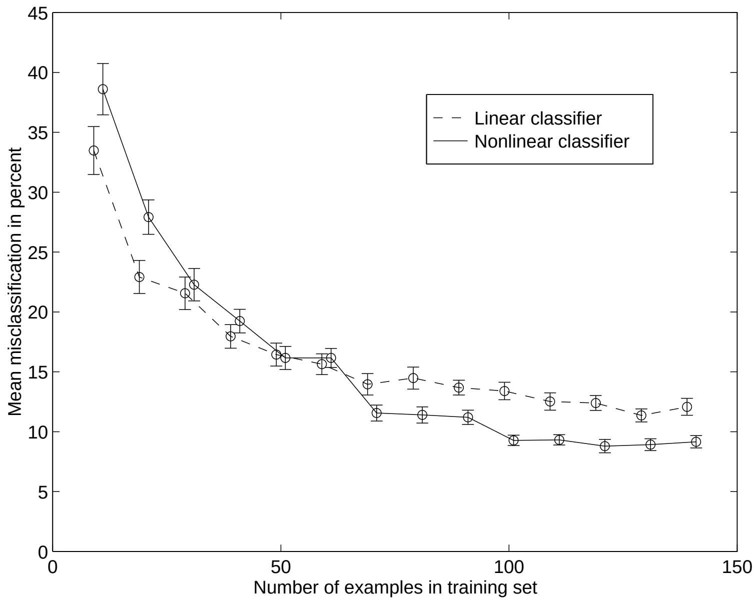

Fig. 3. For an [O^{15}]water PET study of a simple finger opposition task model generalization (measured as the mean misclassification on an independent test set) is plotted as a function of number of observations, N_{train}, used to estimate model parameters. Generalization error decreases with increasing N_{train} for the linear as well as the nonlinear classifiers. However, generalization error decreases more rapidly and settles at a higher level for the linear classifier than for its nonlinear counterpart. Thus, for this task linear classifiers seem optimal for small datasets. As more observations become available we are better off using the more flexible nonlinear classifiers.

seems to generalize worse for small training sets but perform relatively better as N_{train} increases. Indeed, a generalization crossover occurs for training sets with around 60 examples, and as N_{train} increases further generalization error for the nonlinear classifier settles at a lower level than that of its linear counterpart.

For the fMRI dataset Fig. 4 shows a similar picture. Again the learning curves for the linear and nonlinear classifiers cross as the number of observations in the training set is increased. Thus, for small training sets we can not reject the linear model.

6 Discussion

We have introduced a general framework for the analysis of functional activation datasets. In this framework the extremely ill-posed nature of typical datasets

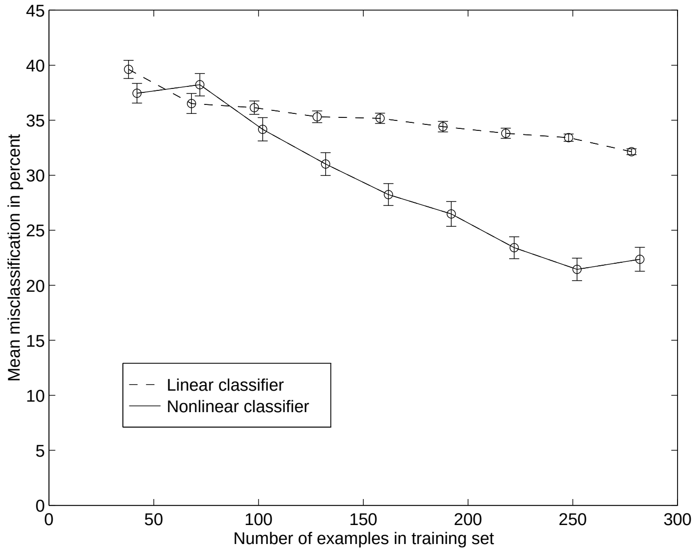

Fig. 4. For an fMRI study of a left-handed finger-to-thumb opposition task model generalization (measured as the mean misclassification on an independent test set) is plotted as a function of number of observations, N_{train}, used to estimate model parameters. Generalization error decreases with increasing N_{train} for the linear as well as the nonlinear classifiers. However, generalization error decreases more rapidly and settles at a higher level for the linear classifier than for its nonlinear counterpart. Again, the linear classifiers can not be rejected for small datasets. As more observations become available we are better off using the more flexible nonlinear classifiers.

imposes an immense computational burden on any modeling procedures. We have shown how a simple coordinate transform reduces data representation without loss of information, thus minimizing the computational load.

The importance of not measuring model performance on the same set of data used to estimate the model parameters has been stressed, and we have sketched how independent test sets provide empirical estimates of generalization. We have hypothesized how generalization error decreases as more observations become available for parameter estimation. Decreasing learning curves satisfying our hypothesis have been demonstrated on two functional activation datasets of PET and fMRI scans of subjects performing simple motor tasks.

By employing model families that differ in flexibility we have further shown the effect of model flexibility on the slope of the learning curves. For the studied tasks we have identified generalization crossovers, at which point enough observations are available to support the use of a more flexible, nonlinear model. We believe this to have implications for the future of modeling in functional neuroimaging; as more and more data become available the support for more sophisticated and flexible models increase. While introducing problems of their own (by e.g. not being linear in their parameters), these models can potentially lead to increased knowledge of the systems that govern information processing in the living, human brain.

7 Acknowledgments

This work has been funded in part by the Human Brain Project grant P20 MH57180, the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center, CONNECT, the Danish Research Council for Medical Science, and the Danish Research Academy.

References

1. H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.
2. C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.
3. J. S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in Neural Information Processing Systems*, 2:211–217, 1990.
4. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.
5. B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability. Chapman & Hall, 1993.
6. J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Journal of Knowledge Discovery and Data Mining*, 1996. In press.
7. K. J. Friston, J.-P. Poline, A. P. Holmes, C. D. Frith, and R. S. J. Frackowiak. A multivariate analysis of PET activation studies. *Human Brain Mapping*, 4:140–151, 1996.
8. B. Hassibi and D. G. Stork. Optimal brain surgeon. *Advances in Neural Information Processing Systems*, 5:164–174, 1992.
9. J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley, 1994.
10. M. Hintz-Madsen, M. W. Pederson, L. K. Hansen, and J. Larsen. Design and evaluation of neural skin classifiers. In Y. Tohkura, S. Katagiri, and E. Wilson, editors, *Proceedings of 1996 IEEE Workshop on Neural Networks for Signal Processing*, pages 223–230, 1996.
11. J. E. Jackson. *A User's Guide to Principal Components.* Wiley Series on Probability and Statistics, John Wiley and Sons, 1991.
12. B. Lautrup, L. K. Hansen, I. Law, N. Mørch, C. Svarer, and S. C. Strother. Massive weight sharing: A cure for extremely ill-posed problems. In H. J. Hermann, D. E. Wolf, and E. P. Pöppel, editors, *Proceedings of Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks, HLRZ, KFA Jülich, Germany*, pages 137–148, 1994.
13. Le Cun, Y., J. S. Denker, and S. Solla. Optimal brain damage. *Advances in Neural Information Processing Systems*, 2:598–605, 1990.
14. K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, 1979.

15. J. R. Moeller and S. C. Strother. A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *Journal of Cerebral Blood Flow and Metabolism*, 11:A121–A135, 1991.
16. 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. *Journal of Cerebral Blood Flow and Metabolism*, 7:649–658, 1987.
17. J. Moody. Prediction risk and architecture selection for neural networks. In V. Cherkassky, J. H. F. H., and H. Wechsler, editors, *From Statistics to Neural Networks, Theory and Pattern Recognition Applications*, pages 147–165. Springer Verlag, 1992.
18. N. Mørch, L. K. Hansen, I. Law, S. C. Strother, C. Svarer, B. Lautrup, U. Kjems, N. Lange, and O. B. Paulson. Generalization and the bias-variance trade-off in models of functional activation. *IEEE Transactions on Medical Imaging*, 1996. Submitted.
19. N. Mørch, U. Kjems, L. K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother, and K. Rehm. Visualization of neural networks using saliency maps. In *Proceedings of 1995 IEEE International Conference on Neural Networks*, volume 4, pages 2085–2090, 1995.
20. N. Murata, S. Yoshizawa, and S.-I. Amari. Network information criterion—determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5:865–872, 1994.
21. M. I. Posner and M. E. Raichle. *Images of Mind.* W. H. Freeman, 1994.
22. 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 scaled subprofile model compared to intersubject averaging and statistical parametric mapping: I. "Functional connectivity" of the human motor system studied with [^{15}O]water PET. *Journal of Cerebral Blood Flow and Metabolism*, 15:738–753, 1995.
23. S. C. Strother, J. R. Anderson, K. A. Schaper, J. J. Sidtis, and D. A. Rottenberg. Linear models of orthogonal subspaces & networks from functional activation PET studies of the human brain. In Y. Bizais, C. Barillot, and R. D. Paola, editors, *Proceedings of the 14th International Conference on Information Processing in Medical Imaging*, pages 299–310. Kluwer Academic Publishers, 1995.
24. C. Svarer, L. K. Hansen, and J. Larsen. On design and evaluation of tapped-delay neural network architectures. In H. R. Berenji et al., editors, *Proceedings of 1993 IEEE International Conference on Neural Networks*, pages 45–51, 1993.
25. J. Talairach and P. Tournoux. *Co-planar stereotaxic atlas of the human brain.* Thieme Medical Publishers Inc., New York, 1988.
26. A. W. Toga and J. C. Mazziotta. *Brain Mapping.* Academic Press, 1996.
27. R. P. Woods, S. R. Cherry, and J. C. Mazziotta. A rapid automated algorithm for accurately aligning and reslicing positron emission tomography images. *Journal of Computer Assisted Tomography*, 16:620–633, 1992.
28. R. P. Woods, J. C. Mazziotta, and S. R. Cherry. Automated image registration. In K. Uemura et al., editors, *Quantification of Brain Function. Tracer Kinetics and Image Analysis in Brain PET*, pages 391–400. Elsevier Science Publishers B. V., 1993.