NeuroImage **9,** 534–544 (1999)
Article ID nimg.1998.0425, available online at http://www.idealibrary.com on IDEAL®

Generalizable Patterns in Neuroimaging: How Many Principal Components?

Lars Kai Hansen,*,[1] Jan Larsen,* Finn Årup Nielsen,* Stephen C. Strother,†,‖ Egill Rostrup,‡ Robert Savoy,§ Nicholas Lange,¶ John Sidtis,‖ Claus Svarer,** and Olaf B. Paulson**

*Department of Mathematical Modeling, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark; †PET Imaging Service, VA Medical Center, and Department of Radiology, and ‖Department of Neurology, University of Minnesota, Minneapolis, Minnesota 55455; ‡Danish Center for Magnetic Resonance, Hvidovre Hospital, DK-2650 Hvidovre, Denmark; §Department of Radiology, Massachusetts General Hospital, Charlestown, Massachusetts 02129; ¶McLean Hospital and Harvard Medical School, Belmont, Massachusetts 02178; and **Neurobiology Research Unit, Rigshospitalet DK-2100, Copenhagen, Denmark

Received June 25, 1998

Generalization can be defined quantitatively and can be used to assess the performance of principal component analysis (PCA). The generalizability of PCA depends on the number of principal components retained in the analysis. We provide analytic and test set estimates of generalization. We show how the generalization error can be used to select the number of principal components in two analyses of functional magnetic resonance imaging activation sets. © 1999 Academic Press

INTRODUCTION

Principal component analysis (PCA) and the closely related singular value decomposition technique are popular tools for analysis of image data sets and are actively investigated in functional neuroimaging (Moeller and Strother, 1991; Friston *et al.*, 1993, 1995; Lautrup *et al.*, 1995; Strother *et al.*, 1995, 1996; Bullmore *et al.*, 1996; Ardekani *et al.*, 1998; Worsley *et al.*, 1997). By PCA the image data set is decomposed in terms of orthogonal "eigenimages" that may lend themselves to direct interpretation. Principal components, the projections of the image data onto the eigenimages, describe uncorrelated event sequences in the image data set.

Furthermore, we can capture the most important variations in the image data set by keeping only a few of the high-variance principal components. By such *unsupervised learning* we discover hidden, linear relations among the original set of measured variables.

Conventionally learning problems are divided into supervised and unsupervised learning. Supervised learning concerns the identification of *functional* relationships between two or more variables as in, e.g., linear regression. The objective of PCA and other unsupervised learning schemes is to capture *statistical* relationships, i.e., the structure of the underlying data distribution. Like supervised learning, unsupervised learning proceeds from a finite sample of training data. This means that the learned components are stochastic variables depending on the particular (random) training set forcing us to address the issue of generalization: How robust are the learned components to fluctuation and noise in the training set, and how well will they fare in predicting aspects of future test data? Generalization is a key topic in the theory of supervised learning, and significant theoretical progress has been reported (see, e.g., Larsen and Hansen, 1997). Unsupervised learning has not enjoyed the same attention, although results for specific learning machines can be found. In Hansen and Larsen (1996) we defined generalization for a broad class of unsupervised learning machines and applied it to PCA and clustering by the K-means method. In particular we used generalization to select the optimal number of principal components in a small simulation example.

The objective of this paper is to expand on the implementation and application of generalization for PCA in functional neuroimaging. A brief account of these results was presented in Hansen *et al.* (1997).

In what follows we describe a general framework for parameter estimation on any given data set and the ensuing generalization errors associated with this parameterization. We then discuss the application of this framework to PCA and how the generalization error can be estimated analytically and empirically. By exam-

[1] To whom correspondence should be addressed. Fax: (+45) 4587 2599. E-mail: lkhansen@imm.dtu.dk.

1053-8119/99 $30.00

ining the generalization error as a function of the number of principal components retained in the model, we can identify the number of principal components that leads to the minimal generalization error. Finally, we apply the framework to the principal component analysis of fMRI data.

MATERIALS AND METHODS

Good generalization is obtained when the model capacity is well matched to sample size solving the so-called bias/variance dilemma (see, e.g., Hastie and Tibshirani, 1990; Geman *et al.*, 1992; Mørch *et al.*, 1997). If the model distribution is too biased it will not be able to capture the full complexity of the target distribution, while a highly flexible model will support many different solutions to the learning problem and is likely to focus on nongeneric details of the particular training set (overfitting).

Here we analyze unsupervised learning schemes that are parametrized smoothly and whose performance can be described in terms of a cost or error function. If a particular data vector is denoted x and the model is parametrized by the parameter vector θ, the associated cost function will be denoted by $\epsilon(x|\theta)$.

A training set is a finite sample $D = \{x_\alpha\}_{\alpha=1}^{N}$ of the stochastic image vector x. Let $p(x)$ be the "true" probability density of x, while the empirical probability density associated with D is given by

$$p_e(x) = \frac{1}{N}\sum_{\alpha=1}^{N} \delta(x - x_\alpha). \tag{1}$$

For a specific model and a specific set of parameters θ we define the training and generalization errors as

$$E(\theta) = \int dx\, p_e(x)\, \epsilon(x|\theta) = \frac{1}{N}\sum_{\alpha=1}^{N} \epsilon(x_\alpha|\theta) \tag{2}$$

$$G(\theta) = \int dx\, p(x)\, \epsilon(x|\theta). \tag{3}$$

Note that the generalization error is nonobservable, i.e., it has to be either estimated from a finite *test set* also drawn from $p(x)$ or estimated from the training set using statistical arguments. In Hansen and Larsen (1996) we show that for large training sets the generalization error for maximum likelihood-based unsupervised learning can be estimated from the training error by adding a complexity term proportional to the number of fitted parameters denoted by $\dim(\theta)$:

$$\hat{G} = E + \frac{\dim(\theta)}{N}. \tag{4}$$

Empirical generalization estimates are obtained by dividing the data set into separate sets for training and testing, possibly combined with resampling (see Stone, 1974; Toussaint, 1974; Hansen and Salamon, 1990; Larsen and Hansen, 1995). Conventionally resampling schemes are classified as cross-validation (Stone, 1974) or bootstrap (Efron, 1983; Efron and Tibshirani, 1993) although many hybrid schemes exist. In cross-validation training and test sets are sampled without replacement while bootstrap is based on resampling with replacement. The simplest cross-validation scheme is *hold-out*, in which a given fraction of the data is left out for testing. V-fold cross-validation is defined by repeating the procedure V times with overlapping or nonoverlapping test sets. In both cases we obtain *unbiased* estimates of the average generalization error. This requires only that test and training sets are independent.

Principal Component Analysis

The objective of principal component analysis is to provide a simplified data description by projection of the data vector onto the eigendirections corresponding to the largest eigenvalues of the covariance matrix (Jackson, 1991). This scheme is well suited to high-dimensional, highly correlated data, as, e.g., found in exploratory analysis of functional neuroimages (Moeller and Strother, 1991; Friston *et al.*, 1993; Lautrup *et al.*, 1995; Strother *et al.*, 1995; Ardekani *et al.*, 1998; Worsley *et al.*, 1997). A number of neural network architectures are devised to estimate principal component subsets without first computing the covariance matrix (see, e.g., Oja, 1989; Hertz *et al.*, 1991; Diamantaras and Kung, 1996). Selecting the optimal number of principal components is a largely unsolved problem, although a number of statistical tests and heuristics have been proposed (Jackson, 1991). Here we suggest using the estimated generalization error to select the number of principal components in close analogy with the approach recommended for optimization of feed-forward artificial neural networks (Svarer *et al.*, 1993). See Akaike (1969), Ljung (1987), and Wahba (1990) for numerous applications of test error methods within system identification.

We follow Hansen and Larsen (1996) in defining PCA in terms of a cost function. In particular we assume that the data vector x (of dimension L, pixels or voxels) can be modeled as a Gaussian multivariate variable whose main variation is confined to a subspace of dimension K. The "signal" is degraded by additive, independent isotropic noise,

$$x = s + \nu. \tag{5}$$

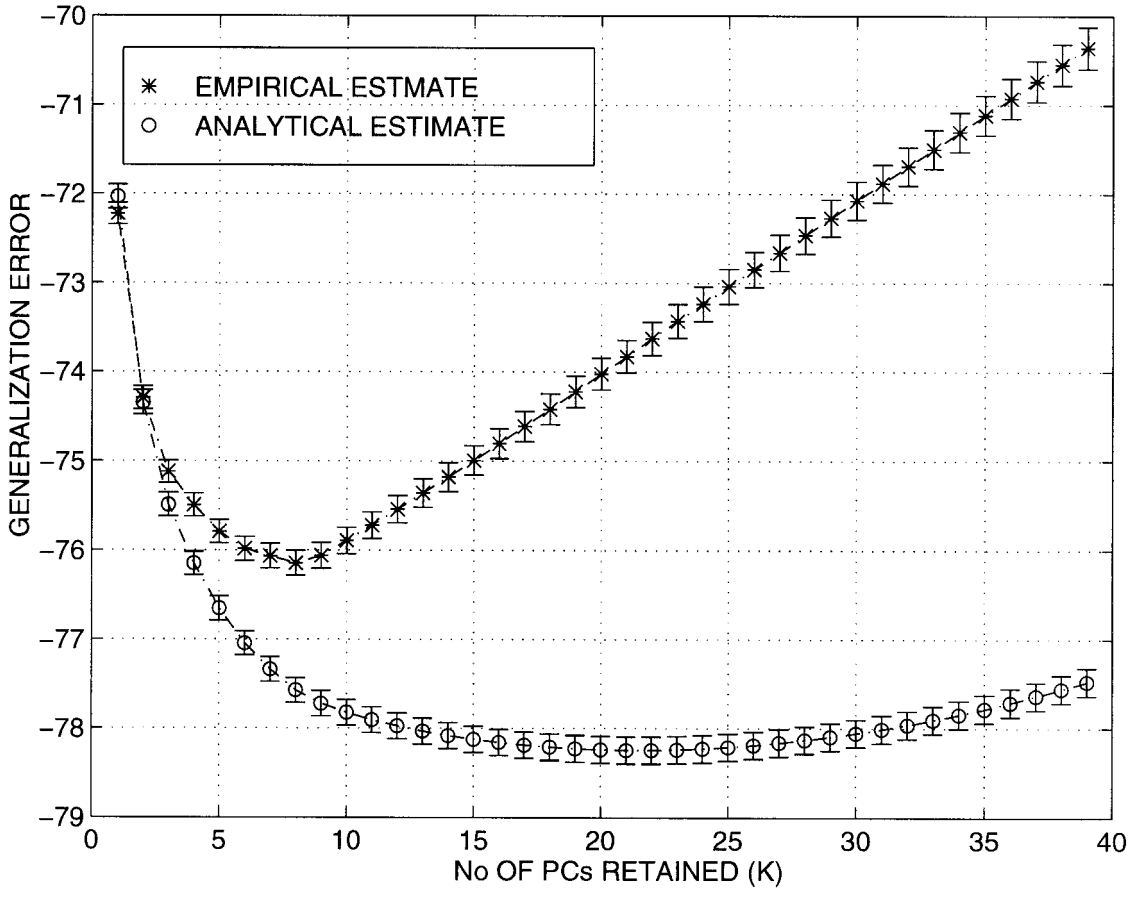

FIG. 1. Data set I. Bias/variance trade-off curves for PCA. The test set (*) generalization error estimate (mean ± the standard deviation of the mean, for 10 repetitions of experiment) and the asymptotic estimate. The empirical test error is an unbiased estimate, while the analytical estimate is *asymptotically* unbiased. The empirical estimate suggests an optimal PCA with $K = 8$ components. Note that the asymptotic estimate is too optimistic about the generalizability of the found PC patterns.

The signal and noise are assumed to be distributed as multivariate Gaussian random variables, $s \sim \mathcal{N}(x_0, \Sigma_s)$, $\nu \sim \mathcal{N}(0, \Sigma_\nu)$.

We assume that Σ_s is singular, i.e., of rank $K < L$, while $\Sigma_\nu = \sigma^2 I_L$, where I_L is a $L \times L$ identity matrix and σ^2 is a noise variance. This "PCA model" corresponds to certain tests proposed in the statistics literature for equality of covariance eigenvalues beyond a certain threshold (a so-called sphericity test) (Jackson, 1991).

Using well-known properties of Gaussian random variables we find

$$x \sim \mathcal{N}(x_0, \Sigma_s + \Sigma_\nu). \tag{6}$$

We use the negative log-likelihood as a cost function for the parameters $\theta \equiv (x_0, \Sigma_s, \Sigma_\nu)$,

$$\epsilon(x|\theta) = -\log p(x|\theta), \tag{7}$$

where $p(x|\theta)$ is the p.d.f. of the data given the parameter vector. Here,

$$p(x|x_0, \Sigma_s, \Sigma_\nu) = \frac{1}{\sqrt{|2\pi(\Sigma_s + \Sigma_\nu)|}} \cdot \exp\left(-\frac{1}{2}\Delta x^\top (\Sigma_s + \Sigma_\nu)^{-1} \Delta x\right), \tag{8}$$

with $\Delta x = x - x_0$.

Parameter Estimation

Unconstrained minimization of the negative log-likelihood leads to the well-known parameter estimates

$$\hat{x}_0 = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha, \quad \hat{\Sigma} = \frac{1}{N}\sum_{\alpha=1}^{N} (x_\alpha - \hat{x}_0)(x_\alpha - \hat{x}_0)^\top. \tag{9}$$

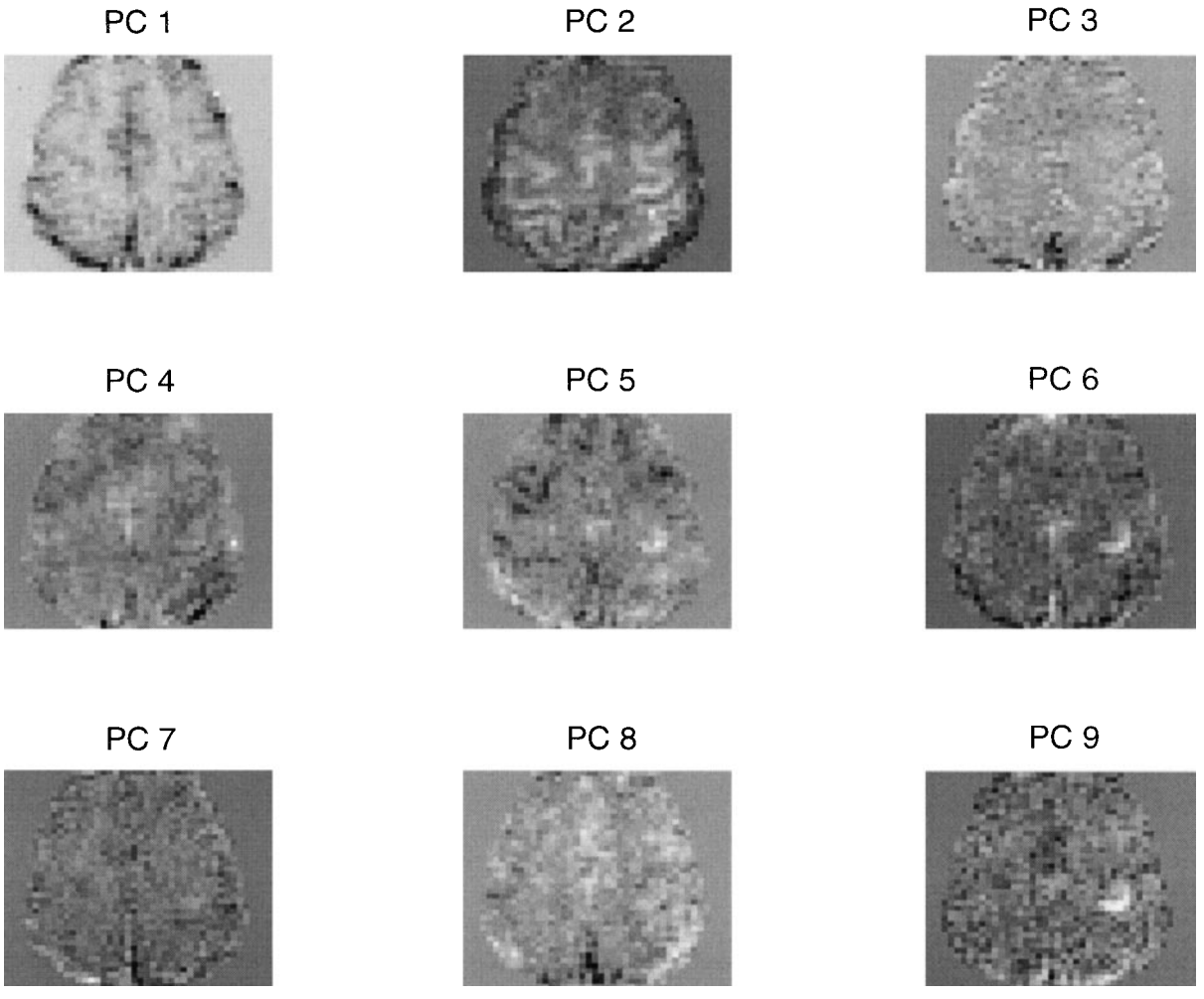

FIG. 2. Data set I. Eigenimages corresponding to the nine most significant principal components. The unbiased generalization estimate in Fig. 1 suggests that the optimal PCA contains eight components. Among the generalizable patterns we find in the sixth component, PC 6, focal activation in the right hemisphere corresponding to areas associated with the primary motor cortex. Also there is a trace of a focal central activation with a possible interpretation as the supplementary motor area.

Our model constraint involved in the approximation $\Sigma = \Sigma_s + \sigma^2 I_L$ is implemented as follows. Let $\hat{\Sigma} = S\Lambda S^\top$, where S is an orthogonal matrix of eigenvectors and $\Lambda = \text{diag}([\lambda_1, \ldots, \lambda_L])$ is the diagonal matrix of eigenvalues λ_i ranked in decreasing order. By fixing the dimensionality of the signal subspace, K, we identify the covariance matrix of the subspace spanned by the K largest PCs by

$$\hat{\Sigma}_K = S \cdot \text{diag}([\lambda_1, \ldots, \lambda_K, 0, \ldots, 0]) \cdot S^\top. \quad (10)$$

The noise variance is subsequently estimated so as to conserve the total variance (viz., the *trace* of the covariance matrix),

$$\hat{\sigma}^2 = \frac{1}{L - K} \text{Trace}\,[\hat{\Sigma} - \hat{\Sigma}_K], \quad (11)$$

hence

$$\hat{\Sigma}_\nu = \hat{\sigma}^2 I_L \quad (12)$$

and

$$\hat{\Sigma}_s = S \cdot \text{diag} \cdot ([\lambda_1 - \hat{\sigma}^2, \ldots, \lambda_K - \hat{\sigma}^2, 0, \ldots, 0]) \cdot S^\top. \quad (13)$$

This procedure is maximum likelihood under the constraints of the model.

Estimating the PCA Generalization Error

When the training set for an adaptive system becomes large relative to the number of fitted parameters

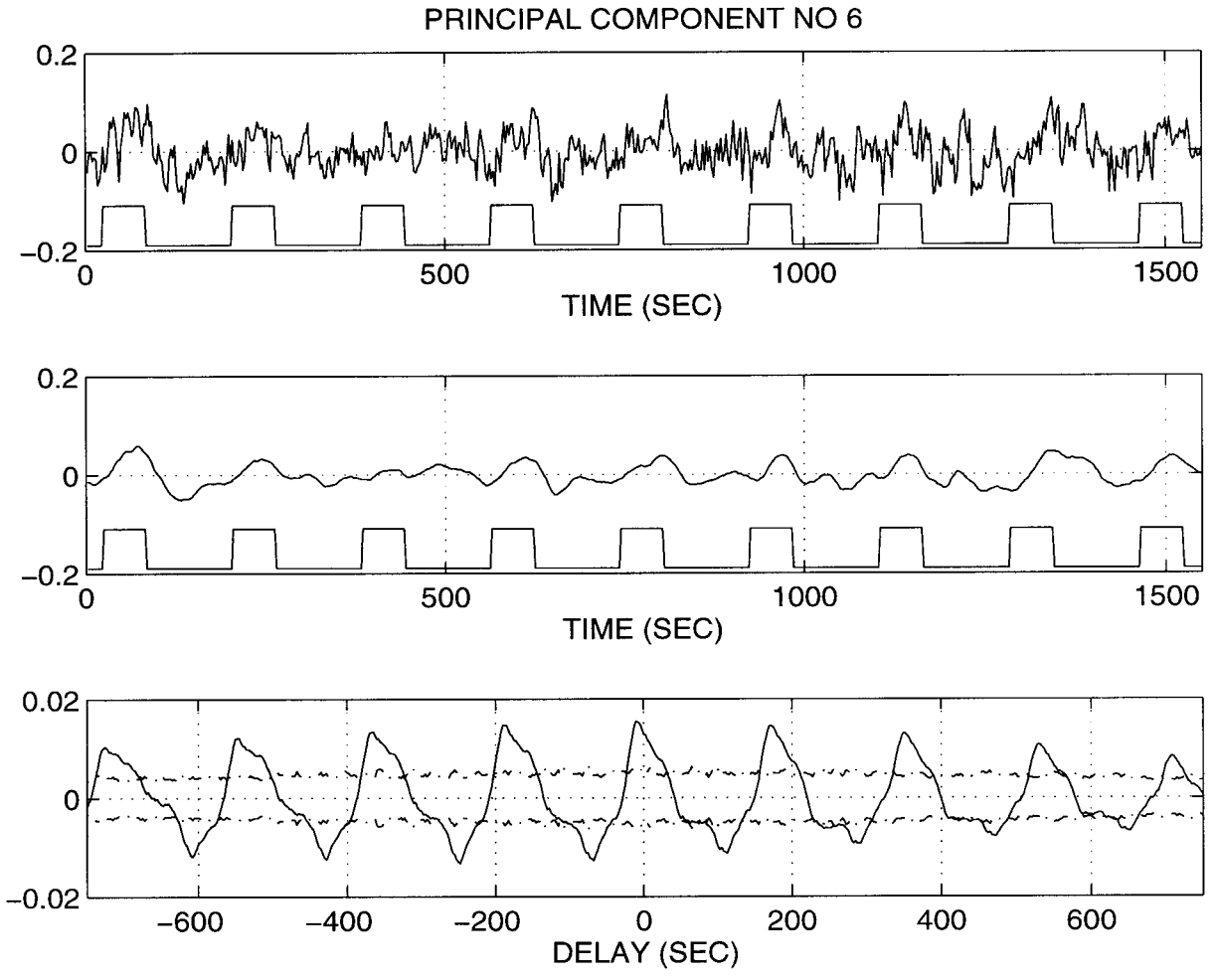

FIG. 3. Data set I. Top: The "raw" time course of principal component 6 (projection of the image sequence onto eigenimage 6 of the covariance matrix); the offset square wave indicates the time course of the binary finger opposition activation. The subject performed finger tapping at time intervals at which this function is high. Middle: Smoothed principal component with a somewhat more interpretable response to the activation. Note that both the delay and the form of the response vary from run to run. Bottom: The cross-correlation function of the reference sequence and the principal component (unsmoothed); the dot–dash horizontal curves indicate the symmetric $P = 0.001$ significance level for rejection of a white noise null hypothesis. The significance level has been estimated using a simple permutation test (Holmes *et al.*, 1996). $1/P = 1000$ random permutations of the principal component sequence were cross-correlated with the reference function and the symmetric extremal values used as thresholds for the given P value.

the fluctuations of these parameters decrease. The estimated parameters of systems adapted on different training sets will become more and more similar as the training set size increases. In fact we can show for *smoothly* parametrized algorithms that the distribution of these parameters—induced by the random selection of training sets—is asymptotically Gaussian with a covariance matrix proportional to $1/N$ (see, e.g., Ljung, 1987). This convergence of parameter estimates leads to a similar convergence of their generalization errors. Hence, we may use the average generalization error (for identical systems adapted on different samples of N) as an asymptotic estimate of the generalization error of a specific realization. Details of such analysis for PCA can be found in Hansen and Larsen (1996), where we derived the relation

$$\hat{G}(\hat{\theta}) \approx E(\hat{\theta}) + \frac{\dim(\hat{\theta})}{N}, \tag{14}$$

valid in the limit as $\dim(\theta/N) \to 0$. The dimensionality of the parametrization depends on the number, $K \in [1, L]$, of principal components retained in the PCA. As we estimate the (symmetric) signal covariance matrix, the L-dimensional vector x_0 and the noise variance σ^2, the total number of estimated parameters is $\dim(\theta) = L + 1 + K(2L - K + 1)/2$.

In real world examples facing *limited* data sets we

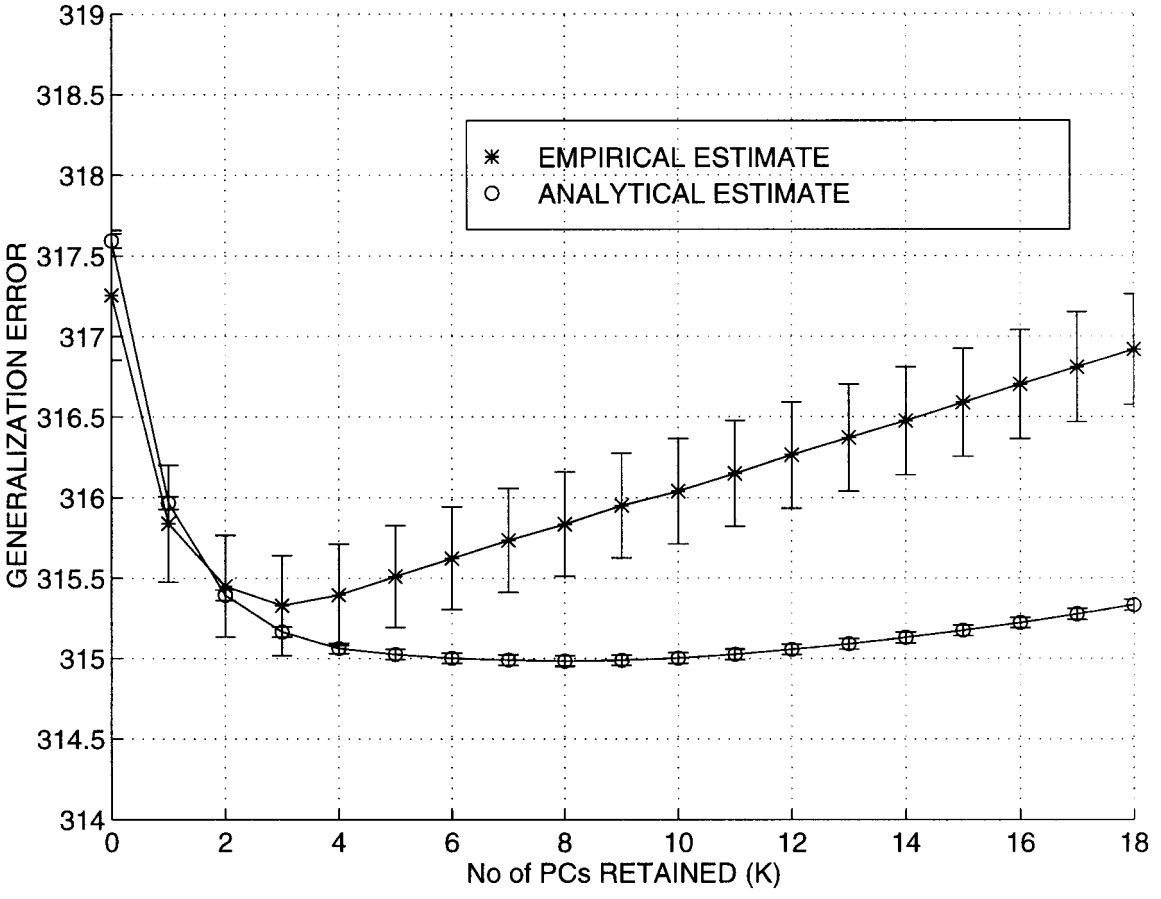

FIG. 4. Data set II. Bias/variance trade-off curves for PCA on visual stimulation sequence. The upper curve is the unbiased test set generalization error (mean of 10 repetitions of the experiment, the error bars indicate 1 standard deviation of the mean). The second curve is the corresponding analytical estimate. The unbiased estimate suggests an optimal PCA with $K = 3$ components. As in the analysis of data set I, we find that the analytical estimate is too optimistic and that it does not provide a reliable model selection scheme.

generally prefer to estimate the generalization error by means of resampling. For a particular *split* of the data set we can use the explicit form of the distribution to obtain expressions for the training and test errors in terms of the estimated parameters,

$$E = \frac{1}{2}\log|2\pi(\hat{\Sigma}_s + \hat{\Sigma}_n)| + \frac{1}{2}\,\mathrm{Trace}\,[(\hat{\Sigma}_s + \hat{\Sigma}_n)^{-1}\hat{\Sigma}_{\mathrm{train}}] \tag{15}$$

$$\hat{G}_{\mathrm{testset}} = \frac{1}{2}\log|2\pi(\hat{\Sigma}_s + \hat{\Sigma}_n)| + \frac{1}{2}\,\mathrm{Trace}\,[(\hat{\Sigma}_s + \hat{\Sigma}_n)^{-1}\hat{\Sigma}_{\mathrm{test}}], \tag{16}$$

where the covariance matrices, $\hat{\Sigma}_{\mathrm{train}}$ and $\hat{\Sigma}_{\mathrm{test}}$, are estimated on the two different sets respectively. In the typical case in functional neuroimaging, the estimated covariance matrix in (9) is rank deficient. Typically, $N \ll L$, hence the rank of the $L \times L$ matrix will be at most N. In this case the we can represent the covariance structure in the reduced spectral form

$$\hat{\Sigma}_{\mathrm{train}} = \sum_{n=1}^{N} \lambda_n s_n s_n^{\top}, \tag{17}$$

where s_n are the N columns of S corresponding to nonzero eigenvalues. In terms of this reduced representation we can write the estimates for the training error,

$$E(\hat{\theta}) = \frac{1}{2}\log 2\pi + \frac{1}{2}\sum_{n=1}^{K}\log\lambda_n + \frac{1}{2}(L-K)\log\hat{\sigma}^2 + \frac{1}{2N\hat{\sigma}^2}\left[\sum_{\alpha=1}^{N} x_\alpha^{\top}x_\alpha - \sum_{n=1}^{K}\frac{\lambda_n - \hat{\sigma}^2}{\lambda_n}\sum_{\alpha=1}^{N}(x_\alpha^{\top}s_n)^2\right], \tag{18}$$

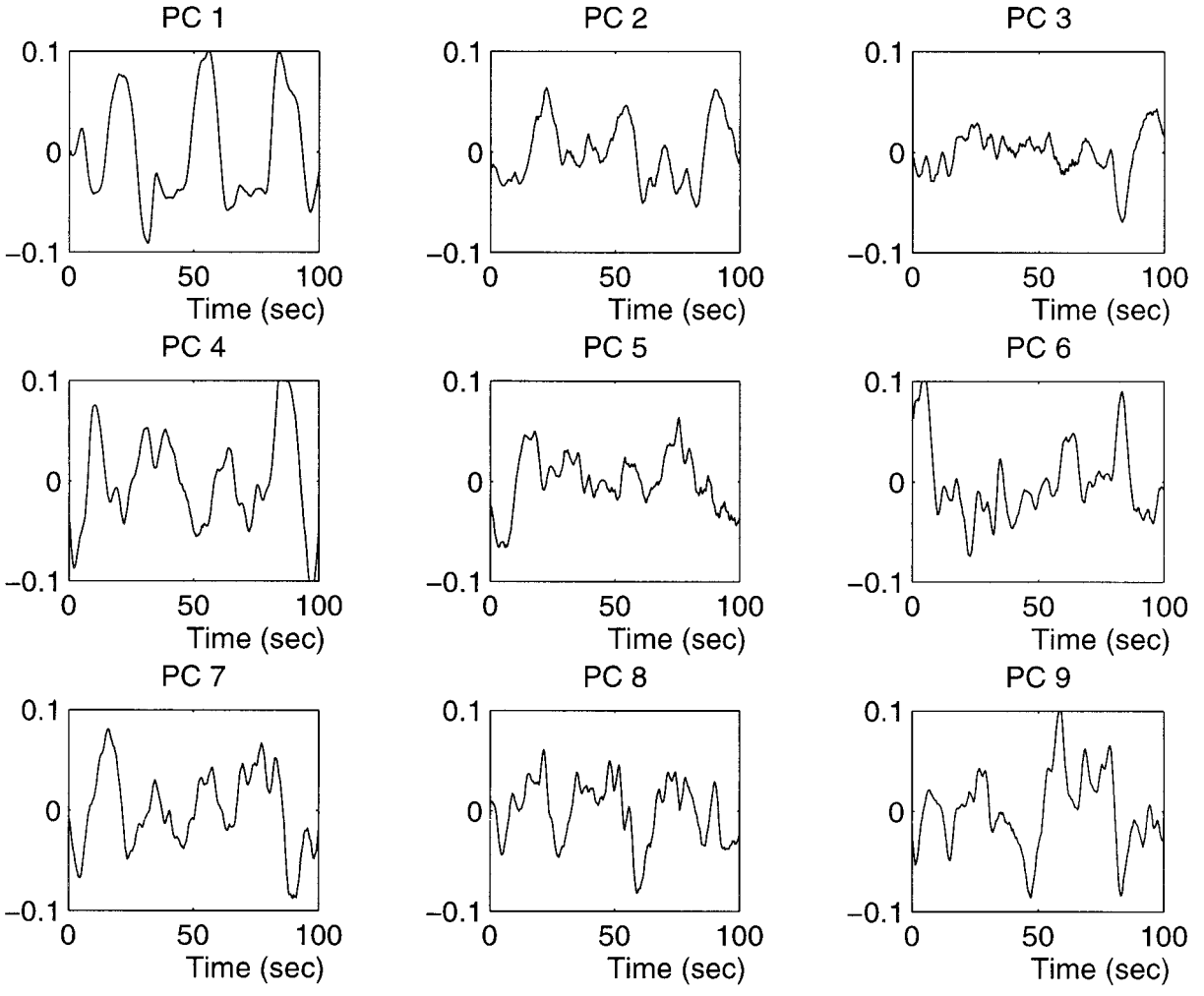

FIG. 5. Data set II. The principal components corresponding to the nine largest covariance eigenvalues for a sequence of 300 fMRI scans of a visually stimulated subject. Stimulation takes places at scan times τ = 8–25 s, τ = 42–59 s, and τ = 75–92 s relative to the start of this three-run sequence. Scan time sampling interval was TR = 0.33 s. The sequences have been smoothed for presentation, reducing noise and high-frequency physiological components. Note that most of the response is captured by the first principal component, showing a strong response to all three periods of stimulation. Using the generalization error estimates in Fig. 4, we find that only the time sequences corresponding to the first three components generalize.

and similarly the estimate of the generalization error for a test set of N_{test} data vectors will be

$$\hat{G}(\hat{\theta}) = \frac{1}{2}\log 2\pi + \frac{1}{2}\sum_{n=1}^{K}\log\lambda_n + \frac{1}{2}(L-K)\log\hat{\sigma}^2 + \frac{1}{2N_{\text{test}}\hat{\sigma}^2} \cdot \left[\sum_{\beta=1}^{N_{\text{test}}} x_\beta^\top x_\beta - \sum_{n=1}^{K}\frac{\lambda_n - \hat{\sigma}^2}{\lambda_n}\sum_{\beta=1}^{N_{\text{test}}}(x_\beta^\top s_n)^2\right]. \quad (19)$$

The nonzero eigenvalues and their eigenvectors can, e.g., be found by *singular value decomposition* of the data matrix $X \equiv [x_\alpha]$. In Eqs. (18) and (19) we have assumed zero mean signals, $x_0 = 0$, for simplicity.

Here we assume that the principal components are *importance* ranked according to their variance contribution, i.e., leading to a simple sequential optimization of K. For each value of K we estimate the generalization error and commend the value that provides the minimal test error. It is interesting to consider more general search strategies for principal component subset selection; however, an exhaustive combinatorial search over the 2^N (there are only N nonzero covariance eigenvalues for $N < L$) possible subsets is out of the question for most neuroimaging problems.

RESULTS AND DISCUSSION

Example I: Motor Study

An fMRI activation image set of a single subject performing a left-handed finger-to-thumb opposition

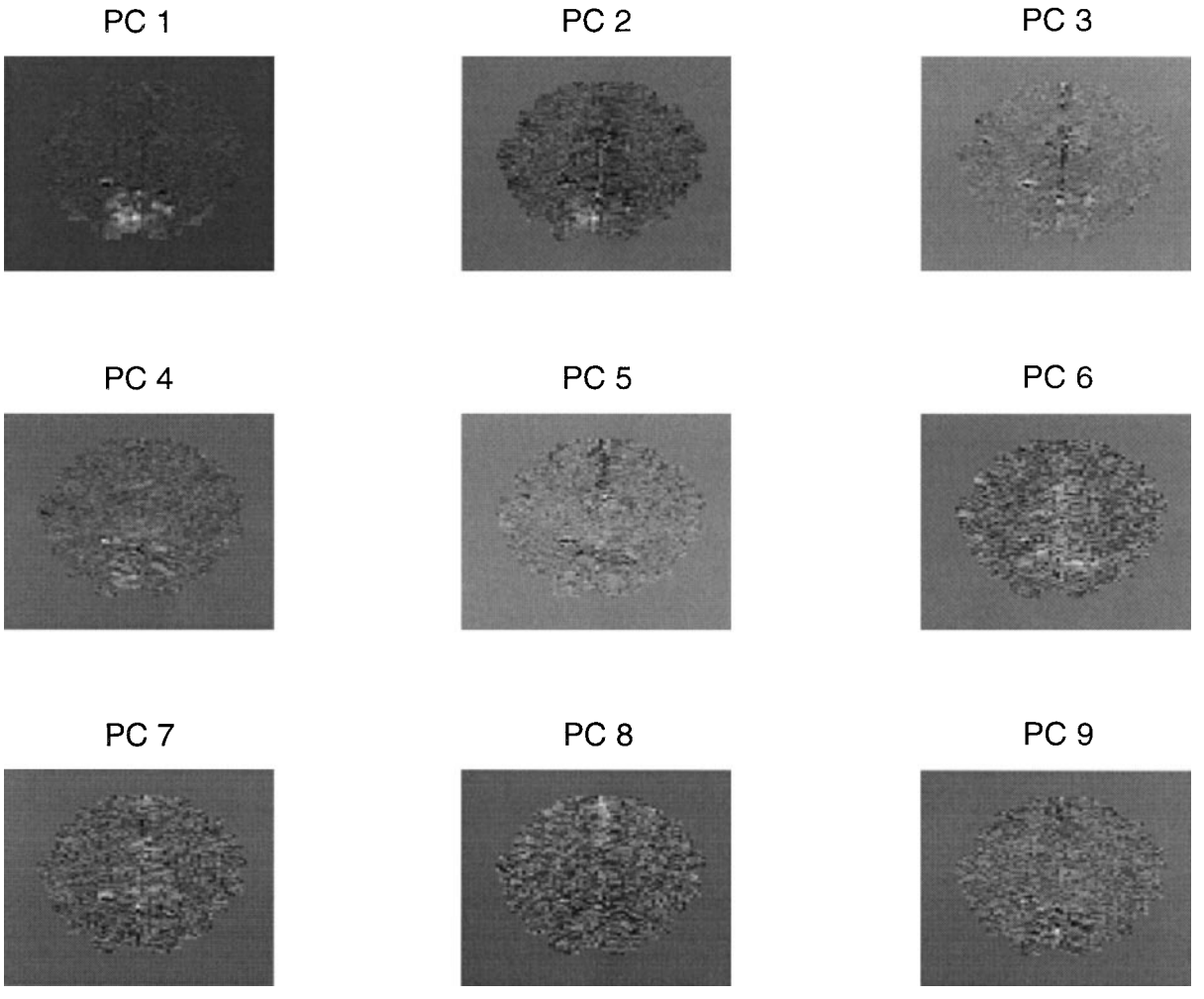

FIG. 6. Data set II. Covariance eigenimages corresponding to the nine most significant principal components. The eigenimage corresponding to the dominating first PC is focused in visual areas. Using the bias/variance trade-off curves in Fig. 4, we find that only the eigenimages corresponding to the first three components generalize.

task was acquired. Multiple runs of 72 2.5-s (24 baseline, 24 activation, 24 baseline) whole-brain echo planar scans were aligned, and an axial slice through primary motor cortex and the supplementary motor area of 42 × 42 voxels (3.1 × 3.1 × 8 mm) was extracted. Of a total of 624 scans, training sets of size $N = 300$ were drawn at random and for each training set the remaining independent set of 324 scans was used as test set. PCA analyses were carried out on the training set. We used Eq. (19) to compute the average negative log-likelihood on the test set using the covariance structure estimated on the training sets and Eqs. (4) and (18) to compute the analytical estimate of the generalization error. Both estimates were then plotted versus size of the PCA subspace; see Fig. 1.

Inspection of the test set curve suggests a model comprising eight principal components. We note that the analytical estimate is too optimistic, presumably because the sample is not large enough for the asymptotic results to hold. It underestimates the level of the generalization error and points to an optimal model with more than 20 components which—as measured by the unbiased estimate—has a generalization error as bad as a model with one or two components.

The covariance eigenimages are shown in Fig. 2. Eigenimages corresponding to components 1–5 are dominated by signal sources that are highly localized spatially (hot spots comprising one to four neighbor pixels). It is compelling to defer these as confounding *vascular* signal sources. Component 6, however, has a somewhat more extended hot spot in the contralateral (right hemisphere) motor area. In Fig. 3 we provide a more detailed temporal analysis of this signal source. At the top the "raw" principal component sequence is aligned with the binary reference function encoding the activation state (high, finger opposition; low, rest). Below, in the center, we give a low-pass filtered version of the principal component sequence. The smoothed

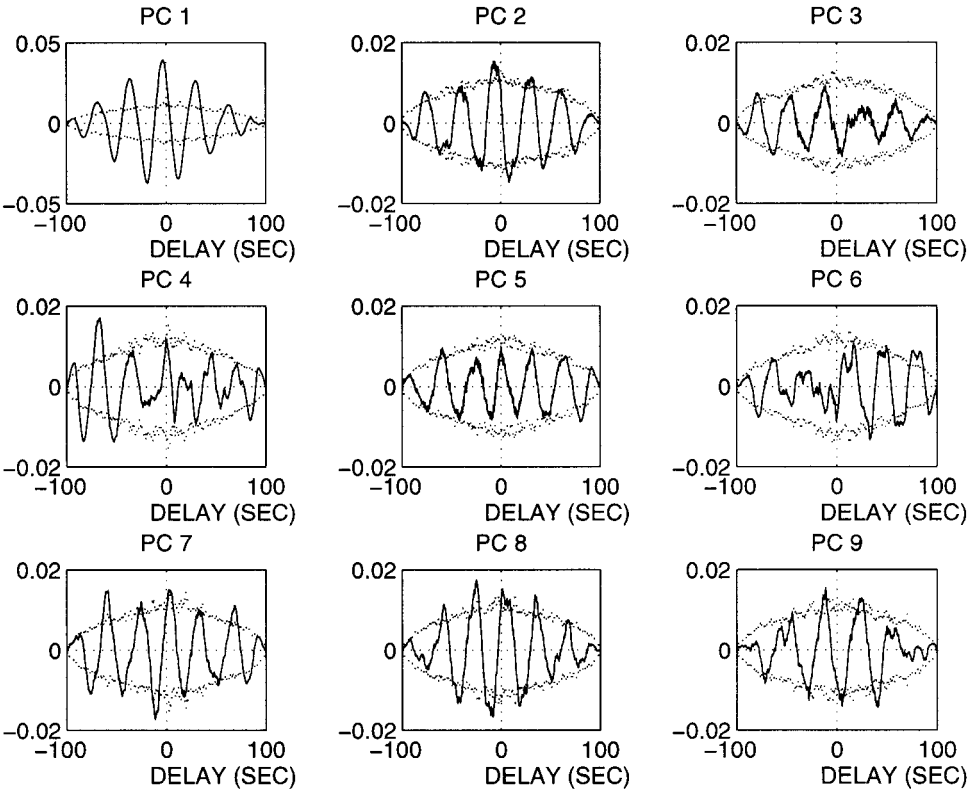

FIG. 7. Data set II. Cross-correlation analyses based on a reference time function taking the value −1 at times corresponding to rest scans and the value +1 for scans taken during stimulation. Each panel shows the cross-correlation of the principal component and the reference function for times $\tau \in [-100\text{ s}, 100\text{s}]$. The dotted horizontal curves are $P = 0.001$ levels in a nonparametric permutation test for significant cross-correlation. These curves are computed in a manner similar to Holmes *et al.* (1996): $1/P = 1000$ random permutations of the principal component sequence were cross-correlated with the reference function and the (two-sided) extremal values used as thresholds for the given P value. The test has not been corrected for simultaneous test of multiple hypotheses (Bonferroni). Of the three generalizable patterns (c.f., Fig. 4) only component 1 shows significant correlation with the activation reference function. The first component is maximally correlated with a reference function delayed by $\tau \approx 10$ scans corresponding to 3.3 s.

component shows a definite response to the activation, though with some randomness in the actual delay and shape of the response. At the bottom we plot the cross-correlation function between the on/off reference function and the unsmoothed signal. The cross-correlation function shows the characteristic periodic sawtooth shape resulting from correlation of a square-wave signal with a delayed square wave. The horizontal dash–dot curves are the symmetric $P = 0.001$ intervals for significant rejection of a white noise null hypothesis. These significance curves were computed as the extremal values after cross-correlating 1000 time-index permutations of the reference function with the actual principal component sequence. The significance level has not been corrected for multiple hypotheses (Bonferroni). Such a correction depends in a nontrivial way on the detailed specification of the null and is not relevant to the present exploratory analysis.

Example II: Visual Stimulation

A single slice holding 128 × 128 pixels was acquired with a time interval between successive scans of TR = 333 ms. Visual stimulation in the form of a flashing annular checkerboard pattern was interleaved with periods of fixation. A run consisting of 25 scans of rest, 50 scans of stimulation, and 25 scans of rest was repeated 10 times. For this analysis a contiguous mask was created with 2440 pixels comprising the essential parts of the slice including the visual cortex. Principal component analyses were performed on a subset of three runs ($N = 300$, runs 4–6) with increasing dimensionality of the signal subspace. Since the time interval between scans was much shorter than in the previous analysis temporal correlations were expected on the hemodynamic time scale (5–10 s). Hence, we used a block-resampling scheme: the generalization error is computed on a randomly selected "hold-out" contiguous time interval of 50 scans (~16.7 s). The procedure was repeated 10 times with different generalization intervals. In Fig. 4 we show the estimated generalization errors as function of subspace dimension. The analysis suggests an optimal model with a three-dimensional signal subspace. In line with our observation for data

set I the analytic estimate is too optimistic about the generalizability of the high-dimensional models.

The first nine principal components are shown in Fig. 5; all curves have been low-pass filtered to reduce measurement noise and physiological signals. The first component picks up a pronounced activation signal. In Fig. 6 we show the corresponding covariance eigenimages. The first component is dominated by an extended focal activity in primary visual cortex (V1). The third component, also included in the optimal model, shows an interesting *temporal* localization, suggesting that this mode is a kind of "generalizable correction" to the primary response in the first component, this correction being active mainly in the final third run. Spatially, the third component is also quite localized, picking up signals in three spots anterior to the primary visual areas.

In Fig. 7 we have performed cross-correlation analyses of all of the nine most variant principal component sequences. The horizontal dash–dot curves indicate $P = 0.001$ significance intervals as above. While the first component stands out clearly with respect to this significance level, the third component does not seem significantly cross-correlated with the reference function in line with the remarks above. This component represents a minor correction to the primary response in the first component active mainly during the third run of the experiment.

CONCLUSION

We have presented an approach for optimization of principal component analyses on image data with respect to generalization. Our approach is based on estimating the predictive power of the model distribution of the our PCA model. The distribution is a constrained Gaussian compatible with the generally accepted interpretation of PCA, namely that we can use PCA to identify a low-dimensional salient signal subspace. The model assumes a Gaussian signal and Gaussian noise appropriate for an exploratory analysis based on covariance. We proposed two estimates of generalization. The first is based on resampling and provides an unbiased estimate, while the second is an analytical estimate which is *asymptotically* unbiased.

The usefulness of the approach was demonstrated on two functional magnetic resonance data sets. In both cases we found that the model with the best generalization ability picked up signals that were strongly correlated with the activation reference sequence. In both cases we also found that the analytical generalization estimate was too optimistic about the level of generalization. Furthermore, the optimal model suggested by this method was severely overparametrized.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for constructive comments that improved the presentation of this paper. This project has been funded by The Danish Research Councils Interdisciplinary Neuroscience Project and the Human Brain Project P20 MH57180 "Spatial and Temporal Patterns in Functional Neuroimaging."

REFERENCES

Akaike, H. 1969. Fitting autoregressive models for prediction. *Ann. Inst. Stat. Mat.* **21:**243–247.

Ardekani, B., Strother, S. C., Anderson, J. R., Law, I., Paulson, O. B., Kanno, I., and Rottenberg, D. A. 1998. On detection of activation patterns using principal component analysis. In *Proceedings of BrainPET'97: Quantitative Functional Brain Imaging with Positron Emission Tomography* (R. E. Carson *et al.*, Eds.). Academic Press, San Diego, in press..

Bullmore, E. T., Rabe-Hasketh, S., Morris, R. G., Williams, S. C. R., Gregory, L., Gray, J. A., and Brammer, M. J. 1996. Functional magnetic resonance image analysis of a large-scale neurocognitive network. *NeuroImage* **4:**16–33.

Diamantaras, K. I., and Kung, S. Y. 1996. *Principal Component Neural Networks: Theory and Applications.* Wiley, New York.

Efron, B. 1983. The jackknife, bootstrap and other resampling plans. SIAM Monogr. 38. CBMS-NSF, Philadelphia.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J. 1993. Functional connectivity: The principal-component analysis of large (PET) data sets. *J. Cereb. Blood Flow Metab.* **13:**5–14.

Friston, K. J., Frith, C. D., Frackowiak, R. S. J., and Turner, R. 1995. Characterizing dynamic brain responses with fMRI: A multivariate approach. *NeuroImage* **2:**166–172.

Geman, S., Bienenstock, E., and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Comput.* **4:**1–58.

Hansen, L. K., and Salamon, P. 1990. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12:**993–1001.

Hansen, L. K., and Larsen, J. 1996. Unsupervised learning and generalization. In *Proceedings of the IEEE International Conference on Neural Networks 1996*, Vol. 1, pp. 25–30. Washington, DC.

Hansen, L. K., Nielsen, F. A. A., Toft, P., Strother, S. C., Lange, N., Morch, N., Svarer, C., Paulson, O. B., Savoy, R., Rosen, B., Rostrup, E., Born, P. 1997. How many principal components? Third International Conference on Functional Mapping of the Human Brain, Copenhagen, 1997. *NeuroImage* **5:**474.

Hastie, T., and Tibshirani, R. 1990. *Generalized Additive Models.* Chapman & Hall, London.

Hertz, J., Krogh, A., and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation.* Addison–Wesley, Redwood City, CA.

Holmes, A. P., Blair, R. C., Watson, J. D. G., and Ford, I. 1996. Non-parametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.* **16:**7–22.

Jackson, J. E. 1991. *A User's Guide to Principal Components.* Wiley, New York.

Larsen, J., and Hansen, L. K. 1995. Empirical generalization assessment of neural network models. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing V*, (F. Girosi, J. Makhoul, E. Manolakos, and E. Wilson, Eds.), pp. 30–39. Piscataway, NJ.

Larsen, J., and Hansen, L. K. 1997. Generalization: The hidden agenda of learning. In *The Past, Present, and Future of Neural*

Networks for Signal Processing (J.-N. Hwang, S. Y. Kung, M. Niranjan, and J. C. Principe, Eds.), IEEE Signal Processing Magazine, November, pp. 43–45.

Lautrup, B., Hansen, L. K., Law, I., Mørch, N., Svarer, C., and Strother, S. C. 1995. Massive weight sharing: A cure for extremely ill-posed problems. In *Supercomputing in Brain Research: From Tomography to Neural Networks* (H. J. Herman et al., Eds.), pp. 137–148. World Scientific, Singapore.

Ljung, L. 1987. *System Identification: Theory for the User.* Prentice Hall, Englewood Cliffs, NJ.

Moeller, J. R., and Strother, S. C. 1991. A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *J. Cereb. Blood Flow Metab.* **11:**A121–A135.

Mørch, N., Hansen, L. K., Strother, S. C., Svarer, C., Rottenberg, D. A., Lautrup, B., Savoy, R., and Paulson, O. B. 1997. Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In *Information Processing in Medical Imaging* (J. Duncan *et al.*, Eds.), *Lecture Notes in Computer Science*, Vol. 1230, pp. 259–270. Springer-Verlag, Berlin/New York.

Oja, E. 1989. Neural networks, principal components, and subspaces. *Int. J. Neural Syst.* **1:**61–68.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictors. *J. R. Stat. Soc. B* **36:**111–147.

Strother, S. C., Anderson, A. R., Schaper, K. A., Sidtis, J. J., Woods, R. P., and Rottenberg, D. A. 1995. Principal component analysis and the scaled subprofile model compared to intersubject averaging of a statistical parametric mapping. I. "Functional connectivity" of the human motor system studied with [^{15}O]PET. *J. Cereb. Blood Flow Metab.* **15:**738–775.

Strother, S. C., Lange, N., Savoy, R. L., Anderson, J. R., Sidtis, J. J., Hansen, L. K., Bandettini, P. A., O'Craven, K., Rezza, M., Rosen, B. R., and Rottenberg, D. A. 1996. Multidimensional state spaces for fMRI and PET activation studies. *Neuroimage* **2**(Pt 2):98.

Svarer, C., Hansen, L. K., and Larsen, J. 1993. On design and evaluation of tapped-delay neural network architectures. In *Proceedings of the 1993 IEEE International Conference on Neural Networks* (H. R. Berenji *et al.*, Eds., Vol. 1, pp. 46–51. IEEE Service Center, Piscataway, NJ.

Toussaint, G. T. 1974. Bibliography on estimation of misclassification. *IEEE Trans. Inf. Theor.* **20:**472–479.

Wahba, G. 1990. Spline models for observational data. In *CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59. SIAM, Philadelphia.

Worsley, K. J., Poline, J.-B., Friston, K. J., and Evans, A. C. 1997. Characterizing the response of PET and fMRI data using multivariate linear models (MLM). *NeuroImage* **6:**305–319.