Generalizable Patterns in Neuroimaging: How Many Principal Components?

Lars Kai Hansen[1], Jan Larsen, Finn Årup Nielsen,

Dept. of Mathematical Modelling, Build. 321,

Technical University of Denmark, DK-2800 Lyngby, Denmark.

Tel: (+45) 4525 3889 – Fax: (+45) 4587 2599

Email: lkhansen@imm.dtu.dk

Stephen C. Strother

PET Imaging Service, VA Medical Center, and

Radiology+Neurology Depts. University of Minnesota,

Minneapolis, Minnesota, USA.

Egill Rostrup

Danish Center for Magnetic Resonance,

Hvidovre Hospital, Denmark,

Robert Savoy

Massachusetts General Hospital,

Boston, USA,

Nicholas Lange

McLean Hospital & Harvard Medical School

Boston, USA,

John Sidtis

Neurology Department University of Minnesota,

Minneapolis, Minnesota, USA.

Claus Svarer, Olaf B. Paulson

Neurobiology Research Unit,

Rigshospitalet, Copenhagen, Denmark

[1]Corresponding author

Running title: *Generalizable patterns*

Abstract

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.

1 Introduction

Principal component analysis (PCA) and the closely related singular value decomposition (SVD) technique are popular tools for analysis of image data sets and are actively investigated in functional neuroimaging [Moeller & Strother 91, Friston et al. 93, Friston et al. 95, Lautrup et al. 95, Strother et al. 95, Bullmore et al. 96, Strother et al. 96, Ardekani et al. 98, Worsley et al. 97]. 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 in 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 & Hansen 98]. Unsupervised learning has not enjoyed the same attention, although results for specific learning machines can be found. In [Hansen & Larsen 96] 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. 97].

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 examining the generalization error as 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.

2 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 & Tibshirani 90, Geman et al. 92, Mørch et al. 97]. 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 non-generic 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

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

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

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

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

Note that the generalization error is non-observable, i.e., it has to be estimated either from a finite *test set* also drawn from $p(x)$, or estimated from the training set using statistical arguments. In [Hansen & Larsen 96] 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 $\dim(\theta)$.

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

Empirical generalization estimates are obtained by dividing the data set into separate sets for training and testing, possibly combined with resampling, see [Stone 74, Toussaint 74, Hansen & Salamon 90, Larsen & Hansen 95]. Conventionally resampling schemes are classified as cross-validation [Stone 74] or bootstrap [Efron 83, Efron & Tibshirani 93] 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 non-overlapping test sets. In both cases we obtain *unbiased* estimates of the average generalization error. This only requires that test and training sets are independent.

2.1 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 91]. This scheme is well-suited for high-dimensional, highly correlated data, as, e.g., found in exploratory analysis of functional neuroimages

[Moeller & Strother 91, Friston et al. 93, Lautrup et al. 95, Strother et al. 95, Ardekani et al. 98, Worsley et al. 97]. A number of neural network architectures are devised to estimate principal component subsets without first computing the covariance matrix, see e.g., [Oja 89, Hertz et al. 91, Diamantaras & Kung 96]. Selecting the optimal number of principal components is a largely unsolved problem, although a number of statistical tests and heuristics have been proposed [Jackson 91]. 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. 93]. See [Akaike 69, Ljung 87, Wahba 90] for numerous applications of test error methods within system identification.

We follow [Hansen & Larsen 96] 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

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

The signal and noise are assumed to be distributed as a 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 91].

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,

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

with $\Delta x = x - x_0$.

2.1.1 Parameter estimation

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

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

Our model constraint involved in the approximation $\Sigma = \Sigma_s + \sigma^2 I_L$ is implemented as follows. Let $\widehat{\Sigma} = S\Lambda S^\top$ where S is an orthogonal matrix of eigenvectors and $\Lambda = \mathrm{diag}([\lambda_1, \cdots, \lambda_L])$ 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

$$\widehat{\Sigma}_K = S \cdot \mathrm{diag}([\lambda_1, \cdots, \lambda_K, 0, \cdots, 0]) \cdot S^\top. \tag{10}$$

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

$$\widehat{\sigma}^2 = \frac{1}{L-K}\mathrm{Trace}[\widehat{\Sigma} - \widehat{\Sigma}_K], \tag{11}$$

hence

$$\widehat{\Sigma}_n = \widehat{\sigma}^2 I_L \tag{12}$$

and

$$\widehat{\Sigma}_s = S \cdot \mathrm{diag}([\lambda_1 - \widehat{\sigma}^2, \cdots, \lambda_K - \widehat{\sigma}^2, 0, \cdots, 0]) \cdot S^\top. \tag{13}$$

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

2.2 Estimating the PCA Generalization Error

When the training set for an adaptive system becomes large relative to the number of fitted parameters 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* parameterized 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 87]. 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 & Larsen 96], where we derived the relation

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

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 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,

$$\begin{aligned} E &= \frac{1}{2}\log|2\pi(\widehat{\Sigma}_s + \widehat{\Sigma}_n)| + \frac{1}{2}\mathrm{Trace}[(\widehat{\Sigma}_s + \widehat{\Sigma}_n)^{-1}\widehat{\Sigma}_{\mathrm{train}}] & (15) \\ \widehat{G}_{\mathrm{testset}} &= \frac{1}{2}\log|2\pi(\widehat{\Sigma}_s + \widehat{\Sigma}_n)| + \frac{1}{2}\mathrm{Trace}[(\widehat{\Sigma}_s + \widehat{\Sigma}_n)^{-1}\widehat{\Sigma}_{\mathrm{test}}], & (16) \end{aligned}$$

where the covariance matrices, $\widehat{\Sigma}_{\mathrm{train}}, \widehat{\Sigma}_{\mathrm{test}}$ 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

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

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

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

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

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

The non-zero eigenvalues and their eigenvectors can, e.g., be found by *singular value decomposition* of the data matrix $X \equiv [x_\alpha]$. In equations (19-20) we have assumed zero mean signals, $x_0 = 0$, for simplicity.

Here we assume that the principal components are *importance* ranked according by 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 providing 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 non-zero covariance eigenvalues for $N < L$) possible subsets is out of the question for most neuroimaging problems.

3 Results and Discussion

3.1 Example I: Motor study

An fMRI activation image set of a single subject performing a left-handed finger-to-thumb opposition task was acquired. Multiple runs of 72 2.5-second (24 baseline, 24 activation, 24 baseline) whole-brain echo planar scans were aligned, and an axial slice through primary motor cortex and the suplementary motor area (SMA) of 42 x 42 voxels (3.1 x 3.1 x 8

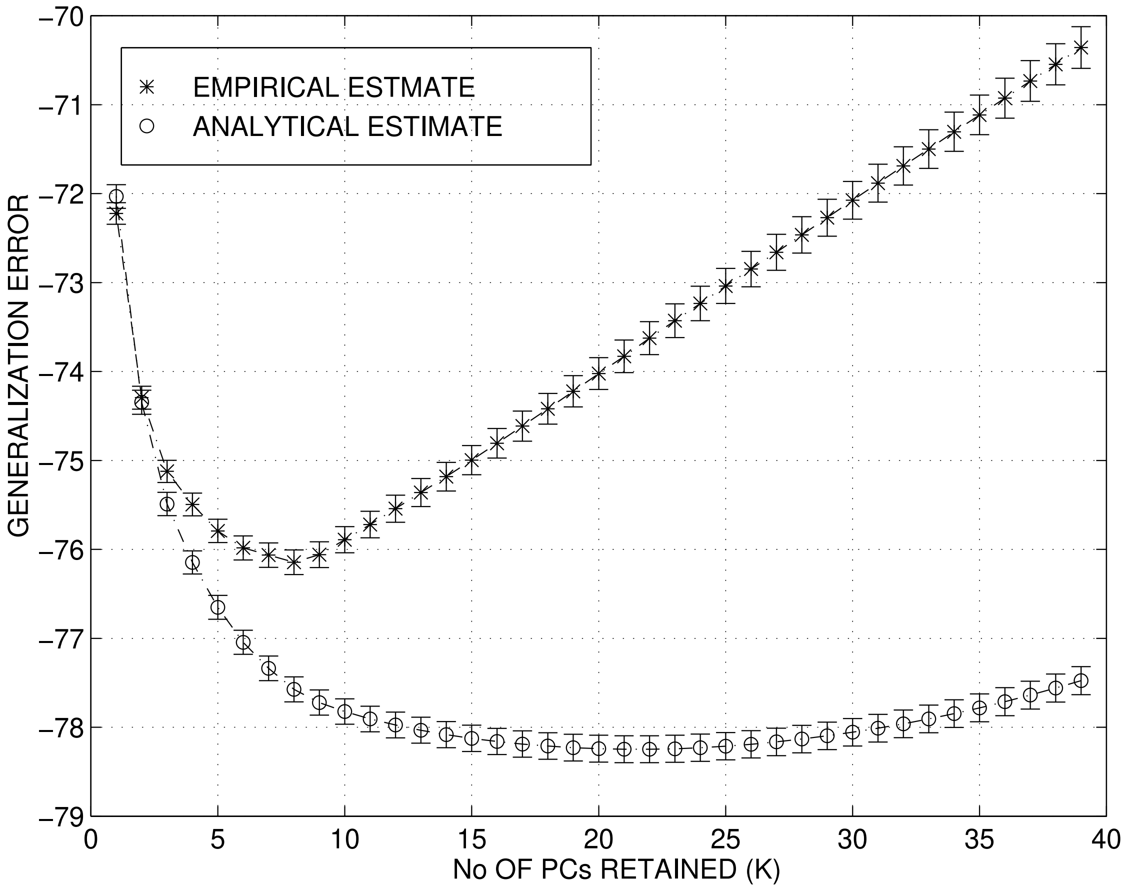

Figure 1: Data set I. Bias/Variance trade-off curves for PCA. The test set ($*$) generalization error estimate (mean plus/minus the standard deviation of the mean, for 10 repetitions of experiment) and the asymptotic estimate. The empirical test error is an unbiased estimate, while the analytical estimate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

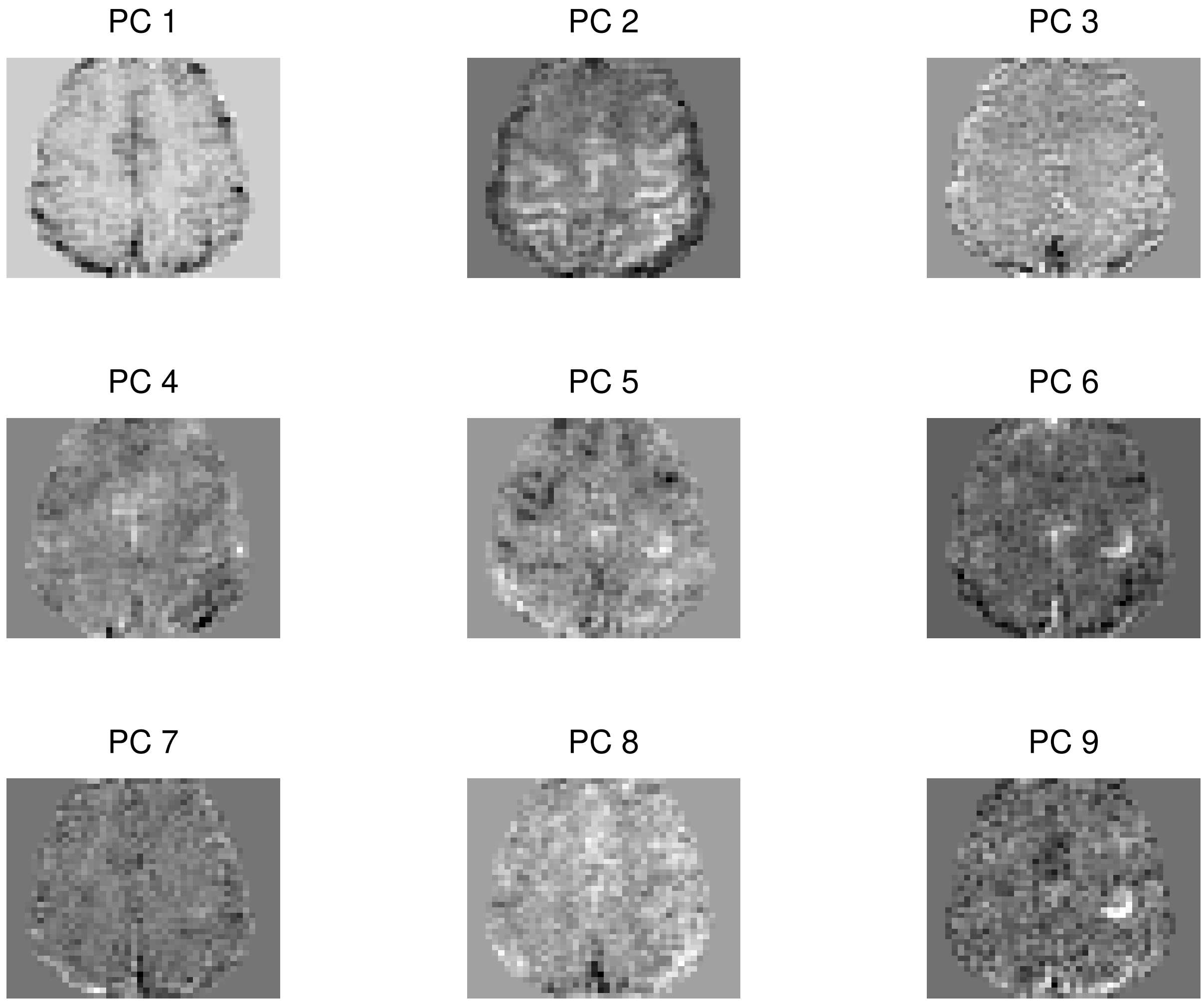

Figure 2: Data set I. Eigenimages corresponding to the nine most significant principal components. The unbiased generalization estimate in figure 1 suggests that the optimal PCA contains eight components. Among the generalizable patterns we find in sixth component, "PC 6", focal activation in the right 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.

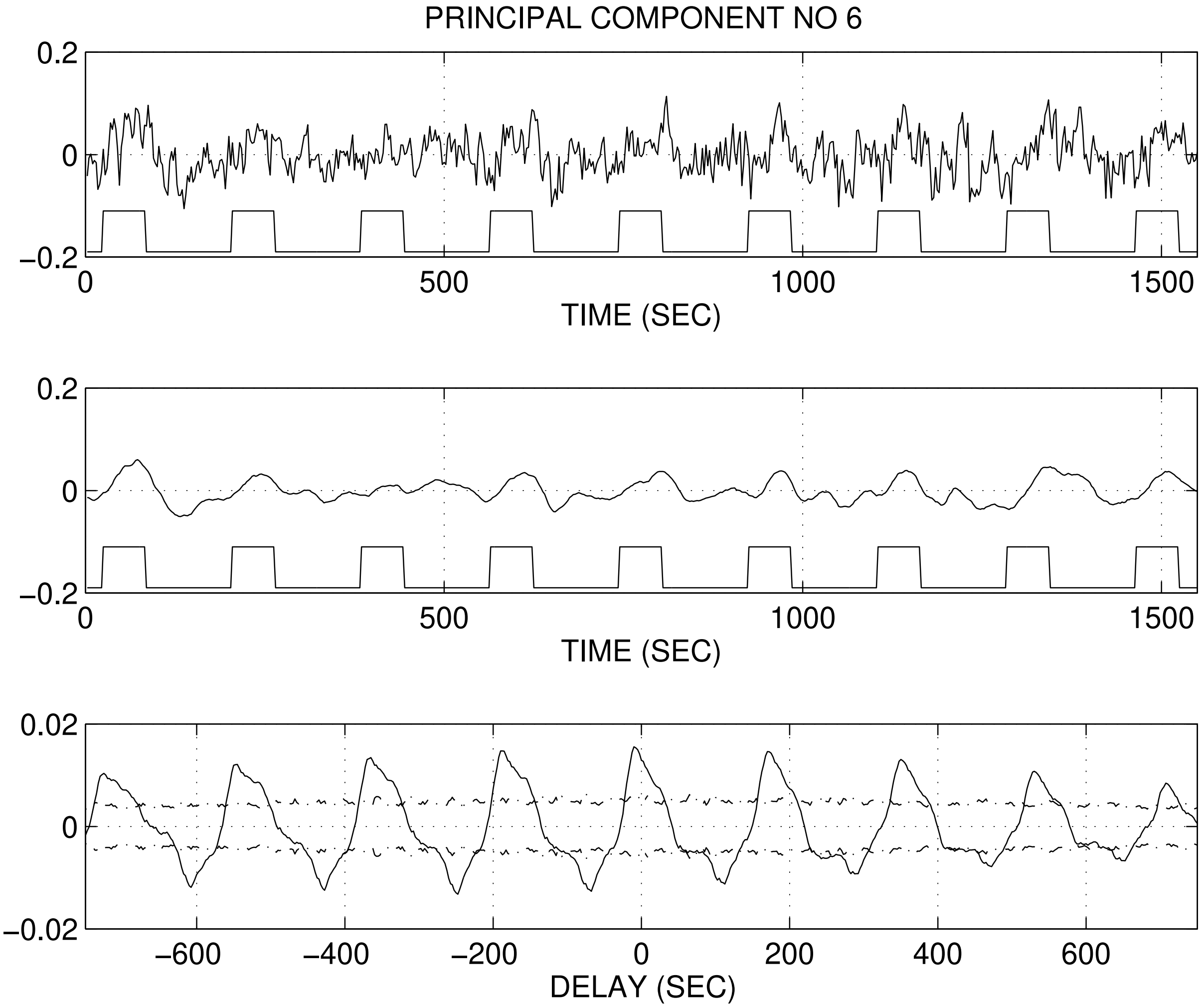

Figure 3: Data set I. Upper panel: The "raw" time course of principal component 6 (projection of the image sequence onto eigenimage number 6 of the covariance matrix), the offset square wave indicates the time course of the binary finger opposition activation. The subject performed finger tapping at time intervals where this function is high. Middle panel: 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. Lower panel: the cross-correlation function of the reference sequence and the principal component (un-smoothed), 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. 96]. $1/p = 1000$ 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.

mm) extracted. Of a total of 624 scans training sets of size $N = 300$ were drawn at random from the pool of scans and for each training set the remaining independent set of 324 scans used as test set. PCA analyses were carried out on the training set. We used equation (19) to compute the average negative log-likelihood on the test set using the covariance structure estimated on the training sets and equations (4) and (18) to compute the analytical estimate of the generalization error. Both estimates were then plotted versus size of the PCA subspace, see figure 1.

Inspection of the test set curve suggest 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 it 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 1-2 components.

The covariance eigenimages are shown in figure 2. Images corresponding to components 1-5 are dominated by signal sources that are highly localized spatially (hot spots comprising 1-4 neighbor pixels). It is compelling to defer these as confounding *vascular* signal sources. Component number six, however, has a somewhat more extended hot spot in the contralateral, (right hemisphere) motor area. In figure 3 we provide a more detailed temporal analysis of this signal source. In the upper panel 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 panel we give a low-pass filtered version of the principal component sequence. The smoothed component shows a definite response to the activation, though with a some randomness in the actual delay and shape of the response. In the bottom panel we plot the cross-correlation function between the on/off reference function and the un-smoothed signal. The cross-correlation function shows the characteristic periodic saw-tooth shape resulting from correlation of a square-wave signal with a delayed square wave. The horizontal dash-dotted 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 non-trivial way on the detailed specification of the null and is not relevant to the present exploratory analysis.

3.2 Example II: Visual stimulation

A single slice holding 128×128 pixels was acquired with a time interval between successive scans of TR = 333 msec. 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 are much shorter than in the previous analysis temporal correlations are expected on the hemodynamic time scale (5-10 sec). Hence, we have used a block-resampling scheme: the generalization error is computed on a randomly selected "hold-out" contiguous time interval of 50 scans ($\sim$ 16.7 seconds). The procedure was repeated 10 times with different generalization intervals. In figure 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 figure 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 figure 6 we show the corresponding covariance eigenimages. The first component is dominated by an extended focal activity in the 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

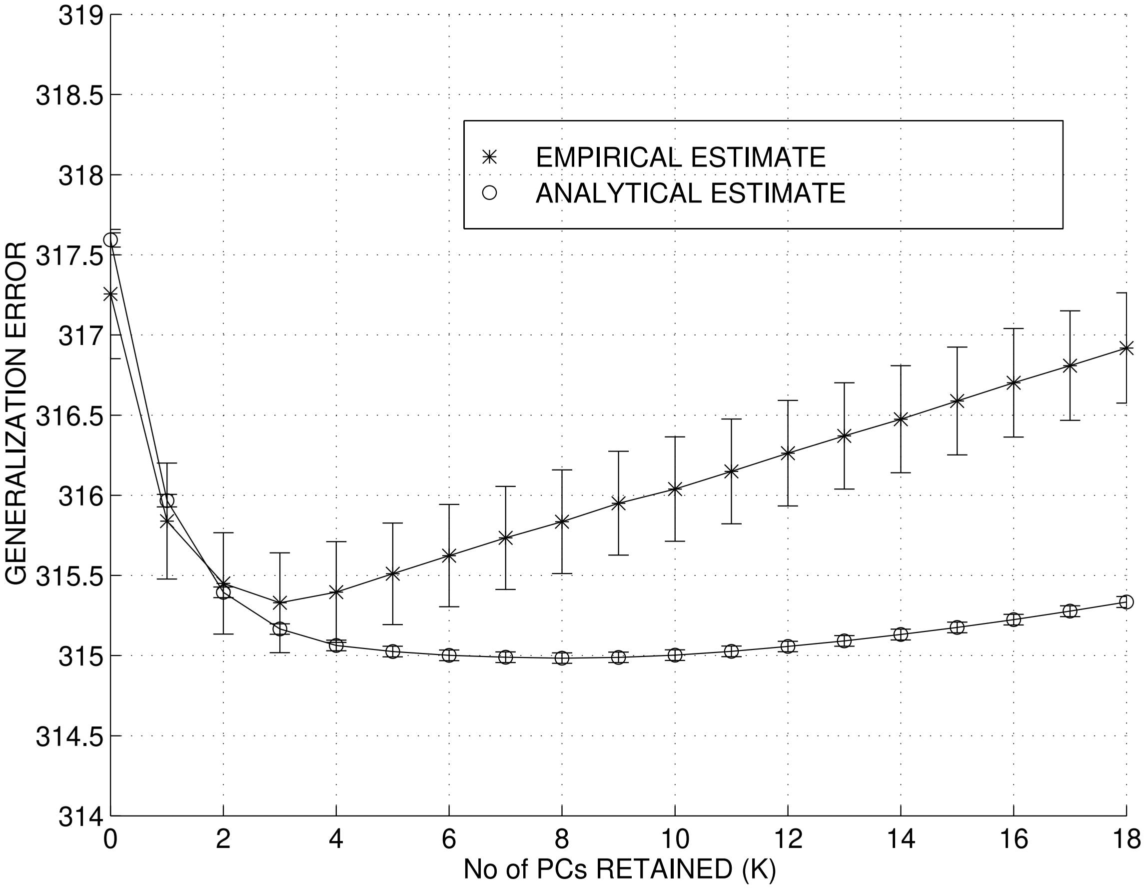

Figure 4: Data set II. Bias/Variance trade-off curves for PCA on visual stimulation sequence. The upper curve is the unbiased test set generalization error (mean of 10 repetitions of the experiment, the error bars indicate one standard deviation of the mean). The second curve is the corresponding analytical 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.

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 figure 7 we have performed cross-correlation analyses of all of the nine most variant principal component sequences. The horizontal dash-dotted 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 first component active mainly during the third run of the experiment.

4 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* ubiased.

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 over-parametrized.

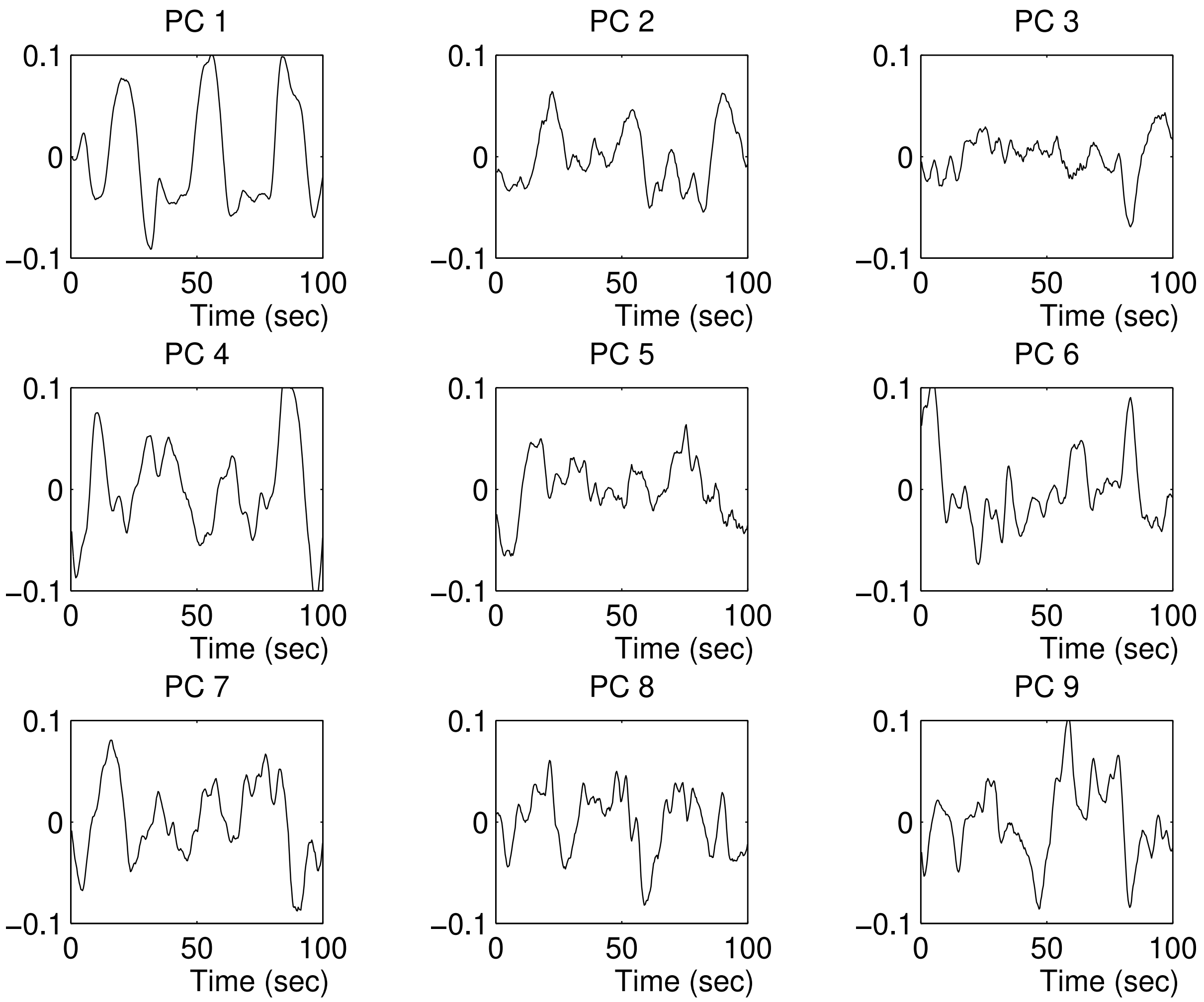

Figure 5: Data set II. The principal components corresponding to the nine largest covariance eigenvalues for a sequence of 300 fMRI scans of a visually stimulated subject. Stimulation takes places at scan times $\tau = 8 - 25$ sec, $\tau = 42 - 59$ sec, and $\tau = 75 - 92$ sec relative to the start of this three run sequence. Scan time sampling interval was $TR = 0.33$ sec. 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 figure 4, we find that only the time sequences corresponding to the three first components generalize.

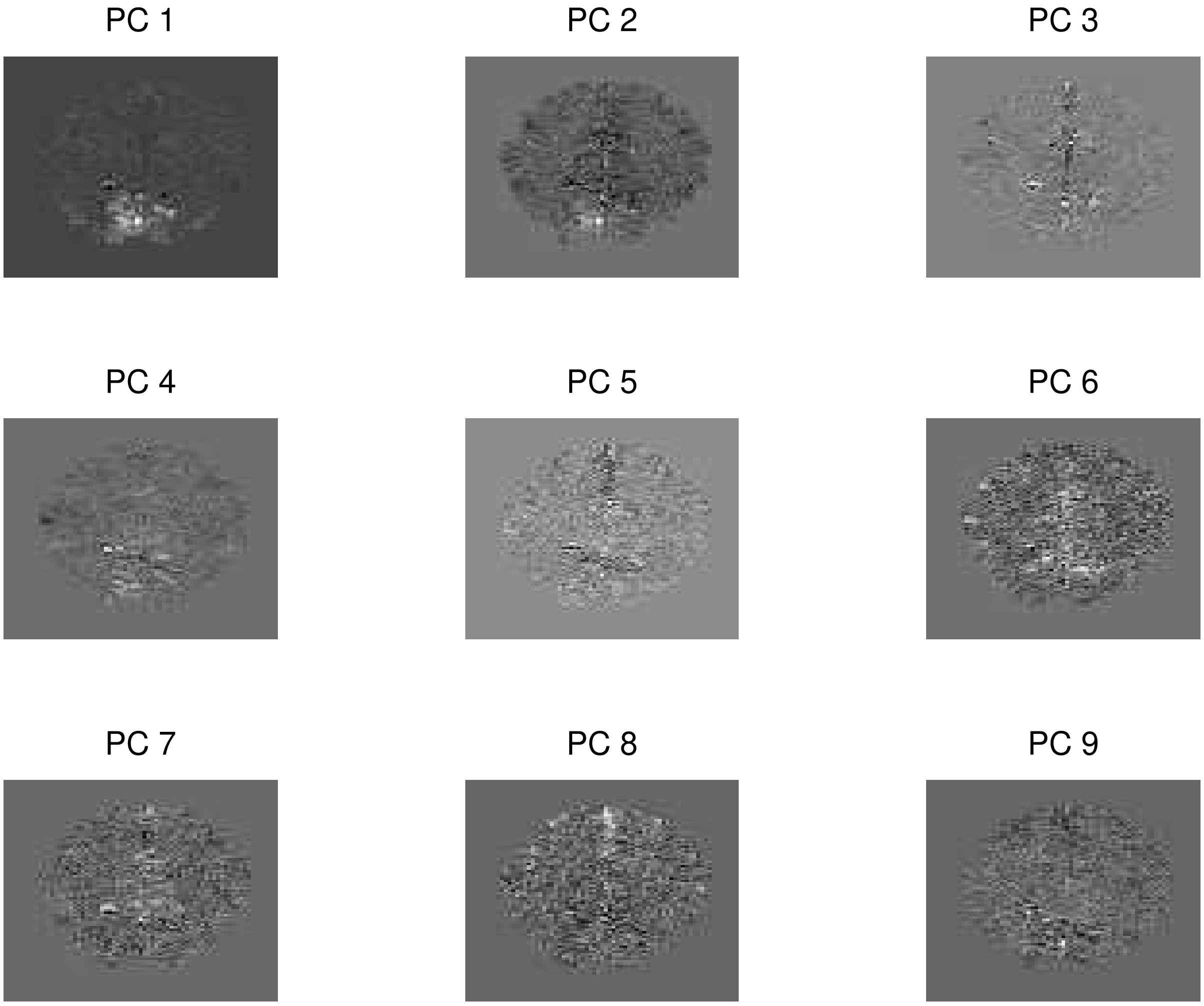

Figure 6: Data set II. Covariance eigenimages corresponding nine most significant principal components. The eigenimage corresponding to the dominating first PC is focused in visual areas. Using the Bias/Variance trade-off curves in figure 4, we find that only the eigenimages corresponding to the three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

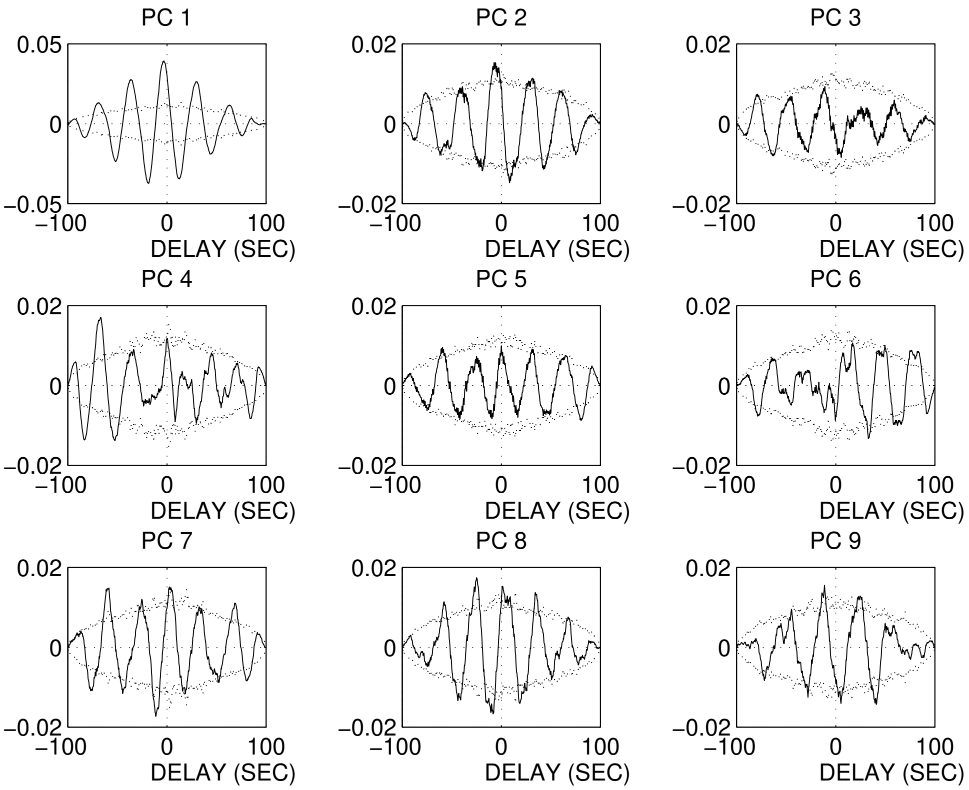

Figure 7: Data set II. cross-correlation analyses based on a reference time function taking value -1 at times corresponding to rest scans and value $+1$ for scans taken during activation. Each panel shows the cross-correlation of the principal component and the reference function for times $\tau \in [-100sec, 100sec]$. The dotted horizontal curves are $p = 0.001$ levels in a non-parametric permutation test for significant cross-correlation. These curves are computed in a manner similar to [Holmes et al. 96]: $1/p = 1000$ 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., figure 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 sec.

5 Acknowledgments

The authors wish to 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 69] H. Akaike: *Fitting Autoregressive Models for Prediction.* Ann. Inst. Stat. Mat. **21**, 243–247 (1969).

[Ardekani et al. 98] B. Ardekani, S.C. Strother, J.R. Anderson, I. Law, O.B. Paulson, I. Kanno, D.A. Rottenberg: *On Detection of Activation Patterns Using Principal Component Analysis.* In Proc. of BrainPET'97. "Quantitative functional brain imaging with Positron Emission Tomography". R.E. Carson et al. Eds. Academic Press, San Diego. In Press (1998).

[Bullmore et al. 96] Bullmore, E.T., Rabe-Hasketh, S., Morris, R.G., Williams, S.C.R., Gregory, L., Gray, J.A., and Brammer, M.J. *Functional magnetic resonance image analysis of a large-scale neurocognitive network.* NeuroImage, 4: 16-33, (1996).

[Efron 83] , Efron, B. The Jackknife, Boostrap and Other Resampling Plans. SIAM Monograph no. 38. CBMS-NSF. Philadelphia (1983).

[Efron & Tibshirani 93] B. Efron & R.J. Tibshirani: *An Introduction to the Bootstrap.* New York: Chapman & Hall (1993).

[Friston et al. 93] K.J. Friston, C.D. Frith, P.F. Liddle, and R.S.J. Frackowiak: *Functional Connectivity: The principal-component analysis of large (PET) data sets.* Journal of Cerebral Blood Flow and Metabolism **13**, 5-14 (1993).

[Friston et al. 95] Friston, K.J., Frith, C.D., Frackowiak, R.S.J., and Turner, R. *Characterizing dynamic brain responses with fMRI: a multivariate approach.* NeuroImage, 2: 166-172, (1995).

[Geman et al. 92] S. Geman, E. Bienenstock, and R. Doursat *Neural Networks and the Bias/Variance Dilemma.* Neural Computation, **4**, 1-58 (1992).

[Hansen & Salamon 90] L.K Hansen and P. Salamon: *Neural Network Ensembles.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **12**, 993-1001 (1990).

[Hansen & Larsen 96] L.K. Hansen and J. Larsen *Unsupervised Learning and Generalization.* In Proceedings of the IEEE International Conference on Neural Networks 1996, Washington DC, vol. 1, 25-30 (1996).

[Hansen et al. 97] Hansen L.K., Nielsen F.AA., Toft P., Strother S.C., Lange N., Morch N., Svarer C., Paulson O.B., Savoy R., Rosen B., Rostrup E., Born P. *How many Principal Components?* Third International Conference on Functional Mapping of the Human Brain, Copenhagen, 1997. NeuroImage **5**, 474, (1997)

[Hastie & Tibshirani 90] Hastie, T., and Tibshirani, R. *Generalized Additive Models*, Chapman & Hall, London (1990).

[Hertz et al. 91] J. Hertz, A. Krogh & R.G. Palmer: *Introduction to the Theory of Neural Computation.* Redwood City, California: Addison-Wesley Publishing Company (1991).

[Holmes et al. 96] A.P. Holmes , R.C. Blair, J.D.G. Watson and I. Ford: *Non-parametric analysis of statistic images from mapping experiments.* J. Cereb. Blood Flow Metab. **16**, 7-22 (1996).

[Jackson 91] J.E. Jackson: *A Users Guide to Principal Components.* John Wiley & Sons Inc., New York (1991).

[Diamantaras & Kung 96] K.I. Diamantaras and S.Y. Kung: *Principal Component Neural Networks: Theory and Applications* John Wiley & Sons Inc., New York (1996).

[Larsen & Hansen 94] J. Larsen & L.K. Hansen: *Generalization Performance of Regularized Neural Network Models.* In J. Vlontzos, J.-N. Hwang & E. Wilson (eds.), Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, Piscataway, New Jersey: IEEE, 42–51, (1994).

[Larsen & Hansen 95] J. Larsen & L.K. Hansen: *Empirical Generalization Assessment of Neural Network Models.* In F. Girosi, J. Makhoul, E. Manolakos & E. Wilson (eds.). Proceedings of the IEEE Workshop on Neural Networks for Signal Processing V, Piscataway, New Jersey: IEEE, 30–39, (1995).

[Larsen & Hansen 98] J. Larsen & L.K. Hansen: *Generalization: The Hidden Agenda of Learning* in J.-N. Hwang, S.Y. Kung, M. Niranjan & J.C. Principe (eds.), The Past, Present, and Future of Neural Networks for Signal Processing, IEEE Signal Processing Magazine, 43-45, Nov. 1997.

[Lautrup et al. 95] B. Lautrup, L.K. Hansen I. Law, N. Mørch, C. Svarer, S.C. Strother: *Massive weight sharing: A cure for extremely ill-posed problems.* In H.J. Herman et al., eds. Supercomputing in Brain Research: From Tomography to Neural Networks. World Scientific Pub. Corp. 137-148 (1995).

[Ljung 87] L. Ljung: *System Identification: Theory for the User.* Englewood Cliffs, New Jersey: Prentice-Hall, (1987).

[Moeller & Strother 91] J.R. Moeller and S.C. Strother: *A regional covariance approach to the analysis of functional patterns in positron emission tomographic data.* J. Cereb. Blood Flow Metab. **11**, A121-A135 (1991).

[Moody 91] J. Moody: *Note on Generalization, Regularization, and Architecture Selection in Nonlinear Learning Systems.* In B.H. Juang, S.Y. Kung & C.A. Kamm (eds.) Proceedings of the first IEEE Workshop on Neural Networks for Signal Processing, Piscataway, New Jersey: IEEE, 1–10, (1991).

[Mørch et al. 97] N. Mørch, L.K. Hansen, S.C. Strother,C. Svarer, D.A. Rottenberg, B. Lautrup, R. Savoy, O.B. Paulson: *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 **1230**, 259-270: Springer-Verlag (1997).

[Murata et al. 94] N. Murata, S. Yoshizawaand & S. 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).

[Oja 89] E. Oja: *Neural Networks, Principal Components, and Subspaces.* International Journal of Neural Systems, **1**, 61-68 (1989).

[Seber & Wild, 89] G.A.F. Seber & C.J. Wild: *Nonlinear Regression.* New York: John Wiley & Sons, (1989).

[Shao & Tu 95] J. Shao & D. Tu: *The Jackknife and Bootstrap.* Springer Series in Statistics, Berlin, Germany: Springer-Verlag (1995).

[Stone 74] M. Stone: *Cross-validatory Choice and Assessment of Statistical Predictors.* Journal of the Royal Statistical Society B, **36**, 111–147 (1974).

[Strother et al. 95] S.C. Strother, A.R. Anderson, K.A. Schaper, J.J. Sidtis, R.P. Woods. D.A. Rottenberg: *Principal Component Analysis and the Scaled Subprofile Model compared to Intersubject Averaging a Statistical Parametric Mapping: I. "Functional Connectivity" of the Human Motor System Studied with [^{15}O]PET.* J. Cereb. Blood Flow Metab. **15**, 738-75 (1995).

[Strother et al. 96] , *Multidimensional state spaces for fMRI and PET activation studies.* 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., Rottenberg D.A. Neuroimage, 3(Part 2):98, (1996).

[Svarer et al. 93] C. Svarer, L.K. Hansen, and J. Larsen: *On Design and Evaluation of Tapped-Delay Neural Network Architectures.* In H.R. Berenji *et al.* (eds.) Proceedings of the 1993 IEEE Int. Conference on Neural Networks, IEEE Service Center, NJ, vol. 1, 46-51 (1993).

[Toussaint 74] G.T. Toussaint: *Bibliography on Estimation of Misclassification.* IEEE Transactions on Information Theory, **20**, 472-479 (1974).

[Wahba 90] G. Wahba: *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM **59**, (1990).

[Worsley et al. 97] K.J. Worsley, J.-B. Poline, K.J. Friston, and A.C. Evans: *Characterizing the Response of PET and fMRI Data Using Multivariate Linear Models (MLM).* NeuroImage **6** 305-319 (1997).