Multivariate strategies in functional magnetic resonance imaging

Lars Kai Hansen

Informatics and Mathematical Modelling,
Technical University of Denmark,
DK-2800 Kgs. Lyngby, Denmark

Abstract

We discuss aspects of multivariate fMRI modeling, including the statistical evaluation of multivariate models and means for dimensional reduction. In a case study we analyze linear and non-linear dimensional reduction tools in the context of a 'mind reading' predictive multivariate fMRI model.

Key words: Generalization, dimensional reduction, Laplacian eigenmap, NPAIRS, fMRI

Introduction

The human brain processes information in a massively parallel network of highly interconnected neuronal ensembles. In order to understand the resulting long range spatio-temporal correlations in functional image sets, we and others have pursued a wide variety of multivariate analysis strategies. To reduce the two important sources of bias in neuroimage analysis, namely model bias and the statistical bias arising from relatively small sample sizes available for modeling.

Neuroimaging experiments are designed to explore the spatio-temporal pattern of information processing in the brain associated with a given behavior or delivery of stimulus. Stimuli may be external, e.g., passive viewing, hearing etc., or may be internally generated by spontaneous cognitive processes or physical activity, e.g., speaking, motor activity; for a review see (Frackowiak et al., 2003). The brain imaging device measures the mesoscopic brain state,

Email address: lkh@imm.dtu.dk (Lars Kai Hansen).

i.e., information processing averaged over small volumes, called voxels. We denote this high-dimensional image measurement by $\mathbf{x}(t)$, with t being the time of acquisition. The corresponding macroscopic cognitive state is denoted by $\mathbf{g}(t)$. The total amount of data is denoted $D = (\mathbf{x}_t, \mathbf{g}_t,)t = 1, ..., T$.

Thus, the neuroimaging agenda is to explore the joint distribution: $p(\mathbf{x}, \mathbf{g})$ based on D and background neuroinformatics knowledge bases.

The joint distribution of brain states and behavior can be modelled directly, or as more frequently done by one of the two equivalent factorizations: $p(\mathbf{x}, \mathbf{g}) = p(\mathbf{x}|\mathbf{g})p(\mathbf{g})$ or $p(\mathbf{x}, \mathbf{g}) = p(\mathbf{g}|\mathbf{x})p(\mathbf{x})$.

By factoring $p(\mathbf{x}, \mathbf{g}) = p(\mathbf{x}|\mathbf{g})p(\mathbf{g})$, we consider the stimulus as the control signal and look for differences in the neuroimage distribution among different cognitive states. This is the most prevalent mode of analysis, dating back to the so-called subtraction paradigm, in which the mean images from two different conditions are subtracted and shown as a measure of *contrast.* This approach has been refined in the several neuroimage analysis tools, e.g., SPM, see (Frackowiak et al., 2003). In SPM the conditional neuroimage distribution is further factorized as $p(\mathbf{x}|\mathbf{g}) \sim \prod_i p(\mathbf{x}_i|\mathbf{g})$, where $\mathbf{x}_i$ are individual voxel measurements. A product of such univariate factors amounts is equivalent to assuming voxel-to-voxel independence, also known as 'naïve Bayes'. Further, by assuming that the univariate conditionals are all Gaussian distributions we arrive at the generative model equivalent to SPM's mass univariate t-test approach, framed in the so-called general linear model, see also (Kjems et al., 2002) for further discussions of the naive Bayes generative model.

On the other hand by factoring the joint pdf as $p(\mathbf{x}, \mathbf{g}) = p(\mathbf{g}|\mathbf{x})p(\mathbf{x})$, we enter was has been dubbed the *mind reading* paradigm, in which the model is set up to infer the instantaneous cognitive state from the concurrent neuroimage. This approach was first developed by in the mid-90s (Lautrup et al., 1994; Mørch et al., 1995, 1997) for functional neuroimaging based on PET and fMRI, and has also more recently gained interest in the machine learning community, see e.g., (Mitchell et al., 2004). As a historical note: the first application of artificial neural networks for classifications of brain scans date back to the 1992 (Kippenham et al., 1992), and concerned the classification of regionally averaged PET data in terms of normal and Alzheimers disease. Univariate and multivariate models including: SPM, cross-correlation analysis, independen component analysis, artificial neural networks, were carefully compared using both simulate and real data in the 1999 NeuroImage paper by (Lange et al., 1999).

Predictive value: The generalizability of a model

We model probability distributions (countable support) or probability density functions (continuous support) using parameterized families, $p(\mathbf{x},\mathbf{g}) \sim p(\mathbf{g},\mathbf{x}|\theta)$, where the parameters θ can be both continuous (like means and variances) or discrete (like model orders, number of components, etc.). A neuroimaging experiment is not able to pin down the parameters, rather we are left with an uncertainty captured by the so-called posterior distribution $p(\theta|D)$. Given the posterior distribution we may predict or simulate future data based on the initial data

$$p(\mathbf{x}_{t+1},\mathbf{g}_{t+1}|D) = \int p(\mathbf{x}_{t+1},\mathbf{g}_{t+1}|\theta)p(\theta|D)d\theta, \tag{1}$$

The closer these predictions are to 'true' pdf's: $p(\mathbf{x}_{t+1},\mathbf{g}_{t+1})$, the more generalizable the model is, i.e., the more similar are the two stochastic processes. Closeness may be measured by one of several costfunctions. Often used costfunctions are the miss-classification rate, the mean square error, and the deviance (Ripley, 1996). The generalization error is defined as the expected cost on a new datum.

The expected deviance is - apart from a an additive constant - identical to the basic information theoretic Kullback-Leibler measure,

$$\mathrm{KL}[p(.),p(.|D)] = \int\int \log\left[\frac{p(\mathbf{x},\mathbf{g})}{p(\mathbf{x},\mathbf{g}|D)}\right] p(\mathbf{x},\mathbf{g})d\mathbf{x}d\mathbf{g} \tag{2}$$

This loss is zero if the predictive distribution is identical to the 'true' distribution and otherwise positive for all other distributions. The generalization error difference between two models may be estimated by a sample of test data

$$\widehat{\Delta\mathrm{KL}} = \widehat{\mathrm{KL}}[p(.),p_1(.|D)] - \widehat{\mathrm{KL}}[p(.),p_2(.|D)] = \frac{1}{T_{test}}\sum_{\tau=1}^{T_{test}} \log\left[\frac{p_2(\mathbf{x}_\tau,\mathbf{g}_\tau|D)}{p_1(\mathbf{x}_\tau,\mathbf{g}_\tau|D)}\right] \tag{3}$$

This test error is an unbiased estimate of the generalization error if the test set is sampled independently from the training data D.

The generalization error typically depends strongly on the amount data in the training set $N = |D|$, and is also strongly dependent on the complexity of the model. Most often the generalization error decreases as a function of N, hence the models generalize better the more data is provided. If the model family can be described by a single complexity parameter, say the number of estimated parameters, there is typically a bias-variance trade-off: The best model is not too simple (biased) and not too complex. If the model is too complex the predictive distribution will be able to adapt to closely to the training data, hence, be highly too different between different training sets (variance).

Hence, one could say that there is always a hidden agenda in modeling from data: We are interested in models that predict well, not simply models that fit well to the training data.

It can be shown that for the deviance loss function the optimal predictive distribution is obtained using so-called Bayesian averaging,

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}, \tag{4}$$

where $p(D|\theta)$ is the likelihood function, and $p(\theta)$ is the 'true' prior distribution. In practise the true prior is not available and most often the likelihood is at best an approximation to the true likelihood. However, it is an empirical finding that even rather crude approximate Bayesian averaging procedures provide better predictive distributions than that obtained from using a point estimate, e.g., a maximum likelihood estimate $p(.) \sim p(.|\theta_{\mathrm{ML}}(D))$, where $\theta_{\mathrm{ML}}(D)$ maximize the likelihood function. The main difficulty with the Bayesian estimators is the often significant computational overheads incurred by parameter averaging. Note that the generalization error and the unbiased estimate can be used to evaluate any model, no matter how the models predictive distribution is obtained.

In neuroimaging we are interested in models that generalize and models that can be interpreted, typically in terms of a 'brain map': an image or volume of local activation involvement, i.e., a statistical parametric map. For multivariate neuroimaging models of the SPM is a function of the model parameters, hence, we are interested in a models for which the parameters are relatively robust.

To investigate the stability of representations we have suggested the NPAIRS framework. NPAIRS is based on split half resampling. For two models adapted to fit two independently sampled data subsets of the same size, i.e., the two sets in split half, any derived parameters should be identical between the two subsets. Thus the difference between SPMs estimated in the two sets is an unbiased estimator of the pixel-by-pixel variance. For additional discussion and examples see (Kjems et al., 2002; Strother et al., 2002). Other resampling strategies may under different assumption provide useful estimators, e.g., bootstrap and jackknife which is closely related to leave-one-out cross-validation, see e.g., (Kjems et al., 2000). In our analysis below we are primarily interested in the predictive value of the models, hence, we use the leave-one-out for estimation of performance.

Models

We characterize models as parametric, non-parametric, and semi-parametric (Bishop, 1995). Parametric models have fixed parameter sets, say mean and co-variance for a normal pdf model. Non-parametric models use the data set as model, e.g., nearest neighbor density models. Semi-parametric models have a general structure but with variable dimensionality of the parameterization, e.g., mixture models where the number of mixture components is determined from data. Examples of parametric models cover, e.g, modelling with normal distributions. The normal distribution is described by the mean vector and the co-variance matrix. Gaussian mixture models (clustering) and artificial neural networks are typical semi-parametric systems (Bishop, 1995). The parametrizations are adapted to the given data set. Nearest neighbor methods are generic non-parametric models. Non-parametric models are extremely flexible and need careful complexity control.

In machine learning a further distinction is made between supervised and unsupervised learning. In supervised learning the aim is to model conditional distributions, e.g., $p(\mathbf{g}|\mathbf{x})$. To adapt models we need supervised data sets with both inputs $\mathbf{x}$) and outputs $\mathbf{g}$. In unsupervised learning we are interested in modeling marginal pdf's, e.g., $p(\mathbf{x})$, which can be adapted from a sample of 'input' data only (Mørch et al., 1997).

Dimensionality reduction

Data sets of neuroimages usually have many more voxels J than image samples ($T << J$). This means that multivariate models involving the image representation can easily be ill-posed. If, e.g., a parameter is used for each image dimension we would thus invoke at least J parameters, which would be estimated on T samples. Thus, we need to consider the representation carefully for example by an initial dimensional reduction step prior to modeling.

The primary objective of dimensional reduction is to create a mapping from the original high dimensional image space to a relevant low dimensional subspace in which we can safely establish the posterior distribution. Principal component analysis (PCA) is a well established scheme for dimension reduction in functional imaging, see e.g., the discussions in (Kjems et al., 2000, 2002; Strother et al., 2002). PCA identifies an orthogonal basis for a subspace which captures the most variance for a given dimensionality. Dimensional reduction using PCA is meaningful if the dominant effects in the data are those induced by the stimulus. More advanced representations based linear projection schemes, some of which are further including information from the reference function are discussed in (Worsley et al., 1997).

It is worth noting that the PCA approach is not jeopardized if the data contains high variance, confounding signal components because these can be suppressed in the subsequent modeling; the requirement is that the effects of interest are present in the subspace. We will therefore be quite liberal in our choice of subspace dimension in the following. PCA is achieved by *singular value decomposition* (SVD), see e.g. (Kjems et al., 2000). The data matrix $\mathbf{D}$ of size $J \times T$ where $T \ll J$, is decomposed into

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{\top}, \tag{5}$$

where $\mathbf{U}$ is a $J \times T$ orthonormal matrix, $\mathbf{S}$ is a $T \times T$ diagonal matrix and $\mathbf{V}$ is $T \times T$ orthonormal matrix using a so-called 'economy size' decomposition where the null space has been removed. The diagonal of matrix $\mathbf{S}$ has nonnegative elements in descending order. These diagonal elements are the singular values that correspond to standard deviations of the input data projected onto the given basis vectors represented by matrix $\mathbf{U}$. The reduced input space is obtained by using only some fixed $K \leq N$ number of the largest principal components. The reduced data matrix is given by

$$\mathbf{X} = \tilde{\mathbf{U}}^{\top}\mathbf{D}, \tag{6}$$

where the transformation matrix from the original input space to the reduced inputs space is given by $\tilde{\mathbf{U}}$, a $F \times K$ sub-matrix of $\mathbf{U}$. Here we use the fact that $\mathbf{U}$ is orthonormal giving $\mathbf{U}^{-1} = \mathbf{U}^{\top}$.

Our focus here is on the prediction and reproducibility performance of models. We will invoke a resampling (cross-validation) scheme. Thus we need to enforce that not only the classifier is generalizable, but also the reduced representation. Using all the data to estimate the transformation matrix $\tilde{\mathbf{U}}$ would introduce dependence between the training set and test set. To avoid this dependence, the PCA is computed using only training data, giving a $\tilde{\mathbf{U}}$ transformation matrix. This matrix is then used to transform the test data to the lower dimensional subspace with

$$\mathbf{X}_{te} = \tilde{\mathbf{U}}^{\top}\mathbf{D}_{te}, \tag{7}$$

which ensures the unbiased nature of the test data. Tools exist for optimizing the PCA signal-to-noise ratio, with respect to the dimensionality K (Hansen et al., 1999; Beckmann et al., 2001).

Linear dimension reduction can be based on more advanced decompositions, such as independent component analysis (ICA) (McKeown et al., 2003). The advantage of ICA is that is not subject to the strong orthogonality constraints of PCA, which means that it can eliminate more subtle artifacts than PCA. For relatively high-dimensional mappings the difference in subspaces spanned by the orthogonal basis by PCA and the non-orthogonal basis of ICA is not likely to be of much significance.

Non-linear dimensional reduction has not been explored nearly as widely as their linear counterparts. (Thirion and Faugeras, 2004) investigated so-called Laplacian eigenmaps (Belkin and Niyogi, 2002) as a dimensional tool, these maps are created as non-linear low-dimensional representations that as closely as possible maintains the topological neighbors in a high-dimensional feature space, similar in spirit to the classical multi-dimensional scaling method (Kruskal and Wish, 1978).

The Laplacian eigenmap is based on a non-Euclidean metric, defined though a kernel function

$$d(j,k) = K(\mathbf{x}_j - \mathbf{x}_k) \tag{8}$$

In many applications the kernel is chosen to be the gaussian function $K(u) = \exp(-u^2/\sigma^2)$. Based on the metric the $N \times N$ matrix $\mathbf{Q}$ is established as the matrix of pairwise distances between data points. Let $\mathbf{r}_j$ be the $j'th$ row sum of $\mathbf{Q}$, and let $\mathbf{R}$ be the diagonal matrix with vector $\mathbf{r}$ in the diagonal. The Laplace eigenmaps are the N-dimensional eigenvectors of matrix $\mathbf{Q} - \mathbf{R}$. The Laplace eigenmap is closely related to so-called spectral clustering (Weiss, 1999), and the interpretation is similar. Two data points are active in a given Laplace eigenmap if there is a 'diffusion path' between the two formed by neighboring data points active in the given map. While linear dimensional reduction based on variance can be expected to work well for modelling strong mean difference effects, the Laplacian eigenmap can form a representation in which warped mean effects are present, say in the presence of a weak non-stationarity or drift of the class means so that they form non-linear trajectories in input space, say as function of time, see e.g., (Ng et al., 2001) for several creative illustrations.

Case Study

The data set used for illustration of multivariate models in this presentation was acquired by dr. Egill Rostrup at Hvidovre Hospital on a 1.5 T Magnetom Vision MR scanner. The scanning sequence was a 2D gradient echo EPI ($T2*$ weighted) with 66 ms echo time and 50 degrees RF flip angle. The images were acquired with a matrix of 128×128 pixels, with FOV of $230mm$, and $10mm$ slice thickness, in a para-axial orientation parallel to the calcarine sulcus. The visual paradigm consisted of a rest period of $20sec$ of darkness using a light fixation dot, followed by $10sec$ of full-field checkerboard reversing at $8Hz$, and ending with $20sec$ of rest (darkness). In total, 150 images were acquired in $50sec$, corresponding to a period of approximately $330msec$ per image. The experiment was repeated in 10 separate runs containing 150 images each. In order to reduce saturation effects, the first 29 images were discarded, leaving 121 images for each run.

Representations

Based on the 1210 scans included in the analysis we form a linear principal component representation and a nonlinear representation based on the Laplacian eigenmaps. The topology of the two representations are illustrated in Figure and Figure respectively. The linear representation shows a pronounced mean difference effect between baseline and activation scans, while the non-linear representation is somewhat more subtle showing difference in the shape of the clusters of baseline and activation scans, revealing potential non-stationarity in the two states.

[Figure 1 here]

[Figure 2 here]

Non-parametric modeling

The next step in modeling is to construct a map representing the relation between brain image and brain state. This have been accomplished using both parametric models, e.g., (Kjems et al., 2002), semi-parametric (Mørch et al., 1997), and non-parametric models e.g. (Mitchell et al., 2003). We here follow the latter and use the k-nearest neighbor model (KNN). This system is typically computing neighbors using an un-weighted Euclidean metric in feature space, and most often classifying by majority voting among the neighbors. It remains to determine the number of neighbors k. We use a leave-one-out resampling scheme in the training set to select k. This can be done a very little overhead. The leave-one-out error can simply be computed by classifying using the neighbors excluding the actual data point. The optimal k can then be used in classifying the training set. If a full Bayes posterior probability is required, the method is easily extended to compute local pdf values (Bishop, 1995).

We use both PCA and Laplacian eigenmap features to illustrate the procedure. We split the dataset in two equal size subsets: Five runs for training and five runs for testing. As the test and training data are independent, the test error estimates are unbiased estimator of performance. We use a simple on-off activation reference function for supervision of the classifier. The reference function is off-set by 4 seconds to emulate the hemodynamic delay.

For each method we estimate k for a number of feature space dimensions $d = 1 : 20$ using the leave-one-out (LOO) procedure on the training set. The LOO-optimal k is then applied to the test set for the same feature space dimensionality and a classification error rate estimated on the test set, the resulting relations between feature space dimensionality, LOO and test errors are indicated in Figure 3. The best (LOO) error is obtained for a $d = 6$ dimensional feature subspace for the Laplacian eigenmap, while the best model using PCA is a $d = 4$. The corresponding unbiased test set classification error rates are 5% and 9%, in favor of the non-linear features.

[Figure 3 here]

[Figure 4 here]

In Figure 4. we show the test set activation time series obtained by the two models. In the non-linear feature the errors basically occurs at the onset and at end of stimulation, while the PCA based linear feature representations also make generalization errors in the baseline, suggesting spurious short burst of activation.

Discussion

For modeling of neuroimaging data we are in general interested in models that are able to extract the relevant generalizable long range dependencies between behavior and brain state, hence, that have predictive power. Because of the high-dimensional representations that are inherent in multivariate models we need to exercise extreme care in model optimization. This includes dimensional reduction and feature selection. We have outlined a resampling based approach framework and demonstrated its implementation for general non-linear models including also non-linear dimensional reduction schemes. In a visual simulation study we have shown that non-linear dimensional reduction and non-parametric modeling led to optimal generalizability.

Acknowledgments

This work is supported Human Brain Project Grant P20 MN57180, and the Danish foundation Lundbeckfonden.

References

Beckmann, C., Noble, J., Smith, S., 2001. Investigating the intrinsic dimensionality of fmri data for ica. In: Proc. Seventh Int. Conf. on Functional Mapping of the Human Brain. NeuroImage. p. 13(6)S76.

Belkin, M., Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering.
URL `citeseer.ist.psu.edu/belkin01laplacian.html`

Bishop, C., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W. (Eds.), 2003. Human Brain Function, 2nd Edition. Academic Press.
URL `http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/`

Hansen, L., Paulson, O., Larsen, J., Nielsen, F., Strother, S., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., 1999. Generalizable Patterns in Neuroimaging: How Many Principal Components? NeuroImage 9, 534–544.

Kippenham, J., Barker, W., Pascal, S., Nagel, J., Duara, R., 1992. Evaluation of a neural network classifier for PET scans of normal and Alzheimers Disease Subjects. Journal Nuclear Medicine 33, 1459–1467.

Kjems, U., Hansen, L., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., Strother, S., 2002. The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves. NeuroImage 15 (4), 772–786.

Kjems, U., Hansen, L. K., Strother, S. C., 2000. Generalizable singular value decomposition for ill-posed datasets. In: NIPS. pp. 549–555.

Kruskal, J. B., Wish, M., 1978. Multidimensional Scaling. Sage Publications, Beverly Hills, California.

Lange, N., Strother, S. C., Anderson, J. R., Nielsen, F. Å., Holmes, A. P., Kolenda, T., Savoy, R., Hansen, L. K., September 1999. Plurality and resemblance in fMRI data analysis. NeuroImage 10 (3), 282–303.
URL `http://www.sciencedirect.com/science/article-/B6WNP-45FCP48-13/2/bd7e7f72099b83540609e24c627a2fc4`

Lautrup, B., Hansen, L., Law, I., Mørch, N., Svarer, C., Strother:, S., 1994. Massive Weight Sharing: A cure for Extremely Ill-posed Problems. In: Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks. Juelich, Germany, pp. 137–144.

McKeown, M., Hansen, L. K., Sejnowski, T. J., oct 2003. Independent component analysis for fmri: What is signal and what is noise? Current Opinion in Neurobiology 13 (5), 620–629.
URL `http://www2.imm.dtu.dk/pubdb/p.php?2878`

Mitchell, T., Hutchinson, R., abd R.S. Niculescu, M. J., Pereira, F., Wang, X., 2003. Classifying instantaneous cognitive states from fmri data. In: American Medical Informatics Association Annual Symposium.

Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to Decode Cognitive States from Brain Images. Machine Learning 57 (1).

Mørch, N., Hansen, L., Strother, S., Svarer, C., Rottenberg, D., Lautrup, B., Savoy, R., Paulson, O., 1997. Nonlinear versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover. In: Proceedings of the 15th International Conference on Information Processing in Medical Imaging, 1997. Vol. Springer Lecture Notes in Computer Science 1230. pp. 259–270.

Mørch, N., Kjems, U., Hansen, L., Svarer, C., Law, I., Lautrup, B., Strother, S., Rehm, K., 1995. Visualization of Neural Networks Using Saliency Maps. In: Proceedings of the 1995 IEEE International Conference on Neural Networks. Vol. 4. pp. 2085–2090.

Ng, A., Jordan, M., Weiss, Y., 2001. On spectral clustering: Analysis and an algorithm.
URL `citeseer.ist.psu.edu/ng01spectral.html`

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Strother, S., Anderson, J., Hansen, L., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework. NeuroImage 15 (4), 747–771.

Thirion, B., Faugeras, O., Apr. 2004. Nonlinear dimension reduction of fMRIdata: the Laplacian embedding approach. In: Proc. 2st Proc. IEEE ISBI. Arlington, VA, pp. 372–375.

Weiss, Y., 1999. Segmentation using eigenvectors: A unifying view. In: ICCV (2). pp. 975–982.
URL `citeseer.ist.psu.edu/weiss99segmentation.html`

Worsley, K. J., Poline, J.-B., Friston, K. J., Evans, A. C., November 1997. Characterizing the response of PET and fMRI data using multivariate linear models. NeuroImage 6 (4), 305–319.
URL `http://www.idealibrary.com/links/citation/1053-8119/6/305`

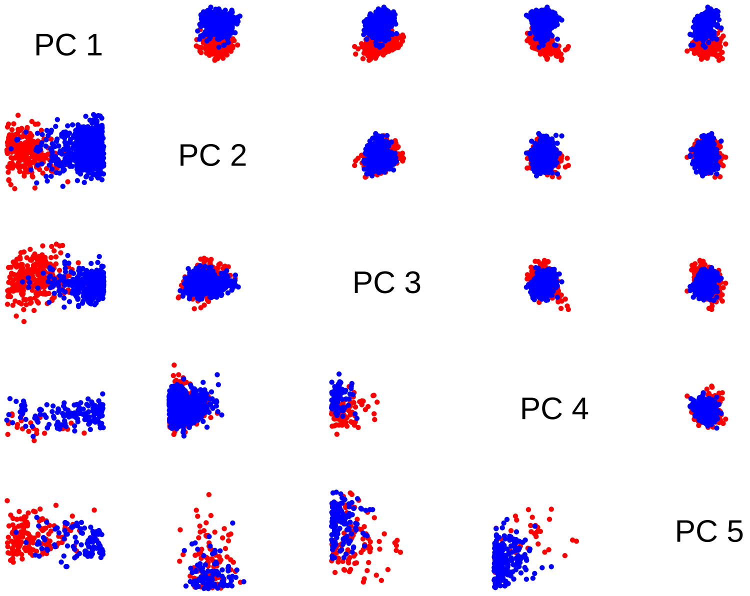

Fig. 1. Linear representation by principal component analysis. The individual sub graphs show scatter plots of the indicated principal component projections of data. The data set consists of 1210 single BOLD fMRI slices (TR = 0.333 sec) acquired in a para-axial orientation parallel to the calcarine sulcus. The subject was exposed visual stimulation in a simple block design in ten runs (stimulated scans marked as red, baseline marked as blue). The representation is determined by linear projections of maximum variance. The basis vectors are subject to strict orthogonality constraints. The baseline-activation difference is seen as a mean shift effect in several scatter plots, e.g., PC1 vs PC2. The scatter plots below the diagonal are zoomed and rotated versions of the corresponding plots above the diagonal for illustration of details in the representation.

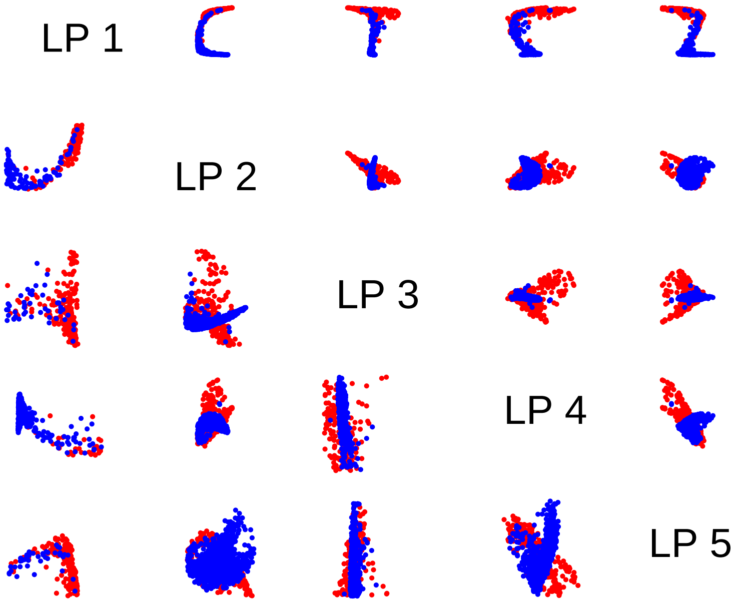

Fig. 2. Non-linear representation by Laplacian eigenmaps. The individual sub graphs show scatter plots of the indicated eigenmap projections of data. The data set consists of 1210 single BOLD fMRI slices (TR = 0.333 sec) acquired in a para-axial orientation parallel to the calcarine sulcus. The subject was exposed visual stimulation in a simple block design in ten runs (stimulated scans marked as red, baseline marked as blue). The projections are determined as eigenvectors of a neighbor diffusion matrix, hence, data is mapped together if there is a 'short-hop path' connecting via neighbors active in the same map. The baseline-activation difference is seen here as a projection *shape difference* as opposed to the mean shift effect of the linear PCA projections illiustrated in Figure 1. As in Figure 1. the scatter plots below the diagonal have been zoomed and rotated relative to the corresponding plots above the diagonal for illustration of details in the representation.

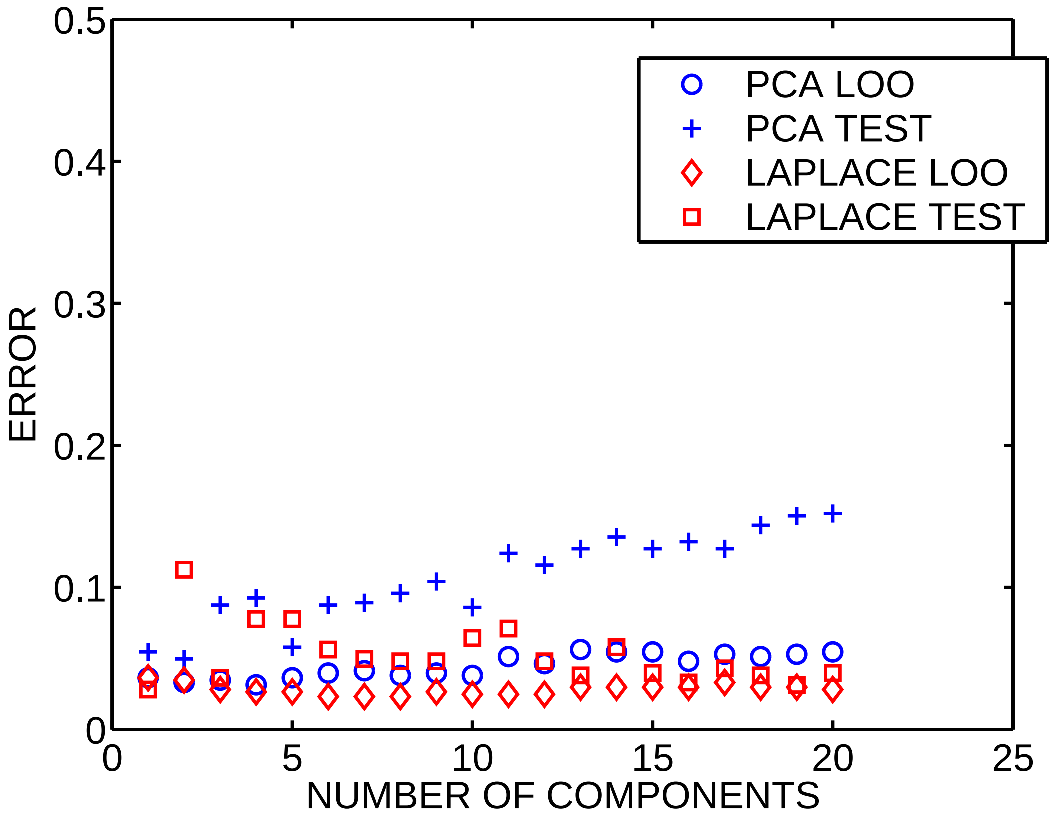

Fig. 3. Error rates for k-nearest neighbor classifiers trained on fMRI data from a single subject and generalizing to data not part of the training set. We train the classifier on the two different representations obtained by PCA and by Laplacian eigenmaps, as shown in Figures 1. and 2. We estimate the optimal number of neighbors in the voting classifier in feature space of dimensionality $d = 1 : 20$. The leave-one-out optimal dimensionality is $d = 4$ for PCA and $d = 6$ for the Laplacian eigenmap. The resulting classifiers obtain unbiased test set classification error rates 5% and 9%, in favor of the non-linear features.

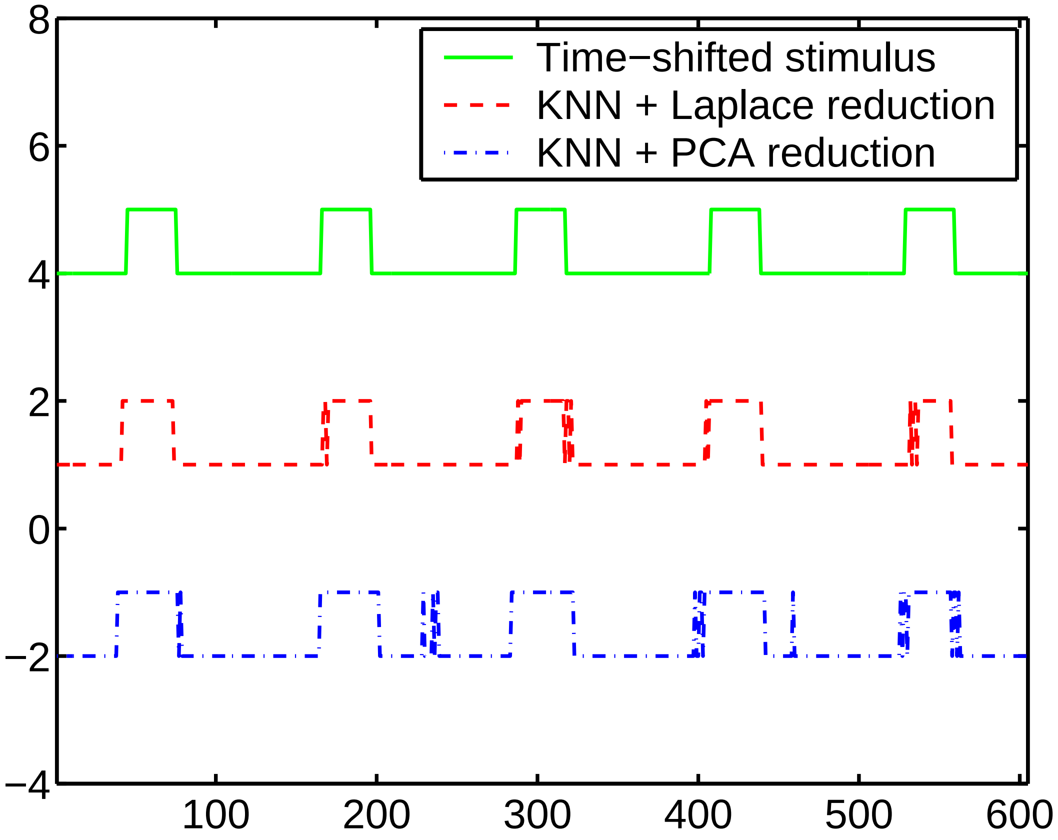

Fig. 4. Test set reference activation time course and activation time courses produced by k-nearest neighbor classifiers based on linear and non-linear feature spaces. The non-linear feature based classifier's errors basically occurs at the onset and at end of stimulation. The KNN model based on the linear feature representation make additional generalization errors in the baseline, where it suggest a few short burst of activation.