Multi-Subject fMRI Generalization with Independent Component Representation

Rasmus E. Madsen[1]

Technical University of Denmark, Kgs. Lyngby DK-2800, Denmark,
rem@imm.dtu.dk,
http://www.imm.dtu.dk/~rem

Abstract. Generalizability in a multi-subject fMRI study is investigated. The analysis is based on principal and independent component representations. Subsequent supervised learning and classification is carried out by canonical variates analysis and clustering methods. The generalization error is estimated by cross-validation, forming the so-called learning curves. The fMRI case story is a motor-control study, involving multiple applied static force levels. Despite the relative complexity of this case study, the classification of the 'stimulus' shows good generalizability, measured by the test set error rate. It is shown that independent component representation leads to improvement in the classification rate, and that canonical variates analysis is needed for making generalization cross multiple subjects.

Keywords: Independent Component Analysis (ICA), functional Magnetic Resonance Imaging (fMRI), Canonical Variates Analysis (CVA), Principal Component Analysis (PCA), Multiple Subjects.

1 Introduction

Biomedical signals, that originate from physiological processes, are in general difficult to measure isolated. Especially when non-invasive measuring techniques are used. The signals measured from the body are often a mixture of signals from different physiological processes, contaminated with noise and artifacts from the data acquisition equipment. This is also the case when we here are analyzing neuroimages, estimated by use of functional magnetic resonance imaging (fMRI). fMRI signals measured from the brain further has the disadvantage of being high dimensional and highly correlated, due to the high degree of connectivity in the brain.

From the neuroimages we seek to reveal knowledge, giving us the opportunity to model the functionality of the brain. To complete this task, it is essential to isolate the interesting macroscopic spatial and temporal patterns of brain activation, to create a reliable model. For this model to be interesting, generalizability across subjects must be adapted into the model, so one group of subjects also can be used to interpret another group of subjects. Due to the problems mentioned, the task of generating reliable generalizable models is a non-trivial task.

Multivariate statistical tools that can help us understand the brain activation patterns is therefore topic of great interest.

Independent Component Analysis (ICA) has been applied to different biomedical signals, but only recently to Functional Magnetic Resonance Imaging (fMRI) [23]. Many experiments where ICA has been applied to fMRI, has been binary experiments, where the subjects are either exposed or not exposed to some stimuli, see eg. [12]. In the experiments presented here, the subjects are exposed to different degree of stimuli. The following classification of the experiment results, therefore falls into multiple classes. At the same time the experiment is performed by multiple subjects, making classification based on group inference. We here examine how the classification generalization performance is affected by choice of an ICA representation instead of the often used representation based on PCA.

2 Functional Magnetic Resonance Imaging

fMRI is a sub-species of Magnetic Resonance Imaging (MRI) techniques. In MRI, the difference in magnetic susceptibility in different tissue in the human body, is used as a non-invasive technique, to localize different body structures. In fMRI, the difference in magnetic susceptibility, in De-oxygenated hemoglobin (HbR) and oxygenated hemoglobin (HbO_2) is used determine changes in blood-flow [11]. The Blood Oxygen Level Dependent (BOLD) contrast is the most common signal, used to determine blood-flow changes, and is therefore a indirect measurement of brain-areas with neural activity.

The measured fMRI signal has many sources, that originate from various physiological processes, including processes that are not related to experiment stimuli. The most prominent confound signal components originates from the cardiac (about 1Hz) and the physiological respiratory signal (about 0.3Hz). Artifacts from eye and body movement also influences the measured signals. The sampling frequency used in this paper and many other fMRI experiments, is well below 1Hz. This implies that some of the confound signal components becomes aliased, resulting in non-trivial temporal behavior for these confounds. On top of physiological confounds, also noise from the data acquisition equipment occur in the data.

Apart from the confounds, the signals we want to measure are not ideal. This is because the response in blood-flow to the neural active areas in the brain is not instant. The response of the blood-flow is described by the Hemodynamic Response Function (HRF) [11]. The HRF is the theoretical impulse response, that BOLD fMRI measures, when a subject is exposed to a very short stimulus. In [7] it is shown that differences in HRF time-to-peak values, varies from 2.7 to 6.2 sec. In [10] it is shown, that the HRF to the same type of subject-stimuli varies. It is also shown that the HRF may vary due to trial, site, stimulus and subject. It is not easy to make a reliable model of the HRF. There has been a lot of effort trying to model the HRF, see eg. [6], but a complete model has not yet been discovered. In this paper we try to eliminate the effect of the HRF on

our experiment signals, by simply removing the samples, where the HRF takes place. It is reasonable to believe that by removing 8 seconds of the samples, before and after subject stimuli, the effect of HRF will be eliminated. Our reason for eliminating the HRF is that neural networks applied to the fMRI data, waste too much effort trying to model the nonlinearities in the HRF, instead of the modelling the underlying stimuli function. It is reasonable to believe that other nonlinear models will have similar problems with the HRF. The confounds and the HRF makes data analysis of fMRI signals, a non-trivial task.

3 Neuroimaging data acquisition and Preprocessing

The fMRI data are acquired during a motor control study, where 16 involved subjects were doing a static force task. In the experiment the subjects were to apply a static force to a pressure gauge, using right hand thumb and index-finger. Following a visual cue, the subjects were to apply five different force levels (200,400,600,800,1000g) to the pressure gauge. The order of the force-levels was randomized. The subjects could visually monitor the force-level on the pressure gauge during the experiment. Between each force level, there was a baseline resting period. The baseline and force periods were approximately 10 TR's (TR = 4 seconds). The experiment was carried out on a Siemens 1.5T clinical scanner (fMRI: EPI BOLD, TR/TE=3986/60 m.sec., FOV=22×22×15cm, slices=30, voxels=3.44×3.44×5.00mm, MRI: T1-weighted 3D FLASH).

The scans from the 16 different subjects has been aligned, using AIR1 and AIR7 six-parameter rigid body transformation with 5th order polynomial warp to a reference MRI [20] [19]. The alignment reduces the inter-subject variance, hence increases the generalizability. The data has following been spatial smoothed with a 2D Gaussian kernel (FWHM = 0 or 6.0 pixels). The voxel time series were de-trended using linear basis of cosine basis functions.

4 Modelling

The brain activation is modelled by the relationship between the subject stimulus and the fMRI response. This is carried out by the joint probability distribution $p(X, G)$ between the microscopic variables X and the macroscopic variables G. The macroscopic variables covers the whole experimental setup that is used during the data acquisition, including the subject stimulus. The microscopic data are the observations measured during the experiment.

When modelling the joint distribution, two approaches can be chosen see eg. [18]. The joint distribution $p(X, G)$ can be factorized into either $p(X|G)p(G)$ or $p(G|X)p(X)$. With $p(X|G)p(G)$, $p(X|G)$ is modelled as a high dimensional conditional density estimate in the space of the macroscopic data. In the other approach, $p(G|X)p(X)$, the dependency $p(G|X)$ is modelled as a low dimensional conditional density estimate. In the ladder approach the dimension of X is reduced, leaving the conditional density estimate in much lower dimension than the first approach. The ladder approach has been used here.

5 Representation and data reduction

In this study, two representation (PCA and ICA) for the fMRI data are used. Both representations are obtained by modelling $p(X)$ by use of unsupervised learning, based on generative models of the form (1).

$$p(X) = \int p(X|S, A)p(S)dS \tag{1}$$

Where $p(X|S, A) = \delta(X - AS)$ is the observation model and $p(S)$ is the source distribution. For PCA, the source distributions 1st and 2nd order moments are uncorrelated. For ICA also higher order moments are uncorrelated.

Principal Component Analysis is carried out by Singular Value Decomposition (SVD). PCA applied to the $V \times T$ matrix X, where V is the number of voxels and T is the time.

$$X = U\Lambda V^T, \qquad X_{m,n} = \sum_{i=1}^{T} U_{m,i}\Lambda_{i,i}(V^T)_{i,n} \tag{2}$$

Where U is a $M \times N$ matrix, and Λ, V are $N \times N$ matrixes. Λ is a diagonal matrix containing the singular values, arranged by size. U contains the eigenvectors corresponding to the eigenvalues of XX^T, in the columns. V contains the eigenvectors corresponding to the eigenvalues of X^TX, in the rows. The dimension of X is reduced from T to K, by simply using only the K first columns of U and the first K rows of V^T as representation.

The ICA can be applied to the fMRI in either spatial or temporal domain, to produce either independent time-series or independent image components. The general ICA decomposition is defined in eq. 3, where X is a $M \times N$ matrix containing the fMRI.

$$X = AS, \qquad X_{m,n} = \sum_{i=1}^{K} A_{m,i}S_{i,n} \tag{3}$$

Where A is a matrix of image columns and S the corresponding matrix of time-series. When doing spatial ICA the columns of A becomes independent, and similarly the rows of S becomes independent when temporal ICA is applied. Temporal ICA is can be defined:

$$Y \equiv U^TX = U^TAS \equiv BS \tag{4}$$

Where Y is the $N \times N$ matrix containing the PCA time-series and S are the independent time-series. On the other hand we can define spatial ICA by the transformation:

$$Y^T \equiv V^TX^T = V^TS^TA^T \equiv (BS)^T \tag{5}$$

Here Y is the $N \times M$ matrix containing the PCA images and S are the independent images. Both transformations (Spatial and Temporal) are simple re-writings of the separation problem, and no loss of generality is generality is introduced.

The spatial and temporal ICA approaches should probably not compete against each other but could be used together. The independent time series should be used to model the paradigm. It is most likely that the independent time series will model the paradigm best. The images that are associated with the independent time series, will model multiple areas in the brain, that are active with the paradigm. The independent images will model volumes in the brain that are independent. These places are most likely isolating the functional different places in the brain, that are used during the experiment. The brain area for vision would fx. follow the experiment paradigm, but would also be influenced by the eye flickering. In [3] fMRI data was analyzed searching for temporal independent activation sequences. Here temporal ICA was able to separate two induced effects and CO_2 inhalation (hypercapnia). Since hypercapnia induces a global spatial effect, temporal independence is more appropriate than spatial independence.

ICA can be carried out by various algorithms. Different assumptions has led to multiple approaches for solving the ICA problem. In [12] three ICA algorithms were compared. The first approach is using De-correlation techniques, which was first proposed by Molgedey and Schuster [14]. The algorithm was later enhanced [15][16][22], eliminating the limitations in the original algorithm and applying a delay estimate. The second approach is the info-max algorithm [2][1]. The info-max algorithm maximizes the information-transfer through a artificial neural network (ANN), thereby separating independent components. This approach can also be seen from a maximum likelihood point of view [5]. The info-max approach needs a probability density function (PDF) estimate of the components to make a correct estimate of the independent components. In the first algorithms, the components was just expected to have the same PDF's, often super-Gaussian. Later the info-max algorithm was extended [24], enabling it to distinguish between either a super- or sub-Gaussian PDF. In the third approach, Dynamic Component Analysis (DCA) [8] [9], the assumptions from the De-correlation and information-maximization algorithms are combined into one single algorithm. The three algorithms were compared using spatial and temporal modes and shown to produce consistent spatial activation maps and corresponding time-series. There are lots of algorithms to choose from, each having different advantages over each other. In the following, the enhanced De-correlation algorithm has been used [21], due to low computational time, which has been important due to size of the fMRI data sets.

6 Subject Inference and Classification

When using PCA and ICA for preprocessing the fMRI data, inference between subjects is not achieved. Especially the ICA algorithm separates the signals associated with different subjects. The ICA algorithm makes the signals independent resulting in independent signals for different subjects. when wanting to make classification of fMRI data with multiple subjects, group inference is needed. Canonical Variates Analysis (CVA), can create group inference based

on labelling. The objective in CVA is to find a linear transformation of two data sets, x and y, so that the transformed data-sets have the largest possible correlation [13]. We here use the CVA to find the largest possible correlation between the labels and the ICA components. In figure 1 lack of group inference for the ICA components is shown together with the CVA components where inter-subject inference is present. The CVA algorithm combines the ICA components

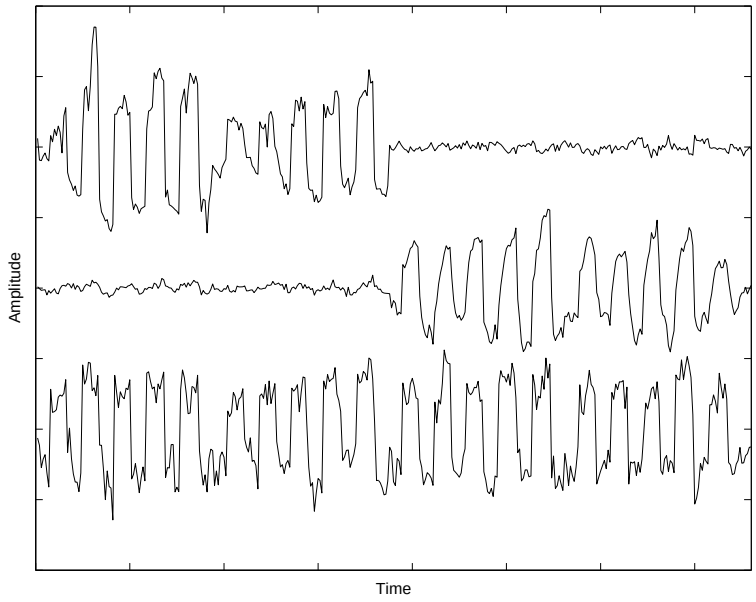

Fig. 1. CVA and ICA components. The ICA algorithm has separated similar components from different subjects. This makes it impossible to make inter-subject learning. The CVA algorithm combines the ICA components into few components that can be used for classification. The CVA algorithm combines the components by maximizing the correlation between the paradigm and ICA components.

by maximizing the correlation between the paradigm and ICA components. Full subject inference can only be achieved for the training data. Subject inference for the test data is only achieved if the training data looks similar to part of the test data. The CVA components can following be used for classification with various clustering algorithms. Due to relative few data points, we here use the N Nearest Neighbors approach, see e.g. [4].

7 Experiments

The fMRI data from the 16 subjects is arranged in a 2D data matrix, where the first dimension is the time and the second dimension is the 3D voxel image

arranged in one dimension. The subjects are stacked in the first dimension, in hope that the subject are activated in the same parts of the brain, during the experiments. The data could also be stacked in the second dimension, if the time activation patterns were expected to be the same. This is not possible in our case, since the static force task is performed in random order. Data reduction is performed by PCA, reducing the fMRI-data from 2984 to 400 components. The no. of relevant components to be found in the fMRI data was first estimated by use of the Bayesian Information Criteria [17]. The result is shown in figure 2.

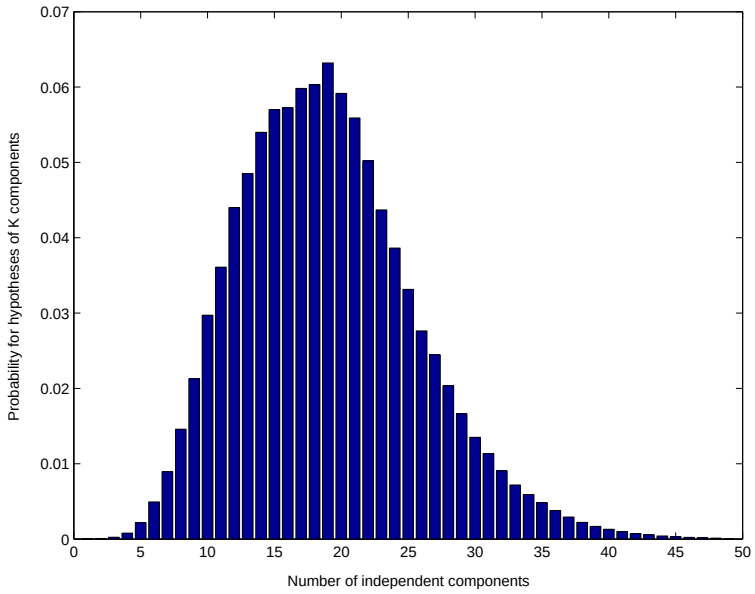

Fig. 2. The Bayesian Information Criteria BIC is applied to determine the no. of relevant components to use from the fMRI data. BIC finds approximately 20 relevant components in the fMRI data. At least 5 times the no. of components found by BIC, must be used to achieve accurate classification.

The BIC settles for about 20 components as the most optimal choice. Unfortunately the no. of components BIC finds optimal, shows very poor performance when CVA and classification is applied. At least 5 times the no. of components found by BIC, must be used to achieve accurate classification. The reason that BIC fails to find the optimal no. of components for classification, should probably be found in the fact that the ICA algorithm separates similar components that emerge in different subjects.

The optimal no. of components is instead found by estimating the bias-variance tradeoff, shown in figure 3.

When using low-dimensional representation of the data the classifier has high error-rates, because the representation is not rich enough, i.e. biased. On

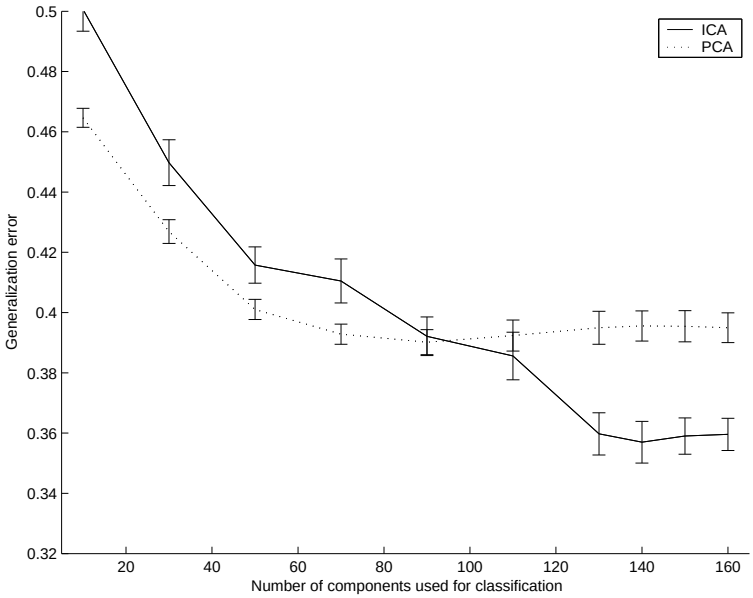

Fig. 3. Bias variance tradeoffs for ICA and PCA. The classification error rates based on test-sets is shown, when the classifier is based on ICA and PCA bases. For low dimensional bases the classifier has high error-rate because the representation is not rich enough, i.e. biased. For to high bases, for PCA D $>$ 90 and ICA D $>$ 140, the test error-rates increases because of the over-fit of the classifier in the high-dimensional representation. When using the best ICA representation, the generalization error is much lower than when using the best PCA representation. It is likely that the ICA algorithm is better suppressing the noise to the lower components, leading to enhanced generalization error.

the other hand, when representation is high-dimensional, the error-rate increase because the classifier over-fits. The best bias-variance tradeoff is approximately 90 components for the PCA representation, and 140 components for the ICA representation.

The generalization error, when using the best ICA representation is much lower than when using the best PCA representation. When using ICA, the components from 90 to 140 can be used for generalization, without over-fitting the model. It is likely that the ICA algorithm is better suppressing the noise to the lower components. This will result in more useful components which can be used to lower the generalization error.

Data spatially smoothed with a 2D gaussian kernel are compared with non-smoothed data. Two different brain warp approaches are also compared [20] [19]. Learning curves [18] for the four combinations are shown in figure 4.

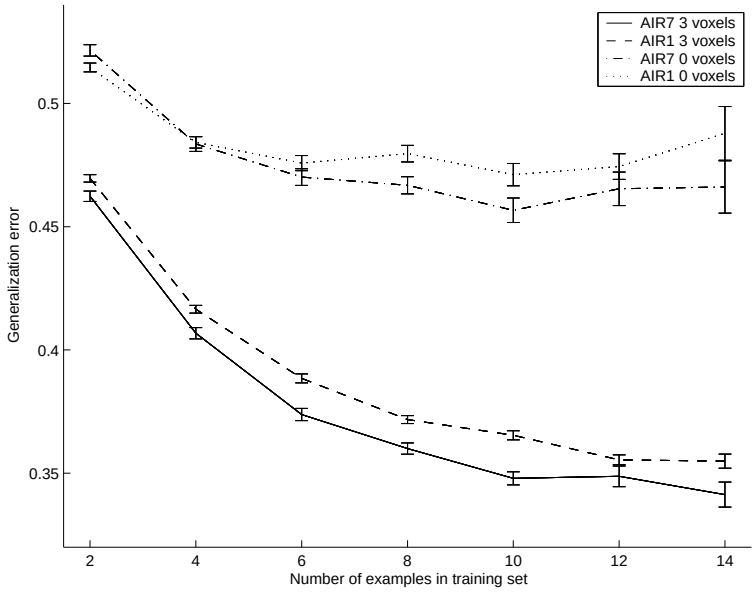

Fig. 4. Learning curves for different smoothing and registration warp method (AIR1 vs. AIR7). Without smoothing the generalization error is high. The effect of the different warp methods is limited, but AIR7 performs better than AIR1. This result was obtained with random re-sampled disjoint training- and test-sets. The classifier is based on CVA with an ICA representation of 140 basis vectors. The error-bars represent the standard deviation of the mean, and are estimated from 100 re-samples.

From figure 4, it is clear that spatial smoothing is a important parameter when making subject inference. The difference in using AIR1 or AIR7 warp technique is not dramatic. The AIR7 warp method is the best though.

Our primary objective is the question of representation. In figure 5 the learning curves for high dimensional PCA and ICA bases are shown. We are considering a six way classifier, making it interesting to see whether the classifier can distinguish between baseline and force, and following to see how well it predicts the actual force level. To be able to tell whether we can make the distinction between baseline and force, the force-level distinction line is introduced in figure 5. The error rate at the force level distinction line is $P = 0.4$, and is calculated by random selection between the force-levels.

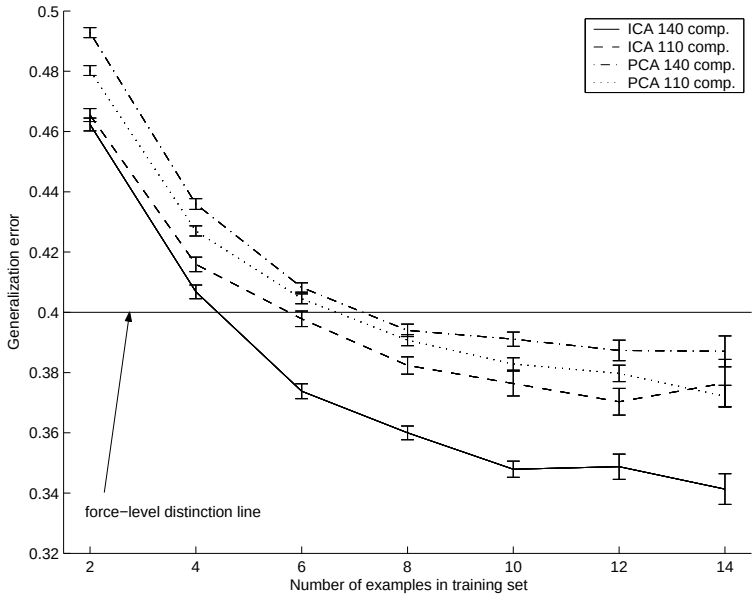

Fig. 5. Learning curves when using ICA and PCA representation with 110 and 140 components. The best performance for the PCA algorithm is found when using 110 components, and for ICA representation using 140 components. The error bars represent the std. of the mean for 100 repetitions. When using the ICA basis, the force-level distinction line is clearly passed. There are still scans though, for which baseline- and force-labels are confused.

Even though the ICA basis clearly passed the force-level distinction line in figure 5 , there are scans for which baseline- and force-labels are confused. All four representations has lower generalization error than the force-level distinction line, when using a sufficient amount of training examples, i.e. some knowledge about the force-level is preserved for all representations. Using the force-level distinction line as offset, the best ICA basis is clearly the best representation.

The distribution of the errors for the ICA representation with 140 components, is further elaborated in figure 6. Here the different types of errors that occurs in the classification experiment are shown. The figure shows that when

the classifier makes an classification error, the correct force level is typically not far away in terms of force level. To specify this, the errors are compared with baseline probabilities assuming that the classifier would only be able to distinguish between force and baseline states.

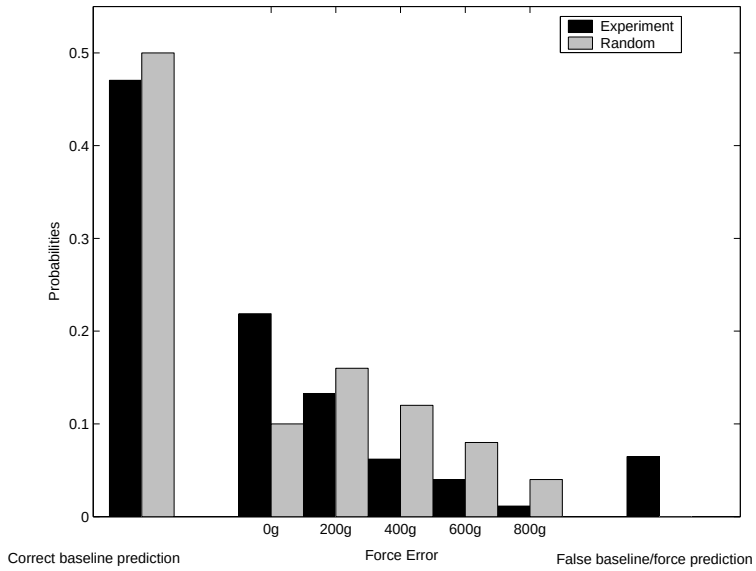

Fig. 6. Error distribution for the multi-subject experiment. The errors are compared with baseline probabilities assuming that the classifier would only be able to distinguish between force and baseline states. The re-sampling experiment is based on a CVA classifier using an ICA basis with 140 vectors. We used 14 training subjects and AIR7 warp with smoothing corresponding to 3 voxels. The 'false force/baseline' distinction is used to indicate scan classifications where the subject is in the resting baseline state but the classifier outputs a force level label or vise versa. The 'correct baseline', are the scans for which the scan is correctly estimated to be resting. The 'force error' is the difference in grams predicted by the classifier, hence zero 'force error' indicates that the force level is estimated correct.

The classifiers ability to predict the force states is further illustrated in figure 7, where the reference activation function (the 'paradigm') is compared with the classifier predictions.

In figure 8, a 3D model of the most salient spatial activation regions in the brain are shown. The regions are based on the two first CVA components. The first component forms the force baseline discriminant, and the second the force-level discriminant. The statistical means of the two CVA components are shown in figure 9

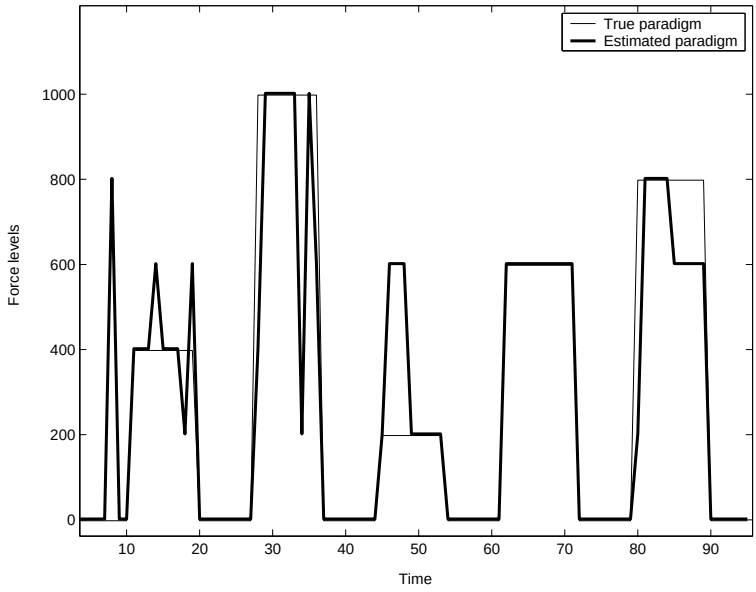

Fig. 7. The predicted activation function, compared with the actual experimental paradigm. The shown predictions are from one of the more easy subjects to predict.

8 Conclusion

Multi subject fMRI analysis based on two different representations, PCA and ICA bases, has been investigated. Based on the two representations, supervised classification been performed using Canonical Variates Analysis. We conclude that ICA allows for higher dimensional representation, providing a less biased estimate, resulting in an improved test-set classification. While the choice of representation is important, spatial smoothing and alignment by warp are still more important determinants for good generalization. We also conclude that it is important apply CVA to the ICA and PCA bases, to get group inference for generalization.

9 Acknowledgement

This work was supported in part by NHI grants NS35273 and MH57180 and by the European Union through the project MAPAWAMO.

References

1. Bell A.J. and Sejnowski T. Blind separation and blind deconvolution: An information-theoretic approach. In *International Conference on Accoustics Speech and Signal Processing (ICASSP).*, volume 5, pages 3415–3418, 1995.

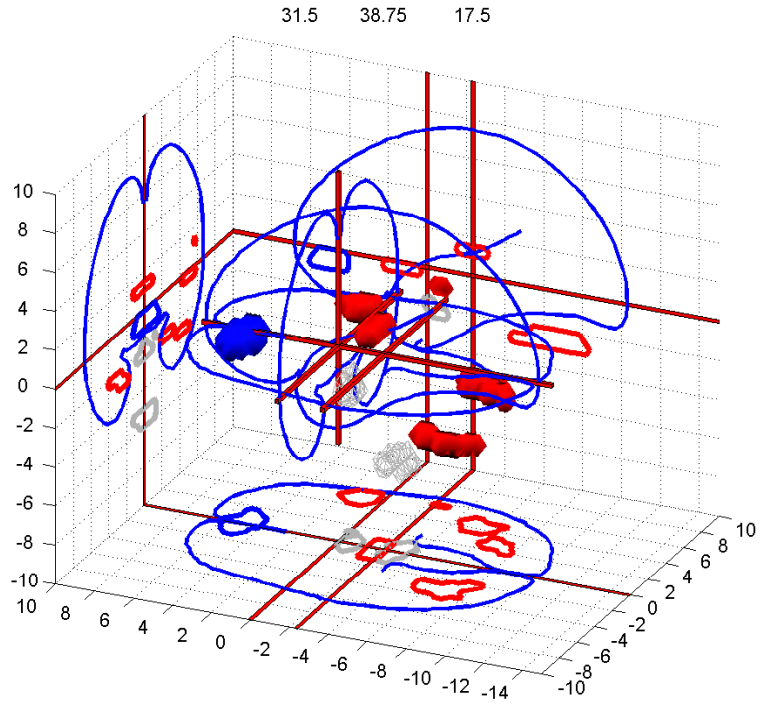

Fig. 8. Corner cube rendering of the most salient spatial pattern found by the CVA model, the dark colored locations are the most active locations for this pattern. These regions primarily encode the discrimination between force application and baseline states. The gray wire-frame regions are the most activated regions for the second most salient CVA patterns. These regions code for the amount of force applied.

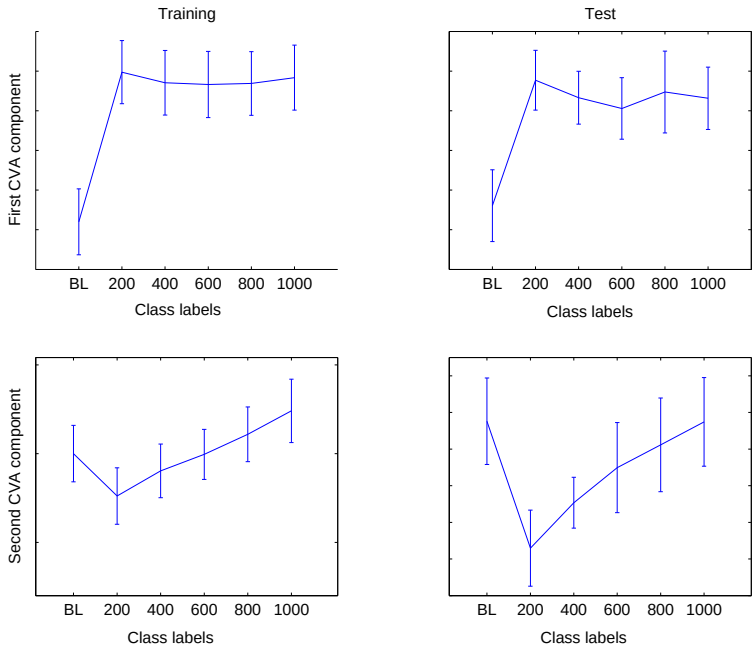

Fig. 9. The mean and std. of two most salient CVA components, for the baseline class and the five force classes. The first component is clearly discriminative between baseline and force activation. The second most salient CVA component encode for the amount of force applied. The CVA components are based on 140 ICA vectors.

2. Bell A.J. and Sejnowski T. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
3. B.B. Biswal and Ulmer J.L. Blind source separation of multiple signal sources of fmri data sets using independent component analysis. *Journal of Computer Assisted Tomography*, 23(2):265–271, 1999.
4. Bishop C.M. *Neural Networks for Patthern Recognition.* Oxford University Press., 1996. Generel book on patthern recognition methods.
5. MacKay D.J.C. Maximum likelihood and covariant algorithms for independent component analysis., 1996.
6. Marrelec G., Benali H., Ciuciu P., and Poline J.B. Bayesian estimation of the hemodynamic response function in functional mri. In Robert Fry, editor, *Bayesian Inference and Maximum Entropy Methods, Baltimore, MD, USA, MaxEnt Workshop, August*, 2001.
7. Aguirre G.K., Zarahn E., and D'Esposito M. The variability of human, bold hemodynamic responses. *Neuroimage*, 8(4):360–369, 1998.
8. Attias H. and Schreiner C.E. Blind source separation and deconvolution by dynamic component analysis. In *IEEE workshop on Neural Networks for Signal Processing*, volume 7, pages 456–465, 1997.
9. Attias H. and Schreiner C.E. Blind source separation and deconvolution: The dynamic component analysis. *Neural Computation*, 10:1373–1424, 1998.
10. Duann J.R., Jung T.P., Kuo W.J., Yeh T.C., Makeig S., Hsieh J.C., and Sejnowski T.J. Single-trial variability in event-related bold signals. *Neuroimage*, 15(4):823–835, 2002. The hemodynamic responce function (HRF) in BOLD fMRI varies. The HRF from the same type of subject-stimuli varies. It is also shown that the HRF may vary due to trial, site, stimulus and subject.
11. Friston K., Jezzard P., and Turner R. The analysis of functional mri time series. *Human Brain Mapping*, 1:153–171, 1994.
12. Petersen K.S., Hansen L.K., Kolenda T., Rostrup E., and Strother S.C. On the independent components of functional neuroimages. In *ICA-2000, Helsinki, Finland, June 22*, 2000.
13. Mardia K.V., Kent J.T., and Bibby J.M. *Multivariate Analysis.* Academic Press Limited, 1979.
14. Molgedey L. and Schuster H. Separation of independent signals using time-delayed correlations. In *Physical Review Letters*, volume 72, pages 3634–3637, 1994.
15. Hansen L.K. and Larsen J. Source separation in short image sequences using delayed correlation. In *NORSIG'98. 3rd IEEE Nordic Signal Processing Symposium*, pages 253–256. Aalborg Univ, 1998.
16. Hansen L.K., Larsen J., and Kolenda T. On independent component analysis for multimedia signals. In *Multimedia Image and VideoProcessing.* CRC Press., 2000.
17. Hansen L.K., Larsen J., and Kolenda T. Blind detection of independent dynamic components. In *International Conference on Accoustics Speech and Signal Processing (ICASSP).*, 2001.
18. Mrch N., Hansen L.K., Strother S.C., Svarer C., Rottenberg D.A., Lautrup B., Savoy R., and Paulson O.B. Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In J. Duncan and G. Gindi., editors, *Proceedings of the 15th International Conference on Information Processing in Medical Imaging*, volume 1230, pages 259–270. Springer Verlag, 1997.
19. Woods R.P., Dapretto M., Sicotte N.L., Toga A.W., and Mazziotta J.C. Creation and use of a talairach-compatible atlas for accurate, automated, nonlinear intersub-

ject registration, and analysis of functional imaging data. volume 8, pages 73–79, 1999.
20. Woods R.P., Grafton S.T., Holmes C.J., Cherry S.R., and Maziotti J.C. Automated image registration: I. general methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, 22(1):139–152, 1998.
21. Kolenda T. Mole ica matlab toolbox. mole.imm.dtu.dk/toolbox/ica/, 2002.
22. Kolenda T., Hansen L.K., and Larsen J. Signal detection using ica: Application to chat room topic spotting. In *ICA'2001, San Diego, USA, December 9-13*, 2001.
23. Jung T.P., Makeig S., McKeown M.J., Bell A., Lee T.W., and Sejnowski T.J. Imaging brain dynamics using independent component analysis. In *Proceedings of the IEEE*, volume 89, pages 1107–1122, 2001.
24. Lee T.W., Girolami M., and Sejnowski T.J. Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):409–433, 1999.