NeuroImage **15,** 772–786 (2002)
doi:10.1006/nimg.2001.1033, available online at http://www.idealibrary.com on IDEAL®

The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves

U. Kjems,*[,1] L. K. Hansen,* J. Anderson,†[,]‡ S. Frutiger,‡[,]§ S. Muley,§
J. Sidtis,§ D. Rottenberg,†[,]‡[,]§ and S. C. Strother†[,]‡[,]§[,]¶

**Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark; †Radiology Department, §Neurology Department, and ¶Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota 55455; and ‡PET Imaging Center, VA Medical Center, Minneapolis, Minnesota 55417*

Received January 20, 2001

Learning curves are presented as an unbiased means for evaluating the performance of models for neuroimaging data analysis. The learning curve measures the predictive performance in terms of the generalization or prediction error as a function of the number of independent examples (e.g., subjects) used to determine the parameters in the model. Cross-validation resampling is used to obtain unbiased estimates of a generic multivariate Gaussian classifier, for training set sizes from 2 to 16 subjects. We apply the framework to four different activation experiments, in this case [^{15}O]water data sets, although the framework is equally valid for multisubject fMRI studies. We demonstrate how the prediction error can be expressed as the mutual information between the scan and the scan label, measured in units of bits. The mutual information learning curve can be used to evaluate the impact of different methodological choices, e.g., classification label schemes, preprocessing choices. Another application for the learning curve is to examine the model performance using bias/variance considerations enabling the researcher to determine if the model performance is limited by statistical bias or variance. We furthermore present the *sensitivity map* as a general method for extracting activation maps from statistical models within the probabilistic framework and illustrate relationships between mutual information and pattern reproducibility as derived in the NPAIRS framework described in a companion paper.

Key Words: **learning curve; multisubject PET and fMRI studies; macroscopic and microscopic models; generalization error; prediction error; mutual information; cross-validation; sensitivity map.**

INTRODUCTION

In this paper we describe a prediction-based scheme for statistical modeling of functional neuroimaging data. In predictive modeling the objective is a generalizable model, i.e., a model whose parameters are estimated on a set of training samples but nevertheless can predict properties of an independently drawn set of test samples. The performance is quantified through an error function. The test performance measured as the average error on the set of test samples is an unbiased estimate of the mean test error. The mean test error is denoted the *generalization* error in the machine learning literature, while it is known as the *prediction* error in applied statistics. The prediction error typically depends on the flexibility of the invoked model and the sample size. The key element of the proposed evaluation scheme is the *learning curve*, which is the test error plotted as a function of size of the training set. In the neuroimaging context the number of data samples will typically be the number of subjects whose data participate in estimating model parameters.

Model performance is determined by the sample size and model flexibility in a manner which can be understood in a bias/variance context. A model that is too simple can fail to implement the "true" mechanisms generating the data. This gives rise to a *bias* contribution to the prediction error. On the other hand, a more flexible and complex model will tend to overfit the data points, giving rise to a *variance* contribution to the error. Using the bias/variance tradeoff we can interpret the learning curves obtained, identifying models dominated by either bias or variance. The reader is referred to Bishop (1995) or Geman *et al.* (1992) for an introduction to bias/variance decompositions and Heskes (1998) for a more general derivation. Evaluation of model generalizability has previously been studied in a functional neuroimaging context by Mørch *et al.* (1996, 1997), Mørch (1998), Hansen *et al.* (1999), and Kustra

[1] To whom correspondence should be addressed. E-mail: uk@oticon.dk.

1053-8119/02 $35.00

and Strother (2001). The approach described here differs from earlier work, most importantly by being aimed at intersubject generalization.

We give examples of *crossing* learning curves for models of variable flexibility. The generic generalization crossover occurs when a flexible model has worse performance than a biased model for small sample sizes, implying that the flexible model is the better choice only for large sample sizes. This leads to the conclusion that model "optimality" can critically depend on the amount of data available.

Similarly, we compare learning curves obtained using the same model and experiment but with different model control parameters or preprocessing procedures. Hence, the prediction framework as used here and in our companion paper is proposed as a tool for resolving model tuning and data preprocessing issues. Such issues are otherwise hard to approach in real data sets in which the activation image "ground truth" is unknown and thus ROC techniques are not available (Skudlarski *et al.*, 1999). This includes issues such as how much smoothing to apply to the image sets prior to analysis, which intra- and intersubject spatial normalization procedure to use, which model to use, and so on.

The learning curves in this paper are obtained by means of the cross-validation technique, in which the available data set is repeatedly divided into disjunct training and test subject sets. The test error is averaged over all cross-validation splits with fixed training set size, to yield an unbiased estimate of the prediction error; see Bishop (1995) for more information on the cross-validation technique. The cross-validated average of the performance is more accurate than the individual estimates because it is independent of any particular split into training/test set. With approximately 150 cross-validations for each training set size, as used in the present work, the method adds significantly to the computational demands of the analysis.

By insisting on unbiased performance estimates we avoid many of the technical difficulties and approximations related to alternative approaches based on hypothesis testing. In the classical approach a hypothesis of no activation is formed for each voxel and a resulting p map is computed, representing the level at which the local null hypothesis is rejected. In contrast, the prediction performance measures express to what degree the data are in accordance with the model: we express what the data *are* rather than what they *are not*. The prediction error framework tests the validity of model assumptions, which are assumed in the hypothesis framework. Using learning curve plots we may also evaluate the relative merits of competing hypotheses.

In the following section we present an introduction to the concepts needed to understand the principles of the prediction approach, i.e., the notion of generalization error applied to statistical models. We will emphasize that prediction error is closely related to the mutual information (Cover and Thomas, 1991) between brain images and brain information processing states (represented as task labels). In the second part of the paper, we will demonstrate learning curves applied to four different [^{15}O]water PET activation experiments and show how model performance is sensitive to various modeling parameters. We have included two technical appendices useful for any multivariate modeling approach, namely on effective use of principal component representations in a cross-validation context and on a general vizualization scheme for multivariate models called the *sensitivity map*.

THEORY

We will use $p(\,)$ to denote a probability distribution or density function for discrete or continuous stochastic variables, respectively. For simplicity we will use the term "distribution" for both. All voxels in a neuroimage can be arranged in the variable vector $\mathbf{x}$ by lexicographical ordering; $p(\mathbf{x})$ is the distribution of $\mathbf{x}$.

While the neuroimage or "scan" $\mathbf{x}$ is sometimes referred to as a *microscopic* variable, the external *macroscopic* control variable $\mathbf{g}$, the task label, quantifies the experimental conditions of the scan. The control variable may be a simple binary indicator baseline/activation or vectors (where we use boldface notation) with a more detailed description of the scan (time label), activation task (e.g., a graduated paradigm), performance data (i.e., behavioral measures of subject task performance), and subject (gender, mood, disease state, etc.). Under the General Linear Model (GLM; Friston *et al.*, 1995b), $\mathbf{g}^{(j)}$ could correspond to the jth row of the design matrix.

The micro- and macroscopic variables are, in general, multivariate stochastic variables. Consider a functional activation data set $\mathcal{D} = \{(\mathbf{x}^{(j)}, \mathbf{g}^{(j)})\}$, consisting of $j = 1, \ldots, N$ observations. A complete analysis of $\mathcal{D}$ requires the investigation of the joint distribution $p(\mathbf{x}, \mathbf{g})$ from which the data are drawn. This analysis is performed by means of a model parameterized by $\boldsymbol{\theta}$, denoted by $p_{\boldsymbol{\theta}}(\,)$. We may of course choose to disregard certain variables so that only (potentially) relevant information is represented in $\mathbf{x}$ and $\mathbf{g}$. For example, the scan should be masked for nonbrain voxels and preprocessed by stereotactic alignment and other normalization procedures. In particular we compute and subtract the mean scan of each subject as part of the preprocessing procedure (this corresponds to including subject block effects in the design matrix in the General Linear Model).

One may argue that treating the control variable $\mathbf{g}$ as a stochastic variable is an unnecessary complication. Typically, the values are decided by the experimental design; for example, the first scan is baseline and the second activation and so forth, $\{g^{(j)} | j = 1, \ldots, N\} = \{A, B, A, B, \ldots\}$; this is clearly not a random

sequence. However, from a modeling perspective it may be useful to treat these parameters as stochastic variables, so that we can contemplate predictive distributions like $p(g|\mathbf{x})$. Such a distribution tells us, for a given scan, which task label is most likely associated with a given brain scan in view of the complete set of data and model assumptions and importantly this may be different from the "true" known label, say, in the transition between two different activation subtasks.

Rather than modeling the joint distribution, $p(\mathbf{x}, \mathbf{g})$, typical functional data analysis schemes model either the predictive distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{g})$ or alternatively $p_{\boldsymbol{\theta}}(\mathbf{g}|\mathbf{x})$, which we will refer to as *micro-* and *macroscopic* models, respectively. The two model types are related through Bayes' formula,

$$p(\mathbf{x}, \mathbf{g}) = p(\mathbf{x}|\mathbf{g})p(\mathbf{g}) = p(\mathbf{g}|\mathbf{x})p(\mathbf{x}). \quad (1)$$

The distributions $p(\mathbf{x})$ and $p(\mathbf{g})$ specify the marginal distributions of scans and labels, respectively. We can assume $p(\mathbf{g})$ to be known when $\mathbf{g}$ describes data which are "certain," for example what the subject was instructed to do. What and how the subject actually *performed* while being scanned are not always available to us as measurement data. If we decide to include measured performance data and other random factors in the $\mathbf{g}$ vector, we cannot claim to know the exact shape of $p(\mathbf{g})$. This is of relevance when we design performance measures for our probabilistic models later on. The general distribution of scans $p(\mathbf{x})$ is on the other hand extremely complex and can be only crudely approximated, for example by the parameterization $p_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{g'} p_{\boldsymbol{\theta}}(\mathbf{x}|g')p(g')$.

We use Bayes' formula, Eq. (1), to convert models formulated in the microscopic domain $p_{\boldsymbol{\theta}}(\mathbf{x}|g)$ into a macroscopic formulation. This can be done when the control variable g is categorical, so that

$$p_{\boldsymbol{\theta}}(g|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, g)}{p_{\boldsymbol{\theta}}(\mathbf{x})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, g)}{\sum_{g'} p_{\boldsymbol{\theta}}(\mathbf{x}, g')} = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|g)p(g)}{\sum_{g'} p_{\boldsymbol{\theta}}(\mathbf{x}|g')p(g')}. \quad (2)$$

The parameterized model can of course be implemented in many ways; we will not even try to give an overview here. The reader is referred to the literature on pattern recognition, e.g., Duda *et al.* (2001), Cherkassky and Mulier (1998), Burges (1998), or Bishop (1999).

Micro- and Macroscopic Models

The vast majority of functional data analysis is based on microscopic type models. The General Linear Model, as invoked in SPM (Friston *et al.*, 1995a), has gained widespread use and is one example of a microscopic model. In its most basic form, the GLM is applied as parallel univariate models, estimating the model parameters independently for each voxel. One of the reasons for the popularity of the microscopic formulation is the close relation to regional hypothesis testing of the activation.

The GLM model encompasses both multiple linear regression and ANOVA/ANCOVA type models. It has a number of features which makes it easy to use for a variety of experimental setups. For example, the model can be configured directly for the known experimental setup through the design matrix, and the parameters estimated by the model can be analyzed to produce statistical maps of activations in relation to each contrast (linear combination of effects) in the design matrix. We will briefly describe the GLM and show that it can be put into a probabilistic framework.

If we form the data matrix $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}]$ and the design matrix $\mathbf{G} = [\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(N)}]^T$ we can write the GLM as the matrix equation

$$\mathbf{X}^T = \mathbf{GB} + \boldsymbol{\epsilon}^T,$$

where each scan is described as a linear function projection of a row in the design matrix $\mathbf{G}$. The maximum likelihood solution is given explicitly by $\hat{\mathbf{B}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}^T$, assuming that $\boldsymbol{\epsilon}$ is a white noise matrix with independent Gaussian elements and a separate variance estimate per voxel, $\epsilon_{ij} \in N(0, \sigma_i^2)$. The GLM assumes the noise is additive, i.e., $\mathbf{x} = \mathbf{s}(\mathbf{g}) + \boldsymbol{\epsilon}$, with $\mathbf{x}$ being the observed scan, $\mathbf{s}(\mathbf{g})$ is the underlying or "true" signal written as a function of the control variable $\mathbf{g}$, and $\boldsymbol{\epsilon}$ is a stochastic white noise process. For each noise element ϵ_{ij} at voxel i and scan number j we have

$$p_{\boldsymbol{\theta}}(\epsilon_{ij}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\epsilon_{ij}^2/2\sigma_i^2}. \quad (3)$$

The parameterized vector $\boldsymbol{\theta}$ contains $\hat{\sigma}_i^2$ and the elements of $\hat{\mathbf{B}}$, so that the probabilistic formulation of the General Linear Model reads

$$p_{\theta}(\mathbf{x}|\mathbf{g}) = \prod_{i=1}^{I} (2\pi\sigma_i^2)^{-1/2} \exp\left[-\sum_i \frac{[\mathbf{x} - \hat{\mathbf{B}}_{\mathbf{g}}^T]_i^2}{2\sigma_i^2} \right]. \quad (4)$$

A direct numerical evaluation of the above expression may not be possible since the normalization factor can be a very small or large number. In the context of the GLM the above product is appropriate for independent voxels, which is equivalent to applying a Bonferroni correction for multivoxel random field thresholding (e.g., Genovese *et al.*, 2001).

Parameter Estimation and Prediction Error

Maximum likelihood estimation of the model parameters is equivalent to maximizing the conditioned prob-

ability. In the macroscopic case this probability can be expressed as $\langle -\log p_{\boldsymbol{\theta}}(g|\mathbf{x})\rangle_{p(\mathbf{x},g)}$, where the angle brackets denote average[2] with respect to the distribution $p(\mathbf{x}, g)$. Since we do not know $p(\mathbf{x}, g)$, we will invoke the *sample distribution* provided by the training set $\mathcal{T}_{tr}$ consisting of N_{tr} examples. The sample distribution is the sum of Dirac delta functions centered at each training example:

$$p^*_{tr}(\mathbf{x}, g) = \frac{1}{N_{tr}} \sum_{\mathcal{T}_{tr}} \delta(\mathbf{x} - \mathbf{x}^{(j)}, g - g^{(j)}). \tag{5}$$

This leads to

$$\begin{aligned}\langle -\log p_{\boldsymbol{\theta}}(g|\mathbf{x})\rangle_{p(\mathbf{x},\mathbf{g})} &\simeq \langle -\log p_{\boldsymbol{\theta}}(g|\mathbf{x})\rangle_{p^*_{tr}(\mathbf{x},g)} \\ &= \int\int -\log p_{\boldsymbol{\theta}}(g|\mathbf{x}) p^*_{tr}(\mathbf{x}, g) d\mathbf{x} dg \\ &= \frac{1}{N_{tr}} \sum_{\mathcal{T}_{tr}} -\log p_{\boldsymbol{\theta}}(g^{(j)}|\mathbf{x}^{(j)}) \\ &= -\log L(\boldsymbol{\theta}; \mathcal{D}),\end{aligned} \tag{6}$$

so that the model parameter vector estimate $\hat{\boldsymbol{\theta}}$ maximizes the likelihood function $L(\boldsymbol{\theta}; \mathcal{D})$, i.e., we estimate the model parameters by maximizing the conditional distribution sampled at the data locations. It is a well-known fact that this maximization is at the risk of overfitting the data points and the model can end up learning the noise and not the underlying rule in the data (Bishop, 1995).

The prediction error $G_{\boldsymbol{\theta}}$ measures the performance on new data and can be understood as the mean conditional probability of observing new data. Put in the above terms, we form another sample distribution $p^*_{te}(\mathbf{x}, \mathbf{g})$, only this time based on the examples in the test set. This produces the prediction error estimate

$$\begin{aligned}G_{\boldsymbol{\theta}} &= \langle -\log p_{\boldsymbol{\theta}}(g|\mathbf{x})\rangle_{p(\mathbf{x},g)} \simeq \langle -\log p_{\boldsymbol{\theta}}(g|\mathbf{x})\rangle_{p^*_{te}(\mathbf{x},g)} \\ &= \frac{1}{N_{te}} \sum_{\mathcal{T}_{te}} -\log p_{\boldsymbol{\theta}}(g^{(j)}|\mathbf{x}^{(j)}),\end{aligned} \tag{7}$$

where $\mathbf{x}^{(j)}$ and $\mathbf{g}^{(j)}$ refer to the jth scan and label of the test set. We note that the prediction error is computed by feeding the examples of the test set through the same cost function expression that was used when optimizing for the model parameters. Likewise, the definition of $G_{\boldsymbol{\theta}}$ for microscopic models uses the average $\langle -\log p_{\boldsymbol{\theta}}(\mathbf{x}|g)\rangle_{p(\mathbf{x},g)}$.

Other test-error measures could be created to compute alternative prediction errors. Consider a classifier (which is a macroscopic type model, $p(g|\mathbf{x})$, where g is discrete and one dimensional): If we use the output class probabilities to classify (i.e., choosing the label with highest probability), we can count the number of false model class decisions. The error rate is then the fraction of incorrectly classified test examples.

While the log-probability measure is consistent with likelihood-based estimation, it can cause problems when the model generates output probabilities of zero or very close to zero. We will regard this as a problem of the model and/or test set sample size rather than the error measure itself; a predicted probability of zero is indeed a very strong statement. The most common problem arises in connection with *outlier samples*, which for some reason (e.g., incorrect measurements) behave completely different from the rest of the samples. Models which use Gaussian assumptions for the noise can be heavily affected by such outliers. We shall not delve deeper into this subject here, but refer to Hintz-Madsen *et al.* (1995) for a treatment of outliers in a neural network classifier context.

The Prediction Error as Mutual Information

We will now address the problem of interpretability of the prediction error measure $G_{\boldsymbol{\theta}}$ of Eq. (7), by demonstrating how it can be interpreted as the *mutual information* between scans and scan labels. The mutual information in the context of classifiers have been pursued by Hertz *et al.* (1995) for analysis of the transmitted information in neuronal single-cell recordings. We will measure performance in units of bits, which corresponds to the information contained in a single, balanced probability, binary decision.

The mutual information (MI) between scans and labels is defined as the Kullback–Leibler divergence between the joint distribution and the product of the marginals:

$$\mathrm{MI} = \left\langle \log_2 \frac{p(\mathbf{x}, g)}{p(\mathbf{x})p(g)} \right\rangle_{p(\mathbf{x},g)}. \tag{8}$$

We can use either the macro- or the microscopic formulations:

$$\begin{aligned}\mathrm{MI}_{\boldsymbol{\theta}} &= \left\langle \log_2 \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, g)}{p_{\boldsymbol{\theta}}(\mathbf{x})p(g)} \right\rangle_{p(\mathbf{x},g)} \\ &= \left\langle \log_2 \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|g)}{p_{\boldsymbol{\theta}}(\mathbf{x})} \right\rangle_{p(\mathbf{x},g)} \\ &= \left\langle \log_2 \frac{p_{\boldsymbol{\theta}}(g|\mathbf{x})}{p(g)} \right\rangle_{p(\mathbf{x},g)}.\end{aligned} \tag{9}$$

In practice we approximate the above expressions by averaging over the test examples. For the macroscopic model we have

[2] This is to understand that $(\mathbf{x}, g)$ are *drawn* infinitely many times from $p(\mathbf{x}, g)$ and the expression inside the angle brackets is averaged.

$$\begin{aligned} \mathrm{MI}_{\boldsymbol{\theta}} &\simeq \frac{1}{N_{te}} \sum_{\mathcal{T}_{te}} \log_2 p_{\boldsymbol{\theta}}(g^{(j)}|\mathbf{x}^{(j)}) - \sum_g p(g)\log_2 p(g) \\ &= -G_{\boldsymbol{\theta}}/\log 2 - \sum_g p(g)\log_2 p(g). \end{aligned} \tag{10}$$

Thus, the $\mathrm{MI}_{\boldsymbol{\theta}}$ is directly related to the prediction error $G_{\boldsymbol{\theta}}$ of Eq. (7) through a simple linear scaling and offset. With $p_{\boldsymbol{\theta}}(\mathbf{x}) = \Sigma_{g'} p_{\boldsymbol{\theta}}(\mathbf{x}|g')p(g')$ we have for microscopic models

$$\begin{aligned} \mathrm{MI}_{\boldsymbol{\theta}} &\simeq \frac{1}{N_{te}} \sum_{\mathcal{T}_{te}} \log_2 \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}|g^{(j)})}{\Sigma_{g'} p_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}|g')p(g')} \\ &= \frac{1}{N_{te}} \sum_{\mathcal{T}_{te}} \log_2 \frac{p_{\boldsymbol{\theta}}(g^{(j)}|\mathbf{x}^{(j)})}{p(g^{(j)})}, \end{aligned} \tag{11}$$

where the last equality shows that the microscopic model actually should be converted into a macroscopic model, recall Bayes' relation in Eq. (1), before the MI is computed.

As discussed earlier we may in fact know $p(g)$, for example, when g describes a univariate activation label determined by the experimental design.[3] The above expressions measure the amount of information about g that the model can extract from a single scan in excess of what we already know. Consider a model which can give us no further information. This model will have $p_{\boldsymbol{\theta}}(g|\mathbf{x}) = p(g)$ and we will in a sense have learned zero bits from the scans. Note that it is possible for the mutual information to be negative,[4] for example a model which assigns too small probabilities to test examples, something that easily occurs in connection with models that overfit the data. A macroscopic model with negative MI can always be regularized so that the performance is at least zero bits of MI by simply regularizing the predictions toward the prior $p(g)$.

On the other hand, a model which is as close to the truth as possible, $p_{\boldsymbol{\theta}}(g|\mathbf{x}) = p(g|\mathbf{x})$, will achieve the upper bound on $\mathrm{MI} = \langle \log_2 p(g|\mathbf{x})/p(g) \rangle_{p(\mathbf{x},g)}$, which is the true mutual information between scans and labels. We cannot in practice compute this upper bound since we don't know the true distribution. However, by inserting $p_{\boldsymbol{\theta}}(g|\mathbf{x}) \leq 1$ into Eq. (9) we see that there is another upper limit,

$$\mathrm{MI}_{\boldsymbol{\theta}} \leq \left\langle \log_2 \frac{1}{p(g)} \right\rangle_{p(\mathbf{x},g)} = -\sum_g p(g)\log_2 p(g), \tag{12}$$

i.e., the label entropy. These bounds are useful for interpretation of measured mutual information along with the result explained below for comparison of mutual information from different label schemes.

Mutual Information for Hierarchical Label Sets

We will now illustrate that the MI measure is directly comparable across different label schemes. For example, consider an activation/baseline type experiment with eight scans per subject. We may choose to label each scan according to baseline/activation state $g \in \{A, B\}$, i.e., *ABABABAB*, or we may choose *agnostic* labeling, i.e., label each scan separately, so that the eight scans are labeled $A', \ldots, H'$ (see, e.g., Strother *et al.*, 1996; Kjems *et al.*, 1999; Frutiger *et al.*, 2000). These two labeling schemes are hierarchically related with the mapping $A' \rightarrow A$, $B' \rightarrow B$, $C' \rightarrow A$, $\ldots$, $H' \rightarrow B$. Yet the mutual information obtained from the two-label modeling experiment can be directly compared to the mutual information obtained from an eight-label experiment.

To see this, consider two different label sets $\{g\}$ and $\{g'\}$ with a mapping $g' \rightarrow g$ represented in the table $g(g')$, with $n(g)$ different g' labels mapping onto the specific label g (in above example, $n(A) = n(B) = 4$). We can now translate a model of g labels into a model of g' labels by stating that with $g = g(g')$, $p(g') = p(g)/n(g)$ and likewise for our model predictions $p_{\boldsymbol{\theta}}(g'|x) = p_{\boldsymbol{\theta}}(g|x)/n(g)$. Inserting this into Eq. (9) we see that MI for the translated model $p_{\boldsymbol{\theta}}(g'|x)$ is

$$\begin{aligned} \mathrm{MI}_{p_\theta(g'|x)} &= \left\langle \log_2 \frac{p_{\boldsymbol{\theta}}(g'|\mathbf{x})}{p(g')} \right\rangle_{p(\mathbf{x},g')} \\ &= \left\langle \log_2 \frac{p_{\boldsymbol{\theta}}(g|\mathbf{x})/n(g)}{p(g)/n(g)} \right\rangle_{p(\mathbf{x},g')} \\ &= \left\langle \log_2 \frac{p_{\boldsymbol{\theta}}(g|\mathbf{x})}{p(g)} \right\rangle_{p(\mathbf{x},g)} \\ &= \mathrm{MI}_{p_\theta(g|x)}. \end{aligned} \tag{13}$$

In other words, the predictions of the two-label model from the above example can be translated into agnostic labels by spreading the predictions of labels A and B evenly on labels A', C', E', G' and B', D', F', H', respectively, and this translated model will have the exact same MI.

We can therefore view the two-label model as a less flexible (more biased) instance of the eight-label model. In addition, as a result of the bias/variance tradeoff the biased two-label model is expected to perform better than the flexible eight-label model for small training set sizes, while the flexible model should be expected to match the performance or better for large training set sizes.

[3] If $p(g)$ cannot be assumed known one would have to estimate the distribution using some model and plug the result into Eq. (10).

[4] Actually there is no lower bound on $\mathrm{MI}_{\boldsymbol{\theta}}$: if $\mathbf{x}$ and $\epsilon > 0$ exist so that $p_{\boldsymbol{\theta}}(g|\mathbf{x}) \rightarrow 0$ and $p(g|\mathbf{x}) > \epsilon$, we will have $\mathrm{MI}_{\boldsymbol{\theta}} \rightarrow -\infty$.

Model Visualization

Once the validity of the model has been established by the learning curve, we address the problem of visualizing the relationship between scan $\mathbf{x}$ and label g as identified by the mathematical model $p_{\theta}(g|\mathbf{x})$ (or $p_{\theta}(\mathbf{x}|g)$). We propose a general procedure for extracting activity maps, namely the *sensitivity map*, which we will define as

$$s_i = \left\langle \left[\frac{\partial p_{\theta}(g|\mathbf{x})}{\partial x_i} \right]^2 \right\rangle_{p(\mathbf{x}, g)}. \quad (14)$$

The sensitivity map measures how much the class predictions change when the ith voxel is modified. As before, the angle brackets denote averaging with respect to the (unknown) true distribution. In practice the sensitivity map is computed using a finite sum over examples (please refer to Appendix B for a derivation for the CVA classifier model),

$$s_i \simeq \frac{1}{N} \sum_j \frac{[\partial p_{\theta}(g^{(j)}|\mathbf{x}^{(j)})]^2}{\partial x_i}. \quad (15)$$

Because of its quadratic form the map will be positive, i.e., $s_i \geq 0$ for all voxels. The objective of this map is to identify the brain regions that are most relevant to the model's predictions. There is no sign information present in the map; therefore it is not possible to distinguish regions with relatively increased or decreased flow, as is possible with the statistical parametric maps in connection with GLM or with the canonical eigenmaps obtained from canonical variate analysis of Eq. (19) (CVA; see Mardia and Kent, 1979; Kjems *et al.*, 1999). Such information can potentially be gathered post hoc, for example, by investigating the particular structure of the signal inside those regions identified by the sensitivity map.

When visualizing using CVA eigenmaps, we obtain $k - 1$ maps when there are k different values of g. The researcher is thus left with the problem of combining (and comprehending) multiple activation maps. The sensitivity map is a principled way to achieve a single map for visualization. In general, the interpretation of the sensitivity map is *the following areas of the brain are found relevant by the model* $p_{\theta}(g|\mathbf{x})$ *for discriminating between brain states* g. The sensitivity map further has the advantage that it is defined for all models in the macroscopic formulation, and the sensitivity maps associated with different models can be readily compared. Using the NPAIRS framework in the companion paper (Strother *et al.*, 2002), the sensitivity map also can be transformed into a Z map, i.e., a map with which we can form regional hypotheses about activation.

METHODS

Subjects

Seventy-four normal right-handed volunteer subjects were each scanned performing one of four left-handed motor tasks after written informed consent was obtained in accordance with a protocol approved by the Minneapolis VA Medical Center's Institutional Review Board. Subjects with a history of substance abuse or of a neurologic, medical, or psychiatric disorder were eliminated from the subject pool. Prior to PET scanning subjects underwent a complete neurologic examination and were administered the Edinburgh Handedness Inventory to verify right-hand dominance. All female subjects of child-bearing age had a prescan serum pregnancy test.

Data Acquisition and Quality Control

All [^{15}O]water PET scans were acquired with a Siemens ECAT 953B-31 scanner operating in its 3D mode (10.8-cm axial field of view, with reconstructed in-plane and axial resolution of 8.5 and 6 mm, respectively, on a 128 × 128 × 31-voxel grid with 3.125 × 3.125 × 3.375-mm^3 voxels). Infusion of a 13-mCi [^{15}O]water bolus initiated task or control trials, which were separated by 7 to 10 min, and a 90-s scan was triggered when radioactivity reached the brain. PET counts were corrected for dead time, randoms, and attenuation and were reconstructed using 3D-filtered back-projection (Liow *et al.*, 1997). After reconstruction scans from each scanning session were visually examined and excluded for image artifacts or poor positioning within the axial field-of-view with inadequate coverage of sensorimotor cortex, anterior parietal area, and superior cerebellum. Compared to the relatively strict empirical criteria used for the allowable brain coverage of the data sets in Strother *et al.* (2002) the criteria were relaxed somewhat to allow larger data sets to pass the quality control screen. Following this initial quality control screen the alignment, thresholding, and smoothing steps were identical to those detailed in Strother *et al.* (2002).

Tasks

Tracing (TR)

Eighteen volunteers were scanned while using a joystick with their left hand to trace a path along the perimeter of a six-pointed star displayed on a rear-projection screen at the foot of the PET scanner couch. Scanning sessions contained 1 baseline scan (no tracing, eyes open viewing the screen, ears plugged, resting quietly), followed by 8 tracing scans and a final baseline scan for 10 scans/session. See Frutiger *et al.* (2000) for further details of this data set.

Finger Opposition (FO)

Twenty volunteers were scanned while performing left-handed sequential opposition of the thumb and successive digits (2, 3, 4, 5, 4, 3, 2, 3, . . .), externally paced with a 1-Hz auditory signal. Scanning sessions contained 4 to 5 alternating baseline (resting quietly with eyes covered and ears plugged) and activation scans for 8 to 10 scans per session. See Kustra (2000) and Kustra and Strother (2001) for related analysis of finger-opposition data sets.

Static Force (SF3)

Eighteen volunteers were scanned with 1 baseline followed by two blocks of 5 static-force activation scans/block and a final baseline scan for 12 scans/session. Activation consisted of static force, exerted on a load cell using the thumb and index finger of the right hand, which controlled the cursor displayed on a rear-projection screen at the foot of the PET scanner couch. Before scanning subjects were practiced to criterion, keeping the cursor (force) within preset limits (lines on the screen) about a target-force level (central line). Target force levels of 200, 400, 600, 800, and 1000 g were each used once in randomized order within each block. See Muley *et al.* (2001) for further details of this data set.

Mirror Tracing (MT)

Eighteen volunteers were scanned while performing a modification of the tracing task described above. Scanning sessions contained 2 standard left-handed tracing scans—after the subject had performed the tracing task six times in the scanner—followed by 8 mirror tracing scans with the vertical cursor–hand movement feedback reversed, for a total of 10 scans/session; subjects performed an additional mirror tracing trial in each 8-min interval between scans. The first 4 scans (2 tracing and 2 mirror tracing) were chosen for this study. Preliminary results from this data set have been reported by Frutiger *et al.* (1998).

Modeling Extremely Ill-Posed Data Sets

Imaging data sets usually have more voxels, I, than there are scans, N. The space of all possible observations $\mathcal{I}$ is of dimension I, while the actual observations in the data set span the *signal space* $\mathcal{S}$ which at most can be of dimension N. Typically $\dim(\mathcal{S}) \ll \dim(\mathcal{I})$, making $\mathcal{S}$ a small subspace of $\mathcal{I}$. This is exactly what characterizes extremely ill-posed data sets. Because the dimension of $\mathcal{S}$ is low we have a correspondingly low number of degrees of freedom available in any subsequent modeling.

It is possible, however to reduce the dimensionality of the observation space using Singular Value Decomposition (SVD), see e.g. Press *et al.* (1986), Strother *et al.* (1995), Golub and Loan (1989), and Lautrup *et al.* (1995). In brief SVD is the matrix decomposition of the $I \times N$ matrix of voxels-by-scans $\mathbf{X}$, $N \ll I$,

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T, \tag{16}$$

where $\mathbf{U}$ contains N orthogonal vectors $\mathbf{U}^T\mathbf{U} = \mathbf{I}_N$, Λ is a diagonal with the singular values $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, and $\mathbf{V}$ is an orthogonal $N \times N$ matrix $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Since the vectors of $\mathbf{U}$ span the same space as the vectors of $\mathbf{X}$, we can choose to work in the basis represented by $\mathbf{U}$, with no loss of information:

$$\mathbf{Q} = \mathbf{U}^T\mathbf{X}. \tag{17}$$

Each column of $\mathbf{Q}$ contains the coordinates of the corresponding column (scan) in $\mathbf{X}$, but does this with only N numbers which represents a significant saving in memory. The large matrix multiplication in Eq. (17) can also be avoided since $\mathbf{Q} = \Lambda\mathbf{V}^T$ is obtained directly from the SVD operation on $\mathbf{X}$.

The cross-validation technique used in this paper requires the division of the data set into independent training and test parts. The procedure outlined above is valid for generating training sets only, since the computation of the basis set $\mathbf{U}$ can be considered part of the training procedure also. This means we cannot simply distribute the columns of the matrix $\mathbf{Q}$ in Eq. (17) in training and test sets—this would violate the independence of the test set. Instead, we should compute the basis on the training set only, with $\mathbf{X}_{tr} = \mathbf{U}_{tr}\Lambda_{tr}\mathbf{V}_{tr}^T$ and the training set projection becomes $\mathbf{Q}_{tr} = \Lambda_{tr}\mathbf{V}_{tr}^T$ while the test set projection onto the basis defined by the training set is $\mathbf{Q}_{te} = \mathbf{U}_{tr}^T\mathbf{X}_{te}$. In Appendix A we explain how to construct such basis projections for the cross-validations in a manner which is both numerically and computationally efficient using a two-step SVD procedure.

CVA Model with Sensitivity Map

For the experiments in this paper we adopt a multivariate Gaussian model, with a Gaussian covariance structure within each group, pooled across groups. This model "lives" in a reduced dimensionality space defined by a CVA. This analysis is again built on top of the P first singular directions ($P \leq N$, we chose $P \sim 4$–12 depending on experiment). With the terminology introduced in Appendix A, we use the top P rows of the matrix $\mathbf{Q}_{tr}$ forming a new matrix $\mathbf{Q}_{tr}^*$ (likewise for the $\mathbf{B}_{0,tr}$ matrix we use the first P columns forming $\mathbf{B}_{0,tr}^*$). The number of dimensions to keep, P, has been chosen identical to the number of scans per subject which is between 4 and 12 in the data sets used here. The reason for making this choice is that we expect the "interesting" structures in signal space to be of lower or the same dimension as the number of scans per sub-

ject. Please note that we expect interesting activation-related effects to appear in the very first singular directions, which is different from most of this earlier work (Strother *et al.*, 1995; Mørch, 1998) in which an N subject study had the first $N - 1$ components dominated by intersubject effects. This is not the case here because each subject's mean scan is subtracted from each scan before the SVD processing.

There is, of course, no guarantee that the signal we are looking for actually appears in the top 4–12 singular directions although our experimental results seem to justify this. Friston *et al.* (1996) have argued for the use of all components whose variances are larger than the average variance. We found that this procedure includes too many components, with decreased model performance as a result. Furthermore, our aim is that the learning curves are created with as few parameters changed as possible as the training set size varies.

The reduced data matrix $\mathbf{Q}^*_{tr}$ contains one scan per column. Counting scans with the same label, $\mathbf{q}_{j,g}$ is the jth column of $\mathbf{Q}^*_{tr}$ (each column represents a scan) which has label g, and let there be N_g scans in total with label g. We next perform a CVA to obtain a linear subspace that separates the $\mathbf{q}_{j,g}$ the most (across groups) while removing within-group covariance: In short (please refer to Mardia and Kent, 1979, Ch. 12, or Kjems *et al.*, 1999, or Worsley *et al.*, 1997, for further details) this proceeds by forming within- and between-group SSP matrices, all of size $P \times P$:

$$\begin{aligned}\mathbf{W} &= \sum_{j,g} (\mathbf{q}_{j,g} - \bar{\mathbf{q}}_g)(\mathbf{q}_{j,g} - \bar{\mathbf{q}}_g)^T,\\ \mathbf{T} &= \sum_{j,g} (\mathbf{q}_{j,g} - \bar{\mathbf{q}})(\mathbf{q}_{j,g} - \bar{\mathbf{q}})^T = \mathbf{W} + \mathbf{B}, \text{ and}\\ \mathbf{B} &= \sum_{g} N_g(\bar{\mathbf{q}}_g - \bar{\mathbf{q}})(\bar{\mathbf{q}}_g - \bar{\mathbf{q}})^T.\end{aligned} \tag{18}$$

The CVA eigendirections are computed as the $K - 1$ eigenvectors $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_{K-1}]$ with nonzero eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$. Since CVA considers the distribution of group means, the K group means will maximally span a $K - 1$ dimensional subspace spanned by the canonical eigenvectors.

Now, through the initial SVD the CVA eigenvectors $\mathbf{l}_k$ each have an associated map in voxel space (see Appendix A),

$$\mathbf{m}_k = \mathbf{U}\mathbf{B}^*_{0,tr}\mathbf{l}_k. \tag{19}$$

As discussed earlier this map can be visualized and given interpretation if held together with the scan's projections onto the canonical directions, the so-called canonical coordinates:

$$\mathbf{c}_{j,g} = \mathbf{L}^T\mathbf{q}_{j,g} = \mathbf{L}^T\mathbf{B}^*_{0,tr}\mathbf{U}^T\mathbf{x}_{j,g}. \tag{20}$$

Returning to our probabilistic model, which we set out to formulate, we now construct a classical Gaussian classifier on top of the CVA. The model is a Gaussian per group, with common covariance, namely the within-group structure $\mathbf{W}$ as identified in the CVA,

$$p_{\boldsymbol{\theta}}(\mathbf{c}|g) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\|\mathbf{c} - \bar{\mathbf{c}}_g\|^2\right], \tag{21}$$

with $\mathbf{c}$ given by Eq. (20). At this point it would most likely be advantageous to reduce the dimensionality of $\mathbf{c}$ by retaining only a subset of the canonical directions, namely those that test significant (cf. Mardia and Kent, 1979, p. 341). Meanwhile, for the sake of simplicity we shall not consider such an approach here.

The model parameter vector $\boldsymbol{\theta}$ contains the canonical vectors $\mathbf{L}$, the basis vectors $\mathbf{B}^*_{0,tr}$, and the group means $\bar{\mathbf{q}}_g$. The covariance matrix in Eq. (21) is a diagonal unit matrix because the canonical coordinates of $\mathbf{c}$ are uncorrelated and have unit variances.

As it stands, Eq. (21) is neither a microscopic nor a macroscopic model because it considers only a small subspace of the input space. We need an additional noise model to model the remaining part of signal space. Assume that there is a mapping $\mathbf{x} \leftrightarrow (\mathbf{n}, \mathbf{c})$ and that the activation signal is independent of the noise,

$$p_{\boldsymbol{\theta}}(\mathbf{x}|g) = p_{\boldsymbol{\theta}}(\mathbf{n})p_{\boldsymbol{\theta}}(\mathbf{c}|g). \tag{22}$$

This expression can be turned into a classifier using Bayes' relation (1),

$$p_{\boldsymbol{\theta}}(g|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{n})p_{\boldsymbol{\theta}}(\mathbf{c}|g)p(g)}{\sum_{g'} p_{\boldsymbol{\theta}}(\mathbf{n})p_{\boldsymbol{\theta}}(\mathbf{c}|g')p(g')} = \frac{p_{\boldsymbol{\theta}}(\mathbf{c}|g)p(g)}{\sum_{g'} p_{\boldsymbol{\theta}}(\mathbf{c}|g')p(g')}, \tag{23}$$

which does not depend on the choice of noise model $p_{\boldsymbol{\theta}}(\mathbf{n})$. This defines the multivariate CVA as it is used as a classifier in the experiments in this paper.

Pattern Reproducibility

While the generalizability as measured by the mutual information can be used to evaluate model performance and to compare different models' performance from the mutual information, we cannot infer the quality of the associated model visualization, say, the quality of the sensitivity map.

Evaluating the generalizability of the visualization requires a ground truth image, hence, is available only for simulated data (see, e.g., Skudlarski *et al.*, 1999). For real data we can evaluate the variance part of the generalizability, by computing the reproducibility of the produced visualization. But there is no direct way of estimating the bias part. The visualization reproducibility can be obtained within the NPAIRS framework

TABLE 1

Labeling Schemes for the Four Modeling Setups

	Setup 1	Setup 2	Setup 3	Setup 4
Mirror tracing	1122 4 PC's	1234 4 PC's	1 . . . 10 4 PC's	1 . . . 10 10 PC's
Static force	1221 4 PC's	1234 4 PC's	1 . . . 12 4 PC's	1 . . . 12 12 PC's
Finger opposition	1212 4 PC's	1234 4 PC's	1 . . . 8 4 PC's	1 . . . 8 8 PC's
Tracing	1221 4 PC's	1234 4 PC's	1 . . . 10 4 PC's	1 . . . 10 10 PC's

described in the companion paper (Strother *et al.*, 2002) by measuring the correlation coefficient (over voxels) of two maps derived from independent data sets of the same size. The subjects are then permuted many times (we used 150 pairs) to get an improved estimate of the reproducibility rate. Because the patterns need to be computed from disjoint examples, the reproducibility learning curves can be computed with at most $N/2$ subjects in the training set, whereas the test error measures were computed with up to $N - 1$ subjects.

Data Sets and CVA Labeling Schemes

The data matrix for each data set was processed by: (1) dividing each voxel value by the average value across all voxels inside the brain mask and then (2) for each voxel in each subject, subtracting the average value across the subject's scans. This preprocessing strategy creates the $I \times N$ matrix of voxel by scans $\mathbf{X}$ ($N \ll I$) and was designed to maximize sensitivity to within-subject effects while removing individual subject effects.

Each of the four data sets was analyzed with the CVA model described under CVA Model with Sensitivity Map and the misclassification error rates and MI were assessed by means of the cross-validation scheme under Modeling Extremely Ill-Posed Data Sets. For each size of the training set we computed 150 (190 for the FO data set) cross-validations by permutation of the selection of subjects for the training and test sets. Learning curves were computed using four different setups on the four data sets; see Table 1.

We created a smaller data set by selecting only four scans per subject with a balanced number of baseline and active scans: In MT we used the first two tracing scans and the first two mirror tracing scans. In SF we used the first baseline, the first two force scans, and the last scan, which was a baseline. For the FO data set we selected the first four alternating scans, and for TR we used the first baseline, the first two tracing scans, and the last baseline scan. These are the same data sets analyzed in Strother *et al.* (2002). The balancing was done because the CVA model uses a common covariance matrix which for some label setups may not match the data well in the presence of strong temporal trends. Such strong trends may inflate the within-group covariance when single-label classes cover multiple scans, causing poorer model performance.

The first two of the four modeling setups used the smaller balanced data set with either baseline/activation labels or *agnostic* labels, i.e., separate labels for the four scans. The last two setups were done with all scans and agnostic labeling, this time varying the number of principal components (PC's) P retained in the basis projection. We used the first four PC's in the third setup and the same number of components as the number of scans per subject in the fourth component.

RESULTS AND DISCUSSION

To illustrate the function of CVA, Fig. 1 presents a scatter plot of the canonical coordinates for the MT experiment. Each scan is represented by a letter or

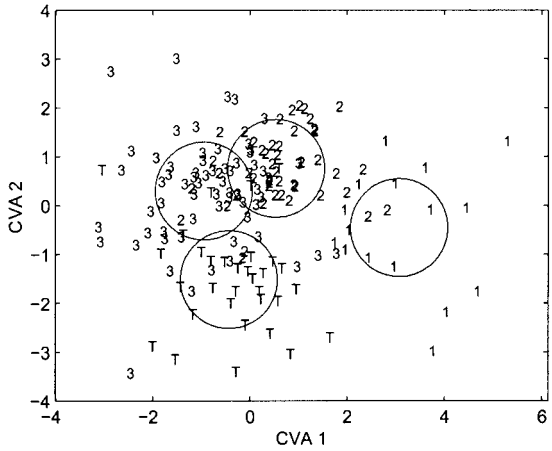

FIG. 1. Scatter-plot visualization of canonical coordinates for the mirror tracing experiment (see Methods), in which the subject traces a star-shaped maze with visual feedback. Each scan is represented in the scatter plot with a letter or digit located at the projection of the scan on to *canonical directions*, i.e., directions that maximize the separation with respect to the scan labels. Each subject's 10 consecutive scans are labeled in four groups as TT12223333; T denotes a tracing scan and digits represent the scans after mirroring of the control signal. Circles illustrate the shape of the model densities of each of the four classes $p_\theta(\mathbf{c}|g)$, i.e., in this case a Gaussian centered at the class mean with variance pooled across groups (the canonical directions remove within-group correlations, so that the within-group variance has circular shape).

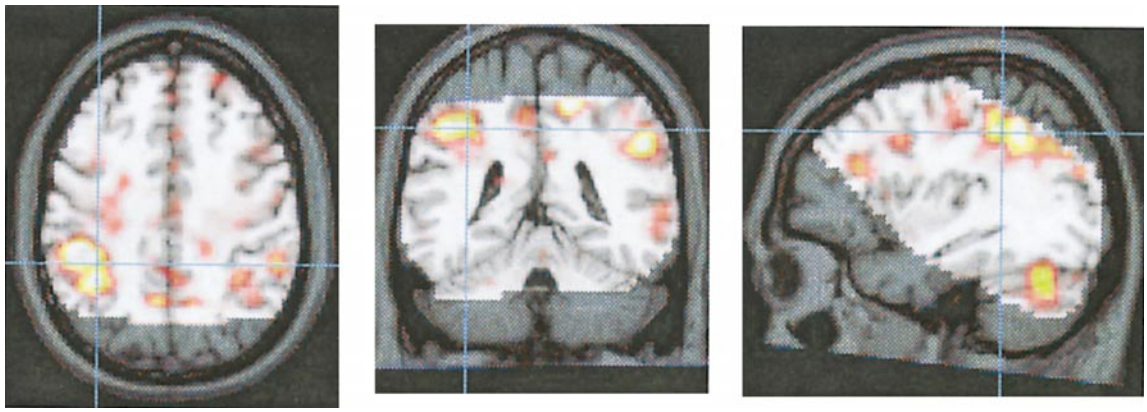

FIG. 2. Orthogonal slices of the sensitivity map for the mirror tracing experiment corresponding to Fig. 1 layered in color on top of an anatomical MR reference. The sensitivity map highlights areas that are important to the model for solving the classification task. Darkened areas were not part of the analysis. Note that the color map in this display is chosen arbitrarily.

digit located at the canonical coordinate in Eq. (20) of the scan. The sensitivity map for the same model is shown in Fig. 2, overlaid with an anatomical reference. The colored regions are found important for the model in order to perform the discrimination in Fig. 1.

The relation between generalizability and visualization reproducibility is illustrated in Fig. 3, in which the four data sets are plotted with average MI versus average pattern reproducibility for different numbers of subjects in the training set. We see that across the four data sets there is a tendency for the MI measure to correlate with the reproducibility, i.e., both tend to increase with the number of subjects in the training set and additionally with the relative difficulty of the modeling task.

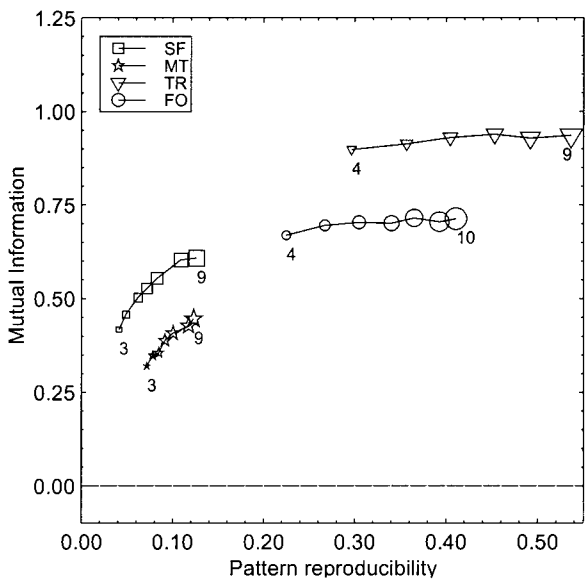

FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both measures indicate improved performance of the model as the number of subjects increases.

Some care should be exercised when comparing reproducibility between different models because it is insensitive to the model bias. However, within a given model and for the same visualization scheme reproducibility does directly address the key issue of the statistical validity of the derived neuroimage.

Figure 4 plots learning curves of MI, misclassification error rates, and pattern reproducibility for each of the four data sets and the four setups described in Table 1. We further analyze the type of error made by the classifiers in Fig. 5 in the so-called confusion matrices, which show predicted label probabilities versus the true label, averaged over all cross-validations. The confusion matrices shown correspond to the agnostic (N-label) N-PC models.

The MI learning curves demonstrate that none of the models are able to discriminate between much more than two brain states, corresponding to 1 bit of label information. Even models which were given the potential to extract far more information (for example in SF the 12 states have a limit of 3.6 bits) fail to do so. This is interesting considering that the experimental designs are made to reduce irrelevant activations and to maximize/isolate the neural activation in question. However, note that this observation does not imply that the more complex models using agnostic labeling schemes equal to the number of scans per subject cannot be used to evaluate temporal trends, as demonstrated in Frutiger *et al.* (2000).

As expected, all learning curves show improved performance as the number of subjects increases. The fact that the curves do show minor nonmonotonicities is attributed to test sample fluctuations, as we have rather limited test set sizes.

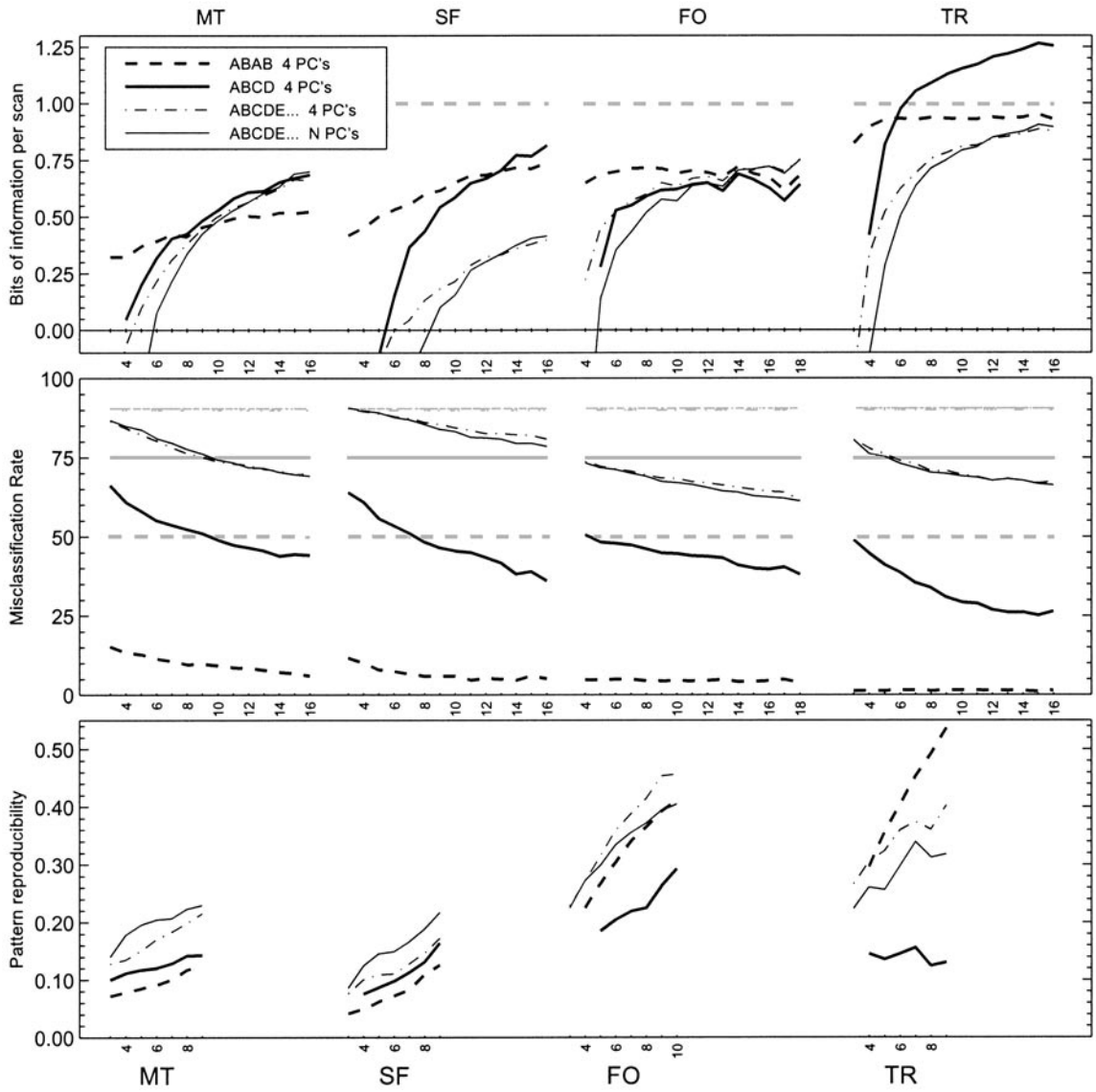

FIG. 4. Learning curves (model performance as function of number of subjects in the training set) for the four activation experiments (MT, mirror tracing; SF, static force; FO, finger opposition; TR, figure tracing). The four curves on each graph represent the different modeling setups of Table 1: Dashed thick line, 4 scans per subject, 4 PC's, 2 labels (activation/baseline AABB or ABAB). Solid thick line, 4 scans per subject, 4 PC's, agnostic labels (ABCD). Thin solid line, all scans per subject used (10, 12, 8, 10, respectively), number of PC's corresponding to number of scans, agnostic labels (ABCDE . . .). Thin dash-dot line, all scans per subject used (10, 12, 8, 10, respectively, only 4 PC's), agnostic labels (ABCDE . . .). The gray levels in the error rate graphs indicate the rates obtained if the model kept guessing the most common label.

Two-label TR and the FO MI curves saturate; i.e., the model predictions do not seem to improve by adding further subjects to the study. It is still possible that another less biased model could improve from a larger training set; see for example the two- versus four-label TR curves. The flattened prediction curves mean that performance is now limited by bias. This argues for selection of statistical models which have the "right" amount of bias for the learning curve to flatten out at a training set size close to the total number of subjects available. Thus, the simple model has better performance for small training set sizes, whereas the flexible model seems to be restrained by variance.

On the other hand, we see that the stability of the extracted activation patterns (as measured by the pattern reproducibility coefficient) improves with additional subjects in most cases. This happens because reproducibility is computed as a spatial average. The major regions relevant for label prediction are identified already with small training set sizes. Increasing

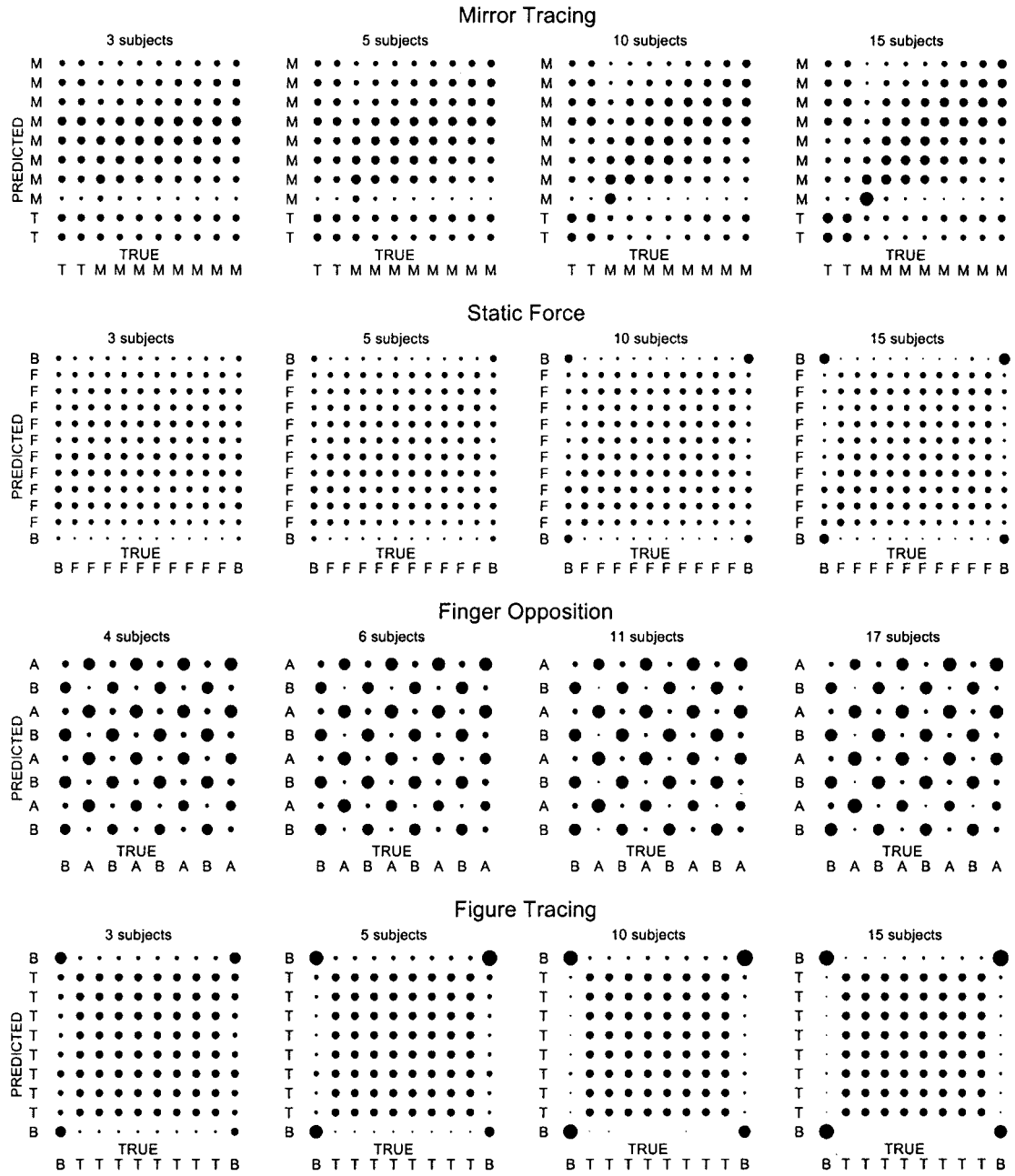

FIG. 5. Probability confusion matrices for the four data sets at different training set sizes. The confusion matrix measures how agnostic labels (Setup 4, Table 1) are confused by the model averaged over 150 subject permutations, averaging predicted class probabilities (rows) versus targets (columns). The averaged probabilities are indicated by the areas of the disks and all columns sum to the same area. These confusion matrices correspond to the *N*-PC curves (thin solid lines) in Fig. 4.

the training set size seems to stabilize the activation map across the entire volume, causing reproducibility and the underlying activation pattern signal-to-noise to keep improving, while label prediction performance does not improve much.

The mutual information learning curves show a tendency to drop very fast when the number of subjects is small. There is a significant penalty in MI if the model for some reason assigns a very small probability to a test example because the MI is derived from the average log of the predicted probability of the true label. The predictive performance is further decreased because the basis vectors $\mathbf{B}^*_{0,tr}$ become more "noisy" for small training set sizes. Error rates do not drop as fast as MI for smaller number of subjects. The misclassification measure is not penalized heavily by low-proba-

bility test examples; correctly and incorrectly classified test examples contribute with equal weight to the error rate. Thus, the misclassification error measure is less sensitive to model overfitting than the MI measure.

We also notice that the models with negative MI still perform as discrete classifiers and the activation patterns are reproducing. In all cases of negative MI, we see misclassification rates below the baseline error rate (error rate obtained by a constant guess at the largest label class). One should interpret a negative MI as an indication that the model is overfitting, i.e., dominated by variance in the bias/variance tradeoff.

There are several examples of crossing MI learning curves indicating the bias/variance tradeoff, for instance the two- and four-label curves in the MT, SF, and TR data sets. The more biased model (the two-label model, dashed thick) performs better on smaller data sets because the bias helps to prevent overfitting. Once enough subjects are available to obtain stable parameter estimates the complex model (four-label, solid thick) eventually outperforms (MT, SF, TR) or matches (FO) the performance of the simple model.

The effect of P, the number of singular directions retained (aka principal components, PCs) can also be understood in the bias/variance context. For all tasks, 4-PC and the less biased N-PC curves seem to meet (perhaps cross). There seems to be no real advantage, in terms of MI, of including more than four principal components (thin dash-dot versus thin solid lines). In all four experiments, the performance of the models with many PC's only just reaches the level of the 4-PC models.

We note that reproducibility values in the sensitivity map are approximately $\frac{1}{2}$ of those in the CVA map (compare with companion paper Strother *et al.* (2002)). The reason for this is that the sensitivity map does not have a sign, i.e., both negatively and positively activated regions map onto a positive sensitivity which reduces the reproducibility correlation coefficient.

The confusion matrix is useful for diagnosing problems faced by the model. In the MT data, the model confuses the 2 initial tracing scans, the first mirror tracing scan is fairly unique, and the last scans have a slight temporal trend that results in a weak diagonal structure in the upper right-hand corner of the matrices. It is also of interest to note that as the subject accommodates to the mirrored control (scans 8–10) the model tends to confuse these scans with the initial 2 tracing scans.

In FO, TR, and SF, we see as expected that the confusion occurs mostly within baseline scans and within active scans. Two of the N-label models (FO, TR) found no temporal effects within the active scans. This corresponds well to the fact that, in these three data sets, none of the N-label models (thin solid lines in Fig. 4) extracted more label information than the two-label models (thick dashed lines). Meanwhile, the four-label models (thick solid lines) do seem to find more than two states in the TR data (and to some extent SF), indicating that the most complex models did not have enough training examples to discover these effects.

CONCLUSION

We have introduced a prediction error-based framework for evaluation of models of functional neuroimage sets. Models were demonstrated that were able to predict the labels of scans in subjects the models had never "seen" before, verifying the validity of the model and model assumptions, something that is just assumed in conventional hypothesis-based analysis framework.

The prediction error was given an interpretation as the mutual information between scans and scan labels, measured in bits. The information rate can be interpreted as the amount of information predicted about the scan label given the scan. It was shown how this performance measure is computed, consistent with micro- and macroscopic type models. Furthermore, we demonstrated that the extracted mutual information is directly comparable across different labeling schemes. Being a linear transformation of the prediction error, the mutual information can be used to expose the bias/variance tradeoff, and it was shown how this can be used to identify models that are either over- or underfitting the data.

The effect of training set size was investigated by computing learning curves, i.e., plots of predictive performance as a function of the number of subjects used to train the model. The learning curves were generated using a cross-validation scheme. The new framework was applied to four [^{15}O]water PET functional activation studies and learning curves from different modeling setups were analyzed using bias/variance considerations.

We introduced the sensitivity map as a new scheme for model visualization used for identification of the parts of the input space important to the model performance. The sensitivity map can be derived for any probabilistic model with a macroscopic formulation.

Having measured the performance of the scan/label predictions, we considered the "quality" of the derived activation map using the pattern reproducibility, as defined in the NPAIRS framework in the companion paper (Strother *et al.,* 2002). We found that pattern reproducibility addresses the variance part of the pattern generalizability and thus is not an unbiased measure of performance. The unbiased generalizability of the visualization requires a ground truth image and is therefore available only for simulated data. For the four real data sets we evaluated the variance part of the generalizability, by computing the reproducibility of the produced visualization. We observed that the visual quality of the extracted activation pattern can-

not always be inferred from the quality of the label predictions. For example, we showed examples in which adding subjects increased the extracted pattern reproducibility without markedly improving the performance in terms of label predictions. For this reason, both label prediction and pattern reproducibility remain key tools for resolving preprocessing issues and verification of the validity of modeling assumptions in real data sets.

APPENDIX A: BASIS PROJECTIONS FOR CROSS-VALIDATION

Consider the problem of obtaining projections of a test set onto an SVD basis computed on an independent training set. We start out with a division of the data matrix in two parts; assume that we can split the columns of the I voxels by N scans data matrix like this: $\mathbf{X} = [\mathbf{X}_{tr}\mathbf{X}_{te}]$. We then compute the SVD on the training data matrix, $\mathbf{X}_{tr} = \mathbf{U}_{tr}\mathbf{\Lambda}_{tr}\mathbf{V}_{tr}^T$ (size $I \times N_{tr}$), and the training set projection becomes $\mathbf{Q}_{tr} = \mathbf{\Lambda}_{tr}\mathbf{V}_{tr}^T$ (size $N_{tr} \times N_{tr}$) while the test set projection onto the basis defined by the training set is $\mathbf{Q}_{te} = \mathbf{U}_{tr}^T\mathbf{X}_{te}$ (size $N_{tr} \times N_{te}$). With such an approach we would require one SVD of a matrix with $I \sim 10{,}000$ rows for every cross-validation iteration, a procedure which is computationally quite demanding. To improve this we perform a two-step SVD which is mathematically equivalent. We compute a single SVD of the entire data matrix once and perform smaller train/test second-level SVD operations for the cross-validations:

1. Compute full SVD: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ and $\mathbf{Q} = \mathbf{\Lambda}\mathbf{V}^T$ (size $N \times N$).
2. Partition the columns of $\mathbf{Q}$ into train and test set (per-subject-wise in PET) $\mathbf{Q} = [\mathbf{Q}_{0,tr}\mathbf{Q}_{0,te}]$.
3. Compute small size SVD, $\mathbf{Q}_{0,tr} = \mathbf{B}_{0,tr}\mathbf{\Lambda}_{0,tr}\mathbf{V}_{0,tr}^T$, and obtain the training set projections $\mathbf{Q}_{tr} = \mathbf{B}_{0,tr}^T\mathbf{Q}_{0,tr} = \mathbf{\Lambda}_{0,tr}\mathbf{V}_{0,tr}^T$ (size $N_{tr} \times N_{tr}$).
4. Obtain test set projections $\mathbf{Q}_{te} = \mathbf{B}_{0,tr}^T\mathbf{Q}_{0,te}$ (size $N_{tr} \times N_{te}$).
5. Compute model performance on $\mathbf{Q}_{tr}$ and $\mathbf{Q}_{te}$.
6. Repeat steps 2–5 for each cross-validation.

It should be noted that the procedure outlined here leads to projections that do not have exactly the same variance in the training and test sets. An alternative algorithm which remedies this problem, called the Generalizable Singular Value Decomposition, has been developed (Kjems *et al.*, 2001) and will not be described here since space does not allow for an elaborate discussion. The results shown in this paper are all based on the conventional SVD as described above.

APPENDIX B: DERIVATION OF SENSITIVITY MAP

Returning to our probabilistic CVA model in Eq. (21) we will derive its sensitivity map. First note the following derivatives for the canonical coordinates: $\partial \log p_{\boldsymbol{\theta}}(\mathbf{c}|g)/\partial\mathbf{c} = -(\mathbf{c} - \bar{\mathbf{c}}_g)$ and $\partial p_{\boldsymbol{\theta}}(\mathbf{c}|g)/\partial\mathbf{c} = -p_{\boldsymbol{\theta}}(\mathbf{c}|g)\,(\mathbf{c} - \bar{\mathbf{c}}_g)$. It follows by direct calculation that

$$\begin{aligned}\frac{\partial \log p_{\boldsymbol{\theta}}(g|\mathbf{c})}{\partial\mathbf{c}} &= \frac{\partial \log \dfrac{p_\theta(\mathbf{c}|g)p(g)}{\sum_{g'} p_\theta(\mathbf{c}|g')p(g')}}{\partial\mathbf{c}} \\ &= \frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{c}|g)}{\partial\mathbf{c}} - \frac{\partial \sum_{g'} \log p_{\boldsymbol{\theta}}(\mathbf{c}|g')p(g')}{\partial\mathbf{c}} \\ &= -(\mathbf{c} - \bar{\mathbf{c}}_g) + \sum_{g'} p_{\boldsymbol{\theta}}(g'|\mathbf{c})(\mathbf{c} - \bar{\mathbf{c}}_g).\end{aligned} \tag{24}$$

We collect the above derivatives in canonical space of each example $((\mathbf{c}^{(j)}, g^{(j)}))$ into a matrix $\mathbf{D} = [\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(N)}]$ with $\mathbf{d}^{(j)} = \partial \log p_\theta(g|\mathbf{c})/\partial\mathbf{c}|_{(\mathbf{c}^{(j)},\mathbf{g}^{(j)})}$. The derivatives with respect to the ith voxel can be found using $\mathbf{c} = \mathbf{L}^T\mathbf{B}_{0,tr}^*\mathbf{U}^T\mathbf{x}$, i.e., $\partial\mathbf{c}^T/\partial\mathbf{x} = \mathbf{U}\,\mathbf{B}_{0,tr}^*\mathbf{L} = \mathbf{M}$, so that

$$s_i \simeq \frac{1}{N}\sum_j \frac{\lceil \partial p_\theta(g^{(j)}|\mathbf{x}^{(j)})\rceil^2}{\partial x_i} = \frac{1}{N}\,\mathrm{diag}(\mathbf{M}\mathbf{D}\mathbf{D}^T\mathbf{M}^T), \tag{25}$$

which is most efficiently computed as the sum of each row of the matrix $1/N\ (\mathbf{M}(\mathbf{D}\mathbf{D}^T)) \odot \mathbf{M}$ ($\odot$ denotes elements-wise multiplication).

Notice that the above sensitivity map takes a particularly simple form for two-state models. In such models the dimensionality of the canonical coordinates is 1 (recall that $\mathbf{c}$ has dimension $k - 1$, with k being the number of classes), so $\mathbf{D}$ is scalar, which again means the sensitivity map equals the squared elements of the canonical eigenvector. This eigenvector is identical to the Fisher linear discriminant when $k = 2$.

In some cases we may not be interested in a global summary map, but rather a map which expresses the uniqueness of one or more of the states g. For example, what characterizes the brain state $g = A$ from $g \neq A$ with no information about the discrimination of states $g \neq A$? Such a map can be obtained by a slight alteration of the definition in Eq. (15) by averaging over only the examples of the state we wish to map, i.e.,

$$s(g)_i \simeq \frac{1}{N}\sum_{j|g^{(j)}=g} \frac{\lceil \partial p_\theta(g^{(j)}|\mathbf{x}^{(j)})\rceil^2}{\partial x_i}. \tag{26}$$

Equation (25) is modified similarly by summing over examples of the investigated state rather than over all states.

ACKNOWLEDGMENTS

This work was supported by Human Brain Project Grant P20 MH57180 and the Danish Research Councils for the Natural and

Technical Sciences through the THOR Center for Neuroinformatics. Finally, the authors thank the anonymous reviewers for many useful questions and suggestions.

REFERENCES

Bishop, C. 1999. Bayesian pca. In *Advances in Neural Information Processing Systems* (M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds.), Vol. 11. MIT Press, Cambridge, MA.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Discovery* **2:** 121–167.

Cherkassky, V., and Mulier, F. 1998. *Learning from Data: Concepts, Theory and Methods.* Wiley, New York.

Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory.* Wiley, New York.

Duda, R. O., Hart, P. E., and Stork, D. G. 2001. *Pattern Classification.* Wiley, New York.

Friston, K., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. S. J. 1995a. Spatial registration and normalization of images. *Hum. Brain Mapping* **2:** 165–189.

Friston, K. J., Holmes, A. P., Worsley, K., Poline, J.-B., Frith, C. D., and Frackowiak, R. 1995b. Statistical parametric maps in functional neuroimaging: A general linear approach. *Hum. Brain Mapping* **2:** 189–210.

Friston, K., Poline, J. B., Holmes, A. P., Frith, C. D., and Frackowiak, R. S. J. 1996. A multivariate analysis of pet activation studies. *Hum. Brain Mapping* **4:** 140–151.

Frutiger, S. A., Anderson, J. R., Daly, D. G., Sidtis, J. J., Arnold, J. B., Strother, S. C., Savoy, R., and Rottenberg, D. A. 1998. PET studies of perceptuomotor learning in a mirror-reversal paradigm. *NeuroImage* **7:** S962.

Frutiger, S. A., Strother, S. C., Anderson, J. R., Sidtis, J. J., Arnold, J. B., and Rottenberg, D. A. 2000. Multivariate predictive relationship between kinematic and functional ativation patterns in a PET study of visiomotor learning. *NeuroImage* **12:** 515–527.

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

Genovese, C. R., Lazar, N. A., and Nichols, T. E. 2001. Threshold determination using the false discovery rate. *Neuroimage* **13:** S124.

Golub, G. H., and Loan, C. F. V. 1989. *Matrix Computations.* John Hopkins Press, Baltimore.

Hansen, L., Larsen, J., Nielsen, F., Strother, S., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., and Paulson, O. 1999. Generalizable patterns in neuroimaging: How many principal components? *NeuroImage* **9:** 534–544.

Hertz, J., Heller, J., Kjær, T., and Richmond, B. 1995. Information spectroscopy of single neurons. In *Neural Networks: From Biology to High Energy Physics,* pp. 123–132. World Scientific, Singapore.

Heskes, T. 1998. Bias/variance decompositions for likelihood-based estimators. *Neural Comput.* **10:** 1425–1433.

Hintz-Madsen, M., *et al.* 1995. Design and evaluation of neural classifiers—Application to skin lesion classification. *1995 IEEE Workshop on Neural Networks for Signal Processing (NNSP'95).*

Kjems, U., Hansen, L. K., and Strother, S. C. 2001. Generalizable singular value decomposition. *Neural Inform. Process. Syst.* **13:** 549–555.

Kjems, U., Strother, S. C., Anderson, J. R., Law, I., and Hansen, L. K. 1999. Enhancing the multivariate signal of [^{15}O]water PET studies with a new non-linear neuroanatomical registration algorithm. *IEEE Trans. Med. Imaging* **18:** 306–319.

Kustra, R. 2000. *Statistical Analysis of Medical Images with Applications to Neuroimaging.* Univ. of Toronto, Toronto. [Ph.D. thesis]

Kustra, R., and Strother, S. C. 2001. Penalized discriminant analysis of [^{15}O]water PET brain images with prediction error selection of smoothing and regularization hyperparameters. *IEEE Trans. Med. Imaging* **20:** 376–387.

Lautrup, B., Hansen, L. K., Law, I., Mørch, N., Svarer, C., and Strother, S. 1995. Massive weight sharing: A cure for extremely ill-posed problems. In *Proceedings of Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks, HLRZ, KFA Jülich, Germany* (H. J. Hermann, D. E. Wolf, and E. P. Pöppel, Eds.), pp. 137–148. World Scientific, Singapore.

Liow, J. S., Strother, S. C., Rehm, K., and Rottenberg, D. A. 1997. Improved resolution for PET volume imaging through three-dimensional iterative reconstruction. *J. Nucl. Med.* **38:** 1623–1631.

Mardia, K. V., and Kent, J. T. 1979. *Multivariate Analysis.* Academic Press, San Diego.

Mørch, N. 1998. *A Multivariate Approach to Functional Neuro Modeling.* Department of Mathematical Modelling, Technical University of Denmark, Lyngby. [Ph.D. thesis]

Mørch, N., Hansen, L. K., Strother, S. C., Law, I., Svarer, C., Lautrup, B., Anderson, J. R., Lange, N., and Paulson, O. B. 1996. Generalization performance of nonlinear vs. linear models for [^{15}O]water PET functional activation studies. *NeuroImage* **3:** 258.

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

Muley, S. A., Strother, S. C., Ashe, J., Frutiger, S. A., Anderson, J. R., Sidtis, J. J., and Rottenberg, D. A. 2001. Effects of changes in experimental design on PET studies of isometric force. *NeuroImage* **13:** 185–195.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. 1986. *Numerical Recipes.* Cambridge Univ. Press, Cambridge, UK.

Skudlarski, P., Constable, R. T., and Gore, J. C. 1999. ROC analysis of statistical methods used in functional MRI: Individual subjects. *NeuroImage* **9:** 311–329.

Strother, S. C., Anderson, J. A., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. 2002. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage* **15:** 747–771.

Strother, S. C., Anderson, J. R., Schaper, K. A., Sidtis, J. J., Liow, J. S., Woods, R. P., and Rottenberg, D. A. 1995. Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping. I. "Functional connectivity" of the human motor system studied with [15-O]water PET. *J. Cereb. Blood Flow Metab.* **15:** 738–753.

Strother, S. C., Lange, N., Savoy, R. L., Anderson, J. R., Sidtis, J. J., Hansen, L. K., Bandettini, P. A., O'Craven, K., Rezza, M., Rosen, B. R., and Rottenberg, D. A. 1996. Multidimensional state spaces for fMRI and PET activation studies. *NeuroImage* **3:** S98.

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