NeuroImage **15,** 747–771 (2002)
doi:10.1006/nimg.2001.1034, available online at http://www.idealibrary.com on IDEAL®

The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework

Stephen C. Strother,*,†,‡,§ Jon Anderson,*,‡ Lars Kai Hansen,¶ Ulrik Kjems,¶ Rafal Kustra,‖ John Sidtis,† Sally Frutiger,†,‡ Suraj Muley,† Stephen LaConte,§ and David Rottenberg*,†,‡

**Department of Radiology, †Department of Neurology, and §Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota 55455; ‡PET Imaging Center, VA Medical Center, Minneapolis, Minnesota 55417; ¶Institute of Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark 2800; and ‖Public Health Sciences, University of Toronto, Toronto, Ontario, M5S-1A8 Canada*

Received January 19, 2001

We introduce a data-analysis framework and performance metrics for evaluating and optimizing the interaction between activation tasks, experimental designs, and the methodological choices and tools for data acquisition, preprocessing, data analysis, and extraction of statistical parametric maps (SPMs). Our NPAIRS (nonparametric prediction, activation, influence, and reproducibility resampling) framework provides an alternative to simulations and ROC curves by using real PET and fMRI data sets to examine the relationship between prediction accuracy and the signal-to-noise ratios (SNRs) associated with reproducible SPMs. Using cross-validation resampling we plot training–test set predictions of the experimental design variables (e.g., brain-state labels) versus reproducibility SNR metrics for the associated SPMs. We demonstrate the utility of this framework across the wide range of performance metrics obtained from [^{15}O]water PET studies of 12 age- and sex-matched data sets performing different motor tasks (8 subjects/set). For the 12 data sets we apply NPAIRS with both univariate and multivariate data-analysis approaches to: (1) demonstrate that this framework may be used to obtain reproducible SPMs from any data-analysis approach on a common *Z*-score scale (rSPM{*Z*}); (2) demonstrate that the histogram of a rSPM{*Z*} image may be modeled as the sum of a data-analysis-dependent noise distribution and a task-dependent, Gaussian signal distribution that scales monotonically with our reproducibility performance metric; (3) explore the relation between prediction and reproducibility performance metrics with an emphasis on bias-variance tradeoffs for flexible, multivariate models; and (4) measure the broad range of reproducibility SNRs and the significant influence of individual subjects. A companion paper describes learning curves for four of these 12 data sets, which describe an alternative mutual-information prediction metric and NPAIRS reproducibility as a function of training-set sizes from 2 to 18 subjects. We propose the NPAIRS framework as a validation tool for testing and optimizing methodological choices and tools in functional neuroimaging. © 2002 Elsevier Science (USA)

Key Words: **multisubject PET and fMRI studies; data analysis; univariate; multivariate; prediction error; reproducibility; cross-validation; resampling.**

INTRODUCTION

A wide range of techniques and software tools has become available with which to process functional neuroimaging data sets. To date this has not been accompanied by the development of a similarly wide range of performance metrics or benchmark data sets with which to evaluate and compare the tools. Moreover, activation patterns obtained from functional neuroimaging studies reflect interactions among a complicated "data chain" of experimental decisions involving the activation task, a wide range of experimental design parameters, and a series of methodological choices including data acquisition, postacquisition processing, and data-analysis model selection. Many researchers focus on extracting "neuroscientifically relevant" results from their data sets, sometimes based on their ability to test explicit hypotheses. However, this is typically done without attempting to optimize and/or understand the relative influence of the experimental design and methodological choices that were made in obtaining the data. The generation of a "plausible result" that can be "linked" to the neuroscientific literature, perhaps through a hypothesis, is often taken as justification of the choices made, providing a systematic bias in the field toward prevailing neuroscientific expectations. Strother *et al.* (1995a) have noted that, "the fact that a data-analytic model can be used to produce regions that *may* be involved in a particular cognitive process or disease state does not constitute sufficient evidence for preferentially selecting that data-analytic model," and recently Skudlarski *et al.* (1999) have also noted the existence of this neuroscientifically

1053-8119/02 $35.00

driven result bias due to the "arbitrariness of the choice of data-analytic strategies."

We certainly do not advocate ignoring the existing neuroscientific knowledge base, but both its implicit and its explicit use needs to be balanced against a concerted effort to define and test the basic validity of the wide range of experimental and methodological techniques used in functional neuroimaging. Our approach to this problem is guided by the rapidly developing field of predictive learning in statistics (e.g., Friedman, 1994; Larsen and Hansen, 1997; Ripley, 1998). We propose that the "validity or quality" of functional neuroimaging results, and the experimental and methodological choices made in obtaining them, should be established by quantitatively measuring and optimizing performance metrics. Our goal is to optimize the ability of the "functional-neuroimaging data chain" to produce data-analytic model parameters, including statistical parametric maps (SPMs) from a *training data set* that (1) can *accurately predict* the values of experimental design parameters (e.g., brain-state labels, performance measures) in an independent *test data set* and (2) can also *reliably reproduce the SPM* image parameters in the same test data set. Such validity defined as optimal prediction accuracy and SPM reproducibility in a test data set is not guaranteed by inferential statistical procedures even when all of the underlying model assumptions are true, unless we have asymptotically large data sets, something that is far from satisfied in functional neuroimaging experiments.

This problem with inferential statistical procedures occurs because they typically focus on obtaining maximum likelihood (ML) parameter estimates, which *only asymptotically* approach normal, unbiased estimates with minimum variance, and for small to moderate sample sizes such estimates are not "efficient" (Papoulis, 1991; Ripley, 1996). This means that for moderate sample sizes there are other estimation procedures that have smaller parameter variance than the ML estimates although these other procedures' parameters converge to biased estimates of the true population values asymptotically, i.e., for real finite data sets there is a bias–variance tradeoff to be considered. This phenomenon is seen in t values which are themselves model parameter estimates subject to sampling noise (Holmes *et al.*, 1996; Svarer *et al.*, 1997). As a result there is evidence that using biased, but more efficient, pooled-variance estimates of t values (i.e., scaled mean-difference SPMs) produces more reproducible SPMs with a better detection signal-to-noise ratio (SNR) than those obtained with single-voxel variance estimates (Strother *et al.*, 1998, and Section 4d, Petersson *et al.*, 1999b). In addition, for prediction metrics Mørch *et al.* (1997) have shown that even though a nonlinear model may have better performance given enough data (i.e., asymptotically), for small data sets the nonlinear model may be outperformed by a more biased linear model, leading to so-called "crossed learning curves" (see also Kjems *et al.*, 2002). Parameter estimation, particularly in finite, high-dimensional (i.e., imaging) data sets, requires choosing a bias–variance tradeoff, which may not be optimized using inferential estimates based on maximum likelihood techniques. As a result resampling procedures, such as nonparametric prediction, activation, influence, and reproducibility resampling (NPAIRS), may be essential for optimizing the functional neuroimaging chain because they provide insight into the bias–variance tradeoffs being made in real, finite data sets.

Many studies have been performed on components of the functional neuroimaging data chain, but it is sometimes difficult to utilize this literature given a new task and experimental design because it may be unclear which results are directly applicable from these earlier studies. In [^{15}O]water PET there has been work on the experimental design issues of group size and its influence on statistical power (e.g., Grabowski *et al.*, 1996; VanHorn *et al.*, 1998; Strother *et al.*, 1998; Petersson *et al.*, 1999b), with attempts to use meta-analysis to identify important factors that influence published PET results (Gold *et al.*, 1997). Unfortunately, such studies have been somewhat limited by the widespread use of a small set of methodological and statistical choices from within the "SPM" software package (http://www.fil.ion.ucl.ac.uk/spm/). The use of a pluralistic data-analytic modeling strategy in PET (e.g., Strother *et al.*, 1995b, 1998; Muley *et al.*, 2001) and in fMRI (e.g., Lange *et al.*, 1999; Tegeler *et al.*, 1999), with the extension to consensus activation patterns from multiple models proposed by Hansen *et al.* (2001), represents one approach to overcoming such model-dependent biases. When this pluralistic strategy is combined with "learning curves"—plots of prediction performance as a function of training-set size—and the activation-pattern reproducibility metrics outlined in this paper, a powerful framework for tuning and testing the functional neuroimaging chain is obtained. In particular, this framework, with its emphasis on tuning adaptive, data-driven models, demonstrates that the methodological choices should be optimized for the task, modality, methodology, and amount of available data (e.g., Mørch *et al.*, 1997; Somorjai *et al.*, 2001; Kjems *et al.*, 2002).

Another approach for testing the validity of choices in the functional neuroimaging chain is the use of signal-detection tools like receiver operating characteristic (ROC) techniques (e.g., Skudlarski *et al.*, 1999), which may be part of a pluralistic framework (Lange *et al.*, 1999). However, unless we assume we are dealing with known brain states (e.g., normal and disease, Liow *et al.*, 2000) ROC curves require simulations of "true" spatial neuroimage signal and noise patterns, and thereby introduce their own collection of assump-

tions and biases that can be difficult to identify and evaluate.

Within the functional neuroimaging literature it is striking that there are few comparisons across multiple tasks (e.g., Petersen *et al.*, 1998) that might allow the general features of real spatial activation patterns to be characterized to design better simulations for ROC studies. To our knowledge there are no studies that systematically compare multiple tasks across experimental design and methodological choices, including multivariate and univariate data-analysis models. We believe that this is a critical issue, for the most generally important experimental design and methodological choices are likely to be those that retain their influence across a wide range of tasks. Multitask comparisons may also be important to quantify and explore the observation that activation signals in "higher order" tasks using nonprimary regions are weaker than those from primary-sensory tasks (e.g., Xiong *et al.*, 1996), an issue that does not seem to have been systematically studied.

In this paper and a companion paper (Kjems *et al.*, 2002), we study data-analysis performance metrics in real PET data sets, as an alternative to using ROC curves based on simulated data. In this work we introduce a specific resampling framework we have labeled NPAIRS that extends the idea of measuring prediction accuracy using training–test-set resampling to include activation-pattern reproducibility metrics and subject influence (see http://neurovia.umn.edu/incweb/npairs_info.html for NPAIRS software and documentation). For both univariate and multivariate data analysis models we (1) demonstrate that our resampling framework may be used to directly measure reproducible activation signal-to-noise ratios from multiple models on a common Z-score scale (rSPM$\{Z\}$); (2) demonstrate that the histogram of a rSPM$\{Z\}$ image volume may be modeled as the sum of a data-analysis-dependent noise distribution and a task-dependent, Gaussian signal distribution that scales monotonically with our reproducibility performance metric; (3) explore the relation between prediction accuracy and pattern reproducibility with an emphasis on bias–variance tradeoffs for flexible, multivariate models; and (4) apply NPAIRS to 12 diverse motor data sets to quantitatively measure their broad spread of reproducibility signal-to-noise ratios and the influence of individual subjects on the results.

THEORY

Testing Models with Cross-validation Resampling

The cross-validation resampling procedure for building unbiased data-analysis models that are well adapted to the available data, D, is illustrated in Fig. 1 (e.g., Stone, 1974; Hansen *et al.*, 1990; Efron and Tibshirani, 1993, 1997; Ripley, 1998). The basic idea is to split the data into independent training and test sets and to use these to test that the model, number of parameters, and estimation techniques being used are as consistent as possible with predictions about D while avoiding excessive model bias or variance as a result of estimating too few or too many parameters, respectively, given the available data. Moreover, if the assumptions associated with the model parameters estimated in the training set are badly mismatched to the data this will be reflected as poor prediction accuracy in the test set. We examine selecting the combination of model and methodology with the highest prediction accuracy (or lowest prediction error) demonstrating that it is the "best" representation of those choices tested for the experimental design and finite data set represented by D (see Mørch *et al.*, 1996, 1997, 1998; Hansen *et al.*, 1999; Kustra, 2000; Ngan *et al.*, 2000; McKeown, 2000; Kustra and Strother, 2001). In addition, as the end goal of our functional neuroimaging experiments is not primarily to build a predictive model of the design matrix, we also focus on the extent to which optimizing prediction accuracy is associated with optimized reproducibility metrics and the signal to noise of the reproducible SPMs that we wish to interpret. Our approach directly addresses the problem discussed by Petersson *et al.* (1999a) of choosing between ". . . including all conceivable explanation variables (effects), and parsimony, using the smallest number of effects to form an adequate model."

Resampling Choices

A variety of cross-validation resampling schemes are possible based on K-fold resampling: (1) for N independent observations, split the data into K roughly equal-sized data sets; (2) at the kth resampling step fit the model to a training set composed of $(K - 1)$ data sets without the kth set, which is used as a test set to calculate prediction errors for the fitted model; (3) repeat steps 1 and 2 for $k = 1, \ldots, K$ (Efron and Tibshirani, 1993). There are no general rules for optimally choosing the K-fold resampling scheme for any particular data-analysis problem (Larsen and Goutte, 1999). At one extreme is twofold cross-validation resampling through to the other extreme of N-fold resampling, which are related to delete-d ($d = N/2$) and leave-one-out jackknife ($d = 1$) resampling techniques, respectively. N-fold cross-validation and leave-one-out jackknife each generate N resampling estimates, but twofold cross-validation provides only two estimates while delete-$N/2$ jackknife generates ${}^{N}C_{N/2}$. We shall refer to the process of repeatedly applying twofold cross-validation with different data splits for up to ${}^{N}C_{N/2}$ training and test sets as "split-half resampling." Although theoretically related to jackknife estimates cross-validation prediction estimates are obtained as

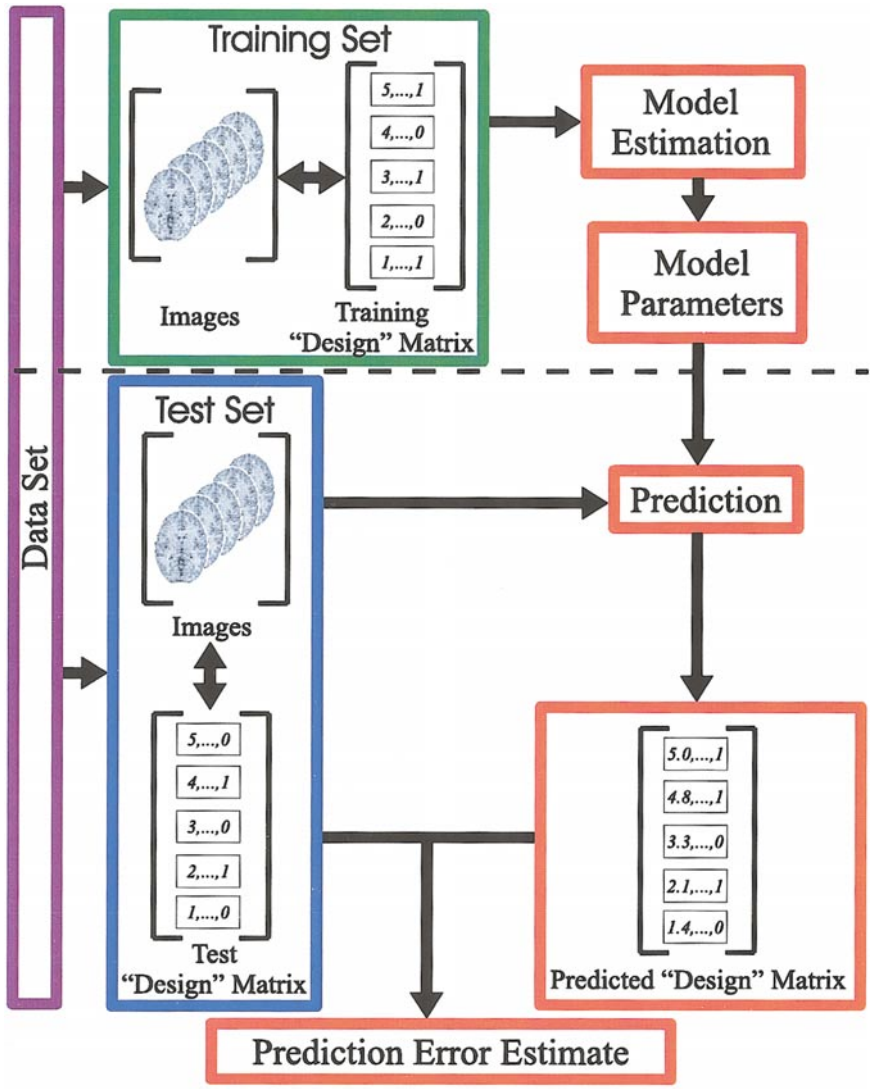

FIG. 1. Measuring the performance of data-analysis models with prediction metrics requires repeatedly splitting the data set (purple) into training (green) and test sets (blue) using cross-validation resampling. Model parameters are then estimated in the training set and used to predict the experimental "design matrix" values (e.g., scan state labels, covariates, etc.) in the independent test set (red). The predictions are then compared to the *known* design matrix values in the test set using one or more prediction error (accuracy) cost functions.

averages of parameters from the "left-out" test-set data, while jackknife estimates are based on averages of model parameter estimates from the training-set data (Efron, 1982). We may directly obtain jackknifed estimates of prediction errors, but this requires two levels of resampling with cross-validation resampling of each of the N jackknifed data sets of $(N - 1)$ observations. As the jackknife may be thought of as a linear approximation to the bootstrap it is probably preferable to consider the more efficient bootstrap estimates of prediction errors (Efron and Tibshirani, 1993). In this regard a technique that is related to two- and threefold cross-validation resampling is "leave-one-out bootstrap," in which bootstrap resampling with replacement is used to choose training and test sets that on average contain about $0.63N$ and $0.37N$ independent observations, respectively (Efron and Tibshirani, 1997); this has been explored recently in a functional neuroimaging context by Kustra and Strother (2001). All of the preceding techniques focus on the goal of obtaining efficient unbiased prediction error estimates for N observations.

In contrast, for NPAIRS we have chosen split-half resampling because we wish to optimize our reproducibility measurements by comparing SPMs from the largest independent groups possible. This split-half resampling choice maximizes the power of each of the independent-group data analyses (i.e., equal-sized training and test sets) while ensuring that their independent error estimates may be directly compared without dealing with bias due to different group sizes. We have adopted this unconventional resampling pro-

cedure because we believe that metrics for *both* prediction accuracy and activation pattern reproducibility are critical in functional neuroimaging. While prediction measurements are unbiased against the assumed truth of the experimental design, they do not directly address the quality of the SPM on which the experimental interpretation is often based. On the other hand, reproducibility measurements directly reflect the reliability and SNR of the SPM measurements, but may contain a significant undetectable bias (i.e., they reflect only SPM variance) because we do not know the true activation pattern.

Prediction Accuracy with Split-Half Resampling

We chose subjects as our basic resampling unit to ensure that the resampled observations were independent. Using this resampling scheme prediction accuracy was obtained between training and test sets for a given split; set designations were swapped and a second prediction accuracy measure was obtained and this was repeated for ${}^{N}C_{N/2}/2$ data splits, where N is the number of subjects per data set. The 35 splits available from each of our eight-subject data sets provided 70 split-half resampling measurements of the prediction accuracy, p, which were summarized by their median ($\tilde{p}$) to avoid sensitivity to outlying p values from influential subjects in individual splits.

Because we are using only half of our data for our training set the resulting prediction error estimates will be larger than estimates for training sets of size $(N - 1)$ in N-fold cross-validation. Prediction errors are larger for split-half resampling because they decrease monotonically as a function of training-set size, forming a learning curve (e.g., Mørch *et al.*, 1997, 1998). In addition, if we use our estimates of test-set prediction accuracy to optimize model output (e.g., by adjusting hyperparameters) they will be biased upward—prediction errors are underestimated—compared to the prediction estimates that would be obtained from a third independent validation data set. When unbiased prediction estimates are required a double-resampling procedure should be used in which the training set is split into learning and validation sets and cross-validation is used within each training set for model optimization (Friedman, 1994; Cherkassky and Mulier, 1998). We have chosen not to use such a double resampling technique because we are primarily interested in the relative, not the absolute values of our prediction estimates. Moreover, the two sources of bias in our prediction accuracy estimates counteract each other and resampling within our small split-half four-subject groups would be computationally very expensive and would generate very noisy estimates—we chose a small amount of bias over very noisy estimates. However, if necessary we could adjust for the bias in our prediction estimates using an approach based on the 0.632^{+} adjustment proposed by Efron and Tibshirani *et al.* (1997) and used in Kustra and Strother (2001) for leave-one-out bootstrap resampling.

A Reproducibility Metric with Split-Half Resampling

Figure 2 illustrates the use of split-half resampling with eight-subject data sets to generate reproducibility histograms. These are based on the Pearson product correlation coefficient (r) of the scatter plots of the resulting pairs of independent statistical parametric maps. The ${}^{N}C_{N/2}/2$ r values from the data splits are then displayed as a histogram, which may be further summarized by its median ($\tilde{r}$) to avoid sensitivity to outlying r values from influential subjects in individual splits. It is equally feasible to randomly choose split-half group sizes from $1, \ldots, N/2$ and examine reproducibility as a function of the number of subjects, a reproducibility learning curve (Strother *et al.*, 1998, 2000; Kjems *et al.*, 2002). When resampling SPMs in multivariate models (e.g., eigenimages from canonical variables analysis (CVA) and principal component analysis (PCA)) we must allow for the fact that these SPMs are defined only up to an arbitrary sign, which will result in both positive and negative r values when the SPMs are compared for independent groups of $N/2$ subjects. To avoid this in the PCA/CVA used in this study we perform a single analysis of all N subjects and use the resulting canonical coordinates (cc; Eq. (12), Appendix) as a reference set to determine if individual canonical dimensions from a single $N/2$-subject training set should be (1) reordered—the ith dimension of the training set may be most highly correlated with the jth ($j \neq i$) dimension of the reference set—and/or (2) reflected—the signs of the training-set cc's (and the associated canonical eigenimages) may be switched to ensure positive correlations with all reference set dimensions. This procedure represents a modified version of the reference set "filtering" for singular value decomposition (or PCA) with a parametric bootstrap, outlined by Milan and Whittaker (1995).

A Model-Independent, Reproducibility Signal-to-Noise Ratio for Activation Patterns

For each of the split-half scatter plots illustrated in Fig. 3 the result of a PCA of the associated two-dimensional correlation matrix is given by

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}, \quad (1)$$

where the two independent SPMs being compared

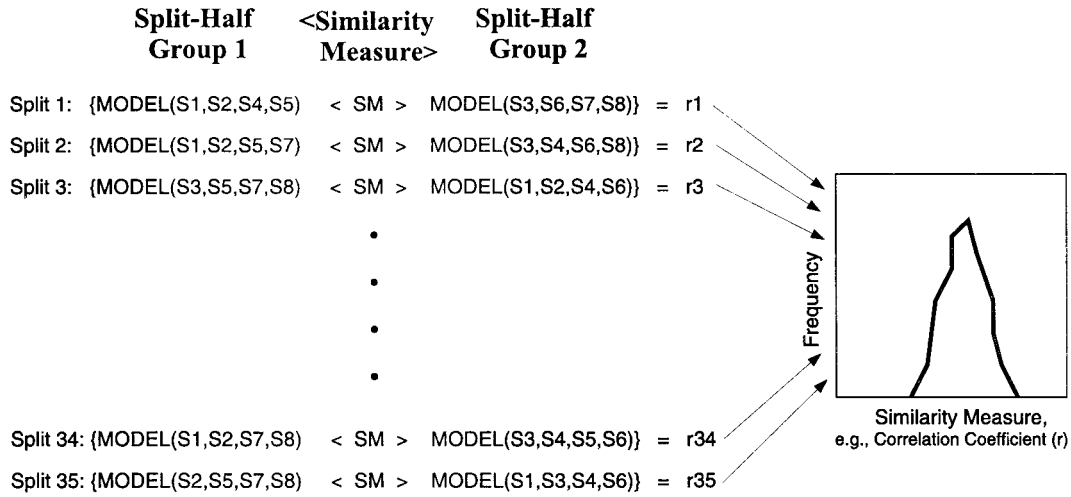

FIG. 2. Measuring the performance of data-analysis models with reproducibility metrics using repeated applications of twofold cross-validation with different splits, i.e., split-half resampling. We illustrate the technique for a group of eight subjects (S1, . . . , S8) that may be split into two independent groups of four subjects in 35 ways. For each of the 35 pairs of independent groups any data-analysis model can be applied to produce two independent statistical parametric maps (SPMs). A pattern similarity measure (SM) is used to summarize the reproducibility of the two SPMs using, for example, the correlation coefficient (r) of SPMs' voxel values. The 35 r values are then plotted as a "reproducibility histogram."

have been normalized by their whole-brain standard deviations (SD). Each point in the scatter plot corresponds to a brain location in Talairach space and is defined by the two independent SPM values obtained from a pair of split-half data sets. Equation (1) demonstrates that a PCA of this scatter plot will produce a principal axis along the line of identity (i.e., direction cosine $= 1/\sqrt{2}$) with variance of $(1 + r)$ and an uncorrelated minor axis with variance $(1 - r)$ for $r \in [0, 1]$. If the two normalized SPMs are very similar with low noise the principal-axis variance is ≈ 2 with minor axis variance ≈ 0, and the scatter plot will be a long thin ellipse along the line of identity. Alternatively, if the two SPMs contain only symmetric noise and no reproducing signal structure (i.e., $r = 0$) the variances along the principal and minor axes are equal, and the scatter plot will be a circular disk centered on the origin (Fig. 3A). Thus the PCA eigenvalue ratio $(1 + r)/(1 - r)$ provides a global summary of the reproducibility SNR with a range of $[1, \infty)$, and r represents a monotonic mapping of this range onto $[0, 1]$.

Let the two vectors of statistic values for the independent, normalized SPMs from a given split be $\mathbf{z}_1$ and $\mathbf{z}_2$, then projection onto the major and minor axes is equivalent to forming $(\mathbf{z}_1 + \mathbf{z}_2)/\sqrt{2}$ and $(\mathbf{z}_1 - \mathbf{z}_2)/\sqrt{2}$, respectively (personal communication, anonymous reviewer). This viewpoint makes it clear that signal and noise estimates may be readily obtained from individual voxels and regions of interest. The use of projections within the PCA framework emphasizes (1) the scatter plot as a visualization tool closely related to the reproducibility SNR provided by the PCA eigenvalue ratio, (2) that the SPMs being compared may be normalized differently or perhaps not at all, (3) the intuitive geometric extension to PCA of three or more independent samples outlined in Tegeler *et al.* (1999), and (4) that the underlying general problem is one of summarizing the structure of the joint density represented by an n-dimensional scatter plot, so that we may replace correlations with, for example, mutual information (Papoulis, 1991) and PCA with any other technique for modeling the joint density distribution. The large literature on outliers in multivariate data is applicable here, with alternate techniques being robust PCA and robust estimation of the correlation coefficient for bivariate data with outlier identification (e.g. see Chapter 7 of Barnett and Lewis, 1994).

The following theory allows us to explore the quantitative relationship between the shape of our rescaled signal histograms and our reproducibility performance metric, r. Our intuition is that r will be quite sensitive to the small number of potentially activated voxels with the largest signal values that determine the structure of the histogram's tails because the correlation coefficient is based on squared signal (and noise) values. We will refer to the reproducible activation image, s, obtained by projecting scatter-plot points onto the principal axis as a reproducibility SPM (rSPM). Let the activation-signal density function of rSPM across the whole brain be $a(s)$. After rescaling the signal axis by the minor-axis SD, $\sqrt{1 - r}$, we may derive an analytic relationship between our reproducibility metric, r, and the spread of the tails of the rescaled rSPM histogram as measured by its confidence intervals ($CI_{1-\alpha}$), given by

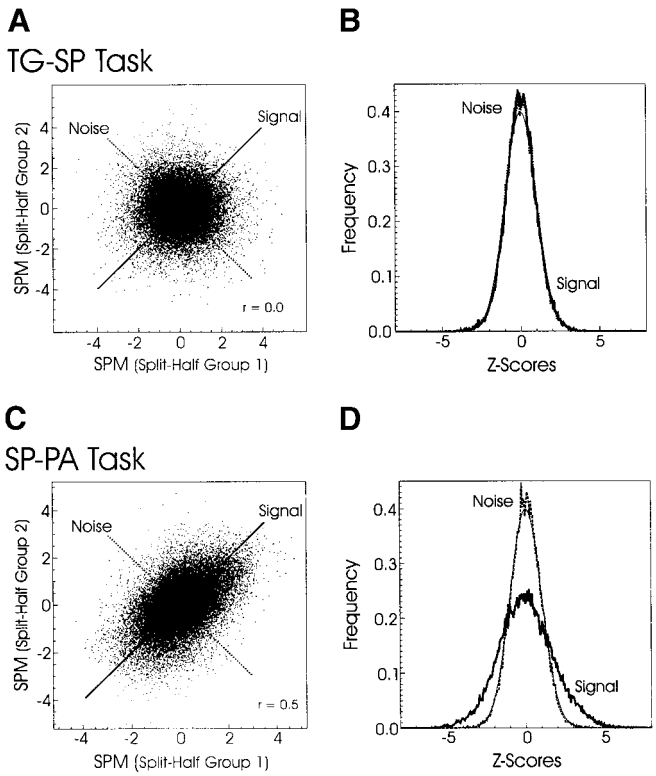

FIG. 3. The scatter plots used to compare independent SPMs and generate reproducibility correlation coefficients (r; left) may also be used to obtain reproducible signal and noise histograms (right). For a single split-half resampling we illustrate comparison of eigenimage SPMs from two-class canonical variate analysis of independent four-subject groups for (A, B) a target interception task (TG-SP) with no reproducibility ($r = 0$) and (C, D) a speech task (SP-PA) with relatively high reproducibility ($r = 0.5$). (Left) Reproducible-signal (solid line) and uncorrelated-noise (dotted line) axes from the principal component analysis (PCA) of the scatter plot after normalizing each SPM by its standard deviation (SD) and (right) reproducible-signal (rSPM{Z}, thick solid line) and noise (dotted line) histograms from projections of the scatter-plot voxel values onto the major and minor PCA axes, respectively, with rescaling of both axes by the SD of the noise-axis histogram to obtain Z scores. The rescaled noise histogram is overplotted with a Gaussian $\sim N(0, 1)$ (thin solid line).

$$\mathrm{CI}_{1-\alpha} = (s_{1-\alpha/2} - s_{\alpha/2})/\sqrt{1-r}. \qquad (2)$$

We shall refer to the rescaled rSPM values of Eq. (2) as rSPM{Z} as we find that the whole-brain noise distribution obtained by projecting scatter-plot values onto the minor PCA axis is often approximated by a Gaussian, $N(0, 1 - r)$ (see Fig. 3).

In addition to the simultaneous estimation of prediction and reproducibility metrics, it is our rescaling steps for each resampled, split-half pair of SPMs that we believe make our approach unique compared to other resampling estimates of the SPM voxel's standard errors; first we rescale the SPMs themselves and then we rescale the resulting reproducible SPM by the uncorrelated noise SD. The same split-half groups could also be used for standard error estimates with a delete-$N/2$ jackknife technique. Such subsampling techniques can help to make jackknife estimates for nonsmooth statistics more efficient (Efron and Tibshirani, 1993; Politis, 1998). As a result of rescaling rSPM by the uncorrelated noise estimate between independent sets of $N/2$ observations our approach may have several advantages over delete-$N/2$ jackknife estimates because: (1) the final rSPM{Z} statistical images based on averaging the rSPM{Z}'s from each of the resampled splits are weighted averages that will be robust to outlying SPM voxel differences between particular split-half groups, (2) if the resampled observations (e.g., subjects) introduce significant random effects these will be explicitly included in the weighted averaging, and (3) the final Gaussianized rSPM{Z} may be tested using the large body of random field theory techniques for both homogeneous and heterogeneous random fields that have been developed during the past decade (Worsley *et al.*, 1996, 1999).

Assume $a(s)$ from the principal axis is also a Gaussian distribution but with variance $(1 + r)$ as in Eq. (1), then from Eq. (2) we have

$$\mathrm{CI}(Z)_{1-\alpha} = (Z_{1-\alpha/2} - Z_{\alpha/2})\left(\frac{1+r}{1-r}\right)^{1/2}, \qquad (3)$$

where $Z \sim N(0, 1)$. Using the series expansion $\ln\{(1 + r)/(1 - r)\} \approx 2r + 2r^3/3 + 2r^5/5$ for $r^2 < 1$, we obtain

$$\log(\mathrm{CI}(Z)_{1-\alpha}) \approx \log(2Z_{1-\alpha/2}) + (\log e)\left(r + \frac{r^3}{3} + \frac{r^5}{5}\right). \qquad (4)$$

Equation (4) demonstrates that the shape of the reproducible SPM histogram obtained by projecting scatter-plot values onto the major PCA axis is composed of the sum of a fixed noise distribution (intercept for r is 0) and a Gaussian signal that scales approximately linearly with r. We are exploring generalizations of Eqs. (3) and (4) to non-Gaussian distributions.

A Probabilistic Framework for Discriminant CVA

For a fuller description of the general framework for probabilistic modeling in functional neuroimaging see the companion paper by Kjems *et al.* (2002) and the work of Mørch *et al.* (1997, 1998) and Hansen *et al.* (1999). For a detailed description of the CVA data-analysis framework and its close relation to MANOVA, penalized discriminant analysis, canonical correlation analysis, and partial least squares see, for example, Nielsen *et al.* (1998), Kustra (2000), and Kustra and Strother (2001). Examples of the use of this general multivariate modeling framework in functional neuroimaging are found in Moeller and Strother (1991),

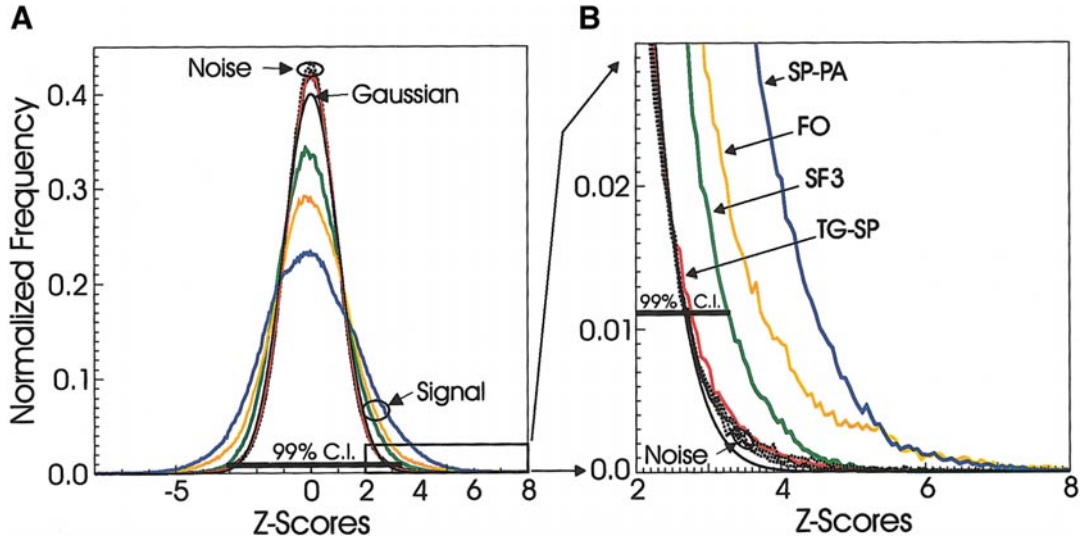

FIG. 4. From 35 split-half scatter plots we illustrate (A) the average reproducible-signal (solid colored lines) and average noise histograms (dotted black lines) for two-class canonical variate analysis of 8-subject groups performing target interception (TG-SP; red), static force (SF3; green), finger opposition (FO; orange), and speech (SP-PA; blue)—the signal and noise histograms were averaged following a principal component analysis (PCA) of each scatter plot and projection of the normalized voxel values onto the PCA axes with rescaling of both axes by the standard deviation of the noise-axis histogram. (B) A zoomed view of the positive tails of the signal and noise distributions in A, showing the similar noise distributions and range of reproducible activation signals from the four tasks. The noise histograms in both A and B are overplotted with a Gaussian $\sim N(0, 1)$ (black solid line). The thick black horizontal bar marks the 99% confidence interval (C.I.) of the average reproducible-signal histogram for the SF3 task.

Clark *et al.* (1991), Azari *et al.* (1993), Friston *et al.* (1995a, 1996), Fletcher *et al.* (1996), Bullmore *et al.* (1996), Rottenberg *et al.* (1996), Strother *et al.* (1995a,b, 1996), McIntosh *et al.* (1996, 1999), Worsley *et al.* (1997), Tegeler *et al.* (1999), Moeller *et al.* (1999), Frutiger *et al.* (2000), Muley *et al.* (2001).

Let $\mathbf{x}^{(j)}$ represent a vector of stochastic variables containing all voxel values from a single scan, j, and $\mathbf{g}^{(j)}$ be a stochastic vector of experimental design and other (e.g., performance) variables associated with scan $j = 1, \ldots, J$. We adopt this somewhat uncommon view of $\mathbf{g}$ because we believe it represents the general functional neuroimaging problem we are dealing with; namely the relation between a rich and largely unknown multivariate set of stochastic variables describing the total experimental environment and the neuroimages measured within that environment. This is the natural viewpoint for flexible associative multivariate techniques such as CCA, CVA, and PLS (e.g., McIntosh *et al.*, 1996; Frutiger *et al.*, 2000; Kustra and Strother, 2001). For example, a stochastic $\mathbf{g}$ is appropriate when (1) using behavioral performance measures, which are inherently stochastic, or (2) when we view the data analysis problem as one of estimating marginal distributions within the joint density function of $p(\mathbf{x}, \mathbf{g})$, as outlined below. Note that this view of $\mathbf{g}$ can incorporate fixed effects as the special case of point density estimates.

Using the nomenclature of the general linear model (GLM; Friston *et al.*, 1995b) we have a data matrix, $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(J)}]$, and a design matrix, $\mathbf{G} = [\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(J)}]$. For a functional activation data set $D = \{(\mathbf{x}^{(j)}, \mathbf{g}^{(j)})\}$ we would like to estimate the joint density function, $p_{\theta}(\mathbf{x}, \mathbf{g})$, using model parameters $\boldsymbol{\theta}$, to completely characterize D. We work with $p(\mathbf{x}|\mathbf{g};\boldsymbol{\theta})$, in the context of the GLM, or $p(\mathbf{g}|\mathbf{x};\boldsymbol{\theta})$ in the context of CVA. It is not obvious that either of these two marginal forms is to be preferred on mathematical grounds as they are both closely related to the joint density as described by the Bayes Theorem,

$$p(\mathbf{x}, \mathbf{g}) = p(\mathbf{x}|\mathbf{g})p(\mathbf{g}) = p(\mathbf{g}|\mathbf{x})p(\mathbf{x}). \tag{5}$$

For CVA we are interested in estimating $p(g|\mathbf{x};\boldsymbol{\theta})$, where g represents a scalar indicator or class variable of the experimentally defined brain state for each scan, $\mathbf{x}$. From Eq. (5) we have

$$p(g|\mathbf{x}) = \frac{p(\mathbf{x}, g)}{p(\mathbf{x})} = \frac{p(\mathbf{n})p(\mathbf{c}, g)}{p(\mathbf{n})p(\mathbf{c})}, \tag{6}$$

where $\mathbf{c}$ spans a signal subspace within $\mathbf{x}$ defined by model parameters, $\boldsymbol{\theta}$, such that $p(\mathbf{x}) \approx p(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{n})p(\mathbf{c}|\boldsymbol{\theta})$ with $p(\mathbf{n}, g) = p(\mathbf{n})$; $\mathbf{n}$ is an independent noise subspace that is factored out to obtain

$$p(g|\mathbf{x};\boldsymbol{\theta}) = \frac{p(\mathbf{c}, g|\boldsymbol{\theta})}{\sum_{g'} p(\mathbf{c}, g'|\boldsymbol{\theta})} = \frac{1}{C}\, p(\mathbf{c}|g;\boldsymbol{\theta})p(g), \tag{7}$$

with C a constant computed to ensure that the posterior probabilities add up to 1 (i.e., each scan must be allocated to some class so that $\Sigma_{g'}\, p(g'|\mathbf{x};\boldsymbol{\theta}) = 1$), and g' is a summation index across all classes, which includes the true experimentally defined class, g, of scan, $\mathbf{x}$.

Using Eq. (7) the CVA model parameters are estimated from a "training" set of data ($\boldsymbol{\theta}_{\text{tr}}$) to define a signal subspace spanned by $\mathbf{c}_{\text{tr}}$. Then $\boldsymbol{\theta}_{\text{tr}}$ is used to estimate the probability of the true class of "test" scans ($\mathbf{x}_{\text{te}}^{(j)}$) that were *not used* to train the model, by projecting them onto $\mathbf{c}_{\text{tr}}$ with CVA defining a probability density function. See the Appendix for the multivariate Gaussian distribution obtained from training and test sets for CVA in the form of Eq. 7. A simple prediction accuracy estimate is obtained by rescaling and offsetting the posterior probability estimates of Eq. (7) to a [0, 1] scale with

$$p_n(g|\mathbf{x};\,\boldsymbol{\theta}) = \frac{p(g|\mathbf{x};\,\boldsymbol{\theta}) - p(g)}{1 - p(g)} \tag{8}$$

and then averaging these normalized estimates of group membership for all classes across all test scans in each class, which we will write as $(\langle p_n(g|\mathbf{x};\boldsymbol{\theta})\rangle_{\text{te}})$. Without this rescaling, in the case of no true class structure in $\mathbf{x}$ despite a distinct set of g' in the experimental design we have $\langle p(g|\mathbf{x};\boldsymbol{\theta})\rangle_{te} = \langle p(g')\rangle_{g'}$, which is equal to the "no information" situation that is obtained by randomly guessing class membership independent of $\mathbf{x}$. It is more popular to work with the log-scale prediction error metric $\langle -\log[p(g|\mathbf{x};\theta)]\rangle_{\text{te}}$, known as the deviance or log-loss generalization error (e.g., Heskes, 1998; Ripley, 1998). Kjems *et al.* (2002) use a rescaled and offset version of generalization error, which is equivalent to the mutual information between $\mathbf{x}$ and g and directly interpretable as an information theory bit rate. Such log-scale prediction metrics place a heavy penalty on small posterior probability values for the correct class. In order to retain an intuitive link with the posterior probabilities of class membership, measure prediction accuracy on a bounded [0, 1] scale, illustrate a different metric, and avoid problems with outliers associated with log metrics we have chosen to work with Eq. (8) in this study.

METHODS

Subjects

Fifty-seven normal right-handed volunteer subjects (27 males, 37 ± 8 years; 30 females, 37 ± 9 years) participated in 88 PET scanning sessions 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 (Oldfield, 1971). All female subjects of child-bearing age had a prescan serum pregnancy test.

Data Acquisition, Preprocessing, and Quality Control

All 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 $128 \times 128 \times 31$-voxel grid with $3.125 \times 3.125 \times 3.375$-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 (Strother *et al.*, 1995b). 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 for the hands, the anterior parietal area and superior cerebellum. These coverage criteria were relaxed for the two static-force data sets (SF2 and SF3) for which the superior cortex was completely covered leading to generally poor cerebellar coverage.

Within each of the remaining scanning sessions all possible 6-parameter rigid-body transformations between pairs of scans were computed using AIR 3.08; this represents an empirical implementation of the analytic consensus approach proposed by Woods *et al.* (1998). The 6-parameter rigid-body transformation matrices between any two scans (T_{ij}) were used to obtain a consensus transformation by averaging the 4×4 homogeneous coordinate products, $T_{ij}^k = T_{ik}T_{kj}$, over all values of k to form $\langle T_{ij}^k\rangle_k$. The average transformation matrix was then converted to a 6-parameter rigid body transformation by using $\langle T_{ij}^k\rangle_k$ to transform an evenly spaced $20 \times 20 \times 20$-point grid covering the average brain mask and then calculating the 6-parameter Procrustes transformation of the original to the transformed grid. Within each session this consensus transformation matrix was used to calculate the centroid and maximum movement per voxel over all brain voxels for each scan (Strother *et al.*, 1994), and any subject with one or more scans exhibiting maximum movement/voxel of $>$4.0 mm was excluded. Scans within each session were aligned to obtain the average scan/session, which was then used to calculate the 12-parameter affine transformation to our simulated PET template volume in Talairach space. In order to ensure that alignment parameters were independent of postreconstruction smoothing choices and to minimize end-slice artifacts, the original reconstructed

scans were further smoothed with a 3D 3 × 3 × 3-voxel boxcar filter (a 2D 3 × 3 filter was used for the end slices) and transformed to Talairach space using the 6-parameter rigid-body and 12-parameter affine transformations combined into a single registration operation (Strother *et al.*, 1995b). Finally, an intracerebral-voxel mask volume was created by thresholding *each slice* at 45% of its maximum value and filling any holes within the boundary.

Tasks and Data Sets

The 88 scanning sessions were obtained from 11 task sets of four male and four age-matched female subjects who each participated in one session while performing a particular motor task. Additionally, the eight-subject set for the target interception task provided 2 data-analysis sets leading to the final 12 unique data-analysis sets reported in this study. While all tasks and scanning sessions involved 8 to 12 scans per session, in order to maximize the number of subjects that passed our strict quality control screen, particularly for movement, we included only 4 scans per session in this study. Therefore, each of the 12 data-analysis sets contained eight sessions (one/subject) with each session contributing 4 scans—2 brain states per subject and 2 scans per state.

Speech

Three speech-task sets were obtained from nine subjects (six subjects performed all three tasks, and three subjects performed only two speech tasks). None of these subjects participated in any of the other tasks. The tasks were SP-PA, eight volunteers (four males 37–52 years; four females 23–54 years; 41 ± 12 years) were scanned while they repeated the syllables *pa, ta,* and *ka* as quickly as possible; SP-LC, eight volunteers (four males 37–54 years; four females 24–54 years; 42 ± 12 years) were scanned while they performed repetitive lip closure (as in producing the syllable *pa* silently) as quickly as possible; SP-PH, eight volunteers (four males 42–52 years; four females 23–54 years; 42 ± 12 years) were scanned with sustained phonation while producing the vowel *ah*. All scanning sessions contained four alternating baseline (resting with eyes covered and ears plugged) and activation scans for eight scans/session. The first four scans from each session were selected for this study. See Sidtis *et al.* (1999) for an analysis of the larger data set from which these scanning sessions were drawn.

Tracing (TR)

Eight volunteers (four males 25–42 years; four females 25–45 years; 34 ± 9 years) 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. The first 3 scans per session and the last baseline scan were selected for this study. See Frutiger *et al.* (2000) for an analysis of the larger data set from which these scanning sessions were drawn.

Finger Opposition (FO)

Eight volunteers (four males 25–44 years; four females 27–47 years; 36 ± 8 years) were scanned while performing sequential opposition of the left thumb and successive digits (2, 3, 4, 5, 4, 3, 2, 3, . . .), paced with a 1-Hz auditory signal. Scanning sessions contained 4 or 5 alternating baseline (resting quietly with eyes covered and ears plugged) and activation scans for 8 or 10 scans per session. The first 4 scans per session were selected for this study. See Strother *et al.* (1995b, 1997, 1998), Ardekani *et al.* (1998), and Kustra and Strother (2001) for analyses of related data sets from which these sessions were drawn.

Finger Tapping

Eight volunteers (four males 34–48 years; four females 27–43 years; 38 ± 6 years) were each scanned during two different sessions while performing high (FT-HI) or low (FT-LO) amplitude tapping with the left index finger, respectively. Each session comprised two blocks of five tapping rates in randomized order. Rates were externally paced with an auditory signal of 0, 2/3, 1, 2, and 3 Hz for 10 scans per session. For this study the 0-, 1-, and 3-Hz scans were selected from the first block and combined with the 0-Hz scan from the second block.

Static Force

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). SF2: Eight volunteers (four males 25–44 years; four females 27–48 years; 36 ± 9 years) were scanned during five alternating baseline (no force exerted, eyes open viewing the screen, ears plugged, resting quietly) and static-force activation scans for 10 scans/session. Target force levels of 100, 200, 400, 800, and 1000 g were used in randomized order across subjects. The first 4 scans (2 control and 2 randomized force levels) per session were chosen for this study. SF3: Eight different volunteers (four males 25–35 years; four females 26–44 years; 33 ±

6 years) were scanned with 1 baseline (as for SF2) followed by two blocks of 5 static-force activation scans/block and a final baseline scan for 12 scans/session. Target force levels of 200, 400, 600, 800, and 1000 g were each used once in randomized order within each block. The first 3 scans (1 control and 2 randomized force levels) and the last baseline scan per session were chosen for this study. See Muley *et al.* (2001) for an analysis of the larger data set from which these scanning sessions were drawn.

Mirror Tracing

Eight volunteers (four males 25–42 years; four females 25–45 years; 34 ± 9 years) 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. The larger data set from which these sessions were drawn is reported in Frutiger *et al.* (1998) and Balslev *et al.* (2001).

Target Interception

Eight volunteers (four males 31–45 years; four females 26–49 years; 37 ± 8 years) were scanned while alternately using a cursor or a button with their left hand to intercept a circular moving target within 6 and 12 o'clock zones of an annular path displayed on a rear-projection screen at the foot of the PET scanner couch. Subjects performed two blocks of scans with each block containing the four conditions of a 2 × 2 factorial design for two levels of interception speed (fast/slow) and response type (linear joystick move/button press) presented in randomized order. The first four scans comprising block 1 with all four conditions were selected for this study. TG-SP: The target-interception-speed data-analysis set was defined by choosing the two brain states for data analysis based on fast and slow interception speeds, irrespective of response type. TG-RE: The target-interception-response data-analysis set was defined by choosing the two brain states for data analysis based on joystick-move and button-press responses, irrespective of interception frequency.

Image Data-Analysis Models

For each of the 12 data-analysis sets a raw data matrix, consisting of rows (32 rows = 8 subjects × 4 scans) and columns (number of voxels within the intersection volume of all subject brain masks) was created. Two preprocessing and data-analysis modeling approaches were used: one multivariate based on a PCA/CVA analysis (Appendix) and the other univariate with a standard GLM regression applied to each voxel (Friston *et al.*, 1995b).

Multivariate: VMN-MSR/CVA

Cell (subject × scan × voxel) residual scores were calculated by: (1) dividing each voxel value by the average value across all voxels/row (i.e., volume mean normalization, VMN) and then (2) for each subject, subtracting the average value across scans from each voxel (mean subject removal, MSR). This VMN-MSR preprocessing strategy was designed to maximize sensitivity to within-subject effects while removing individual subject effects. Principal component analysis was then used to decompose the data matrix into orthogonal eigenvectors and associated eigenimages followed by entry of the first P principal components into a CVA. A single canonical eigenvector and eigenimage (SPM_{CVA}) was calculated, which maximized the variance of the two-class mean difference relative to the within-class noise variance (i.e., between-subject and within-state scan variation). This two-state or two-class CVA is equivalent to a Fisher linear discriminant, which generates a single canonical discriminant function (i.e., linear combination of weighted principal components and associated eigenimages, see Appendix).

Univariate: ANCOVA/GLM

A standard GLM t value for the mean difference between the two brain states was calculated for each voxel—producing an SPM{t} image—together with removal of subject block and ANCOVA global scan effects (Friston *et al.*, 1995b). The design matrix comprised a single brain-state column with (0, 1) class-indicator labels, seven subject-block columns, and a column of scan means.

NPAIRS Analysis

Label-Permutation Noise Distributions for Split-Half Metrics

For each eight-subject data-analysis set 10 "permutation-noise" data sets were generated using label permutations under the null hypothesis that brain-state labels were exchangeable. For each subject the labels for one randomly chosen scan from each of the two brain states were exchanged and this was repeated for all eight subjects, 10 times (Holmes *et al.*, 1996). For both models and each of the 12 task-related data-analysis sets these 10 new permuted data sets were each analyzed using NPAIRS to obtain 10 "permutation-noise" values of the median prediction accuracy ($\tilde{p}_n$), median reproducibility histogram ($\tilde{r}$), and median con-

fidence interval ($\widetilde{CI}(Z)_{1-\alpha}$) metrics from their distributions of 70, 35, and 35 split-half resampling values, respectively. These 120 permutation-noise values (10/data set × 12 data sets) were combined to specify error bounds for the split-half resampling metrics.

Scatter Plots and rSPM{Z} Activation Distributions

For each eight-subject data-analysis set 35 rSPM{Z} signal and whole-brain noise distributions were generated using ANCOVA/GLM and VMN-MSR/CVA with six principal components ($CVA_{P=6}$). The 35 rSPM{Z} signal and the 35 noise histograms were each averaged and compared across tasks to examine the stability and shape of the whole-brain signal and noise distributions generated by the split-half scatter-plot technique.

In order to test for changes in activation patterns as a result of using the split-half technique the 35 rSPM{Z} images from each data-analysis set were averaged to form an $\overline{\text{rSPM}\{Z\}}$ image and compared with the SPM obtained from eight subjects for each data-analysis model. For the GLM the SPM{t} from all eight subjects was plotted versus $\overline{\text{rSPM}\{Z\}}$ forming a scatter plot of all brain-voxel values. A PCA of this scatter plot was performed and the slope of the principal axis (i.e., rSPM{Z}/SPM{t}) and the scatter-plot correlation coefficient were recorded. Similarly, eigenimages (SPM_{CVA}) as a function of number of principal components (P = 6, . . . , 14) were obtained from a CVA of all eight subjects and plotted versus $\overline{\text{rSPM}\{Z\}}$ for $CVA_{P=6}$. The highest correlation coefficient and the number of principal components at which this occurred were recorded for each data-analysis set.

Reproducibility Histograms

To test the task dependence of correlation coefficients from split-half scatter plots as a pattern reproducibility metric we plotted reproducibility histograms for all 12 data-analysis sets. We also tested the dependence of these $\tilde{r}$ values on our choice of resampling technique by comparing the split-half results with those from three to five splits where each split involved randomly choosing independent three- and five-subject groups.

Reproducibility Histograms vs rSPM{Z} Distributions' Tails

To test the quantitative inferences that can be made about rSPM{Z} distributions from measurements of $\tilde{r}$, we plotted $\tilde{r}$ vs $\widetilde{CI}(Z)_{1-\alpha}$ with $\alpha = \{0.1, 0.05, 0.01\}$ for both data-analysis models and each data-analysis set. These results were then compared with the theoretical predictions of $CI(Z)_{1-\alpha}$ as a function of r from Eq. (4). Permutation-noise distributions for $\tilde{r}$ and $\widetilde{CI}(Z)_{1-\alpha}$ were also plotted based on the results from label-permutation data sets.

Prediction Accuracy vs Reproducibility Histograms

To obtain data on the relation between prediction accuracy and reproducibility metrics as a function of task and multivariate-model complexity, each of the 12 data-analysis sets was analyzed with $CVA_{P=2}$, $CVA_{P=6}$, and $CVA_{P=10}$. For each of the three CVA models we measured $\tilde{p}_n$ and $\tilde{r}$ and compared plots of the 12 ($\tilde{r}$, $\tilde{p}_n$) pairs for each value of P. Permutation-noise distributions of ($\tilde{r}$, $\tilde{p}_n$) pairs were also plotted based on the results from label-permutation data sets.

We also compared Bartlett's asymptotic χ^2 values [Appendix, Eq. (14)] as a more traditional multivariate metric to see if it provided performance rankings similar to $\tilde{p}_n$ as a function of task and model complexity. For each of the three CVA models applied to each data-analysis set χ^2 values were measured for each of the 70 split-half four-subject groups and the median ($\tilde{\chi}^2$) value was recorded. The $\tilde{\chi}^2$ values from the 12 data-analysis sets were correlated with the 12 $\tilde{p}_n$ values for each of the three levels of CVA model complexity tested.

Subject Influence

We tested the influence of individual subjects on $\tilde{r}$ to see if subjects were contributing equally to the rSPM{Z} results as a function of the data-analysis model used and the task being performed. For each data-analysis set we identified the pair of independent four-subject groups with the highest split-half correlation coefficient and stored the rSPM{Z} from their scatter plot as a reference image. For each of the other 34 pairs of split-half groups the two independent SPMs produced were correlated with this reference image. For each split-half pair, the four-subject group producing the SPM most highly correlated with the reference image was identified and an integer counter for each subject in that group was incremented by 1. If subjects' scanning sessions are truly interchangeable we expect any particular subject to occur randomly in the identified group, i.e., half the time in 34 splits or about 17 times. Large deviations from this average may indicate that the subject's session is either less or more influential than expected under the null hypothesis that subjects are exchangeable across split-half groups. Therefore, we treated the subject counts as relative influence rankings only. For both models we measured $\tilde{r}$ for the six-subject data-analysis sets obtained by removing (1) the two most influential subjects with the highest counts and (2) the two least influential subjects with the lowest counts. For each of the 12 data-analysis sets we then compared the $\tilde{r}$ values across models for the original eight-subject and the two derived six-subject data-analysis sets.

TABLE 1

Scatter Plot Comparisons of Split-Half Reproducible SPMs, i.e., $\overline{\text{rSPM}\{Z\}}$,[a] from NPAIRS versus Standard Eight-Subject Multivariate and Univariate SPMs, SPM_{CVA},[b] and SPM$\{t\}$,[b] Respectively

Data analysis		Data-analysis sets[c]											
Method	Scatter-plot metrics	TG-SP	TG-RE	MT	SF2	SF3	FT-HI	FT-LO	SP-PH	FO	TR	SP-LC	SP-PA
CVA $\overline{\text{rSPM}\{Z\}}$[a] vs SPM_{CVA}[b]	Correlation coefficient	0.94	0.96	0.98	0.98	0.97	0.94	0.96	0.98	0.99	0.98	0.96	0.98
GLM $\overline{\text{rSPM}\{Z\}}$[a] vs SPM$\{t\}$[b]	Correlation coefficient	0.99	0.92	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00
	PCA slope[d]	0.91	0.93	0.88	0.89	0.88	0.87	0.88	0.92	0.88	0.84	0.80	0.90

[a] Split-half resampling with CVA and GLM was used to produce 35 scatter-plot rSPM$\{Z\}$'s, which were averaged for a single $\overline{\text{rSPM}\{Z\}}$ from each eight-subject data set.

[b] Canonical variates analysis (CVA) and the general linear model (GLM) were used to produce a single SPM_{CVA} and SPM$\{t\}$, respectively, from each eight-subject data set.

[c] Each of the 12 eight-subject data-analysis sets (TG-SP, target interception, speed; TG-RE, target interception, reaction type; MT, mirror tracing; SF2 and SF3, static force, Exp. 2 and Exp. 3; FT-HI, finger tapping, high amplitude; FT-LO, finger tapping, low amplitude; SP-PH, speech, phonation; FO, finger opposition; TR, tracing; SP-LC, speech, lip closure; SP-PA, speech, syllable repetition) was analyzed.

[d] The PCA slope from the principal axis of each scatter plot measures the proportional change of the SPM values across the whole brain as rSPM$\{Z\}$/SPM$\{t\}$—for GLM the mean PCA slope value of 0.88 indicates that the average reproducible Z values produced within NPAIRS are typically about 12% less than the standard SPM$\{t\}$ values—see Discussion.

RESULTS

Scatter Plots and rSPM{*Z*} Activation Distributions

The scatter plots shown in Fig. 3 illustrate the extraction of rSPM$\{Z\}$ activation images using $\text{CVA}_{P=6}$. Each plot depicts a single data split for the data set with the lowest (TG-SP) and highest (SP-PA) $\tilde{r}$ values. In Fig. 3A the scatter plot for TG-SP has the circular shape expected for random noise, with $r = 0.0$ indicating that there is nothing similar about the activation eigenimages from the independent split-half groups. This absence of a reproducing activation signal is also reflected in the identical signal and noise histograms in Fig. 3B. These signal (thick solid line) and noise (thick dotted line) histograms were obtained by projecting the scatter-plot voxel values in Fig. 3A onto the major and minor PCA axes, respectively. Moreover, after rescaling of the projected values by the standard deviation of the noise axis both histograms are similar to the almost obscured Gaussian distribution (thin solid line). In contrast, the scatter plot for SP-PA in Fig. 3C has an elongated elliptical shape, with $r = 0.5$, indicating that eigenimages from the independent split-half groups are similar. In Fig. 3D the signal histogram has extended tails (thick solid line), reflecting the elliptical elongation of the scatter plot. The noise histogram (thick dotted line) is again similar to a Gaussian distribution (thin solid line) and to the histograms in Fig. 3B, supporting our assumptions in the derivation of Eq. (4) and the Z-score labels on the horizontal axes.

In Fig. 4 the average of the signal histograms and the average of the noise histograms from 35 split-half scatter plots are overplotted for 4 of the 12 data-analysis sets. Figure 4A illustrates the consistent, approximately Gaussian noise distributions (dotted lines) obtained together with the widely varying reproducible signal distributions. These average signal histograms (colored lines) reflect the varying rSPM$\{Z\}$ distributions of the underlying split-half scatter plots. In the magnified view of the histograms' right-hand tails in Fig. 4B the noise distributions from the four data sets (dotted black lines) are seen to be very similar and approximately Gaussian (solid black line) with slightly heavy tails. Confidence intervals, such as the 99% CI (i.e., $\text{CI}(Z)_{0.99}$) shown for task SF3, are used to summarize the spread of the rSPM$\{Z\}$ distributions in the following figures.

Table 1 indicates that, using either $\text{CVA}_{P=6}$ or GLM, the average reproducible Z-score image, $\overline{\text{rSPM}\{Z\}}$, is very similar to the SPM_{CVA} or SPM$\{t\}$ images produced with a single analysis of all eight subjects, respectively. For SPM_{CVA} the number of principal components that had the maximum correlation with $\overline{\text{rSPM}\{Z\}}$ for $\text{CVA}_{P=6}$ ranged from a minimum of 6 to a maximum of 10 for FT-LO and TG-SP, respectively; the mode for all tasks was 9. For both models all but three of the scatter-plot correlations are greater than or equal to 0.96 (median = 0.99), demonstrating that voxels retain similar relative ordering for the two SPMs being compared. For GLM, $\overline{\text{rSPM}\{Z\}}$ compared to SPM$\{t\}$ is proportionately reduced by a factor of 0.93 to 0.80 (median = 0.88), which may be caused by multiple effects such as random subject effects and spatially varying noise.

Given the lack of a reproducible SPM across split-half pairs for TG-SP (Fig. 3A) it may at first appear surprising that the resulting $\overline{\text{rSPM}\{Z\}}$'s are highly correlated with eight-subject SPMs, SPM_{CVA}, and SPM$\{t\}$. This may be explained by noting that the normalized SPMs from a given split, $\mathbf{z}_1$ and $\mathbf{z}_2$, will both be some-

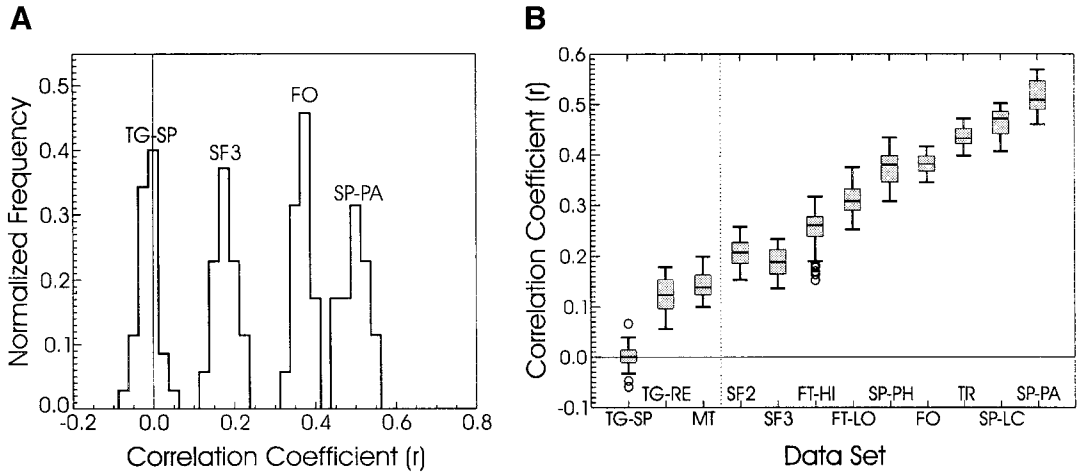

FIG. 5. For 35 split-half scatter plots we illustrate the reproducibility histograms of the 35 correlation coefficients for canonical eigenimages from a two-class canonical variate analysis (see Fig. 2) (A) for the four data sets, TG-SP, SF3, FO, and TR (see Fig. 4), and (B) box–whisker plots for all 12 data sets analyzed (see Methods)—the gray rectangular box represents the lower quartile (lq) to upper quartile (uq) range, which is transected by a black bar at the median. The whiskers reflect the minimum and maximum values within the range [lq − 1.5(uq-lq), uq + 1.5(uq-lq)]. Values outside the whisker range are considered potential outliers and are plotted as individual circles. Note that the horizontal axis in A becomes the vertical axis in B. In B data sets to the left and right of the vertical dotted line used contrasts of active-task scans to active and passive control scans, respectively.

what correlated with the fixed SPM from analyzing all eight subjects, $\mathbf{z}_N$. This will be true even if the input scans contain only random noise. Therefore, $\mathbf{z}_1$ and $\mathbf{z}_2$ may be written as $s_1(\mathbf{z}_N) + \epsilon_1$ and $s_2(\mathbf{z}_N) + \epsilon_2$, respectively, where s_1 and s_2 are scaling factors and ϵ_1 and ϵ_2 are independent samples drawn from a zero-mean random process. The rescaled signal axis from each split can be written as $(\mathbf{z}_1 + \mathbf{z}_2)/(\mathrm{SD}\sqrt{2}) = [(s_1 + s_2)\mathbf{z}_N + (\epsilon_1 + \epsilon_2)]/(\mathrm{SD}\sqrt{2})$, which for random noise processes may be dominated by the zero-mean random noise term $(\epsilon_1 + \epsilon_2)$. However, with averaging over multiple splits the random noise term becomes small, leaving a slightly noisy, scaled version of $\mathbf{z}_N$, resulting in a high correlation between rSPM{Z} and the eight-subject SPMs.

Reproducibility Histograms

Figure 5A illustrates reproducibility histograms for the 35 correlation coefficients from the four data-analysis sets shown in Fig. 4. The $\tilde{r}$ values of the four data sets in Fig. 5A rank themselves in the same order as the reproducible signal tails in Fig. 4, reflecting the monotonic relation between $\tilde{r}$ and the width of the rSPM{Z} distributions, which is analyzed below. Figure 5B illustrates the reproducibility histograms for all 12 data sets depicted as box–whisker plots. The $\tilde{r}$ values range from 0.0 to 0.5 with only one task, FT-HI, with a nonzero median generating outliers (open circles). All of the data sets with values to the right of the vertical dotted line represent contrasts between a primary sensory–motor task and a resting control state, while those to the left of the line represent contrasts between two active task states.

Across the 12 data sets the $\tilde{r}$ values generated by resampling with 3/5 splits are slightly, but significantly, lower than those from the 4/4 splits of split-half resampling (Wilcoxon signed-ranks test for matched pairs (WSRT), $P < 0.01$; paired t test, -0.005 (mean) $\pm$ 0.006 (SD), $P < 0.01$, $n = 12$).

Reproducibility Histograms vs rSPM{Z} Distributions' Tails

Figure 6 illustrates that the predicted relation for $\mathrm{CI}(Z)_{1-\alpha}$ as a function of r in Eq. (4) is quite well matched by the data for both multivariate CVA and univariate GLM model results—note that the plotted curves are theoretical predictions without any free parameters fitted to the experimental data. The error bars ($\pm$2 SD) for the $\widetilde{\mathrm{CI}}(Z)_{1-\alpha}$ and $\tilde{r}$ permutation-noise distributions are plotted in Figs. 6A and 6B based on 110 and 120 median values (10/data set), respectively; the 10 $\tilde{r}$ values for the FT-HI task were excluded as outliers from the SD calculations for the CVA model results. The grand means $\pm$2 SD for the permutation-noise $\tilde{r}$ values are -0.00 ± 0.05 and -0.00 ± 0.06 for the CVA and GLM results, respectively. For the CVA results the grand mean $\pm$2 SD for the permutation-noise values of $\widetilde{\mathrm{CI}}(Z)_{1-\alpha}$ with $\alpha = \{0.1, 0.05, 0.01\}$ are 3.21 ± 0.20, 3.93 ± 0.24, and 5.66 ± 0.34, respectively, and for the GLM results they are 3.26 ± 0.20, 3.94 ± 0.23, and 5.39 ± 0.32, respectively. The theoretical Gaussian values of $\mathrm{CI}(Z)_{1-\alpha}$ for $\alpha = \{0.1, 0.05, 0.01\}$ are 3.29, 3.92, and 5.16, respectively. Comparing the theoretical Gaussian values with the permutation-noise values we see

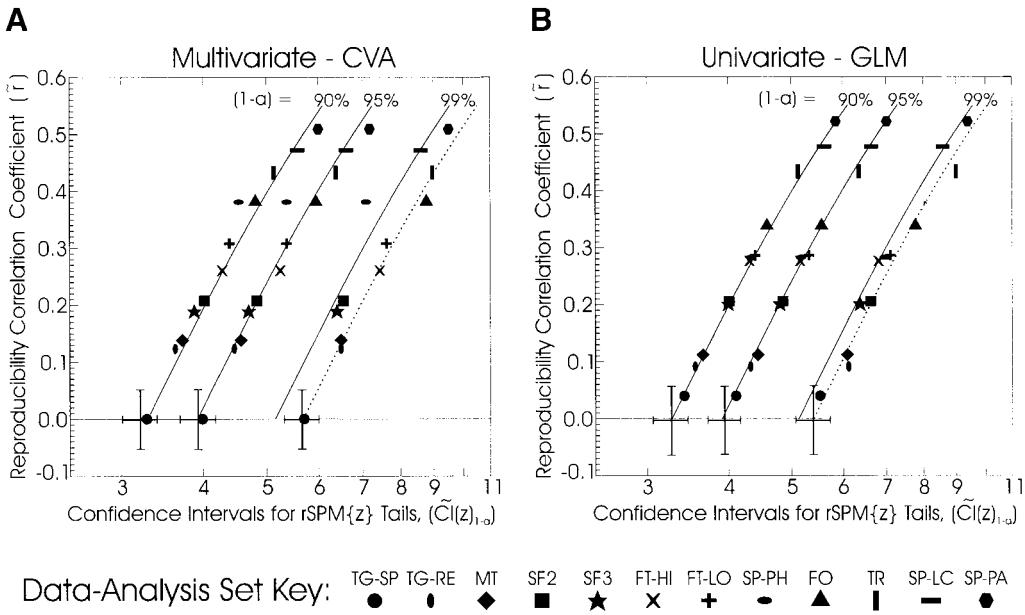

FIG. 6. For split-half resampling of all 12 data sets (see key) we illustrate that the spread of reproducible-signal histograms (from rSPM{Z} in Fig. 3) can be modeled as the sum of a fixed model-dependent noise distribution (at $\tilde{r} = 0$) and a Gaussian signal distribution that scales monotonically with the median reproducibility correlation coefficients ($\tilde{r}$; Figs. 2 and 5). The spread of the tails of the rSPM{Z} histograms are plotted on a log scale as the medians of the distributions of their 90, 95, and 99% confidence intervals, i.e., $\widetilde{CI}(Z)_{1-\alpha}$ for α = 0.1, 0.05, and 0.01. (A) Multivariate discriminant eigenimages from a two-class canonical variate analysis and (B) univariate SPM{t}'s from a general linear model with one design-matrix column for control-activation effects and removal of subject block and ANCOVA global effects. Mean ± 2 SD error bounds for the noise distribution at $\tilde{r} = 0$ were obtained from all 12 data sets with a second-level resampling using label permutations (see Methods). The solid lines are plots of the theoretical relation between $\tilde{r}$ and $log(\widetilde{CI}(Z)_{1-\alpha})$ in Eq. (4), and the dotted lines are the theoretical pairs ($\widetilde{CI}(Z)_{0.99}$, $\tilde{r}$), shifted to the right to intersect the mean value of the permuted noise distributions.

that the noise distributions at $\tilde{r} = 0$ are approximately normal with slightly heavy tails, with CVA having a heavier tail than GLM.

For both CVA and GLM, the experimental ($\widetilde{CI}(Z)_{0.99}$,$\tilde{r}$) pairs follow empirically shifted versions of the theoretical Gaussian curves as illustrated by the dotted lines. These results demonstrate that for each data analysis model and all tasks the spread of the tails of the rSPM{Z} distributions may be modeled as the sum of (1) an approximately Gaussian noise distribution with a slightly heavy tail and (2) a Gaussian signal distribution with a scaling factor determined by $\tilde{r}$. Experimentally we see that (1) the overall deviations from the Gaussian noise model are approximately independent of $\tilde{r}$ across tasks and (2) GLM's local-voxel noise estimates produce a more Gaussian-like noise tail. These observations suggest an underlying mechanism that is spatially dependent and methodological in origin because of its model dependence and task independence, respectively. The most likely methodological candidate is the axial dependence of the spatially varying noise distribution for 3D PET images (e.g., Pajevic *et al.*, 1998).

For CVA in Fig. 6A the ($\widetilde{CI}(Z)_{0.99}$, $\tilde{r}$) pairs from the static force (SF2 and SF3) and speech (SP-PH, SP-LC, SP-PA) tasks fall approximately on the theoretical Gaussian curve (solid line). The ($\widetilde{CI}(Z)_{0.99}$, $\tilde{r}$) pairs for all the other tasks lie almost exactly on the empirically shifted curve for a heavy-tailed noise distribution. These task subgroups in the CVA results may be explained by the presence or absence of sensorimotor and cerebellar activations for the hand appearing in the much noisier 3D image slices near the edge of the axial field of view (FOV). In order to cover both the sensorimotor cortex for the hand and the superior cerebellum with our PET scanner it was necessary to allow these two regions to lie near the edges of our limited 10.8-cm axial FOV. The repositioning of the FOV for the static force data sets (see Methods) moved the primary sensorimotor activations away from the edge of the FOV and the enhanced image noise in those slices. While the FOV for the speech data sets was positioned in the same way as for the majority of the other motor tasks, the primary speech activations do not lie near the edge of the FOV. Therefore, unlike the other tasks the static force and speech tasks do not have primary activations that are impacted by the enhanced image noise near the edge of the FOV. These results indicate that the extraction of a noise distribution within the NPAIRS environment may be sensitive to the local,

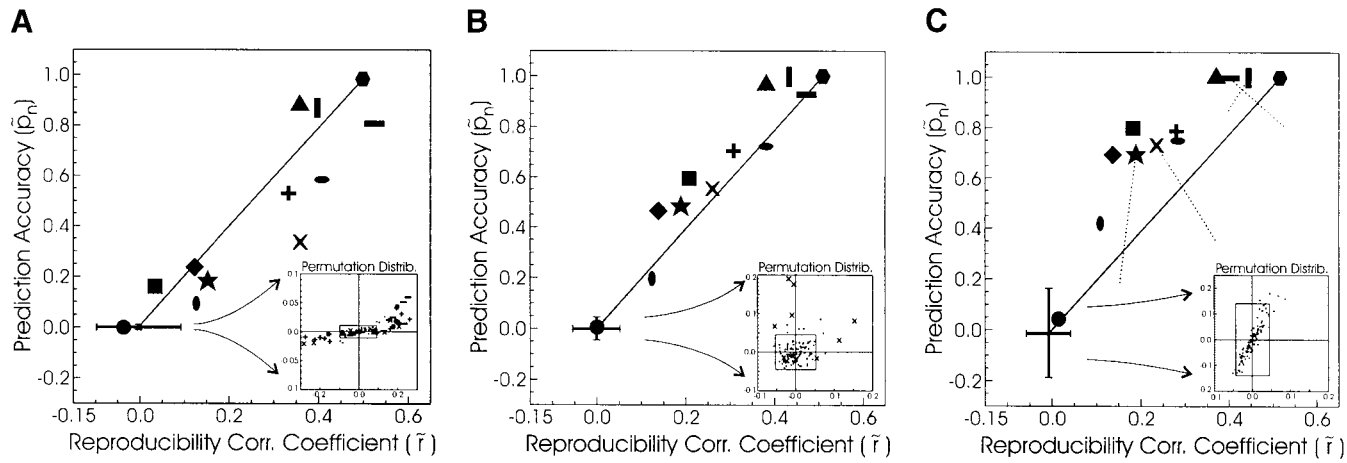

FIG. 7. For split-half resampling of all 12 data sets (see key, Fig. 6) we illustrate the relationship between prediction and reproducibility performance metrics as a function of model complexity. We plot the median prediction accuracy [$\tilde{p}_n$; Eq. (8)] versus the median reproducibility correlation coefficients ($\tilde{r}$; Fig. 2) for discriminant eigenimages from a two-class canonical variate analysis built on a principal components analysis using (A) 2 principal components (least complex), (B) 6 principal components, and (C) 10 principal components (most complex). Mean ± 2 SD error bounds for the noise distribution at $\tilde{r} = 0$ were obtained from all 12 data sets with a second-level resampling using label permutations (see Methods). The solid line provides a reference and joins the stable points defined by the SP-PA speech task and the mean of the permutation noise distributions. The four dotted lines in C illustrate the change in performance metrics from the least to the most complex CVA for tasks: SF3, FT-HI, TR, and SP-LC. The inset permutation distribution plot in each graph shows the 120 permutation values (10 sets of label permutations/data set) with dots representing the values used to generate the error bounds (±2 SD shown as box) and larger task symbols representing outliers that were not included in the permutation error bounds.

model-dependent noise properties of the primary task activations. Such static force and speech subgroups are not seen in the GLM noise tails, providing further support for our interpretation.

Prediction Accuracy vs Reproducibility Histograms

In Figs. 7A, 7B, and 7C we have plotted $\tilde{p}_n$ vs $\tilde{r}$ for all 12 data-analysis sets as a function of $\text{CVA}_{P=2}$, $\text{CVA}_{P=6}$, and $\text{CVA}_{P=10}$, respectively. In general $\tilde{p}_n$ and $\tilde{r}$ increase together across the data sets but the relative difference between any two data sets is strongly dependent on model complexity, which affects $\tilde{p}_n$ and $\tilde{r}$ quite differently. Increasing the number of principal components, P (Figs. 7A–7C) significantly increases $\tilde{p}_n$ for all data sets but TG-SP, which is never distinguishable from noise—the diagonal black line provides a stable reference from graph to graph against which to compare individual data set changes as a function of P. As $\tilde{p}_n$ increases with increasing P, $\tilde{r}$ both increases and decreases depending on the data set. The dotted lines in Fig. 7C indicate relative changes from $P = 2$ (Fig. 7A) to $P = 10$ (Fig. 7C) for (1) two data sets with *decreasing* $\tilde{r}$ as a function of P (FT-HI, X, and SP-LC, horizontal rectangle) and (2) two data sets with *increasing* $\tilde{r}$ as a function of P (SF3, star, and TR, vertical rectangle). Changes in rSPM{Z} voxel values for these four data sets as a function of model complexity are examined in Fig. 8.

Figure 7 displays classic bias–variance tradeoffs for $\tilde{p}_n$ and for the interaction of $\tilde{p}_n$ and $\tilde{r}$ as a function of model complexity. The following interpretation assumes that the permutation error bars displayed at (0, 0) in each graph are approximately valid for all ($\tilde{r}$, $\tilde{p}_n$) data points in that graph. In Fig. 7A, for $P = 2$ there are many biased (low) $\tilde{p}_n$ values measured with high precision (i.e., a small permutation error bar). With increasing P the $\tilde{p}_n$ values increase, becoming less biased, but their precision becomes increasingly worse so that the bias–variance tradeoff for prediction values appears optimal between 6 and 10 PCs (Figs. 7B and 7C). The size and precision of the $\tilde{r}$ values display the opposite behavior as a function of model complexity. However, we must not equate size and bias for the $\tilde{r}$ values because their true values are unknown. In Fig. 7A, for $P = 2$ there are many relatively large $\tilde{r}$ values (i.e., with high rSPM{Z} voxel values) measured with low precision (i.e., a large permutation error bar). With increasing P the $\tilde{r}$ values both increase and decrease, but their precision improves considerably for $P = 6$ with little change for $P = 10$. Therefore, if we focus only on maximizing $\tilde{r}$ as a performance metric the resulting rSPM{Z} values are likely to be unreliable (i.e., low precision) and associated with models that poorly predict the experimental design matrix. If instead we choose models that predict the design matrix well they are likely to have more reliable rSPM{Z} values, which are not necessarily the largest possible, with the caveat that we will be less certain of the reliability of our prediction measurements.

Insets in each graph of Fig. 7 illustrate the 10 median results for the permutation-noise data from each of the 12 data-analysis sets. The results used to calculate the rectangular box representing ±2 SD error bars

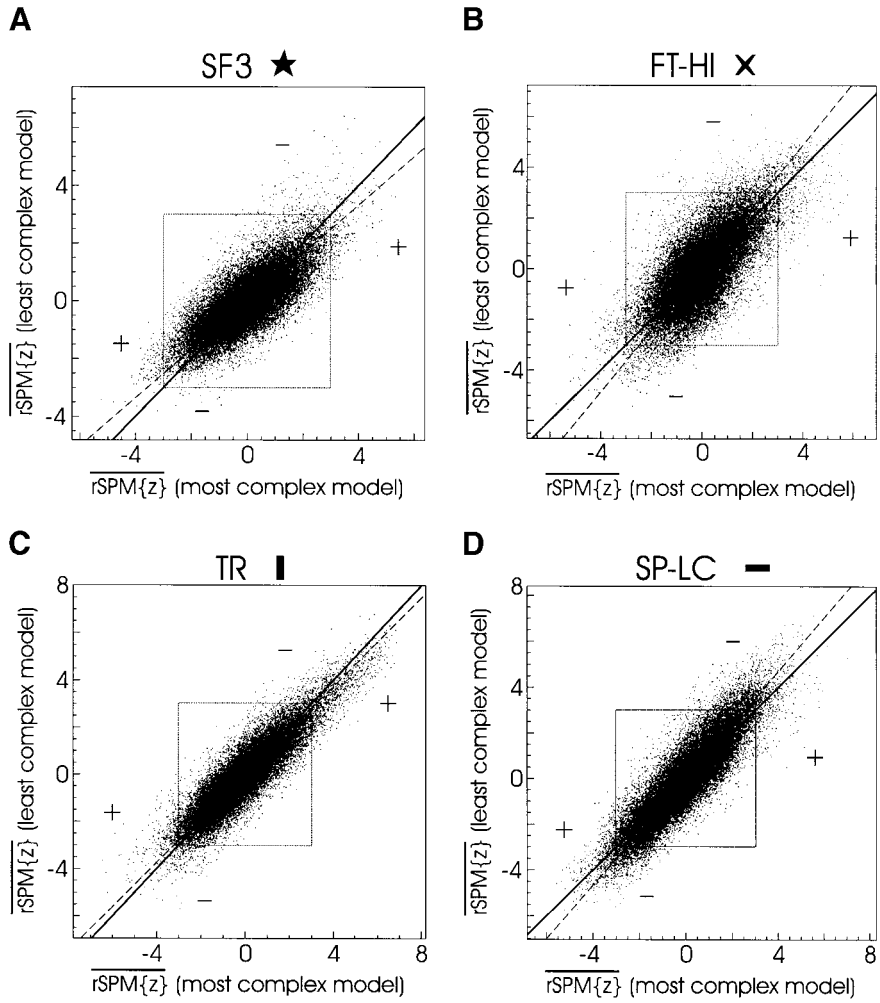

FIG. 8. Scatter plots illustrating both increasing and decreasing voxel Z scores with increased model complexity. For split-half resampling of the four data sets identified in Fig. 7C (dotted lines) the average reproducible eigenimage ($\overline{\text{rSPM}\{Z\}}$) from a two-class CVA built on a 2-component PCA subspace (least complex) is plotted against $\overline{\text{rSPM}\{Z\}}$ from a CVA with a 10-component PCA subspace (most complex): (A) static force 3, (B) finger-tapping with high amplitude, (C) figure tracing, and (D) a lip-closure speech task. For each scatter plot the line of identity (solid line) and principal axis from a PCA regression (dashed line) are overlaid together with a box bounding $Z = \pm 3$. Voxels with proportionately increasing or decreasing Z scores relative to the line of identity are marked by + and −, respectively.

are displayed as small dots, and those from data-analysis sets excluded from the calculation as outliers are displayed as medium-size task symbols. Four tasks were excluded as outliers in calculating the error bounds for Fig. 7A (FT-HI, X; FT-LO, cross; SP-LC, horizontal rectangle; and SP-PA, hexagon), one task was excluded for Fig. 7B (FT-HI), and there were no tasks excluded as outliers for Fig. 7C.

Barlett's $\tilde{\chi}^2$ metric is significantly correlated with $\tilde{p}_n$ across the 12 data sets with correlation coefficients of 0.98, 0.91, and 0.80 for $P = 2$, $P = 6$, and $P = 10$, respectively. The decreasing correlation with increasing model complexity may be partly driven by the fact that while $\tilde{p}_n$ is bounded ($\tilde{p}_n \in [0, 1]$) $\tilde{\chi}^2$ is unbounded ($\tilde{\chi}^2 \in [0, \infty)$). With increasing P, as $\tilde{p}_n \rightarrow 1.0$ for some data-analysis sets, the corresponding $\tilde{\chi}^2$ values also increase but disperse, perhaps reflecting data-set-dependent increases in degrees of freedom ($f \neq P$) with increasing model complexity since the χ^2 mean and variance are f and $2f$, respectively.

In Fig. 8 the scatter plots illustrate changes in $\overline{\text{rSPM}\{Z\}}$ for the least complex ($\text{CVA}_{P=2}$) compared to the most complex ($\text{CVA}_{P=10}$) PCA/CVA models for each of the four data sets with dotted lines displayed in Fig. 7C. Each scatter plot is overlaid with the line of identity (solid line) and the principal axis from a PCA regression (dashed line). Figures 8A and 8C have principal-axis slopes of less than 1.0 and are associated with increases in $\tilde{r}$ in Fig. 7C, while Figs. 8B and 8D have principal axis slopes of greater than 1.0 and are associated with decreases in $\tilde{r}$ in Fig. 7C. Thus an increase or decrease in $\tilde{r}$ is associated with a proportional increase or decrease in whole-brain $\overline{\text{rSPM}\{Z\}}$ values as a function of increasing P, respectively. For

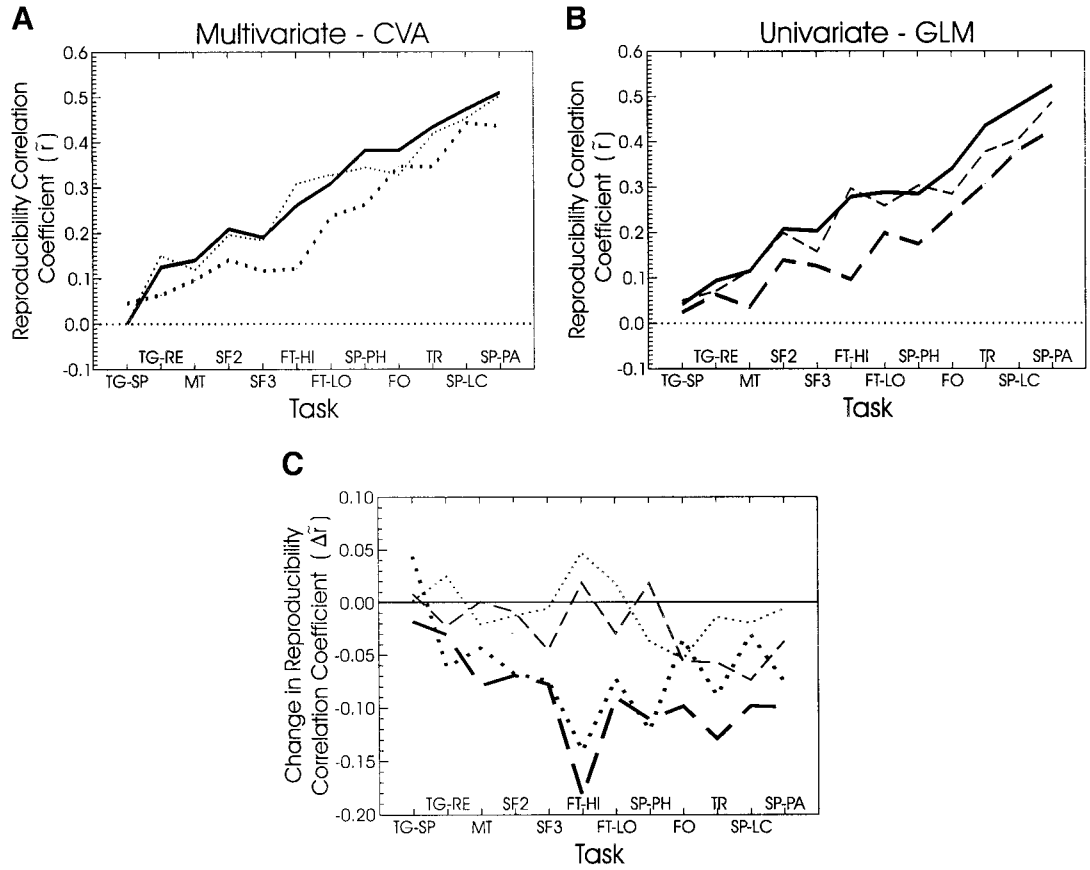

FIG. 9. For split-half resampling of 12 data sets the effects of subject influence are illustrated with plots of the median reproducibility correlation coefficients for eight-subject groups (solid black line), six-subject groups after removal of the two least influential subjects (thin dotted and dashed lines), and six-subject groups after removal of the two most influential subjects (thick dotted and dashed lines). (A) Multivariate discriminant eigenimages from a two-class canonical variate analysis and (B) univariate SPM{t}'s from a general linear model with one design-matrix column for control-activation effects and removal of block-subject and ANCOVA global effects. (C) The change in median values ($\Delta\tilde{r}$) for the two six-subject groups compared to the eight-subject group. Note the differences across models and tasks for FO, TR, SP-LC, and SP-PA compared with the other eight tasks.

the 11 data-analysis sets with $\tilde{r} \neq 0$, changes in $\tilde{r}$ for $P = 10$ compared to $P = 2$ are highly correlated with such proportional changes in whole-brain rSPM{Z} values ($\rho = 0.97$). From this PCA regression a change in $\tilde{r}$ of +0.01 units is associated with a proportional change in rSPM{Z} scores of +1.75%.

In each graph of Fig. 8 all voxels with absolute rSPM{Z} values greater than 3.0 lie outside the square and have either increased (+) or decreased (−) values as a result of increasing model complexity. These results demonstrate that neither the number of voxels with large absolute rSPM{Z} values nor the largest such values should be used to optimize methodological choices. In Fig. 8A the largest rSPM{Z} values occur for $P = 2$, but the principal-axis slope of 0.835 indicates a proportional increase in absolute rSPM{Z} values of 20% for $P = 10$, which is coupled with a large increase in prediction accuracy (see Fig. 7C). The large Z scores for $P = 2$ in Fig. 8A are a result of noisy Z-score estimates, not a better model, and reflect the bias–variance tradeoffs seen in Fig. 7.

Figures 8A and 8C illustrate that the PCA basis may be quite efficient for some tasks. The large increases in prediction accuracy seen in Fig. 7C (dotted lines) for SF3 and TR are obtained by improving overall rSPM{Z} values while even further enhancing a subset of voxels in Figs. 8A and 8C, respectively. In contrast, Figs. 8B and 8D illustrate that the large increases in prediction accuracy seen in Fig. 7C (dotted lines) for FT-HI and SP-LC are obtained by enhancing the rSPM{Z} values of a small subset of voxels while simultaneously decreasing the majority of rSPM{Z} values by 18 and 14%, respectively. As a result the PCA basis may be suboptimal for the FT-HI and SP-LC tasks as it requires a decrease in most rSPM{Z} voxel values in order to enhance a subset of voxels that result in improved prediction accuracy.

Subject Influence

In Fig. 9 we compare the reproducibility of the standard eight-subject data-analysis sets (thick solid lines)

with the two six-subject sets created by removing the two most influential or the two least influential subjects. In Figs. 9A and 9B, for $CVA_{P=6}$ and GLM results, respectively, the $\tilde{r}$ values of the data sets without the two most influential subjects (thick dashed and thick dotted lines) lie significantly below those from the eight-subject data sets or the data sets without the two least influential subjects (thin dashed and thin dotted lines).

The GLM and CVA results for the eight-subject data sets are not significantly different, supporting their use as reference levels in Fig. 9C (WSRT, $P > 0.1$; paired t test, 0.011 ± 0.036, $P > 0.3$, $n = 12$). For the six-subject results the GLM $\tilde{r}$ values are significantly lower than the CVA values for both removal of the two least influential subjects (WSRT single sided, $P < 0.01$; 0.028 ± 0.034, $P < 0.015$, $n = 12$) and removal of the two most influential subjects (WSRT single-sided, $P < 0.005$; 0.038 ± 0.036, $P < 0.004$, $n = 12$). These results indicate that subject influence is somewhat stronger for GLM than for CVA.

Figure 9C indicates that subject influence is different in the first eight data sets with smaller $\tilde{r}$ values compared to the last four data sets with the largest $\tilde{r}$ values. For the *first eight data sets,* the combined CVA and GLM results after removal of the *two least influential* subjects are not significantly different from the *standard eight-subject* results (WSRT, $P > 0.1$; 0.003 ± 0.025, $P > 0.7$, $n = 16$). For the *last four data sets,* the combined CVA and GLM results after removal of the *two most influential* subjects are significantly lower than the *standard eight-subject* results (WSRT one sided, $P < 0.004$; 0.040 ± 0.024, $P < 0.002$, $n = 8$). In addition, for the *last four data sets,* the combined CVA and GLM results after removal of the *two most influential* subjects are significantly lower than those after removal of the *two least influential* subjects (WSRT one sided, $P < 0.008$; 0.042 ± 0.034, $P < 0.010$, $n = 8$). Figure 9C also indicates that for the last four tasks the effects of subject influence are greater in the GLM compared to the CVA results. Overall, these results demonstrate that subject influence is a significant factor in all tasks with $\tilde{r} \neq 0$, but the relative impact of the least and most influential subjects is a function of the data-analysis model, reproducibility SNR, and task.

DISCUSSION

For functional neuroimaging studies from multiple tasks we have demonstrated how to explore and characterize reproducible activation signal structure and the associated noise distributions. We do this without using spatially localized measurements and the model-specific thresholding schemes or neuroscientific hypotheses that often accompany them. We have introduced the NPAIRS testing framework that (1) is dependent only on choices within the functional neuroimaging data chain and (2) initially uncouples the testing of the "quality" of functional neuroimaging results from interpretations based on the neuroscientific knowledge base and associated spatial hypotheses. When optimizing the functional neuroimaging chain our goal is to avoid any systematic bias due to hypotheses based on prevailing neuroscientific expectations. The NPAIRS framework achieves this by defining global metrics that measure the ability of the neuroimaging chain to produce model parameters from a training data set that can accurately predict experimental-design parameters and reproduce the associated activation patterns in an independent test data set. While we have emphasized the application of NPAIRS metrics from whole-brain SPMs the framework can also be applied to obtain spatially localized measurements.

The sums and differences formulation for extracting reproducible signal and noise estimates, respectively, from scatter plots may be applied to two independent SPM measurements of a single voxel (see Theory). However, useful variance estimates of the noise will require at least four or five independent voxel locations and it may be necessary to include a number of "noise-only" spatial locations in order to obtain reproducibility metrics with a useful dynamic range as a function of methodological choices. Such localized estimates may then be compared as a function of data-analysis models, tasks, etc., although they are likely to be much noisier than the whole-brain values studied here. Use of the NPAIRS framework with spatially localized regions is a topic for future research.

We chose split-half resampling for NPAIRS because it incorporates cross-validation resampling of equal-sized training and test sets from groups of even numbers of subjects. Our results indicate that it may be possible to use asymmetric data splits with only a very small loss in power, allowing odd numbers of subjects to be tested within the same framework. NPAIRS allows the entire imaging chain including data-analysis models to be tested and optimized using both prediction and reproducibility metrics within a single resampling framework. For the subproblem of optimizing and selecting among data-analysis models there are analytic alternatives such as the Akaike information criterion (AIC). However, Hansen *et al.* (1999) have shown that AIC estimates of the number of significant components to retain after a PCA are over optimistic compared to resampled cross-validation estimates—see Ripley (1998) for a discussion of the assumptions underlying AIC and related model-selection metrics. Alternative resampling approaches to the problem of statistically characterizing SPMs have used label permutation (Arndt *et al.,* 1996; Holmes *et al.,* 1996; McIntosh *et al.,* 1996; Ardekani *et al.,* 1998) and bootstrap (McIntosh *et al.,* 1999) techniques in order to perform spatial signal detection based on inferential tests of

nonparametric distributions (see the review by Petersson *et al.*, 1999a). Within the NPAIRS framework these techniques may be viewed as a second level of resampling that is part of the data-analysis model specification to be applied to the independent split-half groups. In this way pattern reproducibility may be tested for model specifications that include resampling procedures. While computationally expensive, our results demonstrate that it is quite feasible to perform a second, nested level of resampling based on label permutations. Advantages of such second-level resampling within NPAIRS are the ability to test (1) the accuracy of spatial inferences made with rSPM{Z} values based on their resampled null distributions if it is computationally feasible to generate at least several 100 second-level permutation samples and (2) the exchangeability requirement for exact, unbiased permutation test results (Good, 1994). The second point is illustrated by the median values for the FT-HI data set in the permutation distribution inset in Fig. 7B (crosses); unlike the other data sets, particular permutations of the FT-HI labels result in unusually large prediction and/or reproducibility median values, indicating that the labels may not be treated as exchangeable.

A key feature of the NPAIRS framework is the reproducibility metric based on correlation coefficients of split-half scatter plots. We have shown that these $\tilde{r}$ values summarize a reproducibility SNR based on the PCA of the split-half scatter plots. This allows reproducible activation-signal and uncorrelated noise distributions to be obtained from any model's SPM output and converted to a common statistical scale. This model-independent mechanism for generating signal and noise distributions is particularly useful for the adaptive multivariate models we are most interested in (e.g., CVA eigenimages) and for generic techniques for extracting SPMs from broad classes of models (e.g., the sensitivity map in Kjems *et al.*, 2002), which do not have parametrically defined noise distributions. In addition, the rSPM{Z} and $\overline{\text{rSPM}\{Z\}}$ images produced by NPAIRS may be analyzed using standard random field theory techniques (Worsley *et al.*, 1996, 1999). This allows us to directly compare spatial inferences from t maps with those from eigenimages (e.g., Shaw *et al.*, 2002). The conversion to a common Z-score scale appears to work quite well with some evidence for spatially heterogeneous noise distributions in comparisons of GLM and CVA model results. The shape of the CVA noise distributions from the minor PCA axis seen in Figs. 3, 4, and 6 may be explained by assuming that a small number of voxels have higher than average variance. The pooled variance estimate will then be too large for most voxels, overcorrecting them toward zero to create a higher than Gaussian central peak, and too small for the minority of voxels with high variance, undercorrecting them to leave an extended non-Gaussian tail. The results in Fig. 6 indicate that these high variance voxels probably arise from strong activations that occur in image slices near the edge of the axial FOV.

For each task the average of the rSPM{Z}'s from multiple split-half pairs (i.e., $\overline{\text{rSPM}\{Z\}}$) is very similar to the SPM obtained from a single application of each data-analysis model to each eight-subject data set (Table 1). For the GLM results the $\overline{\text{rSPM}\{Z\}}$ values are 10–20% lower than the t values from the eight-subject SPM{t} image. If the GLM t-test model is valid, the variance of the voxel t values in the split-half SPM{t} images should be stable with a value of $f/(f-2)$, where f represents the degrees of freedom for each four-subject group. This spatially homogeneous variance makes a global variance estimate from the PCA noise axis a good choice for rescaling the scatter-plot comparison of GLM results. We expect additional noise from random subject effects between the independent split-half groups. This will increase the variance of the PCA noise axis and may be the cause of the observed decrease in the $\overline{\text{rSPM}\{Z\}}$ values relative to the eight-subject SPM{t} values. Figure 6 clearly shows that the variance stabilizing effect of the GLM has been effective, creating almost Gaussian noise distributions compared to the longer noise tails from the majority of the CVA results.

The experimental results in Fig. 6 compared with the theoretical predictions of Eq. (4) demonstrate the possibility of developing analytic models of the whole-brain signal and noise structure. We have shown that NPAIRS measures model-dependent noise distributions that are sensitive to the local spatial noise levels of the major activation foci. In addition, values of the reproducibility metric quite accurately summarize the shape of the rSPM{Z} histograms. These results, together with those in Fig. 5 defining separate ranges of reproducibility SNRs for resting and active control state studies, indicate a potentially fruitful research area based on quantitatively summarizing whole-brain activation signal and noise distributions across multiple tasks. These results may be used (1) to guide the selection of optimal combinations of task, experimental design, and data-analysis model; (2) as constraints for developing more realistic simulation studies; and (3) as possible priors for Bayesian analysis techniques. Moreover, because the NPAIRS framework involves comparisons of independent groups of subjects it provides an empirical means of adjusting any model for random subject effects and for testing the utility of particular fixed- versus random-effects model assumptions (e.g., Petersson *et al.*, 1999b). The extracted rSPM{Z} image from any split-half pair of groups has been implicitly adjusted for subject-dependent differences between the groups, which are incorporated into the noise distribution of the PCA noise axis. We have recently shown how to use this feature of split-half resampling to assess significant reproducibility for group comparisons

in which random subject effects are most problematic (Strother *et al.*, 1999; Shaw *et al.*, 2002).

Another key feature of the NPAIRS framework is the ability to define a ROC-like plot for studying optimization of the functional neuroimaging data chain. By plotting prediction accuracy vs pattern reproducibility on [0, 1] scales we obtain a 2D space in which optimization implies moving toward the upper right-hand corner for perfect prediction with an infinite SNR, i.e., toward $(\tilde{r}, \tilde{p}_n) = (1, 1)$. We have demonstrated that it is worthwhile to optimize prediction accuracy (i.e., $\tilde{p}_n \rightarrow 1.0$) because, in addition to providing a better match to the experimental design conditions, the bias–variance trade-offs between $\tilde{p}_n$ and $\tilde{r}$ result in more precise rSPM{Z} values (Fig. 7). However, we have shown that when tuning a PCA basis optimizing $\tilde{p}_n$ may be associated with both decreases and increases in $\tilde{r}$ and the associated rSPM{Z} values (Fig. 7C, dotted lines). Based on Figs. 8B and 8D, $\tilde{r}$ decreases when $\tilde{p}_n$ increases because only a subset of voxels' rSPM{Z} values increase while the SNR of most voxels decreases. In an extreme case a multivariate model need reflect only the single location with the best $\tilde{p}_n$ while becoming insensitive to (e.g., downweighting) values at all other locations. This is the reason that prediction accuracy alone is not a sufficient optimization metric for multivariate techniques. It is an open research issue whether $\tilde{p}_n$ alone becomes a sufficient optimization metric when fitted on a voxel-by-voxel basis (e.g., see Kjems *et al.*, 2002; Goutte *et al.*, 2001) or for individual basis components (McKeown, 2000). Figures 8A and 8C illustrate that for some tasks it is possible to simultaneously increase $\tilde{r}$ and $\tilde{p}_n$ by tuning a PCA basis. For these tasks PCA/CVA can be tuned to provide general noise filtering while simultaneously providing an even larger SNR boost for task-dependent regions. This example demonstrates that by making methodological choices that simultaneously increase $\tilde{p}_n$ and $\tilde{r}$ we may enhance our ability to find subtle activation effects near the detection threshold. These examples demonstrate the potential advantages of using movement toward (1, 1) on a $(\tilde{r}, \tilde{p}_n)$ plot as a metric for selecting optimal methodological choices.

Nevertheless, plots of $(\tilde{r}, \tilde{p}_n)$ pairs do not provide absolute optimization criteria as there are a number of open research issues. Assume we identify two new sets of methodological choices that both simultaneously increase $\tilde{r}$ and $\tilde{p}_n$ relative to the choices in our standard processing chain. Relative to each other, one of the new sets has larger $\tilde{r}$ values and the other set larger $\tilde{p}_n$ values. Which of these should we choose as the optimal set of methodological choices, or equivalently, how should we measure distance from (1, 1) in the $(\tilde{r}, \tilde{p}_n)$ space? We know that the prediction measurements are unbiased against the assumed truth of the experimental design but the larger $\tilde{p}_n$ values are less precise than smaller values. In contrast the reproducibility measurements directly reflect the SNR of the pattern we wish to interpret but they may contain a significant undetectable bias. We do not know what the true activation pattern should look like, and the SNR and resulting reproducibility measures can often be increased by introducing some bias, e.g., with spatial smoothing. One possible approach is to develop the consensus SPM techniques outlined by Hansen *et al.* (2001), in which consensus is taken across subsets of methodological choices identified within the NPAIRS framework using $(\tilde{r}, \tilde{p}_n)$ plots.

Our preliminary results indicate that, in addition to being unbounded, standard multivariate distribution measures such as Bartlett's χ^2 (and by implication Wilke's lambda, e.g., Mardia *et al.*, 1979) may be too variable to be used as quality metrics as proposed by Friston *et al.* (1996) and discussed in Petersson *et al.* (1999b). Bullmore *et al.* (2000) have recently addressed the similar case of using χ^2 as a metric for fitting path analysis models, and because of potential problems with such asymptotic inferential statistics in small data sets have advocated an alternative using resampling techniques in the spirit of the NPAIRS framework. While a number of parametric and nonparametric performance metrics that relate to prediction of the design matrix have been proposed (Ripley, 1998; Kjems *et al.*, 2002), the reproducibility metric outlined in the study appears to be unique and provides important performance information that complements that available from prediction metrics. NPAIRS reproducibility measurements have been used to evaluate changes in experimental design (Muley *et al.*, 2001), detect significant multidimensional results (Strother *et al.*, 1999; Frutiger *et al.*, 2000), evaluate within-subject reproducibility for BOLD fMRI (LaConte *et al.*, 2001), and compare univariate versus multivariate analysis techniques for the analysis of abnormal and normal groups (Shaw *et al.*, 2002).

We have used the NPAIRS framework to measure relative subject influence in terms of changes in $\tilde{r}$, and hence $\overline{\text{rSPM}\{Z\}}$, per subject. The motivating question in terms of a "reproducibility learning curve" plotting $\tilde{r}$ as a function of the number of subjects in a data set is, does the order in which subjects are added to the data set (e.g., experimentally collected) matter? The answer for both GLM and CVA in Fig. 9 is an unqualified yes, with removal of the two most influential subjects in each group being responsible for reductions in $\tilde{r}$ of -0.05 to -0.15 with proportional decreases in rSPM{Z} of 9 to 27%. In contrast, for the eight-subject data sets with the lowest $\tilde{r}$ values and smallest rSPM{Z}'s, removing the two least influential subjects does not significantly change $\tilde{r}$ for each data set. The four data sets with the highest $\tilde{r}$ values and largest rSPM{Z}'s (FO, TR, SP-LC, SP-PA) behave differently from the other eight data sets (Fig. 9C). For these four data sets the $\tilde{r}$ values after removal of the two least or two most in-

fluential subjects are more similar than those of the other eight data sets. This indicates that these four groups are more homogeneous across subjects than the other eight groups, and, in addition, their results are more strongly influenced by the data-analysis model, i.e., GLM is more susceptible to subject influence than CVA. We speculate that these differences across data sets and models represent interactions between basic activation signal structure, signal-to-noise levels, subject-dependent signal variation, and the way each model responds to random subject effects. For example, CVA has an ability to fit subject-dependent signal variation and provides an approximation to a subject random-effects model, which depends on the class structure (Kustra, 2000). The standard GLM has neither of these features, which may explain its susceptibility to subject influence given a large-enough reproducibility SNR. Important issues for future research are (1) the relation between the NPAIRS influence measures and more traditional random-effects models and measures of sample heterogeneity (e.g., Biggerstaff and Tweedie, 1997), (2) how these issues relate to the conjunction analysis proposed by Friston *et al.* (1999), and (3) can the NPAIRS influence measure be used to distinguish subgroups that are relatively homogeneous from subjects that are unique in some way or from subgroups that are heterogeneous.

An important issue that we have not addressed in this paper is the variation of NPAIRS results as a function of the number of subjects. For N subjects NPAIRS is easily applied to split-half group sizes from 1 to $N/2$ subjects (Strother *et al.*, 1997, 1998, 2000). In general we believe that for groups larger than eight subjects we would increase prediction and/or reproducibility metric values, reduce our error bars, and potentially reorder tasks as a function of $\tilde{r}$ in Fig. 5. Moreover, while quantitative details are likely to be a strong function of N (i.e., the number of PCs required to produce a particular increase in $\tilde{p}_n$) our demonstration of model-dependent and task-independent noise structure in Fig. 6 and task-dependent bias–variance tradeoffs in Figs. 7 and 8 are unlikely to depend on group size. However, individual subject influence and loss in power due to asymmetric data splits are likely to decrease with increasing group size. "Learning curves" that plot prediction, reproducibility, and potentially other metrics as a function of the number of subjects in the training set may be used as empirical estimates of these effects. For most models applied to finite data sets analytic power estimates are not available (e.g., Megalooikonomou *et al.*, 2000) and learning curves provide a means of estimating the signal-to-noise impact of collecting additional subjects. For four of the tasks in this paper the variation of reproducibility and prediction metrics across tasks as a function of N and CVA model complexity is addressed in the companion paper by Kjems *et al.* (2002).

CONCLUSIONS

We have introduced NPAIRS, a nonparametric resampling framework that extends the idea of measuring prediction accuracy using cross-validation resampling for independent training and test sets. Our extensions include ensembles of split-half cross-validation estimates to measure reproducibility metrics, activation patterns on a common statistical scale, and subject influence. For both univariate and multivariate data-analysis models applied to 12 [^{15}O]water PET data sets from diverse motor tasks we have used NPAIRS to (1) directly measure reproducible activation signal-to-noise ratios on a common Z-score scale while incorporating spatially varying noise and random subject effects, (2) demonstrate that the histogram of an rSPM$\{Z\}$ image volume may be modeled as the sum of a data-analysis-dependent noise distribution and a task-dependent Gaussian signal distribution, (3) explore the relation between prediction accuracy and activation pattern reproducibility in real data sets as an alternative to ROC curves based on simulations, and (4) quantitatively measure the broad spread of reproducible activation SNRs and the strong influence of individual subjects on the reproducible activation patterns from the 12 data sets. We propose the NPAIRS framework as a validation tool for testing and optimizing methodological choices and tools for data acquisition, preprocessing, data analysis, and extraction of statistical parametric maps in functional neuroimaging.

APPENDIX: CVA ON A PCA BASIS

Let $\mathbf{X}_{\text{tr}}$ be the set of M training scans ($M < S$), where S is the total number of scans available, and test and training sets are chosen by splitting the available number of subjects (N) to ensure that the two groups of scans are truly independent. Using a PCA or equivalently a singular value decomposition, we obtain

$$\mathbf{X}_{\text{tr}} = \mathbf{U}_{\text{tr}}\mathbf{\Lambda}_{\text{tr}}\mathbf{V}_{\text{tr}}^{T} \tag{9}$$

and penalize the subsequent CVA (i.e., control complexity and avoid singular matrices) by using a reduced number of components $P < M$ to obtain the $P \times M$ matrix $\mathbf{Q}_{\text{tr}}^{*} = [\mathbf{q}_1, \ldots, \mathbf{q}_M]$, defined by

$$\mathbf{Q}_{\text{tr}}^{*} = (\mathbf{U}_{\text{tr}}^{*})^{T}\mathbf{X}_{\text{tr}} = \mathbf{\Lambda}_{\text{tr}}^{*}(\mathbf{V}_{\text{tr}}^{*})^{T}. \tag{10}$$

Form the $P \times P$ within- and between-class covariance matrices

$$\mathbf{W} = \sum_{j,k} (\mathbf{q}_{jk} - \bar{\mathbf{q}}_{\cdot k})(\mathbf{q}_{jk} - \bar{\mathbf{q}}_{\cdot k})^T,$$
$$\mathbf{B} = \sum_{k} N_k (\bar{\mathbf{q}}_{\cdot k} - \bar{\mathbf{q}}_{\cdot\cdot})(\bar{\mathbf{q}}_{\cdot k} - \bar{\mathbf{q}}_{\cdot\cdot})^T, \quad (11)$$

where $\mathbf{q}_{jk}$ is the vector of P component values of scan j in class k with N_k scans/group, where $k = (1, \ldots, K)$ and $K < P$. The CVA solution for the data matrix $\mathbf{Q}^*_{\text{tr}}$ and the class structure indexed by k is defined by the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, providing the $K - 1$, P-dimensional canonical eigenvectors $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_{K-1}]$, normalized such that $\mathbf{L}^T(\mathbf{W}/(M - K))\mathbf{L} = \mathbf{I}$ (e.g., Mardia *et al.*, 1979). This provides PCA-like orthogonal canonical coordinates (c), which successively maximize the SNR defined by the between-class mean variance divided by the pooled within-class variance. The training-set scans' canonical coordinates are given by

$$\mathbf{c}_r = (\mathbf{Q}^*_{\text{tr}})^T \mathbf{l}_r, \quad (12)$$

where $r = (1, \ldots, K - 1)$, with $\mathbf{c}_i^T \mathbf{c}_j = 0$ $(i \neq j)$, and $\mathbf{c}_i^T \mathbf{c}_i = (1 + \lambda_i)$, where λ_i is the eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ associated with eigenvector $\mathbf{l}_i$. The associated canonical eigenimages are given by

$$\mathbf{e}_r = \mathbf{U}^*_{\text{tr}} \mathbf{l}_r. \quad (13)$$

To sequentially test for significant dimensions, $r = 0, \ldots, K - 1$, we may use Bartlett's asymptotic χ^2 approximation with the degrees of freedom given by $f = (P - r)(K - r - 1)$.

$$(M - 1 - \tfrac{1}{2}(P + K)) \sum_{i=r+1}^{K-1} \log(1 + \lambda_i) \xrightarrow{\lim N \to \infty} \chi^2_f. \quad (14)$$

If we assume that the combination of subspace selection with P principal components together with the training-set parameter estimates for the CVA model define a canonical coordinate signal subspace and an independent noise subspace, from Eq. (7) we may write

$$p(g^{(j)} | \mathbf{x}^{(j)}_{\text{te}};\ \theta_{\text{tr}}) = \frac{1}{C} \exp\left\{ -\frac{1}{2} \left\| \mathbf{L}^T_{\text{tr}} (\mathbf{U}^*_{\text{tr}})^T (\mathbf{x}^{(j)}_{\text{te}} - \bar{\mathbf{x}}^{(g^{(\cdot)})}_{\text{tr}}) \right\|^2 \right\} p(g^{(j)}). \quad (15)$$

The posterior probability of a test-set scan, $\mathbf{x}^{(j)}_{\text{te}}$, being assigned to the class representing its true brain state, $g^{(j)}$, is governed by the Euclidean distance between the mean training-set scan for that class and the test-set scan after projection through the reduced PCA basis, $\mathbf{U}^*_{\text{tr}}$, onto the canonical coordinate subspace, $\mathbf{L}$. This is a very versatile expression for we may choose any one of a wide range of possible basis sets in place of $\mathbf{U}^*_{\text{tr}}$, e.g., the tensor-product splines in Kustra and Strother (2001) or independent components in Lange *et al.* (1999).

ACKNOWLEDGMENTS

This work was partly supported by NIH Grant NS33179 and Human Brain Project Grant P20 MN57180 and by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center (CONNECT) and the Technology Center Through Highly Oriented Research (THOR). We thank Jeih-San Liow and Dana Daly for their assistance with data collection and analysis; Nick Lange, Jan Larsen, Niels Mørch, and Finn Nielsen for many helpful and enlightening discussions; and the anonymous reviewers for suggestions that significantly improved the paper.

REFERENCES

Ardekani, B. A., Strother, S. C., Anderson, J. R., Law, I., Paulson, O. B., Kanno, I., and Rottenberg, D. A. 1998. On the detection of activation patterns using principal components analysis. In *Quantitative Functional Brain Imaging with Positron Emission Tomography* (R. E. Carson, M. E. Daube-Witherspoon, and P. Herscovitch, Eds.), pp. 253–257. Academic Press, San Diego.

Arndt, S., Cizadlo, T., Andreasen, N. C., Heckel, D., Gold, S., and Oleary, D. S. 1996. Tests for comparing images based on randomization and permutation methods. *J. Cereb. Blood Flow Metab.* **16:** 1271–1279.

Azari, N. P., Pettigrew, K. D., Schapiro, M. B., Haxby, J. V., Grady, C. L., Pietrini, P., Salerno, J. A., Heston, L. L., Rapoport, S. I., and Horwitz, B. 1993. Early detection of Alzheimer's disease: A statistical approach using positron emission tomographic data. *J. Cereb. Blood Flow Metab.* **13:** 438–447.

Balslev, D., Nielsen, F. Å, Frutiger, S. A., Sidtis, J. J., Christiansen, T. B., Svarer, C., Strother, S. C., Rottenberg, D. A., Hansen, L. K., Paulson, O. B., and Law, I. 2001. Cluster analysis of activity time-series in motor learning. *Hum. Brain Mapp.*, in press.

Barnett, V., and Lewis, T. 1994. *Outliers in Statistical Data*, 3rd ed. Wiley, New York.

Biggerstaff, B. J., and Tweedie, R. 1997. Incorporating variability in estimates of heterogeneity in the random-effects model in meta-analysis. *Stat. Med.* **16:** 753–758.

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

Bullmore, E. T., Horwitz, B., Honey, G., Brammer, M., Williams, S., and Sharma, T. 2000. How good is good enough in path analysis of fMRI data? *NeuroImage* **11:** 289–301.

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

Clark, C. M., Ammann, W., Martin, W. R., Ty, P., and Hayden, M. R. 1991. The FDG/PET methodology for early detection of disease onset: A statistical model. *J. Cereb. Blood Flow Metab.* **11:** A96–102.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans.* Soc. Ind. Appl. Math., Philadelphia.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap.* Academic Press, San Diego.

Efron, B., and Tibshirani, R. J. 1997. Improvements on cross-validation: The .632+ bootstrap method. *J. Am. Stat. Assoc.* **92:** 548–560.

Fletcher, P. C., Dolan, R. J., Shallice, T., Frith, C. D., Frackowiak, R. S. J., and Friston, K. J. 1996. Is multivariate analysis of PET data more revealing than the univariate approach? Evidence from a study of episodic memory retrieval. *NeuroImage* **3:** 209–215.

Friedman, J. H. 1994. An overview of predictive learning and function approximation. In *From Statistics to Neural Networks: Theory*

and Pattern Recognition Applications (V. Cherkassky, J. H. Friedman, and H. Wechsler, Eds.), pp. 1–61. Springer-Verlag, Berlin.

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

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

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

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., and Worsley, K. J. 1999. Multisubject fMRI studies and conjunction analyses. *Hum. Brain Mapp.* **10:** 385–396.

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

Gold, S., Arndt, S., Johnson, D., O'Leary, D. S., and Andreasen, N. C. 1997. Factors that influence effect size in ^{15}O PET studies: A meta-analytic review. *NeuroImage* **5:** 280–291.

Good, P. 1994. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer-Verlag, Berlin.

Goutte, C., Nielsen, F. Å., and Hansen, L. K. 2000. Modeling the hemodynamic response in fMRI with smooth FIR filters. *IEEE Trans. Med. Imaging* **19:** 1188–1201.

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

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

Hansen, L. K., Nielsen, F. A., Strother, S. C., and Lange, N. 2001. Consensus inference in neuroimaging. *NeuroImage* **13:** 1212–1218.

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

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

Kjems, U., Hansen, L. K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., and Strother, S. C. 2002. The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. *NeuroImage* **15:** 772–786.

Kustra, R. 2000. *Statistical Analysis of Medical Images with Applications to Neuroimaging.* Univ. of Toronto, Toronto. [Ph.D. thesis. http://www.utstat.utoronto.ca/~rafal/thesis.ps.gz]

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.

LaConte, S., Strother, S. C., Anderson, J., Muley, S., Frutiger, S., Hansen, L. K., Yacoub, E., Hu, X., and Rottenberg, D. A. 2001. Evaluating pre-processing choices in single-subject BOLD–fMRI studies using data-driven performance metrics. *NeuroImage* **13:** S179.

Lange, N., Strother, S. C., Anderson, J. R., Nielsen, F. A., Holmes, A., Kolenda, T., Savoy, R., and Hansen, L. K. 1999. Plurality and resemblance in fMRI data analysis. *NeuroImage* **10:** 282–303.

Larsen, J., and Hansen, L. K. 1997. Generalization: The hidden agenda of learning. In *The Past, Present, and Future of Neural Networks for Signal Processing* (J.-N. Hwang, S. Y. Kung, M. Niranjan, and J. C. Principe, Eds.), *IEEE Signal Processing Magazine,* pp. 43–45. IEEE Press, New York.

Larsen, J., and Goutte, C. 1999. On optimal data split for generalization estimation and model selection. In *Proceedings IEEE Workshop on Neural Networks for Signal Processing IX* (Y. H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds.), pp. 225–234. IEEE Press, New Jersey.

Liow, J. S., Rehm, K., Strother, S. C., Anderson, J. R., Mørch, N., Hansen, L. K., Schaper, K. A., and Rottenberg, D. A. 2000. Voxel based covariance analysis for optimal discrimination of groups of FDG PET scans between normal and HIV-1 seropositive subjects. *J. Nucl. Med.* **41:** 612–621.

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

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., and Grady, C. L. 1996. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage* **3:** 143–157.

McIntosh, A. R., Rajah, M. N., and Lobaugh, N. J. 1999. Interactions of prefrontal cortex in relation to awareness in sensory learning. *Science* **284:** 1531–1533.

McKeown, M. J. 2000. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *NeuroImage* **11:** 24–35.

Megalooikonomou, V., Ford, J., Shen, L., Makedon, F., and Saykin, A. 2000. Data mining in brain imaging. *Stat. Methods Med. Res.* **9:** 359–394.

Milan, L., and Whittaker, J. 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *J. R. Stat. Soc. Ser. C Appl. Stat.* **44:** 31–49.

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

Moeller, J. R., Nakamura, T., Mentis, M. J., Dhawan, V., Spetsieres, P., Antonini, A., Missimer, J., Leenders, K. L., and Eidelberg, D. 1999. Reproducibility of regional metabolic covariance patterns: Comparison of four populations. *J. Nucl. Med.* **40:** 1264–1269.

Mørch, N. 1998. *A Multivariate Approach to Functional Neuromodeling.* Danish Technical Univ., Lyngby. [Ph.D. thesis. http://eivind.imm. dtu.dk/publications/phdthesis.html]

Mørch, N., Hansen, L. K., Strother, S. C., Law, I., Svarer, C., Lautrup, B., Anderson, J., Lange, N., and Paulson, O. B. 1996. Generalization performance of nonlinear vs. linear models for [^{15}O]water PET functional activation studies. *NeuroImage* **3:** S258.

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.

Ngan, S. C., LaConte, S. M., and Hu, X. P. 2000. Temporal filtering of event-related fMRI data using cross-validation. *NeuroImage* **11:** 797–804.

Nielsen, F. A., Hansen, L. K., and Strother, S. C. 1998. Canonical ridge analysis with ridge parameter optimization. *NeuroImage* **7:** S758.

Pajevic, S., Daubewitherspoon, M. E., Bacharach, S. L., and Carson, R. E. 1998. Noise characteristics of 3-D and 2-D PET images. *IEEE Trans. Med. Imaging* **17:** 9–23.

Papoulis, A. 1991. *Probability, Random Variables and Stochastic Processes*, 3rd ed. McGraw–Hill, New York.

Petersen, S. E., van Mier, H., Fiez, J. A., and Raichle, M. E. 1998. The effect of practice on the functional anatomy of task performance. *Proc. Natl. Acad. Sci. USA* **95:** 853–860.

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. 1999a. Statistical limitations in functional neuroimaging. I. Noninferential methods and statistical models. *Philos. Trans. R. Soc. Ser. B Biol. Sci.* **354:** 1239–1260.

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. 1999b. Statistical limitations in functional neuroimaging. II. Signal detection and statistical inference. *Philos. Trans. R. Soc. Ser. B Biol. Sci.* **354:** 1261–1281.

Politis, D. N. 1998. Computer-intensive methods in statistical analysis. *IEEE Signal Process. Mag.* **15:** 39–55.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge, UK.

Ripley, B. D. 1998. Statistical theories of model fitting. In *Neural Networks and Machine Learning* (C. M. Bishop, Ed.), pp. 3–25, Springer-Verlag, Berlin.

Rottenberg, D. A., Sidtis, J. J., Strother, S. C., Schaper, K. A., Anderson, J. R., Nelson, M. J., and Price, R. W. 1996. Abnormal cerebral glucose metabolism in HIV-1 seropositives with and without dementia. *J. Nucl. Med.* **37:** 1133–1141.

Shaw, M., Strother, S. C., McFarlane, A. C., Morris, P., Anderson, J., Clark, C. R., and Egan, G. F. 2002. Abnormal functional connectivity in post-traumatic stress disorder. *NeuroImage*, in press.

Sidtis, J. J., Strother, S. C., Anderson, J. R., and Rottenberg, D. A. 1999. Are brain functions really additive? *NeuroImage* **9:** 490–496.

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.

Somorjai, R. L., Jarmasz, M., and Baumgartner, R. 2001. A fast, two-stage strategy for the exploratory analysis of functional MRI data by temporal fuzzy clustering. *AI Med.*, in press.

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

Strother, S. C., Anderson, J. R., Xu, X.-L., Liow, J.-S., Bonar, D. C., and Rottenberg, D. A. 1994. Quantitative comparisons of image registration techniques based on high resolution MRI of the brain. *J. Comput. Assisted Tomogr.* **18:** 954–962.

Strother, S. C., Kanno, I., and Rottenberg, D. A. 1995a. Principal component analysis, variance partitioning and "functional connectivity." *J. Cereb. Blood Flow Metab.* **15:** 353–360.

Strother, S. C., Anderson, J. R., Schaper, K. A., Sidtis, J. J., Liow, J.-S., Woods, R. P., and Rottenberg, D. A. 1995b. 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.

Strother, S. C., Lange, N., Anderson, J. R., Schaper, K. A., Rehm, K., Hansen, L. K., and Rottenberg, D. A. 1997. Activation pattern reproducibility: Measuring the effects of group size and data analysis models. *Hum. Brain Mapp.* **5:** 312–316.

Strother, S. C., Rehm, K., Lange, N., Anderson, J. R., Schaper, K. A., Hansen, L. K., and Rottenberg, D. A. 1998. Measuring activation pattern reproducibility using resampling techniques. In *Quantitative Functional Brain Imaging with Positron Emission Tomography* (R. E. Carson, M. E. Daube-Witherspoon, and P. Herscovitch, Eds.), pp. 241–246. Academic Press, San Diego.

Strother, S. C., Anderson, J. R., Frutiger, S., Hansen, L. K., Lange, N., Sidtis, J. J., Daly, D., Arnold, J. B., and Rottenberg, D. A. 1999. The reproducibility of activation patterns in patient and control populations: Measurement of group and subject effects. *J. Cereb. Blood Flow Metab.* **19**(Suppl 1): S829.

Strother, S. C., Anderson, J., Frutiger, S., Muley, S., Rottenberg, D., Kjems, U., and Hansen, L. K. 2000. The quantitative evaluation of functional neuroimaging expertiments: The NPAIRS data analysis framework. *NeuroImage* **11:** S592.

Svarer, C., Strother, S. C., Morch, N., Law, I., Hansen, L. K., and Paulson, O. B. 1997. Evaluating statistical parametric mapping (SPM) analysis results using leave-one-out resampling in a [^{15}O]water PET functional activation study. *NeuroImage* **5:** S374.

Tegeler, C., Strother, S. C., Anderson, J. R., and Kim, S.-G. 1999. Reproducibility of BOLD-based functional MRI obtained at 4T. *Hum. Brain Mapp.* **7:** 267–283.

VanHorn, J. D., Ellmore, T. M., Esposito, G., and Berman, K. F. 1998. Mapping voxel-based statistical power on parametric images. *NeuroImage* **7:** 97–107.

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C. 1988. Automated image registration: I. General methods and intrasubject, intramodality validation. *J. Comput. Assisted Tomogr.* **22:** 139–152.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. 1996. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4:** 58–73.

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:** 305–331.

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., and Evans, A. C. 1999. Detecting changes in non-isotropic images. *Hum. Brain Mapp.* **8:** 98–101.

Xiong, J. H., Gao, J. H., Lancaster, J. L., and Fox, P. T. 1996. Assessment and optimization of functional MRI analyses. *Hum. Brain Mapp.* **4:** 153–167.