Published in final edited form as:
*Magn Reson Imaging*. 2014 November ; 32(9): 1102–1113. doi:10.1016/j.mri.2014.07.011.

# Using Copula Distributions to Support More Accurate Imaging-Based Diagnostic Classifiers for Neuropsychiatric Disorders

**Ravi Bansal, Ph.D.**[1], **Xuejun Hao, Ph.D.**[1], **Jun Liu, Ph.D.**[1], and **Bradley S. Peterson, M.D.**[1]
[1]Department of Psychiatry, Columbia College of Physicians & Surgeons New York, NY 10032

## Abstract

Many investigators have tried to apply machine learning techniques to magnetic resonance images (MRIs) of the brain in order to diagnose neuropsychiatric disorders. Usually the number of brain imaging measures (such as measures of cortical thickness and measures of local surface morphology) derived from the MRIs (i.e., their dimensionality) has been large (e.g. >10) relative to the number of participants who provide the MRI data (<100). Sparse data in a high dimensional space increases the variability of the classification rules that machine learning algorithms generate, thereby limiting the validity, reproducibility, and generalizability of those classifiers. The accuracy and stability of the classifiers can improve significantly if the multivariate distributions of the imaging measures can be estimated accurately. To accurately estimate the multivariate distributions using sparse data, we propose to estimate first the univariate distributions of imaging data and then combine them using a Copula to generate more accurate estimates of their multivariate distributions. We then sample the estimated Copula distributions to generate dense sets of imaging measures and use those measures to train classifiers. We hypothesize that the dense sets of brain imaging measures will generate classifiers that are stable to variations in brain imaging measures, thereby improving the reproducibility, validity, and generalizability of diagnostic classification algorithms in imaging datasets from clinical populations. In our experiments, we used both computer-generated and real-world brain imaging datasets to assess the accuracy of multivariate Copula distributions in estimating the corresponding multivariate distributions of real-world imaging data. Our experiments showed that diagnostic classifiers generated using imaging measures sampled from the Copula were significantly more accurate and more reproducible than were the classifiers generated using either the real-world imaging measures or their multivariate Gaussian distributions. Thus, our findings demonstrate that estimated multivariate Copula distributions can generate dense sets of brain imaging measures that can in turn be used to train classifiers, and those classifiers are significantly more accurate and more reproducible than are those generated using real-world imaging measures alone.



Corresponding Author: Ravi Bansal, Room 2410, Unit #74, 1051 Riverside Drive, New York, NY 10032, (W): 212-543-6145, bansalr@nyspi.columbia.edu.

## 1. Introduction

Several machine learning algorithms have been applied to brain imaging measures in order to generate automated classification rules that can aid the clinical diagnosis of patients who have one or more neuropsychiatric disorders. Supervised machine learning algorithms[1–15] use known clinical diagnoses for participants to learn classification boundaries that best separate in feature space the brain imaging measures of participants who have different diagnoses. Among the numerous available classes of machine learning algorithms, perhaps the most commonly used when developing diagnostic algorithms from brain imaging data are support vector machines (SVMs).[16] An SVM identifies a hyperplane that partitions the feature space of imaging measures[17] in such a way that imaging measures on one side of the hyperplane are assigned to one diagnosis and measures on the other side are assigned to another diagnosis. Support vectors for the imaging measures establish the location and orientation of the hyperplane in feature space that optimally separates those disorders. The diagnostic accuracy of the support vectors and their associated hyperplane therefore is sensitive both to how representative the sampling of participants is of the general population of persons who have those disorders and to the errors in the imaging measures of their brains, especially when the measures are available from only a relatively small sample of participants[18].

Measures in feature space from real-world imaging data are usually sparsely populated because of the considerable effort and expense required to recruit, scan, and process data from a large number of participants. A sparsely populated feature space makes estimation of the distribution of brain imaging measures in that space unreliable and sensitive to measurement errors in the raw imaging data that subsequently introduce errors and variability in the support vectors of the SVM hyperplane, which in turn produces errors in diagnostic classification. Several probabilistic SVM methods, including Gaussian process framework[19], Bayesian framework[20], probability product kernel[21], and fuzzy support vector machines[22], have been proposed to reduce the diagnostic consequences of these errors and variability when estimating the distributions of measures in feature space. These methods, however, are simply ad hoc modifications of the SVM and do not model the generative process that yielded the distributions of brain imaging measures upon which those SVMs operate.

The performance of the SVM-based classifications can be improved significantly using advanced generative models -- i.e., multivariate distributions -- that can be sampled to generate sets of brain imaging measures that are sufficiently dense to permit more accurate and stable estimates of multivariate distributions in feature space, which in turn yield more accurate and stable classification algorithms that are built on them. In addition, generative models can yield probabilistic rather than dichotomous diagnostic classifications, and probabilistic classifications may be more useful clinically than dichotomous ones, particularly when making clinical decisions about the risk for illness onset in high-risk individuals, about treatment response, or about the likely course of future illness. Furthermore, estimation of the multivariate distributions of imaging measures would permit integration of available information into the diagnostic algorithms, such as the symptoms endorsed or the provisional phenotypic diagnosis based on clinical interviews, to generate a

more accurate probabilistic or binary diagnosis. Thus, using precise generative models for brain imaging measures can significantly improve the performance of SVM-based and other classification algorithms for diagnosing patients in real-world settings.

Parametric and nonparametric techniques, including histograms and kernel density estimators,[23] can both be used to estimate the multivariate distributions of brain imaging measures. Parametric methods are attractive for this purpose because they require a relatively small number of participants to estimate a small number of parameters, such as mean and variances. These parametric distributions, particularly when estimated using a small number of participants, may differ considerably from the true, unknown distributions of imaging measures, especially in the tails of the distributions, where most of the probability weightings for multivariate distributions aggregate.[24] Relatively small inaccuracies in the parametric estimates of univariate distributions for imaging measures therefore have highly detrimental consequences for the accuracy of parametric estimates of the multivariate distributions of those same measures. In contrast, nonparametric techniques for estimating univariate distributions, such as histograms, empirical distributions, and kernel estimators, do not assume a particular model for the distribution and therefore can provide estimates that are closer to the true unknown distribution of measures. Nonparametric estimators, however, require data from an exorbitantly large number of participants to estimate the multivariate distribution with a high degree of confidence and accuracy. For example, estimating a multivariate Gaussian distribution using a Gaussian-kernel based nonparametric technique in a 10-dimensional feature space, requires the collection of imaging measures from 842,000 participants to ensure that the mean square error is <0.1 at the mode of the distribution. Because imaging data typically are available from, at most, several hundred participants, standard statistical methods cannot be used to estimate the multivariate distributions of imaging measures needed to generate accurate diagnostic classifiers. The use of nonparametric techniques to estimate the univariate distributions of imaging measures, in contrast, require data from only 4 participants to ensure a mean square error of < 0.1 from the mode of an unknown real distribution.[24]

In contrast, multivariate distributions of brain imaging measures that are sparsely distributed in a high dimensional space cannot be estimated accurately using standard parametric or nonparametric techniques. Nevertheless, the independent, univariate probability distribution of each brain imaging measure be estimated with high accuracy using standard techniques,[25] and those univariate distributions for each measure can be coupled[26–29] to estimate their multivariate distribution. The estimated multivariate distribution, in turn, can be sampled as densely as desired to generate brain imaging measures that are distributed according to the univariate distributions of the real-world measures and that have the same correlation structure as specified for the Copula. Copulas are powerful statistical tools because they permit the easy estimation of multivariate distributions using their corresponding marginal distributions and dependence structure. Estimated multivariate distributions therefore can be used to generate machine learning-based classifiers that can diagnosis patients with greater accuracy and with better reliability than would otherwise be possible using sparsely sampled, real-world measures.

# 2. Methods

## 2.0 Overview

We will first briefly review Copula and multivariate Gaussian distributions in the context of measures from magnetic resonance (MR) images of the brain, including measures of cortical thickness and local morphology of the cerebral surface and subcortical nuclei.[30–33] We refer to these as "real-world" imaging measures when they are taken directly from the MRIs of living participants, as opposed to measures that are computer-generated. We then use standard nonparametric statistical techniques to estimate the univariate distribution of each measure that we then couple to generate multivariate distributions. We use the Kolmogorov – Smirnov (KS) statistic to compare (1) the univariate and multivariate distributions of imaging measures sampled from the Copula with (2) the corresponding distributions of measures sampled from their estimated multivariate Gaussian distributions, as well as with the actual distributions of the real-world measures used to estimate the multivariate distributions. In these comparisons, we first assess whether a multivariate Copula better models the real-world data than does a multivariate Gaussian, even if the real-world data are distributed according to a specified multivariate Gaussian distribution. We then assess the effects that increasing the dimensionality of the feature space has on the accuracy of the estimated Copula and multivariate Gaussian distributions.

We next assess how well these various distributions discriminate imaging measures taken from different diagnostic groups. Because we expect the multivariate Gaussian distributions of imaging measures to have a greater probability mass in their tails than will their corresponding Copula distributions, we expect the multivariate Gaussian distributions for imaging measures from two groups of participants to have greater overlap in their support vectors, and therefore we expect the multivariate Gaussian to discriminate diagnostic groups less well than the Copula distributions discriminate them. We plot the receiver operating characteristic (ROC) curve and calculate the area under the ROC curve to assess how well the various distributions separate imaging measures from differing diagnostic groups in feature space. Finally, we apply SVMs to imaging measures sampled from either the Copula or multivariate Gaussian distributions to develop imaging-based diagnostic classifiers for participants who have differing neuropsychiatric illnesses. We calculate the mean and variance of the misclassification rates and the areas under the ROC curve for the diagnostic classifiers to quantify the accuracy and stability of the classification rules that SVMs generate when applied to imaging measures sampled from either the Copula or multivariate Gaussian distributions.

## 2.1 Copulas

Copulas are multivariate distributions of variables whose marginal distributions are uniformly distributed on the unit [0,1] interval. Copulas have been used widely across diverse fields of investigation.[34–38]. A Copula models the correlation among multiple variables independently of their univariate distributions. They provide a way of constructing a multivariate distribution that has a specified correlation structure and in which each of the component variables has a specified marginal distribution. The correlation structure is specified using a parametric model that is estimated from the sparsely populated raw data.

The estimated correlation structure and the univariate distributions of each of the invididual variables define the Copula. The estimated Copula can be sampled to generate a dense set of multivariate measures that are transformed by the inverse cumulative distribution of each of the measures, thereby generating pseudo-random samples of multivariate measures that have the specified correlation structure and the specified marginal distributions. We use the dense set of Copula-based multivariate estimates for our brain imaging measures to train an SVM classifier that is robust to the errors in the sparsely populated imaging measures and that can be used to generate a probabilistic diagnosis for each participant.

A two-dimensional Copula $C(u, v)$ is a function on the domain [0,1]×[0,1] such that $C(u, 0)=0=C(0, v)$, $C(u,1)= u$, $C(1, v)= v$. Furthermore, the function $C(u, v)$ is 2-increasing, that is, for every $u_1$, $u_2$, $v_1$, $v_2$ in [0,1] such that $u_1 \leq u_2$ and $v_1 \leq v_2$, $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$, and therefore the $C$ -volume of any rectangle $[u_1, v_1]\times[u_2, v_2]$ is nonnegative in the domain [0,1]×[0,1].[39] Samples $(u, v)$ generated randomly from the Copula are transformed by applying the inverse univariate distributions, $x=F^{-1}(u)$ and $y=G^{-1}(v)$, to generate densely populated multivariate imaging measures from the multivariate distribution $H(x, y)$ of the sparsely populated real-world measures.

A multivariate Gaussian Copula $C_{\sum}^{Gauss}(\overrightarrow{u})$ in a $d$-dimensional unit cube $[0,1]^d$ is constructed by applying the probability integral transform to a multivariate Gaussian distribution $\Phi_{\sum}$ with correlation matrix $\Sigma \in \Re^d \times \Re^d$. That is, $C_{\sum}^{Gauss}(U)=\Phi_{\sum}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d))$, where $\Phi^{-1}(u_1)$ is the inverse cumulative Gaussian distribution function. Similarly, a $d$-dimensional $t$ Copula $C_{\nu,\sum}^{t}(\overrightarrow{u})$ with $\nu$ degrees of freedom is given by

$$C_{\nu,\sum}^{t}(U)=\int_{-\infty}^{t_v^{-1}(u_1)}\cdots\int_{-\infty}^{t_v^{-1}(u_d)}\frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^d\,|\sum|}}(1+\frac{X^T\sum X}{\nu})^{-\frac{\nu+d}{2}}dX. \quad (1)$$

The correlation matrix $\Sigma$, which measures the dependence structure among the brain imaging measures, is calculated using maximum likelihood estimation. We fit the Copula and sampled imaging measures from the estimated Copula using the “copulafit” and “copularnd” functions in Matlab® Mathworks software[40].

## 2.2 The Multivariate Gaussian Distribution

In addition to using a Copula to estimate the multivariate distribution of sparsely populated brain imaging measures, we also estimate their multivariate Gaussian distribution $\Phi_S$ using an unbiased estimate of the covariance matrix $S=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^T$, where the matrix $nS$ is Wishart distributed with $n-1$degrees of freedom, and $\overline{X}=\frac{1}{n-1}\sum_{i=1}^{n}X_i$ is the mean feature vector of the imaging measures. To generate $M$ feature vectors of the imaging measures from the estimated $d$-dimensional, multivariate Gaussian distribution, we first

randomly generate *M*, d-dimensional feature vectors with independent measures that are distributed according to the zero mean, standard Gaussian distribution *N* (0,1). The mean of the *M* feature vectors is subtracted from each to ensure a zero mean of the generated feature vectors. The covariance matrix *S*, which is a symmetric and positive-definite, is decomposed into a lower triangular matrix *L* such that $L \cdot L^{T=}S$, where the matrix *L* represents an effective square root of the covariance matrix *S*.[41] Each *d*-dimensional feature vector is multiplied by the lower triangular matrix *L*, and the mean vector $\bar{X}$ is added to generate measures from a multivariate Gaussian distribution with the specified covariance matrix *S* and mean $\bar{X}$.

### 2.3 Support Vector Machines (SVM)

An SVM[42,43] is a supervised machine learning algorithm that constructs an optimal linear classification boundary in feature space to separate the feature vectors in the training set $S=\{x_i, y_i\}_{i=1}^{N}$, $y_i \in \{-1,1\}$, $x_i \in \Re^d$, where $x_i$ are the *n*-dimensional vector of the imaging measures, $y_i$ are the class label, and *N* is the total number of available feature vectors. One method for SVM estimates the optimal classification boundary $f(x) = w^T \cdot x - b$ that maximizes the margin between the features from the two diagnoses by the following solving the following constrained optimization

$$\min_{W,b,\xi_i} \left\{ \frac{1}{2} w^T \bullet w + C \sum_{i=1}^{N} \xi_i \right\} \quad (2)$$

subject to the constraint that $y_i(w^T \bullet x_i - b) \geq 1 - \xi_i$, $\xi_i; \geq 0$, where $\xi_i$ are the slacks variable in a soft margin[44] method that permits overlap of imaging measures in the two classes. The classification boundary is a hyperplane (a plane in feature spaces of dimension greater than two) defined by "support vectors" such maximize the distances between the hyperplane and closest feature vectors. Support vector machines have been extended to classifying a feature vector among more than two diagnostic classes and for computing nonlinear separating planes by first using nonlinear mapping to transfer the features to a higher dimensional space and then estimating an optimal hyperplane in the higher dimensional space.[45,46] The accuracy of an SVM depends upon the optimal mapping to the higher dimensional space, with the consequence that the computational complexity of defining a classifier is high.[47] Furthermore, because the decision boundary learned by an SVM is defined by the support vectors only, the classification boundary is highly sensitive to noise in these vectors. In other words, small variations in support vectors can produce large variations in the decision boundary, especially for sparse datasets[18]. We implement the SVM classifier using Weka[48], a Java-based, freely available platform that provides a collection of machine learning algorithms for data mining tasks. Within this platform, the SVM classifier uses an algorithm based on sequential minimal optimization[49] (SMO) to estimate the optimal hyperplane to separate multivariate measures associated with each diagnostic class.

### 2.4 Experiments

We now describe the computer-generated and real-world brain imaging measures that we use to evaluate the accuracy of the multivariate Gaussian and Copulas in estimating the multivariate distributions and accuracies of SVM-based diagnostic classifiers.

**2.4.1 Real-World Datasets—**We acquired high-resolution, T1-weighted MRI scans from several diagnostic groups: (1) 40 healthy adults (22 males, age 32·42±10·7 years)[50], (2) 36 adults with Tourette's Syndrome (TS, 21 males, age 37·34±10·9 years)[50], (3) 26 adults with Bipolar Disorder (BD, 11 males, age 37·65±10·35 years)[51], (4) 65 individuals (31 children) who had a low familial risk (LR) for Major Depressive Disorder (MDD), and (5) 66 individuals (12 children) who had a high familial risk (HR) for MDD[52,53]. The T1-weighted MR images were acquired on a 1·5 Tesla GE scanner using a sagittal spoiled gradient recall sequence (TR=24msec, TE=5msec, 45° flip, frequency encoding S/I, no wrap, 256×192 matrix, FOV=30 cm, 2 excitations, slice thickness=1.2 mm, 124 contiguous slices). This sequence was selected to provide superior signal-to-noise and contrast-to-noise ratios in high-resolution images having nearly isotropic voxels ($1.171\times1.171\times1.2$ $mm^3$). For the participants who were at HR or LR for MDD, imaging data were acquired on a 1.5T Siemens Sonata scanner using a 3D MPRAGE sequence with the same parameters[52].

Brain images were preprocessed to minimize intensity inhomogeneities.[54] The brain was isolated from non-brain tissues using semiautomated tools together with manual editing.[55–58] The brains were rotated into standard Talairach orientation, where the cortical mantle was defined using a combination of automated tools and manual editing in all three orientations of the brain.[52] Experts in neuroanatomy used detailed protocols to delineate various brain regions manually, including the cerebral hemispheres[59], hippocampus[32], amygdala[32], caudate, putamen, globus pallidus[59], and thalamus[60]. The definitions were reviewed for accuracy by a second expert in neuroanatomy before the definitions were committed to our database. Interrater reliability was >0.9 for all subcortical regions and >0.99 for the cerebral surface and cortical thickness. The thickness of the cortical mantle in each brain was measured by applying a 3-dimensional morphological operator that distance-transformed the surface of the white matter to the surface of the cortex.[61,62] We used previously described methods[31,32,52,63,64] to quantify precisely the local variations in morphological features across the surfaces of all brain regions. Conformal mapping then mapped these local variations in morphological features and cortical thickness onto a unit sphere[65]. Measures on the sphere were subjected to spherical wavelet analysis[66,67] that generated scaling coefficients to capture patterns of morphological variation in surface measures for each brain region at decreasing spatial resolutions. The scaling coefficients that differed significantly ($p < 10^{-7}$) between (1) HR and LR individuals, (2) healthy adults and TS adults, and (3) healthy adults and BD adults, were the brain imaging measures used as our feature vectors[30]. Imaging measures used in each of these 3 group comparisons were (1) cortical thickness measures for the HR and LR individuals, (2) surface morphology of the right hippocampus for the TS adults and healthy adults, and (3) surface morphology of the right hemisphere, left hippocampus, and left and right amygdala for the BD adults and healthy adults.[30] We selected these 3 groups of participants because our preliminary results suggested that the imaging measures had large overlap for the HR and LR participants, some

overlap for the healthy and TS adults, and minimal overlap for healthy and BD adults. Selecting these three datasets with varying amounts of overlap allowed us to assess the utility of using estimated multivariate distributions under a large range of conditions for generating classification rules from machine learning algorithms.

**2.4.2 Simulating Gaussian-Distributed Brain imaging Measures—**We assessed whether the distributions of measures sampled from an estimated Copula distribution were closer to the distributions of the real-world imaging measures than those sampled from an estimated multivariate Gaussian distribution, even if the imaging measures were distributed according to a specified multivariate Gaussian. We therefore simulated two sets of 40 and 36 vectors that were distributed as multivariate Gaussian distributions. The first set of 40 simulated measures was generated by first fitting a multivariate Gaussian distribution to the brain imaging measures of 40 healthy participants, and then sampling the fitted multivariate Gaussian to generate 40 vectors for the imaging measures. Similarly, the second set of 36 simulated measures was generated through the same procedures applied to the brain imaging measures of 36 adults with Tourette's Syndrome (TS). Therefore, the two sets of 40 and 36 simulated imaging measures were distributed according to specified multivariate Gaussian distributions with different means and covariance matrices.

We then estimated 4 multivariate distributions: (1 & 2) A Copula- and a multivariate Gaussian distribution from the set of 40 vectors from healthy participants and (3 & 4) another Copula and multivariate Gaussian distributions from the set of 36 vectors of the 36 TS adults. We sampled 2000 vectors of brain imaging measures from each of these 4 distributions and calculated the KS statistic that assessed the distances of their distributions from the distributions of imaging measures drawn from Gaussian distributions of the two sets of original 40 and 36 vectors (Fig. 1). In addition, we plotted the ROC curves and computed the area under the curves (AUCs) for these comparisons to assess the separation of the two estimated Copula distributions from the corresponding two estimated multivariate Gaussian distributions. Finally, we applied procedures for cross validation to the classification rules derived from simulated imaging measures and from the Gaussian Copula and multivariate Gaussian distributions to assess the effects of the differing sampling schemes on the accuracy of diagnostic classifications generated by our machine learning algorithms.

**2.4.3 Increasing the Dimensionality of Feature Space—**The probability mass in any small hypercube in feature space asymptotically approaches zero for increasing dimensionality, and therefore any two multivariate distributions in a feature space larger than 10 dimensions typically will be perfectly distinct. With increasing dimensionality, however, imaging measures from an exponentially larger number of participants is required to estimate multivariate distributions accurately. To understand better the effect of an increasing dimensionality of feature space on the estimated multivariate distributions, we used feature vectors with either 2, 3, or 4 real-world brain imaging measures that were selected at random in our cohort of 40 healthy adults and 36 adults with TS and whose multivariate distributions we then estimated. The distributions for increasing dimensionality were compared using various statistical tools described below.

### 2.4.4 Statistical Assessment of the Distributions and Classifiers

**Kolmogorov-Smirnov (KS) Statistics:** We used the KS statistic[68–70] and its associated P-value to compare the univariate distributions of imaging measures sampled from the Copula with the simple univariate distributions of real-world imaging measures. The KS statistic is defined as the largest distance between two cumulative distributions, and its p-value is computed under the null hypothesis that the two distributions are equal. We used the KS statistics to assess (1) whether the distributions of measures sampled from Copula differed from those of real-world brain imaging measures, and (2) whether Copula-generated imaging measures were better matched in their distributions to the real-world measures than were measures sampled from the multivariate Gaussian distribution.

**Split-Half Cross-Validation of SVM Classifiers:** We used 10 independent cross validations, starting with 2-fold and extending to 11-fold cross validation, to assess the accuracy of the SVMs generated from the sampled brain imaging measures, thereby allowing us to estimate the average and variance in misclassification rates for the SVM-based classifiers generated from the multivariate datasets (Fig. 2). Each of these procedures for n-fold cross validation first divides the set of brain imaging measures into n subsets with equal numbers of brain imaging measures in each, then trains a classifier for the imaging measures in n-1 subsets, and finally tests the learned classifier using the imaging measures in the remaining subset that was not used in training the classifier. The procedure is then repeated, using each one of the n subsets as the test data, to compute the average misclassification rates and the AUC of the classifier. For example, a 2-fold cross validation procedure is identical to 2 repeats of a split-half cross validation, wherein one-half of the data are used to train the classifier and the other half are used to test the performance of the classifier independently on data that were not involved in training of the classifier. The training and testing are then repeated by switching the test and training subsets. Across the various n-fold cross validation procedures (procedures from 2-fold to 11-fold cross validations), the average misclassification rates and their AUCs, and their associated standard deviations, provide a measure of the accuracy and reproducibility of the SVMs. Because no single cross validation procedure may provide an accurate estimate for the misclassification rates and their variability, using 10 independent cross validations allows us to assess the variations in misclassification rates with different instantiations of the SVM classifiers. These measures of accuracy and reproducibility can be compared across classifiers that are generated using either the real-world measures or the measures sampled from the Copula or multivariate Gaussian distributions.

**Receiver Operating Characteristics (ROC):** We used ROC curves and a Fisher linear discriminant[71] function to assess quantitatively how well the various ways of estimating the multivariate distributions were able to discriminate brain imaging measures from the different diagnostic groups. We first calculated the two means and the pooled covariance matrix for brain imaging measures from each of the two diagnostic populations. We then used the Fisher linear discriminant to calculate the direction vector that best separated imaging measures of the two groups. We calculated this direction vector by multiplying the inverse of the pooled covariance matrix with the vector from the mean of one group to the mean of the other. We projected each brain imaging measure on the direction vector and

then varied a discriminant threshold along the vector from a very small value to a very large one. For each value of the discriminant threshold, the projected measures that were smaller than the threshold were labeled with the clinical diagnosis of one group and those greater than the threshold were labeled with the clinical diagnosis of the other group. Because the true diagnosis was known for each measure, we could calculate the sensitivity and the specificity associated with each discriminant threshold. We then generated the ROC curve for this discrimination by plotting the sensitivity and specificity values for each discriminant threshold.

We then calculated the area under each ROC curve. Because each curve was independent of a machine learning algorithm, the curve and the area under the curve (AUC) provided a quantitative measure of how well separated in feature space were the multivariate distributions of imaging measures from the two patient populations: low AUC values would indicate large overlap of the two multivariate distributions, whereas values close to 1.0 would indicate excellent separation of the multivariate distributions in feature space. We used the AUC to assess the degree of overlap between (1) differing Copula distributions, (2) differing multivariate Gaussian distributions, and (3) empirical distributions of real-world imaging measures of healthy and ill participants. In addition, we calculated the standard error[72] for the AUC and used the computed standard error and AUC to assess statistical significance of the difference in AUCs computed for the various ROC curves.

**Stability of the Classifiers Learned Using the Sampled Brain Imaging Measures:** We assessed the performance of classifiers generated using the estimated multivariate distributions. First we used the densely sampled measures to generate our classifiers. We then used these classifiers to assign the real-world brains to one of two diagnoses. This allowed us to calculate the misclassification rates when assigning real-world brains to the diagnoses for classifiers generated from densely sampled brain imaging measures. Next, we independently sampled 10 sets of brain imaging measures and for each generated a classifier, yielding a total of 10 classifiers. We assigned every real-world brain to a neuropsychiatric diagnosis using each of the 10 classifiers. Because the use of a large number of measures sampled from the estimated distributions reduces variability in the learned classifier, we expected that the real-world brains would be similarly classified (either incorrectly or correctly) by each of the 10 classifiers.

## 3. Results

### 3.1 Simulating Gaussian-Distributed Brain Imaging Measures

The distributions of brain imaging measures sampled from both the Gaussian Copula and the multivariate Gaussian distributions were, upon visual inspection, close to the distributions of the simulated brain imaging measures (Fig. 3). The areas under the ROC curves were 0.982 for the simulated measures, 0.992 for the measures sampled from the Gaussian Copulas, and 0.994 for the measures sampled from the estimated multivariate Gaussian distributions. The KS distances between the distributions of the simulated measures and those sampled from the Gaussian Copula were 0.085 (P-value=0.94) in healthy adults and 0.084 (P-value=0.96) in TS adults. The KS distances between the distributions of the simulated measures and those sampled from the estimated multivariate Gaussian distributions were 0.101 (P-

value=0.818) in healthy adults and 0.111 (P-value=0.774) in TS adults (Fig. 4). The misclassification rates for an SVM to diagnose individual brains were (1) for simulated measures: 0.028 in healthy adults and 0.0075±0.012 in TS adults, (2) for measures sampled from the estimated multivariate Gaussian: 0.0069 in healthy adults and 0.0082 in TS adults, and (3) for measures sampled from the estimated Gaussian Copula: 0.0076 in healthy adults and 0.0045 in TS adults. The KS statistics, misclassification rates, and AUCs in our cross validation analyses suggested that the distributions of imaging measures sampled from the estimated Gaussian Copula estimated the distributions of simulated measures for the specified multivariate distribution better than did the distributions of measures sampled from the estimated multivariate Gaussian. Thus, the Gaussian Copula better estimated the distribution of the simulated measures than did another multivariate Gaussian distribution, even when the simulated measures were Gaussian distributed.

### 3.2 Effects of Increasing the Dimensionality of Feature Space

The areas under the ROC curve for brain imaging measures sampled from Gaussian Copula were 0.912±0.001 for feature vectors with two measures, 0.95±0.0005 with three measures, and 0.96±0.0004 with four measures, whereas the AUC for imaging measures sampled from the multivariate Gaussian distribution were 0.906±0.0007 for feature vectors with two measures, 0.93±0.001 with three measures, and 0.95±0.03 with four measures (Table 1). The increasing AUCs for an increasing number of brain imaging measures in the feature vector indicated that the overlap between two multivariate distributions decreased with an increasing dimensionality of feature space. In addition, the larger AUCs for imaging measures sampled from the Gaussian Copula compared with AUCs for imaging measures sampled from the multivariate Gaussian suggested that the multivariate distributions estimated using the Copula were better separated in feature space than were the distributions estimated using the multivariate Gaussian distribution. The better separation of the distributions estimated using the Copula was also demonstrated by the smaller misclassification rates for NC and TS adults, which were 0.1±0.0008 and 0.08±0.001 for feature vectors with two measures, 0.05±0.0007 and 0.044±0.0008 for three measures, and 0.03±0.0005 and 0.03±0.0007 for four measures using the Copula, compared with 0.1±0.0008 and 0.08±0.001 for feature vectors with two measures, 0.07±0.001 and 0.07±0.001 for three measures, and 0.035±0.0006 and 0.046±0.001 for four measures sampled from the estimated multivariate Gaussian distribution (Table 1). KS statistics and their associated P-values showed that the distributions of imaging measures sampled from both the Gaussian Copula and multivariate Gaussian distributions were close to the distribution of the real-world brain imaging measures. The overlap of the distributions of imaging measures decreased with the increasing dimensionality of feature space, thereby increasing the separation between the distributions.

### 3.3 Diagnostic Discrimination in Real-World Imaging Datasets

**Participants at LR or HR for familial MDD—**The AUCs for these brain imaging measures sampled from the Copula were larger than those for the real-world measures and the measures sampled from the multivariate Gaussian distributions, indicating a smaller overlap in the supports of imaging measures sampled from the Copula distributions (Table 2). The AUC for the real-world measures, multivariate Gaussian, and Gaussian Copula were

0.73±0.014, 0.738±0.0007, and 0.82±0.001, respectively, suggesting significantly greater probability mass in the tails of the estimated multivariate Gaussian distribution compared to the Gaussian Copula ($p < 0.0001$). The KS statistics for brain measure sampled from Gaussian Copula were 0.072 (LR) and 0.066 (HR), and those sampled from multivariate Gaussian distribution were 0.079 (LR) and 0.08 (HR), showing that the distribution of imaging measures sampled from the Gaussian Copulas were closer to the distribution of the real-world imaging measures than were distributions of measures sampled from multivariate Gaussian distributions.

The mean and standard deviations of the misclassification rates were 0.23±0.02 (LR) and 0.3±0.025 (HR) using real-world measures, 0.18±0.002 (LR) and 0.29±0.002 (HR) for the multivariate Gaussian distribution, and 0.13±0.0025 (LR) and 0.2±0.001 (HR) for the Gaussian Copula (Table 2). The variance in misclassification rates for the Gaussian Copula were smaller than were those for the real-world imaging measures, suggesting that measures sampled from the estimated Gaussian Copula generate stable classifiers in the presence of variability in those measures and were reproducible when applied to different samples of participants in split-half method for cross validation. Finally, when diagnosing the 66 HR and 65 LR participants, the same HR and LR individuals were misclassified by each SVM trained using the measures sampled from Gaussian Copula, suggesting that the SVMs are reproducible when applied to differing samples of participants.

**Healthy Adults and Adults with Tourette's Syndrome (TS)—**The SVMs generated from Gaussian Copula had the smallest misclassification rates (0.024±0.0004, HA, and 0.023±0.0007, TS) and the largest AUC (0.97±0.0005), whereas the SVMs generated from the multivariate Gaussian distribution had the largest misclassification rates (0.08±0.001, HA, and 0.084±0.001, TS) and the smallest AUC (0.92±0.0004) (Table 2). The means and standard deviations for misclassification rates calculated using real-world imaging measures were 0.05±0.015 (HA) and 0.06±0.024 (TS adults) (Table 2). The misclassification rates for the SVMs and AUCs did not change when a larger number of imaging measures (10,000 each for the healthy and TS adults) were sampled from the fitted distributions to generate classifiers. In fact, systematically increasing the number of sampled measures from 50 to 10,000 decreased only the variance, and not the average, of misclassification rates for the machine learning algorithms. The misclassification rates and the AUCs of the classifiers trained using 2,000 imaging measures were similar to those trained using 10,000 imaging measures because the measures were sampled from the same distributions, thereby indicating that the performance of the classifiers depended on the distributions rather than the imaging measures sampled from those distributions. The variance in the misclassification rates and the AUCs, however, decreased with increasing number of samples. Therefore, both the accuracy and the stability of the SVMs increased significantly when imaging measures sampled from Gaussian Copula were used to train the classifiers. Misclassification rates were the highest and AUCs were the smallest for the multivariate Gaussian distribution, because the plots of the distribution of sampled measures showed that its marginal distributions had the greatest probability masses in their tails as compared to the marginal distributions of the real-world measures and to the measures sampled from the Copula distributions.

For the cohort comprising healthy adults and the adults with Bipolar Disorder (BD), the average AUC was 0.98±0.01 for the real-world measures and a perfect 1.0 for both the multivariate Gaussian and Copula distributions (Table 2). The KS statistic, however, showed that the distributions of measures sampled from the Gaussian Copula were more similar to the distributions of real-world measures than those sampled from the multivariate Gaussian distribution (Table 2). The misclassification rates were 0.02±0.02 (HA) and 0.025 (BD adults) for real-world measures, 0.005 (both HA and BD adults) for the multivariate Gaussian distribution, and 0 (both NC adults and BD adults) for the Copula distribution (Table 2). The variance in misclassification rates for classifiers trained using the real-world measures were much larger than the variance in the misclassification rates for classifiers trained using imaging measures sampled from multivariate distributions (Table 2). Therefore, imaging measures sampled from the Gaussian Copula improved the reproducibility and performance of the learned machine classifier. Finally, the distributions of imaging measures for the Gaussian Copula had more pronounced plateaus and less probability masses in their tails than did distributions of measures sampled from the multivariate Gaussian (Fig. 5). [Figure 5]

## 4. Discussion

We demonstrated that multivariate distributions can be estimated accurately from the available, sparsely distributed, real-world brain imaging measures. The estimated distributions can be sampled to generate dense sets of imaging measures that, in turn, generate classifiers that are more accurate and more robust to the errors in present in measures from real-world datasets and to variations in the sampling of participants. Using simulated and real-world datasets, we showed that the variability in misclassification rates were small for Copula distributions, thereby demonstrating that the classifier learned using Copula distributions will be robust to variations in the individual participants of the sample. Therefore, the classifier will have a greater validity when applied to other individuals in the general population. Furthermore, because the support of the estimated Copula distributions is bounded, the estimated distributions define better the subspace of measures for each neuropsychiatric disorder than do the multivariate Gaussian distributions.

The estimated multivariate Copula can also permit the generation of SVM classifiers that diagnose individuals as having one of several possible diagnoses. The optimal hyperplane estimated using SVMs partitions the entire feature space into two noncompact subspaces. The imaging measures in one subspace are labeled with one diagnosis and measures in the other subspace are labeled with the other diagnosis. Labeling all measures in a non-compact subspace with one diagnosis will produce errors in classification, especially when the imaging measures are far from the real-world measures that were used to learn the hyperplane. Furthermore, partitioning of the feature space using hyperplanes to diagnose three or more disorders requires using an iterative strategy to train a machine learning algorithm, with the complexity of the classifier increasing with the number of disorders that require classification.[73–75] The estimated multivariate Copula distributions, in contrast, are closed and bounded and naturally partition the feature space into compact regions of imaging measures that are assigned to a single diagnosis. Therefore, Copula distributions

provide a natural partitioning of the feature space to diagnose a patient as having one among two or more neuropsychiatric disorders.

Our experiments demonstrated that the distributions of imaging measures sampled from the estimated Gaussian Copula distribution were closer to the distributions of the real-world measures than were the marginal distributions of the estimated multivariate Gaussian distributions. In particular, the tails of the distributions from the Copula had less probability mass and more closely matched those of the real-world measures than did those from the multivariate Gaussian distribution. In addition, the distributions of measures from the Copula were plateau-shaped and visually matched better the shape of the distribution of the real-world measures. In addition, the AUC analyses showed that the distributions of measures sampled from the Gaussian Copula were better separated than those sampled from the multivariate Gaussian distributions. Thus, because the Gaussian Copula match the distributions of real-world data and generate classifiers that have higher accuracy and are robust to variations in real-world measures than the multivariate Gaussian distributions, we suggest using the imaging measures sampled from the estimated Copula distribution in machine learning algorithms to generate rules for diagnostic classification.

Although multivariate distributions can be estimated using either a parametric Gaussian distribution or a semiparametric Copula distribution, the estimated distributions are only as representative of the population distributions as the distributions of the real-world measures from which they derive. If the real-world measures are taken from individuals who are not representative of that class of person in the general population, then the estimated multivariate distributions will be biased and unrepresentative of the population as well. Any classifier trained on data from non-representative participants will be likely to misclassify a new participant from the general population whose data did not contribute to the generation of the classification rule, posing enormous problems for the practical implementation and dissemination of the classification algorithm in clinical settings. Conversely, if participants in a study are representative of that class of person in the general population, then their data can be used to estimate multivariate distributions. Those distributions subsequently can be sampled to generate a dense set of brain imaging measures that can be used in training machine learning-based algorithms to generate classification rules that are robust to individual variability in real-world imaging measures. Thus, although measures from a larger number of participants would be more representative of those in the general population and would generate more accurate real-world distributions of those measures, sampling the multivariate distributions estimated from sparse imaging data can be used to generate compact subspaces of the overall feature space. Those compact subspaces can be used in turn to generate diagnostic classifiers that are robust to small individual variations in the imaging measures.

A Copula distribution can model both unimodal and bimodal distributions equally well (Fig. 6). In contrast, a single mode of the multivariate Gaussian distribution limits its application to the estimation of distributions of data that are unimodal. A unimodal model for the distribution of imaging measures assumes that neuropsychiatric disorders are homogeneous in those measures and, by extension, are homogeneous as well in the disease processes that generated those measures. Nearly all neuropsychiatric disorders, however, are

heterogeneous in their etiologies, and therefore a multimodal distribution should in principle better model the distribution of imaging measures from a population of people who carry that diagnosis. Accurate estimation of a multimodal distribution requires measures from a larger number of participants than those who are generally available in a neuroimaging study.

Our experiments demonstrated that the distributions of imaging measures sampled from Gaussian Copula distributions are closer to the distributions of the real-world measures than are multivariate Gaussian distributions, and the data sampled from those estimated Copula distributions generated machine learning-based classifications that were more accurate and stable than using either the real-world measures or multivariate Gaussian distributions alone. In addition, Gaussian Copula-based distributions are visually closer to the corresponding real-world distributions than are multivariate Gaussian-based distributions, and because the Copula distribution can model multimodal distributions appropriately, Copula distributions better model the multivariate distributions of imaging measures from etiologically heterogeneous disorders, although this latter contention requires further empirical support in future studies. Thus, although data from a large sample of participants are invaluable to improve the validity, reproducibility, and generalizability of the classifiers, estimated Copula distributions can overcome some of the limitations of sparse datasets and generate diagnostic classifiers that are significantly more accurate and stable to variations in the sampling of individuals who provide the imaging measures.

## Acknowledgments

This work was supported by National Institute of Mental Health grants MH036197, MH068318, and K02-74677, the Suzanne Crosby Murphy endowment at Columbia University, and the Tom Klingenstein and Nancy Perlman Family Fund.

## References

1. Lao Z, Shen D, Xue Z, Karacali B, Resnick S, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. NeuroImage. 2004; 21:46–57. [PubMed: 14741641]
2. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RSJ. Automatic Classification of MR Scans in Alzheimer's Disease. Brain. 2008; 131(3):681–689. [PubMed: 18202106]
3. Duchesnay E, Cachia A, Roche A, Rivière D, Cointepas Y, Papadopoulos-Orfanos D, Zilbovicius M, Martinot J-L, Régis J, Mangin JF. Classification Based on Cortical Folding Patterns. IEEE Trans on Medical Imaging. 2007; 26(4):553–565.
4. Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. Detection of prodromal Alzheimer's disease via pattern classification of MRI. Neurobiol Aging. 2008; 29(4):514–523. [PubMed: 17174012]
5. Liu, Y.; Teverovskiy, L.; Carmichael, O.; Kikinis, R.; Shenton, M.; Carter, CS.; Stenger, VA.; Davis, S.; Aizenstein, H.; Becker, J., et al. Discriminative MR Image Feature Analysis for Automatic Schizophrenia and Alzheimer's Disease Classification. Barillot, C.; Haynor, DR.; Hellier, P., editors. Springer-Verlag GmbH; Saint-Malo, France: 2004. p. 393-401.
6. Teipel SJ, Born C, Ewers M, Bokde AL, Reiser MF, Moller HJ. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. NeuroImage. 2007; 38(1):13–24. [PubMed: 17827035]
7. Barnes J, Scahill RI, Boyes RG, Frost C, Lewis EB, Rossor CL. Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates. NeuroImage. 2004; 23:574–581. [PubMed: 15488407]

8. Fan Y, Shen D, Davatzikos C. Classification of structural images via high-dimensional image warping, robust feature extraction, and {SVM}. Med Image Comput Comput Assist Interv Int Conf. 2005; 8:1–8.
9. Kawasaki Y, Suzuki M, Kherif F, Takahashi T, Zhou SY, Nakamura K. Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. NeuroImage. 2007; 34:235–242. [PubMed: 17045492]
10. Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage. 2005; 28:980–995. [PubMed: 16275139]
11. Wahlund LO, Almkvist O, Blennow K, Engedahl K, Johansson AGW. Evidence-based evaluation of magnetic resonance imaging as a diagnostic tool in dementia workup. Top Magn Reson Imagin. 2005; 16:427–437.
12. eFigueiredo RJ, Shankle WR, Maccato A, Dick MB, Mundkur P, Mena I. Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. Proc Natl Acad Sci. 1995; 92:5530–5534. [PubMed: 7777543]
13. Herholz K, Salmon E, Perani D, Baron JC, Holthoff V, Frolich L. Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET. NeuroImage. 2002; 17:302–316. [PubMed: 12482085]
14. Lerch JP, Pruessner J, Zijdenbos AP, Collins DL, Teipel SJ, Hampel H, Evans A. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. Neurobiol Aging. 2006; 29(1):23–30. [PubMed: 17097767]
15. Jack CR, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, Knopman DS. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology. 2004; 62:591–600. [PubMed: 14981176]
16. Vapnik, VN. The nature of statistical learning theory. New York: Springer; 1995. p. xvp. 188
17. Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging. 2011; 32(12): 2322, e19–27. [PubMed: 20594615]
18. Li H-X, Yang J-L, Zhang G, Fan B. Probabilistic support vector machines for classification of noise affcted data. Information Sciences. 2012; 221:60–71.
19. Gao JB, Gunn SR, Harris CJ, Brown M. A probabilistic framework for SVM regression and error bar estimation. Machine Learning. 2002; 46(1–3):71–89.
20. Sollich P. Bayesian methods for support vector machines: Evidence and predictive class probabilities. Machine Learning. 2002; 46(1–3):21–52.
21. Jebara T, Kondor R, Howard A. Probability product kernels. Journal of Machine Learning Research. 2004; 5:819–844.
22. Wang YQ, Wang SY, Lai KK. A new fuzzy support vector machine to evaluate credit risk. Ieee Transactions on Fuzzy Systems. 2005; 13(6):820–831.
23. Scott, DW. Multivariate density estimation : theory, practice, and visualization. New York: Wiley; 1992. p. xiip. 317
24. Silverman, BW. Density estimation for statistics and data analysis. Boca Raton: Chapman & Hall/CRC; 1998. p. ixp. 175
25. Silverman, BW. Density estimation for statistics and data analysis. London ; New York: Chapman and Hall; 1986. p. 175
26. Nelsen, RB. An Introduction to Copulas. New york: Springer; 1999.
27. Cherubini, U.; Luciano, E.; Vecchiato, W. Copula methods in finance. Hoboken, NJ: John Wiley & Sons; 2004. p. xvip. 293
28. Joe, H. Multivariate models and dependence concepts. Boca Raton, FL: Chapman & Hall/CRC; 2001. p. xviiip. 399
29. Balakrishnan, N.; Lai, CD. Continuous bivariate distributions. Dordrecht ; New York: Springer; 2009. p. xxxvip. 684

30. Bansal R, Staib LH, Laine AF, Hao X, Xu D, Liu J, Weissman M, Peterson BS. Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. PLoS One. 2012; 7(12):e50698. [PubMed: 23236384]
31. Peterson BS. Form Determines Function: New Methods for Identifying the Neuroanatomical Loci of Circuit-Based Disturbances in Childhood Disorders. Journal of the American Academy of Child and Adolescent Psychiatry. 2010; 49(6):533–535. [PubMed: 20494263]
32. Peterson BS, Choi HA, Hao X, Amat J, Zhu H, Whiteman R, Liu J, Xu D, Bansal R. Morphology of the Amygdala and Hippocampus in Children and Adults with Tourette Syndrome. Archives General Psychiatry. 2007
33. Peterson BS, Warner V, Bansal R, Zhu H, Hao X, Liu J, Durkin K, Adams PB, Wickramaratne P, Weissman MM. Cortical thinning in persons at increased familial risk for major depression. Proc Natl Acad Sci U S A. 2009; 106(15):6273–8. [PubMed: 19329490]
34. Aas K, Czado C, Frigessi A, Bakken H. Pair-copula constructions of multiple dependence. Insurance Mathematics & Economics. 2009; 44(2):182–198.
35. Low RKY, Alcock J, Faff R, Brailsford T. Canonical vine copulas in the context of modern portfolio management: Are they worth it? Journal of Banking & Finance. 2013; 37(8):3085–3099.
36. Meneguzzo D, Vecchiato W. Copula sensitivity in collateralized debt obligations and basket default swaps. Journal of Futures Markets. 2004; 24(1):37–70.
37. Onken A, Grunewalder S, Munk MHJ, Obermayer K. Analyzing Short-Term Noise Dependencies of Spike-Counts in Macaque Prefrontal Cortex Using Copulas and the Flashlight Transformation. Plos Computational Biology. 2009; 5(11)
38. Strelen, JC. Tools for dependent simulation input with copulas. Brussels, Belgium: 2009.
39. Sklar A. Fonctions de répartition à n dimensions et leurs marges. Publ Inst Statist Univ Paris. 1959; 8:229–231.
40. MATLAB and Statistics Toolbxo Release. The MathWorks, Inc; Natick, Massachusetts: 2012b.
41. Press, WH. Numerical recipes in C++ : the art of scientific computing. Cambridge, UK ; New York: Cambridge University Press; 2002. p. xxviip. 1002
42. Cortes, C.; Vapnik, V. Machine Learning. 1995. Support-Vector Networks; p. 20
43. Vapnik, VN. The Nature of Statistical Learning Theory. Springer; 1999.
44. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995; 20(3):273–297.
45. Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control. 1964; 25:821–837.
46. Boser BE. Pattern-Recognition with Optimal Margin Classifiers. Fundamentals in Handwriting Recognition. 1994; 124:147–171.
47. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998; 2(2):121–167.
48. Frank, EMAH.; Holmes, G.; Kirkby, R.; Pfahringer, B.; Witten, IH. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Springer Verlag; 2005.
49. Platt, JC. Advances in Kernel Methods. Cambridge, MA: MIT Press; 1999. Fast training of support vector machines using sequential minimal optimization; p. 185-208.
50. Peterson BS, Staib L, Scahill L, Zhang H, Anderson C, Leckman JF, Gore JC, Albert J, Webster R. Regional brain and ventricular volumes in Tourette syndrome. Arch Gen Psychiatry. 2001; 58:427–440. [PubMed: 11343521]
51. Blumberg HP, Kaufman J, Martin A, Whiteman R, Zhang JH, Gore JC, Charney DS, Krystal JH, Peterson BS. Amygdala and hippocampal volumes in adolescents and adults with bipolar disorder. Arch Gen Psychiatry. 2003; 60(12):1201–8. [PubMed: 14662552]
52. Peterson BS, Warner V, Bansal R, Zhu H, Hao X, Liu J, Durkin K, Adams PB, Wickramaratne P, Weissman MM. Cortical thinning in persons at increased familial risk for major depression. Proc Natl Acad Sci USA. 2009; 106:6273–6278. [PubMed: 19329490]
53. Weissman MM, Wickramaratne P, Nomura Y, Warner V, Verdeli H, Pilowsky DJ, Grillon C, Bruder G. Families at High and Low Risk for Depression. Arch Gen Psychiatry. 2005; 62:29–36. [PubMed: 15630070]

54. Sled GJ, Zijdenbos AP, Evans AC. A Nonparametric Method for Automatic Correction of Intensity Nonuniformity in MRI Data. IEEE Trans of Medical Imaging. 1998; 17(1):87–97.
55. Kates WR, Abrams MT, Kaufman WE, Breiter SN, Reiss AL. Reliability and validity of MRI measurement of the amygdala and hippocampus in children with fragile X syndrome. Psychiatry Res: Neuroimaging. 1997; 75:31–48.
56. Watson C, Andermann F, Gloor P, Jones-Gotman M, Peters T, Evans A, Olivier A, Melanson D, Leroux G. Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. Neurology. 1992; 42(9):1743–1750. [PubMed: 1513464]
57. Peterson BS, Riddle MA, Cohen DJ, Katz LD, Smith JC, Hardin MT, Leckman JF. Reduced basal ganglia volumes in tourette's syndrome using three-dimensional reconstruction techniques from magnetic resonance images. Neurology. 1993; 43(5):941–949. [PubMed: 8492950]
58. Shattuck DW, Leahy RM. BrainSuite: An Automated Cortical Surface Identification Tool. Medical Image Analysis. 2002; 8(2):129–142. [PubMed: 12045000]
59. Peterson BS, Thomas P, Kane MJ, et al. A Basal ganglia volumes in patients with Gilles de la Tourette syndrome. Arch Gen Psychiatry. 2003; 60:415–424. [PubMed: 12695320]
60. Ivanov I, Bansal R, Hao X, Zhu H, Kellendonk C, Miller L, Sanchez-Pena J, Miller AM, Chakravarty MM, Klahr K, et al. Morphological Abnormalities of the Thalamus in Youths With Attention Deficit Hyperactivity Disorder. Am J Psychiatry. 2010; 167:397–408. [PubMed: 20123910]
61. Haralick, RLS. Computer and Robot Vision. Vol. 1. Addison-Wesley Publishing Company; 1992.
62. Rosenfeld, A.; Kak, AC. Digital Picture Processing. Academic Press, Inc; 1982.
63. Bansal R, Staib LH, Wang Y, Peterson BS. ROC-based assessments of 3D cortical surface-matching algorithms. Neuroimage. 2005; 24:150–162. [PubMed: 15588606]
64. Plessen KJ, Bansal R, Zhu H, Whiteman R, Quackenbush GA, Hugdahl K, Peterson BS. Hippocampus and amygdala morphology in Attention-Deficit/Hyperactivity Disorder. Arch Gen Psychiatry. 2006; 63:795–807. [PubMed: 16818869]
65. Lorensen W, Cline H. Marching Cubes: a High-Resolution 3D surface construction algorithm. Computer Graphics. 1987; 21:163–169.
66. Schroder, P.; Sweldens, W. Spherical Wavelets: Efficiently Representing Function on the Sphere. Proceedings of the 22nd annual conference on Computer graphics and interactive techniques; 1995. p. 161-172.
67. Sweldens, W. The lifting scheme: A custom design construction of biorthogonal wavelets. Department of Mathematics, University of South Carolina; 1994.
68. Stephens MA. Test of fit for the logistic distribution based on the empirical distribution function. Biometrika. 1979; 66(3):591–595.
69. Kolmogorov AN. Sulla determinazione empirica di una legge di distribuzione. G Inst Ital Attuari. 1933; 4:83.
70. Smirnov NV. Tables for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics. 1948; 19:279.
71. Duda, RO.; Hart, PE.; Stork, DG. Pattern classification. New York: Wiley; 2001. p. xxp. 654
72. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
73. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research. 2002; 2(2):265–292.
74. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research. 2005; 6:1453–1484.
75. Joachims T, Finley T, Yu CNJ. Cutting-plane training of structural SVMs. Machine Learning. 2009; 77(1):27–59.

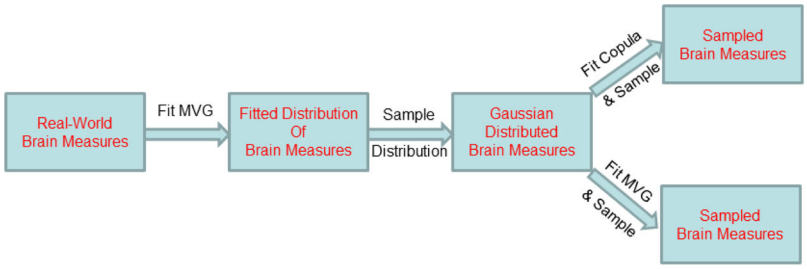


**Figure 1. Generating Simulated from Real-World Brain Imaging Measures**

Simulated measures were generated from real-world imaging measures by first fitting a multivariate Gaussian (MVG) distribution and then sampling that distribution to generate imaging measures that were distributed with a known MVG distribution. We then fitted either a Copula or another MVG distribution to the sampled imaging measures. The fitted distributions were sampled subsequently to generate 2000 samples from either one of the two distributions.

**MVG**=Multivariate Gaussian Distribution.

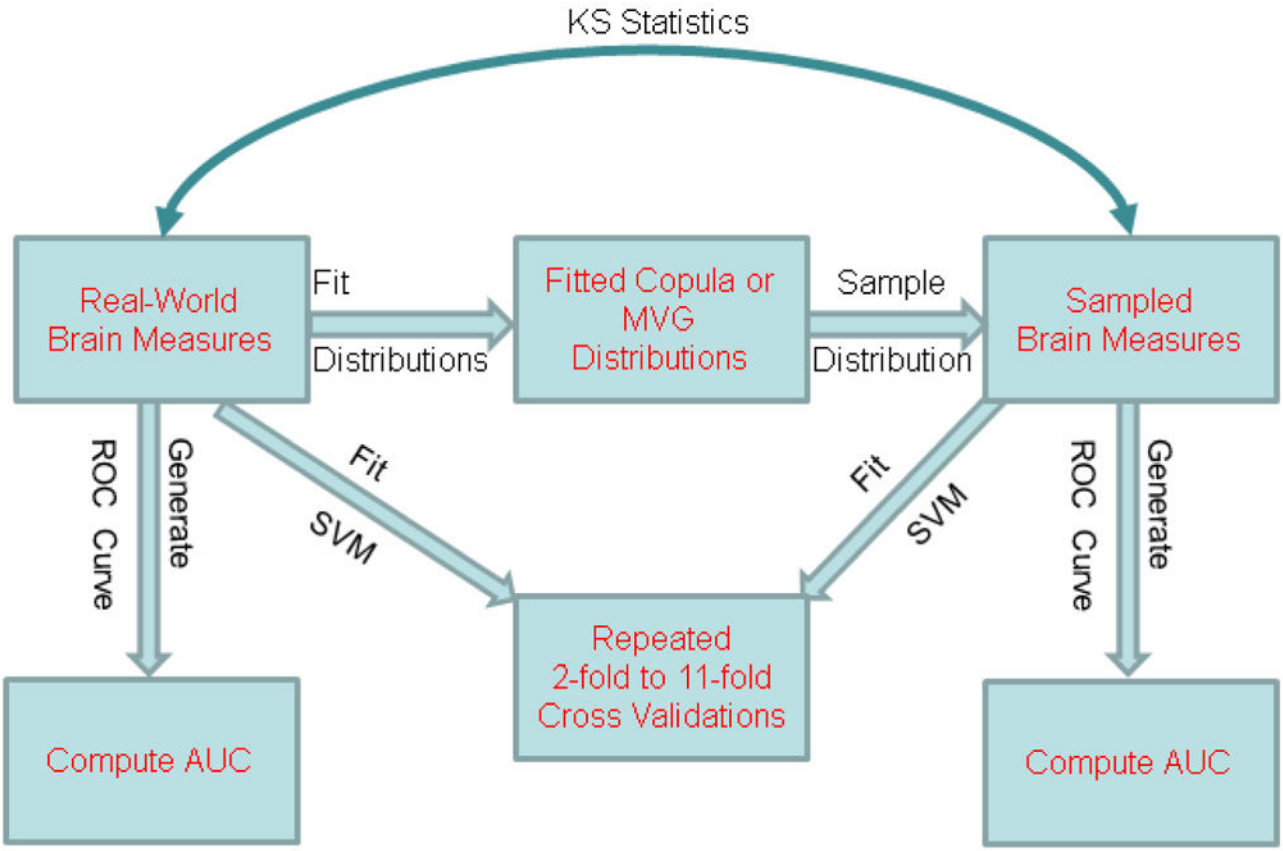


**Figure 2. Statistical Validations of the Sampled Measures and Machine Learning Algorithms**

We fitted either a Copula or a multivariate Gaussian (MVG) distribution to the real-world imaging measures and then sampled the fitted distributions to generate dense sets of brain imaging measures. The sampled imaging measures were compared statistically to the real-world measures using the Kolmogorov-Smirnov (KS) statistic. We then trained a support vector machine (SVM), a machine learning algorithm, using either the real-world or the sampled imaging measures and then used cross validation procedures to assess the accuracy of the machine-based classifications. In addition, we generated receiver operating characteristics (ROC) curves and computed the areas under the curves (AUCs) to assess and compare the fitted distributions with the distributions of the real-world measures. **ROC**=Receiver Operator Characteristic curve; **AUC**=Area under the ROC curve; **KS**=Kolmogorov-Smirnov; **SVM**=Support Vector Machine; **MVG**=Multivariate Gaussian Distribution.

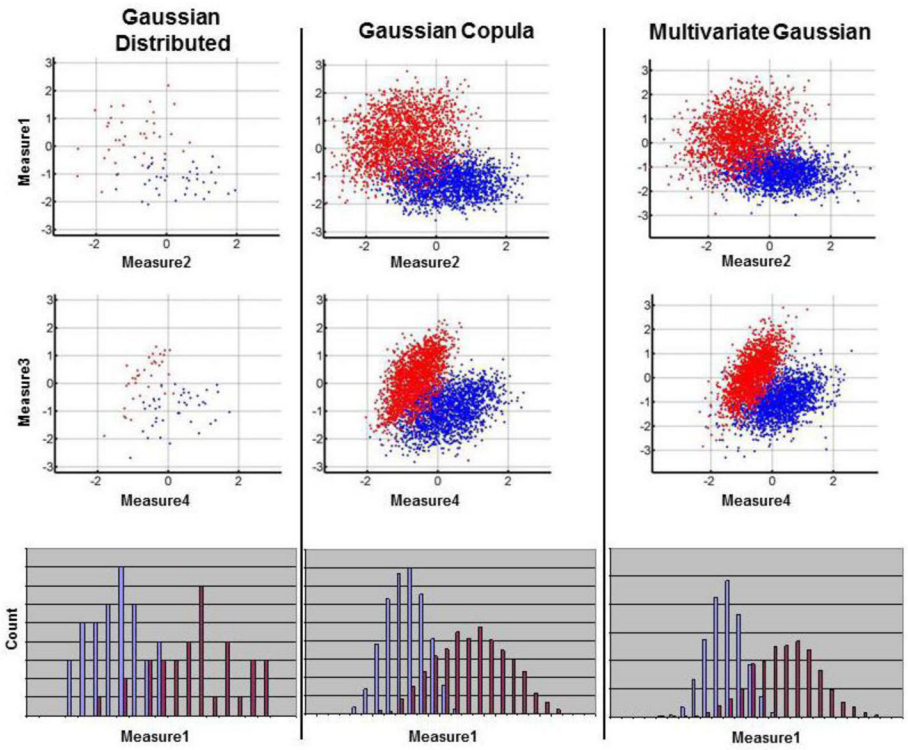


**Figure 3. Scatterplots of Simulated Measures and Measures Sampled from the Estimated Copula**

We fitted two multivariate Gaussian distributions to the real-world feature vectors with 6 brain imaging measures: one distribution to the feature vectors from 40 healthy adults (HA), and the other distribution to the feature vectors from 36 adults with Tourette Syndrome (TS). The first distribution was sampled to generate 40 simulated feature vectors that were representative of measures for healthy adults, and the other was sampled to generate 36 simulated feature vectors that were representative measures for TS adults. Using these simulated feature vectors, we estimated two Copulas and two multivariate Gaussian distributions: one Copula and one multivariate Gaussian for healthy adults, and the other Copula and multivariate Gaussian for TS adults. Each Copula, and each multivariate Gaussian distribution, was sampled to generate 2000 feature vectors for either the healthy adults or the TS adults. We used these sampled brain imaging measures to generate scatterplots and histograms in order to compare visually the distributions of the simulated measures that were Gaussian distributed and the measures that were sampled from the estimated Copula and multivariate Gaussian distributions. *Left Column:* The scatter plots of the four imaging measures and the histogram of one of the simulated measures. The imaging measures for TS adults are color coded in *Red* and those for healthy adults are coded in *Blue. Middle Column*: The scatterplots and the histograms for imaging measures sampled from the estimated Gaussian Copula. *Right Column*: The scatterplots and histograms for imaging measures sampled from the estimated multivariate Gaussian distribution. Visual comparison of the scatterplots shows that the sampled measures are distributed similarly to the simulated measures and that multivariate Gaussian distributions have a greater probability mass in their tails.

**HA**=healthy adults; **TS**=Tourette's Syndrome.

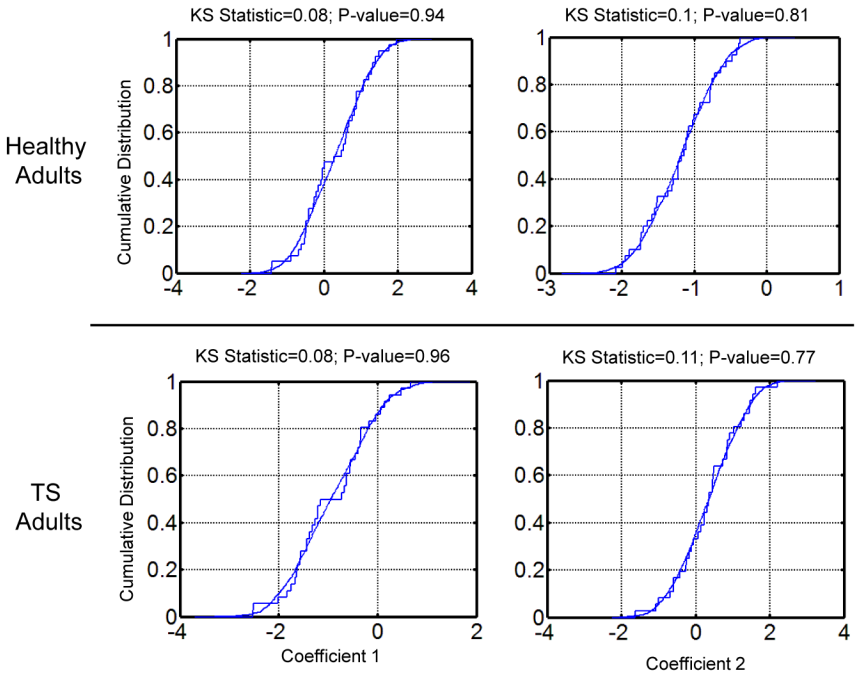


**Figure 4. Kolmogorov-Smirnov (KS) Statistics**

These were calculated to compare univariate distributions of the simulated measures that were Gaussian distributed and those of the measures sampled from the Gaussian Copulas (*left column*) and the multivariate Gaussian distribution (*right column*). The feature vectors of 6 simulated imaging measures each were computed from the brains of 40 healthy adults (HA) and 36 adults with Tourette Syndrome (TS), which were used to estimate the Copula and multivariate Gaussian distributions. The estimated multivariate distributions were sampled independently to generate 2000 feature vectors that were representative of the measures for healthy adults and another 2000 feature vectors that were representative of the measures for TS adults. We computed empirical cumulative distributions for each set of simulated and sampled imaging measures and calculated the KS statistic to compare quantitatively the cumulative distributions of the corresponding real-world and sampled measures. *Top Row*: Example cumulative distributions of two simulated and sampled imaging measures for healthy adults. *Bottom Row*: Example cumulative distributions for two simulated and sampled measures for TS adults. In each of these 4 examples, the P-value for the KS statistic is close to 1.0, indicating that the cumulative distributions of the sampled measures do not differ from those of the simulated ones.

**HA**=healthy adults; **TS**=Tourette's Syndrome.

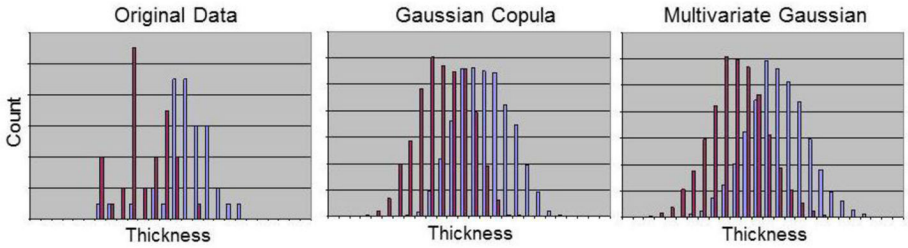


**Figure 5. Visual Assessment of the Distributions of Brain Imaging Measures**

We plotted the histograms of (1) *Left*: the original measures for 40 healthy adults (HA) and 36 adults with Tourette's Syndrome (TS), (2) *Middle*: measures sampled from the estimated Gaussian Copula, and (3) *Right*: measures sampled from the estimated multivariate Gaussian distribution. These plots show that the distributions of measures from the Gaussian Copula have a pronounced plateau and less probability mass in their tails than do the distributions of measures sampled from the multivariate Gaussian distribution.

**HA**=healthy adults; **TS**=Tourette's Syndrome.

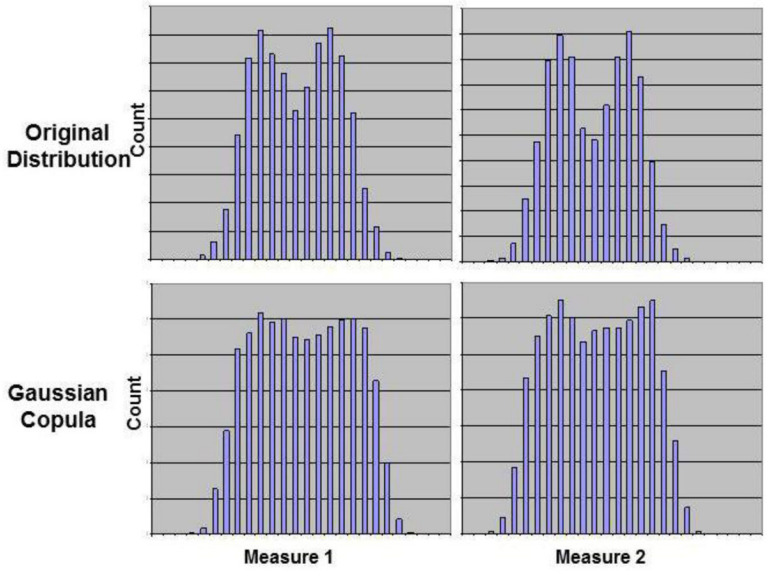


**Figure 6. Using the Copula to Estimate Multivariate Distributions that have Bimodal Marginal Distributions**

We constructed a bivariate distribution of two brain imaging measures that were statistically independent. Each measure was bimodally distributed as a sum of two Gaussian distributions with equal variances but differing means. We sampled 2000 imaging measures from this bivariate distribution and estimated a Gaussian Copula using the sampled measures. We then sampled the estimated Copula distribution to generate 2000 measures and plotted the histogram of the sampled values. Finally, we visually compared the histograms of imaging measures sampled from the bivariate distribution and those of measures sampled from the estimate Copula. *Top Row*: The histograms of the two imaging measures sampled from the bivariate distribution, which clearly shows that the two measures were bimodally distributed. *Bottom Row*: Although the two modes are less distinct, the histograms for the two measures sampled from the estimated Copula show that each measure was bimodally distributed. This example shows that a Gaussian Copula can estimate multimodal distributions better than a multivariate Gaussian distribution, because its marginal distributions are unimodal.

**Table 1**

**The Effect of Increasing Dimensions of the Feature Space on Accuracy of a Machine Learning Algorithm**

estimated using measures sampled from both an estimated Copula and an estimated multivariate Gaussian distribution. We increased the number of real-world brain imaging measures from two to four in the feature vectors for our cohort of 40 healthy adults (HA) and 36 adults with Tourette Syndrome (TS). Using the real-world feature vectors of either 2, 3, or 4 brain imaging measures, we independently estimated Gaussian Copula and multivariate Gaussian distributions for the healthy and TS adults. The estimated distributions were sampled to generate 2000 feature vectors of imaging measures for healthy adults and 2000 feature vectors for TS adults. We plotted Receiver Operating Characteristic (ROC) curves and computed the Kolmogorov-Smirnov (KS) statistics to assess the similarity of the imaging measures sampled from the estimated distributions to those of the real-world imaging measures. The sampled feature vectors were used to train and test a Support Vector Machine (SVM) to diagnose an individual as having TS or not. We evaluated the accuracy of the SVM algorithms using cross validation methods. The table shows that the areas under the ROC curves (AUCs) increased and their standard errors[72] deceased for increasing dimensionality of the feature space, indicating that the extent of the common support for the two distributions, one for healthy adults and the other for TS adults, decreased with the increasing dimensionality of the feature vector. The increase in AUCs was statistically significant at P-value<0.0001 for both the Copula and Multivariate Gaussian distributions when the number of features increased from 2 to 3 or from 3 to 4. In addition, the AUCs were higher for the Copula than for the multivariate Gaussian (P-value < 0.0001) when the number of features was either 3 or 4. Moreover, the AUCs for the measures sampled from multivariate Gaussian distributions were less than those for the Copula distributions, indicating that multivariate Gaussian distributions have greater probability mass in their tails. In addition, the KS statistics show that the univariate distributions of imaging measures sampled from both the Copula and multivariate Gaussian distributions did not differ from those of the real world measures. Finally, the misclassification rates demonstrated that the accuracy of the machine learning algorithm increased with the number of imaging measures in the feature vector. Thus, increasing dimensionality increased the separation between the distributions of imaging measures from the two diagnostic groups, thereby increasing the accuracy of the classification rules that the machine learning algorithms generated.

| Number of Features | Distribution | AUC | Misclassification Rates | | KS Statistic (P-value) | |
|---|---|---|---|---|---|---|
| | | | HA | TS | HA | TS |
| 2 | Copula | 0.912±0.001 | 0.1±0.0008 | 0.08±0.001 | 0.16 (0.22) | 0.11 (0.77) |
| | Multivariate Gaussian | 0.906±0.0007 | 0.1±0.0008 | 0.08±0.001 | 0.165 (0.23) | 0.093 (0.91) |
| 3 | Copula | 0.95±0.0005 | 0.05±0.0007 | 0.044±0.0008 | 0.16 (0.24) | 0.094 (0.91) |
| | Multivariate Gaussian | 0.93±0.001 | 0.07±0.001 | 0.07±0.001 | 0.15 (0.32) | 0.095 (0.9) |
| 4 | Copula | 0.96±0.0004 | 0.03±0.0005 | 0.03±0.0007 | 0.17 (0.18) | 0.094 (0.91) |
| | Multivariate Gaussian | 0.95±0.003 | 0.035±0.0006 | 0.046±0.001 | 0.16 (0.258) | 0.089 (0.93) |

**AUC**=area under the Receiver Operator Characteristic (ROC) curve; **KS**=Kolmogorov-Smirnov; **HA**=healthy adults; **TS**=Tourette's Syndrome; **SVM**=Support Vector Machine.

**Table 2**

**Performance of a Machine Learning Algorithm Using Measures Sampled from Copula or Multivariate Gaussian Distributions**

in our real-world cohort of 40 healthy adults (HA), 36 adults with Tourette Syndrome (TS), 26 adults with Bipolar Disorder (BD), 65 adults at a low risk (LR) for familial depression, and 66 adults at a high risk (HR) for familial depression. Using the real-world brain imaging measures from our participants, we estimated their Gaussian copula, Student's T-Copula, and multivariate Gaussian distributions, and then generated a dense set of imaging measures by sampling the estimated multivariate distributions. The sampled imaging measures were used to train Support Vector Machines (SVMs) to discriminate the brains of (1) LR and HR adults, (2) healthy adults and TS adults, or (3) healthy adults and BD adults. In general, AUCs were greatest for Copula distributions, whereas those for Gaussian distributions were similar to or less than those for the distributions of real-world measures. AUCs were significantly larger for the Copula than for the multivariate Gaussian distribution when discrimination healthy from TS adults and LR from HR adults (P-values<0.0001). Therefore, the support for Copula distributions was more compact and the multivariate Gaussian distributions had greater probability mass in their tails. KS statistics showed that the univariate distributions of imaging measures sampled from both the Copula and multivariate Gaussian were similar to the univariate distributions of real-world measures, with the Copula generating measures that were slightly closer to the real-world imaging measures than were measures sampled from multivariate Gaussian distributions. Furthermore, misclassification rates in our cross validation analyses suggested that SVMs trained using measures sampled from the Copula distributions performed significantly better than did SVMs trained using either the real-world measures or the measures sampled from the multivariate Gaussian distributions. Finally, variability in the misclassification rates was significantly smaller for SVMs trained using samples from the Copula distributions. Thus, brain imaging measures sampled from the estimated Copula yield SVMs that are stable in the presence of variability in the sampling of the participants who provided the imaging measures.

| Cohort | Brain Measures | AUC | Misclassification Rates | | K-S Statistic (P-value) | |
|---|---|---|---|---|---|---|
| | | | LR | HR | LR | HR |
| **LR (65) & HR (66)** | Real-World | 0.73±0.014 | 0.23±0.02 | 0.3±0.025 | | |
| | Multivariate Gaussian | 0.738±0.0007 | 0.18±0.002 | 0.29±0.002 | 0.079 (0.82) | 0.08 (0.79) |
| | Copula | 0.82±0.001 | 0.13±0.0025 | 0.2±0.001 | 0.072 (0.9) | 0.06 (0.94) |
| | | | **HA** | **TS** | **HA** | **TS** |
| **HA (40) & TS (36)** | Real-World | 0.943±0.01 | 0.05±0.015 | 0.06±0.024 | | |
| | Multivariate Gaussian | 0.92±0.0004 | 0.08±0.001 | 0.084±0.001 | 0.16 (0.264) | 0.094 (0.91) |
| | Copula | 0.97±0.0005 | 0.024±0.0004 | 0.023±0.0007 | 0.167 (0.22) | 0.09 (0.876) |
| | | | **HA** | **BD** | **HA** | **BD** |
| **HA (40) & BD (26)** | Real-World | 0.98±0.01 | 0.02±0.02 | 0.025 | | |
| | Multivariate Gaussian | 1 | 0.005 | 0.005 | 0.117 (0.65) | 0.12 (0.931) |
| | Copula | 1 | 0 | 0 | 0.114 (0.68) | 0.087 (0.99) |

**AUC**=area under the Receiver Operator Characteristic (ROC) curve; **KS**=Kolmogorov-Smirnov; **HA**=healthy adults; **BD**=Bipolar Disorder; **TS**=Tourette's Syndrome; **HR**=high risk for familial depression; **LR**=low risk for familial depression; **SVM**=Support Vector Machine.