An Improved Conditional Wasserstein GAN for Synthetic Brain Imaging

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding the individual variability of brain function and its association with behavior is one of the major concerns in modern cognitive neuroscience. To this end, we present a novel generative model trained on real neuroimaging data to synthesize task-dependent functional brain images. Our model is constructed as a conditional generative adversarial network combined with three dimensional convolutions. This architecture enables the modeling of high dimensional brain image tensors with structured spatial correlations. Our results suggest that the proposed model is able to generate high quality synthetic brain images which are diverse and task dependent. We also demonstrate the neuroscientific utility of synthetic brain imaging for prospective power analyses.

1 Introduction

Human brain activity, as measured by functional Magnetic Resonance Imaging (fMRI), varies significantly between individuals. In response, modern analysis now prioritizes understanding the inter-subject variability of brain function [Dubois and Adolphs, 2016, Geerligs et al., 2017]. Our work is motivated by the view that generative models provide a useful tool for understanding this variability as they enable the synthesis of a variety of plausible brain images representing different hypothesized individuals and high-quality generative models can be analyzed to posit potential mechanisms that explain this variability [Horn et al., 2008]. This manuscript provides – to our knowledge for the first time, positive results suggesting that it is indeed possible to generate high quality, diverse, and task dependent functional brain images. Beyond providing a model for individual variability, high quality brain image synthesis addresses pressing data issues in cognitive neuroscience. Progress in the computational neurosciences is stifled by the difficulty of obtaining brain data [Poldrack and Gorgolewski, 2014]. For the computational neuroscientist, generated images deliver *unlimited* quantities of high quality brain imaging data that can be used to develop state of the art tools before application to real subjects and/or patients [Varoquaux and Thirion, 2014]. This approach of using modern generative models to synthesize data, which in turn accelerates scientific study, has already proven useful in fields such as particle physics and astronomy [Castelvecchi et al., 2017]. Our work represents a first application of this approach to neuroscience.

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] are a popular family of generative modeling techniques. Wasserstein GANs (WGANs) [Arjovsky et al., 2017] formulate the GAN objective using the Wasserstein distance. Recently, training of WGANs [Gulrajani et al., 2017] has been improved by applying an additional gradient penalty to the discriminator objective that avoids pathological behavior caused by weight clipping. We propose the Improved Conditional Wasserstein GAN (ICW-GAN), a 3D conditional GAN developed to synthesize fMRI brain imaging data. The ICW-GAN can be described as the Improved Wasserstein GAN augmented with 3D inputs and label conditioning. The proposed model is evaluated using a series of image classification tasks. Results show that augmenting training datasets of downstream classifiers using synthetic

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

data generated by our models can greatly improve test classification accuracy of brain volumes. Additionally, we qualitatively assess the quality and diversity of generated brain volumes. Our results suggest that the proposed models are able to generate high-quality, task-dependent, and diverse three-dimensional brain images with direct neuroscientific application to prospective power analyses. We further demonstrate that by exploiting known relationships between experiments, our generative neuroimaging model can synthesize imaging results for experiments which are not present in the training set.

We consider the utility of our generative neuroimaging model for low-cost but reliable statistical power analysis. Statistical power is the conditional probability of a test rejecting the null hypothesis given that the alternative hypothesis is true. Prospective power analyses are a standard step in scientific experiment design to maximize statistical power and ensure resources are not wasted on unnecessary subjects or futile experiments. In a traditional power analysis, real data is collected and used to determine the sample size needed to achieve a desired experimental power. Unfortunately, MRI scanning costs $500+ an hour[1]. Because of the insuffiency and acquisition cost of data, cognitive neuroscientists often attempt to obtain stand-in data from other research groups or mine effect sizes and variance estimates from previously published studies. However, published articles are biased towards significant activation and high power estimates, leading to the design of new experiments that are often severely underpowered Lakens and Albers [2017]. Power analysis results computed on this synthetic data suggest that it can be a reliable and low-cost substitute for real data used in experiment design. To our knowledge, we are the first to explore implicit generative models for statistical power analysis and these results may be of independent interest. All the code for our analysis is provided[2].

2 Background and Related Work

We are unaware of any published work using neural networks to generate brain imaging data. However, neural networks have been used for classifying brain imaging data. [Firat et al., 2014] and [Koyamada et al., 2015], used 2D deep nets to extract features of fMRI brain images to classify brain states. [Nathawani et al., 2016] applied both 2D and 3D neural networks to classify fMRI brain data. [Svanera et al., 2017] decoded fMRI data of video stimuli and classified data into visual categories. Similarly, [Nathawani et al., 2016] extracted features from 4D fMRI data and used deep learning methods for discrimination of cognitive processes.

To learn a distribution over data $\boldsymbol{x}$, a GAN [Goodfellow et al., 2014] formulates a 2-player non-cooperative game between two deep nets. The *generator* G takes a random noise vector $\boldsymbol{z}$ sampled from a prior distribution $P_{\boldsymbol{z}}(\boldsymbol{z})$ as input and produces an image $G(\boldsymbol{z})$. The generator G is trained to fool the discriminator D, which receives either synthetic data or real data and is trained to differentiate between them. In a conditional GAN [Mirza and Osindero, 2014], both G and D are conditioned on some extra information $\boldsymbol{y}$, for instance, class labels or other features. This conditioning can be accomplished by feeding an encoding of $\boldsymbol{y}$ into both the generator and the discriminator. This approach is successful in one-to-many mappings such as image labeling with many tags.

The Wasserstein GAN (WGAN) [Arjovsky et al., 2017] objective uses the *Wasserstein-1* distance:

$$\min_{\theta} \max_{w \in W} \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[D_w(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[D_w((G_\theta(\boldsymbol{z})))], \tag{1}$$

where $\{D_w\}_{w \in W}$ denotes a set of functions that are K-Lipschitz for some K. Intuitively, the Wasserstein metric between distributions measures the minimum cost of transporting mass to transform the distribution P_{data} into the distribution $P_{\boldsymbol{z}}$. This loss is continuous everywhere and its gradient with respect to its input has been found to be more stable than its classical GAN counterpart.

[Gulrajani et al., 2017] argues in the Improved Wasserstein GAN (IWGAN) that weight clipping of the critic in WGANs inevitably causes the gradient to either vanish or to explode. To address this issue, [Gulrajani et al., 2017] proposes an alternative penalty term in the critic loss based on the gradient norm when optimizing:

$$\mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[(1 - D_w(G_\theta(\boldsymbol{z})))] - \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[D_w(\boldsymbol{x})] + \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{\boldsymbol{x}}}}[(\parallel \nabla_{\hat{\boldsymbol{x}}} D_w(\hat{\boldsymbol{x}}) \parallel_2 - 1)^2]. \tag{2}$$

In this optimization problem formulation, $\hat{\boldsymbol{x}}$ is a convex combination of real data and artificial samples, i.e., $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon) G(\boldsymbol{z})$ with ϵ drawn from a uniform distribution ($\epsilon \sim U[0, 1]$).

[1] http://fmri.research.umich.edu/users/billing.php

[2] Code link anonymized for review, will published upon acceptance.

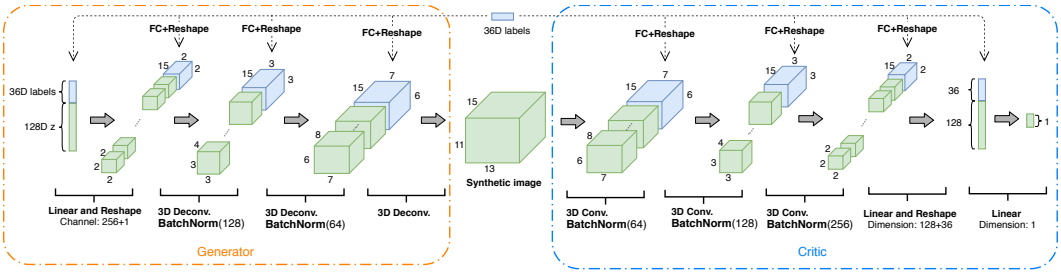


Figure 1: The architecture of our ICW-GAN. For the generator (in the orange box), the 128 dimensional encoding z is drawn from a multivariate normal distribution. The label vector is a binary encoding. It is concatenated to input and hidden layers and for each of these layers, fully connected layers transform the label vector to volumes. Our stride in the de-convolutional layers is [1,2,2,2,1] in the batch, height, width, length and feature map dimension. Batch normalization is leveraged in the feature map dimension, sizes of which are in parentheses, i.e. 64 and 128. The structure of the critic is illustrated in the blue box, which is almost a mirrored generator in terms of the structure, convolutional strides, and use of labels.

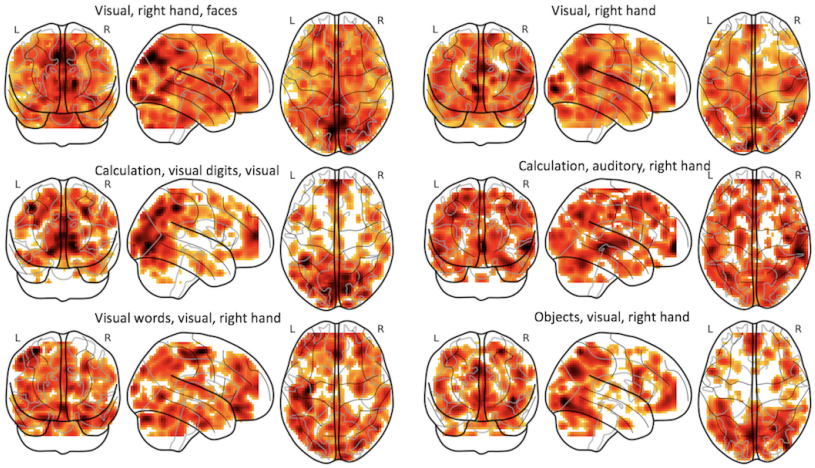


Figure 2: 2D projections of synthetic brain volumes generated using the proposed ICW-GAN. The tag above each image sequence indicates the class to which the data belongs.

3D-GANs [Wu et al., 2016] extend GANs to 3D object generation. Different from classical GANs, 3D-GANs apply the three dimensional convolution in both the generator and the discriminator. By learning deep object representations, 3D GANs can generate visually appealing yet variable 3D object volumes.

3 Our Approach

In the following, we introduce our 3D Improved Conditional Wasserstein GAN (ICW-GAN) model for fMRI data generation. The proposed ICW-GAN differs from existing GAN models in structure and use of label information. Similar to classical generative adversarial networks (GANs), ICW-GANs are formulated as a non-cooperative two-player game between two adversaries: (1) a generator $\hat{\boldsymbol{x}} = G_\theta(z)$, which generates artificial samples $\hat{\boldsymbol{x}}$ from randomly drawn latent encodings z via a transformation using a deep net parameterized by θ; and (2) a discriminator $D_w(\boldsymbol{x})$ represented via the logit obtained from a deep net parameterized by w.

We construct our models to minimize the improved Wasserstein distance in Equation 2. Our model extends the IW-GAN in two ways. First, because fMRI data is three dimensional, 3D convolution and deconvolution are used to capture the spatial structure of the voxel information. Second, both the discriminator D_w and the generator G_θ are conditioned on available labels. As a result, our ICW-GAN integrates the advantages of C-GANs, IW-GANs and 3D-GANs.

The overall model architecture is illustrated in Figure 1. Our generator consists of three fully convolutional layers with kernels of size $4 \times 4 \times 4$ and stride 2, batch normalization and Leaky ReLU layers added between, and a $\tanh$ layer at the end. The critic architecture is a mirrored generator, which also employs the labels as additional information. The critic differs in that the final layer uses linear activation.

To include label information, we concatenate labels to the input and hidden layers. At the input of the generator, binary labels are combined with the brain vector. Then, for each of the intermediate layers, we use a fully connected layer followed by a $\tanh$ activation to transform the binary vector to a volume of appropriate size, i.e., $15 \times 2 \times 2$ for the first hidden layer, and $15 \times 3 \times 3$ for the next. We empirically found that the model is not sensitive to the choice of volume size. We concatenate the label volume to intermediate volumes and pass the joint to the next deconvolutional layer. We follow the same procedure in the architecture of the discriminator. We note in passing that we experimented with concatenating labels in various layers and did not observe significant differences. Thus, we chose to implement the model using full concatenation of labels.

The objective function of our ICW-GAN model is as follows:

$$L = \mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))] - \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[D(\boldsymbol{x}|\boldsymbol{y})] + \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{\boldsymbol{x}}}}[(\| \nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}|\boldsymbol{y}) \|_2 - 1)^2], \quad (3)$$

where $\boldsymbol{y}$ denotes the volume labels, and $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon) G(\boldsymbol{z})$, $\epsilon \sim U[0, 1]$. λ is a gradient penalty coefficient. We optimize both the discriminator and generator loss using an Adam optimizer [Kingma and Ba, 2014].

Table 1: Classification results. Input represents the input of a classifier: Real means only use real training data, while 'Real+Synth.' means real training data plus synthetic data which is produced by ICW-GAN. Two classifiers are utilized: SVM and neural networks (NN). Results show that augmenting real data with synthetic data improves classification performance.

Input	Classifier	Accuracy	Macro F1	Precision	Recall
Real	SVM	0.797	0.797	0.813	0.797
Real	NN	0.802	0.802	0.817	0.802
Real+Synth.	SVM	0.806	0.803	0.823	0.807
Real+Synth.	NN	**0.819**	**0.817**	**0.830**	**0.819**

3.1 Downstream Classifiers for ICW-GAN

We consider classification with data augmentation to quantitatively investigate the hypothesis that the generated images accurately reproduce the conditional image statistics. We note that support vector machines (SVMs) and neural network classfiers are state of the art for fMRI applications [Pereira et al., 2009] and we consider it sufficient to employ them for evaluation. To this end, we compare the SVM and the 3D deep net classifier trained with real brain images ('Real') or real plus synthetic brain images ('Real+Synth.'). Accuracy, macro F1, precision and recall metrics are used to measure the results.

SVM: We use a simple linear SVM to classify test data and do not extract any intermediate features. Instead, as is common in whole brain classification literature Pereira et al. [2009], Varoquaux and Thirion [2014], we use raw brain data, vectorized to a 1-dimensional vector.

Deep Net: The deep net structure is broadly similar to the discriminator with a 3 dimensional structure and identical number of convolution layers with Leaky ReLU activations. Unlike the discriminator, the classifier doesn't concatenate intermediate and input data with any label information.

3.2 Synthetic Power Analyses

In a traditional power analysis, real data is collected and used to determine the sample size needed to achieve a desired experimental power $P^* = \mathbb{E}[\mathbb{1}\{p < \alpha\}]$, where p represents the p-value in statistical hypothesis testing and α is the pre-defined rate of type 1 error.

In our simulated experiments, we mimic a power analysis with both real and synthetic data by considering samples from true underlying distributions. To determine the sample size required to achieve the desired experimental power P^*, we simply find the sample size n with power greater than

or equal to P^*. If at every sample size the power of a test computed between synthetic distributions is similar to the power of a test computed between real distributions, then using synthetic data as a stand-in for real data will yield a similar sample size estimate to what we would have achieved with real data. Further details on the power analysis experiments are provided in the appendix.

4 Experiments

This section illustrates the qualitative and quantitative evaluation of the proposed ICW-GAN. First, we present detailed qualitative results via examples of generated 3D volumes. Next, we present quantitative results for 3D volume classification via training of downstream classifiers on mixtures of real and synthetic data. We further employ a Gaussian Mixture Model (GMM) and simple data augmentation methods as baselines – showing that simpler generative models do not achieve the same results. Finally we show that the ICW-GAN can be used to synthesize reliable prospective power analysis results for neuroimaging data.

Neurovault [Gorgolewski et al., 2015] is currently the largest open database of preprocessed neuroimaging data. We evaluate the performance of the proposed ICW-GAN on the three largest Neurovault functional brain image collections 1952, 2138 and 503[3]. Due to limited space, we only show results using collection 1952. Following standard preprocessing procedures, the brain image is downsampled[4] to $13 \times 15 \times 11$ using the nilearn python package[5]. Additional results, i.e. experiments with additional datasets and various resolutions, are provided in the Appendix.

4.1 Results

Collection 1952 was obtained from OpenfMRI, the Human Connectome Project, and Neurospin research center. With 6573 brain images and 45 classes with a total number of 19 sub-classes, i.e. a multi-label encoding, collection 1952 is designed to map a wide set of cognitive functions. The labels are the set of cognitive processes associated with the image e.g. 'visual', 'language', and 'calculate'. We observe that only 45 of the possible unique label combinations are observed. Over the dataset, classes with more than than 100 images are split into training, validation and testing subsets with a ratio 7:1:2. For classes with less than 100 images but more than 30 images, we use a 3:1:2 data split. Nine classes with less than 30 samples are ignored. This leaves a total of 36 classes with at least 30 examples. We train the ICW-GAN for 2500 epochs with a learning rate of 1e-4 and 50% exponential decay for each of 3000 iterations with a batch size of 50.

Visualization of synthetic images 2D projections of several brain volumes generated by the ICW-GAN are illustrated in Figure 2. Note that before projecting, we upsample the images to the original spatial resolution. Images are plotted after thresholding values smaller than 0.48 to highlight areas with the largest activation. Note that thresholding is standard practice when presenting neuroimaging data. The projections in Figure 2 are synthesized brain images conditioned on the accompanying cognitive process labels. Visual examination of the generated images by neuroscience experts suggests high quality and high diversity. In particular, experts report high activation in the appropriate brain regions e.g. the motor cortex for motor labels, and the visual cortex for visual labels.

Classification results To further assess the quality of the generated data, we evaluate the performance of downstream classifiers. Note that the test data for classification is always composed of real images. The classification results are shown in Table 1. The first column indicates the type of training data we use for the classifier: only using real data ('Real'), or using the mixed data of real and generated volumes ('Real+Synth'). The second column denotes the classifier type, i.e., an SVM or a deep neural net ('NN'). We use the validation dataset to choose the best training models and use these models to classify the test data. We observe that including the synthetic data is generally beneficial for classifier training. We also observe that the deep net classifier generally outperforms SVMs.

[3]Collections are publicly available online, e.g., https://neurovault.org/collections/503

[4]For the interested reader, we note that downsampling and parcellation is a common practice in fMRI analysis, as standard preprocessing renders fMRI images to be spatially smooth [Horn et al., 2008]. In addition, there is significant evidence in the fMRI literature that downsampling has limited effect on classifier performance (particularly cross-subject brain alignment and subsequent smoothing, see Figure 3 of [Dubois and Adolphs, 2016]). This is not necessarily true for other kinds of brain imaging. For example, structural brain images can be reliably analyzed at higher resolution.

[5]http://nilearn.github.io

Table 3: Comparison between GMMs, simple data augmentation and ICW-GAN. We list 9 training data strategies: in the first four rows, we only use synthetic data to train the deep net classifier while in the last two rows, we mix real and synthetic data together to train the same classifier. The 5th to 7th rows present the results of adding 'real training data' which are disturbed by Gaussian noises.

Training data	Accuracy	F1	Precision	Recall
Synth. data from GMM (20 images/class)	0.203	0.309	0.309	0.202
Synth. data from GMM (500 images/class)	0.720	0.725	0.765	0.720
Synth. data from ICW-GAN (20 images/class)	0.458	0.433	0.537	0.458
Synth. data from ICW-GAN (500 images/class)	0.783	0.776	0.805	0.783
Real+'Real data' with Gaussian noise(var=0.1)	0.782	0.781	0.807	0.782
Real+'Real data' with Gaussian noise(var=0.05)	0.810	0.814	0.832	0.811
Real+'Real data' with Gaussian noise(var=0.01)	0.813	0.809	0.829	0.807
Real+Synth. (from GMM)	0.793	0.798	0.824	0.793
Real+Synth. (from ICW-GAN)	**0.819**	**0.817**	**0.830**	**0.819**

Variance of cross-validated performance in ICW-GAN We tested our model with various cross-validation settings and calculated the variances of the evaluation metrics (Table 2). Except for the number of folds used when partitioning the test dataset, the training strategy for 5-fold and 10-fold cross-validation is similar to that of 3-fold cross-validation. The small variances suggest that the reported accuracy differences, while small, are indeed significant.

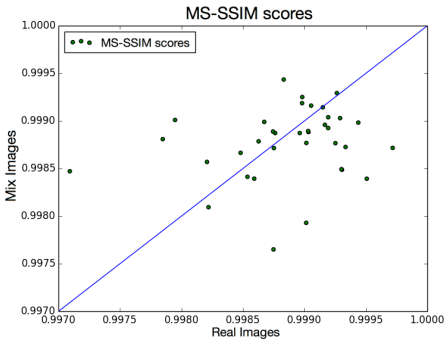
Table 2: Accuracy, F1, Precision, Recall and their variance (column 3,5,7,9) for 3-fold, 5-fold and 10-fold cross validation. We conducted this experiment with the training data of mixed 'Real+Synth.' data.

N-fold	Accuracy	Var_Acc	F1	Var_F1	Precision	Var_P	Recall	Var_R
3-fold	0.819	0.0002782	0.817	0.0002543	0.830	0.0001843	0.819	0.0002432
5-fold	0.837	0.0002432	0.841	0.0001985	0.862	0.0001643	0.837	0.0002434
10-fold	0.843	0.0002782	0.857	0.0003027	0.894	0.0004785	0.843	0.0003789

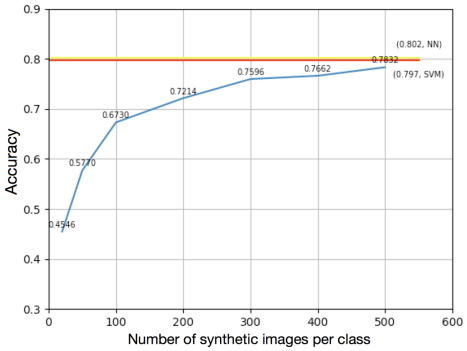
MI-SSIM score Inspired by Odena et al. [2016], we compute the multi-scale structural similarity (MS-SSIM) [Wang et al., 2003] to examine the intra-class diversity of the data generated by ICW-GAN. MS-SSIM score values are within $[0, 1]$. Higher MS-SSIM values correspond to more similar images. We measure the mean MS-SSIM score with 100 randomly chosen volumes of mixed data with a given class, and ones of real training data. We find that 22 classes with mixed data have a lower MS-SSIM score than only with real training data. In other words, 61.1% classes with mixed data have sample variability that exceeds those only with real training data. See Figure 3(a) for details.

Results of only using generated data for training To further evaluate the quality of the generated data, we trained a classifier using only generated data and tested with real data. In this experiment, we used the deep net classifier for evaluation and varied the number of input samples for each class. Figure 3(b) shows that in general, the test accuracy improves as the amount of artificial training data increases. The red and yellow lines show the accuracy obtained when training the classifiers only with real images. The curves in Figure 3(b) also suggest that using sufficient numbers of synthetic data can perform well as real images.

Comparing to other baselines We also compare our ICW-GAN model to the following baseline methods: data augmentation by adding Gaussian noise and Gaussian Mixture Models (GMMs). First, a simple data augmentation method is to add noise to the original data. We mix together the real data and 5000 'real images' with additive Gaussian noise. Remember that the original brain images were normalized to [-1,1]. To ensure that the data remains in a similar range, we chose to add Gaussian



(a) Mean MS-SSIM scores between pairs of images within a given class, which were calculated on real data and synthetic data (using the ICW-GAN). Each point represents an individual class. Values in horizontal axis are MS-SSIM scores computed on real images, while values on the vertical axis are calculated using mixed images.



(b) We only use generated data from the ICW-GAN to train the deep net classifier (blue curve). Red and yellow lines mean only using real data with the two classifiers

Figure 3: MS-SSIM scores and the accuracy curve for the classification when training only with synthetic data.

noise with mean value 0 and variance 0.01, 0.05 and 0.1. Next, we trained a Gaussian Mixture Model (GMM) to generate brain images. We trained a separate GMM for each label. When synthesizing a new image, we first chose a specific class and then randomly sampled from the trained GMM. When training the deep net classifier only with synthetic images, 20 (or 500) images are used for each class (20 or 500 $\times$number of classes in all). The classification results are shown in Table 3. The evaluation scores of the ICW-GAN are significantly higher than the performance obtained for all augmentations with these baseline methods, particularly when only using synthetic data to train the classifier.

Taken together, our results illustrate that the ICW-GAN achieves much improved performance on the dataset. The proposed approach consistently improves accuracy metrics – which is by far the most popular metric for evaluating multi-class classification performance. The ICW-GAN also outperformed simple data augmentation methods and GMMs as a generative model baseline – clearly illustrating that not all data augmentation and generative models are suitable to for the task of synthetic functional brain image generation.

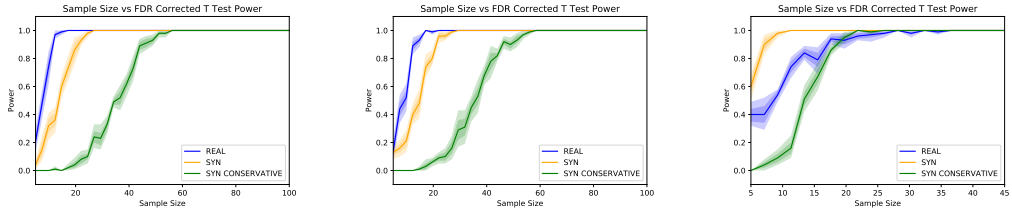
4.2 Synthetic Power Analyses

We provide empirical evidence to suggest that synthetic functional magnetic resonance images could serve as a low-cost and statistically reliable replacement for real data used in prospective power analyses. We emphasize that since the power analysis is used to select the sample size, it is especially important that the power curve is conservative. Thus our aim is to ensure that the synthetic data power curves always lie below the real data power curve.

Conservative power analysis overestimates the number of samples required, so we may collect unnecessary samples but the study is guaranteed to have the desired power. Compared to aggressive analysis which underestimates sample size and leads to underpowered research, we prefer our synthetic power analysis results to be conservative in their sample size recommendations. We implement the standard test and also a conservative test ensuring conservative power analysis. For each experiment, the result is averaged over 10 trials by comparing 20,000 samples. We only report neuroimaging application results here. Experimental details and additional synthetic power analysis experiments are presented in the appendix.

4.2.1 Individual Cognitive Process Tags

Figure 4(a) illustrates the power curve of simulated traditional, synthetic, and conservative-adjusted synthetic power analyses comparing samples of brain images with the 'visual' cognitive process tag to brain images without the 'visual' cognitive process tag. Similarly, Figure 4(b) illustrates the



(a) Individual cognitive process tag 'Visual' vs NO 'Visual'

(b) Individual cognitive process tag 'Auditory' vs NO 'Auditory'

(c) New cognitive process tag combination 'Visual, Auditory' vs NO 'Visual, Auditory'

Figure 4: False discovery rate corrected t-test power analyses for individual tags and tag combinations. Each figure includes Traditional power analysis (using real data, not availiable in practice) compared to synthetic brain imaging-based power analysis, and our proposed additional conservative adjustment for the synthetic analysis.

power curve of simulated traditional, synthetic, and conservative-adjusted synthetic power analyses comparing samples of brain images with the 'auditory' cognitive process tag to brain images without the 'auditory' cognitive process tag. Note that in each experiment, the real and synthetic power curves are very similar. As the number of samples increases, all three power curves converge to 100% power. In 4(a) and 4(b), the conservative adjustment is not necessary as the raw synthetic data already yields a conservative estimate. However, as illustrated in 4(c), the raw synthetic data is not guaranteed to be conservative and it is therefore important that the conservative power curve still yields a reasonable power estimate.

4.2.2 Novel Cognitive Process Tag Combinations

Our ICW-GAN model can synthesize cognitive process tag combinations that are not present in the original training set, in effect simulating results of experiments that have never been done before. To validate the statistical properties of such data, we train our ICW-GAN model after removing all examples of any tag combination of interest "x". We then generate synthetic examples of tag combination "x" and compute a power analysis with the left-out real and generated synthetic examples of tag combination "x". Figure 4(c) presents the power curve of simulated traditional, synthetic, and conservative-adjusted synthetic power analyses comparing samples of brain images with the 'visual, auditory' cognitive process tag combination to brain images without the 'visual, auditory' cognitive process tag combination. The conservative analysis maintains a tight underestimate of power, which is a desirable property in neuroscientific application. These results suggest that the synthetic data is a reliable replacement for real data used in prospective power analyses.

5 Conclusion

Generative models provide a useful tool for understanding the individual variability of brain images. The results of this manuscript show – to our knowledge for the first time, that 3D conditional GANs, in particular our proposed ICW-GAN, can generate high quality diverse and task dependent brain images. We hope our results inspire additional research on generative models for brain imaging data. Beyond qualitative evaluation, we evaluate quantitative performance by using the generated images as additional training data in a predictive model – mixing synthetic and real data to train classifiers. The results show that our synthetic data augmentation can improve classification accuracy. The ICW-GAN can easily outperform the generative baselines of GMMs and data augmentation with Gaussian noise, which illustrates not all data augmentation methods are suitable for the task of image generation. We further demonstrate the neuroscientific utility of high quality synthetic brain imaging as statistically reliable stand-in data in prospective power analyses – a novel application of implicit generative models that may be of independent interest. Future work will focus on additional qualitative evaluation of the generated images by neuroscience experts and exploration of additional applications. We also plan to more throughly investigate the trained models to explore what it may contribute to the science of individual variability in neuroimaging. Finally, we plan to expand our models to combine data across multiple studies – each of which use different labels, by exploring techniques for merging labels based on the underlying cognitive processes [Poldrack, 2006].

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GANs. *arXiv preprint arXiv:1701.07875*, 2017.

D. Castelvecchi et al. Astronomers explore uses for AI-generated images. *Nature*, 2017.

L. J. Chang, P. J. Gianaros, S. B.Manuck, A. Krishnan, and T. D. Wager. A sensitive and specific neural signature for picture-induced negative affect. *PLoS biology*, 2015.

J. Dubois and R. Adolphs. Building a science of individual differences from fMRI. *Trends in cognitive sciences*, 2016.

O. Firat, L. Oztekin, and F. Vural. Deep learning for brain decoding. In *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014.

L. Geerligs, K. A. Tsvetanov, and R. N. Henson. Challenges in measuring individual differences in functional connectivity using fMRI: The case of healthy aging. *Human Brain Mapping*, 2017.

I. J. Goodfellow, J. A. Pouget, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, 2014.

KJ Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwarz, SS Ghosh, C. Maumet, VV. Sochat, T. E-Nichols, Russell A Poldrack, J-B. Poline, et al. Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, 9, 2015.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

J. D. Van Horn, S. T. Grafton, and M. B. Miller. Individual variability in brain activity: a nuisance or an opportunity? *Brain imaging and behavior*, 2008.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Koyamada, Y. Shikauchi, K. Nakae, M. Koyama, and S. Ishii. Deep learning of fmri big data: a novel approach to subject-transfer decoding. *arXiv preprint arXiv:1502.00093*, 2015.

D. Lakens and C. Albers. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. 2017.

M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

D. D. Nathawani, T. Sharma, and Y. Yang. Neuroscience meets deep learning, 2016.

A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 2016.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 2009.

R. A. Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, 2006.

R. A. Poldrack and K. J. Gorgolewski. Making big data open: data sharing in neuroimaging. 2014.

M. Svanera, S. Benini, G. Raz, T. Hendler, R. Goebel, and G. Valente. Deep driven fMRI decoding of visual categories. *arXiv preprint arXiv:1701.02133*, 2017.

G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 2014.

Z. Wang, E.-P. Simoncelli, and A.-C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, 2003.

J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, 2016.

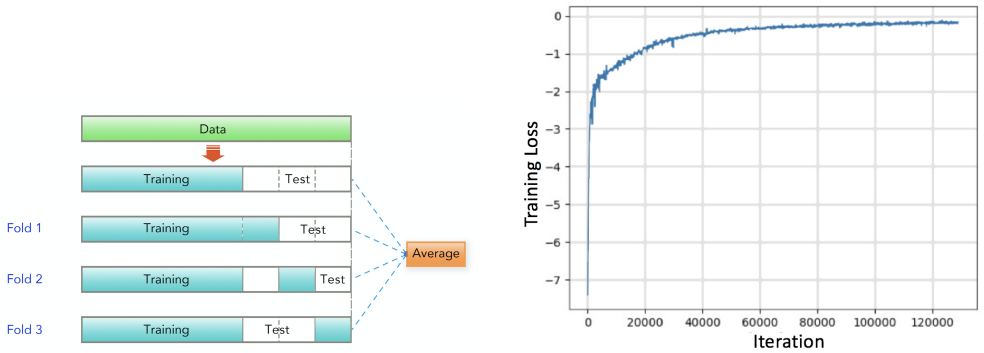
Appendix A Additional details and results

In Section A.1 We show our cross validation strategy and training loss curve of the ICW-GAN. We present additional results from collection 1952 and all results for collections 2138 and 503 in A.2. Note that all visualized images have the same processing procedure as described by Figure 2.

A.1 Cross validation strategy and loss curve

Cross-validation strategy Our cross-validation strategy for mixed real/generated data is illustrated in Figure 5(a). For 3-fold cross-validation, the data is first partitioned into training and test sets. Next, we partition the validation data into three subsets. For each fold, the model is trained on the concatenation of the training data, and one of the validation data subsets (potentially augmented by generated data), and tested on the remainder. This strategy ensures sufficient training data, and ensures that the test set consists of only real data. All presented results are averaged over the three rounds.

Training loss curve of the ICW-GAN Figure 5(b) illustrates the ICW-GAN training loss curve from one fold. This result is presented to emphasize the training stability of the proposed ICW-GAN.



(a) Visual illustration of our 3-fold cross validation strategy.

(b) Training loss curve of the ICW-GAN which shows the stability of our model.

Figure 5: Cross validation strategy and training loss curve.

A.2 Additional results on Neurovault collection 1952, 2138 and 503

Dataset 1: Collection 1952 We present the classification results for downsampled $2\times$ data ($26 \times 31 \times 23$ spatial resolution) in Table 4.

Table 4: Classification results for downsampled $2\times$ data. 'Input' describes the training set of a classifier: 'Real' indicates only real training data while 'Real+Synth' indicates real training data mixed with synthetic data produced by the ICW-GAN model. Two classifiers are evaluated: SVM and neural networks (NN). Results show that augmenting real data with synthetic data improves classification performance.

Input	Classifier	Accuracy	Macro F1	Precision	Recall
Real	SVM	0.855	0.857	0.867	0.857
Real	NN	0.863	0.863	0.872	0.863
Real+Synth.	SVM	0.860	0.863	0.860	0.857
Real+Synth.	NN	**0.891**	**0.894**	**0.906**	**0.891**

Dataset 2: Collection 2138 The collection 2138 dataset includes data from the Individual Brain Charting (IBC) project; developed to collect high resolution fMRI to map 12 subjects that undergo a

large number of tasks: the HCP tasks, the ARCHI tasks, a specific language task, video watching, low-level visual stimulation etc. There are 1847 brain images, 61 classes and 50 labels in collection 2138. Because of the small size of the dataset, we randomly choose 70% of the brain images as training data and leave 30% as test data. In this case, we do not have development data to supervise the training process. Thus, we train our models for 1000 epochs in several runs and record the best classification results which are summarized in Table 5.

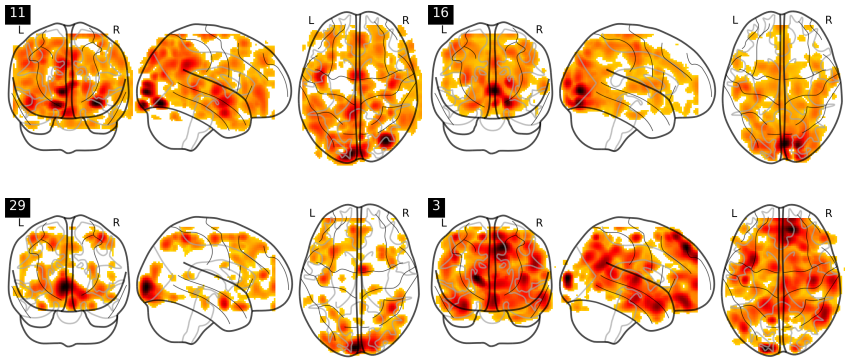


Figure 7: Brain images in collection 503 generated by ICW-GAN and a left top tag represents the stimulating picture for subjects.

Table 5: Results on collection 2138. As before, results show that augmenting real data with synthetic data improves classification performance.

Downsampling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
8.0×	Real	SVM	0.523	0.480	0.497	0.523
	Real	NN	0.530	0.517	0.545	0.530
	Real+Synth.	SVM	0.531	0.493	0.510	0.533
	Real+Synth.	NN	**0.562**	**0.539**	**0.568**	**0.563**
4.0×	Real	SVM	0.555	0.507	0.517	0.533
	Real	NN	0.723	0.712	**0.737**	**0.723**
	Real+Synth.	SVM	0.562	0.517	0.527	0.563
	Real+Synth.	NN	**0.737**	**0.715**	0.727	**0.723**

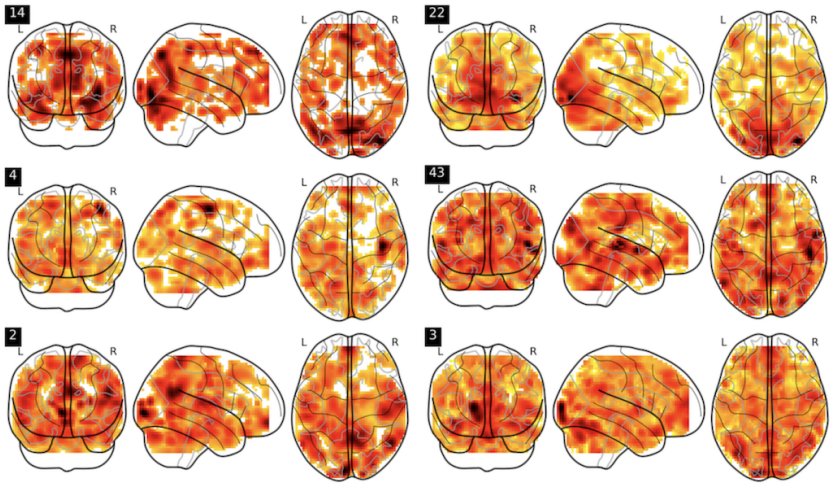


Figure 6: Brain images in collection 2138 generated by the ICW-GAN. The top-left tag is the class label associated with the generated brain image.

Table 6: Results on collection 503. As before, results show that augmenting real data with synthetic data improves classification performance.

Downsampling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
6.0×	Real	SVM	0.212	0.210	0.220	0.210
	Real	NN	0.293	0.287	0.302	0.293
	Real+Synth.	SVM	0.225	0.227	0.24	0.223
	Real+Synth.	NN	**0.299**	**0.298**	**0.321**	**0.298**
3.0×	Real	SVM	0.233	0.230	0.233	0.233
	Real	NN	0.358	**0.386**	0.390	0.358
	Real+Synth.	SVM	0.231	0.227	0.230	0.230
	Real+Synth.	NN	**0.373**	0.380	**0.410**	**0.372**

In this collection, we downsample brains by a factor of 8 and 4 to volume sizes of $13 \times 15 \times 11$ and $26 \times 31 \times 22$ respectively. Similar to the earlier reported results, we observe deep nets outperform SVMs, but more importantly, generated data during training was again suitable to improving the classifier performance on real data. 2D projections of several brain volumes generated by the ICW-GAN model are shown in Figure 6.The corresponding categories of the classes are as follows:

- **14:** visual form recognition, feature comparison, response selection, response execution, relational comparison, visual pattern recognition
- **22:** response execution, working memory, body maintenance, visual body recognition
- **4:** response selection, response execution, punishment processing
- **43:** motion detection
- **2:** response selection, response execution, animacy perception, animacy decision, motion detection
- **3:** response selection, response execution, motion detection

Dataset 3: Collection 503 Subjects in collection 503 are required to respond to 30 images from the International Affective Picture Set [Chang et al., 2015]. These 30 images were used to train the Picture Induced Negative Emotion Signature. Collection 503 contains 5067 brain images. Each of the 30 images is used to represent a task label. This collection requires additional preprocessing because brain volumes are available with two shapes: $79 \times 95 \times 68$ and $91 \times 109 \times 91$. Similarly, all experiments are at the two levels of resolution, $13 \times 15 \times 11$ and $26 \times 31 \times 22$. Classification results are summarized in Table 6 and again follow the trend reported earlier. Samples of the generated brain images are shown in Figure 7.

Appendix B Synthetic Power Analyses

B.1 Experiment Procedure

B.1.1 Neuroimaging and Conservative Power Analysis

We consider the case when our power analysis results computed between synthetic distributions are not identical to the power analysis results computed between real distributions. If our synthetic power analysis underestimates the sample size requirement (i.e. overestimates power), then our real study will be underpowered. Lower power means less opportunity to detect an effect and less confidence that the reported statistically significant effect is the truth. On the other hand, if our synthetic power analysis overestimates the number of samples required, we collect unnecessary samples but the study is guaranteed to have the desired power. Thus, not all errors are equally favorable. We often prefer our synthetic power analysis results to be conservative in their sample size recommendations. To probabilistically guarantee a conservative power analysis for the false discovery rate corrected T test, we add an adjustment β to our synthetic T test's p-value. The experiment procedure is as follows:

1. Train two ICW-GAN models to individually recover estimates $\hat{D_1}$ and $\hat{D_2}$ of the underlying data distributions D_1 and D_2.
2. For trial k from $1...K$:

(a) For sample size n from $1...N$:

i. Sample n bootstrap replicates $\mathcal{S}_1$, $\mathcal{S}_2$ from the real distributions D_1 and D_2, and $\hat{\mathcal{S}}_1$, $\hat{\mathcal{S}}_2$ from the synthetic distributions $\hat{D}_1$ and $\hat{D}_2$.

ii. Compute statistics for tests distinguishing between the real and synthetic bootstrap replicates and calculate the p-value as $p_{n,k}^{real}$ and $p_{n,k}^{syn}$.

iii. Compute a corrected p-value $p_{n,k}^{c-syn} = p_{n,k}^{syn} + \beta$ that is guaranteed to provide a conservative estimate of power.

(b) Estimate the power P_k^{real} of the test distinguishing between $\mathcal{S}_1$, $\mathcal{S}_2$, and P_k^{syn} and P_k^{c-syn} of the tests distinguishing between $\hat{\mathcal{S}}_1$, $\hat{\mathcal{S}}_2$ by $P_k = \frac{1}{N}\sum_n \mathbb{1}\{p_n < \alpha\}$.

We find the conservative adjustment β that ensures our synthetic power analyses are conservative for each cognitive process label in our dataset. In practice, we use the median β_{med} of all such β values and apply it to synthetic power analyses of cognitive process tag combinations that are not present in the original dataset. To estimate β_{med} in neuroimaging experiments, we perform the following procedure:

1. Train ICW-GAN on samples drawn from D to recover an estimated distribution $\hat{D}$ of D.
2. For every label z from $1...Z$ in the samples from D
   (a) For sample size n in $1...N$
      i. Draw n samples $\mathcal{S}_z$ from D with label z and $\mathcal{S}_{\neg z}$ from D without label z. Similarly draw $\hat{\mathcal{S}}_z$ and $\hat{\mathcal{S}}_{\neg z}$ from synthetic distribution $\hat{D}$.
      ii. Necessary β adjustment for data with label z and size n is

$$\beta_{n,z} = \min\left(\{\beta_0 | \frac{1}{K}\sum_{i=1}^{K}(\mathbb{1}\{p_{n,z,k}^{real} < \alpha\} - \mathbb{1}\{p_{n,z,k}^{syn} + \beta_0 < \alpha\}) \geq 0, \beta_0 \in [0, \alpha)\}\right)$$

      , where K is the number of trials and $p_{n,z,k}$ is the p-value for distinguishing $\mathcal{S}_z$, $\mathcal{S}_{\neg z}$ from D and $\hat{\mathcal{S}}_z$, $\hat{\mathcal{S}}_{\neg z}$ from $\hat{D}$ in the k-th trial.
   (b) Necessary β adjustment for data with label z is $\beta_z = \max_n \{\beta_{n,z}\}$.
3. Select $\beta_{med} = median(\{\beta_z | t \in [1, Z]\})$ as conservative adjustment.

To guarantee a conservative power analysis for the false discovery rate corrected t-test traditionally used in neuroscience, we compute an adjustment β_{med} that we add to our synthetic T test's p-value. We run traditional and synthetic power analyses over 16 individual cognitive process tags with 5 repeated trials each comparing 20,000 samples. We compute the median β_{med} of all β values calculated on real data and apply it as a conservative adjustment to the power analyses generated on future synthetic data tag combinations that may or may not have ever been collected before.

B.1.2 Univariate and Multivariate

In the univariate and multivariate experiments that follow, we employ a three layer fully-connected neural network architecture in both the generator and critic of an improved conditional wasserstein GAN. There are 64 neurons in the hidden layer with leaky relu activations after the input and hidden layers. The generator network includes dropout of 50% after the input and hidden layer. We train for 200,000 steps with batches of size 64, an input noise vector of length 64, and a gradient penalty λ of 1. Both the critic and generator networks are optimized with an Adam optimizer Kingma and Ba [2014] and the critic network is updated 5 times per step.

In our simulated experiments, we mimic a power analysis with both real and synthetic data by considering samples from the true underlying distributions D_1 and D_2. The experiment procedure is as follows:

1. Train two generative models to recover estimates $\hat{D}_1$ and $\hat{D}_2$ of the underlying data distributions D_1 and D_2.
2. For sample size n from $1...N$:
   (a) For trial k from $1...K$:

i. Sample bootstrap replicates of size n from the real distributions D_1 and D_2 and the synthetic distributions $\hat{D}_1$ and $\hat{D}_2$.

ii. Compute statistics for tests distinguishing between the real and synthetic bootstrap replicates and calculate the p-value as $p_{n,k}^{real}$ and $p_{n,k}^{syn}$.

(b) Estimate the power $P_{real}^{(n)}$ of the test distinguishing between D_1 and D_2 and $P_{syn}^{(n)}$ of the test distinguishing between $\hat{D}_1$ and $\hat{D}_2$ with sample size n by $P^{(n)} = \frac{1}{K}\sum_k \mathbb{1}\{p_{n,k} < \alpha\}$.

If at every sample size the power of a test computed between synthetic distributions is similar to the power of a test computed between real distributions, then using synthetic data as a stand-in for real data will yield a similar sample size estimate to what we would have achieved with real data.

B.2 Univariate

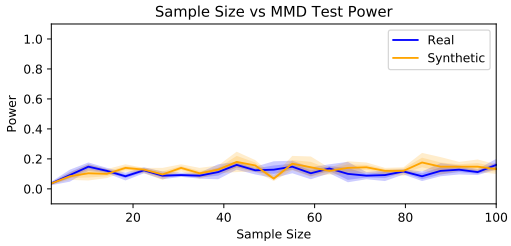


Figure 8: [NULL] $\chi^2(9)$ vs $\chi^2(9)$

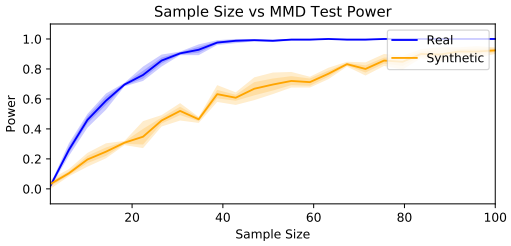


Figure 9: [ALTERNATIVE] $\chi^2(9)$ vs Exp(9)

Figure 8 summarizes the resulting MMD test power curves of 5 simulated traditional and synthetic power analyses comparing 20,000 samples from the $\chi^2(9)$ univariate distribution to the $\chi^2(9)$ univariate distribution under the null hypothesis. Note that the real and synthetic power curves stay close to 0% power as the number of samples grows, which is consistent with the null hypothesis that the two distributions have no difference.

Figure 9 summarizes the resulting MMD test power curves of 5 simulated traditional and synthetic power analyses comparing 20,000 samples from the $\chi^2(9)$ univariate distribution to the $Exp(9)$ univariate distribution. Note that both the real and synthetic power curves grow toward 100% as the number of samples grows, which is consistent with the alternative hypothesis that the two distributions are different.

B.3 Multivariate

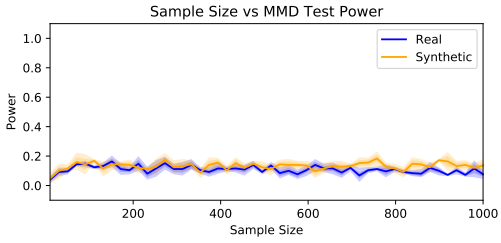


Figure 10: [NULL]
$\mathbb{N}_5(\mu = \mathbf{0}, \Sigma = \mathbb{I})$ VS $\mathbb{N}_5(\mu = \mathbf{0}, \Sigma = \mathbb{I})$

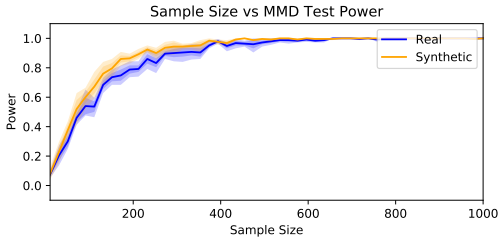


Figure 11: [ALTERNATIVE]
$\mathbb{N}_5(\mu = \mathbf{0}, \Sigma = \mathbb{I})$ VS $\mathbb{N}_5(\mu = \mathbf{1}, \Sigma = \mathbb{I})$

Figure 10 summarizes the resulting power curves of 5 simulated traditional and synthetic power analyses comparing 20,000 samples from the $\mathbb{N}_5(\mu = \mathbf{0}, \Sigma = \mathbb{I})$ multivariate distribution to the $\mathbb{N}_5(\mu = \mathbf{0}, \Sigma = \mathbb{I})$ multivariate distribution. Note that both the real and synthetic power curves stay close to 0% power as the number of samples grows, which is consistent with the null hypothesis that the two distributions have no difference.

Figure 11 summarizes the resulting power curves of 5 simulated traditional and synthetic power analyses comparing 20,000 samples from the $\mathbb{N}_5(\mu = \mathbf{0}, \Sigma = \mathbb{I})$ multivariate distribution to the $\mathbb{N}_5(\mu = \mathbf{1}, \Sigma = \mathbb{I})$ multivariate distribution. Note that both the real and synthetic power curves grow towards 100% power as the number of samples grows, which is consistent with the alternative hypothesis that the two distributions are different.