Optimal Design for Discovery Science: Applications in Neuroimaging

Shangsi Wang, Zhi Yang, Xi-Nian Zuo, Michael Milham, Cameron Craddock,
Consortium for Reliability and Reproducibility, Carey E. Priebe, Joshua T. Vogelstein

May 14, 2016

I Introduction

The data age is enabling us as a society to answer both age old questions and brand new ones. Fundamentally, to obtain obtain quantitative answers to any inquiry requires making two decisions: (i) "how should the data be collected?", and (ii) "how should the data be processed?" Optimally addressing experimental design decisions can yield dramatic improvements in both the financial and inferential costs (**?**). When the downstream inference task is known, a priori, the theory of experimental design tells us how to proceed (**?**). Recently, across industry, governmental, and academic settings, certain datasets become benchmark or reference datasets. Such datasets can then be used for a wide variety of different inferential problems. For example, the Sloan Digital Sky Survey (SDSS) data has been used to discover new quasars (**?**), and the human genome project data has revealed genetic insights into a many different diseases (**?**). Collecting and processing these datasets requires massive institutional investments, and choices related to questions (i) and (ii) above have dramatic effects on all subsequent analyses. Therefore, a theoretical framework to enable investigators to select from a set of possible design decisions using pilot data could reap great rewards.

The goal of such a framework would be to enable data set construction to be optimized for *any* subsequent downstream inference task, including all those currently not yet even conceived (**?**). This is not possible, because different inference tasks have different sources of information. Instead, however, we can show that one can choose from amongst a set of alternatives to maximize a lower bound on all subsequent inference tasks. Both recent and historical work in a discipline of study called "Generalizability Theory" has related goals. More specifically, it aims to analyze and quantify the various sources of error. Intra-class correlation (ICC) (**?**) and Kendall's tau (**?**) are parametric and non-parametric methods, respectively, for evaluating the "reliability" of a particular univariate statistic (such as magnitude or direction). More recently, I2C2 was proposed (**?**) as a multivariate extension to ICC. These methods, while quite powerful, also have certain limitations. For example, ICC and Kendell's tau only operate on univariate data. The motivating example for this work, however, is high-dimensional neuroimaging data. The high-dimensional generalization of ICC, I2C2, makes certain parametric assumptions, and therefore is inappropriate in certain settings of interest, such as sparsity. Moreover, we are not interested in quantifying the relative sources of error explicitly, rather, we have supervised learning problem; we desire to optimize reliability.

To this end, we have developed a novel formal definition of reliability, a non-parametric statistical property of a joint distribution in a hierarchical model, to reflect the trade-off across within and between entity variance. We prove that our notion of reliability (which may be more aptly called repeatability), provides a lower bound on predictive accuracy for any downstream inference task, even if we have never seen the covariates to predict. We then derive an estimator of reliability, and derive our estimator's asymptotic distribution, demonstrating that it is unbiased. Numerical simulations demonstrate the utility of our reliability estimator in a variety of settings. Finally, we apply our approach to choose amongst a set of choices one must make when designing a neuroimaging study to specifically study connectomics. We start by finding the maximally reliable threshold for converting correlation matrices (sometimes called functional connectomes), into binary edge lists (or graphs). Indeed, consistent with out theoretical and simulated results, maximizing the reliability of our datasets also maximizes performance on a suite of different downstream inference tasks. We then ask about a series of pre-processing steps: should one motion correct or not, and should one implement global signal regression or not, etc. We find the maximally reliable strategy amongst

a combinatorial set of options. Moreover, the pre-processing strategy that maximizes reliability also maximizes predictive accuracy on a suite of tests. Finally, we apply our method to determine which dataset is more reliable by comparing multiple different datasets. We conclude that a particular dataset is more reliable than the others, and that predictive accuracy from that dataset is improved relative to the others.

Thus, in total, our reliability analysis is a powerful tool for making decisions about how to collect and analyze datasets designed for discovery science. We expect this method to be useful in a wide variety of applications, and therefore have made all the code open source and available from http://openconnecto.me.

II Results

II.A Illustrative Simulation

II.B Problem Statement

Consider the following generative process. For each sample i, there exists some true physical property $\boldsymbol{v}_i$. In our running example throughout this article, $\boldsymbol{v}_i$ is the true connectome of subject i. Unfortunately, we do not get to directly observed $\boldsymbol{v}_i$, rather, we measure it with some device, that transforms the truth from $\boldsymbol{v}_i$ to $\boldsymbol{w}_i$ via $f_{\boldsymbol{\phi}}$. The parameter $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ characterizes all options in the measurement, including, for example, which scanner to use, which resolution, the number of images, sampling rate, etc. The output of $f_{\boldsymbol{\phi}}$ is the "raw" observation data $\boldsymbol{w}_i$, but it is corrupt in various ways, including movement or intensity artifacts introduced by the measurement process. Therefore, rather than operating directly on $\boldsymbol{w}_i$, we intentionally "pre-process" the data, in an effort to remove a number of nuisance variables. This pre-processing procedure further transforms the data from $\boldsymbol{w}_i$ to $\boldsymbol{x}_i$ via $g_{\boldsymbol{\psi}}$. The parameter $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ indexes all pre-processing options, including whether to perform motion correction, which motion correction, deconvolution, etc. More specifically, the entire code base, including dependencies, and even the hardware the pre-processing is running on, could count as $\boldsymbol{\psi}$. For brevity, we define $\boldsymbol{x}_i := g_{\boldsymbol{\psi}}\big(f_{\boldsymbol{\phi}}(\boldsymbol{v}_i)\big)$.

In addition to $\boldsymbol{v}_i$, there are other properties of sample i of interest; we call all of them $\boldsymbol{y}_i \in \mathcal{Y}$. These may include, for example, the phenotype of the subject, including personality tests, demographic information, and genetic data. The goal of experimental design, in this context, is to choose $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ to maximize some function of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

To quantify the performance of our choice, we introduce some assumptions. First, assume that each $(\boldsymbol{V}_i, \boldsymbol{Y}_i)$ pair is sampled independently and identically from some distribution, $(\boldsymbol{V}_i, \boldsymbol{Y}_i) \overset{iid}{\sim} F_{V,Y}$. For simplicity, let us assume that our goal is regression, using $\boldsymbol{X}_i$ as the *predictor* variables, and a single dimension of $\boldsymbol{Y}_i$ as the *target* variable (we will generalize these assumptions below). Moreover, let $h\colon \mathcal{X} \to \mathcal{Y}$ be the regression function (which we will typically learn from the data). The loss function, ℓ quantifies the error of our predictions, $\ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i) \in \mathbb{R}_+$. In the above described scenario, we can characterize our desiderata as minimizing the risk, or expected loss, under the above assumptions:

$$\underset{\boldsymbol{\phi}\in\boldsymbol{\Phi},\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\text{minimize}} \quad \mathbb{E}[\ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i)], \tag{1}$$

where the expectation is taken with respect to the joint distribution of X and Y, which is induced by the joint distribution of $(\boldsymbol{V}, \boldsymbol{Y})$, and transformed by $f(\cdot)$ and $g(\cdot)$; that is, $F_{X,Y} := F_{f(g(V),Y}$.

Problem (1) is an experimental design question. However, in the scenario of interest, $\boldsymbol{y}$ is not observed. Therefore, we cannot directly optimize Problem (1), and instead, we must consider some surrogate function.

II.C Proposed Solution

We propose to borrow an idea from generalizability theory (or G theory), in particular, by introducing "facets". A facet is a particular source of variance. For simplicity, consider the following illustrative scenario (we will generalize this idea below). In particular, assume that our goal is to optimize the pre-processing routine, encoded by $\boldsymbol{\psi}$. Let i denote the sample's unique *identity* (hereafter, referred to as the *subject*) and t denote the trial number. Thus, there is a single $\boldsymbol{v}_i$ for subject i, but we have $\boldsymbol{w}_{i,t}$, which is the t^{th} trial,

implicitly also a function of ϕ, which encodes all the details of the measurement. If both f and g together do not introduce too much noise, then we would expect that $\boldsymbol{x}_{i,1}$ and $\boldsymbol{x}_{i,2}$ are *closer* to one another than either are to any other subject's data, $\boldsymbol{x}_{i',t}$. Define δ to be a metric computing the distance between two data points, $\delta \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Formally, we expect that $\delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'}) \leq \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''})$, for any i, i', t, t', t''. For brevity, let $\delta_{i,t,t'} := \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i,t'})$ and $\delta_{i,i',t,t'} := \delta(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{i',t''}$. This intuition leads to our definition of reliability:

$$R(\boldsymbol{\psi}, \boldsymbol{\phi}, F) = \mathbb{P}[\delta_{i,t,t'} \leq \delta_{i,i',t,t'}] \tag{2}$$

Now, we can consider a surrogate objective function:

$$\underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}}{\text{minimize}} \quad \mathbb{E}[\ell(\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2})], \tag{3}$$

a useful definition of reliability (R)

Theorem 1. *reliability bounds predictive accuracy*

II.D Estimator/Test Statistic

- an estimator of R, called Rhat
- proof that our estimator is unbiased (in a model free setting), $E(Rhat) = R$.
- proof that our estimator asymptotically converges to truth (in a model free setting), $\hat{R}_n \to R$

II.E Some Simulations

II.F Some Real Data Experiments

III Discussion

Summary

Recommendations

- Prior to embarking on any new large experimental data collection and analysis, we recommend generating a pilot dataset amenable to reliability analysis.
- Upon designing a new experiment or analysis, we recommend reliability, as defined in Eq. (**??**), as the criterion upon which to optimize decisions.
- Because data analytics now have a large number of software dependencies, to maximize reliability of analytics workstreams, we recommend packaging all necessary software into portable deployable systems, such as vagrant
- For Human multimodal MRI connectomics, assuming the hardware system and scanner are sufficiently similar to those studies in this manuscript, we recommend using the reference pipelines
- Upon designing a new experiment or analysis, we recommend reliability, as defined in Eq. (**??**), as the criterion upon which to optimize decisions.
- Because data analytics now have a large number of software dependencies, to maximize reliability of analytics workstreams, we recommend packaging all necessary software into portable deployable systems, such as vagrant

Next Steps

A Introduction

In many problems arising in the data science, data preprocessing is the first step toward statistical inference. In this era, data usually comes with high dimensionality and in complicated forms which make directly working with raw data almost impossible. Therefore, necessary preprocessing must be performed to prepare the data ready for subsequent inference task. However, this crucial first step is sometimes done in an arbitrary or subjective fashion which lacks proper guidance and theoretical justification. In our current work, we investigate the relationship between data preprocessing and supervised learning. In particular, we propose that data preprocessing should be done to maximize reliability of processed data. We demonstate theoretically and with real data experiments that under supervised learning setting this approach produces data with small prediction error.

Reliability refers to the overall consistency of a measure. For example, if a subject is measured twice under the same conditions, two measures should be close to each other given the measure is reliable. The same spirit can be applied to data preprocessing. Good data preprocessing method should be reliable which means the processed data should have samples from the same subject close to each other. Actually, our analysis reveals that optimal data preprocessing is achieved when the reliability of processed data is maximized. In order to compare and select among preprocessing methods, we first need to be able to measure the reliability of processed data.

Many successful attempts have been made toward measuring reliability. Previous works include but not limit to Cohen's Kappa, Intraclass Correlation Coefficient (ICC), Bland Altman Test and Image intraclass correlation coefficient (I2C2). The key limitation the first three measure have is that they can only measure reliability of one-dimensional data, which makes them not appropriate to apply in multidimensional data preprocessing circumstances. I2C2 is designed to capture signal-to-noise ratio of high dimensional data. However, I2C2 assumes data has additive noise and lies in Euclidean space equiped with L_2 norm. These assumptions make it hard to generalize and apply for real data. To overcome the disadvantages of previous measure, we propose to use mean normalized rank (MNR) as a new measure of reliability.

Let us first introduce some notations and the definition of normalized rank. Suppose there are n subjects and each subject is measured s times. Denote O_{ij}, for $1 \leq i \leq n$, $1 \leq j \leq s$ the jth observation on subject i. The normalized rank R_{ijk}, $k \neq j$ counts the number of observations from other subjects are closer to observation O_{ij} compared to O_{ik}, and the count is then normalized by dividing the total number of observations from other subjects which is $(n-1)s$. That is,

$$R_{ijk} := \frac{\sum_{p=1, p\neq i}^{n} \sum_{q=1}^{s} I\{\|O_{ij} - O_{pq}\| < \|O_{ij} - O_{ik}\|\}}{(n-1)s}$$

The normalized rank R_{ijk} can be also understood as the rank of distance between O_{ij} and O_{ik}, among the distances between observations of other subjects and O_{ij}, but the rank is rescaled to make sure it lies between 0 and 1. The definition of R_{ijk} above can be easily generalized to metric space by replacing the norm $\|\cdot - \cdot\|$ with ditance $d(\cdot, \cdot)$. There are two things to notice. First, when computing the rank of $\|O_{ij} - O_{ik}\|$, the distances between other observations of the subject i and O_{ij} are excluded. Second, when obeservations are discrete and there are ties, it is recommended to add $0.5 \times I\{\|O_{ij} - O_{pq}\| = \|O_{ij} - O_{ik}\|\}$ to the nominator. Mean normalized rank (MNR) R of a data set is defined to be the avergare of all normalized ranks.

$$R := \frac{\sum_{i=1}^{n} \sum_{j=1}^{s} \sum_{k=1, k\neq j}^{s} R_{ijk}}{ns(s-1)}$$

Intuitively, in a reliable data set distances between observations of same subject are small compared to distances between observations from different subjects; therefore, reliable data set should have a small

MNR.

Our paper is structured as follows. In Section 2, we analyze properties of MNR theoretically. In particular, we demonstrate it bounds error probability of Bayes classifier. In Section 3, we demonstrate utility of MNR through simulation and real data expeiments. Reliability and Prediction error of two human connectome data sets are studied and compared. We conclude in section 4 with implications of our work and possible future extensions.

B Theoretical Analysis

II.A Reliability Measured by Mean Normalized Rank

As discussed in the previous section, we propose to use MNR as a measure of reliability. The next lemma is to show the intuition behind the definition of MNR. In particular, MNR can be understood as the probability that distance between two observations from different subjects are smaller than distances between two observations from one subject.

Lemma 1. *With the same notations above, assume all Os follow the same distribution then,*

$$E(R) = P(\|O_{ij} - O_{pq}\| < \|O_{ij} - O_{ik}\|)$$

Here, $p \neq i$ and $j \neq k$.

This lemma justifies validity of using MNR as a measure of reliability. For a reliable data set, samples from the same subject are close to each other compared to samples from other subjects. Then, $\|O_{ij} - O_{ik}\|$ is more likely to be smaller than $\|O_{ij} - O_{pq}\|$. As a consequence, MNR of the reliable data set should be small. The lemma above concerns the expectation of MNR. If we analyze the variance of MNR carefully, we can see the variance of MNR is less than $\frac{1}{n}$. This implies that MNR converges in probability to its expectation as the number samples goes to infinity.

Lemma 2. *As $n \to \infty$,*

$$R \xrightarrow{p} P(\|O_{ij} - O_{pq}\| < \|O_{ij} - O_{ik}\|)$$

Here, $p \neq i$ and $j \neq k$.

The two lemmaa above requires very few assumptions. There is no requirement on distribution of obervations. It also doesn't assume the noise is additive or data lies in Euclidean space. Furthermore, one may notice that n and s have no effect on the expectation of MNR. Therefore, MNR can also be applied to compare reliability across different data sets, in which the number of subjects and the number of oservations per subject may differ. In real data experiment section, this idea will be demonstrated by comparing reliability of two human connectome data sets through MNR. The next two subsections are going to study properties of data set with small MNR.

II.B Reliability and Signal-to-Noise Ratio

In this subsection, we investigate relationship between reliability and signal-to-noise ratio (SNR). Specifically, we show that SNR is lower bounded by a decreasing function of MNR. Consequently, preprocessing data to maximize reliability can be also recognized as to maximize a lower bound on SNR. Before presenting the result, we introduce the model and some notations.

As in classical regression analysis, we assume the observed sample O_{ij} is a sum of two variables.

$$O_{ij} = X_i + Z_{ij}$$

X_i denote the true value of measurment for subject i and Z_{ij} is the observational error or measurement error. Furthermore, X and Z are assumed to be independent of each other. From here to the end of this

section, we assume X and Z lie in Euclidean space equipped with L_2 norm. In addition, we assume

$$\begin{aligned} 0 &= E(Z_{ij}) \\ \mu &= E(X_i) \\ \sigma_1^2 &= Trace(Cov(Z_{ij}, Z_{ij})) \\ \sigma_2^2 &= Trace(Cov(X_i, X_i)) \\ E(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||^2) &\geq a^2\sigma_1^2 \\ \frac{E^2(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||^2)}{E(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||)^4} &\geq b \\ \lambda &= \frac{\sigma_2}{\sigma_1} \end{aligned}$$

Here, λ is defined to be the signal-to-noise ratio which measures the strength of true values to the strength of errors. Actually, Image Intraclass Correlation Coefficient (I2C2) is defined to be precisely $\frac{\lambda^2}{1+\lambda^2}$. MNR and I2C2 will be discussed in further detail in data experiment section. The next lemma demonstrates realtionship between MNR and SNR under additive noise setting.

Lemma 3.

$$\frac{\sigma_2}{\sigma_1} \geq \frac{a}{\sqrt{2}}(1 - (\frac{2-2D}{b})^{1/3})$$

Statistical inference based on large SNR data is usually easier. More technics can be applied to yield useful result. Furthermore, we are more likely to reach correct inference results and the results are more statistically significant. As a consequnce, by preprocessing to maximize reliability, processed data with large SNR should be optimal for miscellaneous inference tasks. For examples, these tasks can be classification, regression or clustering. There is no need to preprocess differently for different task. This idea will be illuminated in the data experiment section.

II.C Reliability and Classification Error

In this subsection, we focus on a specific inference task which is classification and the relationship between reliability and classification error is analyzed. In particular, theorem 1 shows that under suitable regularity conditions, Bayes error of classification is bounded by reliability measured by MNR. Let us introduce the classification setting and some notations.

Throughout this subsection, we assume $O_{ij} = X_i + \frac{1}{\lambda}Z_{ij}$, here Z and X does not depend on constant λ. Despite O_{ij}, another class label Y_i is observed for each subject i. In the classification setting, $Y_i \in \{0, 1\}$ and our inference task is to predict the value of class label Y_i based on features namely O_{ij}. It is well-known this task is best done by Bayes classifier g_B and denote $L_\lambda = P(g_B(O) \neq Y)$ the error rate of Bayes classifier. Furthermore, the following are assumed.

$$\begin{aligned} 0 &= E(Z_{ij}) \\ \pi_i &= P(Y = i) \\ \mu_i &= E(X|Y = i) \\ \Sigma_i &= \mathsf{Var}(X|Y = i) \\ \Sigma &= \mathsf{Var}(Z) \end{aligned}$$

A straight forward application of Mahalanobis distance yields the following lemma.

Lemma 4. *With the notations above,*

$$L_\lambda \leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1(\mu_0 - \mu_1)^T\Sigma_\lambda^{-1}(\mu_0 - \mu_1)}$$

Here, $\Sigma_\lambda = \pi_0\Sigma_0 + \pi_1\Sigma_1 + \frac{1}{\lambda^2}\Sigma$

It is important to notice that the right hand side is a decreasing function of λ. This can be seen by taking derivative of the denominator.

$$\begin{aligned} & \frac{d}{d\lambda}(\mu_0 - \mu_1)^T\Sigma_\lambda^{-1}(\mu_0 - \mu_1) \\ = \quad & \frac{2}{\lambda^3}(\mu_0 - \mu_1)^T\Sigma_\lambda^{-1}\Sigma\Sigma_\lambda^{-1}(\mu_0 - \mu_1) \\ \geq \quad & 0 \end{aligned}$$

The lemma shows that when noise is relatively small compared to true values, the Bayes error is also small. Lemma 3 and Lemma 4 can be combined to establish the relationship between MNR and Bayes error. We need one more assumption to make sure the definitions of λ in both lemmas are consistent, that is $Trace(\text{Var}(X)) = Trace(\text{Var}(Z))$. This can be quarnateed by properly scaling variable Z.

Theorem 2. *Let* $\lambda_* = \frac{a}{\sqrt{2}}(1 - (\frac{2-2D}{b})^{1/3})$*, under suitable regularity assumptions Bayes classification error* L *of* (O, Y) *satisfies,*

$$L \leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1(\mu_0 - \mu_1)^T\Sigma_{\lambda_*}^{-1}(\mu_0 - \mu_1)}$$

The theorem shows that maximizing reliability yields data with small Bayes error. Although Bayes error is usually not achievable in practice, it is clear that the performance of many widely used classifiers is closely related to Bayes error. If the Bayes error is small, classifiers include but not limit to k-NN, LDA and SVM tend to also have small classification error. If the primary inference task is classification, this theorem justifies maximizing reliability of processed data improves prediction accuracy.

C Simulated Data Experiment

III.A Simulation Experiment: Convergence of MNR

In Lemma 2, we prove that MNR converges to $P(\|O_{ij} - O_{pq}\| < \|O_{ij} - O_{ik}\|)$ as $n \to \infty$. If we define the reliability as the probability of distance between 2 observations of different subject less than 2 observations of the same subject, the Lemma 2 indicates that our sample approximated reliability MNR converges to the population reliability. In this experiment, we demonstrate this idea with simulated data. We generate data according to additive Guassian noise setting, which is $O_{ij} = X_i + Z_{ij}$ and $X_i \sim \text{N}(0, 1)$,$Z_{ij} \sim \text{N}(0, 1)$. 2 observations are generated for each one of n subject. For each fixed sample size, we repeated generate data and compute reliability 100 times. The figure below shows how the sample approximated reliability distributed, when we vary the sample size. Under this setting, the population reliability is 0.3850 which is computed through numerical integration. We can see clearly from the figure as sample size increases, sample approximated reliability converges to its mean.

III.B Simulation Experiment: Comparing MNR and I2C2

In this subsection, we use simulated data to demonstrate the utility of MNR. In addition, comparison is made between MNR and another reliability measure I2C2. I2C2 is another reliability measure introduced in paper (**). Under additive noise assumption, I2C2 is defined to be one minus the ratio of within subject variance over total variance. That is given $O_{ij} = X_i + Z_{ij}$,

$$I2C2 = 1 - \frac{\text{Trace}(\text{Cov}(Z_{ij}))}{\text{Trace}(\text{Cov}(O_{ij}))}$$

In practice, $\text{Cov}(Z_{ij})$ and $\text{Cov}(O_{ij})$ need to be estimated from data. Compared to I2C2, MNR has a few advantages. First, MNR is a more generalized and non-parametric measure of reliability. MNR can be easily extended to non-Euclidean metric and still be able to interpreted as a probability. Due to the fact that MNR rely on fewer assumptions and only estimate the probability, it turns out that relative standard error of

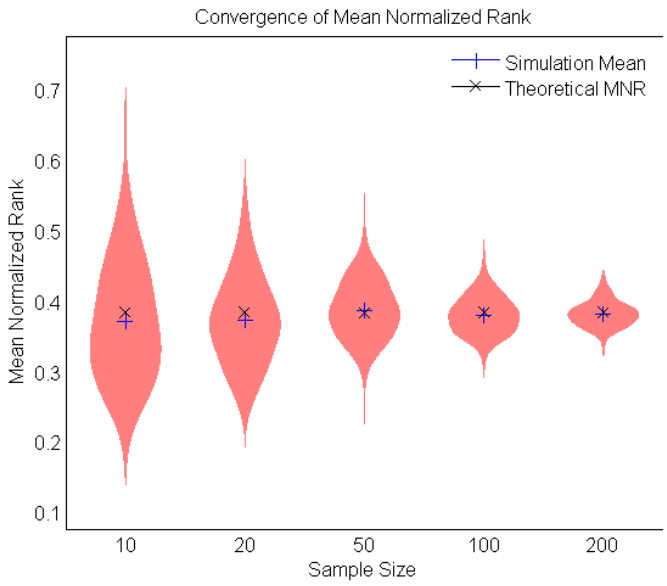

Figure 1: Simulation results

MNR estimate is usually less than I2C2. Furthermore, non-additive noise and non-continuous data cannot be properly handled by I2C2. Even in continuous and additive noise setting, I2C2 is only determined by first two moments of data and is not sensitive to higher moments or correlations between noises.

In the coming experiment, we consider four data generation scheme. We assume there are only two subjects under study, and 50 observations in $\mathbb{R}^2$ are generated for each subject. Then we compute ICC of the first dimension, ICC of the second dimension, I2C2 and MNR. For each data generation scheme, 100 repetitions are done. The figure 2 shows the result.

Scheme 1: The noise is additive independent multivariate normal distribution. That is $O_{ij} = X_i + Z_{ij}$, where $X_1 = [1, 1]$, $X_2 = [1, -1]$ and $Z_{ij} \sim$ MVN$(0, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix})$.

Scheme 2: The data is generately essentially the same as scheme 1, but rotated $\frac{\pi}{4}$. That is $O_{ij} = X_i + Z_{ij}$, where $X_1 = [\sqrt{2}, 0]$, $X_2 = [0, -\sqrt{2}]$, $Z_{ij} \sim$ MVN$(0, \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix})$.

Scheme 3: The noise is still additive but generated from Laplacian distribution. That is $O_{ij} = X_i + Z_{ij}$, where $X_1 = [1, 1]$, $X_2 = [1, -1]$. The first dimension of Z_{ij} is generated from Laplace$(0, \frac{1}{\sqrt{2}})$, and the second dimension is generated from Laplace$(0, 1)$. The parameters are chosen to make sure two subject conditional distributions has the same mean and covariance matrix as in scheme 1.

Scheme 4: The noise is multiplicative generated from log-normal distribution. That is $O_{ij} = X_i * Z_{ij}$, where $X_1 = [1, 1]$, $X_2 = [1, -1]$ and $Z_{ij} \sim$ LN$([\ln 2, \ln 3], \begin{pmatrix} \ln 2 & 0 \\ 0 & \ln 3 \end{pmatrix})$. The parameters are chosen to make sure two subject conditional distributions has the same mean and covariance matrix as in scheme 1.

In figure 1, the scatter plots at the top are observations from two subjects, and the box plots at the bottom are the corresponding reliability measures. First, it is not surprising that one dimensional reliability measure ICC does not capture reliability of higher dimensional data. Comaring scheme 1 and scheme 2, we expect that simple rotation does not affect reliability; however, ICCs of two dimensions change substantially. Sec-

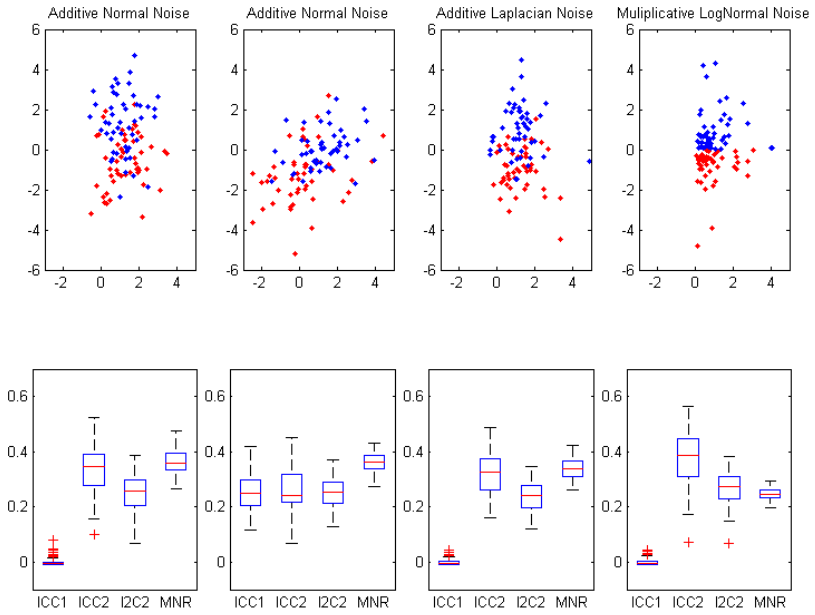

Figure 2: Simulation results

ond, we can see I2C2 does not fully reflect the reliability of data. In all four cases, I2C2 are roughly the same due to the fact we fix the first and second moments of data. However, we can see from the scatter plots that observations of scheme 3 is slightly more separated than observations of scheme 1, and observations of scheme 4 is mostly separated. At last, I2C2 clearly has larger estimation variance than MNR. This is partly due to the fact that I2C2 have more parameters to estimate. In order to compute I2C2, we must first estimate diagonal terms of two covariance matrices which leads to large variance in I2C2 estimate.

In this simple setting, we can quantify the degree of separation between two subjects by using Bayes error. Two subjects can be treated as two classes, and the Bayes errors of four schemes are 0.2395, 0.2395, 0.1839 and 0 respectively. It can be seen that MNR respects this ordering and reflects the degree of separation between subjects. If we are interested in predicting a phenotype, our intuition is that the error rate will more likely to be small, when there is more separation between subjects. We conclude here that MNR is a better reliability measure.

III.C Simulation Experiment: Maximizing Reliability

In this experiment, we demonstrate reliability can be used to select parameter in data preprocessing. As in the last section, we consider only 2 subjects, each with s observations in $\mathbb{R}^2$. Again, we consider additve noise setting. The means of two subjects are $X_1 = [1, 0]$ and $X_2 = [-1, 0]$ respectively. For the noise, we consider two cases. The first case is $Z_{ij} \sim \text{MVN}([0, 0], \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix})$ and the second case is $Z_{ij} \sim \text{MVN}([0, 0], \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix})$.

Now, assuming we want to linearly project the observations into 1 dimensional space. To achieve this, we sample possible lines to project uniformly from sphere, and compute reliability of projection. Then, we

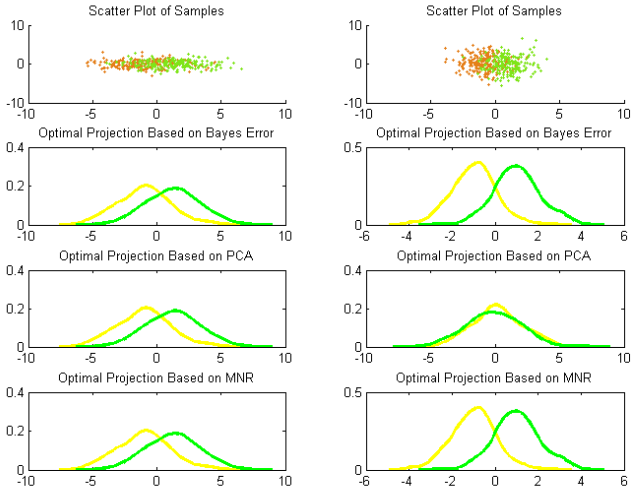

Figure 3: Simulation results

choose the projection which has maximum MNR. Also, PCA is performed for comparison. Because of the way that data is generated, we see the optimal linear projection should be projecting observations onto x-axis, and Bayes error is the same before and after this projection. The figure 3 shows the result. From the figure, we see both MNR PCA give the optimal projection in the first case. However, in the second case only MNR gives the optimal projection. In this case, the second dimension has larger variance than the first dimension; therefore, PCA choose y-axis to project to.

D Real Data Experiment

IV.A Human Connectome Data: NKI

The first real data set analyzed is a human connectome data set NKI. There are 176 subjects in the data set and each subject's brain is scanned twice ($n = 176$, $s = 2$). The original time series data is parcellized and the brain is divided into 131 and 165 regions. We first compute a correlation matrix which record correlations between different regions of the brain. The correlation matrix is then thresholded by t. Specifically, there is an edge between two regions if and only if the absolute correlation between two regions are greater than t. In this way, we construct a graph from each brain scan.

Despite brain scans, demographic information including age and gender of subjects is also collected. The inference task is to predict age and gender of a subject based on the graph derived from his or her brain scan. K-Nearest Neighbor rule is applied to predict both age and gender. In this experiment, we are interested in how to correctly select the threshold t and how the threshold affects inference results. Figure 4 and 5 summarize the results.

From figure 4 and 5, we can see the processed data is most reliable when the threshold t is around 0.16. Meanwhile, mean squared error of age prediction and classification error of gender prediction are also close to their minimums. Gender and age are two uncorrelated variables in the data set. Therefore, this experiment demonstrates that processed data of maximum reliability is optimal for two independent inference tasks.

Data Set	Mean Rank	S.E.
NKI131	0.0151	0.0034
NKI165	0.0215	0.0044
HCP25	0.1393	0.0030
HCP50	0.1077	0.0029
HCP100	0.0753	0.0024
HCP200	0.0843	0.0026
HCP300	0.0944	0.0027

Table 1: Reliability Comparison

IV.B Human Connectome Data: HCP

HCP is another human connectome data set we analyze. In the data set, 476 subjects are each scanned 4 times. Scans are then parcellized into 25, 50, 100, 200 and 300 regions. Again we compute the correlation matrix and threshold it to construct a graph for each brain scan. Age and gender are then predicted based on the constructed graph.

Figure 6-10 show similar results. The most reliable processed data has smallest classification error. However, mean square error of age prediction seems to just fluctuate randomly. This may due to the fact that age prediction is a much harder task and the K-NN is not good enough to capture the change of differentiability of processed data.

A natural question to ask next is that among these 7 connectome data sets which one is most reliable. The question can be answered by comparing MNR. For each data set, we first choose the threshold t which maximizes reliability and then compare the corresponding MNR across data set. A standard error estimate is also calculated by assuming normalized ranks are independent. The assumption leads to underestimate the standard error due to mild positive correlations between normalized ranks. To test whether two data sets are of same reliability, we can construct to confidence intervals of MNR and check whether they overlap or not.

Table 1 below summarizes the results. Two NKI data sets are remarkably more reliable compared to 5 HCP data sets. Reliability of NKI131 tends to be better than reliability of NKI165, but the difference is not statistical significant. Among 5 HCP data sets, HCP100 tends to be the most reliable one. By looking at HCP data sets alone, we discover another interesting fact that reliability demonstrates some kinds of bias-variance tradeoff. When the brain is partitioned into very few regions, the data set is not reliable possibly due to large bias. When the brain is partitioned into a large number of regions, the data set is also not reliable due to large variance.

E Conclusion

In the sections above, we demonstrate that maximizing reliability helps improving prediction accuracy. Our work provides guidance and direction to scientists on data preprocessing. Particularly, when facing decisions which are hard to make in data preprocessing, scientists should consider preprocessing method which maximizes reliability of processed data.

The current work can be extended in both theoretical and practical directions. From theoretical aspect, we only show Bayes error rate varies with reliability. It is also interesting to see the classification error of a specific classifier varies with reliability. Secondly, assumptions of theoretical analysis should be relaxed. Assumptions include additive noise, independence between noise and true values, and Euclidean distance maybe relaxed to better model real data. Thirdly, we only analyze the relationship between classification

and reliability. More inference tasks should be considered. These inference tasks can be but not limit to hypothesis testing, regression and clustering. Their relationship to reliability should be investigated.

From the aspect of practice, the whole data preprocessing procedure should be studied. In the experiments above, we focus on only one step of the data preprocessing. We may investigate reliability of data preprocessing step as a whole or even starting at data collection step. Secondly, more data experiments should be done in different setting to prove utility of reliability in data preprocessing. Data of more complicated form or with non-Euclidean distance can be investigated in a similar manner. The points mentioned in this and last paragraph can all extend our current work and improve its applicability. Nevertheless, our primary purpose is to demonstrate the relationship between reliability and predictability.

F Proof of Lemmas

Proof of Lemma 1. By definition of R,

$$E(R) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{s}\sum_{k=1,k\neq j}^{s} E(R_{ijk})}{ns(s-1)}$$

The expectation of R_{ijk} can be simplified,

$$\begin{aligned} & E(R_{ijk}) \\ = & \frac{\sum_{p=1,p\neq i}^{n}\sum_{q=1}^{s} E(I\{\|O_{ij}-O_{pq}\| < \|O_{ij}-O_{ik}\|\})}{(n-1)s} \\ = & \frac{\sum_{p=1,p\neq i}^{n}\sum_{q=1}^{s} P(\|O_{ij}-O_{pq}\| < \|O_{ij}-O_{ik}\|)}{(n-1)s} \\ = & P(\|O_{ij}-O_{pq}\| < \|O_{ij}-O_{ik}\|) \text{ for any } p \neq i \end{aligned}$$

The last equality is due to the fact that Os have the same distribution. Notice the symmetry of R_{ijk} in i, j and k we have,

$$E(R) = E(R_{ijk}) = P(\|O_{ij}-O_{pq}\| < \|O_{ij}-O_{ik}\|)$$

□

Proof of Lemma 2. By definition of R,

$$\begin{aligned} R &= \frac{\sum_{i,j,k,k\neq j} R_{ijk}}{ns(s-1)} \\ &= \frac{\sum_{i,j,k,k\neq j,p,p\neq i,q} I_{ijkpq}}{ns(s-1)(n-1)s} \end{aligned}$$

Here, $I_{ijkpq} := I\{\|O_{ij}-O_{pq}\| < \|O_{ij}-O_{ik}\|\}$. We show in the previous lemma that $E(R) = P(\|O_{ij}-O_{pq}\| < \|O_{ij}-O_{ik}\|)$. To demonstrate that R converges to its epxectation in probability, it is suffice to show that $\mathrm{Var}(R) \to 0$. Since then, by Chebyshev's inequality,

$$P(|R-E(R)| \geq \epsilon) \leq \frac{\mathrm{Var}(R)}{\epsilon^2} \to 0$$

If we expand the variance of R,

$$\mathsf{Var}(R) = \frac{\sum\limits_{i,j,k,p,q} \sum\limits_{i',j',k',p',q'} \mathsf{Cov}(I_{ijkpq}, I_{i'j'k'p'q'})}{(ns(s-1)(n-1)s)^2}$$

There are $(ns(s-1)(n-1)s)^2$ terms in the sum at the nominator; however, some of them are actually 0. I_{ijkpq} is a function of O_{ij}, O_{ik} and O_{pq}; therefore, is independent of any observations of subjects other than i and p. This implies I_{ijkpq} is independent of $I_{i'j'k'p'q'}$ as long as $\{i,p\} \cap \{i',p'\} = \emptyset$. As a consqeunce, there are $(4n-6)(s(s-1)s)2 = ns(s-1)(n-1)s - (n-2)s(s-1)(n-3)s$ combinations of $i'j'k'p'q'$ such that covariance between I_{ijkpq} and $I_{i'j'k'p'q'}$ maybe non-zero. Furthermore, the covariance must be less $\frac{1}{4}$ due to the fact that they are indicator random variables. Therefore, we have

$$\begin{aligned}
\mathsf{Var}(R) &= \frac{\sum\limits_{i,j,k,p,q} \sum\limits_{i',j',k',p',q'} \mathsf{Cov}(I_{ijkpq}, I_{i'j'k'p'q'})}{(ns(s-1)(n-1)s)^2} \\
&\leq \frac{\sum\limits_{i,j,k,p,q} (4n-6)(s(s-1)s)}{4(ns(s-1)(n-1)s)^2} \\
&= \frac{(ns(s-1)(n-1)s)(4n-6)(s(s-1)s)}{4(ns(s-1)(n-1)s)^2} \\
&= \frac{(4n-6)}{4n(n-1)} \\
&\leq \frac{1}{n} \\
&\to 0 \text{ , as } n \to \infty
\end{aligned}$$

As discussed above, this concludes the proof for the lemma. □

Proof of Lemma 3. By Lemma 1, we only need to bound $P(\|O_{ij} - O_{ik}\| < \|O_{ij} - O_{pq}\|)$. Under additive noise setting,

$$\begin{aligned}
& P(\|O_{ij} - O_{ik}\| < \|O_{ij} - O_{pq}\|) \\
= \; & P(\|Z_{ij} - Z_{ik}\| < \|X_i + Z_{ij} - X_p - Z_{pq}\|) \\
\leq \; & P(\|Z_{ij} - Z_{ik}\| < \|X_i - X_p\| + \|Z_{ij} - Z_{pq}\|) \\
= \; & P(\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\| < \|X_i - X_p\|) \\
= \; & \frac{1}{2}P(\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\| < \|X_i - X_p\| | \|Z_{ij} - Z_{ik}\| < \|Z_{ij} - Z_{pq}\|) + \\
& \frac{1}{2}P(\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\| < \|X_i - X_p\| | \|Z_{ij} - Z_{ik}\| > \|Z_{ij} - Z_{pq}\|) \\
= \; & \frac{1}{2} + \frac{1}{2}P(\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\| < \|X_i - X_p\| | \|Z_{ij} - Z_{ik}\| > \|Z_{ij} - Z_{pq}\|) \\
= \; & \frac{1}{2} + \frac{1}{2}P(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\|| < \|X_i - X_p\|) \\
= \; & 1 - \frac{1}{2}P(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\|| > \|X_i - X_p\|)
\end{aligned}$$

To bound the probability above, we bound the two terms on different sides of equation separately. We start with the term at the left side and compute its expectation.

$$\begin{aligned}
& E(\|X_i - X_p\|^2) \\
= \; & E(X_i^T X_i + X_p^T X_p - 2X_i^T X_p) \\
= \; & 2\sigma_2^2
\end{aligned}$$

We then use Markov's Inequality to this term,

$$P(\|X_i - X_p\| < t) \geq 1 - \frac{2\sigma_2^2}{t^2}$$

Assume,

$$E(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||^2) \geq a^2\sigma_1^2$$
$$\frac{E^2(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||^2)}{E(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||)^4} \geq b$$

Then, we apply Paley-Zygmund Inequality,

$$P(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||^2 > t^2) \geq b(1 - \frac{t^2}{a^2\sigma_1^2})^2$$

Understand the fact that Xs and Zs are independent, we can combine the two inequalities and get a bound on $P(\|O_{ij} - O_{ik}\| < \|O_{ij} - O_{pq}\|)$.

$$\begin{aligned}
& P(\|O_{ij} - O_{ik}\| < \|O_{ij} - O_{pq}\|) \\
\leq\ & 1 - \frac{1}{2}P(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\|| > \|X_i - X_p\|) \\
\leq\ & 1 - \frac{1}{2}P(|\|Z_{ij} - Z_{ik}\| - \|Z_{ij} - Z_{pq}\||^2 > t^2)P(\|X_i - X_p\|^2 < t^2) \\
\leq\ & 1 - \frac{1}{2}b(1 - \frac{t^2}{a^2\sigma_1^2})^2(1 - \frac{2\sigma_2^2}{t^2})
\end{aligned}$$

Assume $a^2\sigma_1^2 \geq 2\sigma_2^2$ and set $k = \sqrt{2}a\sigma_1\sigma_2$,

$$P(\|O_{ij} - O_{ik}\| < \|O_{ij} - O_{pq}\|) \leq 1 - \frac{1}{2}b(1 - \frac{\sqrt{2}\sigma_2}{a\sigma_1})^3$$

Denote $P(\|O_{ij} - O_{ik}\| < \|O_{ij} - O_{pq}\|)$ by D, we can solve for and lower bound $\frac{\sigma_2}{\sigma_1}$.

$$\frac{\sigma_2}{\sigma_1} \geq \frac{a}{\sqrt{2}}(1 - (\frac{2 - 2D}{b})^{1/3})$$

□

Proof of Lemma 4. To compute Mahalanobis distance, we first compute weighted covariance matrix.

$$\begin{aligned}
\mathsf{Var}(O|Y = i) &= \mathsf{Var}(X|Y = i) + \mathsf{Var}(Z|Y = i) \\
&= \Sigma_i + \frac{1}{\lambda^2}\Sigma
\end{aligned}$$

Therefore, the weighted covariance matrix of Os

$$\begin{aligned}
\Sigma_\lambda &= \pi_0\mathsf{Var}(O|Y = 0) + \pi_1\mathsf{Var}(O|Y = 1) \\
&= \pi_0\Sigma_0 + \pi_1\Sigma_1 + \frac{1}{\lambda^2}\Sigma
\end{aligned}$$

Then, apply Devijver and Kittler's result,

$$L_\lambda \leq \frac{2\pi_0\pi_1}{1 + \pi_0\pi_1(\mu_0 - \mu_1)^T\Sigma_\lambda^{-1}(\mu_0 - \mu_1)}$$

□

G Bibliography