Unpaired Data Empowers Association Tests

Mingming Gong[1,2], Peng Liu[1], Frank C. Sciurba[1], Petar Stojanov[2], Dacheng Tao[3], George C. Tseng[1], Kun Zhang[2], and Kayhan Batmanghelich[*1]

[1]University of Pittsburgh
[2]Carnegie Mellon University
[3]University of Sydney

Abstract

To achieve a holistic view of the underlying mechanisms of human diseases, the biomedical research community is moving toward harvesting retrospective data available in Electronic Healthcare Records (EHRs). The first step for causal understanding is to perform association tests between types of potentially high-dimensional biomedical data, such as genetic, blood biomarkers, and imaging data. To obtain a reasonable power, current methods require a substantial sample size of individuals with both data modalities. This prevents researchers from using much larger EHR samples that include individuals with at least one data type, limits the power of the association test, and may result in higher false discovery rate. We present a new method called the Semi-paired Association Test (SAT) that makes use of both paired and unpaired data. In contrast to classical approaches, incorporating unpaired data allows SAT to produce better control of false discovery and, under some conditions, improve the association test power. We study the properties of SAT theoretically and empirically, through simulations and application to real studies in the context of Chronic Obstructive Pulmonary Disease. Our method identifies an association between the high-dimensional characterization of Computed Tomography (CT) chest images and blood biomarkers as well as the expression of dozens of genes involved in the immune system.

Increasingly, data from Electronic Health Records (EHRs) in hospitals are becoming available to clinical researchers. Such massive collections contain various types of data from sources such as high-resolution imaging, genome sequencing, and physiological metrics. By studying such a large and diverse data, researchers can provide a holistic view of the underlying mechanisms of human diseases. For example, while a large proportion of human diseases are influenced by genetic variants [1–3], their mechanisms are not well understood [4–6]. To understand the mechanism, measuring other variables such as gene expression is required. Unfortunately, it is unlikely that all patients in the EHR have all measurement modalities. For example, due to the high cost of image acquisition and specimen maintenance, hospitals order those only when they are needed. Consequently, only the record of a few patients contains all data modalities, which reduces the power of association tests and increases the chance of false discovery.

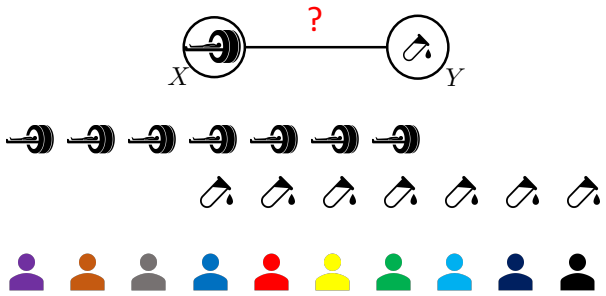

Figure 1: X and Y represent two modalities. Current approaches only use paired data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$. Assuming that the total number of samples of X (M) and Y (N) is more than the paired data, we aim to find out how the control of the false discovery and the power of association tests can be improved by the unpaired data $\{\mathbf{x}_i\}_{i=n+1}^{M}$ and $\{\mathbf{y}_i\}_{i=n+1}^{N}$.

Furthermore, a multidimensional phenotype can offer better sensitivity to the clinical and genetic underpinning of human diseases than a one-dimensional scalar phenotype [7–9]. For instance, high-dimensional features can be computed to summarize the folding pattern of the brain structure in Magnetic Resonance (MR) imaging [10], or the texture and distribution of the lung tissue destruction can be measured and summarized by Computed Tomography (CT) imaging [11, 12]. Those metrics are highly predictive of the diseases (e.g., Alzheimer's disease [9, 13] and bipolar disorder [14, 15] for MR, and COPD [12] for CT). Relating that high-dimensional phenotype to genetic and genomic measurements provides more evidence for understanding the etiology of the disease.

In this paper, we present a new method to formally test the association between two types of potentially high-dimensional data that allows incorporating unpaired samples, i.e., samples with one data modality (see Fig. 1 for a schematic illustration). Our approach provides better control of false-positive and, under some mild assumptions, increases the statistical power of the test. Unpaired data enables us to better estimate the null distribution, which results in more accurate control of the false positive rate. Furthermore, it allows us to leverage the underlying structure of the high-dimensional measurements, which consequently increases the power of the test. The proposed method, the Semi-paired Association Test (SAT), falls in the kernel machine framework [16–25]. More specifically, two variants of our method generalize the Variance Component Score Test (VCST) [16–20] and the Kernel Independent Test (KIT) [21–25] such that they can exploit unpaired data. The VCST is commonly used to test for heritability of a phenotype [16–20] and is implemented in popular software such as GCTA [26]. The KIT is widely used for statistical independence test in various scenarios [22, 25, 27, 28]. We provide a connection between those methods. Our proposed test makes unpaired data, previously wasted, available for discovering novel associations in massive uncontrolled datasets such as EHRs. Unearthing unnoticed associations assists in understanding the underlying mechanism of human diseases.

This paper makes two contributions. First, it provides a statistically grounded method for the inclusion of unpaired data. The extensive simulation, as well as theoretical study, supports the hypothesis that the unpaired data is beneficial to control the false discovery and if the conditions are satisfied, can improve the power. Second, we apply our method to two different real studies. In the first experiment, we show that unpaired data can discover a new association between the

high-dimensional radiographic measurements of Chronic Obstructive Pulmonary Disease (COPD) and peripheral blood biomarkers that play a role in the immune system. In this dataset, only a subset of the cohort has blood samples. In the second experiment, we apply our approach to genotype-phenotype data from the General Population Cohort from Uganda [29]. In this dataset, all subjects have genotype data but only one-fourth of them have phenotypes. Our method is able to find more heritable phenotypes.

Results

Theoretical Framework

We propose a method to test the association between two potentially high dimensional datasets. In addition to paired data, our method is able to exploit unpaired data, meaning the data from subjects that have only one modality. One can view the distance between the joint distribution and the product of the marginals as the strength of the association. To test association formally, the null distribution for the distance should be estimated. Our first theorem (Theorem 1) shows that the null distribution can be estimated more accurately using unpaired data. Hence, our method results in better control of the type I error. In addition to type I error, we show, in Theorem 2, that power can be improved if the data (of at least one modality) live on a lower dimensional space. Such an assumption is mostly the case for real data. For example, previous studies have shown that the space of Magnetic Resonance images of the brain can be modeled by a relatively low-dimensional manifold [30–32]. A similar assumption has been explored to model the low-dimensional space of gene expression for single-cell expression analysis [33, 34]. Unpaired data help us to estimate the low-dimensional space more accurately. Our new test statistic, which is a random variable calculated from sample data, exploits the low-dimensional assumption by taking advantage of the unpaired data.

Our method, the Semi-paired Association Test (SAT), generalizes two popular methods for association testing, the Variance Component Score Test (VCST) and the Kernel Independent Test (KIT), which are commonly used in statistical genetics to test the heritability of traits [17, 20] or gene-level associations [19, 28, 35]. More specifically, a variant of our method, SAT-fx, generalizes the VCST, which assumes that one of the modalities is not a random variable (i.e., fixed). For example in heritability analysis, the effect size is random but the genotype is given. The second variant, SAT-rx, generalizes KIT, which assumes that both modalities are random variables.

Simulation Results

To evaluate our method's improvement of type I and type II errors, we mimic the data missingness mechanism by conducting two levels of simulations:

i We synthesize both X and Y. In this simulation, we evaluate both variants of our method, including SAT-fx and SAT-rx.

ii Following the literature of population genetics in which testing for the heritability of traits is a topic of interest, we use genotype data as X and synthesize Y. We only evaluate SAT-fx because the genotype data is fixed.

In simulation (i), to generate X, we first generate N low-dimensional (dim=10) data points from a Gaussian distribution and then map them to high-dimensional X using a linear transformation

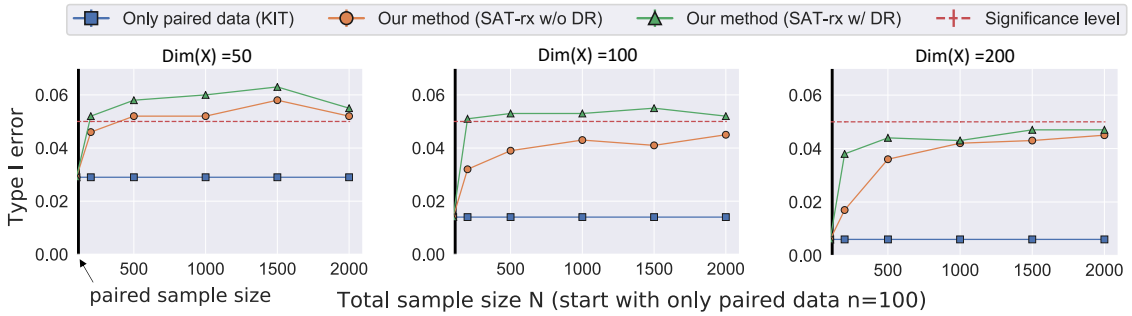

Figure 2: Evaluation of SAT-rx type I error rate control on the simulated data generated by procedure (i) in the random X setting. The blue line (KIT) is the result of using only paired data; hence it does not change with addition of unpaired data. KIT only uses the $n = 100$ paired data points. Our methods (green and orange) start with n pairs and gradually adds unpaired data to improve type I error control. False-positive rates for both variants of our method SAT-rx are well controlled around the nominal value (DR:Dimension Reduction).

plus independent Gaussian noise. To generate Y, we first generate low-dimensional data according to the variance components model (see Eq. 1 in the Method section) and then map them to high-dimensional Y using another linear transformation plus independent Gaussian noise.

In simulation (ii), we use the real genotype data from the COPDGene cohort as X. COPDGene is a multi-centered study of the genetic epidemiology of Chronic Obstructive Pulmonary Disease (COPD) that enrolled individuals aged 45-80 years with at least a 10 pack-year history of smoking [36]. We generate Y using the same procedure used in simulation (i).

In all the simulations, we create 1000 simulation replicates to evaluate the type I error rate and test power. Type I error rates and powers are calculated using the percentage of p-values smaller than a given significance level ($\alpha = 0.05$) under null models and alternative models, respectively. We set the heritability $h^2 = 0$ for the evaluation of type I error rates and $h^2 = 0.1$ for the evaluation of power. To show the benefits of incorporating unpaired data, we compare the type I errors of the VCST/KIT as a baseline with two variants of both SAT-fx and SAT-rx: with and without Dimensionality Reduction (DR). VCST and KIT only use n paired data points, while SAT-fx and SAT-rx use n paired data points together with an additional $N - n$ unpaired data points. For evaluation, we have access to the oracle where we can apply VCST and KIT using all N data points as paired, which is the best we can achieve. We set $n = 100$ for simulation (i) and $n = 3000$ for simulation (ii).

Fig. 2 and Fig. 3 report the type I error rates and power in simulation (i), respectively. Here we only show results for random X. The results for fixed X have similar trends and are available in Section S1 of the Supplementary Information. The results in Fig. 2 demonstrate that the type I error rates of our proposed method approach the predefined significance level (0.05) as we add more unpaired data. In addition, Fig. 3 shows that our method's test power increases when adding unpaired data. Though our method has lower power than the oracle method which has access to all the paired data, it consistently outperforms the baseline KIT method that uses only paired data.

Fig. 4 and Fig. 5 report the type I error rates and powers of all the methods evaluated in simulation (ii). Again, we can see from Fig. 4 that the type I error rates of our proposed

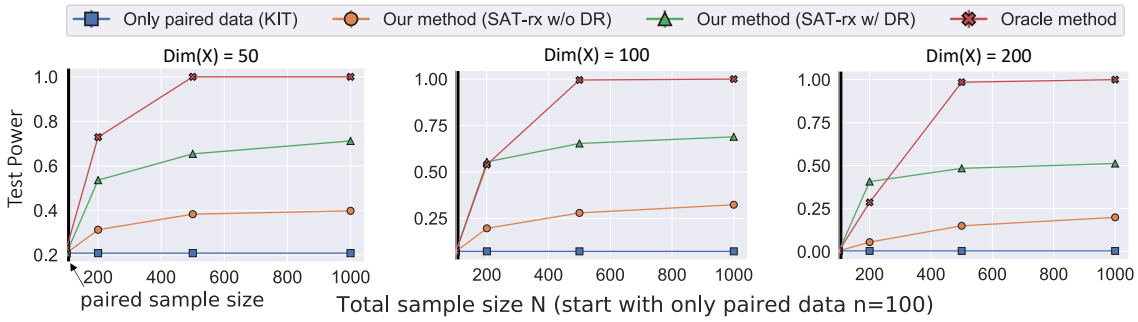

Figure 3: Evaluation of SAT-rx test power on the simulated data generated by procedure (i) in the random X setting (DR:Dimension Reduction). The results for heritability values $h^2 = 0.1$ and dimensionality $dim(X) = dim(Y) = 50, 100, 200$ are shown. KIT only uses the $n = 100$ paired data points. Our methods start with n pairs and gradually add unpaired data to improve test power.

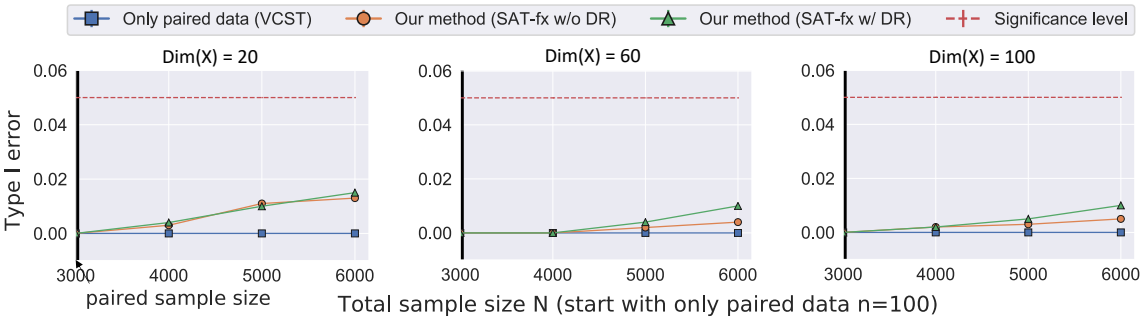

Figure 4: Evaluation of SAT-fx type I error rate control on the data generated in simulation (ii). VCST only uses the $n = 3000$ paired data points. Our method SAT-fx starts with n pairs and gradually adds unpaired data to improve type I error control.

methods approach the significance level (0.05) as we add more unpaired data. However, because the dimensionality of the genotype is very high, the test is still very conservative even after adding unpaired data. Nevertheless, our method's power exceeds that of VCST and increases as we add unpaired data.

COPD: Imaging Data and Peripheral Blood Biomarkers

In this experiment, we investigate whether the high-dimensional radiographical measurement from Computed Tomography (CT) imaging is associated with peripheral blood biomarker signature of emphysema. COPD is a highly heterogeneous disease and involves many subprocesses, including emphysema [37]. CT imaging is increasingly used for emphysema diagnosis because it directly characterizes anatomical variation introduced by the disease [38]. Currently, Low Attenuation Area (LAA) is used to quantify the emphysema [39, 40]. However, LAA is based on a single intensity threshold value and cannot characterize variation in the texture of the lung parenchyma due to

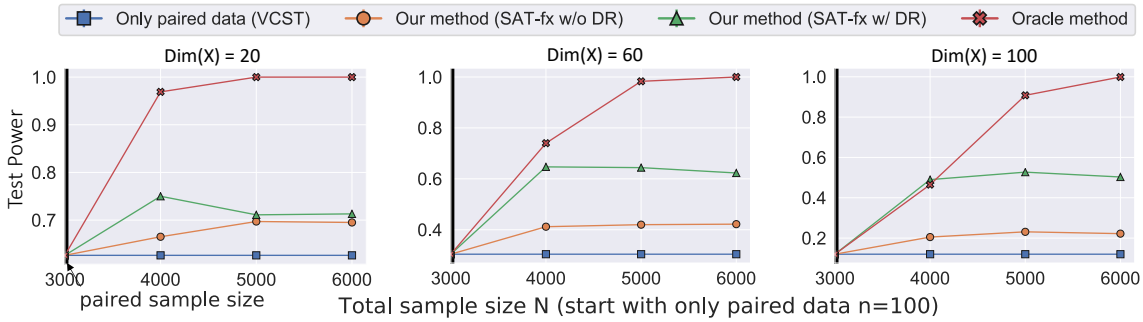

Figure 5: Evaluation of SAT-fx test power on the data generated in simulation (ii). VCST only uses the $n = 3000$ paired data points. Our method SAT-fx starts with n pairs and gradually adds unpaired data to improve test power.

different disease subtypes [41]. Over the past year, researchers have proposed various generic and specific local image descriptors that extract higher order statistical features from CT images [11, 42, 43]. However, it is not clear whether such high-dimensional measurements are considered phenotypes, and whether the relationship to the causal biological processes is maintained.

We test the association between one of these multidimensional phenotypes and peripheral blood biomarkers. We use the method proposed by Schabdach *et al.* [11] that computes the similarity between 4629 patients and associates a 100-dimensional vector to each patient (see Supplementary Section S5 for details). Only 377 patients have both the blood biomarker and imaging data. We correct for the effects of covariates including age, sex, BMI (body mass index), and pack-year smoking history. Fig. 6 (a) reports the $-\log_{10}$(p-value) of different methods with respect to size of the unpaired imaging data. The results show that our method takes advantage of unpaired data and detects an association between high-dimensional imaging phenotypes and blood biomarkers that was not detected by the baseline method using only paired data.

COPD: Imaging Data and Peripheral Blood Genes

Although smoking is a major risk factor for COPD, not all smokers develop debilitating disease, which suggests that COPD is a systemic disease and other factors might be involved in its development. Bahr *et al.* identified a set of genes whose expression is associated with two measurements used to diagnose COPD: percent predicted Forced Expiratory Volume in one second (FEV1) and the ratio of FEV1 to forced vital capacity (FEV1/FVC) [44]. These genes in Peripheral Blood Mononuclear Cells (PBMC) play a role in the immune system, inflammatory responses, and sphingolipid metabolism. Similar to the previous experiment, we investigate whether the multidimensional imaging phenotype is associated with systemic measurements. In this dataset, 90 subjects have both phenotype and gene expression measurements while more than 4539 subjects only have imaging phenotypes. We use the same covariates as the previous experiment. Fig. 6 (b) shows that our method exploits the unpaired data and results in lower p-values, suggesting an association between the imaging phenotypes and PBMC gene expression (p-value < 0.05) while the p-values of the baseline method using only paired data fails to pass the significance level.

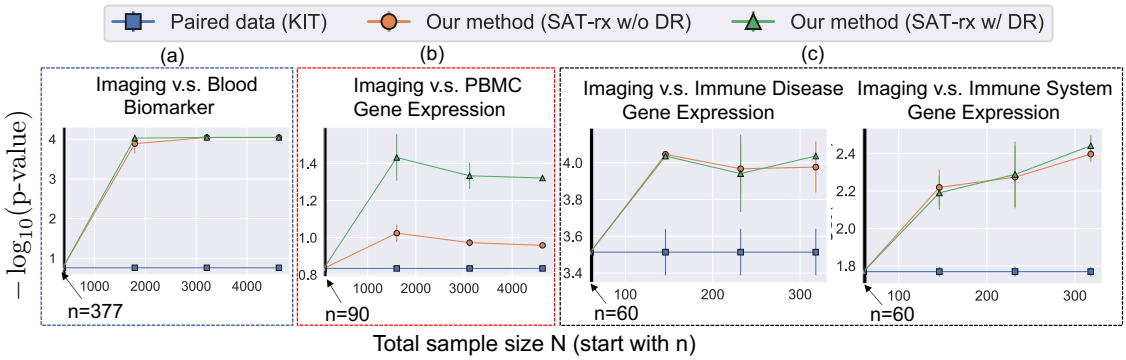

Figure 6: Experiments on three real imaging and genetics datasets. (a) Test an association between multidimensional imaging features and plasma biomarkers. (b) Test an association between imaging features and peripheral blood mononuclear cell gene expression data. (c) Test an association between imaging features and gene expression of genes in immune system pathway of the disease. In all the experiments, we start with n paired data points and show the behavior of our methods when adding unpaired data, with and without dimensionality reduction (DR).

COPD: Imaging Data and Immune System Gene Expression

In this experiment, we apply our method again in the context of COPD but on a different dataset. We investigate the hypothesis that anatomical changes manifested on images are related to auto immune pathways. More specifically, we chose the "immune disease" and "immune system" gene pathways in the KEGG database [45]. We apply our method to imaging phenotypes and gene expression data containing 319 subjects from several sources (gene expression data from the GEO repository, imaging and clinical information from the Lung Genomics Research Consortium) [46]. Because only 60 patients have imaging phenotypes, we have a number of unpaired gene expression data. We compare our method with the baseline method that does not use the unpaired gene data and the results are shown in Fig. 6 (c). We can see that our method finds more significant associations as we add more unpaired data.

Heritable Phenotype Discovery

In this section, we use the General Population Cohort (CPC), Uganda [29], to establish genotype-phenotype associations in Genome-Wide Association Studies (GWAS), and show that our method can benefit from unpaired data.

GWAS have discovered many genetic risk variants of common diseases [2, 3]. Before performing GWAS, one should test the hypothesis that a given phenotype is "heritable" or not. Given the observation of a phenotype in a population of subjects, so-called narrow sense heritability is defined as an additive genetic portion of the phenotypic variance [47, 48]. A linear mixed model (LMM), which is a form of multivariate regression, is used to estimate the heritability (h^2). Testing for the null hypothesis of $\mathcal{H}_0 : h^2 = 0$ can be done using VCST and the power of the test is affected by the sample size.

We apply our method to study the heritability of a set of phenotypes from the General Population Cohort (GPC), Uganda. More specifically, it contains 37 phenotypes, including anthropometric

Table 1: P-values on Uganda General Population Cohort. The newly found associations by our method at the significance level 0.05 are marked as bold. Since we mimic the missingness for phenotypes in the top part of the table, we are able to compare our performance with the oracle. In the bottom part of the table, a subset of the subjects has a missing phenotype; hence, the oracle columns are empty.

	KIT		SAT-rx (w/o DR)		SAT-rx		Oracle	
	p-value	p-value (Bonf)	p-value	p-value (Bonf)	p-value	p-value (Bonf)	p-value	p-value (Bonf)
SBP	0.293	1.000	0.224	1.000	0.128	0.897	0.010	0.195
DBP	0.091	0.928	**0.031**	0.537	**7.25e-03**	0.138	< 1.00e-05	< 1.90e-04
BMI	0.101	0.907	**0.035**	0.528	**0.011**	0.214	< 1.00e-05	< 1.90e-04
WHR	0.249	1.000	0.171	0.901	0.119	0.810	0.033	0.630
Weight	0.057	0.819	**0.012**	0.235	**1.63e-03**	**0.031**	< 1.00e-05	< 1.90e-04
Height	0.031	0.532	3.81e-03	0.072	1.74e-04	**3.31e-03**	< 1.00e-05	< 1.90e-04
HC	0.095	0.930	**0.031**	0.503	**0.010**	0.196	< 1.00e-05	< 1.90e-04
WC	0.127	0.928	0.057	0.662	**0.022**	0.345	1.20e-05	2.28e-04
ALT	0.204	0.920	0.172	0.646	0.106	0.617	1.76e-03	0.033
Albumin	0.117	0.983	**0.046**	0.593	**0.024**	0.395	< 1.00e-05	< 1.90e-04
ALP	0.442	1.000	0.419	1.000	0.318	1.000	0.261	1.000
AST	0.293	1.000	0.322	1.000	0.276	0.875	0.187	1.000
Bilirubin	0.046	0.629	0.027	0.390	8.43e-03	0.160	< 1.00e-05	< 1.90e-04
Cholesterol	0.024	0.448	2.25e-03	**0.043**	1.96e-04	**3.72e-03**	< 1.00e-05	< 1.90e-04
GGT	0.307	1.000	0.290	0.801	0.265	0.800	0.039	0.734
HDL	0.063	0.717	**0.017**	0.326	**4.76e-03**	0.090	< 1.00e-05	< 1.90e-04
LDL	0.012	0.222	6.10e-04	**0.012**	2.20e-05	**4.18e-04**	< 1.00e-05	< 1.90e-04
Triglycerides	0.242	1.000	0.164	1.000	0.126	0.880	6.76e-04	0.013
HbA1c2	6.23e-03	0.118	3.66e-04	**6.95e-03**	1.80e-05	**3.42e-04**	< 1.00e-05	< 1.90e-04
WBC	6.95e-03	0.139	< 1.00e-05	< **2.00e-04**	< 1.00e-05	< **2.00e-04**		
RBC	0.011	0.219	4.40e-05	**8.80e-04**	< 1.00e-05	< **2.00e-04**		
Hemoglobin	0.041	0.815	1.18e-03	**2.36e-02**	1.40e-05	**2.80e-04**		
HCT	0.025	0.508	3.36e-04	**6.72e-03**	< 1.00e-05	< **2.00e-04**		
MCV	1.47e-03	0.029	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		
MCH	2.50e-03	0.050	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		
MCHC	<1.00e-05	<2.00e-04	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		
RDW	6.70e-03	0.134	< 1.00e-05	< **2.00e-04**	< 1.00e-05	< **2.00e-04**		
PLT	3.00e-03	0.060	< 1.00e-05	< **2.00e-04**	< 1.00e-05	< **2.00e-04**		
MPV	1.00e-05	2.00e-04	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		
NEUPr	0.015	0.304	7.80e-05	**1.56e-03**	< 1.00e-05	< **2.00e-04**		
LYMPHPr	3.30e-03	0.066	< 1.00e-05	< **2.00e-04**	< 1.00e-05	< **2.00e-04**		
MONOPr	7.48e-03	0.150	< 1.00e-05	< **2.00e-04**	< 1.00e-05	< **2.00e-04**		
EOSPr	1.13e-01	1.000	**0.017**	0.331	**7.08e-04**	**0.014**		
BASOPr	9.60e-04	0.019	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		
LYMPH	4.10e-04	0.008	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		
NEU	0.062	1.000	**3.31e-03**	0.066	**6.00e-05**	**1.20e-03**		
MONO	0.012	0.236	2.40e-05	**4.80e-04**	< 1.00e-05	< **2.00e-04**		
EOS	0.212	1.000	0.080	1.000	**6.78e-03**	0.136		
BASO	1.20e-05	2.40e-04	< 1.00e-05	< 2.00e-04	< 1.00e-05	< 2.00e-04		

indices, blood factors, glycemic control, blood pressure, lipid tests, and liver function tests (see the complete list of phenotypes in Supplementary section S6). Initially, 5000 individuals were genotyped using the Illumina HumanOmni 2.5M BeadChip array, out of which 4778 samples pass the quality control. We follow Heckerman *et al.* [49] exactly for quality control including the Hardy-Weinberg equilibrium (HWE) test, exclusion of Single Nucleotide Polymorphisms (SNPs) with low Minor Allele Frequency (MAF), and computation of the related matrix.

Among all the phenotypes, 18 phenotypes were measured for all the subjects, while the remaining 19 phenotypes were recorded for only 1423 subjects. Thus we conduct two sets of experiments for these two sets of phenotypes. For the 18 phenotypes measured for all individuals, we conduct experiments to mimic the random missingness of phenotypes. We subsample 3000 individuals as unpaired data and allocate the rest as paired data. We compare the p-values of the KIT as a baseline with two variants of SAT-rx, with and without dimensionality reduction. In this experiment, we are mimicking the missingness, hence we have access to the oracle, i.e., applying KIT using all data as paired, which is the best we can achieve and which we also compare with our method. The upper half of Table 1 reports the p-values generated by different methods for all evaluated phenotypes. We can see that the oracle produces much smaller p-values in general, while the baseline KIT method can hardly find significant associations. Our SAT-rx method clearly outperforms the KIT method and approaches the performance of the oracle on some phenotypes. Among the 18 phenotypes, our method finds 5 more heritable phenotypes than the baseline method at significance level 0.05.

For the other 19 phenotypes, 1415 individuals have both genotype and phenotype values, and the remaining individuals are considered unpaired (only genotype). We compare the p-values of the KIT as a baseline with two variants of SAT-rx, with and without dimensionality reduction. The lower half of Table 1 reports the p-values for all methods evaluated on these phenotypes. Among the 19 phenotypes, our method identifies 12 more heritable phenotypes than the baseline method at significance level 0.05.

Discussion

Establishing the associations between various types of biomedical variables is essential for an understanding of disease mechanisms. When the biomedical variables are high dimensional, for example SNPs and imaging phenotypes, a huge sample size is required to guarantee enough statistical power. Also, a small sample size increases the chance of false discovery. However, due to the high cost of data collection, the available sample size is typically not sufficiently large, and the missingness in the data can cause a further reduction of sample size. To alleviate that problem, the biomedical research community is increasingly turning toward other sources (e.g., EHRs) where a massive amount of data is available. However, there is no guarantee that all subjects have all data modalities.

Here we address one type of missingness that frequently happens in the current association testing. When studying the relationships between two variables, one problem researchers usually encounter is that many data points have observations of only one variable (unpaired data), because the data for these variables may be collected initially for other purposes. We aim at being less "wasteful" by exploiting data points that have this type of missingness. Our method is based on a technical assumption that the data missing mechanism is independent of the association relationship, i.e., missing completely at random (MCAR). If this assumption is violated (e.g., we are only given biased paired data), our method cannot recover the original association. A future direction could be to extend our method to deal with more general missing mechanisms. Another primary assumption is that the distributions of the paired and unpaired data points are the same.

For example, our method cannot be directly applied if gene expression data in the paired samples is collected by one platform while unpaired data is collected using a different platform. In this case, pre-processing is required to ensure that the platform bias is removed and the marginals of the paired and unpaired data are the same.

Although we showed the applicability of our method for univariate phenotypes (Table 1), the main focus of this paper is on multivariate phenotypes [10, 50]. Multivariate phenotypes provide more information than univariate ones, especially to study complicated phenomena such as the effect of cortical brain folding on the onset and progression of neurological diseases [9, 14, 51] and morphological traits in evolutionary biology [52]. Unpaired data enables us to discover the linear or non-linear relationship between variables in the multivariate phenotype.

Our method can take advantage of the unpaired data in two ways. First, our method improves the null distribution estimation by using unpaired data, which offers better control of the false discovery rate. As shown in Theorem 1, the estimation error of the null distribution depends on the total sample size, suggesting that incorporating unpaired data can readily improve the estimation of the null distribution. Second, we construct a new test statistic that explores the low-dimensional structure from unpaired data. We showed that under mild conditions, the new test has higher test power. It is worth noting that higher power can only be obtained if at least one type of data has a low-dimensional structure. Otherwise, our method can possibly have lower power than the baseline methods due to the removal of useful information. The low-dimensional assumption is a reasonable assumption for a multivariate phenotype. Measurements from highly structured data such as imaging are usually modeled as data points on a low dimensional manifold. For example, Gerber *et al.* [31] constructed a low dimensional brain manifold to study clinical variables such as age. Schabdach *et al.* [11] used manifold learning techniques to model information extracted from CT images of the lungs and showed that the low dimensional representation is highly correlated with the severity of the disease. Shi *et al.* [53] assumed gene expression data lay on a manifold and used a nonlinear dimensionality reduction method to capture biologically relevant structures in the data.

Our approach is closely related to the linear mixed effect model. The model is widely applied in statistical genetics to estimate the additive genetic effects of a univariate phenotype. Discovering a non-linear effect requires a much larger sample size, which might not be practical in terms of collecting enough data, at least when working with genotype data. However, there is no limitation in our approach's ability to account for non-linear effects for other types of data. The linear effect assumption is equivalent to using linear kernels; different choices of the kernels (e.g., Radial Basis Function) can model non-linear effects. While the definition of heritability is well-defined for a univariate phenotype, it is less clear for a multivariate phenotype. We adopted the notion proposed by Ge *et al.* [10] (see Eq. 5). Their definition has several advantages. First, it is nicely connected to the mixed effect model and generalizes univariate heritability. Second, it allows us to incorporate unpaired data and derive the null distribution, which is required to compute the p-value efficiently, due to a mathematically appealing link with KIT and VCST. Deriving the null distribution for other definitions of multivariate heritability while incorporating the unpaired data (e.g., Zhou *et al.* [54]) is not as straightforward.

We conducted intensive simulation studies, evaluating various aspects of our proposed method. We generated synthetic data using the linear mixed effect model. To distinguish between SAT-fx and SAT-rx, we fix X in all the iterations for evaluating SAT-fx and randomly generate X in each replication for evaluating SAT-rx. Also, we provided a more practical simulation that uses real genotype data as X and only synthesizes Y. We set the heritability level (h^2) to 10%, which is a

modest heritability value. The higher values of h^2 are less challenging than $h^2 = 0.1$ since the X and Y are more strongly related. The results in all the simulations demonstrate that our method better controls the type I error and has higher power than the baseline methods that ignore unpaired data. Fig. 4 suggests that when the dimensionality of X is large (e.g., X is genotype data) and the effect size is modest, our method controls the type I error conservatively and requires a larger sample size to reach the nominal level of type I error (i.e., 5%). Nevertheless, Fig. 5 shows that the gain in power using unpaired data is significant.

We also applied our method to the multivariate phenotype extracted from lung CT scans of patients with COPD. Two significant components of COPD are airway remodeling and alveolar destruction (emphysema) [55, 56]. Many pathogenic processes, such as chronic inflammation, contribute to the disease or cause anatomical variation [57, 58]. There is some evidence that the inflammatory process [58] and autoimmune response [59] are involved in emphysema. Researchers showed an association between various molecular signatures for COPD and emphysema in the peripheral blood mononuclear cells (PBMCs) [44, 60]. For example, Bowler *et al.* [60] investigated whether Interleukin-16, which is associated with autoimmune disease, is associated with COPD. Carolan *et al.* [61, 62] investigated the association of various blood biomarkers (e.g., adiponectin) with clinical and radiologic COPD phenotypes. However, those studies used FEV1 or LAA as surrogates for the disease severity, both of which are aggregate measures and cannot characterize a subprocess involved in the disease. For example, LAA is insufficient to distinguish emphysema visual subtypes because it merely counts the number of pixels in the lung region of CT images with intensity values lower than a single threshold. Since more sophisticated imaging descriptors (e.g., texture) are shown to be effective for emphysema sub-typing [63–65], we hypothesize that such descriptors are also associated with the systemic characterization of the disease such as PBMCs and gene expression. In this paper, we used a previously developed method [11] that uses image texture and intensity value and constructs a multivariate vector for each patient. The patient vector is shown to be more potent than LAA to characterize the disease severity [11]. We also construct a multivariate phenotype from rather traditional imaging measurements shown to be informative in previous studies [46]. We study the relationship between the imaging phenotype with the various blood measurements that are correlated with the activities of the immune system. Figure 6 reports that the classical approach in both experiments cannot detect the dependence while our method can. It is important to note that we test the dependency between a group of genes involved in the immune pathway, and our method does not identify specific causal genes contributing to the destruction of the tissue, which needs further investigation.

When computing the test statistic, our method adds additional computational load compared to the baselines VCST and KIT, due to the eigendecomposition of the kernel matrix on the paired and unpaired data. We generated samples from the null distribution to calculate p-values for both the baseline methods and our method, which is computationally expensive if a high precision p-value is required. Methods such as gamma approximation can be used to speed up the computation, which will of course introduce approximation errors.

Method

In this section, we first give a brief review of the variance component score test (VCST) and the kernel independence test (KIT). We then discuss the connections between them and show that the differences between them lead to different ways to utilize unpaired data. Finally, we detail our SAT method by demonstrating how unpaired data can be incorporated to improve both VCST and KIT.

Variance Component Score Test (VCST)

We start with the variance component model (a.k.a. the random effect model), which is widely used in statistical genetics for genetic association studies [10,16,26,66]. We use the same nomenclature where $Y \in \mathbb{R}^p$ is a p-dimensional phenotype and $X \in \mathbb{R}^d$ is genotype. However, our method is general and can be applied elsewhere. Given a paired sample containing n observations $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n$, we consider the following multidimensional variance component model [10]:

$$y_{ik} = \mu_{ik} + g_k(\mathbf{x}_i) + \epsilon_{ik}, \tag{1}$$

where y_{ik} is the k-th element of $\mathbf{y}_i$, g_k is a nonparametric function in a reproducing kernel Hilbert space (RKHS) associated with kernel $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$, μ_{ik} is the offset term, and ϵ_{ik} is the error term. (1) can be rewritten in matrix form:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{G} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^{n \times p}$ is the phenotypic matrix of the n observations (subjects) with i-th row $\mathbf{y}_i^\intercal$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \otimes \mathbf{1}_n$ is a matrix of offsets ($\mathbf{1}_n$ is an $n \times 1$ vector of ones), $\mathbf{G} \in \mathbb{R}^{n \times p}$ is the matrix of the aggregate genetic effects, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times p}$ is a matrix of residual effects. We have the following distributional assumptions:

$$\text{vec}(\mathbf{G}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_g \otimes \mathbf{K}), \quad \text{vec}(\boldsymbol{\epsilon}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_n), \tag{3}$$

where $\text{vec}(\cdot)$ is the matrix vectorization operator that converts a matrix into a vector by stacking its columns, $\otimes$ is the Kronecker product of matrices, $\mathbf{I}_n$ denotes an $n \times n$ identity matrix, $\boldsymbol{\Sigma}_g$ is the genetic covariance matrix, $\boldsymbol{\Sigma}_\epsilon$ is the residual covariance matrix, and $\mathbf{K}$ is the kernel matrix with ij-th element $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. For example, in the context of statistical genetics, $\mathbf{K}$ denotes identity-by-state (IBS) kernel [17,67,68], where $[\mathbf{K}]_{ij}$ represents the relatedness between individual i and j.

To test whether Y and X are associated (whether Y is heritable if X is the genotype), we can test the variance components as $\mathcal{H}_0 : tr(\boldsymbol{\Sigma}_g) = 0$ versus $\mathcal{H}_1 : tr(\boldsymbol{\Sigma}_g) > 0$ using the following score test statistic derived from model (1):

$$\hat{\mathcal{S}}_n(\mathbf{K}, \mathbf{L}) = \frac{1}{n^2} tr(\mathbf{K}\mathbf{H}_n\mathbf{L}\mathbf{H}_n) - \frac{1}{n^3} tr(\mathbf{H}_n\mathbf{L}) tr(\mathbf{H}_n\mathbf{K}), \tag{4}$$

where $tr(\cdot)$ computes the trace of a matrix, $\mathbf{L} = \mathbf{Y}\hat{\boldsymbol{\Sigma}}_Y^{-2}\mathbf{Y}^\intercal$ and $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal$, and $\hat{\boldsymbol{\Sigma}}_Y$ is the empirical covariance matrix of Y. The derivation details are provided in the Supplementary Information. The exact fraction of phenotype variability attributed to genetic variation is defined as heritability. There are various ways to define heritability for a multivariate phenotype (e.g., [10,54]). We adopt the definition by Ge *et al.* [10] that closely related to the VCST and subsumes the definition of the heritability for the univariate phenotype, which can be calculated as follows [10]:

$$h^2 = \frac{tr(\boldsymbol{\Sigma}_g)}{tr(\boldsymbol{\Sigma}_g) + tr(\boldsymbol{\Sigma}_\epsilon)}. \tag{5}$$

Kernel Independence Test (KIT)

Kernel independence tests are a class of nonparametric methods which are also widely used for genetic association studies [22,28]. Here we briefly review the Hilbert-Schmidt Independence Criterion (HSIC)-based independence test [22], which provides a general framework for many association

tests [25]. Let $\mathcal{F}_y$ be a RKHS associated with the kernel function $l(\mathbf{y}, \mathbf{y}') = \langle \psi(\mathbf{y}), \psi(\mathbf{y}') \rangle$. HSIC tests $\mathcal{H}_0 : \mathbb{P}_{YX} = \mathbb{P}_X \mathbb{P}_Y$ versus $\mathcal{H}_1 : \mathbb{P}_{YX} \neq \mathbb{P}_X \mathbb{P}_Y$ by testing $\mathcal{H}_0 : I = 0$ versus $\mathcal{H}_1 : I > 0$, where I is defined as follows:

$$I = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}[k(X, X')l(Y, Y')] + \mathbb{E}_X\mathbb{E}_{X'}\mathbb{E}_Y\mathbb{E}_{Y'}[k(X, X')l(Y, Y')] - 2\mathbb{E}_{XY}[\mathbb{E}_{X'}[k(X, X')]\mathbb{E}_{Y'}[l(Y, Y')]]. \quad (6)$$

Given paired data of n subjects, an unbiased estimator of I is the following [69]:

$$\hat{I}_n(\mathbf{K}, \mathbf{L}) = \frac{1}{n(n-3)}\Big[tr(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) + \frac{\mathbf{1}_n^\intercal \tilde{\mathbf{K}} \mathbf{1}_n \mathbf{1}_n^\intercal \tilde{\mathbf{L}} \mathbf{1}_n}{(n-1)(n-2)} - \frac{2\mathbf{1}_n^\intercal \tilde{\mathbf{K}}\tilde{\mathbf{L}}\mathbf{1}_n}{n-2}\Big], \quad (7)$$

where $\tilde{\mathbf{K}} = \mathbf{K} - diag(\mathbf{K})$ and similarly for $\tilde{\mathbf{L}}$ and $\mathbf{L}_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$. To test for statistical independence, one can use characteristic kernels, e.g., the radial basis function $\mathbf{K}_{ij} = \exp\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right)$, such that I can be zero only when X and Y are independent [70].

Connections between VCST and KIT

Now we discuss the similarities and differences between VCST and KIT. Supplementary Table 2 displays the test statistics and null distributions of VCST and KIT.

Test statistic It can be seen from Supplementary Table 2 that the biased statistics of VCST and KIT are identical to each other, if setting $\psi(\mathbf{y}) = \hat{\mathbf{\Sigma}}_Y^{-1}\mathbf{y}$. The unbiased test statistics of VCST and KIT differ. This is because VCST tests for random effects but assumes that the covariate inducing the random effect (X) and the corresponding kernel matrix ($\mathbf{K}$) are fixed while KIT assumes X is random, leading to different ways to correct for the bias.

Null distribution Let η_j ($\hat{\eta}_j$) be the eigenvalues (empirical) of the covariance of $\phi(X)$ and let λ_i ($\hat{\lambda}_j$) be the eigenvalues (empirical) of the covariance of $\psi(Y)$. As shown in Table **??**, the null distributions for VCST and KIT have exactly the same forms, except that VCST uses $\hat{\eta}_j$ while KIT uses η_j. This is also because of their respective fixed or random X assumptions. In practice, because λ_i and η_j are both unknown, we need to replace them with $\hat{\lambda}_i$ and $\hat{\eta}_i$. Therefore, the empirical null distributions of VCST and KIT are identical if only given n paired examples. However, they are inherently different because the null distribution of KIT is derived from asymptotic theory, while the null distribution of VCST is derived from the Gaussian error terms in the variance component model (2). This subtle difference is significant when using unpaired data, which is described as follows.

Unpaired data The main difference between VCST and KIT is that X ($\mathbf{K}$) is considered fixed or random respectively. When given unpaired data, VCST cannot make use of the unpaired data of X due to the fixed X assumption, while KIT can benefit from unpaired data of both X and Y. More specifically, unpaired data can only be used to improve the estimation of λ_i in VCST but they can be used to improve the estimation of both η_j and λ_i in KIT.

Semi-paired Association Test

In this section, we present our SAT method that incorporates unpaired data to improve test power. In addition to the n paired data, suppose we also have access to an unpaired sample $\{\mathbf{x}_i\}_{i=n+1}^N$ and an unpaired sample $\{\mathbf{y}_i\}_{i=n+1}^M$. Without loss of generality, we assume $N = M$ and replace M with

N for notational simplicity. We will show two ways that unpaired data can improve the association test: 1) better control of type I error by improving the estimation of null distributions and 2) improved test power by devising a new test statistic under the intrinsic low-dimension assumption of high-dimensional data. We show how unpaired data are used for both VCST and KIT, resulting in two variants of our method, SAT-fx and SAT-rx.

Enhancing Type I Error Control. To calculate p-values, we need to estimate the parameters λ_i and η_j in the null distributions from empirical data. Because λ_i and η_j are the eigenvalues of the covariance of $\psi(Y)$ and $\phi(X)$, respectively, the estimation does not require paired Y and X examples. Therefore, we can readily make use of unpaired data to obtain more accurate estimation of λ_i or η_j involved in the null distribution.

For SAT-fx, we add unpaired Y data to estimate the covariance of $\psi(Y)$ and its eigenvalues λ_i from both paired and unpaired data $\{\mathbf{y}_i\}_{i=1}^N$, while η_j should be estimated from only $\{\mathbf{x}_i\}_{i=1}^n$ in the paired sample. For SAT-rx, we can further incorporate unpaired X data and use all the X data $\{\mathbf{x}_i\}_{i=1}^N$ to estimate η_j.

The following theorem shows that 1) the empirical null distribution convergences to the true (asymptotic) distribution and 2) the variance of the empirical null distribution converges to the variance of the true (asymptotic) null distribution with rate $1/\sqrt{m}$, where m is the sample size of available data for estimating λ_i and η_j.

Theorem 1 (Informal). *Let* $\check{I} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\lambda_i\eta_j(z_{ij}^2-1)$ *and* $\check{I}_m = \sum_{i=1}^{m}\sum_{j=1}^{m}\hat{\lambda}_i\hat{\eta}_j(z_{ij}^2-1)$.
1) As $m\to\infty$, $\check{I}_m$ *converges in distribution to* $\check{I}$.
2) For all $\mathbb{P}_{XY}$, $\mathbb{E}(\check{I}_m)=\mathbb{E}(\check{I})$ *and* $\mathbb{V}(\check{I}_m)$ *converges in probability to* $\mathbb{V}(\check{I})$ *with rate* $1/\sqrt{m}$.

The theorem is developed for SAT-rx and a similar theorem for SAT-fx can be considered as a special case of the above theorem. From the theorem, we can see that if only using paired data, $m=n$; if further using unpaired data, $m=N$. Because $N>n$, incorporating unpaired data to estimate λ_i and η_j leads to lower estimation error and provides more accurate estimation of the null distribution. The proof details of Theorem 1 are given in Section 6 of the Supplementary Information.

Improving Test Power Unpaired data contribute to a better estimation of the null distribution, resulting in better control of type I error. It can also improve test power. Specifically, if X or Y data (approximately) lie in a low-dimensional space, we show that unpaired data can be used to construct a new test statistic with improved test power. To devise the new test statistics, we first learn the low-dimensional space of X or Y by applying the kernel Principal Component Analysis (PCA) algorithm on both paired and unpaired data. Second, we project the paired data to the learned low-dimensional space and obtain the test statistics of our SAT-fx and SAT-rx by estimating the test statistics of VCST and KIT on the projected data. Due to the use of the kernel trick, calculating the test statistic of SAT-fx and SAT-rx requires only the kernel matrices $\mathbf{K}_N$ and $\mathbf{L}_N$ which are calculated on all the data, paired and unpaired.

In SAT-fx, because we do not consider X as random as does VCST, we can only incorporate unpaired Y data to learn the low-dimensional structure of Y. In SAT-rx, we further use unpaired data X to learn the low-dimensional space of X. The proposed new test statistics of SAT-fx and SAT-rx have the same form as that of VCST (4) and KIT (7), respectively. We only need to change the kernel matrices in the test statistics. Specifically, the new test statistic for SAT-fx is defined as $\hat{\mathcal{S}}_n(\mathbf{K},\mathbf{L}')$, where

$$\mathbf{L}' = \bar{\mathbf{L}}^\intercal\mathbf{U}\mathbf{\Lambda}_y^{-1}\mathbf{U}^\intercal\bar{\mathbf{L}}. \tag{8}$$

In $\mathbf{L}'$, $\bar{\mathbf{L}}$ is the matrix comprised of the first n columns of $\mathbf{L}_N$, $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_{r_Y})$ and $\mathbf{\Lambda}_y = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_{r_Y})$ are the top r_Y eigenvectors and eigenvalues of $\mathbf{L}_N$.

Similarly, the new test statistic of SAT-rx that considers X as random is $\hat{I}_n(\mathbf{K}', \mathbf{L}')$, where

$$\mathbf{K}' = \bar{\mathbf{K}}^\intercal \mathbf{V} \mathbf{\Lambda}_x^{-1} \mathbf{V}^\intercal \bar{\mathbf{K}}. \tag{9}$$

In $\mathbf{K}'$, $\bar{\mathbf{K}}$ is the matrix composed of the first n columns of $\mathbf{K}_N$, $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_{r_X})$ and $\mathbf{\Lambda}_x = \text{diag}(\hat{\eta}_1, \cdots, \hat{\eta}_{r_X})$ are the top r_X eigenvectors and eigenvalues of $\mathbf{K}_N$. The asymptotic null distributions of the proposed $\hat{\mathcal{S}}'_n$ and $\hat{I}'_n$ have the same forms as the null distributions of $\hat{\mathcal{S}}_n$ and $\hat{I}_n$, but using only the top eigenvalues $\{\lambda_i\}_{i=1}^{r_Y}$ and $\{\eta_j\}_{j=1}^{r_X}$, respectively. The derivation details are provided in Section 7 of the Supplementary Information.

The following theorem shows that the power of the new test statistic of SAT-rx is greater than the classical one that only uses paired data.

Theorem 2 (**Informal**). *Assuming that data from X and Y lie in a low-dimensional manifold, the test power of the proposed SAT-rx is higher than that of the KIT method, which only uses paired data.*

SAT-fx follows similar properties as SAT-rx and can be considered as a special case of SAT-rx. The proof details of Theorem 2 are given in Section 8 of the Supplementary Information.

Author Contributions

K.B. envisioned the project. M.G., K.B., and K.Z. developed the statistical method. M.G. implemented the method and performed the analyses. M.G. and K.B. wrote the paper. P.L. processed and performed analysis on the COPD imaging and immune system gene expression data. P.S. helped interpret the experimental results on medical data. D.T. provided assistance in theoretical analysis of the method. G.T. helped proofread biostatistical aspects of the paper.

Competing Interests

The authors declare no competing interests.

References

[1] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *science*, 322(5903):881–888, 2008.

[2] Georg B Ehret, Patricia B Munroe, Kenneth M Rice, Murielle Bochud, Andrew D Johnson, Daniel I Chasman, Albert V Smith, Martin D Tobin, Germaine C Verwoert, Shih-Jen Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103, 2011.

[3] Jacob Gratten, Naomi R Wray, Matthew C Keller, and Peter M Visscher. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature neuroscience*, 17(6):782, 2014.

[4] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274, 2013.

[5] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197, 2015.

[6] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[7] John G Csernansky, Sarang Joshi, Lei Wang, John W Haller, Mokhtar Gado, J Philip Miller, Ulf Grenander, and Michael I Miller. Hippocampal morphometry in schizophrenia by high dimensional brain mapping. *Proceedings of the National Academy of Sciences*, 95(19):11406–11411, 1998.

[8] Emilie Gerardin, Gaël Chételat, Marie Chupin, Rémi Cuingnet, Béatrice Desgranges, Ho-Sung Kim, Marc Niethammer, Bruno Dubois, Stéphane Lehéricy, Line Garnero, et al. Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging. *Neuroimage*, 47(4):1476–1486, 2009.

[9] Xiaoying Tang, Dominic Holland, Anders M Dale, Laurent Younes, Michael I Miller, and Alzheimer's Disease Neuroimaging Initiative. Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and alzheimer's disease: detecting, quantifying, and predicting. *Human brain mapping*, 35(8):3701–3725, 2014.

[10] Tian Ge, Martin Reuter, Anderson M Winkler, Avram J Holmes, Phil H Lee, Lee S Tirrell, Joshua L Roffman, Randy L Buckner, Jordan W Smoller, and Mert R Sabuncu. Multidimensional heritability analysis of neuroanatomical shape. *Nature communications*, 7:13291, 2016.

[11] Jenna Schabdach, William M Wells, Michael Cho, and Kayhan N Batmanghelich. A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies. In *International Conference on Information Processing in Medical Imaging*, pages 170–183. Springer, 2017.

[12] Sumedha Singla, Mingming Gong, Siamak Ravanbakhsh, Frank Sciurba, Barnabas Poczos, and Kayhan N Batmanghelich. Subject2vec: Generative-discriminative approach from a set of image patches to a vector. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 502–510. Springer, 2018.

[13] JG Csernansky, L Wang, J Swank, JP Miller, M Gado, D McKeel, MI Miller, and JC Morris. Preclinical detection of alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. *Neuroimage*, 25(3):783–792, 2005.

[14] Jaeuk Hwang, In Kyoon Lyoo, Stephen R Dager, Seth D Friedman, Jung Su Oh, Jun Young Lee, Seog Ju Kim, David L Dunner, and Perry F Renshaw. Basal ganglia shape alterations in bipolar disorder. *American Journal of Psychiatry*, 163(2):276–285, 2006.

[15] Carrie E Bearden, Paul M Thompson, Rebecca A Dutton, Benício N Frey, Marco AM Peluso, Mark Nicoletti, Nicole Dierschke, Kiralee M Hayashi, Andrea D Klunder, David C Glahn, et al. Three-dimensional mapping of hippocampal anatomy in unmedicated and lithium-treated patients with bipolar disorder. *Neuropsychopharmacology*, 33(6):1229, 2008.

[16] Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

[17] Lydia Coulter Kwee, Dawei Liu, Xihong Lin, Debashis Ghosh, and Michael P Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, 2008.

[18] Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

[19] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[20] Tian Ge, Thomas E Nichols, Phil H Lee, Avram J Holmes, Joshua L Roffman, Randy L Buckner, Mert R Sabuncu, and Jordan W Smoller. Massively expedited genome-wide heritability analysis (megha). *Proceedings of the National Academy of Sciences*, 112(8):2479–2484, 2015.

[21] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[22] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[23] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

[24] K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

[25] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2291, 2013.

[26] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[27] Wen-Yu Hua and Debashis Ghosh. Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics*, 71(3):812–820, 2015.

[28] Changshuai Wei and Qing Lu. A generalized association test based on u statistics. *Bioinformatics*, page btx103, 2017.

[29] Gershim Asiki, Georgina Murphy, Jessica Nakiyingi-Miiro, Janet Seeley, Rebecca N Nsubuga, Alex Karabarinde, Laban Waswa, Sam Biraro, Ivan Kasamba, Cristina Pomilla, et al. The general population cohort in rural south-western uganda: a platform for communicable and non-communicable disease studies. *International journal of epidemiology*, 42(1):129–141, 2013.

[30] Samuel Gerber, Tolga Tasdizen, Sarang Joshi, and Ross Whitaker. On the manifold structure of the space of brain images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–312. Springer, 2009.

[31] Samuel Gerber, Tolga Tasdizen, P Thomas Fletcher, Sarang Joshi, Ross Whitaker, Alzheimers Disease Neuroimaging Initiative, et al. Manifold modeling for brain population analysis. *Medical image analysis*, 14(5):643–653, 2010.

[32] El-ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana Pe'er. visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 31(6):545, 2013.

[33] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10):886, 2011.

[34] Laleh Haghverdi, Florian Buettner, and Fabian J Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.

[35] Arnab Maity and Xihong Lin. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics*, 67(4):1271–1284, 2011.

[36] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

[37] Jørgen Vestbo, Suzanne S Hurd, Alvar G Agustí, Paul W Jones, Claus Vogelmeier, Antonio Anzueto, Peter J Barnes, Leonardo M Fabbri, Fernando J Martinez, Masaharu Nishimura, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American journal of respiratory and critical care medicine*, 187(4):347–365, 2013.

[38] Joyce D Schroeder, Alexander S McKenzie, Jordan A Zach, Carla G Wilson, Douglas Curran-Everett, Douglas S Stinson, John D Newell Jr, and David A Lynch. Relationships between airflow obstruction and quantitative ct measurements of emphysema, air trapping, and airways in subjects with and without chronic obstructive pulmonary disease. *American Journal of Roentgenology*, 201(3):W460–W470, 2013.

[39] Naoki Sakai, Michiaki Mishima, Koichi Nishimura, Harumi Itoh, and Kenshi Kuno. An automated method to assess the distribution of low attenuation areas on chest ct scans in chronic pulmonary emphysema patients. *Chest*, 106(5):1319–1325, 1994.

[40] AD Rames and PW Jones. Tesra (treatment of emphysema with a selective retinoid agonist) study results. *Am J Respir Crit Care Med*, 183:A6418, 2011.

[41] Katashi Satoh, Takuya Kobayashi, Takahiko Misao, Yoshimi Hitani, Yuka Yamamoto, Yoshihiro Nishiyama, and Motoomi Ohkawa. Ct assessment of subtypes of pulmonary emphysema in smokers. *Chest*, 120(3):725–729, 2001.

[42] Carlos S Mendoza, George R Washko, James C Ross, AA Diaz, David A Lynch, James D Crapo, Edward K Silverman, Begoña Acha, Carmen Serrano, and R San José Estépar. Emphysema quantification in a multi-scanner hrct cohort using local intensity distributions. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 474–477. IEEE, 2012.

[43] Lauge Sorensen, Mads Nielsen, Pechin Lo, Haseem Ashraf, Jesper H Pedersen, and Marleen De Bruijne. Texture-based analysis of copd: a data-driven approach. *IEEE transactions on medical imaging*, 31(1):70–78, 2012.

[44] Timothy M Bahr, Grant J Hughes, Michael Armstrong, Rick Reisdorph, Christopher D Coldren, Michael G Edwards, Christina Schnell, Ross Kedl, Daniel J LaFlamme, Nichole Reisdorph, et al. Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *American journal of respiratory cell and molecular biology*, 49(2):316–323, 2013.

[45] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[46] SungHwan Kim, Jose D Herazo-Maya, Dongwan D Kang, Brenda M Juan-Guardela, John Tedrow, Fernando J Martinez, Frank C Sciurba, George C Tseng, and Naftali Kaminski. Integrative phenotyping framework (ipf): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC genomics*, 16(1):924, 2015.

[47] Peter M Visscher, Sarah E Medland, Manuel AR Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, 2(3):e41, 2006.

[48] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[49] David Heckerman, Deepti Gurdasani, Carl Kadie, Cristina Pomilla, Tommy Carstensen, Hilary Martin, Kenneth Ekoru, Rebecca N Nsubuga, Gerald Ssenyomo, Anatoli Kamali, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113(27):7377–7382, 2016.

[50] Michael L Collyer, David J Sekora, and Dean C Adams. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity*, 115(4):357, 2015.

[51] Shantanu H Joshi, Katherine L Narr, Owen R Philips, Keith H Nuechterlein, Robert F Asarnow, Arthur W Toga, and Roger P Woods. Statistical shape analysis of the corpus callosum in schizophrenia. *Neuroimage*, 64:547–559, 2013.

[52] Simon M Huttegger and Philipp Mitteroecker. Invariance and meaningfulness in phenotype spaces. *Evolutionary Biology*, 38(3):335–351, 2011.

[53] Jinlong Shi and Zhigang Luo. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Computers in biology and medicine*, 40(8):723–732, 2010.

[54] Jin J Zhou, Michael H Cho, Peter J Castaldi, Craig P Hersh, Edwin K Silverman, and Nan M Laird. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *American journal of respiratory and critical care medicine*, 188(8):941–947, 2013.

[55] JC Hogg, JL Wright, BR Wiggs, HO Coxson, A Opazo Saez, and PD Pare. Lung structure and function in cigarette smokers. *Thorax*, 49(5):473–478, 1994.

[56] Hironi Makita, Yasuyuki Nasuhara, Katsura Nagai, Yoko Ito, Masaru Hasegawa, Tomoko Betsuyaku, Yuya Onodera, Nobuyuki Hizawa, and Masaharu Nishimura. Characterisation of phenotypes based on severity of emphysema in chronic obstructive pulmonary disease. *Thorax*, 62(11):932–937, 2007.

[57] Joshua D Campbell, John E McDonough, Julie E Zeskind, Tillie L Hackett, Dmitri V Pechkovsky, Corry-Anke Brandsma, Masaru Suzuki, John V Gosselink, Gang Liu, Yuriy O Alekseyev, et al. A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide ghk. *Genome medicine*, 4(8):67, 2012.

[58] James C Hogg, Fanny Chu, Soraya Utokaparch, Ryan Woods, W Mark Elliott, Liliana Buzatu, Ruben M Cherniack, Robert M Rogers, Frank C Sciurba, Harvey O Coxson, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 350(26):2645–2653, 2004.

[59] Laima Taraseviciene-Stewart, Ivor S Douglas, Patrick S Nana-Sinkam, Jong D Lee, Rubin M Tuder, Mark R Nicolls, and Norbert F Voelkel. Is alveolar destruction and emphysema in chronic obstructive pulmonary disease an immune disease? *Proceedings of the American Thoracic Society*, 3(8):687–690, 2006.

[60] Russell P Bowler, Timothy M Bahr, Grant Hughes, Sharon Lutz, Yu-Il Kim, Christopher D Coldren, Nichole Reisdorph, and Katerina J Kechris. Integrative omics approach identifies interleukin-16 as a biomarker of emphysema. *Omics: a journal of integrative biology*, 17(12):619–626, 2013.

[61] Brendan J Carolan, Yu-il Kim, André A Williams, Katerina Kechris, Sharon Lutz, Nichole Reisdorph, and Russell P Bowler. The association of adiponectin with computed tomography phenotypes in chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 188(5):561–566, 2013.

[62] Brendan J Carolan, Grant Hughes, Jarrett Morrow, Craig P Hersh, Wanda K O?Neal, Stephen Rennard, Sreekumar G Pillai, Paula Belloni, Debra A Cockayne, Alejandro P Comellas, et al.

The association of plasma biomarkers with computed tomography-assessed emphysema phenotypes. *Respiratory research*, 15(1):127, 2014.

[63] Lauge Sørensen, Saher B Shaker, and Marleen De Bruijne. Texture classification in lung ct using local binary patterns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 934–941. Springer, 2008.

[64] Lauge Sorensen, Saher B Shaker, and Marleen De Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE transactions on medical imaging*, 29(2):559–569, 2010.

[65] Jie Yang, Elsa D Angelini, Pallavi P Balte, Eric A Hoffman, John HM Austin, Benjamin M Smith, Jingkuan Song, R Graham Barr, and Andrew F Laine. Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: The mesa copd study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 116–124. Springer, 2017.

[66] Arnab Maity, Patrick F Sullivan, and Jun-ing Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic epidemiology*, 36(7):686–695, 2012.

[67] Daniel J Schaid. Genomic similarity and kernel methods i: advancements by building on mathematical and statistical foundations. *Human heredity*, 70(2):109–131, 2010.

[68] Daniel J Schaid. Genomic similarity and kernel methods ii: methods for genomic information. *Human heredity*, 70(2):132–140, 2010.

[69] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *NIPS 20*, pages 585–592, Cambridge, MA, 2008. MIT Press.

[70] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Supplementary Information for "Unpaired Data Empowers Association Tests"

S1. Simulation Results of SAT-fx

Fig. 1 and Fig. 2 report the type I error rates and power of SAT-fx, i.e., the method in the fixed X setting, in simulation (i), respectively.

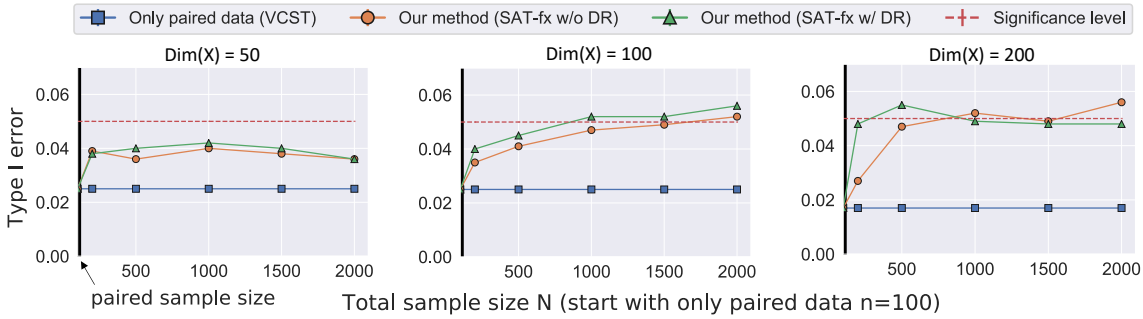

Figure 1: Evaluation of type I error rate control on the simulated data generated by simulation procedure (i) in the fixed X setting. The blue line (VCST) is the result of using only paired data; hence it does not change with addition of unpaired data. VCST only uses the $n = 100$ paired data points. Our methods (green and orange) start with n pairs and gradually adds unpaired data to improve type I error control. False-positive rates for both variants of our method SAT-fx are well controlled around the nominal value. (DR-Dimension Reduction)

S2. Details about the Imaging Feature

We adopt the feature extraction strategy proposed by Schabdach *et al.* [1]. They proposed an efficient method that summarizes the CT images of patients to a low dimensional representation. They applied their method on a large cohort with COPD disease and showed that the low-dimensional representation is highly predictive of the disease severity. The general idea is to use a non-parametric method to first compute the similarity between pairs of patients, then construct the low dimensional representation which can be used to predict disease severity. In the following, we first explain the image pre-processing pipeline followed by the method used to compute the patient-patient similarity and the patient-level low-dimensional representation.

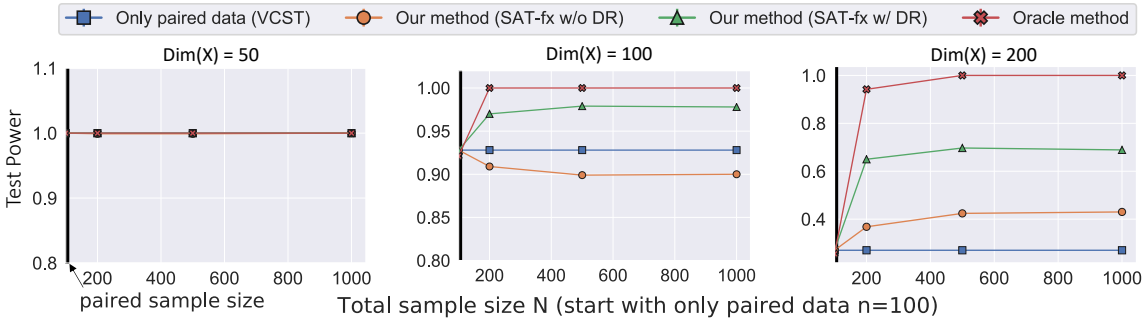

Figure 2: Evaluation of test power on the simulated data generated by procedure (i) in the fixed X setting. The results for heritability values $h^2 = 0.1$ and dimensionality $dim(X) = dim(Y) = 50, 100, 200$ are shown. VCST only uses the $n = 100$ paired data points. Our methods start with n pairs and gradually add unpaired data to improve test power.

Pre-processing Pipeline: We apply the method to lung Computer Tomography (CT) images of 7,292 subjects from the COPDGene study [2]. The general pipeline is shown in Figure 3. First, we use SLIC [3] to over-segment the lung volume into the spatially homogeneous region, which is called super-voxelization. We extract texture and intensity features from each super-voxel. For the texture features, we use a method proposed by Liu *et al.* [4] called Spherical Histogram of the Gradients. It uses spherical harmonics to compute the histogram of gradients of pixels belonging to a super-voxel on a unit sphere. For the histogram features, we extract a 32-bin histogram of the intensity values of the pixels in a super-voxel. The intensity value of the CT images is shown to be highly informative for characterizing emphysema [5]. In summary, we model each patient as a bag-of-words [6] where the words are d-dimensional ($d = 60$) features extracted from super-voxels of lung CT image of a patient.

Computing Pairwise Similarities and Patient-Level Low Dimensional Representation: Let us denote $S_i = \{x_{i1}, \cdots, x_{iN}\}$ ($S_j = \{x_{j1}, \cdots, x_{jN}\}$) be a set of all features from patient i (j) where N (M) represents total number of super-voxels of subjects i (j). We view x_{ik} as random variable drawn from an unknown patient-specific probability density p_i (*i.e.*, $x_{ik} \sim p_i$). Schabdach *et al.* [1] proposed to use Kullback-Leibler divergence (KL) between p_i and p_j as a measure of pairwise *dissimilarity* between image data of patient i and j. The KL divergence is defined as

$$\mathrm{KL}(p_i \| p_j) = \int_{\mathbb{R}^d} \log \frac{p_i(z)}{p_j(z)} p_i(x) dz. \tag{1}$$

Instead of assuming an explicit parametrization, we follow Poczos *et al.* [7] that use a non-parametric estimator for KL divergence that is consistent and unbiased. Instead of global parametrization for p_i and p_j, they parametrize the probability densities *locally*. Let $\rho_{k,S_i}(x) \triangleq \min_{v \in S_i} \|v - x\|_2$ denote the $1 - NN$ distance from x in a set S_i. Poczos *et al.* [7] proposed the following estimator for the KL,

$$\widehat{\mathrm{KL}(p_i \| p_j)} = \frac{d}{|S_i|} \sum_{v \in S_i} \log \frac{\rho_{k,S_i}(v)}{\rho_{k,S_j}(v)} + \log \frac{|S_j|}{|S_i| - 1},$$

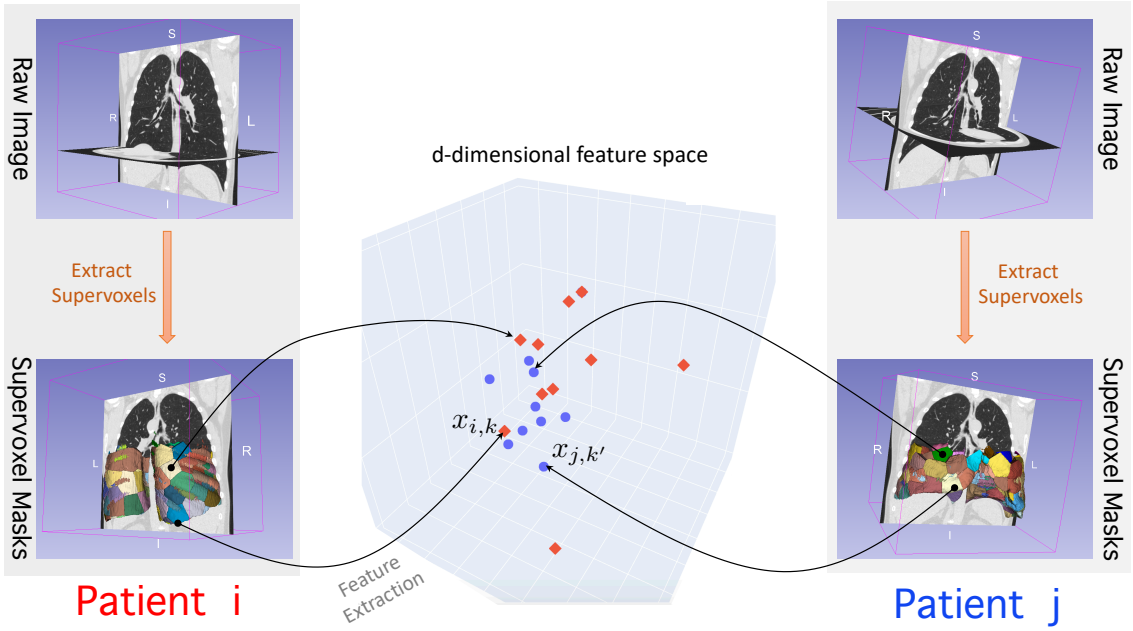

Figure 3: The schematic of feature extraction pipeline on lung imaging data. Each volumetric CT image is over-segmented (Extract Super-voxels). We extract a d-dimensional feature from each super-voxels. In this schematic, the blue circles (red rectangles) represent features extracted from super-voxels of the subject j (i).

where $|S_i|$ and $|S_j|$ are the sizes of the sets S_i and S_j and d is the dimensionality of the input features.

The similarity kernel matrix is a Positive Semi-Definite (PSD) matrix. However, the KL divergence is neither symmetric nor a proper metric. First, we compute a matrix where the entry in row i and column j is

$$[\tilde{L}]_{ij} = \exp\left(-\frac{1}{\sigma^2}\left(\widehat{\mathrm{KL}(p_i\|p_j)} + \widehat{\mathrm{KL}(p_j\|p_i)}\right)\right). \tag{2}$$

The σ is set to median of KL divergences (so-called median trick [8, 9]). Then, we project this matrix on the PSD cone to construct the kernel, $L = \mathrm{Proj}_{\mathrm{PSD}}(\tilde{L})$, where $\mathrm{Proj}_{\mathrm{PSD}}$ computes the Singular Value Decomposition of the input matrix and set the negative singular values to zero. Finally, we use the pairwise similarity kernel and apply Locally Linear Embedding (LLE) [10] to reduce the dimensionality and compute the patient-level feature representation.

S3. Details about the Phenotypes in the Uganda Cohort

Table 1 explains the detailed information about each phenotye in the Uganda Cohort.

Table 1: A description of the phenotypes measured in the Uganda cohort.

Phenotype	Category	Description
SBP	Blood pressure	Systolic blood pressure
DBP	Blood pressure	Diastolic blood pressure
BMI	Anthropometric index	Body mass index
WHR	Anthropometric index	Waist-hip ratio
Weight	Anthropometric index	Weight
Height	Anthropometric index	Height
HC	Anthropometric index	Hip circumference
WC	Anthropometric index	Waist circumference
ALT	Liver function	Alanine aminotransferase test
Albumin	Liver function	Serum albumin test
ALP	Liver function	Alkaline phosphatase test
AST	Liver function	Aspartate aminotransferase test
Bilirubin	Liver function	Bilirubin
GGT	Liver function	Gamma-glutamyl transpeptidase test
Cholesterol	Lipid test	Total cholesterol
HDL	Lipid test	High-density lipoprotein
LDL	Lipid test	Low-density lipoprotein
Triglycerides	Lipid test	Triglycerides
HbA1c2	Glycemic control	HbA1c2
WBC	Blood factor	White blood cell count
RBC	Blood factor	Red blood cell count
Hemoglobin	Blood factor	Hemoglobin
HCT	Blood factor	Hematocrit test
MCV	Blood factor	Mean corpuscular volume
MCH	Blood factor	Mean corpuscular hemoglobin
MCHC	Blood factor	Mean corpuscular hemoglobin concentration
RDW	Blood factor	Red blood cell distribution width
PLT	Blood factor	Platelet count
MPV	Blood factor	Mean platelet volume
NEUPr	Blood factor	Neutrophil percentage
LYMPHPr	Blood factor	Lymphocyte percentage
MONOPr	Blood factor	Monocyte percentage
EOSPr	Blood factor	Eosinophil percentage
BASOPr	Blood factor	Basophil percentage
EOS	Blood factor	Eosinophil count
LYMPH	Blood factor	Lymphocyte count
NEU	Blood factor	Neutrophil count
MONO	Blood factor	Monocyte count
BASO	Blood factor	Basophil count

S4. Comparison of VCST and KIT

Table 2 compares the test statistics and null distributions of VCST and KIT.

Table 2: Comparison of VCST and KIT. TS: Test Statistic. ND: Null distribution.

	Unbiased TS	Unbiased ND	Biased TS	Biased ND	Unpaired X	Unpaired Y
VCST	$\hat{\mathcal{S}}_n(\mathbf{K},\mathbf{L})$	$\lambda_i\hat{\eta}_j(z_{ij}^2-1)$	$\frac{1}{n^2}tr(\mathbf{K}\mathbf{H}_n\mathbf{L}\mathbf{H}_n)$	$\lambda_i\hat{\eta}_j z_{ij}^2$	✗	✓
KIT	$\hat{I}_n(\mathbf{K},\mathbf{L})$	$\lambda_i\eta_j(z_{ij}^2-1)$	$\frac{1}{n^2}tr(\mathbf{K}\mathbf{H}_n\mathbf{L}\mathbf{H}_n)$	$\lambda_i\eta_j z_{ij}^2$	✓	✓

S5. Derivation of VCST Test Statistic $\hat{\mathcal{S}}_n$ and the Null Distribution

Let us define $\vec{\mathbf{y}} = (Y_{11},\ldots,Y_{n1},\ldots,Y_{1p},\ldots,Y_{np})^\intercal$, $\vec{\boldsymbol{\mu}} = (\mu_1,\ldots,\mu_p)^\intercal\otimes\mathbf{1}_n$, $\vec{\epsilon} = (\epsilon_{11},\ldots,\epsilon_{n1},\ldots,\epsilon_{1p},\ldots,\epsilon_{np})^\intercal$, and $\vec{G} = (G_{11},\ldots,G_{n1},\ldots,G_{1p},\ldots,G_{np})^\intercal$, we can write the multivariate variance component model (Eq. 4 in the main text) as

$$\vec{\mathbf{y}} = \vec{\boldsymbol{\mu}} + \vec{G} + \vec{\boldsymbol{\epsilon}}, \tag{3}$$

where $\vec{G}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_g\otimes\mathbf{K})$, $\vec{\boldsymbol{\epsilon}}\sim\mathcal{N}(0,\tilde{\boldsymbol{\Sigma}}_\epsilon)$, and $\tilde{\boldsymbol{\Sigma}}_\epsilon = \boldsymbol{\Sigma}_\epsilon\otimes\mathbf{I}_n$. Therefore, we have $\vec{\mathbf{y}}\sim\mathcal{N}(\vec{\boldsymbol{\mu}},\mathbf{V})$, where $\mathbf{V} = \boldsymbol{\Sigma}_g\otimes\mathbf{K}+\tilde{\boldsymbol{\Sigma}}_\epsilon$. The corresponding restricted maximum likelihood (REML) is

$$L_{REML} = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|(\mathbf{I}_p\otimes\mathbf{1}_n)^\intercal\mathbf{V}^{-1}(\mathbf{I}_p\otimes\mathbf{1}_n)| - \frac{1}{2}(\vec{\mathbf{y}}-\vec{\boldsymbol{\mu}})^\intercal\mathbf{V}^{-1}(\vec{\mathbf{y}}-\vec{\boldsymbol{\mu}}). \tag{4}$$

According to previous studies [11,12], the score statistic evaluated at H_0 can be defined as

$$\begin{aligned}\hat{\mathcal{S}}_n &= tr\left(\frac{\partial L_{REML}}{\partial\boldsymbol{\Sigma}_g}\right)\Big|_{\boldsymbol{\Sigma}_g=0,\mu_i=\hat{\mu}_i,\boldsymbol{\Sigma}_\epsilon=\hat{\boldsymbol{\Sigma}}_Y}\\ &= \frac{1}{2}\left\{(\vec{\mathbf{y}}-\vec{\boldsymbol{\mu}})^\intercal\tilde{\boldsymbol{\Sigma}}_\epsilon^{-1}(\mathbf{I}_p\otimes\mathbf{K})\tilde{\boldsymbol{\Sigma}}_\epsilon^{-1}(\vec{\mathbf{y}}-\vec{\boldsymbol{\mu}}) - tr(\mathbf{P_0}(\mathbf{I}_p\otimes\mathbf{K}))\right\}\Big|_{\mu_i=\hat{\mu}_i,\boldsymbol{\Sigma}_\epsilon=\hat{\boldsymbol{\Sigma}}_Y},\end{aligned} \tag{5}$$

where $\mathbf{P}_0 = \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} - \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1}(\mathbf{I}_p\otimes\mathbf{1}_n)((\mathbf{I}_p\otimes\mathbf{1}_n)^\intercal\tilde{\boldsymbol{\Sigma}}_\epsilon^{-1}(\mathbf{I}_p\otimes\mathbf{1}_n))^{-1}(\mathbf{I}_p\otimes\mathbf{1}_n)^\intercal\tilde{\boldsymbol{\Sigma}}_\epsilon^{-1}\big|_{\boldsymbol{\Sigma}_\epsilon=\hat{\boldsymbol{\Sigma}}_Y}$. Equivalently, $\hat{\mathcal{S}}_n$ can be reformulated as follows

$$\hat{\mathcal{S}}_n(\mathbf{K},\mathbf{L}) = \frac{1}{2n^2}tr(\mathbf{K}\mathbf{H}\mathbf{L}\mathbf{H}) - \frac{1}{2n^3}tr(\hat{\boldsymbol{\Sigma}}_Y^{-1})tr(\mathbf{H}\mathbf{K}). \tag{6}$$

where $\mathbf{L} = \mathbf{Y}\hat{\boldsymbol{\Sigma}}_Y^{-2}\mathbf{Y}^\intercal$. To derive the null distribution of $\hat{\mathcal{S}}_n$, we reformulate the score statistic as $\hat{\mathcal{S}}_n(\mathbf{K},\mathbf{L}) = \frac{1}{2n^2}tr(\tilde{\mathbf{y}}^\intercal\tilde{\mathbf{K}}\tilde{\mathbf{y}}) - \frac{1}{2n^3}tr(\hat{\boldsymbol{\Sigma}}_\epsilon^{-1})tr(\mathbf{H}\mathbf{K})$, where $\tilde{\mathbf{y}} = \tilde{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}}(\vec{\mathbf{y}}-\vec{\boldsymbol{\mu}})\sim\mathcal{N}(0,\mathbf{I}_{np})$ and $\tilde{\mathbf{K}} = \tilde{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_p\otimes\mathbf{K})\tilde{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}}$. Let $\tilde{\eta}_1,\ldots,\tilde{\eta}_{np}$ be the eigenvalues of $\tilde{\mathbf{K}}/n$. The eigenvalues can be calculated from the eigenvalues of $\mathbf{K}$ and $\tilde{\boldsymbol{\Sigma}}_\epsilon$ by $\tilde{\eta}_{((j-1)*p+i)} = \lambda_i\hat{\eta}_j$, where λ_i are the eigenvalues of $\hat{\boldsymbol{\Sigma}}_Y^{-1}$. We then have $n\hat{\mathcal{S}}_n(\mathbf{K},\mathbf{L}) = \sum_{i=1}^p\sum_{j=1}^n\lambda_i\hat{\eta}_j(z_{ij}^2-1)$.

S6. Proof of Theorem 1

We first give a more formal statement of Theorem 1 in our main paper.

Theorem 1 (Formal). *Let* $\check{I} = \sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_i\eta_j(z_{ij}^2-1)$ *and* $\check{I}_m = \sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i\hat{\eta}_j(z_{ij}^2-1)$.

(1) Assume $\sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_i^{1/2}\eta_j^{1/2} < \infty$. *Then, as* $m\to\infty$, $\check{I}_m \xrightarrow{D} \check{I}$.

(2) $\mathbb{E}(\check{I}_m) = \mathbb{E}(\check{I})$. *For* $m > 1$ *and all* $\delta > 0$, *with probability* $1-\delta$, *for all* $\mathbb{P}_{XY}$,

$$|\mathbb{V}(\check{I}_m) - \mathbb{V}(\check{I})| \le \sqrt{\frac{864\max(\kappa_Y^2,\kappa_X^2)\log\frac{12}{\delta}}{m}}, \tag{7}$$

where η_j *be the eigenvalues of* C_X *(covariance of* $\phi(X)$*),* $\hat{\eta}_j$ *be the eigenvalues of* $\hat{C}_X$ *(empirical covariance of* $\phi(X)$*),* λ_i *be the eigenvalues of the* C_Y *(covariance of* $\psi(Y)$*), and* $\hat{\lambda}_i$ *be the eigenvalues of* $\hat{C}_Y$ *(empirical covariance of* $\psi(Y)$*), respectively, in descending order.* κ_Y, κ_X *are constants.*

Proof. (1) The proof of (1) in our Theorem 1 can be obtained by extending the proof of Theorem 1 in [13]. To prove $\check{I}_m \xrightarrow{D} \check{I}$, it suffices to prove

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{\lambda}_i\hat{\eta}_j - \lambda_i\eta_j)z_{ij}^2 \to 0 \tag{8}$$

and

$$tr(\hat{C}_Y)tr(\hat{C}_X) \to tr(C_Y)tr(C_X) \tag{9}$$

in probability as $m\to\infty$. The convergence of the covariance trace operator has been proved in [13], i.e., $tr(\hat{C}_Y)\to tr(C_Y)$ and $tr(\hat{C}_X)\to tr(C_X)$. According to the continuous mapping theorem [14], we can immediately obtain (9). To prove (8), we can first get an upper bound

$$\begin{aligned}
|\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{\lambda}_i\hat{\eta}_j - \lambda_i\eta_j)z_{ij}^2| \le& |\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2}(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})z_{ij}^2| \\
&+ |\sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_i^{1/2}\eta_j^{1/2}(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})z_{ij}^2| \\
\le& \left\{\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i\hat{\eta}_j z_{ij}^4\right\}^{1/2}\left\{\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})^2\right\}^{1/2} \\
&+ \left\{\sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_i\eta_j z_{ij}^4\right\}^{1/2}\left\{\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})^2\right\}^{1/2}.
\end{aligned} \tag{10}$$

According to Chebyshev's inequality, $\sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_i\eta_j z_{ij}^4$ is of $O_p(1)$. Since $\hat{\lambda}_i$, $\hat{\eta}_j$, and z_{ij} are independent, $\mathbb{E}\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i\hat{\eta}_j z_{ij}^4\right] = \sum_{i=1}^{p}\sum_{j=1}^{q}\mathbb{E}(\hat{\lambda}_i)\mathbb{E}(\hat{\eta}_j)\mathbb{E}(z_{ij}^4) = \mathbb{E}(tr(\hat{C}_Y))\mathbb{E}(tr(\hat{C}_X))\mathbb{E}(z_{ij}^4)$. Because $\mathbb{E}(tr(\hat{C}_Y))$ and $\mathbb{E}(tr(\hat{C}_X))$ are bounded, we also have that $\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{\lambda}_i\hat{\eta}_j z_{ij}^4$ is of $O_p(1)$ according to Chebyshev's inequality. The proof is complete if we show

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})^2 = o_p(1). \tag{11}$$

From $(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})^2 \leq |\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2}|(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} + \lambda_i^{1/2}\eta_j^{1/2}) = |\hat{\lambda}_i\hat{\eta}_j - \lambda_i\eta_j|$, we have

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{\lambda}_i^{1/2}\hat{\eta}_j^{1/2} - \lambda_i^{1/2}\eta_j^{1/2})^2 \leq & \sum_{i=1}^{p}\sum_{j=1}^{q}|\hat{\lambda}_i\hat{\eta}_j - \lambda_i\eta_j| \\
\leq & \sum_{i=1}^{p}\sum_{j=1}^{q}|\hat{\lambda}_i\hat{\eta}_j - \hat{\lambda}_i\eta_j| + \sum_{i=1}^{p}\sum_{j=1}^{q}|\hat{\lambda}_i\eta_j - \lambda_i\eta_j| \\
= & \sum_{i=1}^{p}\hat{\lambda}_i\sum_{j=1}^{q}|\hat{\eta}_j - \eta_j| + \sum_{j=1}^{q}\eta_j\sum_{i=1}^{p}|\hat{\lambda}_i - \lambda_i| \\
\leq & tr(\hat{C}_Y)\|\hat{C}_X - C_X\|_1 + tr(C_X)\|\hat{C}_Y - C_Y\|_1, \qquad (12)
\end{aligned}$$

where $\|\cdot\|_1$ is the trace norm. The last inequality makes use of generalized Hoffmann-Wielandt inequality. According to [15, Proposition 12], $\|\hat{C}_Y - C_Y\|_1 \to 0$ and $\|\hat{C}_X - C_X\|_1 \to 0$ in probability, then the proof completes.

(2) To prove (2) in our Theorem 1, we first introduce the following Theorem on deviation bounds for U-statistics [16], which was obtained by applying a bound from [17, p. 25].

Theorem S1. (**Deviation bound for U-statistics**). *A one-sample U-statistics is defined as the random variable:*

$$u := \frac{1}{(m)_r}\sum_{\mathbf{i}_r^m} g(x_{i1}, \ldots, x_{ir}), \qquad (13)$$

where g is the kernel of the U-statistic. If $a \leq g \leq b$, then for all $t > 0$ the following bound holds:

$$\mathbb{P}(u - \mathbb{E}_u[u] \geq t) \leq \exp\left(-\frac{2t^2\lceil m/r\rceil}{(b-a)^2}\right). \qquad (14)$$

Now we are ready to prove the following Lemma.

Lemma S1. *Given a random variable X with covariance in RHKS $C_X = \mathbb{E}_X[\phi(X)\otimes\phi(X)]$ and its empirical estimation $\hat{C}_X = \frac{1}{m-1}\sum_{i=1}^{m}\phi(x_i)\otimes\phi(x_i)$ from a sample $S_x = \{x_1, \ldots, x_m\}$. For all $\epsilon > 0$, we have*

$$\mathbb{P}(\big|\|\hat{C}_X\|_{HS}^2 - \|C_X\|_{HS}^2\big| \geq \epsilon) \leq 3\exp(-m\epsilon^2/54). \qquad (15)$$

Proof. We first write $\|C_X\|_{HS}^2$ in terms of kernels:

$$\|C_X\|_{HS}^2 = \mathbb{E}_{XX'}k(X,X')^2 + [\mathbb{E}_{XX'}k(X,X')]^2 - 2\mathbb{E}_X[\mathbb{E}_{X'}k(X,X')]^2. \qquad (16)$$

Similarly, we can also write $\|\hat{C}_X\|_{HS}^2$ in terms of kernels:

$$\begin{aligned}
\|\hat{C}_X\|_{HS}^2 = & \underbrace{\frac{1}{(m-1)^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i,x_j)^2}_{\text{ⓐ}} + \underbrace{\frac{1}{(m-1)^2m^2}[\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i,x_j)]^2}_{\text{ⓑ}} \\
& \underbrace{-\frac{2}{(m-1)^2m}\sum_{i=1}^{m}[\sum_{j=1}^{m}k(x_i,x_j)]^2}_{\text{ⓒ}}. \qquad (17)
\end{aligned}$$

We first expand $\mathbb{E}_{S_x}$ⓐ into

$$\frac{1}{(m-1)^2}\mathbb{E}_{S_x}\left[\sum_i K_{ii}^2 + \sum_{(i,j)\in \mathbf{i}_2^m} K_{ij}^2\right] = O(m^{-1}) + (1+O(m^{-1}))\mathbb{E}_{XX'}k(X,X')^2, \tag{18}$$

expand $\mathbb{E}_{S_x}$ⓑ into

$$O(m^{-1}) + \frac{1}{(m-1)^2m^2}\mathbb{E}_{S_x}\left[\sum_{(i,j,q,r)\in \mathbf{i}_4} K_{ij}K_{qr}\right] = O(m^{-1}) + (1+O(m^{-1}))[\mathbb{E}_{XX'}k(X,X')]^2, \tag{19}$$

and expand $\mathbb{E}_{S_x}$ⓒ into

$$\begin{aligned}&\frac{1}{m(m-1)^2}\mathbb{E}_{S_x}\left[\sum_i K_{ii}^2 + \sum_{(i,j)\in \mathbf{i}_2^m}(K_{ii}K_{ij}+K_{ij}K_{jj})\right] + \frac{1}{m(m-1)^2}\mathbb{E}_{S_x}\left[\sum_{(i,j,r)\in \mathbf{i}_3^m} K_{ij}K_{jr}\right]\\ &= O(m^{-1}) + (1+O(m^{-1}))\mathbb{E}_X[\mathbb{E}_{X'}k(X,X')]^2,\end{aligned} \tag{20}$$

where $\mathbf{i}_r^m$ is the set of all r-tuples drawn without replacement from {1,...,m} and $\mathbb{E}_{S_x}$ denotes the expectation w.r.t. m independent copies x_i drawn from $\mathbb{P}_X$. By omitting the terms that decay as $O(m^{-1})$ or faster, we have

$$\begin{aligned}&\mathbb{P}\{\|\hat{C}_X\|_{HS}^2 - \|C_X\|_{HS}^2 \geq \epsilon\}\\ &\leq \mathbb{P}\left\{\mathbb{E}_{XX'}k(X,X')^2 - \frac{1}{(m)_2}\sum_{\mathbf{i}_2^m} K_{ij}^2 \geq \frac{1}{3}\epsilon\right\}\\ &+ \mathbb{P}\left\{\mathbb{E}_X[\mathbb{E}_{X'}k(X,X')]^2 - \frac{1}{(m)_3}\sum_{(i,j,r)\in \mathbf{i}_3^m} K_{ij}K_{jr} \geq \frac{1}{6}\epsilon\right\}\\ &+ \mathbb{P}\left\{[\mathbb{E}_{XX'}k(X,X')]^2 - \frac{1}{(m)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4} K_{ij}K_{qr} \geq \frac{1}{3}\epsilon\right\}\end{aligned} \tag{21}$$

$$\begin{aligned}&\leq \exp(-m\epsilon^2/9) + \exp(-m\epsilon^2/54) + \exp(-m\epsilon^2/18), \text{ (using Theorem S1).}\\ &\leq 3\exp(-m\epsilon^2/54).\end{aligned} \tag{22}$$

Now we are ready to prove (2) in our Theorem 1. From the definition of $\check{I}$ and $\check{I}_m$, we have $\mathbb{E}(\check{I}) = \mathbb{E}(\check{I}_m) = 0$, $\mathbb{V}(\check{I}_m) = 2\sum_{i=1}^p\sum_{j=1}^q \hat{\lambda}_i^2\hat{\eta}_j^2 = 2\|\hat{C}_Y\|_{HS}^2\|\hat{C}_X\|_{HS}^2$, and $\mathbb{V}(\check{I}) = 2\sum_{i=1}^p\sum_{j=1}^q \lambda_i^2\eta_j^2 = 2\|C_Y\|_{HS}^2\|C_X\|_{HS}^2$. Then we have

$$\begin{aligned}\mathbb{V}(\check{I}_m) - \mathbb{V}(\check{I}) &= 2(\|\hat{C}_Y\|_{HS}^2\|\hat{C}_X\|_{HS}^2 - \|C_Y\|_{HS}^2\|C_X\|_{HS}^2)\\ &= 2(\|\hat{C}_Y\|_{HS}^2\|\hat{C}_X\|_{HS}^2 - \|\hat{C}_Y\|_{HS}^2\|C_X\|_{HS}^2 + \|\hat{C}_Y\|_{HS}^2\|C_X\|_{HS}^2 - \|C_Y\|_{HS}^2\|C_X\|_{HS}^2)\\ &= 2\|\hat{C}_Y\|_{HS}^2(\|\hat{C}_X\|_{HS}^2 - \|C_X\|_{HS}^2) + 2\|C_X\|_{HS}^2(\|\hat{C}_Y\|_{HS}^2 - \|C_Y\|_{HS}^2).\end{aligned} \tag{23}$$

Thus,

$$\begin{aligned}\mathbb{P}(\mathbb{V}(\check{I}_m) - \mathbb{V}(\check{I}) \geq \epsilon) &= \mathbb{P}(2\|\hat{C}_Y\|_{HS}^2(\|\hat{C}_X\|_{HS}^2 - \|C_X\|_{HS}^2) + 2\|C_X\|_{HS}^2(\|\hat{C}_Y\|_{HS}^2 - \|C_Y\|_{HS}^2) \geq \epsilon) \\ &\leq \mathbb{P}(2\|\hat{C}_Y\|_{HS}^2(\|\hat{C}_X\|_{HS}^2 - \|C_X\|_{HS}^2) \geq \epsilon/2) + \mathbb{P}(2\|C_X\|_2^2\big|\|\hat{C}_Y\|_2^2 - \|C_Y\|_2^2\big| \geq \epsilon/2) \\ &\leq \mathbb{P}(\|\hat{C}_X\|_{HS}^2 - \|C_X\|_{HS}^2 \geq \epsilon/(4\kappa_Y)) + \mathbb{P}(\|\hat{C}_Y\|_{HS}^2 - \|C_Y\|_{HS}^2 \geq \epsilon/(4\kappa_X)), \\ &\leq 3\exp(-m\epsilon^2/(864\kappa_Y^2)) + 3\exp(-m\epsilon^2/(864\kappa_X^2)) \\ &\leq 6\exp(-m\epsilon^2/(864\max(\kappa_Y^2, \kappa_X^2))). \end{aligned} \tag{24}$$

where $\|\hat{C}_Y\|_2^2 \leq \kappa_Y$ and $\|C_X\|_2^2 \leq \kappa_X$. By setting $6\exp(-m\epsilon^2/(864\max(\kappa_Y^2, \kappa_X^2))) = \delta$, we can solve for ϵ:

$$\epsilon = \sqrt{\frac{864\max(\kappa_Y^2, \kappa_X^2)\log\frac{6}{\delta}}{m}}. \tag{25}$$

Therefore, we have that with probability at least $1 - \delta$,

$$|\mathbb{V}(\check{I}_m) - \mathbb{V}(\check{I})| \leq \sqrt{\frac{864\max(\kappa_Y^2, \kappa_X^2)\log\frac{12}{\delta}}{m}}, \tag{26}$$

then the proof completes. □

S7. Derivation of the Null Distributions of Our SAT Test Statistics

Recall that our method SAT has two variants SAT-fx and SAT-rx, which are the extensions of VCST and KIT, respectively. Our SAT method basically project the original data into a subspace learned from unpaired data, and then plug in the projected data into the original VCST or KIT test statistics. The null distributions of our SAT-fx or SAT-rx test statistics follow the same forms as VCST or KIT, and differ in the number of χ^2 terms in the summation. In the following, we will derive the null distributions of SAT-fx and SAT-rx test statistics separately.

SAT-fx Let U be a the projection matrix containing first r_Y columns of the eigenvector matrix of $\hat{\mathbf{\Sigma}}_Y$, in which the eigenvectors are sorted according to their corresponding eigenvalues in a descent order. Let $\mathbf{Y}' = \mathbf{Y}U \in \mathbb{R}^{n\times r_Y}$. Here we define $\vec{\mathbf{y}}' = (Y'_{11}, \ldots, Y'_{n1}, \ldots, Y'_{1r_Y}, \ldots, Y'_{nr_Y})^\intercal$, $\vec{\boldsymbol{\mu}}' = (\mu'_1, \ldots, \mu'_{r_Y})^\intercal \otimes \mathbf{1}_n$, $\vec{\epsilon'} = (\epsilon'_{11}, \ldots, \epsilon'_{n1}, \ldots, \epsilon'_{1r_Y}, \ldots, \epsilon'_{nr_Y})^\intercal$. The covariance of Y' is defined as $\mathbf{\Sigma}'_Y$ and the covariance of $\vec{\epsilon'}$ is defined as $\tilde{\mathbf{\Sigma}}'_\epsilon = \mathbf{\Sigma}'_\epsilon \otimes \mathbf{I}_n$. According to the derivation of the null distribution of $\hat{\mathcal{S}}_n(\mathbf{K}, \mathbf{L})$ in Section S1, we reformulate the score statistic of our SAT-fx as $\hat{\mathcal{S}}_n(\mathbf{K}, \mathbf{L}') = \frac{1}{2n^2}tr(\tilde{\mathbf{y}}'^\intercal\tilde{\mathbf{K}}'\tilde{\mathbf{y}}') - \frac{1}{2n^3}tr(\hat{\tilde{\mathbf{\Sigma}}'}_\epsilon^{-1})tr(\mathbf{HK})$, where $\tilde{\mathbf{y}}' = \tilde{\mathbf{\Sigma}}'^{-\frac{1}{2}}_\epsilon(\vec{\mathbf{y}}' - \vec{\boldsymbol{\mu}}') \sim \mathcal{N}(0, \mathbf{I}_{nr_Y})$ and $\tilde{\mathbf{K}}' = \tilde{\mathbf{\Sigma}}'^{-\frac{1}{2}}_\epsilon(\mathbf{I}_{r_Y} \otimes \mathbf{K})\tilde{\mathbf{\Sigma}}'^{-\frac{1}{2}}_\epsilon|_{\mathbf{\Sigma}'_\epsilon = \mathbf{\Sigma}'_Y}$. Let $\tilde{\eta}'_1, \ldots, \tilde{\eta}'_{nr_Y}$ be the eigenvalues of $\tilde{\mathbf{K}}'/n$. The eigenvalues can be calculated from the eigenvalues of $\mathbf{K}$ and $\tilde{\mathbf{\Sigma}}'_\epsilon$ by $\tilde{\eta}_{((j-1)*r_Y+i)} = \lambda_i\hat{\eta}_j$, where $\lambda_1, \ldots, \lambda_{r_X}$ are the smallest r_Y eigenvalues of $\hat{\mathbf{\Sigma}}_Y^{-1}$. We then have $n\hat{\mathcal{S}}_n(\mathbf{K}, \mathbf{L}') = \sum_{i=1}^{r_Y}\sum_{j=1}^{n}\lambda_i\hat{\eta}_j(z_{ij}^2 - 1)$.

SAT-rx According to Mercer's theorem [18], the kernel functions $\tilde{k}(X, X')$ and $\tilde{l}(Y, Y')$ can be represented using eigenfunctions and eigenvalues defined in connection with them:

$$\tilde{k}(X, X') = \sum_{i=1}^{q} \eta_i \phi_i(X)\phi_i(X'), \quad \tilde{l}(Y, Y') = \sum_{i=1}^{p} \lambda_i \psi_i(Y)\psi_i(Y'). \tag{27}$$

To derive the asymptotic null distribution of $\hat{I}_n(\mathbf{K}', \mathbf{L}')$, it suffices to derive the null distribution of $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{k}'(\mathbf{x}_i, \mathbf{x}_j)\tilde{l}'(\mathbf{y}_i, \mathbf{y}_j)$, where

$$\tilde{k}'(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{r_X} \eta_k \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j), \quad \tilde{l}'(\mathbf{y}_i, \mathbf{y}_j) = \sum_{k=1}^{r_Y} \lambda_k \psi_k(\mathbf{y}_i)\psi_k(\mathbf{y}_j). \tag{28}$$

Using (28), we can rewrite U_n as

$$U_n = \frac{1}{n-1} \sum_{i=1}^{r_Y} \sum_{j=1}^{r_X} \lambda_i \eta_j \left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \phi_j(\mathbf{x}_k)\psi_i(\mathbf{y}_k) \right)^2 - \frac{1}{n-1} \sum_{i=1}^{r_Y} \sum_{j=1}^{r_X} \lambda_i \eta_j \frac{1}{n} \sum_{k=1}^{n} (\phi_j(\mathbf{x}_k)\psi_i(\mathbf{y}_k))^2. \tag{29}$$

Let $S^{ij} = \mathbb{E}_{XY}\phi_j(X)\psi_i(Y)$ and $S_n^{ij} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \phi_j(X_k)\psi_i(Y_k)$, $T^{ij} = \mathbb{E}_{XY}(\phi_j(X)\psi_i(Y))^2$, and $T_n^{ij} = \frac{1}{n} \sum_{k=1}^{n} (\phi_j(X_k)\psi_i(Y_k))^2$, where $\{(X_k, Y_k)\}_{k=1}^{n}$ are *i.i.d.* variables with the same distribution as (X, Y). Under the null hypothesis, $S^{ij} = 0$, the expectation of S_n^{ij} is

$$\mathbb{E}(S_n^{ij}) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \mathbb{E}_{X_k Y_k}[\phi_j(X_k)\psi_i(Y_k)] = \sqrt{n}\mathbb{E}_{XY}[\phi_j(X)\psi_i(Y)] = \sqrt{n}S^{ij} = 0, \tag{30}$$

and the variance of S_n^{ij} is

$$\mathbb{V}(S_n^{ij}) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_{X_k Y_k}[\phi_j(X_k)\psi_i(Y_k) - \sqrt{n}S^{ij}]^2 = [1 - \sqrt{n}(S^{ij})^2] = 1. \tag{31}$$

Similarly, the expectation of T_n^{ij} is

$$\mathbb{E}(T_n^{ij}) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_{X_k Y_k}[(\phi_j(X_k)\psi_i(Y_k))^2] = \mathbb{E}_{XY}[(\phi_j(X)\psi_i(Y))^2] = 1, \tag{32}$$

and the variance of T_n^{ij} is

$$\mathbb{V}(T_n^{ij}) = \frac{1}{n^2} \sum_{k=1}^{n} \{\mathbb{E}_{X_k Y_k}[\phi_j(X_k)\psi_i(Y_k)]^4 - 1\} = O(\frac{1}{n}). \tag{33}$$

Thus, we have $T_n^{ij} \xrightarrow{D} 1$ and $S_n^{ij} \xrightarrow{D} z_{ij}$, where z_{ij} are standard normal variables. Therefore,

$$nU_n \xrightarrow{D} \sum_{i=1}^{r_Y} \sum_{j=1}^{r_X} \lambda_i \eta_j (z_{ij}^2 - 1), \tag{34}$$

so does $\hat{I}_n(\mathbf{K}', \mathbf{L}')$.

S8. Proof of Theorem 2

Here we give a formal version of Theorem 2 in our main paper.

Theorem 2 (Formal). *We assume the following data generating process for X and Y:*

$$\begin{aligned}\phi(X) &= \mathbf{A}Z_x + n_x \\ \psi(Y) &= \mathbf{B}Z_y + n_y,\end{aligned} \tag{35}$$

where $\phi(X) \in \mathbb{R}^q, \psi(Y) \in \mathbb{R}^p, Z_x \sim \mathcal{N}(0, \mathbf{I}_{r_X}), Z_y \sim \mathcal{N}(0, \mathbf{I}_{r_Y}), \mathbf{A} \in \mathbb{R}^{q\times r_X}$, $\mathbf{B} \in \mathbb{R}^{p\times r_Y}$, $n_x \sim \mathcal{N}(0, \sigma_X^2\mathbf{I}_q)$, $n_y \sim \mathcal{N}(0, \sigma_Y^2\mathbf{I}_p)$, $r_X \le q$, $r_Y \le p$, and n_x is independent of n_y. Under the alternative hypothesis, we have

$$\mathbb{P}'(n\hat{I}'_n > q'_{1-\alpha}) \le \mathbb{P}(n\hat{I}_n > q_{1-\alpha}), \tag{36}$$

where $q_{1-\alpha}$ and $q'_{1-\alpha}$ are the $1-\alpha$ quantiles for the null distributions of $n\hat{I}_n$ and $n\hat{I}'_n$, respectively.

Proof. We first give the results on the asymptotic distribution of $\hat{I}_n$ in the following lemma, which can be obtained by applying [19, Theorem 5.5.1 (A)].

Lemma S2. *Let $\tilde{k}(X, X')$ and $\tilde{l}(Y, Y')$ be the centered kernel functions of $k(X, X')$ and $l(Y, Y')$, respectively. Assume that $\zeta_2 = \mathbb{V}[\tilde{k}(X, X')\tilde{l}(Y, Y')] < \infty$ and $\zeta_1 = \mathbb{V}\{\mathbb{E}_{X'Y'}[\tilde{k}(X, X')\tilde{l}(Y, Y')]\} > 0$. Under the alternative hypothesis ($I > 0$), we have $\sqrt{n}(\hat{I}_n - I) \xrightarrow{D} \mathcal{N}(0, 4\zeta_1)$.*

Now we derive the asymptotic distribution of $\hat{I}'_n$ under the alternative hypothesis. Our method can be considered as the original KIT method with new kernel functions $\tilde{k}'(X, X')$ and $\tilde{l}'(Y, Y')$ defined on the dimension-reduced inputs. Therefore, we have $\sqrt{n}(\hat{I}'_n - I') \xrightarrow{D} \mathcal{N}(0, 4\zeta'_1)$, where $I' = \mathbb{E}[\hat{I}'_n]$. According to Mercer's theorem [18], the kernel functions $\tilde{k}(X, X')$ and $\tilde{l}(Y, Y')$ can be represented using eigenfunctions and eigenvalues defined in connection with them:

$$\tilde{k}(X, X') = \sum_{i=1}^{q} \eta_i \phi_i(X)\phi_i(X'), \quad \tilde{l}(Y, Y') = \sum_{i=1}^{p} \lambda_i \psi_i(Y)\psi_i(Y'). \tag{37}$$

Similarly, the new kernels in the test statistic $\hat{I}'_n$ of our SAT-rx method can be represented as

$$\tilde{k}'(X, X') = \sum_{i=1}^{r_X} \eta_i \phi_i(X)\phi_i(X'), \quad \tilde{l}'(Y, Y') = \sum_{i=1}^{r_Y} \lambda_i \psi_i(Y)\psi_i(Y'). \tag{38}$$

By using the representations of the kernels, we can express I as

$$\begin{aligned} I &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'} \sum_{i=1}^{q} \eta_i \phi_i(X)\phi_i(X') \sum_{i=1}^{p} \lambda_i \psi_i(Y)\psi_i(Y') \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} \lambda_i \eta_j \mathbb{E}_{XY}\mathbb{E}_{X'Y'} \phi_j(X)\phi_j(X')\psi_i(Y)\psi_i(Y') \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} \lambda_i \eta_j [\mathbb{E}_{XY} \phi_j(X)\psi_i(Y)]^2. \end{aligned} \tag{39}$$

Similarly, $I' = \sum_{i=1}^{r_Y} \sum_{j=1}^{r_X} \lambda_i \eta_j [\mathbb{E}_{XY} \phi_j(X) \psi_i(Y)]^2$. According to the connection between principal component analysis and the factor analysis model [20,21], the first r_X principal component solutions correspond to the subspace spanned by $\mathbf{A}$ and the eigenfunctions of C_X associated with the $q - r_X$ smallest eigenvalues map the input X to the noise term n_x. Similarly, the eigenfunctions of C_Y associated with the $p - r_Y$ smallest eigenvalues map the input Y to the noise term n_y. Because n_x and n_y are independent, we have $\sum_{i=r_Y+1}^{p} \sum_{j=r_X+1}^{q} \lambda_i \eta_j [\mathbb{E}_{XY} \phi_j(X) \psi_i(Y)]^2 = 0$, which implies $I = I'$.

By using the same representation, we can derive the relation between ζ_1 and ζ_1'. Using (37), ζ_1 can be expanded as

$$\begin{aligned} \zeta_1 &= \mathbb{V}\{\mathbb{E}_{X'Y'}[\tilde{k}(X, X')\tilde{l}(Y, Y')]\} \\ &= \mathbb{V}\{\sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i \eta_j \phi_j(X) \psi_i(Y) \mathbb{E}_{X'Y'}[\phi_j(X')\psi_i(Y')]\}. \end{aligned} \tag{40}$$

Similarly, using (38), we have

$$\begin{aligned} \zeta_1' &= \mathbb{V}\{\mathbb{E}_{X'Y'}[\tilde{k}'(X, X')\tilde{l}'(Y, Y')]\} \\ &= \mathbb{V}\{\sum_{i=1}^{r_Y} \sum_{j=1}^{r_X} \lambda_i \eta_j \phi_j(X) \psi_i(Y) \mathbb{E}_{X'Y'}[\phi_j(X')\psi_i(Y')]\}. \end{aligned} \tag{41}$$

Because $r_X \leq q$, $r_Y \leq p$, we have $\zeta_1' \leq \zeta_1$. The power of the baseline KIT and our SAT methods at the significance level α can be calculated as $\mathbb{P}(n\hat{I}_n > q_{1-\alpha}) = \Phi(nI - \frac{q_{1-\alpha}}{2}\sqrt{n\zeta_1})$ and $\mathbb{P}'(n\hat{I}_n' > q_{1-\alpha}') = \Phi(nI - \frac{q_{1-\alpha}'}{2}\sqrt{n\zeta_1'})$, respectively, where $\Phi(\cdot)$ is the CDF of a standard normal distribution, $q_{1-\alpha}$ is the $1-\alpha$ quantile of $\sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i \eta_j (z_{ij}^2 - 1)$, and $q_{1-\alpha}'$ is the $1-\alpha$ quantile of $\sum_{i=1}^{r_Y} \sum_{j=1}^{r_X} \lambda_i \eta_j (z_{ij}^2 - 1)$. Because $r_X \leq q$ and $r_Y \leq p$, we have $q_{1-\alpha} \geq q_{1-\alpha}'$, and further because of $\zeta_1' \leq \zeta_1$, $\mathbb{P}'(n\hat{I}_n' > q_{1-\alpha}') \geq \mathbb{P}(n\hat{I}_n > q_{1-\alpha})$. The proof completes.

□

References

[1] Jenna Schabdach, William M Wells, Michael Cho, and Kayhan N Batmanghelich. A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies. In *International Conference on Information Processing in Medical Imaging*, pages 170–183. Springer, 2017.

[2] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

[3] Markus Holzer and Rene Donner. Over-segmentation of 3d medical image volumes based on monogenic cues. *Proceedings of the CVWW*, 14:35–42, 2014.

[4] Kun Liu, Henrik Skibbe, Thorsten Schmidt, Thomas Blein, Klaus Palme, Thomas Brox, and Olaf Ronneberger. Rotation-invariant hog descriptors using fourier analysis in polar and spherical coordinates. *International Journal of Computer Vision*, 106(3):342–364, 2014.

[5] Lauge Sorensen, Mads Nielsen, Pechin Lo, Haseem Ashraf, Jesper H Pedersen, and Marleen De Bruijne. Texture-based analysis of copd: a data-driven approach. *IEEE transactions on medical imaging*, 31(1):70–78, 2011.

[6] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):591–606, 2008.

[7] Barnabs Poczos, Liang Xiong, and Jeff Schneider. Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions. *Uncertainty in Artificial Intelligence*, 2011.

[8] Kenji Fukumizu, Arthur Gretton, Gert R Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in neural information processing systems*, pages 1750–1758, 2009.

[9] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

[10] Zhenyue Zhang and Jing Wang. Mlle: Modified locally linear embedding using multiple weights. In *Advances in neural information processing systems*, pages 1593–1600, 2007.

[11] Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

[12] Arnab Maity, Patrick F Sullivan, and Jun-ing Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic epidemiology*, 36(7):686–695, 2012.

[13] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS 23*, pages 673–681, Cambridge, MA, 2009. MIT Press.

[14] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[15] Moulines Eric, Francis R Bach, and Zaïd Harchaoui. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2008.

[16] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[17] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[18] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209:415–446, 1909.

[19] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[20] Gale Young. Maximum likelihood estimation and factor analysis. *Psychometrika*, 6(1):49–53, 1941.

[21] Peter Whittle. On principal components and least square methods of factor analysis. *Scandinavian Actuarial Journal*, 1952(3-4):223–239, 1952.