Techniques to Produce and Evaluate Realistic Multivariate Synthetic Data

John Heine[1*], Erin E.E. Fowler[1], Anders Berglund[2], Michael J. Schell[2], Steven Eschrich[2]

Moffitt Cancer Center & Research Institute, 12901 Bruce B. Downs Blvd, Tampa FL, 33612 (MCC)
[1]Cancer Epidemiology Department, MCC
[2]Department of Biostatistics and Bioinformatics, MCC
[*]Corresponding Author

Email Addresses
JH: john.heine@moffitt.org
EEEF: erin.fowler@moffitt.org
AB: anders.berglund@moffitt.org
MJS: michael.schell@moffitt.org
SE: steven.eschrich@moffitt.org

Abstract

Background: Proper data modeling in biomedical research requires sufficient data for exploration and reproducibility purposes. A limited sample size can inhibit objective performance evaluation.

Objective: We are developing a synthetic population (SP) generation technique to address the limited sample size condition. We show how to estimate a multivariate empirical probability density function (pdf) by converting the task to multiple one-dimensional (1D) pdf estimations.

Methods: Kernel density estimation (KDE) in 1D was used to construct univariate maps that converted the input variables (X) to normally distributed variables (Y). Principal component analysis (PCA) was used to transform the variables in Y to the uncoupled representation (T), where the univariate pdfs were assumed normal with specified variances. A standard random number generator was used to create synthetic variables with specified variances in T. Applying the inverse PCA transform to the synthetic variables in T produced the SP in Y. Applying the inverse maps produced the respective SP in X. Multiple tests were developed to compare univariate and multivariate pdfs and covariance matrices between the input (sample) and synthetic samples. Three datasets were investigated (n = 667) each with 10 input variables.

Results: For all three datasets, both the univariate (in X, Y, and T) and multivariate (in X, Y, and T) tests showed that the univariate and multivariate pdfs from synthetic samples were statistically similar to their pdfs from the respective samples. Application of several tests for multivariate normality indicated that the SPs in Y were approximately normal. Covariance matrix comparisons (in X and Y) also indicated the same similarity.

Conclusions: The work demonstrates how to generate multivariate synthetic data that matches the real input data by converting the input into multiple 1D problems. The work also shows that it is possible to convert a multivariate input pdf to a form that approximates a multivariate normal, although the technique is not dependent upon this finding. Further studies are required to evaluate the generalizability of the approach.

Keywords: synthetic data, multivariate probability density function estimation, principal component analysis, kernel density estimation, and differential evolution

Introduction

Data modeling in biomedical research requires sufficient data for exploration and reproducibility purposes. Unfortunately, data can be limited for variety of reasons: the study of low-incidence diseases or underserved/underrepresented subpopulations [1]; the inability to share data across facilities [2]; cost of molecular tests; and study timeframes.

Biomedical modeling encompasses a wide area of applications. Examples of such modeling include risk prediction [3], response therapy [4] and benign malignant classification [5-7]. In general, there is a common path for developing models intended for practical use. The initial phase of model building often includes analyzing data from the target population. This phase can be used for a variety of objectives with different degrees of endpoint focus, for example: exploration purposes by studying many different data characteristics searching for correlations and patterns; testing the feasibility of a hypothesis by investigating univariate or multivariate relationships; selecting the most suitable model; or determining model parameters with the appropriate optimization procedure. These analyses can effectively deplete the sample. In the final validation phase, the fully specified model (i.e., the model with its parameters fixed) is then evaluated with new data to demonstrate its generalizability. This last step depends critically on having an adequate sample size.

Determining sample size sufficiency in the multivariate setting has many nuances and is a difficult task [8]. The dimensionality (d) and samples size (n) are the two important components in this process. For a two sample test with normally distributed data and common covariance, Hotelling's T^2 is appropriate for large samples (i.e. large n relative to d) when comparing mean vectors [9], noting that this test is not optimal for problems that do not meet the sample size requirement or when the endpoint analysis is something other than comparing multivariate means. Moreover, sample size adequacy can be a function of the modeling technique as well as the number of variables considered. For example, a logistic regression model has on the order of d adjustable parameters when considering d covariates. In contrast, partial least squares (PLS) regression [10] when taken to the limit has roughly d^2 parameters, and deep neural network architectures can require large sample sizes for a given design (many adjustable parameters) for training, testing, and validation [11] (see related table in [11] for examples). The covariance structure can also influence the needed sample size as well. In a broad sense, when the variables under consideration tend to higher correlation, the needed sample size decreases and vice versa. It is our premise that the limiting sample size for a given multivariate prediction problem that permits independent validation is not well-defined in general. Our objective is to develop a synthetic data generation approach to address the problem where n is small relative to d by generating synthetic samples. Such an approach could be used to augment an inadequate sample size or to study sample size requirements for a given covariance structure.

Synthetic data applications in health-related research use a variety of techniques. Some methods are used for generating samples from large populations [12-15]. These approaches include hidden Markov models [14], techniques that reconstruct time series data coupled with sampling the empirical probability density function of the relevant variables [12], and methods that estimate pdfs from the data, not accounting for variable correlation [15]. Other work used moment matching to generate synthetic data but does not consider relative frequencies in the comparison analysis [16]. Comments on synthetic data techniques [16, 17] indicate that the small-sample size condition has received little attention.

We are developing a methodology to produce multivariate synthetic populations (SPs) [18] specifically to address the small sample size condition. The approach is based estimating the underlying multivariate probability density function for arbitrarily distributed data, where normal approximations fail to hold. Under this scenario, access to the population is not possible or practical. Our initial work, based on multivariate kernel density estimation (KDE) with an unconstrained bandwidth, was illustrated with a 5-dimensional problem [18]. Multivariate KDE has noted efficiency problems when considering high dimensionality [19, 20]. As the dimensionality increases, the sample size requirement apparently becomes exceeding large, although this area is under investigation. Although categorizing a problem as high or low dimensionality may be dependent on many factors, reasonable arguments suggest that a *high dimensionality* may be defined as $3 < d \leq 50$, where d is the number of variables considered [21]; in that, density estimators should be able to accommodate this range.

In this report, we present modifications to our methodology to mitigate the KDE efficiency problem and address synthetic data generation in higher dimensionality (i.e., $d = 10$). A key feature of this modified approach is that it converts a multivariate KDE problem to multiple univariate KDE problems while retaining the covariance structure. There are two closely related parallel characteristics investigated in this report. First, will show how to generate multivariate synthetic data relatively efficiently (characteristic 1), but it does not require a fully specified multivariate kernel (i.e., a kernel with a determined bandwidth matrix). The lack of kernel specification is not required for generating synthetic data to address the small sample problem directly with our modified approach but could have a drawback relative to KDE. Kernel specification may be useful for theoretical manipulations because it (the kernel) contains expressions for continuous variables that can be manipulated (e.g., integration and differentiation can be performed with kernel expressions). Secondly, the approach appears capable of transforming and multivariate probability density function with an arbitrary form to one that approximates a multivariate normal form (i.e., transforming the problem to parametric model). This transformation overcomes the theoretical limitations associated with the lack of multidimensional kernel specification and understanding when this approximation holds is referred to as characteristic 2. These two characteristics are essentially embedded, but characteristic 1 does not depend upon characteristic 2 with our approach. We have developed/applied multiple statistical tests to evaluate both characteristics. Our innovation includes (1) conditioning the input variables with univariate transforms so that known techniques can be applied to generate synthetic data, and (2) developing multiple statistical tests for assessing similarity.

2. Methods

2.1 Overview

A brief outline of our SP generation approach is provided followed by an in-depth treatment of the methods.

Definitions: *Population* is used to define a hypothetical collection of virtually *unlimited* number of either real or synthetic individuals except when explaining differential evolution (DE) optimization [22], where the DE-population is limited and defined specifically. *Sample* defines a collection of n real individuals (observations) with d attributes defined as measured variables. A column vector is designated with a lower-case bold letter. For example, attributes are referred to as $\mathbf{x}$, a column vector. Vector components are designated with lower-case subscripted letters. For example, the components of $\mathbf{x}$ are referenced as x_j for $j = 1, 2, 3, \ldots, d$. The associated multivariate probability function (pdf) for $\mathbf{x}$ is designated as $p(\mathbf{x})$, where the components of $\mathbf{x}$ are assumed to be continuous. Variables are measured in the X representation (i.e., realizations of $\mathbf{x}$ are measured in X). We assume $p(\mathbf{x})$ exists at the population level although access to it may be limited. In practice, we evaluated normalized histograms (empirical pdfs) that are also referred to as pdfs for brevity. The one-dimensional (1D) marginal pdfs for $p(\mathbf{x})$ are expressed as $p_j(x_j)$. Matrices are designated with upper case bold lettering. For example, $\mathbf{X}$ is the $n \times d$ matrix with n realizations of $\mathbf{x}$ in its rows (i.e., the j^{th} column of $\mathbf{x}$ contains n realizations of x_j). To reference individual realizations, we use double subscripts to indicate both the realizations and component indices. For example, x_{ij} is the j^{th} component for the i^{th} realization in X (also the indexing for $\mathbf{X}$). The work involves mapping $\mathbf{x}$ to the Y representation. In Y, this creates the corresponding (1) $\mathbf{y}$ (column vectors) with d components, (2) the multivariate pdf $g(\mathbf{y})$ and its marginal pdfs, and (3) the matrix $\mathbf{Y}$ that are all defined the same way as the entities in the X representation. We also work in the T representation (explained below), where $\mathbf{t}$, t_j, and $\mathbf{T}$ are defined in the same way. The multivariate pdf in T is expressed as $r(\mathbf{t})$ with marginal pdfs expressed analogously. We define the cumulative probability functions (i.e., the indefinite integral of a univariate pdf) for $p_j(x_j)$ and $g_j(y_j)$ as $P_j(x_j)$ and $G_j(y_j)$, respectively. Covariance quantities are estimated from the normal multivariate form: $E(w - m_w)(v - m_v)$, where E is the expectation operator, w and v are random variables with means m_w and m_v. The corresponding covariance matrices in each representation are expressed as $\mathbf{C}_x$, $\mathbf{C}_y$, and $\mathbf{C}_t$, respectively.

Approach Summary: From experience, an individual characteristic, such as body mass index (BMI), or measurements taken from image data can have right-skewed pdfs because they are often positive valued and not inclusive of zero (see Figure 1-3 for examples). Measures of this type can bear varying levels of correlation (see lower parts of Table 1-3). Thus, p(**x**) in general does not lend itself to parametric probability modeling in the X representation. To render this data into a representation that overcomes the KDE dimensionality problem, a series of steps were taken to condition **X** from the sample into a form that permits generating multivariate data without multivariate KDE as illustrated in Figure 4. The process starts with the input data in X (Figure 4, top-left). The sample processing flow (X-Y-T) is illustrated in the top row of Figure 4 and the reversed SP generation flow (T_s-Y_s-X_s) in the bottom row.

Step 1: Univariate maps were constructed, referenced as $m_j(x_j)$, to transform x_j to univariate zero mean and unit variance normally distributed [i.e., $y_j = m_j(x_j)$], producing the Y representation (Figure 4, top-middle).

Step 2: Principal component analysis (PCA) was used to transform the variables in Y to the T representation (Figure 4, top-right), where t_j, formed by linear combinations of y_j, are uncorrelated and distributed as $r_j(t_j)$.

Step 3: Synthetic data was generated (Figure 4, bottom-left) with a standard random number generator by approximating $r_j(t_j)$ as univariate zero mean normal pdfs with specified variances creating t_j defining the SP in T (T_s).

Step 4: The inverse PCA transform (Figure 4, bottom-middle) was applied to synthetic t_j creating synthetic y_j defining the SP in Y (Y_s) and restoring the covariance relationships as observed in Y that were removed in Step 2.

Step 5: Synthetic y_j were mapped to the X representation, inverse maps with $x_j = m_j^{-1}(y_j)$, (Figure 4, bottom right) reversing Step 1 creating synthetic x_j defining the SP in X (X_s) and restoring the covariance relationships as observed in X.

In Step 1, univariate maps were generated to transform the observed data pdfs to zero mean, unit variance normal pdfs. Note that, univariate KDE was used to generate synthetic x_j, augmenting the sample size by filling in gaps, to address the small sample size problem. Synthetic x_j generated in this fashion do not maintain the covariance structure and thus were not used further; only sample x_j were mapped. Augmenting the sample size, ensured sample $g_j(y_j)$ were smooth and identically distributed as standardized normal pdfs (i.e., producing a set of identical univariate marginals). KDE was not used beyond this point. After the mapping, characteristic 2 emerges. Here, we have a set of univariate normal marginals. There is no guarantee this set of pdfs will combine to form a multivariate normal pdf, although the reverse is always true (a multivariate normal has univariate marginals). In Step 2, the normalized y_j were linearly transformed with PCA to form t_j (the T representation). In the T representation, sample t_j are uncorrelated. In Step 3, we assumed that $r(t_j)$ were normal with variances given by the j^{th} eigenvalue of $\mathbf{C}_y$. To generate synthetic data, the columns of the matrix $\mathbf{T}_s$ were generated as normally distributed random variables with the specified variances (i.e., the SP in T). In Step 4, the inverse PCA transform was applied to $\mathbf{T}_s$ producing $\mathbf{Y}_s$. $\mathbf{Y}_s$ is the SP in Y, where each row represents a synthetic individual with d attributes. In Step 5 to complete the analysis (characteristic 1), the variables were inverse mapped to X creating $\mathbf{X}_s$ and the respective entities [$p_j(y_j)$, $p(\mathbf{x})$ and $\mathbf{C}_x$] were compared from the sample with those generated from synthetic samples. The SP in X ($\mathbf{X}_s$) has the same synthetic realization arrangement as $\mathbf{Y}_s$.

We evaluated how well the inverse PCA transformation preserved the covariance in Y ($\mathbf{C}_y$) and the method's capability of restoring the univariate/multivariate pdfs in both X and Y. To address characteristic 1, we tested $r_j(t_j)$ for normality and compared $g_j(y_j)$, $g(\mathbf{y})$ and $\mathbf{C}_y$ from the sample with those created with synthetic samples to assess the fidelity of the SP. To address characteristic 2, $g(\mathbf{y})$ was tested for multivariate normality. For reference, the multivariate normal pdf (all components with unit variance) with zero mean for **y** is expressed as

$$g_n(\mathbf{y}) = (2\pi)^{-\frac{d}{2}} |\mathbf{C}_y|^{-\frac{1}{2}} \exp - \frac{1}{2} [\mathbf{y}^T \mathbf{C}_y^{-1} \mathbf{y}], \qquad \text{Eq. [1]}$$

where $\mathbf{C}_y$ is the covariance matrix formed with sample y_j. The fidelity of characteristic 1 is not dependent upon characteristic 2 but the reverse is true. For characteristic 1, if synthetic samples are poor replicas of the sample, then g(y) will be a poor approximation of multivariate normality (i.e., characteristic 2). We believe that the Y representation is best suited for modeling because the marginals are normal. It is common practice in univariate/multivariate modeling to adjust variables (univariately) to a standard form (zero mean and unit variance) or apply transforms to remove skewness.

2.2 Data and measurements

Data used for this work was derived from two sources: (1) mammograms and other data from women undergoing mammography (n = 667), and (2) measurements from dried beans (n = 13611) [23]. For this report, the clinical and technical aspects of these data are not relevant. Data was assembled for generalization purposes. From mammograms, we considered two sets of measurements each with d = 10 variables (i.e., x_j) referred to as Dataset 1 (DS1) and Dataset 2 (DS2). DS1 has 8 double precision measurements from the Fourier power spectrum of two-dimensional mammograms in addition to age, and body mass index (BMI), both captured as integer variables. The Fourier measures are from a set of measurements described previously [24]; the first 8 measurements from this set are labeled as P_i for i = 1 to 8. The Fourier measures are consecutive following an approximate functional form [25] and thus represent components in **x** that are very different than the components in DS2 and DS3. DS1 represents variables with relatively high correlation and varied dynamic range. To cite the covariance quantities and correlation coefficients, we used a modified covariance table format. In these tables, entries below the diagonal are the respective correlation coefficients, whereas the other elements (diagonal and above) are conventional covariance quantities. The modified covariance table for DS1 is shown in Table 1a. DS2 contains 8 double precision measurements also derived from mammograms: mean, standard deviation (SD); high-pass (HP) filter output, low-pass (LP) filter output; local SD summarized; P_{20} Fourier measure (from the set described for DS1 measurements); local correlation summarized; and breast area measured in cm^2 [26]. Age and BMI (from DS1) were also included in this dataset. The modified covariance table (same format as Table 1a) is shown in Table 2a. Selected measures from the dried bean dataset [23] are referred to as Dataset 3 (DS3). The bean dataset has 17 measurements (floating point) related to 7 bean types. We selected 10 measures at random to make the dataset dimensionality compatible with the other datasets giving this set of variables: area (1); minor axis (5); eccentricity (6); convex area (7); equivalent diameter (8); extent (9); solidarity (10); roundness (11); shape factor 3 (15), and shape factor 4 (16). The parenthetical references give the variable number as listed in the respective resource (see [23]). We selected the bean type at random (Sira, with n = 2636) and selected n = 667 observations at random. The modified covariance table is shown in Table 3a. The analysis of three datasets supports the evaluation of the optimization scheme under varied scenarios.

2.3 KDE and mapping

The maps from X to Y are expressed as $y_j = m_j(x_j)$, where m_j is the j^{th} map that operates on x_j. One- dimensional (1D) KDE was used to generate realizations from $p_j(x_j)$ with our approach described previously [18], reduced to 1D. Synthetic realizations were used to augment the sample size to ensure that the maps were smooth (i.e., to remove staircasing effects). Univariate KDE for a given x_j is expressed as

$$p_j(x_j) = \frac{1}{k_j n}\sum_{i=i}^{n} \exp - \frac{(x_j - x_{ij})^2}{2h_j^2} = \frac{1}{n}\sum_{i=1}^{n} K(x_j) \qquad \text{Eq. [2]}$$

where x_j is the respective synthetic variable, k_j is the normalization factor, and h_j is the univariate bandwidth parameter used for synthesizing x_j that must be estimated. Differential evolution optimization [22] was used to determine h_j as described in our previous work [18]; here we apply the univariate version of this approach for each h_j using x_{ij} from the sample. To form generation -zero, each DE-population (n = 1000) was initialized randomly (uniformly distributed) within these bounds: (0, 4 × the variance of x_j]. For a given x_j, this population contains candidate h_j. As a given DE population evolves over several generations, it becomes more fit as measured by the fitness function. As a modification from our previous work, we used the Kolmogorov Smirnov (KS) test [27] as the fitness function. This is a nonparametric test that can be used to compare two numerically derived cumulative probability functions or compare a numerically derived curves to a reference [27]. Here we compare the respective numerical univariate cumulative probability functions derived from synthetic samples with those derived from the sample. The test measure, D_j, for the KS test is the absolute maximum difference between the two cumulative probability functions. The DE optimization is based on minimizing D_j for each variate. Each DE process was allowed evolve for 30 generations (i.e., the terminal generation). For the mapping, we solve $P_j(x_j) = G_j(y_j)$ [i.e., the inverse probability transform problem] for x_j and y_j with interpolation methods described previously [28, 29] to determine $m_j(x_j)$. The corresponding inverse maps are expressed as: $x_j = m_j^{-1}(y_j)$.

2.4 Synthetic data generation

Synthetic data was generated in the uncoupled T representation derived from applying a PCA transformation to sample **y**. After applying the mappings, $y_j = m_j(x_j)$, y_i are assembled into the **Y** matrix **(**analogous to **X**), where the d components of the **y** vectors form the rows of **Y**. The PCA transform can be expressed as

$$\mathbf{T} = \mathbf{YP}, \qquad \text{Eq. [3]}$$

where **P** is a d×d matrix with uncorrelated normalized columns (i.e., normalized eigenvectors of the covariance matrix of $\mathbf{C}_y$). The rows of **T** are components of **t** (i.e., the components of the column vector for the ith individual in the T representation). The covariance matrix, $\mathbf{C}_t$, is diagonal, with elements $c_{jj} = \sigma_j^2(t)$ (i.e., the ordered eigenvalues of $\mathbf{C}_y$). Thus, we make the approximation that r(**t**) (from the sample) has the multivariate normal form

$$r(\mathbf{t}) = (2\pi)^{-\frac{n}{2}} |\mathbf{C}_t|^{-\frac{1}{2}} \exp -\frac{1}{2}[\mathbf{t}^\mathbf{T} \mathbf{C}_t^{-1} \mathbf{t}] \qquad \text{Eq. [4]}$$

If the multivariate normality approximately holds in Y, it should also hold in T. The Eq. [4] approximation will be evaluated. Synthetic t_j were generated as independent normally distributed random variables using the Box-Muller method with zero mean and variance = $\sigma_j^2(t)$. Sample t_j were compared with normality (comparison with synthetic t_j). Synthetic t_j were then assembled in $\mathbf{T}_s$. Synthetic variables in the Y representation were derived by applying the inverse PCA transform substituting $\mathbf{Y}_s$ for **Y** giving

$$\mathbf{Y}_s = \mathbf{T}_s \mathbf{P}^\mathbf{T} \qquad \text{Eq. [5]}$$

where $\mathbf{Y}_s$ is the SP in Y. For synthetic data, $g(\mathbf{y}) = g_n(\mathbf{y})$ by definition. Synthetic and sample y_j were compared. In the final analysis, synthetic y_j were mapped to the X representation (SP in X) and compared with the sample. An Example is provided to illustrate that the approach filled the multivariate space, in contrast with the sparse sample.

2.5 Statistical Analysis

The main goal is to produce synthetic data that are statistically similar to the respective samples based on univariate/multivariate pdf and covariance matrix comparisons. We also evaluated the univariate and multivariate normal assumptions with multiple techniques. Two tests were used to evaluate the similarity of covariance matrices between the sample and synthetic samples.

Univariate pdf comparisons: The KS test was used for all univariate pdf comparisons. In the X representation, $p_j(x_j)$ from the sample were compared with normality and with the respective pdfs derived from synthetic samples of the same size ($n = 667$). In the Y representation, $g_j(y_j)$ determined from the mapped sample were compared with the respective pdfs derived from synthetic samples ($n = 667$) created from the inverse PCA transform of synthetic t_j; this implicitly evaluated univariate normality in Y, because synthetic y_j were derived from a multivariate normal process in the T representation (i.e., its marginals are normal). In the T representation, we compared $r_j(t_j)$ from the sample, after mapping x_j to y_j followed by the PCA transform, with the respective pdfs derived from zero mean normally distributed random variables (i.e., synthetic t_j) with variances $= \sigma_j^2(t)$. For all univariate pdf comparisons, we selected the KS test threshold based on the 5% significance level as the critical value. For each component, the sample was compared with 1000 synthetic samples, and the percentage of times that measured D less than the critical test value was tabulated.

Multivariate pdf comparisons: To test whether the sample and synthetic samples were drawn from the same distribution, we used the maximum mean discrepancy (MMD) test (nonparametric) developed by Gretton et al. [30]. This is a kernel-based (normal kernel) analysis that computes the difference between every possible vector combination between and within two samples (excluding same vector comparisons). To determine the kernel parameter for these tests, we used the median heuristic [30, 31]. This test is based on the critical value ($\mathrm{MMD_c}$) at the 5% significance level and the test statistic ($\mathrm{MMD}_\mathrm{u}^2$). Both quantities are calculated from the two samples under comparison This test has acceptance region given the two distributions are the same: $\mathrm{MMD}_\mathrm{u}^2 < \mathrm{MMD}_\mathrm{c}$ (see Theorem 10 and Corollary in reference [30]). This test was also applied to data from in X, Y, and T. Note, when applying this test in T and Y, it is implicitly testing the sample's similarity with multivariate normality in both representations. In X, Y, and T, 1000 synthetic samples were compared with the sample. The test acceptance percentage was tabulated. $\mathrm{MMD}_\mathrm{u}^2$ and MMD_c values are provided as averages over all trials because they change per comparison.

Additional tests were used to evaluate multivariate normality in the sample and in synthetic samples. These were applied in T and Y. Each test evaluates different distribution characteristics. The first test is based on random projections, as explained generically. The vector **w** with d components is multivariate normal if the scalar random variable, $z = \mathbf{u}^T\mathbf{w}$, is univariate normal, where **u** is a d component vector with unit magnitude [32, 33]. Thus, **u** can be considered a projection vector. As suggested by Zhou and Saho [33], to test **y** and **t** from the samples and synthetic samples, the vector **u** was generated randomly 1000 (i.e. projections in random directions) times in each representation and z was compared with normality using the KS test. The percentage of the times that the null hypothesis was not rejected was tabulated for both representations; in this report, we refer to this as the *random projection test.* This procedure was performed once for samples and 1000 times for synthetic samples and averaged. Secondly, the Mardia test was applied. This test uses multivariate skewness and kurtosis (two components of the test) for evaluating deviations from normality [32]. This test produces a deviation measure for each component as well as a combined measure; we cite the findings from the skewness and kurtosis tests.

Covariance matrix comparisons: Two methods were used to compare the covariance matrices between the real and synthetic samples in the X and Y representations. The first method was based on comparing covariance matrix elements with confidence intervals (CIs), where we assumed the sample and synthetic samples were drawn from the same distribution. We used the elements from each sample's $\mathbf{C}_y$ as point estimates in both the Y and X representations. One thousand synthetic samples (n = 667) were used to calculate 1000 covariance matrices (in X and Y). For each matrix element, the pdf was formed, and 95% CIs were determined. This procedure was repeated 1000 times. The percentage of times the sample's point estimate (for each element in $\mathbf{C}_y$ and $\mathbf{C}_x$) was within the synthetic elements CIs was tabulated. The second test was based on comparing eigenvalues of the covariance matrix $\mathbf{C}_y$ determined with PCA under two conditions with the eigenvalues from the sample's $\mathbf{C}_y$ as the reference. For the first condition, the PCA transform determined with the sample was applied to synthetic **y** (sample/Syn test); synthetic eigenvalues were estimated be calculating the variances of synthetic t_j. For the second condition, the PCA transform was determined with synthetic $\mathbf{C}_y$ and applied to the sample's $\mathbf{y}$ (Syn/sample test); eigenvalues were estimated by calculating the variances of t_j. For both conditions, each eigenvalue was compared to its respective reference (sample) using the F-test.

3. Results

X representation data and mapping optimization

Figures 1-3 show the univariate pdfs (solid) for each dataset in the X representation, respectively. Note, many of these pdfs are observably non-normal, usually right-skewed. When making comparisons with normality, all variables in DS1 showed significant deviation from normality ($p < 0.0001$) except for x_9. In DS2, x_1 and x_9 (x_9 is the same in DS1) did not show significant deviations from normality, whereas the other variables exhibited significant deviations ($p < 0.0003$) with the exception of x_3 ($p = 0.0144$). In DS3, x_1 through x_5 and x_9 did not show significant deviations from normality, whereas the other for variables exhibited significant deviations ($p < 0.002$).

Figure 5 shows an example of the X-Y map, $y_9 = m(x_9)$, in the left-pane and the associated inverse map, $x_9 = m^{-1}(y_9)$, in the right-pane. Red-dashed lines are maps constructed without using synthetic data; note the staircasing effects particularly in the tail region where sample density is low. Black lines are maps constructed with synthetic data: $n = 667$ (the sample) plus $n = 10^6$ synthetic realizations produced with optimized KDE. Note, the staircasing effect was removed when incorporating synthetic data, which was common to all maps (not shown).

The DE optimization process used to develop X-Y maps is illustrated in Figure 6 for x_9 from DS1. The optimization process determined the univariate bandwidth parameters (h_j) used in

KDE. Fitness in the optimization was determined by a KS-test (D_9) performed on cumulative probability function from the sample (x_9) and the respective cumulative probability function derived from a synthetic pdf generated with candidate h_9. The scatter plot between candidate h_9 and respective D_9 for DE generation = 0 is shown in the left-pane, and the scatter plot for the DE termination generation = 30 is shown in the middle-pane. Note in the terminal generation, the solution space (middle-pane) is tightly clustered indicating DE convergence. A closer view of this cluster is shown in the right-pane of Figure 6. This scatter plot characteristic was common among all datasets and variables (not shown). The average h_9 from the terminal population of solutions (and all respective solutions for each variable) was used to generate synthetic samples with KDE.

Univariate comparisons

T representation findings are summarized first because it is the origin of the SPs. Table 4 shows the variances in the T for each dataset (i.e., eigenvalues from the samples). Due to the Y data normalization and that d = 10, multiplying a given, $\sigma_j^2(t)$ by 10% gives the percentage of the total variance that the respective t_j explains. The corresponding pdfs are shown in Figure 8, 9 and 10 for each dataset. Sample pdfs (solid lines) are compared with the corresponding synthetic pdfs (dashed). Table 5 shows the KS test findings for univariate tests. The univariate normal model was in agreement for all variables in DS1, did not hold for t_{10} in DS2, and did not hold for variables t_7, t_8, t_9, and t_{10} in DS3 (Table 5, left column for all datasets). In DS2, t_{10} explains about 0.2% of the total variance. Similarly, in DS3, the sum of the variances of the four variables (t_7, t_8, t_9, and t_{10}) constitutes about 0.14% of the total variance.

Figure 10, 11, and 12 show the Y representation pdfs resulting from the mapped sample (i.e., mapped x_i) for all three datasets respectively (solid) compared with the respective synthetic normal pdfs (dashed). For all datasets, the Y representation (Table 5, middle column) comparisons indicate agreement with univariate normality. The DS2 and DS3 findings indicate that substituting normal pdfs in T whenever sample t_j values were not normally distributed had little influence; this may be because these respective variables in total or isolation explained a minute portion of the variance in the respective PCA models.

The univariate pdf comparisons in the X representation are shown in Figure 1-3. The pdfs from the sample (solid) are compared with the corresponding synthetic pdfs (dashed). The pdfs from sample are in agreement with the corresponding pdfs from synthetic samples for all datasets (Table 5, right-column for all datasets). The parenthetical entries in Table 5 show the test findings without using synthetic data to augment the maps (using the sample only). These indicate that complementing the map construction with KDE is a necessary component of this methodology, although the degree of variation from the KS test varied across datasets.

Multivariate comparisons

The MMD tests indicated the sample and synthetic samples were drawn from the same distributions (X, Y, and T) in all datasets. In 100% of the trials, the measured MMD_u^2 values were less than the critical MMD_c. The MMD test quantities are provided in Table 6.

The multivariate normality tests are shown in Table 7 for all datasets, performed in the Y and T representations. The random projection test indicates that the sample did not deviate from normality for at least 95% (approximately) of the tests in all datasets for both **y** and **t**, whereas synthetic sample showed no deviation. The Mardia test showed significant deviations (both skewness and kurtosis) from normality when evaluating the sample for both **y** and **t** for all datasets. In contrast, synthetic samples did not show deviation from normality in either **y** or **t** for all datasets. The MMD tests in Y and T were also tests for sample-normality due to the way synthetic **y** and **t** were generated.

Sample sparsity and synthetic space filling

An illustration is provided to show that this approach fills in the multidimensional space with synthetic individuals derived from the sparse sample. We selected a synthetic individual at random from DS1 with the vector: $\mathbf{x}^T = [4.20, 2.08, 1.61, 1.15, 0.85, 0.67, 0.54, 0.44, 52.0, 23.6]$. We selected x_1 and x_8 as the scatter plot variables: for the other 8 components, all synthetic individuals within $x_{ij} \pm ½\, \sigma_j$ (the standard deviation for x_j) were selected and viewed in the x_1 - x_8 plane as a scatter plot. The same vector and limits were used to select individuals from the sample. The plots are provided in Figure 13 for comparison. The sample (left-pane) produced n = 24 individuals and the SP (right-pane) produced n = 36,398 realizations. These plots illustrate that the sample is relatively sparse, and the synthetic approach generates realizations that did not exist in the sample.

Covariance matrix comparisons

Covariance matrices with confidence intervals are provided in Table 1-Table 3 for the respective datasets. Part a of these tables shows the X representation quantities, and part b shows the corresponding quantities Y representation. For DS1 (Table 1), covariance references (sample) were within the CIs of the synthetic data for 100% of the trials in both X and Y representations. In DS2 (Table 2), the vast majority of the references were in agreement with the synthetic data except for two entries in the X representation [Table 2a]. From the 1000 trials, x_2-x_3 covariance was out of tolerance for 0.1% of the instances and the x_5-x_{10} covariance was out for 18.7% of instances. For DS3, all covariance references were within tolerance except the x_7-x_{10} covariance that was out of tolerance for 100% of the instances. We note that in this case, the reference covariance was approximately zero. The eigenvalues for the two conditions are also shown in Table 5. For both **y** (upper table) and **t** (lower table), the F-tests where not significant ($p > 0.05$) for all comparisons with the reference eigenvalues (top row for each dataset) for all three datasets.

4. Discussion and Conclusions

A hybrid KDE approach was presented to generate multivariate synthetic data. This method transforms the multivariate input space into multiple 1D marginal pdfs. This approach relies on developing univariate maps from the input space to an intermediate Y representation, where the marginals pdfs are standardized (normally distributed, zero mean with unit variance). This approach separates the problem into the covariance structure and one-dimensional marginal pdfs (i.e., parametric models in Y). The covariance structure is captured numerically (nonparametrically) in the PCA matrix (**P** and its inverse) that was used to transform the Y representation data to the T representation, where synthetic data was generated. This essentially decouples the marginal pdfs from the covariate structure. Our decoupling approach is similar to the objective of Copula modeling that follows from Sklar's work [34, 35]. Copula modeling allows specification of marginal pdfs and the correlation structure independently [36]. In contrast with Copula modeling, which is flexible, the covariance (or correlation) structure in our approach is fixed by the normal form.

There are several points worth noting about this work. A stochastic optimization approach was used to determine the univariate KDE bandwidth parameters to construct the X-Y maps. The fitness function was based on minimizing the maximum absolute difference between two numerically integrated pdfs. The number of generations in the optimization was fixed based on experience. This can be easily modified to a variable termination based on achieving the critical threshold of the KS test. In contrast, there are plug-in kernel bandwidth parameters that can be used statically. For example, plug-in parameters are derived by minimizing the asymptotic behavior of the mean integrated square error or mean squared error [37]. We explored such parameters [38], but they did not perform as well as the optimization based on KS test comparisons. The KS test has limitations as it is more sensitive to the median of the distribution rather than the tails; as an alternative that could be explored, the Anderson Darling test is a variant of the KS test that is sensitive to the distribution tails [27]. The mapping from X to Y, standardized the problem at the univariate level, but in general there is no guarantee that the process is multivariate normal in the Y representation. Generating synthetic t_j as a normal multivariate process with a diagonal covariance structure provides a basis for testing the samples' similarities to a multivariate normal process in the Y and T representations. Univariate comparisons showed the sample (**y** and **t**) did not deviate significantly from normality. When the pdf for **x** is approximately multivariate normal, generating synthetic data based on PCA is a practiced technique; our approach addresses the case when this approximation fails to hold.

The MMD test in X, Y, and T showed the sample and synthetic samples were statistically similar addressing characteristic 1 for the pdf comparisons. For characteristic 2, the multivariate normality tests in T and Y showed mixed results. The random projection test indicated the sample did not significantly deviate from normality, whereas synthetic samples in T and Y showed essentially no deviation from normality. The Mardia tests showed sample deviated from normality in both T and Y, whereas synthetic samples did not deviate significantly. The MMD findings indicated that both real and synthetic samples in X, Y, and T are drawn from the same distributions, indicting normality in Y and T (i.e., synthetic **y** and **t** are multivariate normal). The MMD tests applied in the X representation also indicated synthetic samples were similar to the samples. Although some deviations were noted, the example datasets in this report were approximated as multivariate normal processes in the Y and T representations, noting the approach is not dependent upon this characteristic. When this normality approximation holds, it implies that the original multivariate pdf estimation problem in the X representation was converted to a parametric normal model described by Eq. [1]. If this conversion process generalizes to other datasets (at least in part), it implies that at least some small-sample problems can be studied with simulations by altering, n, d, and the covariance matrix. In the T representation, samples were well approximated as univariate normal processes for the three datasets, although there were noted variations. The T representation may provide a method for comparing datasets or evaluating similarity. For example, and as noted in the T representation, about 99% of the total variance comes from the first four variables in DS1 (see Table 4). In comparison for DS2 and DS3, the first four variables account for 90-92% of the total variance, indicating their similarity in comparison with DS1. This shows DS1 is more compressible than the other two datasets. Although not the purpose of this report, the amount of compression is a likely metric for estimating the effective dimensionality (d_e) for $d_e \leq d$, which could be useful for estimating sample size. When viewing the PCA transform through the NIPALS algorithm [10], it is clear when the total variance is explained by a number of components << d, the remaining components are residue (noise, chatter, rounding errors). This effect could explain why deviations from normality in the T representation in DS2 and DS3 did not influence the Y representation that approximates a multivariate normal process. In these three datasets, we did not encounter non-normal variables (from the samples) in the T representation that explained a significant portion of the total variance. Future work will investigate causes for normality in the T representation i.e., possibly due to forced normality in Y, characteristics of the X representation data, or the PCA transform. If required, the technique could be modified to accommodate non-normality in the T representation. For example, KDE could be investigated for generating univariate non-normal distributions in the T representation with the same method used to augment the X representation data for the map constructions. In this scenario, the Y representation description may deviate from the Eq. [1] description. We have used the analytical covariance form that is suitable for normally distributed variables. In the Y representation, this form is likely appropriate. We used the same form in the X representation as well, noting that this form may not be optimal because covariance relationships are not preserved over non-linear transformations. It is our contention that the Y representation is the most suitable format for data modeling as it is common practice to standardize variables. If the variable interpretation is not important, modeling can also be performed in the T representation. The X-Y transformation converted each variable to unit variance; when the natural variation for a given x_j is important, the mapping can be easily modified to preserve the variance. The relative efficiency of the approach is an important attribute in that it only requires multiple 1D KDE applications.

The methodology in this report partially addresses the small sample problem. The feasibility of this approach was evaluated with three datasets. The approach will require further evaluation on different datasets to understand its general applicability and to understand when the univariate mapping from X to Y forms a multivariate normal pdf. When this approximation holds, synthetic data can be generated efficiently in the T representation and the original nonparametric problem can be converted to a parametric model. Multiple methods were explored to evaluate multivariate normality. These tests indicated that the sample approximated normality in both the Y and T representations in some respects but also showed some deviation from normality. The interpretation of these findings in the context of data modeling may aid in understanding the limits of both the multivariate synthetic population and normality approximations in the data studied in this report. For example, determining the limiting percentage of the random projection tests may be informative in the modeling context. In summary, we offer a definition for an insufficient sample size in the context of synthetic data. When considering a given sample with d attributes and specified covariance structure, a sample size that does not permit reconstructing the hypothetical population can be considered as insufficient. Future, work includes applying the methods in this report to understand the minimum sample size, relative to d and given covariance, that permits recovering the population.

5. Declarations

Ethics and consent to participate: Mammography data was collected under protocol #Ame13_104715, Institutional Review Board (IRB), University of South Florida, Tampa, FL.

Consent of publication: The work does not contain personal identifiers.

Availability of data: The link to publicly available data is provided in text. Mammography summary data can be obtained upon request.

Competing interests: None to repot.

Funding: The work was in part supported by Moffitt Cancer Center grant #17032001 (Miles for Moffitt) and National Institutes of Health Grants R01CA166269 and U01CA200464.

Author contributions: JH is the corresponding author, conceived the plan and methods; EF is a coauthor, developed the computer code and assisted in the plan and methods development; AB is a coauthor and provided statistical and principal component analysis expertise; MJS is a coauthor and provided statistical expertise; SE is a coauthor and assisted in the plan and methods developments.

6. Figure Captions

Figure 1. Marginal Probability Density Functions (pdfs) for DS1 in the X representation: this shows one dimensional marginal pdfs for DS1 (solid) compared with corresponding pdfs derived from synthetic data (dashes). Labeling on the x-axis cites the variable name from its respective dataset and its index name parenthetically (x_j).

Figure 2. Marginal Probability Density Functions (pdfs) for DS2 in the X representation: this shows one dimensional marginal pdfs for DS2 (solid) compared with corresponding pdfs derived from synthetic data (dashes). Labeling on the x-axis cites the variable name from its respective dataset and its index name parenthetically (x_j).

Figure 3. Marginal Probability Density Functions (pdfs) for DS3 in the X representation: this shows one dimensional marginal pdfs for DS2 (solid) compared with corresponding pdfs derived from synthetic data (dashes). Labeling on the x-axis cites the variable name from its respective dataset and its index name parenthetically (x_j).

Figure 4. Processing Flow: The top row shows the processing flow for the sample. The reversed processing flow for the SP generation is shown in the bottom row.

Figure 5. Univariate Mapping Illustration: this shows the forward (left) and inverse (right) maps for age (x_9) in DS1 and DS2. Maps using the sample only (red-dashes) are compared with the maps complemented with synthetic data (black-sold).

Figure 6. KDE Optimization Illustration: This exemplifies differential evolution optimization to determine h_9 (for x_9, age). These show scatter plots for the measured differences (measured D_9 quantities for entire generation) from the Kolmogorov Smirnov (KS) tests and the h_9 quantities for two generations: generation = 1 (left); terminal generation = 30 (middle); and exploited terminal generation (right). The terminal generation shows the solutions for h_9 are tightly clustered (compare left pane with middle and right panes). The average h_9 from the terminal generation = 30 was used as the solution.

Figure 7. Marginal Probability Density Functions (pdfs) for DS1 in the T representation: This shows the marginal pdfs for DS1. Measures from the sample (solid) are compared with synthetic pdfs (dashes). pdfs from the sample (solid) were formed by transforming the measures from Y to T with the PCA transform. These are compared with the corresponding pdfs from synthetic data [normal zero mean with variance $=\sigma_j^2(t)$].

Figure 8. Marginal Probability Density Functions (pdfs) for DS2 in the T representation: This shows the marginal pdfs for DS1. pdfs from the sample (solid) are compared with pdfs from synthetic samples (dashes). pdfs from the sample (solid) were formed by transforming the measures from Y to T with the PCA transform. These are compared with pdfs derived from synthetic samples [normal, zero mean random variables with variance $= \sigma_j^2(t)$].

Figure 9. Marginal Probability Density Functions (pdfs) for DS3 in the T representation: This shows the marginal pdfs for DS3. pdfs from the sample (solid) are compared with synthetic pdfs (dashes). pdfs from the sample (solid) were formed by transforming the measures from Y to T with the PCA transform. These are compared with the corresponding pdfs from synthetic data [normal zero mean with variance $=\sigma_j^2(t)$].

Figure 10. Marginal Probability Density Functions (pdfs) for DS1 in the Y representation: This shows the marginal pdfs for DS1 (solid). Measures from the sample (solid) were formed by mapping the measures from X to Y (normal zero mean and unit variance). These are compared with the corresponding pdfs derived from the synthetic measures (dashes). Synthetic data was derived from transforming the multivariate synthetic normal system r(**t**) to Y with the inverse PCA transform.

Figure 11. Marginal Probability Density Functions (pdfs) for DS2 in the Y representation: This shows the marginal pdfs for DS2 (solid). Measures from the sample (solid) were formed by mapping the measures from X to Y (normal zero mean and unit variance). These are compared with the corresponding pdfs derived from synthetic measures (dashes). Synthetic data was derived from transforming the multivariate synthetic normal system r(**t**) to Y with the inverse PCA transform.

Figure 12. Marginal Probability Density Functions (pdfs) for DS3 in the Y representation: This shows the marginal pdfs for DS3 (solid). Measures from the sample (solid) are compared with synthetic measures (dashes). Measures from the sample (solid) were formed by mapping the measures from X to Y (normal zero mean and unit variance). These are compared with the corresponding pdfs derived from the synthetic measures (dashes). Synthetic data was derived from transforming the multivariate synthetic normal system r(**t**) to Y with the inverse PCA transform.

Figure 13. Sample and Synthetic Data Scatter Plot Illustration: A synthetic realization was selected randomly from DS1 giving this measurement vector: $\mathbf{x}^T$ = [4.20, 2.08, 1.61, 1.15, 0.85, 0.67, 0.54, 0.44, 52.0, 23.6]. x_1 and x_8 were selected for the scatter plot variables. For the other 7 components, all synthetic realizations within ± ½ (the respective standard deviation) were selected and viewed as a scatter plot in the x_1x_8 plane. Using the same vector and limits, individuals were selected from the sample in the same manner. The scatter plot for the sample (n = 24) is shown in the left-pane and from the synthetic data (n = 36,398) in the right-pane.

7. Tables

1a

	P_1 (x_1)	P_2 (x_2)	P_3 (x_3)	P_4 (x_4)	P_5 (x_5)	P_6 (x_6)	P_7 (x_7)	P_8 (x_8)	Age (x_9)	BMI (x_{10})
P_1 (x_1)	2.27E+02 (1.69E+02, 2.88E+02)	7.27E+01 (5.38E+01, 9.44E+01)	3.61E+01 (2.54E+01, 4.83E+01)	1.97E+01 (1.47E+01, 2.58E+01)	1.18E+01 (9.17E+00, 1.53E+01)	8.16E+00 (6.16E+00, 1.04E+01)	5.76E+00 (4.30E+00, 7.19E+00)	4.27E+00 (3.21E+00, 5.26E+00)	-4.13E+01 (-5.15E+01, -2.20E+01)	-1.98E+01 (-2.59E+01, -1.18E+01)
P_2 (x_2)	0.9150	2.78E+01 (1.97E+01, 3.62E+01)	1.41E+01 (9.42E+00, 1.89E+01)	7.78E+00 (5.55E+00, 1.01E+01)	4.68E+00 (3.52E+00, 5.96E+00)	3.21E+00 (2.35E+00, 4.06E+00)	2.25E+00 (1.64E+00, 2.81E+00)	1.66E+00 (1.23E+00, 2.05E+00)	-9.34E+00 (-1.49E+01, -4.84E+00)	-5.59E+00 (-7.08E+00, -2.12E+00)
P_3 (x_3)	0.8570	0.9600	7.82E+00 (4.78E+00, 1.10E+01)	4.30E+00 (2.89E+00, 5.73E+00)	2.58E+00 (1.87E+00, 3.35E+00)	1.77E+00 (1.25E+00, 2.31E+00)	1.22E+00 (8.75E-01, 1.59E+00)	8.95E-01 (6.61E-01, 1.16E+00)	-4.22E+00 (-6.83E+00, -1.60E+00)	-2.29E+00 (-2.92E+00, -2.31E-01)
P_4 (x_4)	0.8230	0.9270	0.9660	2.53E+00 (1.81E+00, 3.28E+00)	1.52E+00 (1.15E+00, 1.93E+00)	1.04E+00 (7.74E-01, 1.32E+00)	7.18E-01 (5.41E-01, 9.14E-01)	5.32E-01 (4.09E-01, 6.73E-01)	-1.65E+00 (-3.25E+00, -3.05E-01)	-1.33E+00 (-1.63E+00, -9.54E-02)
P_5 (x_5)	0.8020	0.9060	0.9410	0.9760	9.59E-01 (7.51E-01, 1.20E+00)	6.48E-01 (5.03E-01, 8.07E-01)	4.47E-01 (3.52E-01, 5.63E-01)	3.32E-01 (2.65E-01, 4.17E-01)	-9.87E-01 (-1.96E+00, -1.37E-01)	-8.59E-01 (-1.03E+00, -6.19E-02)
P_6 (x_6)	0.8030	0.9040	0.9380	0.9720	0.9810	4.54E-01 (3.49E-01, 5.67E-01)	3.12E-01 (2.42E-01, 3.90E-01)	2.33E-01 (1.84E-01, 2.89E-01)	-6.05E-01 (-1.23E+00, 2.02E-02)	-5.83E-01 (-6.93E-01, -2.55E-02)
P_7 (x_7)	0.8100	0.9040	0.9230	0.9570	0.9690	0.9820	2.22E-01 (1.73E-01, 2.78E-01)	1.65E-01 (1.30E-01, 2.04E-01)	-4.09E-01 (-8.82E-01, -8.18E-03)	-4.28E-01 (-4.93E-01, -2.36E-02)
P_8 (x_8)	0.8000	0.8900	0.9050	0.9450	0.9590	0.9750	0.9900	1.25E-01 (9.99E-02, 1.54E-01)	-2.62E-01 (-5.91E-01, 5.92E-02)	-3.10E-01 (-3.64E-01, -1.00E-02)
Age (x_9)	-0.2330	-0.1510	-0.1290	-0.0884	-0.0859	-0.0765	-0.0739	-0.0631	1.38E+02 (1.30E+02, 1.57E+02)	-4.74E+00 (-1.09E+01, 7.87E-01)
BMI (x_{10})	-0.1950	-0.1580	-0.1220	-0.1240	-0.1300	-0.1290	-0.1350	-0.1300	-0.0600	4.52E+01 (3.72E+01, 5.31E+01)

1b

	P_1 (y_1)	P_2 (y_2)	P_3 (y_3)	P_4 (y_4)	P_5 (y_5)	P_6 (y_6)	P_7 (y_7)	P_8 (y_8)	Age (y_9)	BMI (y_{10})
P_1 (y_1)	1.00E+00 (9.03e-01, 1.12e+00)	9.48E-01 (8.52e-01, 1.06e+00)	9.07E-01 (8.11e-01, 1.02e+00)	8.78E-01 (7.85e-01, 9.86e-01)	8.59E-01 (7.67e-01, 9.66e-01)	8.49E-01 (7.54e-01, 9.56e-01)	8.43E-01 (7.47e-01, 9.46e-01)	8.31E-01 (7.34e-01, 9.35e-01)	-2.35E-01 (-3.18e-01, -1.60e-01)	-2.29E-01 (-3.08e-01, -1.53e-01)
P_2 (y_2)	0.9480	1.00E+00 (9.01e-01, 1.11e+00)	9.71E-01 (8.70e-01, 1.09e+00)	9.49E-01 (8.52e-01, 1.06e+00)	9.34E-01 (8.38e-01, 1.05e+00)	9.22E-01 (8.25e-01, 1.03e+00)	9.16E-01 (8.17e-01, 1.03e+00)	9.03E-01 (8.08e-01, 1.01e+00)	-1.78E-01 (-2.56e-01, -1.02e-01)	-1.54E-01 (-2.33e-01, -8.02e-02)
P_3 (y_3)	0.9070	0.9710	1.00E+00 (9.01e-01, 1.12e+00)	9.78E-01 (8.79e-01, 1.09e+00)	9.68E-01 (8.69e-01, 1.08e+00)	9.61E-01 (8.63e-01, 1.07e+00)	9.54E-01 (8.55e-01, 1.07e+00)	9.43E-01 (8.45e-01, 1.05e+00)	-1.45E-01 (-2.26e-01, -6.76e-02)	-1.00E-01 (-1.77e-01, -2.53e-02)
P_4 (y_4)	0.8780	0.9490	0.9780	1.00E+00 (9.01e-01, 1.11e+00)	9.85E-01 (8.86e-01, 1.10e+00)	9.79E-01 (8.82e-01, 1.09e+00)	9.74E-01 (8.75e-01, 1.09e+00)	9.65E-01 (8.67e-01, 1.08e+00)	-1.05E-01 (-1.85e-01, -2.75e-02)	-9.38E-02 (-1.73e-01, -1.72e-02)
P_5 (y_5)	0.8590	0.9340	0.9680	0.9850	1.00E+00 (9.00e-01, 1.11e+00)	9.86E-01 (8.87e-01, 1.10e+00)	9.82E-01 (8.81e-01, 1.09e+00)	9.74E-01 (8.73e-01, 1.08e+00)	-9.71E-02 (-1.76e-01, -2.39e-02)	-9.27E-02 (-1.73e-01, -1.88e-02)
P_6 (y_6)	0.8490	0.9220	0.9610	0.9790	0.9860	1.00E+00 (9.01e-01, 1.11e+00)	9.89E-01 (8.89e-01, 1.10e+00)	9.82E-01 (8.80e-01, 1.10e+00)	-8.20E-02 (-1.62e-01, -7.43e-03)	-8.90E-02 (-1.72e-01, -1.48e-02)
P_7 (y_7)	0.8430	0.9160	0.9540	0.9740	0.9820	0.9890	1.00E+00 (8.95e-01, 1.11e+00)	9.89E-01 (8.83e-01, 1.10e+00)	-8.48E-02 (-1.65e-01, -1.01e-02)	-9.07E-02 (-1.72e-01, -1.52e-02)
P_8 (y_8)	0.8310	0.9030	0.9430	0.9650	0.9740	0.9820	0.9890	1.00E+00 (8.96e-01, 1.11e+00)	-6.30E-02 (-1.43e-01, 8.14e-03)	-8.84E-02 (-1.68e-01, -9.24e-03)
Age (y_9)	-0.2350	-0.1780	-0.1450	-0.1050	-0.0971	-0.0820	-0.0848	-0.0630	1.00E+00 (8.88e-01, 1.11e+00)	-6.64E-02 (-1.44e-01, 1.88e-02)
BMI (y_{10})	-0.2290	-0.1540	-0.1000	-0.0938	-0.0927	-0.0890	-0.0907	-0.0884	-0.0664	1.00E+00 (8.87e-01, 1.11e+00)

Table 1. **Covariance and Correlation for Dataset 1**: In both tables, entries on the diagonal and above give covariance quantities. Entries below the diagonals provide the respective Pearson correlation coefficients. Table 1a gives the X representation quantities and 1b the Y representation quantities. The covariance quantities were generated from the respective sample. Parenthetically, 95% confidence intervals generated from synthetic samples are cited below the respective covariance quantity.

2a

	Mean (x_1)	SD (x_2)	HP Filter (x_3)	LP Filter (x_4)	Local SD (x_5)	P_{20} (x_6)	Local Correlation (x_7)	Br Area (x_8)	Age (x_9)	BMI (x_{10})
Mean (x_1)	1.53E+04 (1.37e+04, 1.71e+04)	2.80E+03 (2.43e+03, 3.54e+03)	1.23E+00 (1.07e+00, 1.41e+00)	1.73E+02 (1.57e+02, 1.99e+02)	2.02E+03 (1.70e+03, 2.36e+03)	7.09E+00 (6.09e+00, 8.35e+00)	-8.77E-02 (-1.38e-01, -3.02e-02)	1.20E+03 (7.05e+02, 2.28e+03)	-2.30E+02 (-3.31e+02, -1.03e+02)	1.29E+02 (9.72e+01, 2.24e+02)
SD (x_2)	0.5170	1.91E+03 (1.50e+03, 2.28e+03)	4.40E-01 (3.71e-01, 5.21e-01)	2.55E+01 (2.55e+01, 3.94e+01)	1.05E+03 (9.20e+02, 1.32e+03)	1.40E+00 (1.32e+00, 2.14e+00)	-1.07E-01 (-1.39e-01, -9.85e-02)	-4.02E+02 (-6.93e+02, -2.05e+02)	-8.94E+01 (-1.56e+02, -7.61e+01)	-3.38E+01 (-4.56e+01, -2.62e+00)
HP Filter (x_3)	0.7500	0.7610	1.75E-04 (1.57e-04, 1.98e-04)	1.46E-02 (1.30e-02, 1.73e-02)	2.74E-01 (2.42e-01, 3.27e-01)	7.72E-04 (6.64e-04, 9.15e-04)	-3.54E-05 (-4.16e-05, -2.91e-05)	-2.17E-01 (-2.64e-01, -1.08e-01)	-1.50E-02 (-2.80e-02, -2.94e-03)	-1.37E-02 (-1.60e-02, -3.26e-03)
LP Filter (x_4)	0.8540	0.3560	0.6760	2.68E+00 (2.36e+00, 2.97e+00)	1.97E+01 (1.79e+01, 2.61e+01)	1.01E-01 (8.51e-02, 1.17e-01)	4.61E-04 (-3.44e-04, 1.08e-03)	2.00E+01 (1.40e+01, 3.57e+01)	-2.59E+00 (-4.11e+00, -1.20e+00)	1.47E+00 (1.21e+00, 2.99e+00)
Local SD (x_5)	0.5890	0.8670	0.7460	0.4340	7.70E+02 (6.53e+02, 8.85e+02)	1.15E+00 (9.51e-01, 1.44e+00)	-9.14E-02 (-1.04e-01, -7.88e-02)	-4.69E+02 (-5.89e+02, -2.84e+02)	-8.12E+01 (-1.12e+02, -6.19e+01)	-3.36E+01 (-3.89e+01, -1.15e+01)
P_{20} (x_6)	0.7750	0.4340	0.7910	0.8320	0.5630	5.46E-03 (4.15e-03, 6.39e-03)	-8.46E-05 (-1.08e-04, -4.57e-05)	-2.26E-01 (-4.08e-01, 4.41e-01)	-4.60E-02 (-1.27e-01, 8.43e-03)	-1.95E-02 (-1.87e-02, 5.41e-02)
Local Correlation (x_7)	-0.1250	-0.4310	-0.4720	0.0497	-0.5800	-0.2020	3.22E-05 (2.93e-05, 3.65e-05)	2.41E-01 (1.99e-01, 2.82e-01)	1.58E-02 (1.11e-02, 2.21e-02)	1.56E-02 (1.20e-02, 1.91e-02)
Br Area (x_8)	0.1230	-0.1170	-0.2090	0.1560	-0.2160	-0.0390	0.5420	6.15E+03 (5.50e+03, 7.01e+03)	5.26E+00 (-7.34e+01, 6.33e+01)	3.25E+02 (2.84e+02, 4.00e+02)
Age (x_9)	-0.1580	-0.1740	-0.0966	-0.1350	-0.2490	-0.0530	0.2380	0.0057	1.38E+02 (1.30e+02, 1.57e+02)	-4.74E+00 (-1.17e+01, 8.82e-01)
BMI (x_{10})	0.1550	-0.1150	-0.1540	0.1330	-0.1800	-0.0393	0.4070	0.6170	-0.0600	4.52E+01 (3.81e+01, 5.34e+01)

2b

	Mean (y_1)	SD (y_2)	HP Filter (y_3)	LP Filter (y_4)	Local SD (y_5)	P_{20} (y_6)	Local Correlation (y_7)	Br Area (y_8)	Age (y_9)	BMI (y_{10})
Mean (y_1)	1.00E+00 (8.96e-01, 1.11e+00)	6.03E-01 (5.18e-01, 6.83e-01)	7.56E-01 (6.62e-01, 8.46e-01)	8.86E-01 (7.85e-01, 9.91e-01)	6.24E-01 (5.32e-01, 7.04e-01)	8.39E-01 (7.39e-01, 9.37e-01)	-1.25E-01 (-1.96e-01, -4.65e-02)	1.53E-01 (7.79e-02, 2.34e-01)	-1.48E-01 (-2.24e-01, -7.16e-02)	1.97E-01 (1.21e-01, 2.73e-01)
SD (y_2)	0.6030	1.00E+00 (8.95e-01, 1.10e+00)	8.14E-01 (7.20e-01, 9.08e-01)	4.90E-01 (4.06e-01, 5.70e-01)	9.47E-01 (8.41e-01, 1.05e+00)	5.83E-01 (5.00e-01, 6.68e-01)	-5.64E-01 (-6.50e-01, -4.70e-01)	-1.49E-01 (-2.24e-01, -7.63e-02)	-2.46E-01 (-3.24e-01, -1.68e-01)	-9.86E-02 (-1.74e-01, -2.62e-02)
HP Filter (y_3)	0.7560	0.8140	1.00E+00 (8.89e-01, 1.10e+00)	7.05E-01 (6.09e-01, 7.94e-01)	7.96E-01 (6.99e-01, 8.87e-01)	8.44E-01 (7.44e-01, 9.43e-01)	-4.89E-01 (-5.75e-01, -3.98e-01)	-1.86E-01 (-2.61e-01, -1.09e-01)	-9.96E-02 (-1.77e-01, -2.22e-02)	-1.14E-01 (-1.87e-01, -4.23e-02)
LP Filter (y_4)	0.8860	0.4900	0.7050	1.00E+00 (8.91e-01, 1.12e+00)	5.12E-01 (4.25e-01, 5.94e-01)	8.84E-01 (7.80e-01, 9.91e-01)	3.63E-02 (-3.58e-02, 1.16e-01)	2.00E-01 (1.21e-01, 2.85e-01)	-1.43E-01 (-2.13e-01, -6.31e-02)	2.01E-01 (1.21e-01, 2.79e-01)
Local SD (y_5)	0.6240	0.9470	0.7960	0.5120	1.00E+00 (8.90e-01, 1.10e+00)	6.27E-01 (5.38e-01, 7.11e-01)	-6.51E-01 (-7.43e-01, -5.54e-01)	-2.18E-01 (-2.96e-01, -1.42e-01)	-2.79E-01 (-3.55e-01, -2.01e-01)	-1.52E-01 (-2.33e-01, -7.82e-02)
P_{20} (y_6)	0.8390	0.5830	0.8440	0.8840	0.6270	1.00E+00 (8.93e-01, 1.11e+00)	-2.07E-01 (-2.83e-01, -1.27e-01)	1.29E-03 (-7.33e-02, 8.16e-02)	-7.42E-02 (-1.53e-01, 9.28e-04)	3.86E-02 (-3.55e-02, 1.12e-01)
Local Correlation (y_7)	-0.1250	-0.5640	-0.4890	0.0363	-0.6510	-0.2070	1.00E+00 (8.89e-01, 1.11e+00)	5.36E-01 (4.50e-01, 6.33e-01)	2.50E-01 (1.70e-01, 3.29e-01)	4.22E-01 (3.35e-01, 5.08e-01)
Br Area (y_8)	0.1530	-0.1490	-0.1860	0.2000	-0.2180	0.0013	0.5360	1.00E+00 (8.90e-01, 1.12e+00)	-5.66E-03 (-8.17e-02, 6.83e-02)	6.57E-01 (5.63e-01, 7.54e-01)
Age (y_9)	-0.1480	-0.2460	-0.0996	-0.1430	-0.2790	-0.0742	0.2500	-0.0057	1.00E+00 (9.00e-01, 1.11e+00)	-6.70E-02 (-1.48e-01, 7.79e-03)
BMI (y_{10})	0.1970	-0.0986	-0.1140	0.2010	-0.1520	0.0386	0.4220	0.6570	-0.0670	1.00E+00 (8.91e-01, 1.11e+00)

Table 2. **Covariance and Correlation for Dataset 2**: In both upper and lower tables, entries on the diagonal and above give covariance quantities. Entries below the diagonals provide the respective Pearson correlation coefficients. Table 2a gives the X representation quantities and 2b the Y representation quantities. The covariance quantities were generated from the respective sample. Parenthetically, 95% confidence intervals generated from synthetic samples are cited below the respective covariance quantity.

3a

	Area (x_1)	Minor Axis Length (x_2)	Eccentricity (x_3)	Convex Area (x_4)	Equivalent Diameter (x_5)	Extent (x_6)	Solidity (x_7)	Roundness (x_8)	Shape Factor 3 (x_9)	Shape Factor 4 (x_{10})
Area (x_1)	1.94E+07 (1.79e+07, 2.21e+07)	3.22E+04 (2.83e+04, 3.63e+04)	5.28E+01 (4.45e+01, 6.85e+01)	1.97E+07 (1.81e+07, 2.23e+07)	5.19E+04 (4.79e+04, 5.88e+04)	-1.46E+01 (-2.49e+01, 3.36e+00)	7.32E-01 (-1.64e-01, 1.76e+00)	-2.23E+01 (-3.73e+01, -1.93e+01)	-6.35E+01 (-8.26e+01, -5.40e+01)	-1.35E+00 (-2.35e+00, -4.87e-01)
Minor Axis Length (x_2)	0.8060	8.22E+01 (7.42e+01, 9.12e+01)	-7.09E-02 (-9.35e-02, -4.78e-02)	3.25E+04 (2.87e+04, 3.68e+04)	8.61E+01 (7.59e+01, 9.72e+01)	7.35E-03 (-1.95e-02, 4.00e-02)	2.49E-03 (6.19e-04, 4.68e-03)	3.75E-02 (1.57e-02, 5.04e-02)	8.55E-02 (5.75e-02, 1.13e-01)	-1.31E-03 (-3.37e-03, 2.79e-04)
Eccentricity (x_3)	0.3710	-0.2420	1.04E-03 (9.46e-04, 1.19e-03)	5.38E+01 (4.44e+01, 6.82e+01)	1.41E-01 (1.17e-01, 1.82e-01)	-2.09E-04 (-2.97e-04, -8.59e-05)	-5.90E-06 (-1.31e-05, 1.13e-06)	-4.60E-04 (-5.93e-04, -4.24e-04)	-1.24E-03 (-1.42e-03, -1.14e-03)	-1.17E-05 (-1.74e-05, -4.76e-06)
Convex Area (x_4)	1.0000	0.8040	0.3740	1.99E+07 (1.84e+07, 2.26e+07)	5.24E+04 (4.85e+04, 5.94e+04)	-1.55E+01 (-2.67e+01, 1.89e+00)	3.40E-01 (-4.52e-01, 1.48e+00)	-2.43E+01 (-3.93e+01, -2.07e+01)	-6.48E+01 (-8.27e+01, -5.41e+01)	-1.62E+00 (-2.55e+00, -6.21e-01)
Equivalent Diameter (x_5)	0.9990	0.8070	0.3720	0.9990	1.38E+02 (1.28e+02, 1.57e+02)	-3.97E-02 (-6.87e-02, 6.85e-03)	2.04E-03 (-3.43e-04, 4.86e-03)	-5.85E-02 (-9.90e-02, -5.13e-02)	-1.70E-01 (-2.19e-01, -1.42e-01)	-3.62E-03 (-6.25e-03, -1.23e-03)
Extent (x_6)	-0.0765	0.0187	-0.1500	-0.0801	-0.0777	1.88E-03 (1.75e-03, 2.03e-03)	1.55E-05 (5.88e-06, 2.61e-05)	2.82E-04 (1.78e-04, 3.54e-04)	2.66E-04 (1.12e-04, 3.68e-04)	5.62E-06 (-6.97e-06, 1.10e-05)
Solidity (x_7)	0.0565	0.0934	-0.0623	0.0260	0.0589	0.1220	8.62E-06 (6.71e-06, 1.04e-05)	3.92E-05 (2.42e-05, 4.13e-05)	1.17E-05 (1.57e-06, 1.90e-05)	5.32E-06 (2.90e-06, 5.03e-06)
Roundness (x_8)	-0.1990	0.1630	-0.5610	-0.2140	-0.1960	0.2560	0.5250	6.46E-04 (4.70e-04, 7.81e-04)	5.77E-04 (5.28e-04, 7.41e-04)	2.10E-05 (1.25e-05, 2.56e-05)
Shaped Factor 3 (x_9)	-0.3720	0.2440	-0.9960	-0.3760	-0.3730	0.1590	0.1030	0.5870	1.50E-03 (1.38e-03, 1.70e-03)	1.96E-05 (9.74e-06, 2.54e-05)
Shape Factor 4 (x_{10})	-0.1130	-0.0533	-0.1350	-0.1340	-0.1140	0.0480	0.6710	0.3060	0.1880	7.29E-06 (5.41e-06, 8.93e-06)

3b

	Area (y_1)	Minor Axis Length (y_2)	Eccentricity (y_3)	Convex Area (y_4)	Equivalent Diameter (y_5)	Extent (y_6)	Solidity (y_7)	Roundness (y_8)	Shape Factor 3 (y_9)	Shape Factor 4 (y_{10})
Area (y_1)	1.00E+00 (8.98e-01, 1.11e+00)	7.94E-01 (6.95e-01, 8.98e-01)	3.86E-01 (3.06e-01, 4.71e-01)	9.98E-01 (8.96e-01, 1.11e+00)	9.99E-01 (8.97e-01, 1.11e+00)	-5.43E-02 (-1.32e-01, 1.73e-02)	6.56E-02 (-1.04e-02, 1.42e-01)	-2.66E-01 (-3.50e-01, -1.89e-01)	-3.87E-01 (-4.71e-01, -3.08e-01)	-1.30E-01 (-2.09e-01, -4.61e-02)
Minor Axis Length (y_2)	0.7940	1.00E+00 (8.94e-01, 1.10e+00)	-2.38E-01 (-3.15e-01, -1.62e-01)	7.97E-01 (6.98e-01, 8.99e-01)	7.98E-01 (6.98e-01, 9.00e-01)	2.91E-02 (-4.55e-02, 1.03e-01)	1.05E-01 (2.92e-02, 1.79e-01)	1.51E-01 (7.92e-02, 2.27e-01)	2.37E-01 (1.61e-01, 3.15e-01)	-6.76E-02 (-1.44e-01, 1.37e-02)
Eccentricity (y_3)	0.3860	-0.2380	1.00E+00 (8.92e-01, 1.11e+00)	3.81E-01 (3.03e-01, 4.66e-01)	3.81E-01 (3.02e-01, 4.66e-01)	-1.39E-01 (-2.17e-01, -6.71e-02)	-7.12E-02 (-1.46e-01, 7.79e-03)	-6.64E-01 (-7.61e-01, -5.72e-01)	-9.98E-01 (-1.11e+00, -8.92e-01)	-1.40E-01 (-2.15e-01, -6.29e-02)
Convex Area (y_4)	0.9980	0.7970	0.3810	1.00E+00 (8.96e-01, 1.11e+00)	9.99E-01 (8.96e-01, 1.11e+00)	-6.17E-02 (-1.40e-01, 9.80e-03)	4.14E-02 (-3.58e-02, 1.19e-01)	-2.80E-01 (-3.62e-01, -2.02e-01)	-3.84E-01 (-4.68e-01, -3.05e-01)	-1.44E-01 (-2.24e-01, -5.87e-02)
Equivalent Diameter (y_5)	0.9990	0.7980	0.3810	0.9990	1.00E+00 (8.97e-01, 1.11e+00)	-5.74E-02 (-1.36e-01, 1.46e-02)	6.85E-02 (-8.13e-03, 1.47e-01)	-2.65E-01 (-3.48e-01, -1.86e-01)	-3.82E-01 (-4.66e-01, -3.03e-01)	-1.27E-01 (-2.07e-01, -4.24e-02)
Extent (y_6)	-0.0543	0.0291	-0.1390	-0.0617	-0.0574	1.00E+00 (8.90e-01, 1.11e+00)	1.35E-01 (6.08e-02, 2.15e-01)	2.62E-01 (1.88e-01, 3.42e-01)	1.43E-01 (7.14e-02, 2.23e-01)	1.40E-02 (-5.91e-02, 9.98e-02)
Solidity (y_7)	0.0656	0.1050	-0.0712	0.0414	0.0685	0.1350	1.00E+00 (8.95e-01, 1.11e+00)	4.83E-01 (3.99e-01, 5.75e-01)	9.87E-02 (2.02e-02, 1.73e-01)	5.42E-01 (4.52e-01, 6.35e-01)
Roundness (y_8)	-0.2660	0.1510	-0.6640	-0.2800	-0.2650	0.2620	0.4830	1.00E+00 (8.94e-01, 1.12e+00)	6.79E-01 (5.87e-01, 7.77e-01)	3.13E-01 (2.33e-01, 3.97e-01)
Shaped Factor 3 (y_9)	-0.3870	0.2370	-0.9980	-0.3840	-0.3820	0.1430	0.0987	0.6790	1.00E+00 (8.94e-01, 1.11e+00)	1.82E-01 (1.04e-01, 2.58e-01)
Shape Factor 4 (y_{10})	-0.1300	-0.0676	-0.1400	-0.1440	-0.1270	0.0140	0.5420	0.3130	0.1820	1.00E+00 (8.92e-01, 1.12e+00)

Table 3. **Covariance and Correlation for Dataset 3**: In both tables, entries on the diagonal and above give covariance quantities. Entries below the diagonals provide the respective Pearson correlation coefficients. Table 3a gives the X representation quantities and 3b the Y representation quantities. The covariance quantities were generated from the respective sample. Parenthetically, 95% confidence intervals generated from synthetic samples are cited below the respective covariance quantity.

		t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
DS1	Sample (ref)	7.6178	1.0682	0.9634	0.2208	0.0544	0.0244	0.0166	0.0142	0.0116	0.0085
	Sample/Syn	7.6142	1.0700	0.9641	0.2213	0.0546	0.0244	0.0166	0.0145	0.0117	0.0087
	Syn/Sample	7.6189	1.0987	0.9377	0.2183	0.0520	0.0246	0.0162	0.0138	0.0109	0.0090
DS2	Sample (ref)	4.8873	2.3806	1.1083	0.6618	0.3575	0.2579	0.1576	0.1064	0.0590	0.0235
	Sample/Syn	4.8857	2.3797	1.1072	0.6590	0.3565	0.2567	0.1647	0.1068	0.0596	0.0239
	Syn/Sample	4.9237	2.2832	1.1094	0.6908	0.3691	0.2564	0.1714	0.1079	0.0634	0.0248
DS3	Sample (ref)	4.1932	2.6215	1.4759	0.9790	0.4877	0.2286	0.0116	0.0015	0.0009	0.0001
	Sample/Syn	4.1901	2.6125	1.4802	0.9837	0.4888	0.2294	0.0127	0.0015	0.0010	0.0001
	Syn/Sample	4.2638	2.6817	1.3871	0.9368	0.4851	0.2312	0.0118	0.0014	0.0009	0.0001

Table 4. **Eigenvalue Comparisons**. Within each dataset, the upper row (sample) gives the eigenvalues for the sample's $\mathbf{C}_y$ used as the reference (ref) for comparisons. The Sample/Syn (condition 1) rows gives the eigenvalues (variances) determined by applying the PCA transform derived from the sample to a synthetic (Syn) sample's $\mathbf{y}$. Syn/Sample (condition 2) rows give the eigenvalues (variances) determined by applying the PCA transform derived from synthetic $\mathbf{C}_y$ applied to the sample's $\mathbf{y}$. An F-test was used to compare Sample/Syn t_j with the reference (ref) t_j and to compare Syn/Sample t_j with the reference t_j. In all tests, the null hypothesis was not rejected (i.e. $p > 0.05$).

variable index	Dataset 1			Dataset 2			Dataset 3		
	t_i	y_i	x_i	t_i	y_i	x_i	t_i	y_i	x_i
1	99.2% (0.0%)	100% (0.0%)	99.0% (0.0%)	99.8% (51.6%)	99.4% (98.6%)	99.6% (99.2%)	99.5% (99.6%)	99.5% (99.4%)	99.7% (99.6%)
2	98.3% (92.2%)	100% (0.0%)	99.6% (0.0%)	89.9% (94.6%)	99.6% (0.00%)	99.6% (0.0%)	95.7% (79.6%)	99.7% (98.2%)	99.6% (97.3%)
3	94.5% (90.3%)	99.8% (0.0%)	99.9% (0.0%)	99.3% (99.1%)	99.4% (64.9%)	99.5% (62.8%)	83.0% (1.6%)	99.2% (97.1%)	99.1% (96.5%)
4	99.8% (4.0%)	99.6% (0.0%)	99.7% (0.0%)	98.6% (78.1%)	99.1% (26.9%)	99.3% (26.3%)	94.8% (10.9%)	99.6% (99.3%)	99.7% (99.5%)
5	85.9% (0.0%)	99.9% (0.0%)	99.6% (0.0%)	98.8% (61.4%)	99.4% (0.0%)	99.6% (0.0%)	84.6% (0.0%)	99.6% (98.5%)	99.6% (99.2%)
6	79.7% (0.0%)	99.6% (0.0%)	99.7% (0.0%)	94.9% (76.1%)	99.9% (0.0%)	99.6% (0.0%)	71.3% (64.2%)	99.7% (0.7%)	99.7% (0.7%)
7	97.2% (0.0%)	99.8% (0.0%)	99.8% (0.0%)	79.7% (60.8%)	99.8% (32.0%)	99.7% (29.1%)	0.0% (0.0%)	98.9% (0.0%)	99.0% (0.0%)
8	95.9% (0.0%)	99.7% (0.0%)	99.7% (0.0%)	72.1% (0.9%)	99.6% (47.0%)	100% (41.7%)	0.0% (0.0%)	99.0% (27.4%)	99.4% (29.8%)
9	96.0% (0.0%)	98.3% (87.4%)	97.6% (73.5%)	68.9% (70.6%)	99.8% (88.0%)	98.3% (84.4%)	0.0% (0.0%)	99.7% (99.7%)	99.6% (99.2%)
10	99.2% (0.0%)	99.5% (3.1%)	99.4% (0.5%)	2.8% (8.7%)	99.2% (4.2%)	99.7% (3.6%)	1.2% (0.0%)	98.1% (0.0%)	98.0% (0.0%)

Table 5. **Univariate Kolmogorov Smirnov (KS) Tests**: In each representation, probability density functions (pdfs) derived from synthetic samples were compared with the respective pdfs derived from sample for each variable. The KS test was applied 1000 times for each comparison. The number of times the null hypotheses was not rejected for a given test was tabulated as a percentage. Due to the experimental arrangements for y_j and t_j, each test is equivalent to testing for normality as well. Parenthetical entries show the results when not using kernel density estimation to supplement the X-Y map constructions.

MMD Test			
x	DS 1	MMD_c	0.2673
		MMD_u^2	0.0008
	DS 2	MMD_c	0.2667
		MMD_u^2	-0.0002
	DS 3	MMD_c	0.2681
		MMD_u^2	-0.0001
y	DS 1	MMD_c	0.2673
		MMD_u^2	0.0005
	DS 2	MMD_c	0.2667
		MMD_u^2	-0.0008
	DS 3	MMD_c	0.2681
		MMD_u^2	-0.0007
t	DS 1	MMD_c	0.2673
		MMD_u^2	-0.0005
	DS 2	MMD_c	0.2667
		MMD_u^2	-0.0006
	DS 3	MMD_c	0.2681
		MMD_u^2	0.0006

Table 6. **Multivariate MMD Tests**: The MMD test was used to compare the sample with 1000 synthetics samples, selected randomly. The MMD test quantities are averages over 1000 trials.

			Random Projection	Mardia Skewness	Mardia Kurtosis
y	DS 1	Sample	99.4%	<0.0001	<0.0001
		Syn	100.0%	0.9445	0.4287
	DS 2	Sample	96.8%	<0.0001	<0.0001
		Syn	100.0%	0.4520	0.4374
	DS 3	Sample	95.0%	<0.0001	<0.0001
		Syn	100.0%	0.2865	0.0859
t	DS 1	Sample	99.5%	<0.0001	<0.0001
		Syn	100.0%	0.1620	0.2143
	DS 2	Sample	94.5%	<0.0001	<0.0001
		Syn	100.0%	0.8516	0.1933
	DS 3	Sample	94.5%	<0.0001	<0.0001
		Syn	100.0%	0.6390	0.4010

Table 7. **Multivariate Normality Tests**: For the random projection test, 1000 random projections were applied to the sample and compared against normality. For the synthetic (syn) data, 1000 random projections were applied to one synthetic sample and tested against normality, which was repeated for 1000 trials. The finding are averages over these trials. Percentages reflect when the deviation was less than the KS critical value. The Mardia tests was applied to the sample and to a synthetic sample.

8. References

[1] J. C. Erves, T. L. Mayo-Gamble, A. Malin-Fair, A. Boyer, Y. Joosten, Y. C. Vaughn, L. Sherden, P. Luther, S. Miller, and C. H. Wilkins, “Needs, Priorities, and Recommendations for Engaging Underrepresented Populations in Clinical Research: A Community Perspective,” *Journal of community health*, vol. 42, no. 3, pp. 472-480, 2017.

[2] R. Foraker, D. L. Mann, and P. R. O. Payne, “Are Synthetic Data Derivatives the Future of Translational Medicine?,” *JACC Basic to translational science*, vol. 3, no. 5, pp. 716-718, 2018.

[3] M. H. Gail, and R. M. Pfeiffer, “Breast Cancer Risk Model Requirements for Counseling, Prevention, and Screening,” *Journal of the National Cancer Institute*, 2018.

[4] A. C. Garrido-Castro, and E. P. Winer, “Predicting breast cancer therapeutic response,” *Nature medicine*, vol. 24, no. 5, pp. 535-537, 2018.

[5] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, “Automated computerized classification of malignant and benign masses on digitized mammograms,” *Acad Radiol*, vol. 5, no. 3, pp. 155-68, Mar, 1998.

[6] C. Lei, W. Wei, Z. Liu, Q. Xiong, C. Yang, M. Yang, L. Zhang, T. Zhu, X. Zhuang, C. Liu, Z. Liu, J. Tian, and K. Wang, “Mammography-based radiomic analysis for predicting benign BI-RADS category 4 calcifications,” *Eur J Radiol*, vol. 121, pp. 108711, Dec, 2019.

[7] D. V. Nguyen, and D. M. Rocke, “Tumor classification by partial least squares using microarray gene expression data,” *Bioinformatics*, vol. 18, no. 1, pp. 39-50, 2002.

[8] K. Siddiqui, “Heuristics for sample size determination in multivariate statistical techniques,” *World Applied Sciences Journal*, vol. 27, no. 2, pp. 285-287, 2013.

[9] Y. Wu, M. G. Genton, and L. A. Stefanski, “A multivariate two□sample mean test for small sample size and missing data,” *Biometrics*, vol. 62, no. 3, pp. 877-885, 2006.

[10] P. Geladi, and B. R. Kowalski, “Partial least-squares regression: a tutorial,” *Analtica Chimca*, vol. 185, pp. 1-17, 1986.

[11] G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal, S. Kadoury, and A. Tang, “Deep Learning: A Primer for Radiologists,” *Radiographics : a review publication of the Radiological Society of North America, Inc*, vol. 37, no. 7, pp. 2113-2131, 2017.

[12] A. L. Buczak, S. Babin, and L. Moniz, “Data-driven approach for creating synthetic electronic medical records,” *Bmc Medical Informatics and Decision Making*, vol. 10, Oct 14, 2010.

[13] J. Q. Chen, D. Chun, M. Patel, E. Chiang, and J. James, “The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 44, Mar 14, 2019.

[14] J. Dahmen, and D. Cook, “A Synthetic Data Generation System for Healthcare Applications,” *Sensors (Basel)*, vol. 19, no. 5, Mar 1, 2019.

[15] A. R. Goncalves, A. P. Sales, P. Ray, and B. Soper, *NCI Pilot 3-Synthetic Data Generation Report* Lawrence Livermore National Lab. (LLNL): LLNL-TR-747902, 2018.

[16] B. M. Bogle, and S. Mehrotra, “A Moment Matching Approach for Generating Synthetic Data,” *Big data*, vol. 4, no. 3, pp. 160-78, 2016.

[17] D. S. Quintana, “A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation,” *eLife*, vol. 9, 2020.

[18] E. E. Fowler, A. Berglund, T. A. Sellers, S. Eschrich, and J. Heine, “Empirically-Derived Synthetic Populations to mitigate Small Sample Sizes ” *Journal of Biomedical Informatics*, vol. 105, 2020.

[19] D. W. Scott, “Feasibility of multivariate density estimates,” *Biometrika*, vol. 78, no. 1, pp. 197-205, 1991.

[20] J.-N. Hwang, S.-R. Lay, and A. Lippman, “Nonparametric multivariate density estimation: a comparative study,” *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2795-2810, 1994.

[21] Z. Wang, and D. W. Scott, “Nonparametric density estimation for high□dimensional data—Algorithms and applications,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 4, pp. e1461, 2019.

[22] K. V. Price, R. M. Storn, and J. A. Lampinen, *Differential evolution : a practical approach to global optimization*, Berlin ; New York: Springer, 2005.

[23] M. Koklu, and I. A. Ozkan, “Multiclass classification of dry beans using computer vision and machine learning techniques,” *Computers and Electronics in Agriculture*, vol. 174, pp. 105507, 2020.

[24] E. E. E. Fowler, A. Smallwood, C. Miltich, J. Drukteinis, T. A. Sellers, and J. Heine, “Generalized breast density metrics,” *Physics in Medicine and Biology*, vol. 64, no. 1, Jan, 2019.

[25] J. J. Heine, and R. P. Velthuizen, “Spectral analysis of full field digital mammography data,” *Med Phys*, vol. 29, no. 5, pp. 647-61, May, 2002.

[26] E. E. E. Fowler, C. Hathaway, F. Tillman, R. Weinfurtner, T. A. Sellers, and J. Heine, “Spatial correlation and breast cancer risk,” *Biomedical Physics & Engineering Express*, vol. 5, no. 4, pp. 045007, 2019/05/22, 2019.

[27] W. H. Press, and Numerical Recipes Software (Firm), *Numerical recipes in C*, 2nd ed., v2.0. ed., Cambridge England ; New York, N.Y.: Cambridge University Press,, 1992.

[28] H. Oh, M. S. Rice, E. T. Warner, K. A. Bertrand, E. E. Fowler, A. H. Eliassen, B. A. Rosner, J. J. Heine, and R. M. Tamimi, “Early-Life and Adult Anthropometrics in Relation to Mammographic Image Intensity Variation in the Nurses' Health Studies,” *Cancer Epidemiol Biomarkers Prev*, vol. 29, no. 2, pp. 343-351, Feb, 2020.

[29] R. Velthuzen, P, and L. Clarke, P, "Digitized mammogram standardization for display and CAD." pp. 179-187.

[30] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723-773, 2012.

[31] D. Garreau, W. Jitkrittum, and M. Kanagawa, “Large sample analysis of the median heuristic,” *arXiv preprint arXiv:1707.07269*, 2017.

[32] M. Zhou, and Y. Shao, “A Powerful Test for Multivariate Normality,” *J Appl Stat*, vol. 41, no. 2, pp. 351-363, Jan 1, 2014.

[33] Y. Shao, and M. Zhou, “A characterization of multivariate normality through univariate projections,” *J Multivar Anal*, vol. 101, no. 10, Nov, 2010.

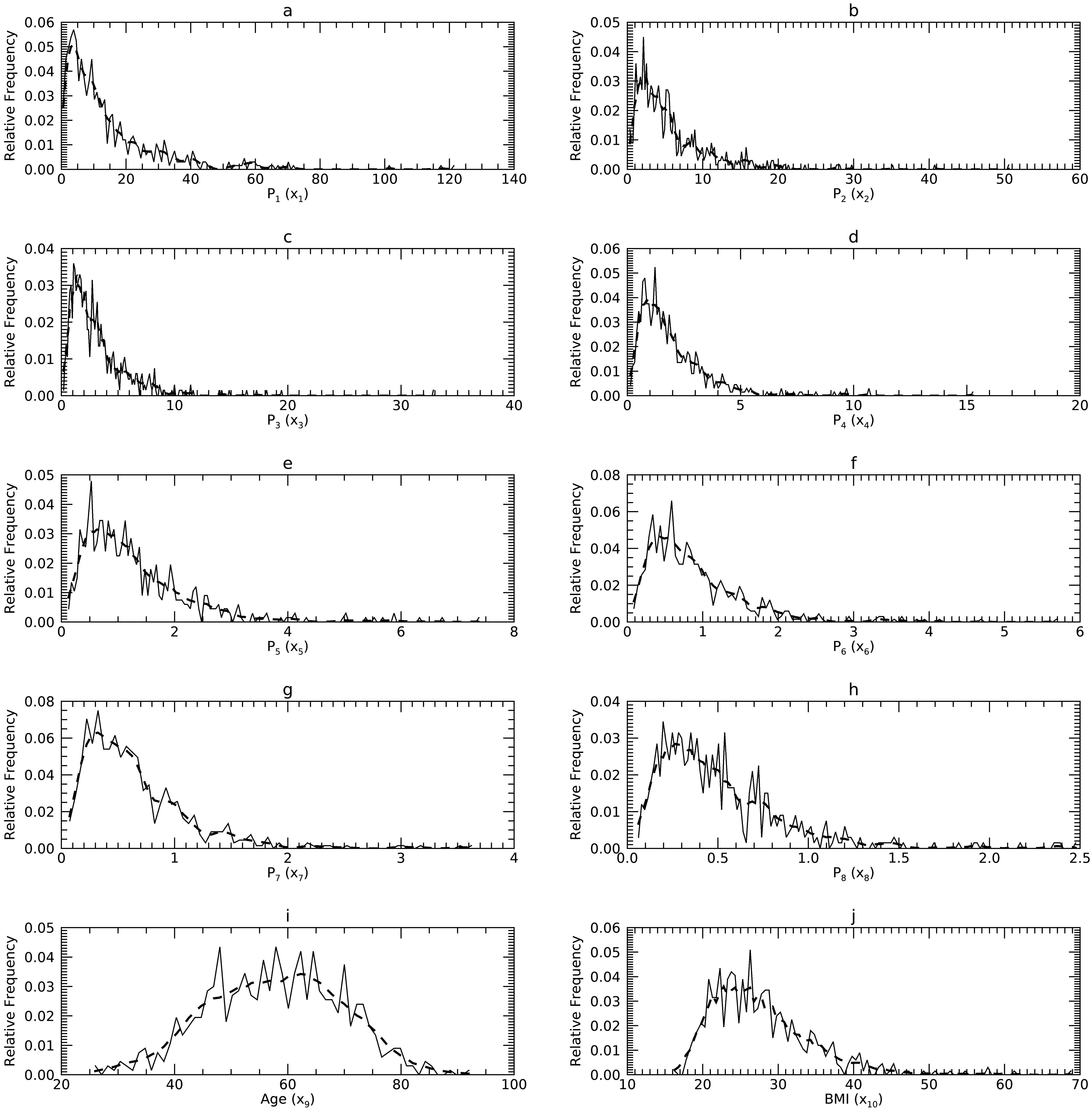
[34] M. Haugh, “An introduction to copulas,” *IEOR E4602: quantitative risk management. Lecture notes. Columbia University*, 2016.

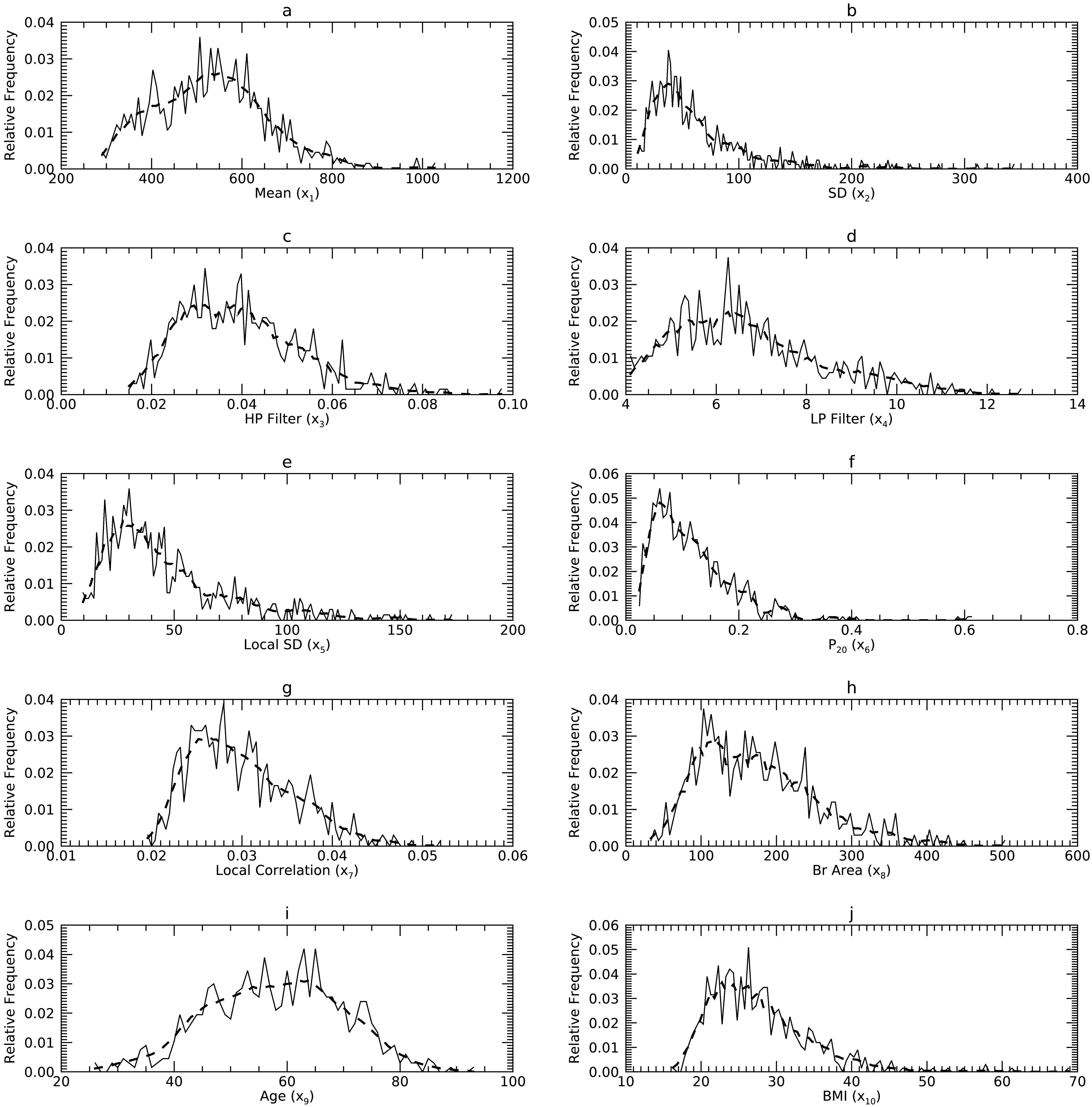
[35] F. Durante, J. Fernández-Sánchez, and C. Sempi, "How to prove Sklar’s Theorem," *Aggregation Functions in Theory and in Practise*, pp. 85-90: Springer, 2013.

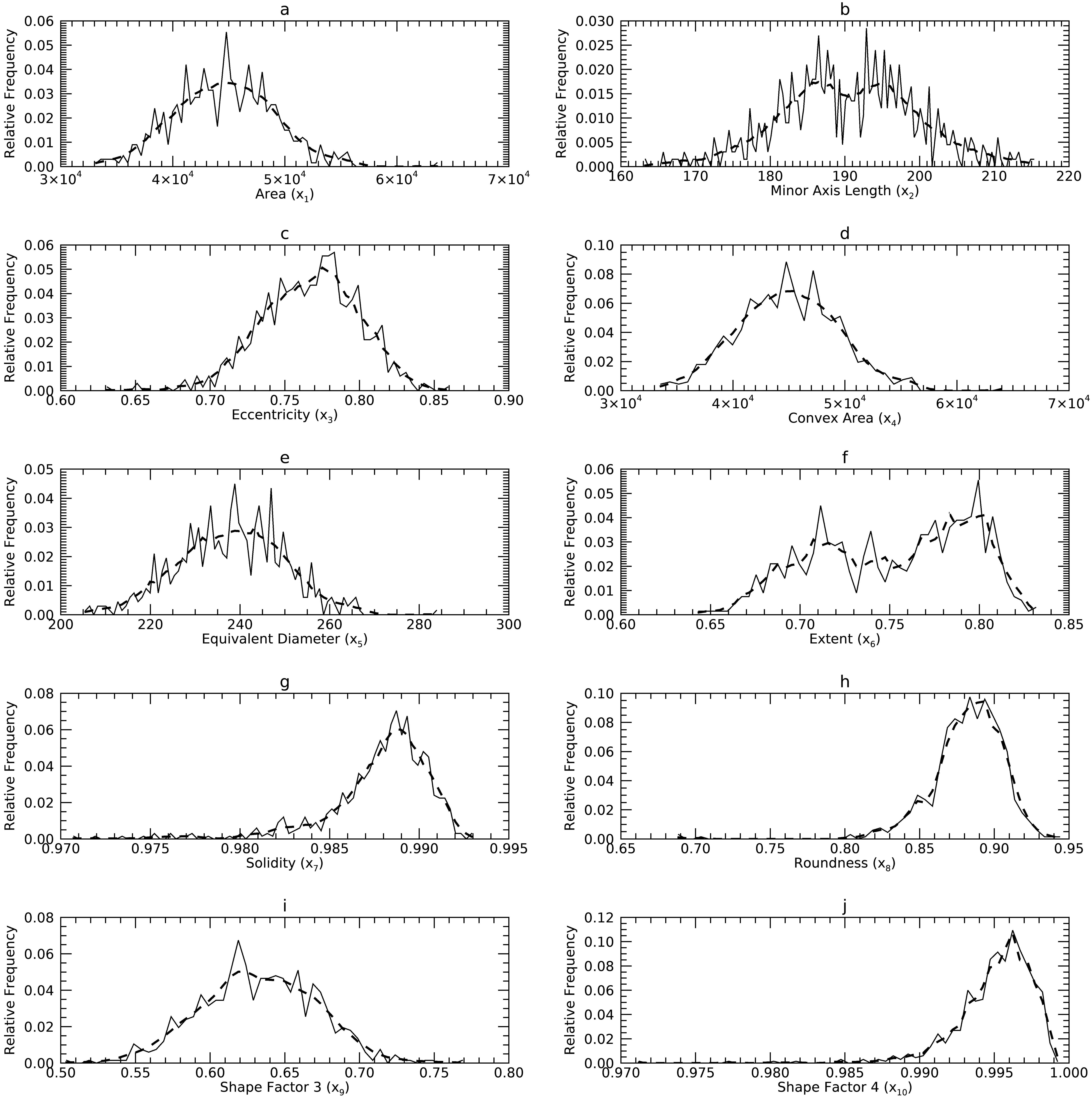
[36] D. Schirmacher, and E. Schirmacher, *Multivariate dependence modeling using pair-copulas*, Technical Report, the Society of Actuaries, 2008.

[37] M. C. Jones, J. S. Marron, and S. J. Sheather, “A brief survey of bandwidth selection for density estimation,” *Journal of the American statistical association*, vol. 91, no. 433, pp. 401-407, 1996.

[38] A. Gramacki, *Nonparametric kernel density estimation and its computational aspects*, Cham, Switzerland: Springer International Publishing AG, 2018.

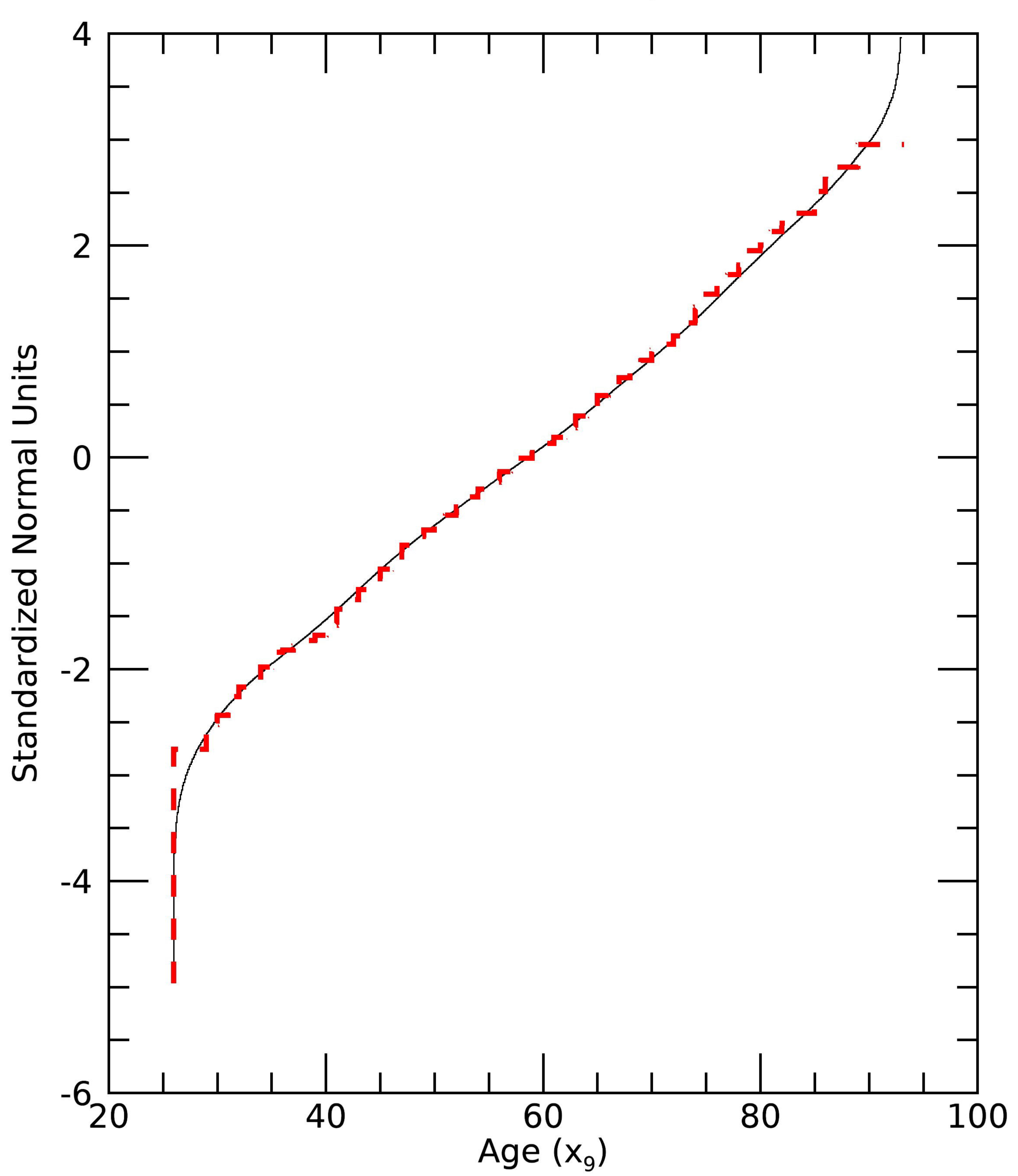

a
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
0 20 40 60 80 100 120 140
P_1 (x_1)
b
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
0 10 20 30 40 50 60
P_2 (x_2)
c
Relative Frequency
0.00 0.01 0.02 0.03 0.04
0 10 20 30 40
P_3 (x_3)
d
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
0 5 10 15 20
P_4 (x_4)
e
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
0 2 4 6 8
P_5 (x_5)
f
Relative Frequency
0.00 0.02 0.04 0.06 0.08
0 1 2 3 4 5 6
P_6 (x_6)
g
Relative Frequency
0.00 0.02 0.04 0.06 0.08
0 1 2 3 4
P_7 (x_7)
h
Relative Frequency
0.00 0.01 0.02 0.03 0.04
0.0 0.5 1.0 1.5 2.0 2.5
P_8 (x_8)
i
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
20 40 60 80 100
Age (x_9)
j
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
10 20 30 40 50 60 70
BMI (x_{10})


a
Relative Frequency
0.00
0.01
0.02
0.03
0.04
200
400
600
800
1000
1200
Mean (x_1)
b
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0
100
200
300
400
SD (x_2)
c
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.00
0.02
0.04
0.06
0.08
0.10
HP Filter (x_3)
d
Relative Frequency
0.00
0.01
0.02
0.03
0.04
4
6
8
10
12
14
LP Filter (x_4)
e
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0
50
100
150
200
Local SD (x_5)
f
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
P_{20} (x_6)
g
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.01
0.02
0.03
0.04
0.05
0.06
Local Correlation (x_7)
h
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0
100
200
300
400
500
600
Br Area (x_8)
i
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
20
40
60
80
100
Age (x_9)
j
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
10
20
30
40
50
60
70
BMI (x_{10})

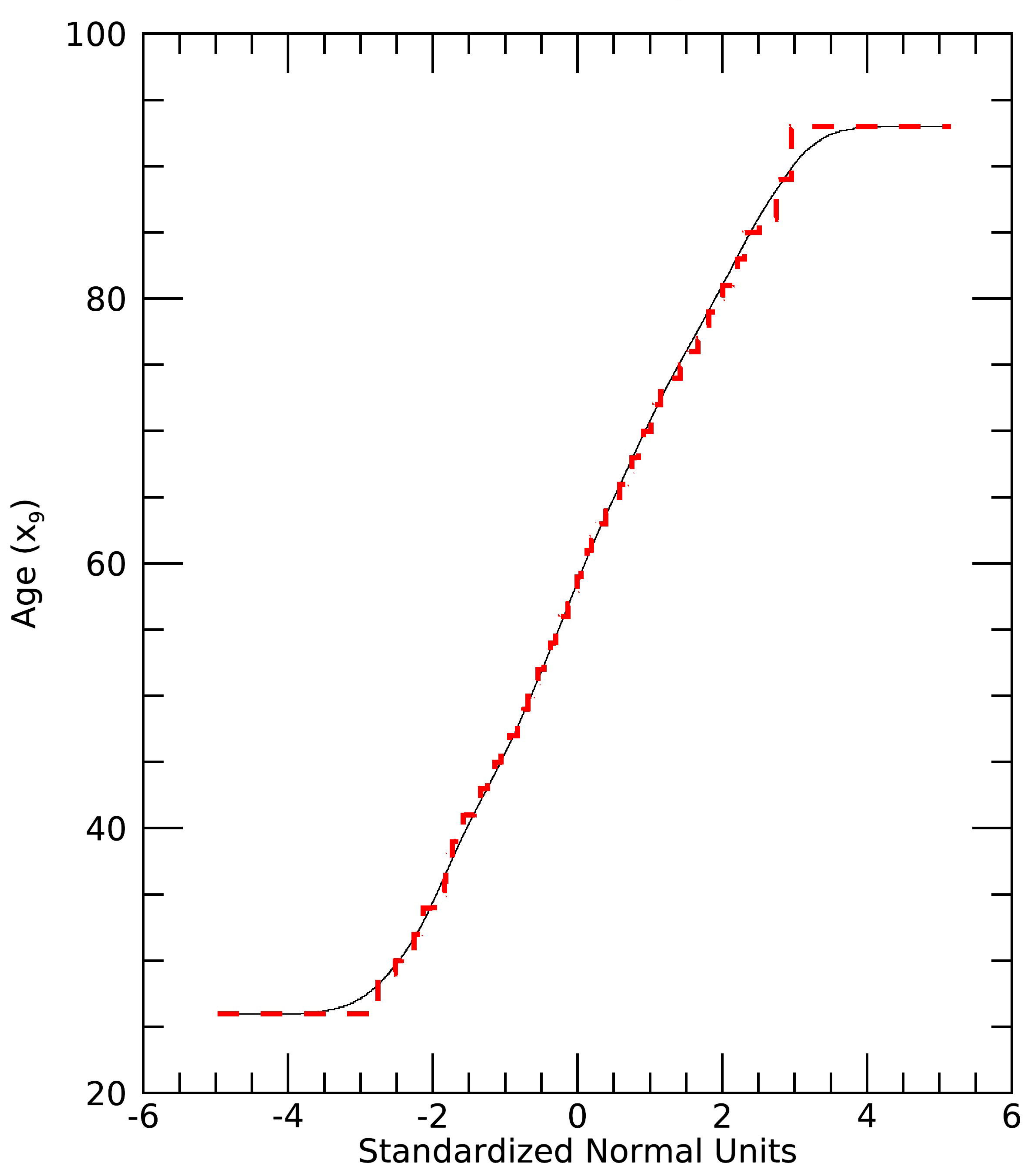

a
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
3×10⁴
4×10⁴
5×10⁴
6×10⁴
7×10⁴
Area (x1)
b
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0.030
160
170
180
190
200
210
220
Minor Axis Length (x2)
c
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Eccentricity (x3)
d
Relative Frequency
0.00
0.02
0.04
0.06
0.08
0.10
3×10⁴
4×10⁴
5×10⁴
6×10⁴
7×10⁴
Convex Area (x4)
e
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
200
220
240
260
280
300
Equivalent Diameter (x5)
f
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.60
0.65
0.70
0.75
0.80
0.85
Extent (x6)
g
Relative Frequency
0.00
0.02
0.04
0.06
0.08
0.970
0.975
0.980
0.985
0.990
0.995
Solidity (x7)
h
Relative Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.65
0.70
0.75
0.80
0.85
0.90
0.95
Roundness (x8)
i
Relative Frequency
0.00
0.02
0.04
0.06
0.08
0.50
0.55
0.60
0.65
0.70
0.75
0.80
Shape Factor 3 (x9)
j
Relative Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.970
0.975
0.980
0.985
0.990
0.995
1.000
Shape Factor 4 (x10)

Input
X
univariate mapping
Intermediate
Y
PCA transform
Uncoupled
T

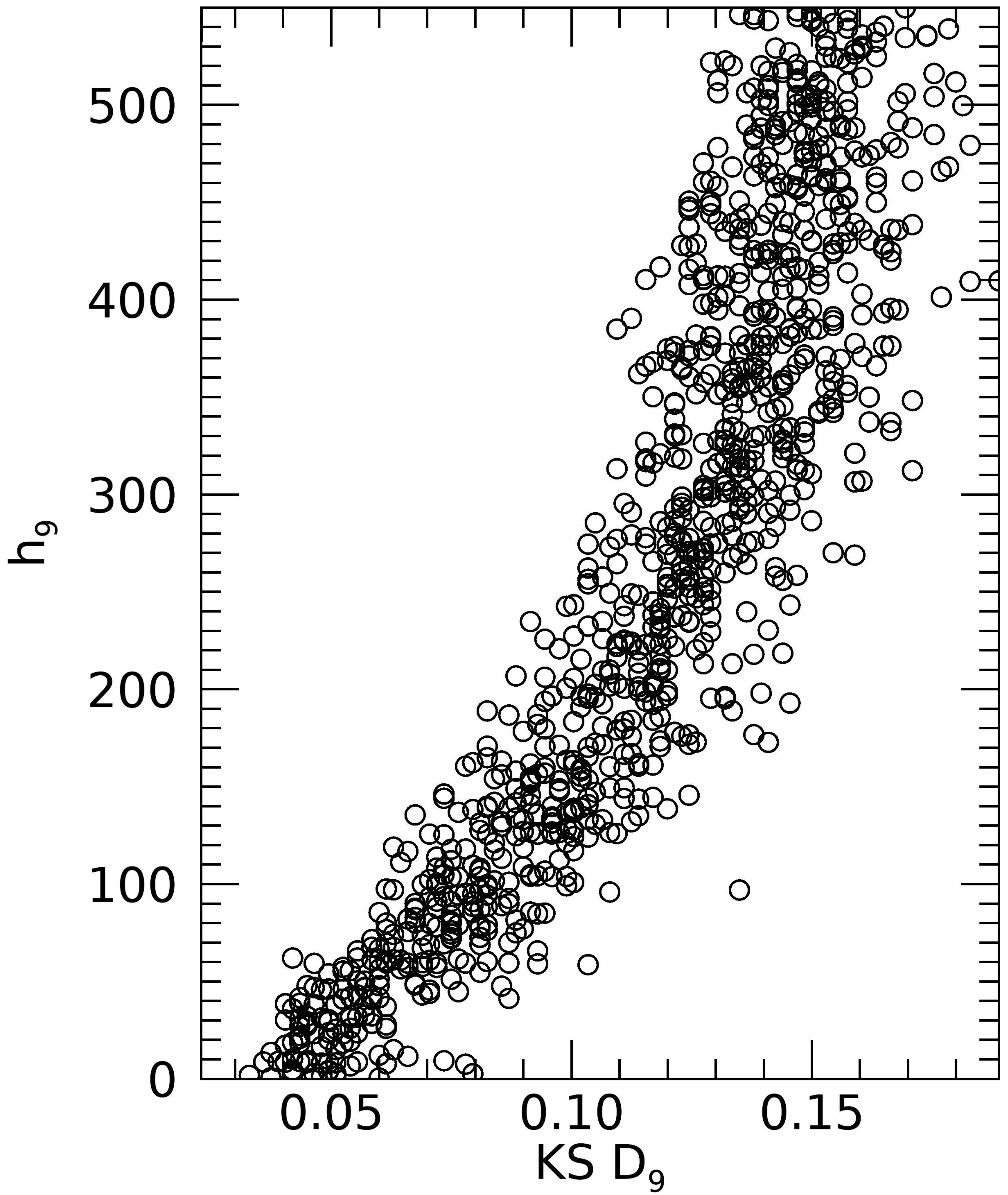
SP
T_s
inverse PCA transform
SP
Y_s
inverse univariate mapping
SP
X_s

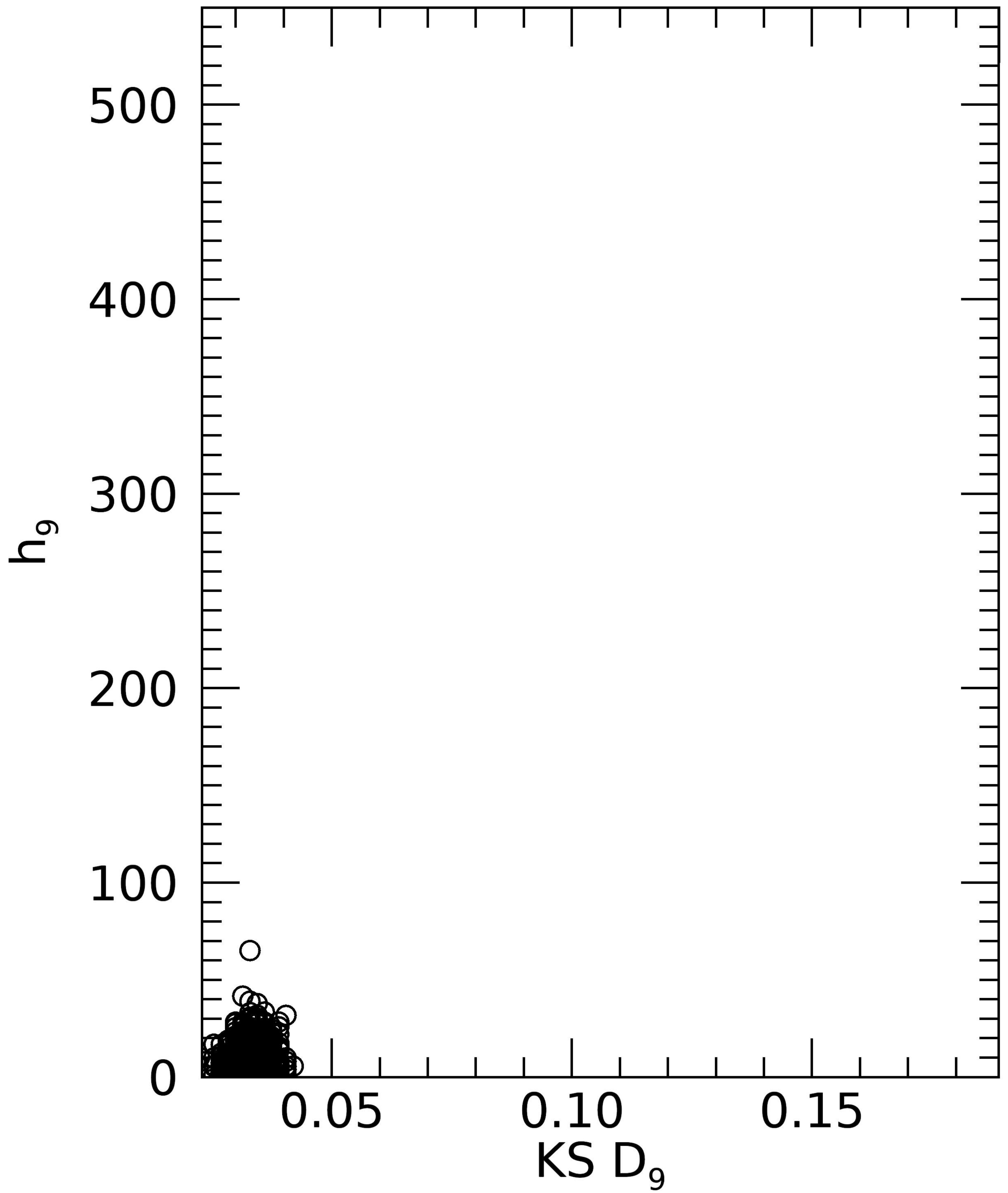


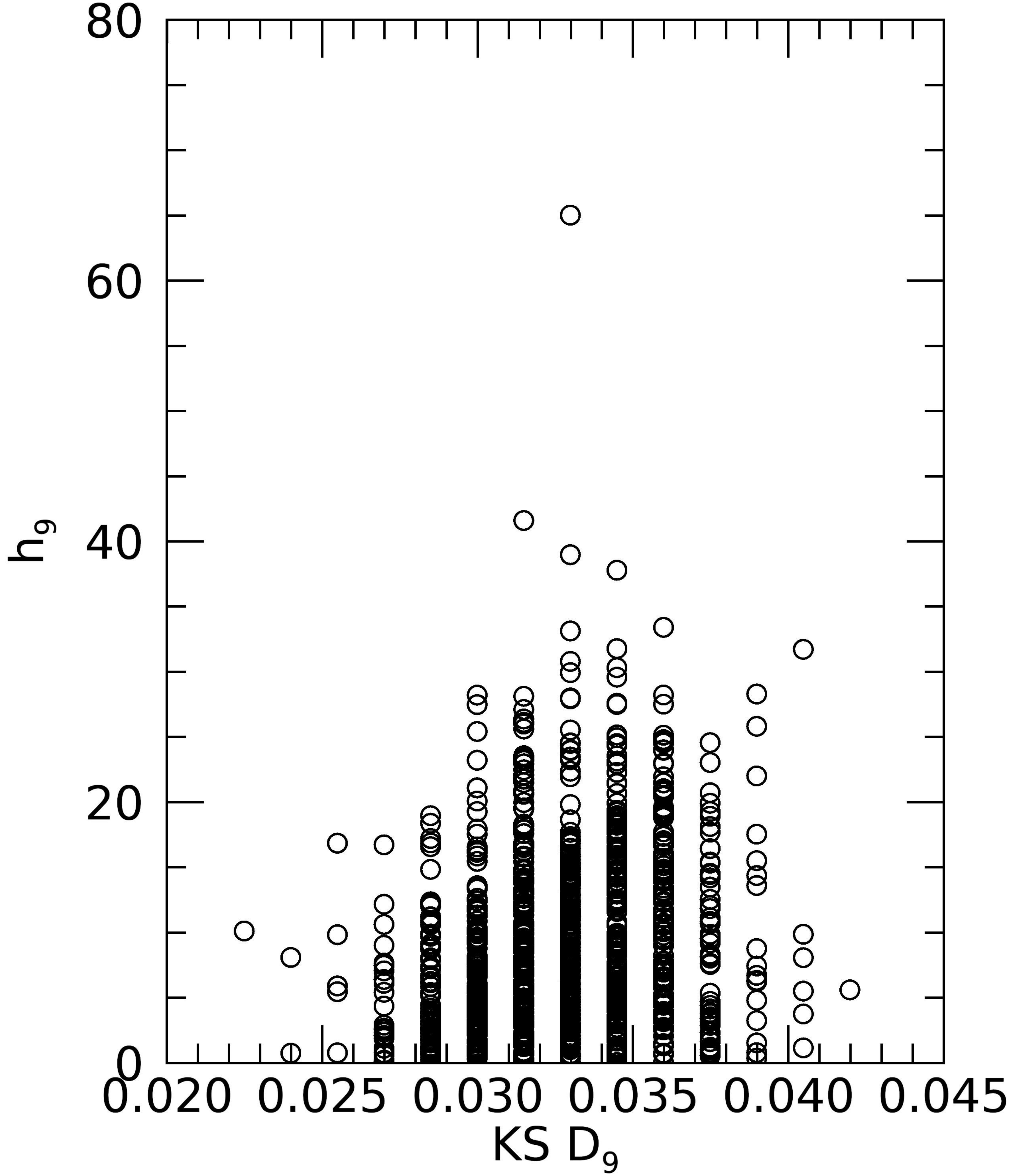
Forward Map

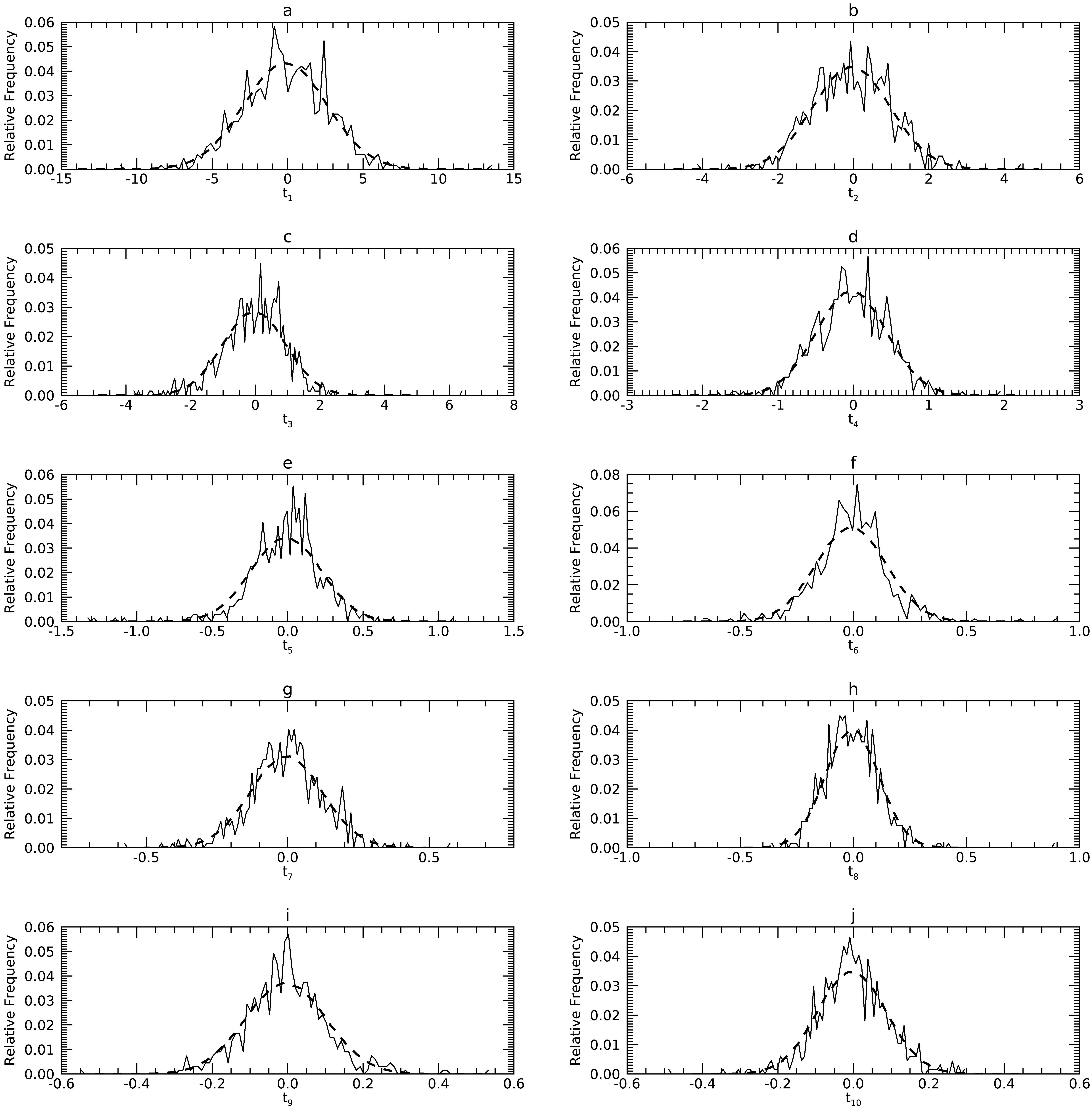


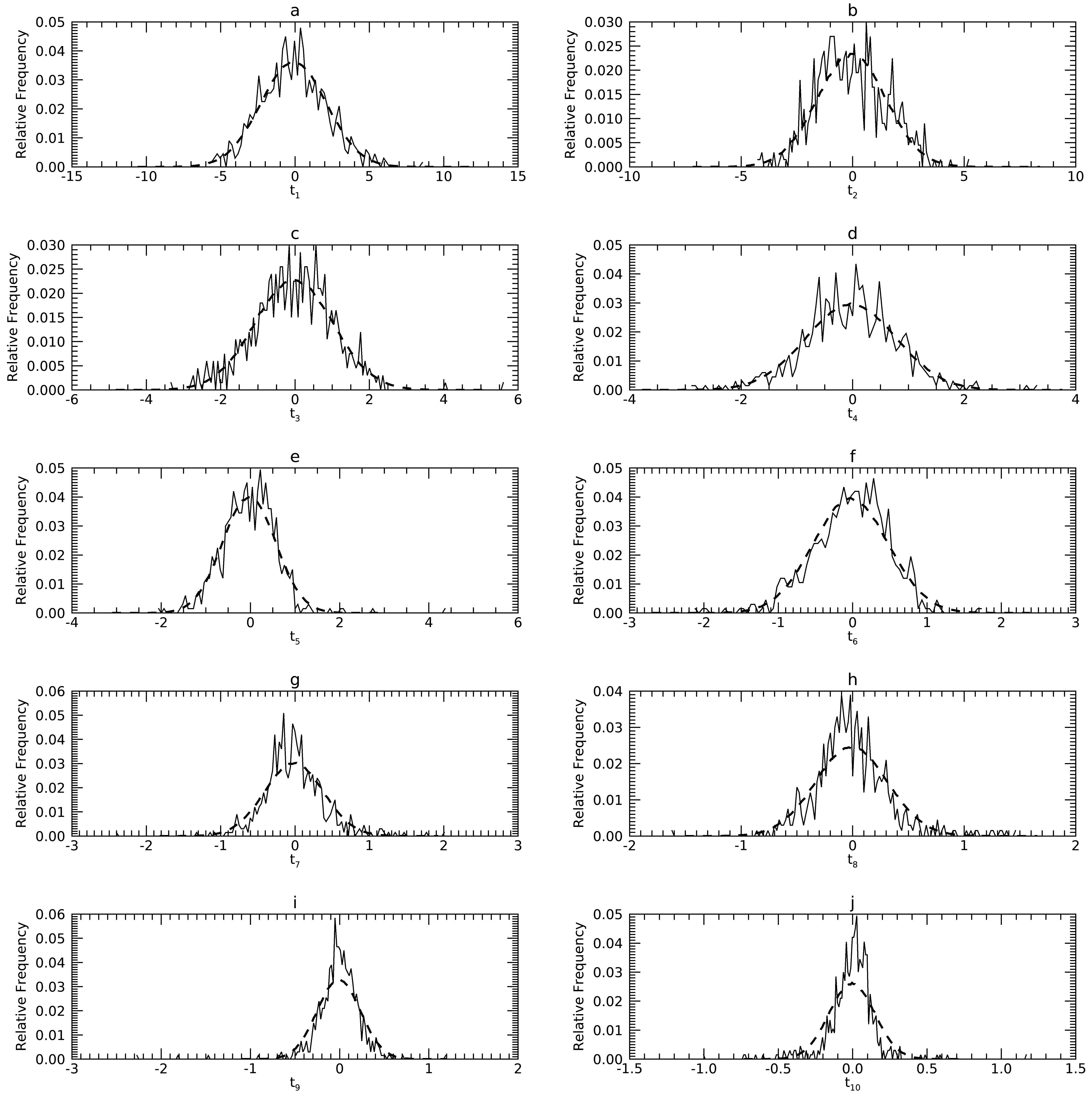
Inverse Map

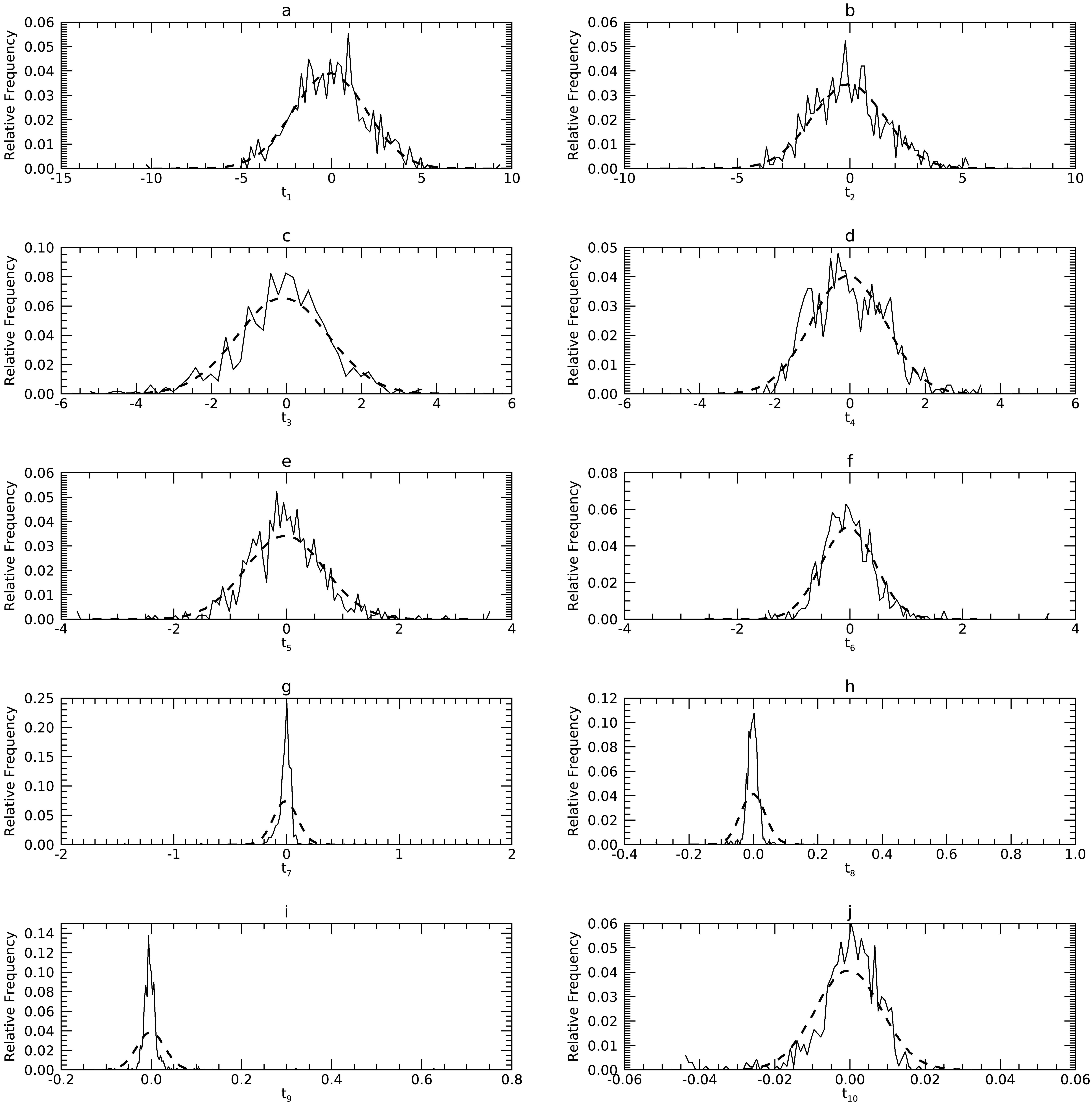

h9
500
400
300
200
100
0
0.05
0.10
0.15
KS D9

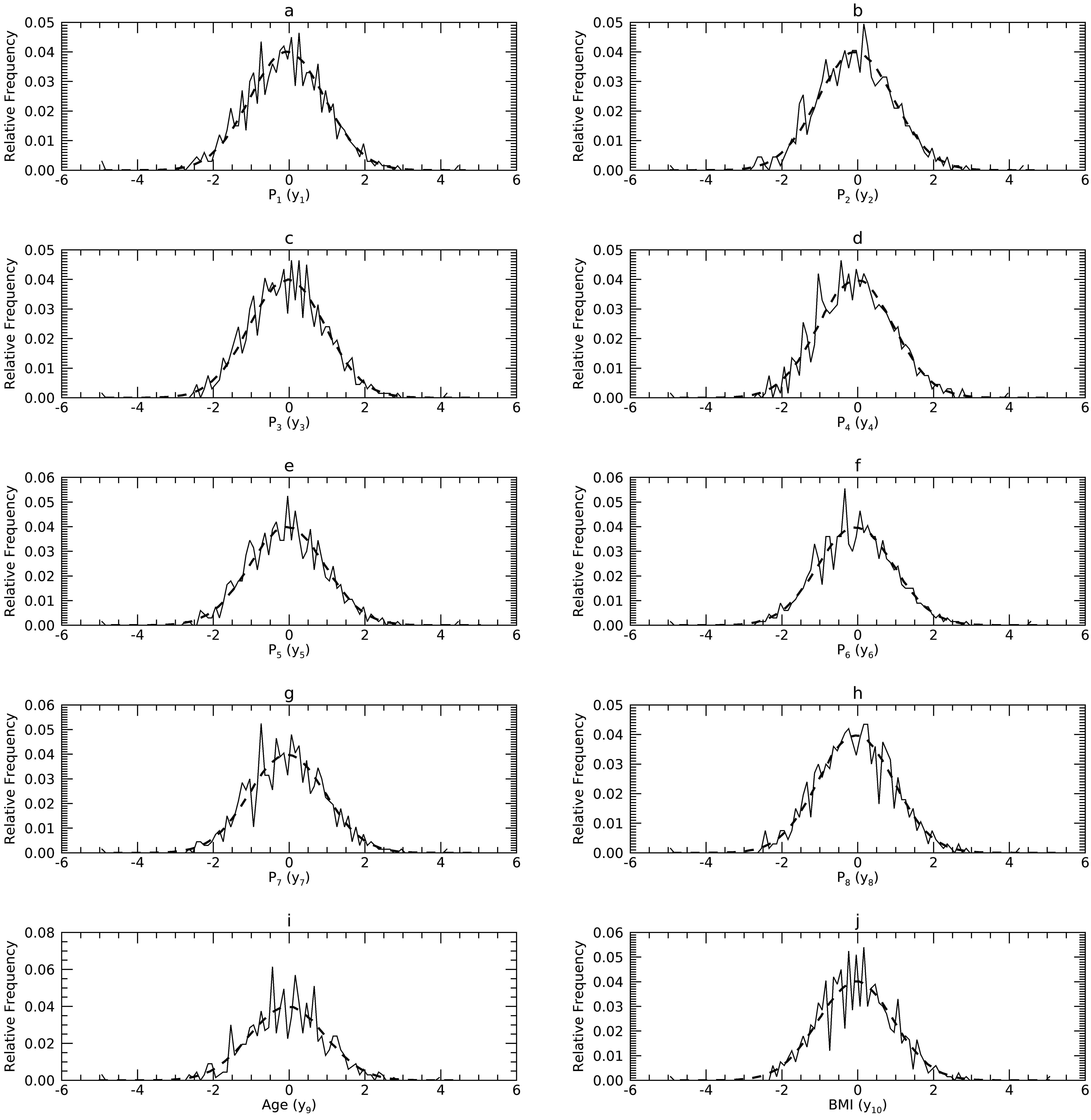

h9
500
400
300
200
100
0
0.05
0.10
0.15
KS D9

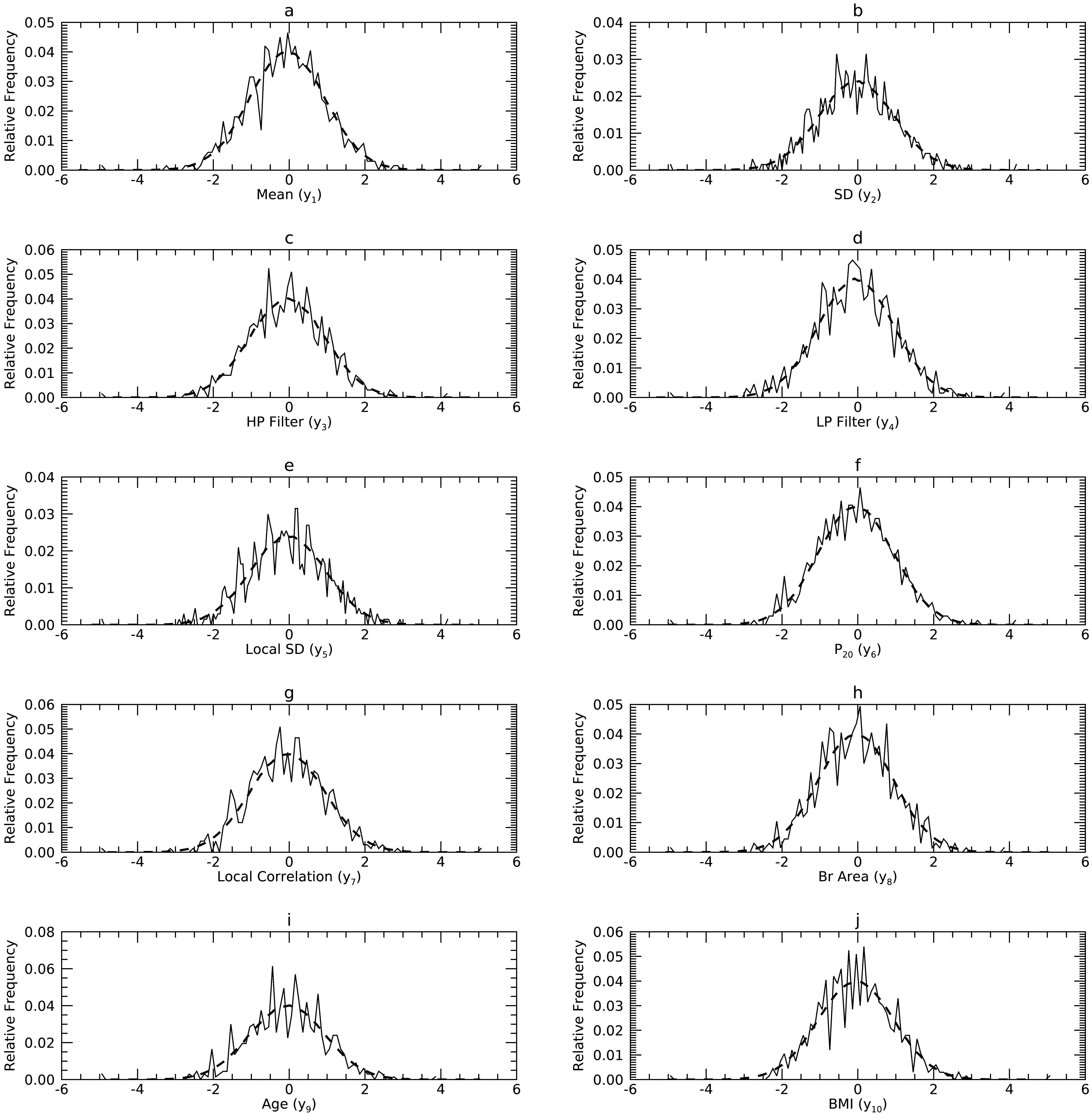

h9
80
60
40
20
0
0.020
0.025
0.030
0.035
0.040
0.045
KS D9

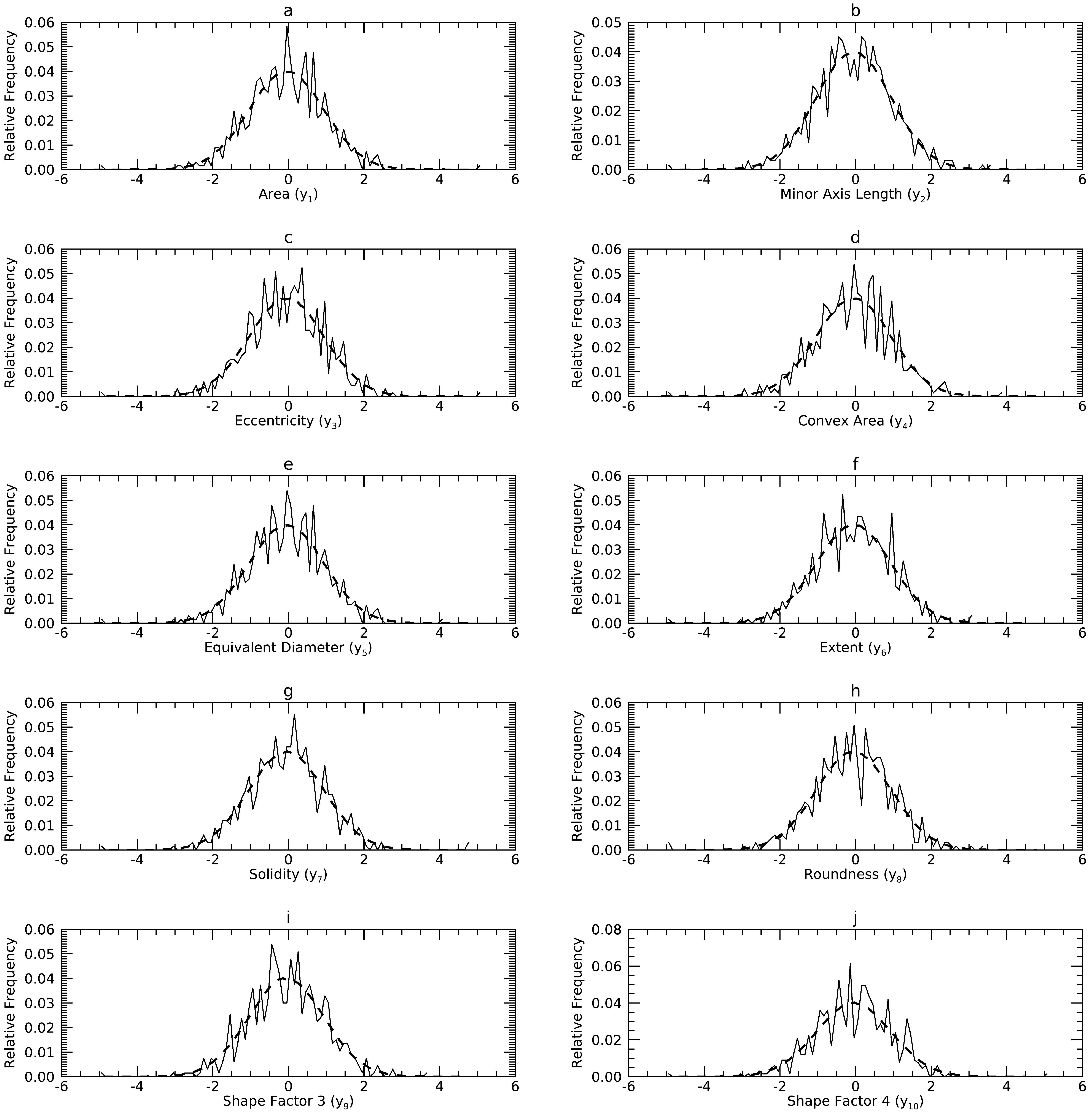

a
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-15 -10 -5 0 5 10 15
t_1
b
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
-6 -4 -2 0 2 4 6
t_2
c
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
-6 -4 -2 0 2 4 6 8
t_3
d
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-3 -2 -1 0 1 2 3
t_4
e
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
t_5
f
Relative Frequency
0.00 0.02 0.04 0.06 0.08
-1.0 -0.5 0.0 0.5 1.0
t_6
g
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
-0.5 0.0 0.5
t_7
h
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
-1.0 -0.5 0.0 0.5 1.0
t_8
i
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6
t_9
j
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6
t_{10}

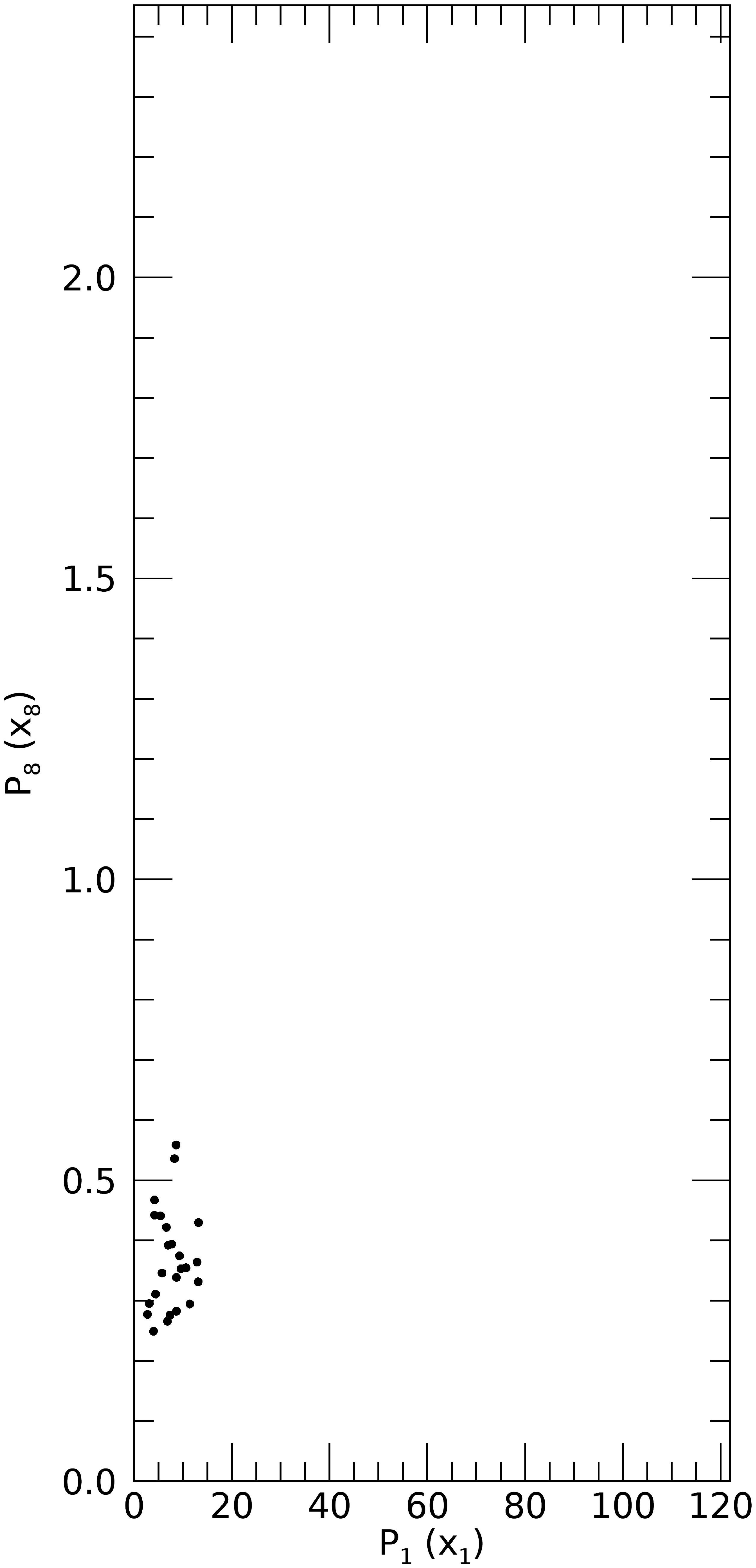
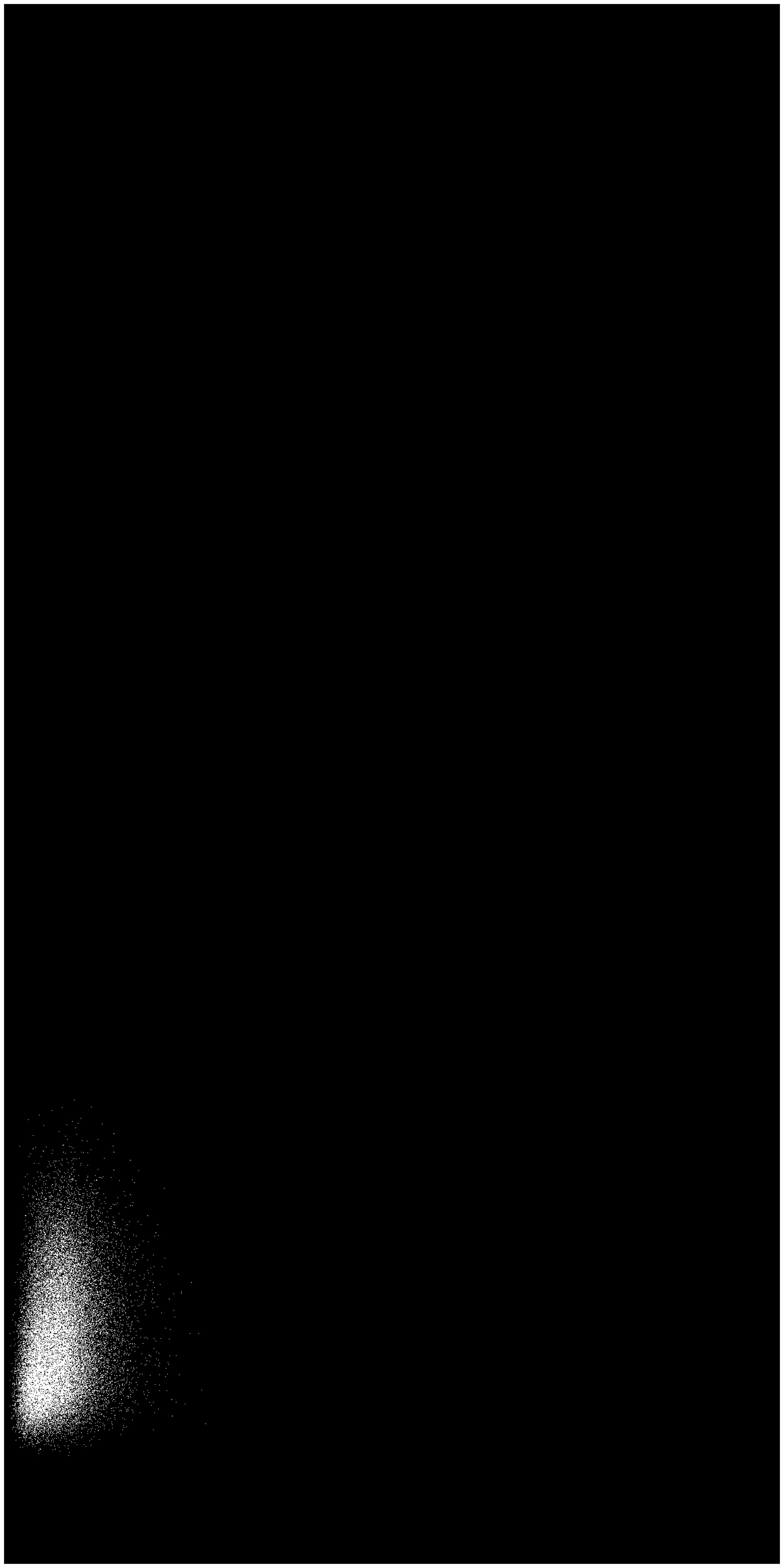

a
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-15
-10
-5
0
5
10
15
t_1
b
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0.030
-10
-5
0
5
10
t_2
c
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0.030
-6
-4
-2
0
2
4
6
t_3
d
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-4
-2
0
2
4
t_4
e
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-4
-2
0
2
4
6
t_5
f
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-3
-2
-1
0
1
2
3
t_6
g
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
-3
-2
-1
0
1
2
3
t_7
h
Relative Frequency
0.00
0.01
0.02
0.03
0.04
-2
-1
0
1
2
t_8
i
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
-3
-2
-1
0
1
2
t_9
j
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
t_{10}


a
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-15 -10 -5 0 5 10
t_1
b
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-10 -5 0 5 10
t_2
c
Relative Frequency
0.00 0.02 0.04 0.06 0.08 0.10
-6 -4 -2 0 2 4 6
t_3
d
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05
-6 -4 -2 0 2 4 6
t_4
e
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-4 -2 0 2 4
t_5
f
Relative Frequency
0.00 0.02 0.04 0.06 0.08
-4 -2 0 2 4
t_6
g
Relative Frequency
0.00 0.05 0.10 0.15 0.20 0.25
-2 -1 0 1 2
t_7
h
Relative Frequency
0.00 0.02 0.04 0.06 0.08 0.10 0.12
-0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
t_8
i
Relative Frequency
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
-0.2 0.0 0.2 0.4 0.6 0.8
t_9
j
Relative Frequency
0.00 0.01 0.02 0.03 0.04 0.05 0.06
-0.06 -0.04 -0.02 0.00 0.02 0.04 0.06
t_{10}


a
b
c
d
e
f
g
h
i
j
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.08
-6
-4
-2
0
2
4
6
P_1 (y_1)
P_2 (y_2)
P_3 (y_3)
P_4 (y_4)
P_5 (y_5)
P_6 (y_6)
P_7 (y_7)
P_8 (y_8)
Age (y_9)
BMI (y_{10})


a
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
-6
-4
-2
0
2
4
6
Mean (y_1)
b
0.00
0.01
0.02
0.03
0.04
SD (y_2)
c
0.06
HP Filter (y_3)
d
LP Filter (y_4)
e
Local SD (y_5)
f
P_{20} (y_6)
g
Local Correlation (y_7)
h
Br Area (y_8)
i
0.08
Age (y_9)
j
BMI (y_{10})


a
b
c
d
e
f
g
h
i
j
Relative Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.08
-6
-4
-2
0
2
4
6
Area (y_1)
Minor Axis Length (y_2)
Eccentricity (y_3)
Convex Area (y_4)
Equivalent Diameter (y_5)
Extent (y_6)
Solidity (y_7)
Roundness (y_8)
Shape Factor 3 (y_9)
Shape Factor 4 (y_{10})


P$_8$ (x$_8$)
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
120
P$_1$ (x$_1$)