Published in final edited form as:

Stat Med. 2013 February 20; 32(4): 685–696. doi:10.1002/sim.5533.

Application of Copulas to Improve Covariance Estimation for Partial Least Squares

Gina M. D'Angelo¹ and Lisa A. Weissfeld²

¹Division of Biostatistics, Washington University School of Medicine, St. Louis, MO

²Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA

Abstract

Dimension reduction techniques, such as partial least squares, are useful for computing summary measures and examining relationships in complex settings. Partial least squares requires an estimate of the covariance matrix as a first step in the analysis, making this estimate critical to the results. In addition, the covariance matrix also forms the basis for other techniques in multivariate analysis, such as principal component analysis and independent component analysis. This paper has been motivated by an example from an imaging study in Alzheimer's disease where there is complete separation between Alzheimer's and control subjects for one of the imaging modalities. This separation occurs in one block of variables and does not occur with the second block of variables resulting in inaccurate estimates of the covariance. We propose the use of a copula to obtain estimates of the covariance in this setting, where one set of variables comes from a mixture distribution. Simulation studies show that the proposed estimator is an improvement over the standard estimators of covariance. The methods are illustrated from the motivating example from a study in the area of Alzheimer's disease.

Keywords

copula; covariance; correlation; complete separation; neuroimaging; partial least squares

1 Introduction

Dimension reduction techniques are used in a wide variety of applications to summarize information from multiple sources to a smaller, more easily understood dimension, and to identify important structures. An important first step in the implementation of a dimension reduction technique, such as principal component analysis (PCA) [1, 2], independent component analysis (ICA) [3], factor analysis [4] and partial least squares (PLS) [5–8], is the estimation of the covariance matrix. Since dimension reduction techniques rely on some form of partitioning of the covariance matrix, accurate estimation of the covariance is critical to the outcome of the analysis. When the sample consists of subgroups with varying distributions, the standard estimator of the covariance may produce inconsistent results and is likely to be subject to bias, thus affecting the results of the dimension reduction techniques. One such case is the setting where the sample consists of two or more different subgroups within the data. An example of this situation occurs in the area of neuroimaging,

Corresponding Author: Gina M. D'Angelo, Washington University School of Medicine, Division of Biostatistics, 660 South Euclid Avenue, Box 8067, St. Louis, MO 63110-1093, Phone: (314) 362-3606, Fax: (314) 362-2693, gina@wubios.wustl.edu. Gina M. D'Angelo, PhD, Washington University School of Medicine, Division of Biostatistics, 660 South Euclid Avenue, Box 8067, St. Louis, MO 63110-1093, Phone: (314) 362-3606, Fax: (314) 362-2693, gina@wubios.wustl.edu. Lisa A. Weissfeld, PhD, Department of Biostatistics, University of Pittsburgh Graduate School of Public Health, 130 DeSoto St, 326 Parran Hall, Pittsburgh, PA 15261, lweis@pitt.edu

specifically, in the analysis of positron emission tomography (PET) data. In a study examining the use of a new radioligand, Pittsburgh compound B (PiB), for imaging in the area of Alzheimer's disease (AD) [9–15] one goal is to assess the potential usefulness of PiB PET (Pittsburgh Compound B positron emission tomography) relative to the standard FDG PET (fluorodeoxyglucose positron emission tomography) for potential diagnosis of AD. Another goal is to understand the relationship between the two PET modalities, namely, PiB PET and FDG PET and correlational analyses play a role in describing this relationship.

The standard sum of squares and cross-products matrix is the estimator most commonly used for an estimate of the covariance. This estimator, which is derived as a maximum likelihood estimate of the covariance from a multivariate normal distribution where the observations are assumed to be identically distributed, is expected to perform well when these assumptions hold. However, there are many settings where the assumptions underlying the covariance estimator are violated. Several authors have attempted to address this through the development of new estimators based on robust covariance estimators [16, 17], the EM algorithm [18] and the development of nonparametric estimators [19]; however, these estimators are not widely used in the area of dimension reduction.

The estimation of the covariance would be straightforward except when there is some or complete separation in the data; that is, when the groups have different covariances. This has a direct impact on the estimation of the covariance matrix needed for a dimension reduction analysis, since the data are from distinctly different groups. Development of methods for covariance estimation that account for this separation in groups should result in an improved estimator. For example, in the AD study the PiB PET data provide complete group separation in the sample space when comparing control and AD PiB data. However, in this study FDG PET does not have complete separation in the sample space when comparing the groups. Therefore, in this data set the PiB data has a bimodal distribution and the FDG data has a unimodal distribution. It is common for PiB values of control and Alzheimer's disease subjects to demonstrate complete separation in the sample space. The data set for the motivating example is not unique, and is a common issue in other neuroimaging and psychometric studies in the AD and depression area. As new compounds for imaging amyloid are developed, this method will be of great use in evaluating the usefulness of these compounds.

We propose covariance estimation with copulas and examine the properties of this new estimator when the data are not identically distributed and the covariances are heterogeneous across groups. Copulas are a flexible approach and have the advantage of being a function of the marginal distributions. It will be demonstrated that the copula can provide an improved covariance estimator when one set of variates is from a mixture distribution and the other set is from a single distribution.

The goal of this work is to develop and discuss covariance estimation approaches when one of the variates is sampled from a mixture. Three approaches are examined: 1) a copula approach with a mixture distribution where the covariance is estimated directly from the copula, 2) estimation of the covariance from a mixture of two normal densities, and 3) standard covariance estimation without accounting for group status. These methods are illustrated in the AD PiB PET/FDG PET data set.

2 Methodology

Let the data be denoted by (X_i, Y_i, G_i) where $X_i = (X_{i1}, \dots, X_{iK})$ is a K -dimensional vector, $Y_i = (Y_{i1}, \dots, Y_{iJ})$ is a J -dimensional vector, and G_i denotes group status for the i th subject where $i = 1, \dots, n$ and n is the sample size. Let $G = \{0, \dots, M\}$ be a variable for group that denotes $(M+1)$ groups and $m = 0, \dots, M$ is a subscript indicating group status. Let $k = 1, \dots,$

K be a subscript indicating the k th variable of X , i.e. X_k , and $j = 1, \dots, J$ be a subscript indicating the j th variable of Y , i.e. Y_j . The goal is to estimate the covariance structure between X and Y . As motivated by the example in the introduction, an assumption being made is that X is unimodal and Y is bimodal where the distribution of Y differs by group status. This is an important assumption since X is unimodal and has a single distribution and Y is bimodal and has a mixture distribution. We will describe a few covariance estimators that take into account group status.

2.1 Existing Covariance Methods

The first approach is the standard estimate of a covariance (STAN) calculated by the sum of squares and cross-products matrix which is the appropriate approach when the data are i.i.d. multivariate normal. The standard covariance calculation is of the form

$$Cov_{STAN}(X, Y) = (1/n) \sum_{i=1}^n (x_i - \bar{x})^T (y_i - \bar{y}). \quad (1)$$

The second approach is a covariance from a mixture of two normal densities (MIX). It is appropriate to use a mixture model when various groups from the data have different distributions. The form of the covariance from a mixture model is [18, 20]

$$\begin{aligned} Cov_{MIX}(X, Y) &= \sum_{m=0}^M p_m Cov(X, Y)_m \\ &= \sum_{m=0}^M p_m \sum_{i=1}^n \frac{1}{n} [X_i - E(X|g=m)]^T [Y_i - E(Y|g=m)]. \end{aligned} \quad (2)$$

2.2 Covariance From Copula

A copula is a mathematical function of a set of marginal distributions that can be used to define a single multivariate distribution [21, 22] with a small number of dependence parameters. Let $C_{\delta_{kj}}(u_k, v_j)$ denote a copula where C is the copula function, δ_{kj} is the dependence parameter, and $u_k = F(x_k)$ and $v_j = F(y_j)$ are the marginal distributions of X_k and Y_j . The margins and dependence structure are estimated separately reducing the computing time and easing the solution. The objective is to estimate the dependence structure from the copula and use this dependence structure to estimate the covariance directly from the copula. The bivariate distributions are calculated from the copula approach. Define X_k to have a single distribution and Y_j to have a mixture of normal distributions where

$v_j = F(y_j) = \sum_{m=0}^M p_m F_m(y_j)$, $p_m = \sum_{i=1}^n I(g_i=m)/n$, and $F_m(y_j)$ is the distribution of Y_j for the m th group. It is statistically valid to place a mixture distribution into a copula because the statistical properties of a distribution hold for the mixture distribution.

We consider three copula families that can address positive and negative dependence that are stable for all parameter values: Gaussian, Frank, and Gumbel copulas. The Clayton and Joe copulas can handle both negative and positive dependence; however, they are unstable when the parameters are close to the boundaries and we are unable to use them for our data. We also consider selecting the copula that best captures the dependency structure for each pair of variables where only the Gaussian, Frank, and Gumbel copulas can be considered due to the reasons just stated. Frank's copula family (Frank) is [21]

$$C_{\delta_{kj}}(u_k, v_j) = -\frac{1}{\delta_{kj}} \ln \left[1 + \frac{(e^{-\delta_{kj}u_k} - 1)(e^{-\delta_{kj}v_j} - 1)}{(e^{-\delta_{kj}} - 1)} \right], \quad -\infty \leq \delta_{kj} \leq \infty. \quad (3)$$

The Gaussian copula family (Gauss) is [21]:

$$C_{\delta_{kj}}(u_k, v_j) = \Phi_{\delta_{kj}}(\Phi^{-1}(u_k), \Phi^{-1}(v_j)), \quad -1 \leq \delta_{kj} \leq 1, \quad (4)$$

where Φ is the $N(0, 1)$ cdf and Φ^{-1} is the inverse function of the normal CDF. The Gumbel copula family (Gumbel) is [21]

$$C_{\delta_{kj}}(u_k, v_j) = \exp \left[-\left(\tilde{u}_k^{\delta_{kj}} + \tilde{v}_j^{\delta_{kj}} \right)^{1/\delta_{kj}} \right], \quad 1 \leq \delta_{kj},$$

where $\tilde{u}_k = -\log u_k$ and $\tilde{v}_j = -\log v_j$. For a negative dependence with the Gumbel copula, [23] the distribution functions of a 90 degree rotation, $C_{90, \delta_{kj}}(u_k, v_j)$, and 270 degree rotation, $C_{270, \delta_{kj}}(u_k, v_j)$, are:

$$C_{90, \delta_{kj}}(u_k, v_j) = v_j - C_{\delta_{kj}}(1 - u_k, v_j) \quad (5)$$

and

$$C_{270, \delta_{kj}}(u_k, v_j) = u_k - C_{\delta_{kj}}(u_k, 1 - v_j). \quad (6)$$

A goodness of fit test using the AIC criterion (from the R CDVine package written by Schepsmeier and Brechman [24]) can determine which rotation is necessary when a negative dependence is present. The R CDVine package written by Schepsmeier and Brechman [24] can be used to find the MLE of the association parameters for all the copula families. From the copula, the (k, j) element of the covariance matrix can be estimated directly from Hoeffding's lemma [25–27]:

$$\begin{aligned} \text{Cov}_{EC}(X_k, Y_j) &= E(X_k Y_j) - E(X_k)E(Y_j) \\ &= \int \int [C_{\delta_{kj}}(F(x_k), F(y_j)) - F(x_k)F(y_j)] dx_k dy_j, \end{aligned} \quad (7)$$

yielding the copula estimator of the covariance. Recall that the distribution for y_j is a mixture distribution. This integral can be approximated using adaptive numerical integration [28, 29]. We consider estimating the covariance with 3 different copulas and a combination of copulas where we use: 1) Frank copula (Frank) for all pairs of variables, 2) Gaussian copula (Gauss) for all pairs of variables, 3) Gumbel copula (Gumbel) for all pairs of variables, and 4) the best copula (Cop_Mix) for that pair of variables, (X_k, Y_j) , with the copula choices of Frank, Gaussian, and Gumbel. We use a goodness of fit test with the AIC criterion to select the best copula family for each pair of variables.

3 Simulation Studies

Simulation studies were performed to assess the small sample properties for each of the approaches proposed in this paper. We compared six methods for the simulation studies: the Frank copula approach (Frank), the Gaussian copula approach (Gauss), the Gumbel copula approach (Gumbel), the best copula approach (Cop_Mix), the mixture covariance method (MIX), and the standard covariance (STAN). We calculated the bias and the square root of

the mean squared error, $\sqrt{MSE}$, of the covariance matrices for each method to determine the properties of the different estimators.

Each simulation study consisted of 1000 replications. We generated 10 variables for X and 10 variables for Y . Different sample sizes of 50 and 200 were generated with 50% of the cohort in each group G where $G \sim \text{bin}(n, .5)$. We considered the scenario of misclassification of group status with varying means, standard deviations, and correlations in the simulation studies. The first and second simulation had 20% and 40% misclassification for group status. For the third and fourth simulation studies the misclassification rate for group status was 15% and 30% respectively.

The values used to generate the data can be found in Tables 1 and 2. The data were generated by group status as $(X, Y)_m \sim N(\mu_m, \Sigma_m)$ where $m = \{0, 1\}$ denotes the subscript representing group status. For the mean structure we assumed that both X and Y had constant mean within group, i.e. $\mu_m = (\mu_x, \mu_y)_m$. For the covariance structure, $\Sigma_m = \sigma_x \sigma_y$, where σ_x was constant for X and σ_y was constant for Y within group. In addition we assumed that $\text{Cov}(X_j, X_k)$ was constant for X and $\text{Cov}(Y_j, Y_k)$ was constant for Y within group. Refer to Table 1 for the means and variances. The correlation structure was generated using the uniform distribution fixing the minimum and maximum value, i.e. $\text{Corr}(X, Y) \sim U(a, b)$ where a was the minimum correlation value and b was the maximum correlation value (refer to Table 2). The mean, correlation and covariance were kept fixed for each simulation study. We assumed the true covariance of (X, Y) was given by equation [20]

$$\Sigma = \sum_{m=0}^1 p_m \Sigma_m. \quad (8)$$

Each covariance was compared to equation 8. For each of the 100 covariance parameters we graphed boxplots of the bias for sample sizes 50 and 200 (Figures 1 and 3) and $\sqrt{MSE}$ for sample sizes 50 and 200 (Figures 2 and 4) to determine the range of bias and $\sqrt{MSE}$ of all methods. We determined the largest frequency of estimates with the smallest bias and smallest $\sqrt{MSE}$ and report the frequencies in Table 3. We found the copula approaches to have the largest frequency of estimates with smallest bias and smallest $\sqrt{MSE}$. The Gauss and Cop_Mix approaches had the largest frequency of estimates with smallest bias, where the Cop_Mix was the least biased. The Frank and Gumbel approaches always had the largest frequency of estimates with the smallest $\sqrt{MSE}$. The mixture and standard approaches never had the smallest MSE and had a few instances with the smallest bias but always less than the Gauss and Cop_Mix approaches. As the sample size increased, the copula approaches increased in the frequency of estimates with the smallest bias where a majority were with the Gauss and Cop_Mix approaches. Also, as the sample size increased, the copulas still had the largest frequency of the smallest MSE with the Frank and Gumbel mainly having the smallest MSE and the Gauss and Cop_Mix occasionally having the smallest MSE. The interquartile range for the bias was similar across all methods except for the Frank and Gumbel approach where it is much wider suggesting that there may be more bias across the parameter estimates. The interquartile range for the $\sqrt{MSE}$ is similar across all methods except is a bit wider for the Gumbel approach. The copula approaches are recommended as they do tend to have smallest bias and smallest MSE.

We also considered simulation studies with no misclassification. The studies not presented here, have shown similar results (in supplementary pages). There is a slight difference where there is an increase in the frequency of smallest bias for the MIX and STAN approaches. This indicates the copula approaches are especially useful when there is misclassification. In

addition, we considered scenarios where the distributional assumptions are incorrect and found the results to differ among the copula families (results not shown). The Gauss and Gumbel approaches performed better than the STAN and MIX approach, where the Gumbel usually had the smallest bias and smallest MSE. The Frank and Cop_Mix approaches either did slightly better or about the same as the STAN and MIX approach. The Gumbel and Gaussian copulas are robust to incorrect distributional assumptions. These simulation studies suggest that a copula approach might be utilized to estimate the covariance in the presence of complete separation of group status.

4 Human Pet Example

With the aging of the general population, the prevalence of diseases such as Alzheimer's disease (AD) is expected to increase. This poses a challenge to the medical care system since there is currently no in vivo diagnostic tool for AD and no method for early detection of the disease. The current state of the art in diagnosis relies on the results of neuropsychological tests, magnetic resonance imaging (MRI) of the brain and/or positron emission tomography (PET) of the brain to measure glucose metabolism. None of these methods measure the deposition of amyloid in the brain, which is thought to be important in the natural history of AD and a potential treatment target for the disease. Development of imaging techniques that can map and detect amyloid deposition have been the focus of much research in this area with Pittsburgh compound B (PiB) showing great promise as an imaging tool [9–15]. While PiB shows great promise as a research tool, it cannot be used in the commercial setting, since it must be mixed on site. Currently, there are multiple commercial compounds that show great promise for the imaging of amyloid using PET. At this point in time, there are large numbers of subjects that have been imaged using PiB providing a database that can be used for evaluating these new compounds. As a result of its wide use, PiB will provide a “gold standard” for assessing new compounds. In addition, the commercial compounds may not yield the complete separation between AD and control subjects that are seen in this example, so that these methods will be helpful in the evaluation of new compounds for imaging amyloid.

One approach to summarizing imaging data is the use of region of interest (ROI) results. The standardized uptake value ratios (SUVR) were computed based on the hand-drawn regions of interest as outlined in Cohen et al. [30]. The values presented are based on both the left and right side of each region of the brain and include the following regions: frontal cortex (ventral and dorsal), anterior cingulate cortex (subgenual and pregenual), mesial temporal cortex (includes hippocampus and amygdala), precuneus/posterior cingulate cortex (ventral, middle and dorsal thirds), parietal cortex, lateral temporal cortex, occipital cortex (calcarine and pole), and cerebellum. This approach provides a set of summary outcomes for 16 regions of the brain; resulting in 16 correlated outcomes. In this setting, partial least squares (PLS) has been used as a dimension reduction technique for the analysis of PET imaging data [8]. An initial step of the PLS analysis requires the computation of the covariance between the ROI data obtained from the different modalities, namely, PiB PET and FDG PET. This covariance matrix then forms the basis for the PLS analysis highlighting the need for accurate estimators of this measure.

The comparison of PiB PET and FDG PET measures would be straightforward; however, the PiB PET data results in complete group separation in the sample space when comparing control and AD PiB data. This phenomenon has a direct impact on the estimation of the covariance matrix needed for the PLS analysis, since the data are from distinctly different groups. Development of methods that account for this separation should result in an improved estimator.

The motivating example for this work is based on PET imaging data from a group of subjects that included subjects with mild-to-moderate AD, subjects with mild cognitive impairment (MCI), and control subjects. It should be noted that some institutions do not use the MCI diagnosis and these “MCI” subjects would be diagnosed as “uncertain”. Subjects were recruited through the University of Pittsburgh Alzheimer Disease Research Center (ADRC) and through a linked study of amyloid deposition in normal aging. All subjects underwent an extensive neuropsychological battery that included the standard mini-mental state examination (MMSE) [31] which ranges from 0–30. The data set includes 12 AD subjects (age mean $\pm$ SD: 71.3 $\pm$ 10.2 yrs; MMSE range: 18–28), 15 MCI subjects (age mean $\pm$ SD: 69.7 $\pm$ 8.8 yrs; MMSE range: 23–29), and 37 control subjects (age mean $\pm$ SD of 71.7 $\pm$ 11.2 yrs; MMSE range: 26–30). Details of the PiB PET and FDG PET methodology are provided in previous publications [10, 11, 14, 32].

We computed the PLS scores based on the six methods for covariance estimation and compared both the percent of covariance explained and classification based on the scores for each method. For the PLS from each of the covariance estimates, we computed a set of

scores, $\sum_{k=1}^{16} A_k X_{ki}$ and $\sum_{k=1}^{16} B_k Y_{ki}$ [7], that represented a linear combination of weights, (A, B) , and the raw data (X, Y) , for the regional PET data (for FDG and PiB, respectively). These scores were based on the minimal set of variables that related the PiB PET and FDG PET results and were obtained from the singular value decomposition (SVD) of the covariance

$$D_k = \frac{d_k^2}{\sum_{k=1}^{16} d_k^2} * 100\%$$

matrix. The summary measure, D_k , where $d_k, k = 1, \dots, 16$, denotes the eigenvalues obtained from the SVD, was used to determine the latent variables of importance with values of 3% or greater determining the number of scores that was to be used in analyses. Using linear discriminant analysis, these scores were then used as covariates to predict group membership (AD, MCI, control) with the results being presented in Table 4. Cohen's Kappa was also computed to add a summary measure of agreement for classification analysis. Also, an overall Kappa was computed to assess the overall performance of each method on the FDG and PIB data. For the copula approach we considered treating only PiB as a mixture (Table 4 Copula* results) and also treating all the data as a mixture (Table 4 Copula** results), since FDG has some overlap, to determine if any of these methods would improve classification rates.

Based on the summary measure, D_k , the results in Table 4 indicate that one PLS component was retained for each of the methods with the exception of the MIX and Frank (with all data mixture) approaches using two PLS components. The percent of covariance ranged from 81% to 98% for the first PLS component, from 1% to 18% for the second PLS component, and from <1% to 1% for the third PLS component, with MIX having the smallest value for the first PLS component. For the FDG results, the copula methods performed the best with kappas ranging from .30–.36, while the results for the remaining methods were much lower. All methods performed similarly for the PiB data with kappas ranging from 0.53 to 0.58. When treating the data all as a mixture the only copula that changed was Frank where its classification improved for FDG but was slightly worse for PiB. The overall Kappas ranged from .18–.41 with Gauss having the highest Kappa followed by Frank and Gumbel. These results confirm that the copula approaches do improve covariance estimation and classification.

5 Discussion

The neuroimaging data from our example motivated this study of covariance structures which are an important component of the PLS analysis. Covariance estimation is not limited to PLS and should motivate exploration in other dimension reduction methods. Specifically, the complete separation of the neuroimaging data in control and AD subjects for the PiB PET data pointed to the need for improvements in correlation estimates for this case. We have shown that care should be taken when describing the correlation structure between two sets of multiple variables in the setting where individuals are from different populations with results differing depending on the methods that are used for the estimation.

Our simulation studies have indicated that the covariance structure is more accurately calculated using the copula approach. The copula approaches have the smallest mean squared error and bias when compared to the other methods. The Frank and Gumbel copulas tend to have the smallest MSE while the Gauss and Cop_Mix approaches tend to have the smallest bias. The results held in the presence of misclassification. The Gaussian and Gumbel copulas are the most robust to incorrect distributional assumptions. In addition, the copula methods classify the subjects better than the other approaches for the example, pointing to the usefulness of these approaches.

The standard covariance and mixture covariance are straightforward to estimate. It should be noted that the mixture covariance is simple to estimate when the group membership is known. However, a disadvantage to both of these approaches is they do not estimate the covariance as well when there is complete separation in the data and when the distributional assumptions are incorrect. The advantage to the copula approach is in its flexibility in being able to input any marginal distribution to estimate dependence. Another advantage to using copulas is they can directly measure dependence. An additional advantage to the copula is it being more robust to incorrect distributional assumptions and misclassification. However, they are limited to inputting a small number of variables. Furthermore, very few copula families are available when negative dependence exists which limits their usefulness. If those available copula families do not fit the data well then they may not estimate the dependency any better than other standard approaches. Whereas the standard covariance and mixture covariance have a simple formula, integration is required of the copula approach. Although, adaptive integration is not a computational barrier anymore with the powerful computing available to most. Even though there are more steps required of the copula approach it is computationally simple. Regardless of the disadvantages of the copula, we have shown that the copula approach improves covariance estimation and classification.

For practical use in a data analysis setting, the simulation results lead one to favor the use of a copula-based method for the estimation of covariance. It is also reasonable to assume that data will often be misclassified and the incorrect distribution may be selected leading one to favor the copula estimators for most data analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract/grant sponsor: The National Institutes of Health (K25 AG035062, 5T32 CAU82087-03, R01 AG018402, P50 AG005133, K02 AG001039, R01 AG020226, K02 AG027998, R01 MH070729, K01 MH001976, R37 AG025516, P01 AG025204, R01 MH64821, K24 RR18192, P50 AG005681, P01 AG003991), The Alzheimer's Association (TLL-01-3381), The U.S. Department of Energy (DE-FD02-03 ER63590), and GE Healthcare, Inc.

We thank the University of Pittsburgh Alzheimer Disease Research Center and the PET Facility team including Scott K. Ziolko, Chester A. Mathis, William E. Klunk, Steven T. DeKosky, and Julie C. Price.

References

- Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. 1933; 24:417–441.
- Jolliffe, IT. *Principal Component Analysis*. Springer-Verlag; New York: 1986.
- Comon P. Independent component analysis, a new concept? *Signal Processing*. 1994; 36:287–314.
- Brown, TA. *Confirmatory Factor Analysis for Applied Research*. Guilford Press; New York: 2006.
- Wold, H. Soft modelling by latent variables: the nonlinear iterative partial least squares (NIPALS) approach. In: Gani, J., editor. *Perspectives in Probability and Statistics, In Honor of M S Bartlett*. Applied Probability Trust; 1975.
- Garthwaite PH. An interpretation of partial least squares. *Journal of the American Statistical Association*. 1994; 89:122–127.
- McIntosh AR, Bookstein FL, Haxby JV, Grady CL. Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*. 1996; 3:147–157.
- McIntosh AR, Lobaugh NJ. Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage*. 2004; 23:S250–S263. [PubMed: 15501095]
- Klunk WE, Wang Y, Huang GF, Debnath ML, Holt DP, Shao L, Hamilton RL, Ikonovic MD, DeKosky ST, Mathis CA. The binding of 2-(4'-methylaminophenyl)benzothiazole to postmortem brain homogenates is dominated by the amyloid component. *Journal of Neuroscience*. 2003; 23:2086–2092. [PubMed: 12657667]
- Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, Bergstrom M, Savitcheva I, Huang GF, Estrada S, Ausen B, Debnath ML, Barletta J, Price JC, Sandell J, Lopresti BJ, Wall A, Koivisto P, Antoni G, Mathis CA, Langstrom B. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Annals of Neurology*. 2004; 55:306–319. [PubMed: 14991808]
- Lopresti BJ, Klunk WE, Mathis CA, Hoge JA, Ziolko SK, Lu X, Meltzer CC, Schimmel K, Tsopelas ND, DeKosky ST, Price JC. Simplified quantification of Pittsburgh Compound B amyloid imaging PET studies: a comparative analysis. *Journal of Nuclear Medicine*. 2005; 46:1959–1972. [PubMed: 16330558]
- Mathis CA, Wang Y, Holt DP, Huang GF, Debnath ML, Klunk WE. Synthesis and evaluation of ¹¹C-labeled 6-substituted 2-arylbenzothiazoles as amyloid imaging agents. *Journal of Medicinal Chemistry*. 2003; 46:2740–2754. [PubMed: 12801237]
- Price JC, Klunk WE, Lopresti BJ, Meltzer CC, Ziolko SK, Hoge JA, DeKosky ST, Mathis CA. Quantitative PIB PET imaging studies of amyloid deposition in subjects with Alzheimer's disease, mild cognitive impairment and healthy controls. *Neuroimage*. 2004; 22 (suppl 2):T49.
- Price JC, Klunk WE, Lopresti BJ, Lu X, Hoge JA, Ziolko SK, Holt DP, Meltzer CC, DeKosky ST, Mathis CA. Kinetic modeling of amyloid binding in humans using PET imaging and Pittsburgh Compound-B. *Journal of Cerebral Blood Flow & Metabolism*. 2005; 25:1528–1547. [PubMed: 15944649]
- Mintun MA, LaRossa GN, Sheline YI, Dence CS, Lee SY, Mach RH, Klunk WE, Mathis CA, DeKosky ST, Morris JC. [¹¹C]PIB in a nondemented population: potential antecedent marker of Alzheimer disease. *Neurology*. 2006; 67:446–452. [PubMed: 16894106]
- Brown, CL.; Breich, RF.; Zoubir, AM. Adaptive m-estimators for robust covariance estimation. *Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*; Brest, France. May 2005;
- Wang N, Raftery AE. Nearest-neighbor variance estimation (NNVE): robust covariance estimation via nearest-neighbor cleaning. *Journal of the American Statistical Association*. 2002; 97:994–1006.
- McLachlan, G.; Peel, D. *Finite Mixture Models*. John Wiley & Sons, Inc; NY: 2000.
- Hall P, Fisher NI, Hoffmann B. On the nonparametric estimation of covariance functions. *The Annals of Statistics*. 1994; 22:2115–2134.
- Cohen AC. Estimation in mixtures of two normal distributions. *Technometrics*. 1967; 9:15–28.

21. Joe, H. Multivariate Models and Dependence Concepts. Chapman & Hall; London, UK: 1997.
22. Nelsen, RB. An Introduction to Copulas. Springer-Verlag; New York: 1999.
23. Brechmann, EC.; Schepsmeier, U. Modeling dependence with C- and D-vine copulas: The R-package CDVine. 2011. URL <http://CRAN.R-project.org/package=CDVine>
24. Schepsmeier, U.; Brechmann, EC. R package version 2.8.1. 2011. CDVine: Statistical inference of C- and D-vine copulas.
25. Hoeffding W. Masstabinvariante Korrelations-theorie. Schr Math Inst Univ Berlin. 1940; 5:181–233.
26. Block HW, Fang Z. A multivariate extension of Hoeffding's Lemma. The Annals of Probability. 1988; 16:1803–1820.
27. Cuadras CM. Correspondence analysis and diagonal expansions in terms of distribution functions. Journal of Statistical Planning and Inference. 2002; 103:137–150.
28. Davis, PJ.; Rabinowitz, P. Methods of Numerical Integration. 2. Academic Press; Orlando: 1984.
29. Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical Recipes in C The Art of Scientific Computing. 2. Cambridge University Press; New York: 1992.
30. Cohen AD, Price JC, Weissfeld LA, James J, Rosario BL, Bi W, Nebes RD, Saxton JA, Snitz BE, Aizenstein HA, Wolk DA, DeKosky ST, Mathis CA, Klunk WE. Basal cerebral metabolism may modulate the cognitive effects of A β in mild cognitive impairment: an example of brain reserve. The Journal of Neuroscience. 2009; 29:14770–14778. [PubMed: 19940172]
31. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12:189–198. [PubMed: 1202204]
32. Ziolko SK, Weissfeld LA, Klunk WE, Mathis CA, Hoge JA, Lopresti BJ, DeKosky ST, Price JC. Evaluation of voxel-based methods for the statistical analysis of PIB PET amyloid imaging studies in Alzheimer's disease. Neuroimage. 2006; 33:94–102. [PubMed: 16905334]

Figure 1.
Boxplots of bias for simulation studies with misclassification and $n=50$

Figure 2.
Boxplots of $\sqrt{\text{MSE}}$ for simulation studies with misclassification and $n=50$

Figure 3.
Boxplots of bias for simulation studies with misclassification and $n=200$

Figure 4.
Boxplots of $\sqrt{\text{MSE}}$ for simulation studies with misclassification and $n=200$

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

Mean and variance values used to generate data for misclassification

Sim	misclass	μ_x	σ_x^2	$Cov(x_h, x_k)^*$	μ_{y0}	σ_{y0}^2	σ_{y1}^2	$Cov(y_p, y)_{h0}^*$	$Cov(y_p, y)_{h1}^*$
1	40%	1	.1	.05	1	2	.4	.05	.1
2	20%	1	.1	.05	1	2	.4	.05	.1
3	30%	1	.2	.05	1	2	.3	.05	.1
4	15%	1	.2	.05	1	2	.3	.05	.1

Note: Sim=simulation,

* $h \quad k, j \quad l$

\$watermark-text

\$watermark-text

\$watermark-text

Table 2

Values used for Uniform distribution to generate correlation matrix

Sim	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
1										
X	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$
	$b=-.3$	$b=-.3$	$b=.05$	$b=-.15$	$b=-.25$	$b=-.3$	$b=-.15$	$b=-.1$	$b=-.2$	$b=-.3$
2, 3, & 4										
X	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$	$a=0$
	$b=-.4$	$b=-.3$	$b=.05$	$b=-.15$	$b=-.25$	$b=-.3$	$b=-.15$	$b=-.1$	$b=-.2$	$b=-.3$

Note: $Corr(X, Y_j) \sim U(a, b), X = (X_1, \dots, X_{10}), j = 1, \dots, 10, \text{Sim} = \text{simulation}$

Correlation is randomly generated separately for each group 20% of correlated values are simulated to be negative

Table 3

Results of misclassification simulation studies: frequency of covariance estimates with smallest bias and smallest $\sqrt{\text{MSE}}$ that were observed for a given method

	Frank	Gauss	Gumbel	Cop_Mix	MIX	STAN
N=50						
Simulation 1						
Freq of Est with Smallest Bias	13	15	2	38	11	21
Freq of Est with Smallest $\sqrt{\text{MSE}}$	90	0	10	0	0	0
Simulation 2						
Freq of Est with Smallest Bias	13	15	2	39	16	15
Freq of Est with Smallest $\sqrt{\text{MSE}}$	90	0	10	0	0	0
Simulation 3						
Freq of Est with Smallest Bias	10	22	1	32	23	12
Freq of Est with Smallest $\sqrt{\text{MSE}}$	42	0	58	0	0	0
Simulation 4						
Freq of Est with Smallest Bias	9	23	1	32	22	13
Freq of Est with Smallest $\sqrt{\text{MSE}}$	41	0	59	0	0	0
N=200						
Simulation 1						
Freq of Est with Smallest Bias	4	28	3	37	14	14
Freq of Est with Smallest $\sqrt{\text{MSE}}$	50	0	47	3	0	0
Simulation 2						
Freq of Est with Smallest Bias	4	29	3	37	13	14
Freq of Est with Smallest $\sqrt{\text{MSE}}$	49	0	48	3	0	0
Simulation 3						
Freq of Est with Smallest Bias	4	26	4	50	7	9
Freq of Est with Smallest $\sqrt{\text{MSE}}$	37	4	55	4	0	0

\$watermark-text

\$watermark-text

\$watermark-text

	Frank	Gauss	Gumbel	Cop_Mix	MIX	STAN
Simulation 4						
Freq of Est with Smallest Bias	3	25	4	51	9	9
Freq of Est with Smallest $\sqrt{MSE}$	37	4	55	4	0	0

Table 4

For each method with AD data: Percentage of covariance for 1st–3rd components of PLS and classification rates for discriminant analysis

Component									
	1st	2nd	3rd	kappa	FDG Classif. rate (Cont/MCI/AD)	kappa	PIB Classif. rate (Cont/MCI/AD)	Overall Kappa	
Copula *									
Frank	96%	2%	1%	0.30	0.59/0.40/0.67	0.58	0.78/0.47/1.00	0.39	
Gauss	97%	2%	<1%	0.36	0.59/0.53/0.67	0.58	0.78/0.47/1.00	0.41	
Gumbel	96%	2%	1%	0.33	0.59/0.47/0.67	0.58	0.78/0.47/1.00	0.39	
Cop Mix	96%	2%	1%	0.31	0.57/0.47/0.67	0.58	0.78/0.47/1.00	0.38	
MIX	81%	18%	1%	0.11	0.27/0.33/0.75	0.53	0.78/0.47/0.83	0.18	
STAN	98%	1%	<1%	0.23	0.57/0.27/0.67	0.58	0.78/0.47/1.00	0.34	
Copula **									
Frank	96%	3%	1%	0.36	0.59/0.53/0.67	0.55	0.81/0.47/0.83	0.41	
Gauss	97%	2%	<1%	0.36	0.59/0.53/0.67	0.58	0.78/0.47/1.00	0.41	
Gumbel	96%	2%	1%	0.33	0.59/0.47/0.67	0.58	0.78/0.47/1.00	0.39	
Cop Mix	96%	2%	1%	0.31	0.57/0.47/0.67	0.58	0.78/0.47/1.00	0.38	

Note:

* Only PiB is treated as mixture in copula

** Both PiB and FDG are treated as mixtures in copula

Classif. rate=Classification rate, cont=control