# Introduction

So I explored the properties of the gaussian copula further.

I looked at three cases with varying values for the unidentifiable correlations. For each set of unidentifiable correlations, the identifiable correlations are varied to see what the effect is on the association between delta S and delta T.

The association is assessed by the regression of delta T on delta S, and by the individual causal association. The individual causal association is quantified with two measures. One based on information theory. The second one, Rho delta, directly follows form the copula parameters.

# Joint model

The joint distribution of the vector of potential outcomes is modelled with a gaussian copula model. So the association between the potential outcomes is captured by the correlation matrix. Only two of the six correlations in Sigma are identifiable. The marginal distributions are fully identifiable. So in principle, the best fitting marginal distribution for each potential outcome can be used.

# Marginal distributions

In all that follows, I used marginal Weibull distributions with a positive overall treatment effect on both the surrogate and the true endpoint.

# Case 1

For the first case, I considered small positive unidentifiable correlations. The identifiable correlations were then varied from zero to 0.95.

The black lines correspond to the regression of delta T on delta S. The red lines correspond to the 95% prediction interval. In the plot titles, the values for the identifiable correlations are given. So clearly, if the identifiable correlation increases the slope of the regression line increases and the prediction interval becomes more narrow.

# Case 2

For the second case, I considered larger positive unidentifiable correlations. The identifiable correlations now vary from zero to 0.94, and not 0.95 because that results in a negative definite matrix.

For the regression lines, the same holds as for case 1 although the prediction intervals are somewhat more narrow as compared to case 1.

# Case 3

For the third case, I considered negative unidentifiable correlations. The identifiable correlations now again vary from 0 to 0.94 because 0.95 results in a negative definite matrix.

The same general conclusions hold for the regression lines, although the prediction intervals are wider than in case 1 and 2.

# Measures for ICA: R\_h

The first measure for the individual causal association is based on information theory. So this measure is defined as the informational coefficient of correlation, the same what has been proposed for different types of outcomes.

To compute this measures, the mutual information is needed which is computed by numerical integration. This is quite computer intensive, because the integral in the definition of the mutual information is computed numerically, but also the distributions of the causal effects. This is I think an important disadvantage for the practical use.

In the table, the information coefficient of correlations are given for the different cases. These correspond to the observations for the regression lines and prediction intervals.

# Measure for ICA: rho\_delta

A second possible measure for the ICA is the Pearson correlation between the individual causal effects on the standard normal transformed outcomes. This directly follows from the correlation parameters of the gaussian copula. So this requires no numerical integration, which is a big advantage is compared to the measure based on information theory. This measure also results in the same ranking for the strength of association for the different cases, except for when the identifiable correlations are zero.

Although, the interpretation is here less clear because a positive individual causal effect on the surrogate does no necessarily imply a positive causal effect on the normal transformed surrogate endpoints.

In this tables, this measure is given for the different situations. The value of this measures is generally a bit larger than the first measure, but as I said, the relative ranking is the same.

# To Do

So what I intend to do further is to implement a sensitivity analysis in which the unidentifiable parameters are varied. Then I think I can apply this on a real data set. I think it would be best to use a dataset on which other approaches for evaluating surrogacy have been applied. In that way the results can be compared with these other approaches.

The issue of time-orderings is not addressed with this model, but I think this model is still suitable for cases when the surrogate is almost never censored by the true endpoint. In essence, when the surrogate events happen much earlier than the true endpoint events.

# Time orderings

To explicitly include the aspect of time orderings, I think modifying the gaussian copula model would be difficult. I think we could use a illness-death type of model. So we model the hazard for progression and death; and the hazard for death given that progression has occurred.

Association within treatment groups can be induces with a shared frailty.

Association between the treatment groups can also be induced by a frailty. But this is of course not identifiable. So the we could do a sensitivity analysis in which we sample the parameters of this frailty distribution, and then fix these parameters and fit the model with these parameters fixed.