The Gaussian Graphical Model in Cross-Sectional and Time-Series Data

Gaussian Graphical Model (CGM: an undirected network of partial correlation coefficients). The goal of the GGM is *exploratory data analysis*. The GGM shows *which variables predict one-another,* *and may highlight potential causal relations.* GGM was applied to 1) independent consecutive cases (cross-sectional data), 2) temporally ordered datasets (experience sampling method), and 3) a mixture of the two.

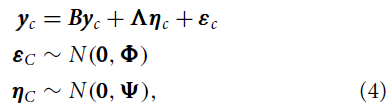
Three network models can be obtained using the residual structure of VAR model; the temporal and contemporaneous network, and the between-subjects network. The last one is a *novel* contribution of the paper. This one is applied to data 3): a time-series measurements of multiple subjects.

The GMM estimates a network of *partial correlation coefficients*, the correlation between two variables after all other variables have been conditioned out.

# The GGM

The GGM in its simplest form: Y\_C ~ N(0, Sigma) is a multidimensional vector, where C represents a case (which can be anything, one subject with multiple (repeated or different) measurements, multiple subjects in the same year, etc.). Assuming MV normality, Sigma encapsulates how all variables are related to one another. Taking the inverse reveal the precision matrix K. This can be standardized into partial correlation coefficients k\_ij. These k\_ij are between -1 and 1 and can be graphically displayed in an undirected weighted graph. Graphs can be sparsified by thresholding |k\_ij|. The graph can be interpreted as modeling conditional associations.

The edges on a GMM can be interpreted as a 1) *predictive effect,* 2) *indicative of causal effects*, although the direction of the effect is lost, 3) *causal generating model*.



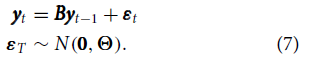
B is causal matrix, lambda is factor-loading matrix, Phi is diagonal matrix, Psi can be modelled in different ways. B is for causal, Lambda is for factor analysis. If Lambda is 0, then estimating only B causes a problem, as there are m variables for m(m-1) parameters in B, making the model *under identified*. Multiple models are not statistically distinguishable.

Using K rather than B, we do not have these issues. An edge in a GGM emerges as a result of a direct causal effect between the variables, or as a result of the fact that both variables have a common effect on a third variable.

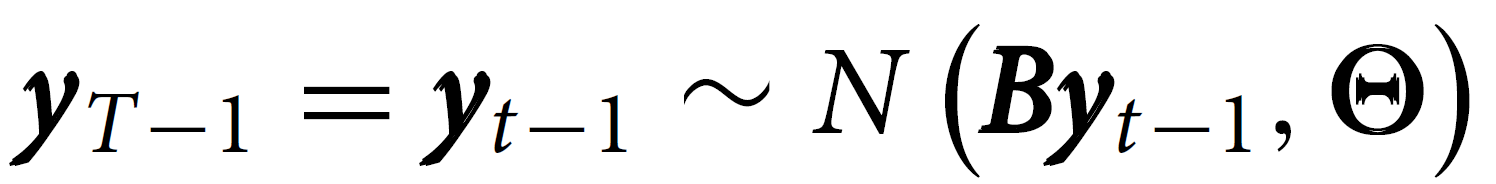
# Estimating GMM

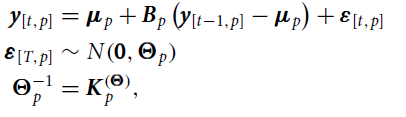
For data with independent cases (e.g. cross-sectional data), we have y\_p ~ N(0, Sigma). Sigma can be estimated using MLE, to sparsify the network, regularization techniques can be used, such as (G)LASSO.

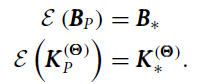
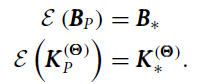
For temporally ordered data of a single subject, we have e.g. a VAR model:



The matrix B is the *temporal* network. The matrix K = inv(Theta) is the *contemporaneous* network. An edge in the *temporal* network indicates that a node predicts another node (or itself in the common case of selfloops) at the next measurement occasion, after controlling for all other variables at the previous measurement occasion.

Interesting perspective of VAR model: , which is equivalent to (4) and is identical to the GMM model when B = 0, so the VAR model is an inclusion of temporal effects on a GGM. An edge in the *contemporaneous* network indicates a conditional association, just as in the GMM.

For temporally ordered data of multiple subjects, we can estimate a GVAR model (7) per subject: 

Where now there is also a vector mu\_p ~ N(0, Omega). We can investigate the individual networks at a second level:

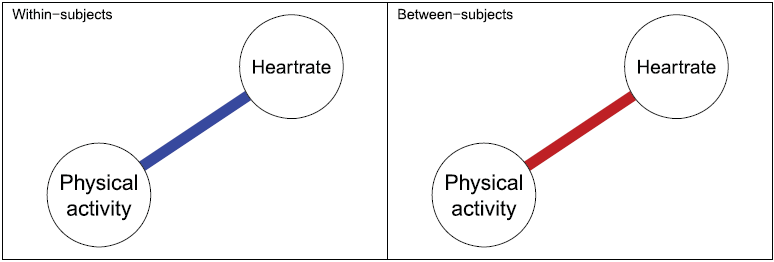
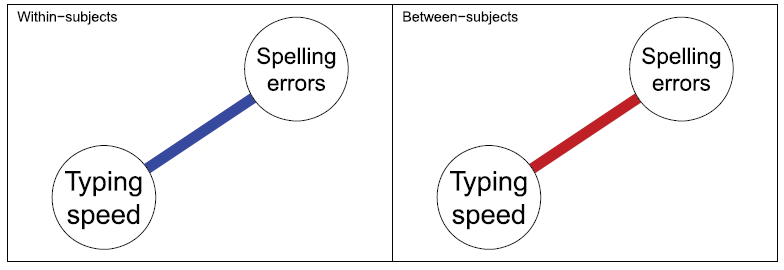
B\* is the temporal network, K(theta) is the contemporaneous network, K(omega) = inv(Omega) is the between-subjects network.

Estimation is not in the digested literature.

# Interpreting GMM from between-subject data

Cross-sectional data analysis, where we have multiple subjects with no repeated measurements, have a within-subject network and a between-subject network. This can, however, not be properly separated because there is only one measurement. The GMM estimate becomes K = inv(Sigma\* + Omega), but the between and within subject networks cannot be disentangled.

When there are multiple measurements, we can separate between and within:



Blue is positive partial correlation, red is negative. Within a subject, typing faster than your average typing speed results in more spelling errors. Within a subject, increasing your physical activity raises your heart rate.

Between a subject, people who type faster on average, tend to have fewer spelling errors (because they are more skilled, e.g. a cleric or translator). Between a subject, people who are more physically active, tend to have a lower heart rate on average. These subtle interpretation details are important.

If we are looking at finding causal relations due to interventions, we should look at the within-subjects level, the between-subjects is misleading. However, temporal, structural changes are to be found in the between-subjects network, as the within-subjects network only model deviations from a stationary mean. Hence, evidence for the effect of such interventions may arise at either the within- or the between subjects level depending on the nature of the intervention.

# Examples

Examples of how the GVAR is used in data from Mõttus et al. (4.1) and data from Bringmann et al (4.2).

# Discussion

Multi-level estimation is not yet feasible for large datasets, requires PSD matrices, and is computationally infeasible for > 8 variables.

The VAR modeling assumptions (y\_t | y\_t-1 is equal distr. for all *t*) give rise to issues. Furthermore, interpreting temporal coefficients (causal / predictive / etc.) is not without discussion.

VAR also assumes normality which may easily be violated. When the process itself is nonnormal, such as skewed residuals, the entire modeling framework does not correctly capture the likelihood. Var also assumes linearity in effects.

The interpretation of edges do not confirm causal relations; merely hypotheses. The networks show how variables predict each other over time (temporal), within time (contemporaneous), and on average (between-subjects).

# Conclusion

This paper provides a methodological overview of how the GGM can be used in various kinds of psychological data. The GGM can be used to map out unique variance in cross-sectional data or at the contemporaneous and between-subjects’ levels of time-series analysis. While losing information on the direction of effect, estimating GGMs offers an attractive alternative in that these models are uniquely identified, well parameterized, closely related to causal models and also offer exploratory insight on predictive observational and experimental data The GGM thus provides a powerful addition to the exploratory toolbox in behavioral research.