Text S1: Proof of Convexity of Sparse CGGM Optimization Problem

Proposition 1. *The optimization problem for learning a sparse CGGM given as below is convex:*

$$\operatorname*{argmin}_{\mathbf{\Theta_{yy}},\mathbf{\Theta_{xy}}} L(\mathbf{\Theta_{xy}},\mathbf{\Theta_{yy}};\mathbf{X},\mathbf{Y})+\lambda_1||\mathbf{\Theta_{xy}}||_1+\lambda_2||\mathbf{\Theta_{yy}}||_1, \tag{1}$$

where $L(\mathbf{\Theta_{xy}},\mathbf{\Theta_{yy}};\mathbf{X},\mathbf{Y}) = 1/2\,\mathrm{tr}\big(\mathbf{Y}\mathbf{\Theta_{yy}}\mathbf{Y}^T\big) + \mathrm{tr}\big(\mathbf{X}\mathbf{\Theta_{xy}}\mathbf{Y}^T\big) + \sum_i \log Z(\mathbf{\Theta_{xy}},\mathbf{\Theta_{yy}},\mathbf{x}^i)$ *is the negative log-likelihood of data.*

Proof. The L_1 penalty is convex and we only need to show that the negative data log-likelihood $L(\mathbf{\Theta_{xy}}, \mathbf{\Theta_{yy}};\mathbf{X},\mathbf{Y})$ in Eq. (1) is convex. To prove the convexity of $L(\mathbf{\Theta_{xy}},\mathbf{\Theta_{yy}};\mathbf{X},\mathbf{Y})$, we use the standard approach for proving the convexity of a function by showing that the second derivative of the given function is positive definite [30]. We notice that a CGGM is a special case of a log-linear model. In addition, it is a standard result that the second derivative of negative data log-likelihood for a log-linear model is an expected feature covariance matrix [29]. Since a covariance matrix is always positive definite, the negative data log-likelihood for CGGM is convex. $\square$