Topics on associations among random processes

by

Michael Patrick McAssey

B.A. (San Francisco State University) 1985
M.A. (San Francisco State University) 2004

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA at DAVIS

Approved:

_____

Fushing Hsieh, Chair

_____

Francisco J. Samaniego

_____

Emilio Ferrer-Caja

Committee in Charge
2011

Acknowledgements

I am deeply grateful to my thesis advisor, Fushing Hsieh, for his wisdom and guidance throughout the research phase of my graduate studies in statistics at UC Davis. Our frequent—almost daily—meetings to discuss the progress of our joint research, to consider various ways to cleverly approach a problem, and to critique my results have been invaluable. Through these sessions I have learned how to think like a statistician, which has often meant breaking away from established methods and models in order to analyze data sets that do not lend themselves readily to traditional approaches. Professor Hsieh has taught me the importance of building collaborative relationships with researchers in other disciplines, and has impressed upon me that the most important role of the statistician is to help scientists use their data to solve real problems. Above and beyond his professional role, he has also been a good friend whose gentle encouragement has helped me persevere.

I would also like to express my appreciation to our collaborators, most especially Emilio Ferrer-Caja, whose psychological data formed the basis for several of my research projects, and Anne Smith, who worked with us to obtain and analyze the EEG

data from the neuroscientists at UCSF. I am grateful to Loren Frank and Margaret Carr of the Department of Physiology at the University of California, San Francisco, for generously allowing us to use their hippocampal EEG data, and the corresponding filtered signals. I also thank Kari Kuulasmaa for making the WHO MONICA data available, and Alexandre Patriota for sharing that data with me.

While the entire statistics faculty at UC Davis have contributed greatly to my learning, my relationship with Frank Samaniego has had the most profound impact on my professional development. Through my work with him both as his teaching assistant and as a typesetter of his books, articles and conference talks, I have had the ongoing benefit of his mentorship and his example. I also acknowledge the contributions of my instructors and advisors within the statistics program: George Roussas, Wolfgang Polonik, Prabir Burman, Hans Müller, Jane-Ling Wang, Rudy Beran, Bob Shumway, Peter Hall, Duncan Temple Lang, Ed Mack, Debashis Paul, Alexander Aue, Jie Peng, Debashis Paul, Rahman Azari and Soma Roychowdery. Among the staff, I am especially grateful to our graduate coordinator, Pete Scully, to our Management Services Officer, Pat Aguilera, and to our information technician, Larry Tai, for their support.

ABSTRACT

This dissertation focuses on innovative techniques for the statistical analysis of concurrent random phenomena. In particular, new methods are presented for estimating measures of the relationships among multiple random variables or processes based on data generated from a common population. Of special interest are multiple random processes, in which the relationships among data sources are in constant transition among several states or different levels of the same state. Such a scenario is considered in the analysis of EEG data recorded from the brain of a rat, in which a technique is developed to measure instants of coupling between pairs of signals generated from different brain regions, and model the evolution of brain activity in terms of several instantaneous coupling states. A similar idea is employed to measure the evolution of synchrony in physiological measures among couples while performing assigned tasks. An innovative method is designed for measurement of the linear association between two variables in the presence of heteroscedastic measurement error, and in the process a clever test for the misapplication of linear models is presented. The benefits of the improved model are illustrated using public health data. This idea is also extended to the aforementioned study of synchrony in physiological measures. These methodologies provide new tools for scientists engaged in research well beyond the applications to neuroscience, public health and psychology highlighted in this dissertation. Finally, a study involving a specified mode of signal transmission

in networks is presented.

The first chapter consists of the study of the EEG data. An important goal in neuroscience is to identify instances when components of EEG signals are coupled. A method is developed to measure the coupling strength between gamma signals on a short time scale as the maximum cross-correlation over a range of time lags within a sliding variable-width window. Instances of coupling states among several signals are also identified, using a mixed multivariate beta distribution to model the coupling strength across multiple gamma signals with reference to a common base signal. The EM algorithm is implemented to determine the number of states and their respective model parameters. The variable-window method is first applied to simulated signals and its performance is compared to a fixed-window approach. The method is then applied to gamma signals recorded in two regions of the rat hippocampus. The results indicate that this may be a useful method for mapping coupling patterns among signals in EEG datasets.

In the second chapter, the line-segment parametrization of the structural measurement error model is extended to situations in which the error variance on both variables may not be constant over all observations. Under these conditions, a method-of-moments estimate of the slope is developed, and its asymptotic variance is derived. An accurate estimator of the variability of the slope estimate based on sample data is then derived in a rather general setting. Simulations are performed which validate these results and demonstrate the superiority of these estimates when the measure-

ment error variance is not small. Lastly, this estimation approach is illustrated using real data involving heteroscedastic measurement error, and its robustness against influential points is demonstrated. However, while the estimated trend based on the line-segment approach agrees well with intuitive expectations, the estimates based on conventional models which incorporate elements of the simple linear model do not. A simulation is employed to illustrate the danger of estimating a linear trend using the structure of the simple linear model without justification.

The third chapter presents two new approaches for identifying synchrony between the physiological signals of individuals in dyads. The approaches are adaptations of two recently-developed techniques, depending on the nature of the physiological time series. For respiration and thoracic impedance, which are measured continuously, the Empirical Mode Decomposition is employed to extract low-frequency components of their noisy nonstationary signals. Then the maximum cross-correlation between the two denoised signals is computed within consecutive overlapping time windows of fixed width throughout each of the experimental tasks, and the relative frequency of large values of this measure during each task is calculated. For the heart rate, which is output discretely, the structural linear regression model that takes into account heteroscedastic measurement error on both the variables is used. The results indicate that these methods are effective in detecting synchrony between physiological measures and can be used to examine emotional coherence in dyadic interactions.

In the fourth chapter, a mode of signal transmission within a network is specified,

and it is shown that this mode results in one of two possible permanent states based on the configuration of the network. Both geometric and algebraic criteria are developed to allow one to determine which state a network will reach. Then, given a random network in which the edges have specified probabilities of occurring and a budget which limits the sum of these probabilities, the simulated annealing algorithm is employed to find optimal allocations of probabilities among the edges which conform to the budget and maximize the probability that a network which possesses desired properties under this mode of signal transmission will be realized.

Contents

List of Tables

List of Figures

Chapter 1

Coupling among electroencephalogram (EEG) gamma signals on a short time scale

1.1 Introduction

Current neuroscience research is focused not only on identification of brain regions associated with particular cognitive tasks but also on how those regions interact during the execution of the these tasks on a short time scale [4, 49, 37]. This chapter consists of an investigation of methods for identifying brief instances in time when groups of

gamma-band signals (40–100 Hz) extracted from electroencephalogram (EEG) recordings become synchronized. Methods are developed to address the specific problem of analyzing EEG recordings from the rat hippocampal formation. Recent studies [49] have demonstrated dynamic coordination at these frequencies between the dentate gyrus, CA1 and CA3 during tasks with high cognitive demand and during REM sleep.

Typically in situations where a signal is suspected to be changing with time, methods from signal analysis, such as the short-time Fourier transform, can be applied quite successfully. For pairs of signals, *coherency*, a measure of signal coherence in the frequency domain, can be computed with confidence bounds, using for example a multi-taper [59]. In both of these cases software has been developed to compute these quantities. For example, the `spectrogram` function in `Matlab` [45] computes the short-time Fourier transform along segments of an individual signal, and the `coherence` function in the `Chronux` package [48] computes the windowed coherence between pairs of signals. However, for signals with synchrony lasting just a few cycles at a time, frequency-domain methods can be less sensitive and may yield unrealistically large confidence bounds. Alternative methods for detecting synchronization of neuroelectric signals are based on transient phase-locking [62, 41]. For the case of many EEG signals, other approaches include Granger causality [25], autoregressive modeling [22], and Bayesian networks [57]. However, these may be limited in their ability to detect only brief instances of synchrony.

One of the challenges in analysis of biological rhythms is that the signal frequency can be quite variable. A method is developed here that computes synchrony among

multiple signals on the time scale of milliseconds and which yields, not just pairwise computations, but a joint result for all signals. This is accomplished by making use of a small time-varying sliding window to compute pairwise cross-correlations, and then by using the Expectation-Maximization (EM) algorithm [13] applied to a mixed multivariate beta model to identify groups of gamma signals that are highly synchronized at any instant.

This technique consists of two steps. First, one of the available gamma signals is selected as the basis for providing the sliding measurement window and the temporal axis for computation of the instantaneous coupling between that dimension and each of the other dimensions. Second, the EM algorithm is implemented to classify the collection of instantaneous coupling measurement vectors into a fixed number of states representing different occasions of gamma-band binding among brain regions.

The effectiveness and reliability of these methods are tested on simulated data. The technique is then applied to a nine-channel EEG data set recorded from tetrodes implanted in the Medial Entorhinal Cortex (MEC) and the CA1 cell layer of the hippocampus in a rat's brain.

1.2 Methods

Assume that subsets of gamma signals extracted from EEG recordings using a band-pass filter are subject to instants of synchrony on the order of a few cycles, after which they become unsynchronized. This phenomenon is termed *instantaneous coupling* (IC), and a method for quantifying it is developed.

1.2.1 Computation of instantaneous coupling between two EEG signals

Since synchrony is sought on very short time scales and since biological signals are prone to variability, the focus here is on finding an appropriate time scale for computing coupling between pairs of signals that adjusts over time based on the changing frequency of the signals.

The first goal is to compute a sequence of IC estimates between two bandpassed oscillating time signals in the gamma range (40–100 Hz), say $X = \{x_t\}_{t=1}^{T}$ and $Y = \{y_t\}_{t=1}^{T}$, throughout a given epoch consisting of T time points. The approach here is similar to the procedure given in [38], but using a time-varying window. The amplitude, frequency, and phase of each signal vary from one instant to the next, not necessarily independently. Choose one of the signals, say X, as the base signal. Then partition the entire epoch based on the N zero crossings of X, which are denoted by $Z_1, Z_2, \ldots, Z_N$, with $N << T$. Regard each interval $[Z_i, Z_{i+2}]$, $i = 1, 2, \ldots, N-2$, as a full cycle, and thus $[Z_i, Z_{i+1}]$ is a half-cycle, $i = 1, 2, \ldots, N-1$. The duration of a cycle may thus vary significantly throughout the epoch.

Let the integer w indicate the window size, i.e., the number of half-cycles of the base signal to be used in determining the duration of an "instant" when estimating the instantaneous coupling between X and Y. Choosing a small value for w not only enables one to discuss the dynamics of signal coupling among brain regions on a very short time scale, but also provides approximate stationarity in the signals. However, if w is chosen too small, the IC may be overestimated due to the sparsity of information. If w is chosen too large, the IC may be underestimated. This issue is

explored in the simulation study.

Next, define the *estimated instantaneous coupling* between X and Y, *with respect to* the base signal X, during the interval $[Z_i, Z_{i+w}]$, where $i = 1, 2, \ldots, N - w$, as the maximum of the cross-correlation between X and Y over this interval across a range of lags h, i.e.,

$$IC_{X,Y}([Z_i, Z_{i+w}]) = \max_h \frac{\sum_{t=Z_i}^{Z_{i+w}} (x_t - \overline{x}_i)(y_{t+h} - \overline{y}_i)}{\sqrt{\sum_{t=Z_i}^{Z_{i+w}} (x_t - \overline{x}_i)^2 \sum_{t=Z_i}^{Z_{i+w}} (y_{t+h} - \overline{y}_i)^2}} ,$$

where

$$\overline{x}_i = \frac{\sum_{t=Z_i}^{Z_{i+w}} x_t}{Z_{i+w} - Z_i} \qquad \text{and} \qquad \overline{y}_i = \frac{\sum_{t=Z_i}^{Z_{i+w}} y_{t+h}}{Z_{i+w} - Z_i} ,$$

for $i = 1, 2, \ldots, N - w$. In the data analysis here, the `ccf()` function in the `R` package [56] is used to compute each cross-correlation over the function's default lag range. In general, this range should run a little more than one half-cycle of the base signal in each direction, i.e., $\approx \pm(Z_{i+w} - Z_i)/w$, for each window $[Z_i, Z_{i+w}]$. This will translate a strong negative correlation into a strong positive one, and also ensures that the maximum cross-correlation between the signals will be positive, or very close to zero if negative.

Besides choosing a value for w, one must also decide how much overlap between consecutive intervals to allow. Different choices for the overlap parameter will affect the degree of smoothness for the computed IC along the temporal axis, but not its value. One could choose no overlap, so that $IC_{X,Y}([Z_i, Z_{i+w}])$ is computed on consecutive adjacent windows $[Z_1, Z_{1+w}]$, $[Z_{1+w}, Z_{1+2w}]$, ... On the other hand, one may choose overlapping windows by selecting some positive integer m, with $1 \leq m < w$, so that $IC_{X,Y}([Z_i, Z_{i+w}])$ is computed on overlapping intervals $[Z_1, Z_{1+w}]$,

$[Z_{1+m}, Z_{1+m+w}]$, $[Z_{1+2m}, Z_{1+2m+w}]$, ... In the real data analysis here, a window size of $w = 6$ and an increment size $m = 2$ are selected, so that each consecutive window pair overlaps by two cycles. This choice is made empirically based on consideration of the apparent duration of synchrony among the plotted signals.

This method also requires the selection of a base signal. When two gamma signals are synchronized, their time scales are roughly the same, so that they are at about the same frequency. Hence the IC between X and Y should be approximately symmetric, so that one will obtain roughly the same IC estimate within each time window, irrespective of which signal is chosen as the base. Any difference will only be due to slightly different measurement windows. The rationale for choosing the base signal depends on whether or not one wishes to analyze the coupling between one specific brain location and several additional locations. If a neuroscientist is only interested in the evolution of synchrony from the perspective of one brain location, then the choice for the base is clear. But knowing the IC between base signal X and signal Y and the IC between X and a third signal Y' does not provide any information about the IC between Y and Y'. To determine the latter using this approach, one must choose either Y or Y' as the base signal and proceed accordingly. To obtain a complete analysis of the evolution of coupling among all pairs in a set of gamma signals, one would have to repeat the computations with each gamma signal taking its turn as the base. This level of analysis is not explored in the present study.

Approximate confidence bounds for the true value of the IC corresponding to each estimated IC measurement may be computed as follows. Let ρ^* denote this true value,

and let r^* denote its estimate. For simplicity, assume that both values occur at the same lag h. If one applies Fisher's Z-transformation

$$\zeta = \frac{1}{2}\log\left[\frac{1+\rho^*}{1-\rho^*}\right], \qquad z = \frac{1}{2}\log\left[\frac{1+r^*}{1-r^*}\right],$$

and uses the established result that, when the observations are independent, the distribution of $\sqrt{n-1}(z-\zeta)$ approaches that of the standard Gaussian for large values of n, one may then compute approximate confidence bounds for ζ, and hence for ρ^*, although the independence condition is violated here. That is, an approximate $(1-\alpha)100\%$ confidence interval for the true value of the IC is

$$\left(\frac{\mathrm{e}^{2L}-1}{\mathrm{e}^{2L}+1}, \quad \frac{\mathrm{e}^{2U}-1}{\mathrm{e}^{2U}+1}\right),$$

where

$$L = \frac{1}{2}\log\left[\frac{1+r^*}{1-r^*}\right] - \frac{z_{\alpha/2}}{\sqrt{n-1}} \quad \text{and} \quad U = \frac{1}{2}\log\left[\frac{1+r^*}{1-r^*}\right] + \frac{z_{\alpha/2}}{\sqrt{n-1}},$$

and $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Potential problems with this method for obtaining confidence bounds, and an alternative method using extreme value theory that addresses these problems, are discussed in [37]. Further work must be done in this area, but the following developments do not involve confidence bounds.

1.2.2 Identification of instantaneous coupling states among multiple EEG gamma signals

Now consider the setting in which one has obtained $J+1$ EEG gamma signals recorded from tetrodes implanted in different brain regions. After obtaining individual esti-

mates of the IC between the selected base signal X and each of J signals $Y_1, \ldots, Y_J$ throughout an epoch, an algorithm is employed to identify neurological states in which particular subsets of these signals are more synchronous with the base signal at any given instant. Essentially, this approach assumes that during each instant of time different subsets of gamma signals are more synchronized with the base signal than are the remaining signals, based on the interaction among the brain regions in which the electrodes are implanted. The algorithm is used to estimate which subsets of gamma signals are most synchronous with the base signal in each instant, and also to estimate parameters that describe the distribution of the IC estimates among the signals pertaining to each subset. It is emphasized that this will provide a model of the synchrony among signals *from the perspective of the selected base signal* only. One may choose additional base signals to obtain models from multiple perspectives, and then combine the results. The multiple-perspective angle is not explored in this study.

Since the coupling measure considered here is the maximum cross-correlation between a pair of signals over a range of lags, and hence falls within a bounded interval, one may model the distribution of these maxima with a univariate beta distribution. It is natural to implement a multivariate generalization of the beta distribution to model the joint distribution of the IC among any collection of J gamma signals with respect to any base gamma signal. Specifically, the multivariate beta (MVB) distribution is chosen, with parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J, \theta_{J+1})$, $\theta_j > 0$ for $j = 1, \ldots, J+1$. The derivation of this distribution is given in Section 1.5 at the end of this chapter.

Given the ith observation $\mathbf{U}_i = (U_{i1}, \ldots, U_{iJ})$, with $0 < U_{ij} < 1$ for $j = 1, \ldots, J$, the joint density of $\mathbf{U}_i$ under the MVB distribution is

$$f_{\mathbf{U}_i}(u_{i1}, \ldots, u_{iJ}) = \frac{\Gamma\left(\sum_{j=1}^{J+1} \theta_j\right)}{\prod_{j=1}^{J+1} \Gamma(\theta_j)} \left(\prod_{j=1}^{J} \frac{u_{ij}^{\theta_j - 1}}{(1 - u_{ij})^{\theta_j + 1}}\right) \left(1 + \sum_{j=1}^{J} \frac{u_{ij}}{1 - u_{ij}}\right)^{-\sum_{j=1}^{J+1} \theta_j}, \tag{1.2.1}$$

where $\Gamma(x) = \int_0^\infty t^{x-1} \mathrm{e}^{-t}\, \mathrm{d}t$, $x > 0$, is the gamma function.

To implement this model with the IC estimates for J gamma signals with respect to a common base signal, first replace any non-positive values of the estimated IC with a very small positive value, e.g., 0.00001, so that $U_j > 0$ for each j. Likewise, if any estimated IC value equals one, replace it with 0.99999 so that $U_j < 1$ for each j. In the analysis of real data presented here, very few non-positive values are computed among the IC estimates, and those that are negative are all very close to zero. To avoid computational errors in the evaluation of (1.2.1), due either to large arguments to the gamma function or to computation of the product of very small values, one may compute instead the logarithm of the density in (1.2.1), then exponentiate the result.

To estimate the $J + 1$ components of the parameter vector $\boldsymbol{\theta}$, given the $N' = N - w$ computed IC vectors $\mathbf{U}_i = (U_{i1}, \ldots, U_{iJ})$, $i = 1, \ldots, N'$, using the method of

maximum likelihood, it is necessary maximize the log-likelihood function

$$\begin{aligned}\ell(\boldsymbol{\theta} \mid \mathbf{U}_1, \ldots, \mathbf{U}_{N'}) &= N' \left[\log \Gamma \left(\sum_{j=1}^{J+1} \theta_j \right) - \sum_{j=1}^{J+1} \log \Gamma(\theta_j) \right] \\ &\quad + \sum_{i=1}^{N'} \sum_{j=1}^{J} \left[(\theta_j - 1) \log U_{ij} - (\theta_j + 1) \log(1 - U_{ij}) \right] \\ &\quad - \left(\sum_{j=1}^{J+1} \theta_j \right) \sum_{i=1}^{N'} \log \left(1 + \sum_{j=1}^{J} \frac{U_{ij}}{1 - U_{ij}} \right)\end{aligned}$$

over all $\boldsymbol{\theta} \in (0, \infty)^{J+1}$. This computation is carried out using the Expectation Maximization (EM) algorithm [13].

Next, it is postulated that the N' IC estimates may be grouped into distinct IC states. The concept of *IC states* refers to occasions in which specific subsets of the J gamma signals have an IC with the base signal which is relatively high, perhaps above some threshold. If one thinks of each signal as being either coupled or not coupled with the base signal during any instant, based on some threshold, there would be 2^J possible IC states. However, it is expected that tetrodes located near each other should tend to exhibit relatively equivalent IC levels with respect to any base, so that far fewer distinct states actually occur.

If there are $p \geq 1$ such IC states among the IC estimates $\mathbf{U}_1, \ldots, \mathbf{U}_{N'}$, then a clustering procedure may be to estimate parameters that describe the distinct states and to classify the individual $\mathbf{U}_i$ among them. One possible clustering procedure is the k-means algorithm, in which the vectors $\mathbf{U}_1, \ldots, \mathbf{U}_{N'}$ are randomly assigned among p clusters, and the mean vector for each cluster is computed. The algorithm then reassigns each vector, if necessary, to the cluster whose mean is nearest, in terms of some distance measure. Since the clusters are usually altered by the reassignments,

the algorithm recomputes the mean vector of each cluster, and then conducts any necessary reassignments of vectors to nearer cluster means. This process continues iteratively until no reassignments are necessary (or a maximum number of iterations is reached). The `R` version of the k-means algorithm implements by default the method described in [28]. This method ensures that p clusters are returned. Because the algorithm is sensitive to the initial allocation, it should be run multiple times, and the best result should be chosen in terms of minimum error sum of squares. However, a more sophisticated approach that takes into account the apparent MVB distribution of the IC estimates is implemented in this study.

Since it is assumed that the representative IC vectors $\mathbf{U}_1, \ldots, \mathbf{U}_{N'}$ follow a MVB distribution, a mixture model is implemented which assumes that each $\mathbf{U}_n$ arises independently from one of p IC states, each of which follows its own MVB distribution with its own parameter vector $\boldsymbol{\theta}_k = (\theta_{k,1}; \theta_{k,2}; \ldots; \theta_{k,J+1})$ for $k = 1, \ldots, p$. This model uses the MVB density to assign each $\mathbf{U}_i$ to that IC state for which the probability that it belongs to that state is largest. These probabilities are latent parameters which must be estimated along with the parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ corresponding to the respective IC states. The EM algorithm [13] is suited to this purpose. The independence assumption is approximate here, but when the sample size N' is quite large this estimation procedure should be robust to departures from the assumption. This robustness must be checked in further work.

To implement the EM algorithm, let $P_{\boldsymbol{\theta}_k}(\mathbf{u}_i)$ denote the value of the MVB density corresponding to the kth state evaluated at IC estimate $\mathbf{u}_i = (u_{i1}, \ldots, u_{iJ})$, where

$i = 1, \ldots, N'$, and let π_k denote the probability that a randomly selected IC estimate belongs to IC state k. In the E-step of the EM algorithm, the *responsibility* of the kth state for $\mathbf{U}_i$ at the qth iteration, $q = 1, 2, \ldots$, is determined by computing

$$r_{k,i}^{(q)} = \frac{\pi_k^{(q-1)} P_{\boldsymbol{\theta}_k^{(q-1)}}(\mathbf{u}_i)}{\sum_{l=1}^{p} \pi_l^{(q-1)} P_{\boldsymbol{\theta}_l^{(q-1)}}(\mathbf{u}_i)}\,, \tag{1.2.2}$$

for $i = 1, \ldots, N'$ and $k = 1, \ldots, p$. Then in the M-step, the mixing parameters and the distribution parameters at the qth iteration are estimated by computing

$$\pi_k^{(q)} = \frac{1}{N'} \sum_{i=1}^{N'} r_{k,i}^{(q)}$$

and

$$\boldsymbol{\theta}_k^{(q)} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N'} r_{k,i}^{(q)} \log P_{\boldsymbol{\theta}}(\mathbf{u}_i)$$

for $k = 1, \ldots, p$. Then continue to iterate between the E-step and the M-step until the parameter estimates converge within a pre-specified tolerance. In the real data analysis, the `constrOptim()` procedure in `R` is used with appropriate settings in order to find the optimal value of $\boldsymbol{\theta}$ in the M-step.

The success of the EM algorithm is sensitive to the selection of the initial estimates $\pi_k^{(0)}$ and $\boldsymbol{\theta}_k^{(0)}$. In the data analysis here, the k-means algorithm is used to first obtain p clusters of the IC estimates, then maximum likelihood estimates (MLEs) of the univariate beta parameters are computed for each individual dimension of the J-dimensional IC vector for each of the p clusters. These univariate estimates are combined into one parameter estimate $\boldsymbol{\theta}_k$ for each cluster, and the `constrOptim()` procedure is used to determine the MLE of $\boldsymbol{\theta}_k$ for each cluster. Then $\pi_k^{(0)}$ is the proportion of IC estimates assigned to the kth cluster, and $\boldsymbol{\theta}_k^{(0)}$ is the computed MLE,

for $k = 1, \ldots, p$. In general, this approach consistently provides initial estimates that lead to eventual convergence to an optimal solution.

In order to implement the above procedure, it is necessary to determine the number p of IC states that occur among the gamma signals. The Bayes Information Criterion (BIC) is one common tool used to estimate the true number of clusters p that are represented in the data, if in fact the data are clustered. To use this tool, perform the EM algorithm for $p = 2, 3, \ldots$, and for each value of p obtain the optimal parameter estimates π_k^* and $\boldsymbol{\theta}_k^*$ for $k = 1, \ldots, p$. Then compute the *mixture log-likelihood* (see [19]) at these optimal parameter estimates as follows

$$\ell_{\text{mix}} = \ell\left(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_p^*, \pi_1^*, \ldots, \pi_p^* \mid \mathbf{u}_1, \ldots, \mathbf{u}_{N'}\right) = \sum_{i=1}^{N'} \log\left[\sum_{k=1}^{p} \pi_k^* P_{\boldsymbol{\theta}_k^*}(\mathbf{u}_i)\right] . \quad (1.2.3)$$

The BIC for the model is then computed from:

$$\text{BIC}_p = -2\ell_{\text{mix}} + [p(J+2) - 1]\log(N') , \quad (1.2.4)$$

where J is the number of gamma signals. (Note that for each of the p states one must estimate a $(J+1)$-dimensional parameter vector $\boldsymbol{\theta}_k$ and a mixing probability π_k, for a total of $p(J+2)$ estimated parameters. However, once $\pi_1, \ldots, \pi_{p-1}$ are estimated, the value of π_p is then fixed, since the mixing probabilities sum to one, resulting in $p(J+2)-1$ independent estimated parameters.) Finally, choose that mixture model, over a suitable range of values for p, for which the value of the BIC is minimized. That is, conclude that the IC with the base signal among the other J signals varies among p^* distinct states, where

$$p^* = \underset{p}{\text{argmin}}\, \text{BIC}_p . \quad (1.2.5)$$

Then one may form an IC state sequence corresponding to the zero crossings of the base signal which estimates which sets of EEG gamma signals are synchronized with the base signal within each measurement window as time unfolds.

To summarize, the procedure is as follows:

1. Choose one signal as the base among $J + 1$ gamma signals recorded at tetrodes implanted in different brain regions.

2. Compute the estimated IC between the base signal and the other J signals within a sliding variable-length window throughout an epoch of activity. This results in a time-ordered sequence of J-dimensional IC estimates.

3. For any block of the sequence, use the EM algorithm to maximize the log-likelihood based on a mixed multivariate beta model involving p distinct clusters, where p ranges over a reasonable set of values.

4. For each value of p, use the optimized log-likelihood to assign each IC estimate to one of p IC states and to obtain estimates for the model parameter vector corresponding to each IC state.

5. Select the model for which the BIC is minimized. This model yields an IC state sequence which represents the evolution in the coupling of the different brain regions among the p IC states from moment to moment from the perspective of the selected base signal.

6. Since the sequence of IC estimates, and the subsequent IC state sequence, correspond to the zero crossings of the base signal, the IC states may be mapped

back to the time scale of the EEG gamma signals.

1.3 Results

1.3.1 Simulation studies

To demonstrate the effect of computing estimates of the instantaeous coupling between two EEG gamma signals using a variable window, generate two signals

$$X(t) = \sin\{2\pi[70 + 10\sin(0.5\pi t)]t\} \quad \text{and} \quad X(t) = \sin\{2\pi[50 + 10\sin(0.5\pi(t-2))]t\}\ ,$$

where t varies from 0 to 20 seconds at a resolution of 1500 points per second. This resolution mimics that of the real EEG data. A plot of the first two seconds of this signal pair is shown in Figure 1.1.

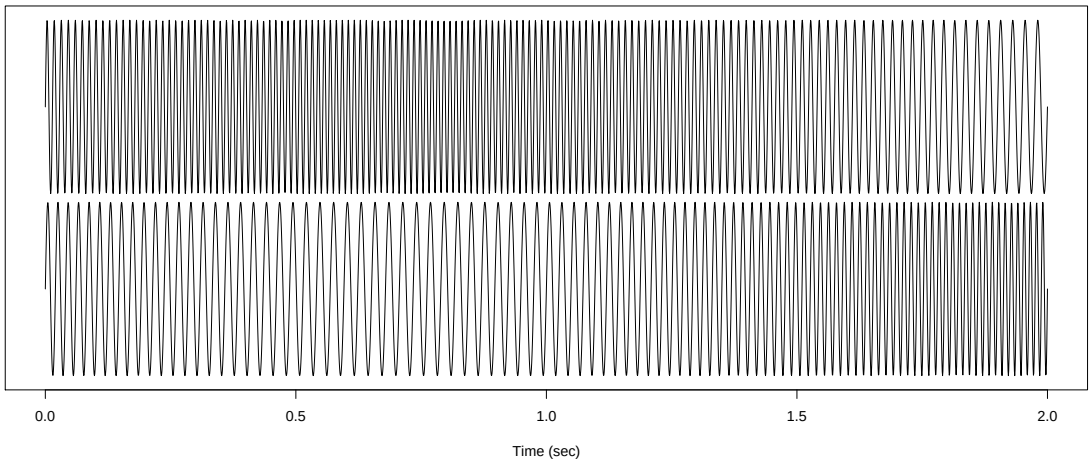


Figure 1.1: Simulated EEG gamma signals $X(t)$ (top) and $Y(t)$ (first two seconds).

Signal X has an instantaneous frequency of $70 + 10\sin(0.5\pi t)$ that oscillates between 60Hz and 80Hz, while the instantaneous frequency $50 + 10\sin(0.5\pi(t-2))$

of signal Y oscillates between 40Hz and 60Hz. Both signals achieve the instantaneous frequency of 60Hz simultaneously. However, the *effective frequency* of each signal — that is, the number of cycles the signal actually experiences per time unit — within any interval is much different from its instantaneous frequency when this instantaneous frequency is not constant, as here. To approximate the effective frequency of either signal at time t, count the number of zero crossings in the interval $[t - 0.05, t + 0.05]$. Since every two zero crossings represents one cycle, divide this count by two. Dividing by 0.1 gives the average frequency of the signal at time t. For the purposes in mind here, this value is not needed for the first or last 0.05 seconds of the 20-second epoch. Next, subtract the average frequency of Y from that of X, and divide by the largest difference in order to place the difference in average frequency on a scale of -1 to 1. A plot of this normalized frequency difference is displayed in Figure 1.2. The IC between X and Y should equal one at points where this plot crosses the horizontal axis, as it is at these points that the two simulated gamma signals become synchronized. The IC should fall to zero elsewhere.

Now apply the variable window technique described in Section 1.2.1 to estimate the IC between signals X and Y, with X chosen as the base signal, using three different values for w, where w is the number of half-cycles of the base signal to determine each measurement window. In each case, choose $m = w/3$ as the increment size. The results are mapped back to the time scale of the signals, and displayed in Figure 1.3 along with the normalized frequency difference from Figure 1.2. With $w = 3$, the estimated IC is close to one whenever the frequency difference is near zero, as desired,

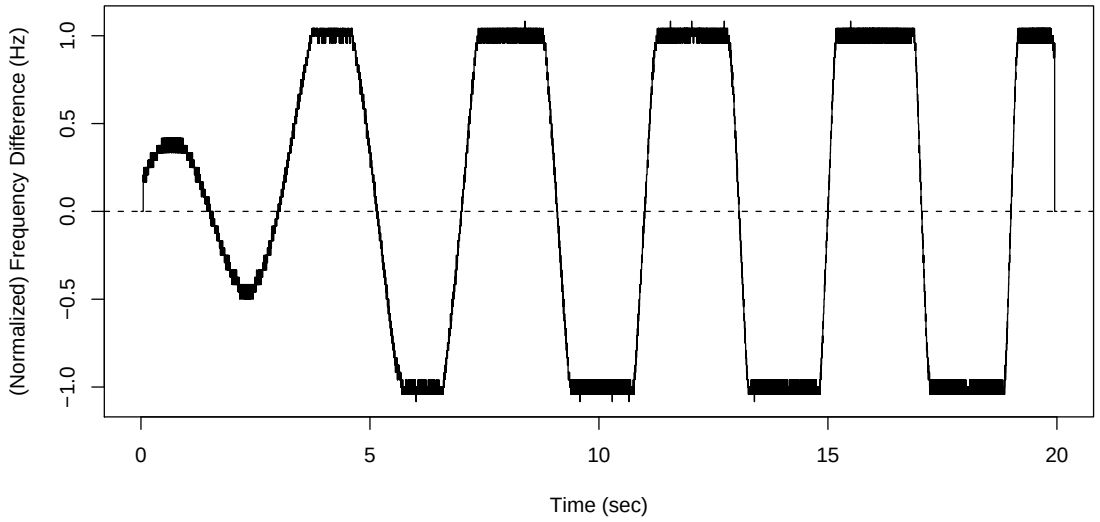


Figure 1.2: Scaled difference in average frequency between X and Y throughout the epoch.

but it oscillates around 0.5 in several intervals where it should be near zero (Figure 1.3, green curve). When $w = 18$, the estimated IC is near zero when the signals are asynchronous, but is not close to one when they are synchronized (Figure 1.3, blue curve). When $w = 6$, the result is most promising, since the estimated IC is close to zero whenever the frequency difference is large, and close to one when the two signals are synchronized (Figure 1.3, red curve).

For comparison, the IC between X and Y is estimated in the same manner, but using windows of fixed width (Figure 1.4). When the window size is 18 time points, the estimated IC is close to one when the signals are synchronized, but is not anywhere near zero in the first few intervals in which the signals are asynchronous (Figure 1.4, green curve). Also, when it does approach zero, it does not drop sharply, but falls off gradually. When the window size is 210 time points, the estimated IC is near

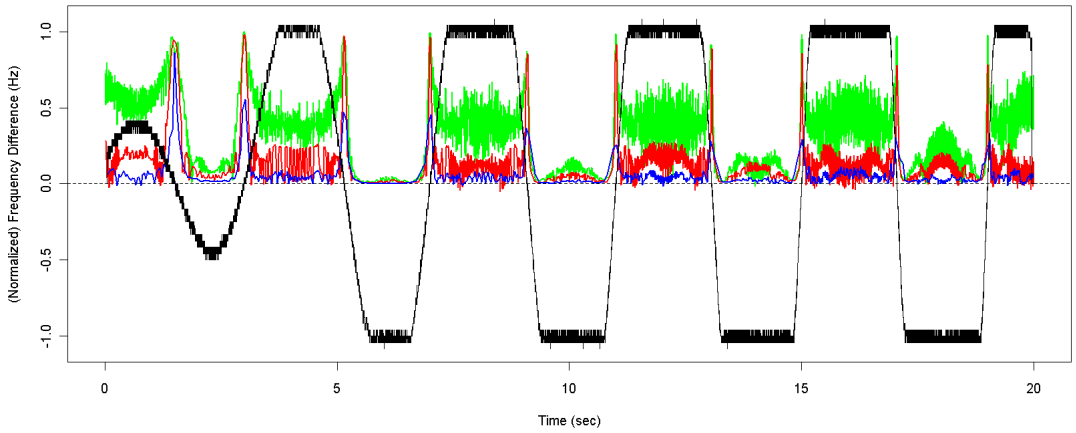


Figure 1.3: Estimated IC when $w = 3$ (green), $w = 6$ (red), and $w = 18$ (blue), with the scaled difference in average frequency between X and Y throughout the epoch.

zero when it should be, but it is generally much too small when it should be close to one (Figure 1.4, blue curve). When the window size is 90 time points, the best IC estimate is obtained (Figure 1.4, red curve). But at several time points where it should be close to one it is closer to 0.5.

This simulation demonstrates that whether a variable window or a fixed window is used, estimation of the IC between two signals is highly sensitive to the choice of the parameter that affects the window width. But the variable window approach has the advantage of adaptability to the local frequency, so that once a good choice is made for the tuning parameter, the IC estimate will be consistently reliable as the signal frequency varies throughout an epoch. A choice for the fixed window width may work well for a specific frequency range, but will not perform well outside of that range. In the simulated signals, the range for the average effective frequency of the base signal X grows as the time increases from 0 to 20 seconds. Hence the performance of

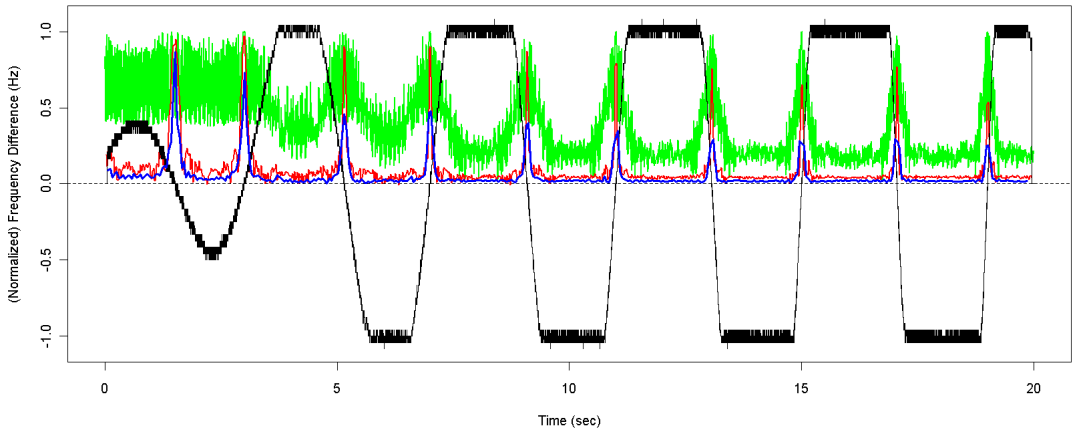


Figure 1.4: Estimated IC with fixed windows of 18 (green), 90 (red), and 210 (blue) time points, with the scaled difference in average frequency between X and Y throughout the epoch.

the IC estimate when $w = 18$ improves as time passes (Figure 1.4, green), while the performance of the IC estimates when $w = 90$ and $w = 210$ diminishes (Figure 1.4, red and blue, respectively). With the variable-window approach (Figure 1.3), it is seen that the performance of each IC estimate remains consistent throughout the epoch even though the range of the average frequency changes. Hence, once a good choice for w is identified, it can be employed in the study of any oscillating signal. Based on this analysis, it seems that $w = 6$ is a good choice.

1.3.2 Experimental data: EEG gamma signals from the rat hippocampus

The method developed above is demonstrated using an analysis of EEG data recorded from nine tetrodes located in the hippocampal formation of rats before, during and after they perform an exercise on a track. Tetrodes were placed in the medial entorhinal cortex (MEC; four electrodes) and CA1 region (five electrodes). This analysis focuses

on the EEG during a "rest" epoch, when the rat is in its cage (but not necessarily inactive) after one exercise epoch and before the next.

A typical data set consists of approximately 15 to 20 minutes of EEG data recorded at 1.5 kHz. This investigation focuses on the identification of instants when the gamma rhythms in both regions become synchronized on a short time scale. As an initial preprocessing step, the raw EEG signals were filtered in the 40–100 Hz range using the `filtfilt.m` routine in `Matlab`. This process was used to extract the gamma signal from the raw EEG signal. Figure 1.5 shows nine raw and nine filtered signals, respectively, from the first 2/3 second of a recording epoch. Note that the cycles for the filtered signals do not always cross the horizontal axis. Hence, for any gamma signal that might be selected for a base signal, its cycles are not necessarily identifiable by zero crossings. However, the Hilbert transform is applied to each filtered signal to obtain the Hilbert phase at each time point, so that one may identify the cycles by locating the points where the phase is approximately $\pm\pi/2$.

In this analysis, the gamma signal extracted from Tetrode 1, located in the MEC, is selected as the base signal. Figure 1.6 displays plots of the first 100 estimated IC values between this base gamma signal and the gamma signals at Tetrodes 6 and 26, respectively, at the beginning of a recording epoch, using $w = 6$ and $m = 2$. Note that the horizontal axis is not transformed back to the time scale, but is given in terms of the sequence of measurement windows. Observe that the estimated IC alternates between values above 0.6 and values below 0.4 in each plot. Frequent instants of high synchrony are expected in the left plot since Tetrodes 1 and 6 are both located within

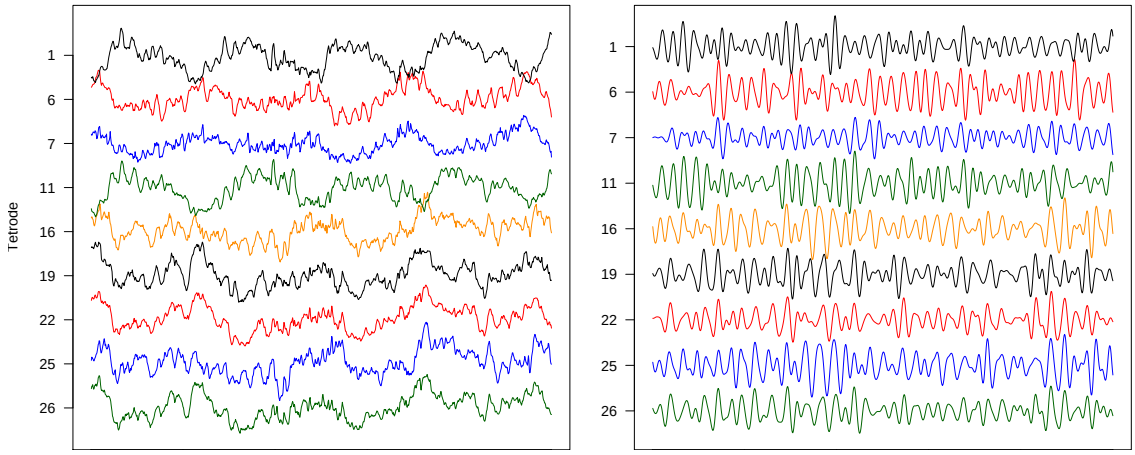


Figure 1.5: Raw EEG signals and filtered gamma signals from MEC (1,6,7,11) and CA1 (16,19,22,25,26) of rat during first 2/3 second of a rest epoch.

the MEC, albeit in different parts, while instants of high synchrony should be less frequent in the right plot, since Tetrode 26 is not in the MEC. For any given timespan

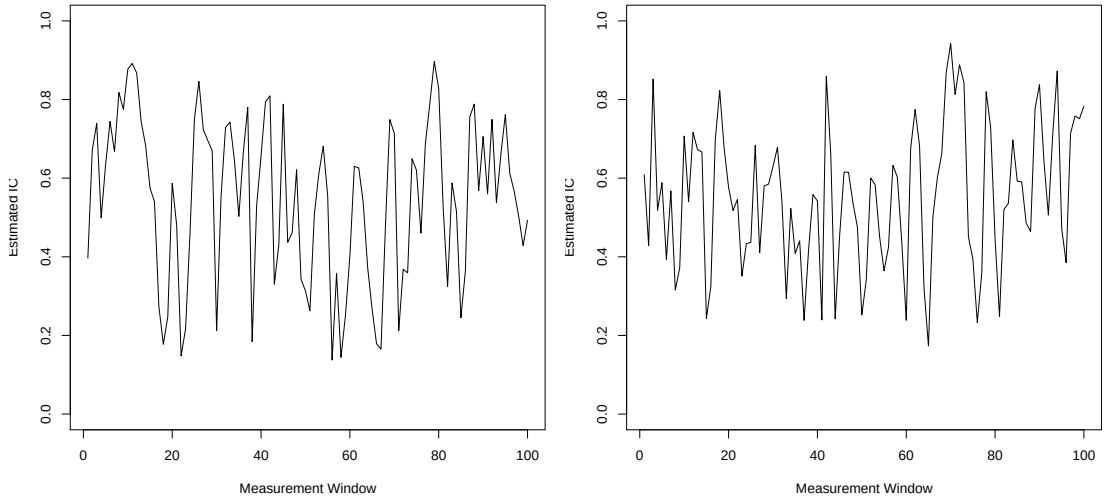


Figure 1.6: First 100 Instantaneous Coupling estimates between gamma signals at Tetrodes 1 (base signal) and 6 (left), and Tetrodes 1 (base signal) and 26 (right), at the beginning of an epoch.

one may choose any of the nine available gamma signals as the base signal and use the remaining signals to compute an eight-dimensional time series of IC estimates corresponding to that base signal, using any appropriate values of w or m. This section explores whether the time series can be partitioned into specific IC states.

Figure 1.7 shows eight histograms of the distributions of the $74,490$ estimated IC values between Tetrode 1 and each of the other eight tetrodes during a twenty-minute rest epoch. Almost all of the estimated IC values are distributed between zero and one, with a negligible remainder just slightly below zero. Hence the MVB distribution should be appropriate for modeling these data, as discussed in Section 1.2.2.

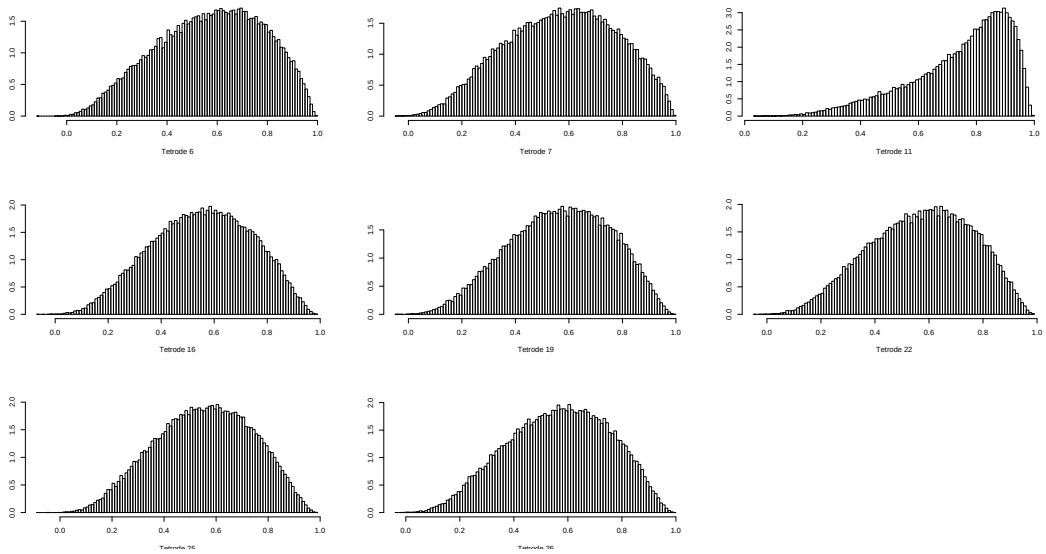


Figure 1.7: Distributions of estimated instantaneous coupling between gamma signals at Tetrode 1 and each of the other eight tetrodes during a rest epoch.

The goal is to reduce this eight-dimensional time series to a single dimension by identifying distinct IC states when different subsets of signals are coupled with the

base. This single dimension — the IC state — will take one of the values in $\{1, \ldots, p^*\}$ in each measurement window, where p^* is defined in (1.2.5). The histograms in Figure 1.7 may be considered superpositions of the histograms of the IC estimates corresponding to these distinct states, each of which is modeled by a MVB distribution. The EM algorithm is implemented as described in Section 1.2.2 to determine the parameters of these distributions and to determine the state membership of each IC estimate, using the `R` statistical package. To illustrate the method conveniently, a ten-second block from the full recorded time series is considered, corresponding to 746 consecutive, overlapping three-cycle windows of the base signal. Moreover, only four dimensions are selected rather than all eight, by choosing representatives from each location in the brain in which multiple EEG signals are recorded. Using the IC estimates between the base signal at Tetrode 1 and the signals at Tetrodes 6, 11, 25 and 26, computed during an interval when the rat is in its cage, the EM algorithm is implemented.

The EM algorithm is run for values of p in the range $2 \leq p \leq 8$. Once the algorithm converges, the parameter estimates are used to compute the mixture log-likelihood according to (1.2.3), and then to obtain the BIC according to (1.2.4). The algorithm is implemented simultaneously for the seven values of p, using several 3.06 gHz machines. Model estimation required from one to eleven days as p increased in value. Note that the BIC decreases as p increments from 2 to 4, is relatively constant for p in the range of 4 to 6 states, and then increases thereafter (Figure 1.8). It may be concluded that the instantaneous coupling between the base signal recorded in the

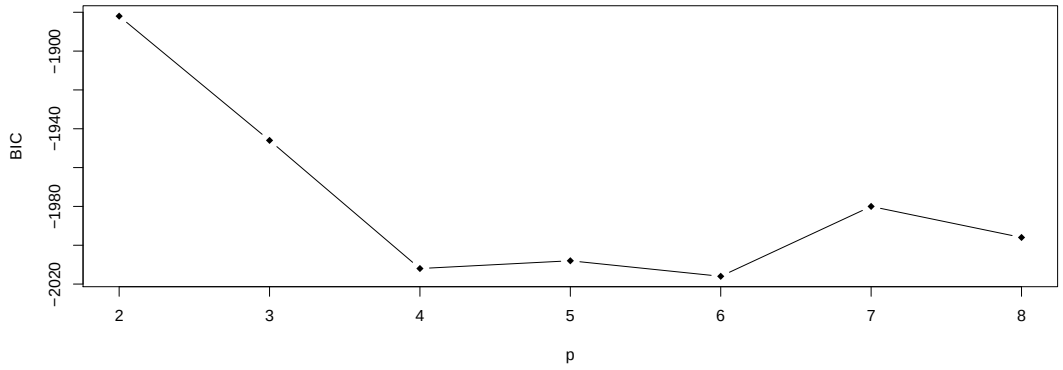


Figure 1.8: Progression of BIC as p increases, for EEG application.

MEC and the signals from the four other selected locations transition among four to six distinct IC states during the chosen ten-second block. Since the simplest model is preferred, one should adopt a model consisting of four IC states.

The parameter estimates corresponding to each of these four IC states, resulting from 108 iterations of the EM algorithm, are given in Table 1.1. To assign each IC

Table 1.1: Parameter estimates corresponding to each of these four IC states, resulting from 108 iterations of the EM algorithm.

State	π	θ_1	θ_2	θ_3	θ_4	θ_5
1	0.16	8.4	4.7	3.1	2.9	2.9
2	0.40	3.2	13.6	2.8	2.8	2.7
3	0.07	2.8	39.2	2.6	2.2	2.8
4	0.38	2.1	3.8	3.2	3.2	1.9

estimate to its appropriate state, inspect the corresponding vector of responsibilities

returned by the EM algorithm after its convergence (see (1.2.2) with $p = 4$), and note the position of the largest value. That is, assign the ith IC estimate to state k, where

$$k = \underset{j}{\text{argmax}}\{r_{j,i} \mid j = 1, 2, 3, 4\} \ ,$$

for $i = 1, \ldots, 746$. Then obtain the mean vector of the subsets of IC estimates assigned to each of the four states, along with their corresponding standard deviations, as shown in Table 1.2. The group means are plotted in Figure 1.9.

Table 1.2: Mean vector of the subsets of IC estimates assigned to each of the four states, along with their corresponding standard deviations.

	Mean (Standard Deviation)			
State	Tetrode 6	Tetrode 11	Tetrode 25	Tetrode 26
1	0.77 (0.08)	0.64 (0.16)	0.52 (0.16)	0.50 (0.15)
2	0.54 (0.17)	0.84 (0.07)	0.51 (0.17)	0.50 (0.16)
3	0.48 (0.17)	0.94 (0.03)	0.46 (0.16)	0.43 (0.19)
4	0.52 (0.17)	0.66 (0.16)	0.63 (0.17)	0.64 (0.17)

If one uses 0.6 as the threshold for distinguishing coupling from non-coupling, then State 1, which comprises 16.2% of the block of IC estimates, represents instants when the base signal is coupled with the other two MEC signals, but not with the CA1 signals. State 4, comprising 37.6% of the IC estimates, represents occasions when the

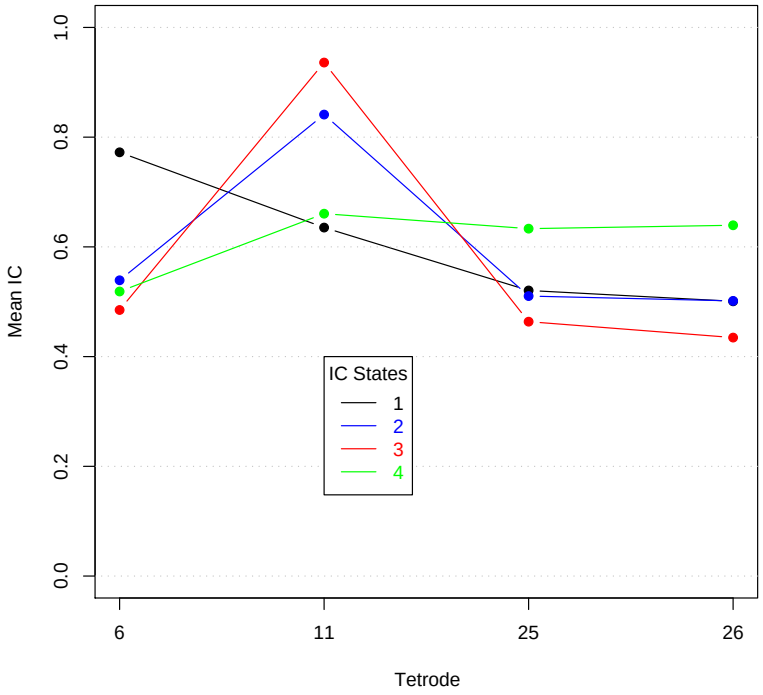


Figure 1.9: Means of the subsets of IC estimates corresponding to each of the four IC states, with the base signal at Tetrode 1, based on optimized mixed MVB model.

base signal is coupled with the signal at nearby Tetrode 11 in the MEC and with the signals at Tetrodes 25 and 26 in the CA1, but not with the signal at Tetrode 6 in another part of the MEC. Hence State 4 indicates synchronization between the MEC and the CA1 cell layer in the rat's brain. States 2 and 3, comprising 39.6% and 6.6% of the IC estimates, respectively, represent instants when the base signal is only coupled with the gamma signal at neighboring Tetrode 11, although at different levels. Since the mean vectors for the subsets of IC estimates corresponding to these two states appear to be very similar, and less than 7% of the estimates are assigned to State 3,

one may consider combining them. However, what distinguishes the two states is not the mean vectors of the corresponding subsets of IC estimates, but the parameter estimates for the two states. The estimates of the parameter θ_2 for States 2 and 3 are very different, which implies that the IC estimates for these two states are distributed quite differently. Moreover, the BIC criterion clearly recommends a minimum of four distinct IC states.

Since this analysis has been performed only on one block of IC estimates taken from a rest epoch, it cannot be assumed that these results apply to the entire epoch or to any other time period. One may repeat the estimation using a block of similar size from an interval during which the rat is motionless in its cage, and may be asleep, or when the rat is performing tasks on the track. Moreover, these procedures may be applied when the EEG signal at any tetrode is selected as the base signal, so that occasions of coupling between any two brain regions can be identified. This approach thus enables a neuroscientist to identify instants when different brain regions are synchronized, and investigate the correspondence between the frequency of such instants and behavioral covariates. Figure 1.10 shows the gamma signals extracted from the EEG recordings at Tetrodes 1, 6, 11, 25 and 26 at the beginning of the sample block on which this method was applied, along with a colored bar above each window associated with three cycles of the base signal at Tetrode 1. The colored bars indicate into which of the four IC states the signals are assigned by this method at each instant. The colors and their corresponding states match those designated in Figure 1.9. That is, black bars indicate coupling between the Tetrode 1 signal

and the signals at Tetrodes 6 and 11 in the MEC, blue bars and red bars denote coupling between the Tetrode 1 signal and the nearby signal at Tetrode 11, and green bars denote coupling between the Tetrode 1 signal and the signals at Tetrodes 11, 25 and 26. Both the sequence and frequency of IC states within any interval may be meaningful in future investigations.

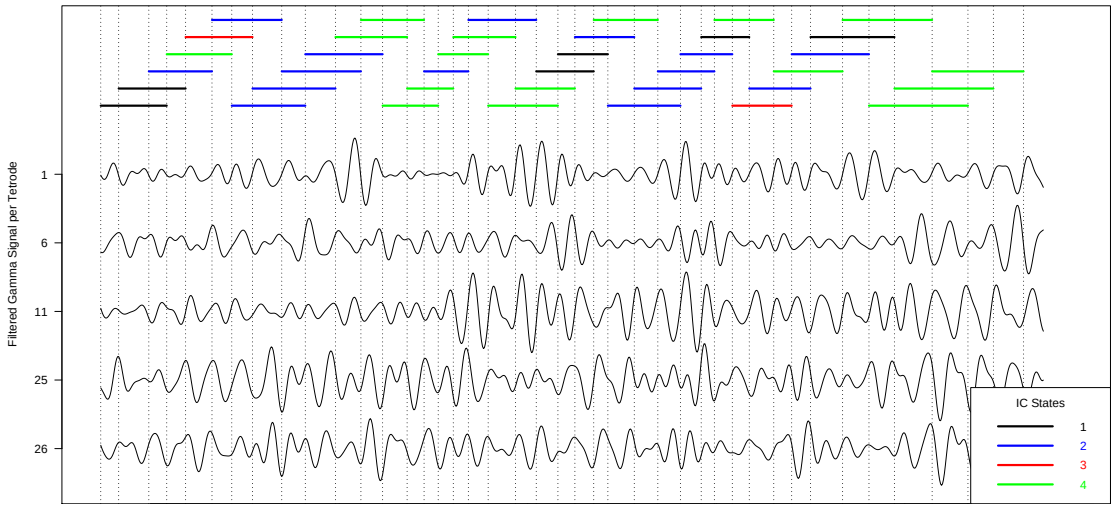


Figure 1.10: Gamma signals during a 2/3-second block of a recording epoch, with IC state designation indicated for each window when the base signal is at Tetrode 1 and four IC states are modeled. Vertical dotted lines correspond to computed cycles of the base signal.

1.4 Discussion

A computational method has been presented for estimating the short time-scale coupling between gamma signals filtered from two EEG recordings, along with confidence bounds on the estimate. This computation requires the selection of one signal as the base, and partitioning the recording epoch based on the cycles of that base. The

instantaneous coupling (IC) on any measurement window is defined as the maximum over a range of lags of the cross-correlation between the base and the other signal during that window. This value is computed throughout the epoch on a sliding window consisting of three cycles of the base, incrementing one cycle at a time so that consecutive measurements come from overlapping windows. At instants when the coupling between signals is strong, the IC estimate should be in the 0.6 to 1.0 range. A simulation study confirms that this procedure is sufficiently accurate in identifying instants of low and high synchronization.

When one gamma signal is chosen as the base and the IC estimate is computed between that signal and the other gamma signals at each of the other tetrodes, a distribution of IC estimates is obtained, which may be jointly modeled using the multivariate beta (MVB) distribution. The parameters of this distribution may be estimated using the EM algorithm. One can examine the structure of the set of IC estimates by implementing a mixture model with a pre-selected number of clusters representing IC states among the estimates. If the estimates naturally fall into a particular number of clusters, the Bayes Information Criterion will recommend that number. This enables one to identify instants in which subsets of the gamma signals are synchronized with the chosen base signal.

The method presented here allows neuroscientists to detect the evolution of IC on a short time scale among multidimensional EEG gamma waves. This evolution reveals the trajectory of local synchronous patterns and could be used to identify binding between separate parts of the brain. These evolving patterns of local and

global synchrony may provide a platform for scientists to map out moment-by-moment progression of signal transmission pathways among distinct regions of the brain. This methodology addresses several algorithmic challenges. First, the chosen measurement window of three cycles of a base gamma signal seems to reasonably capture such instantaneous synchronization, which typically lasts for fewer than five cycles. It is proffered that this is a natural timescale for such synchronization manifested through gamma rhythms. Secondly, the mixed MVB model with parameters estimated by the EM algorithm provides an effective and reliable tool for identifying IC states that signify the occurrence of synchronization among different regions of the brain.

One outstanding issue is the computational burden of the EM algorithm, which is known to converge slowly. In this EEG example the analysis was limited to a ten-second block and only four of the eight dimensions, and it took up to eleven days to converge. However, the computational efficiency would be greatly improved by using custom code, so that the method could be applied effectively to more dimensions and a much longer time span. Alternate methods for model fitting, including Markov chain Monte Carlo techniques or the simulated annealing algorithm, may also be developed. The focus in this chapter is illustration of the method, rather than optimizing the computational efficiency. This task will be undertaken in future work.

1.5 Derivation of the Multivariate Beta density

The derivation of the multivariate beta (MVB) density follows the approach of Olkin and Liu [51]. A full derivation is presented here due to typographical errors found in

the formula for the multivariate case presented in their original publication.

Let $X_1 \sim \Gamma(\theta_1, 1), \ldots, X_J \sim \Gamma(\theta_J, 1)$ and $Y \sim \Gamma(\theta_{J+1}, 1)$ be independent gamma random variables, where $\Gamma(\theta_j, 1)$ denotes the standard gamma distribution with shape parameter θ_j. Hence the joint density of $X_1, \ldots, X_J, Y$ is

$$f(x_1, \ldots, x_j, y) = \prod_{j=1}^{J} x_j^{\theta_j - 1} y^{\theta_{J+1} - 1} \exp\left\{ - \left(\sum_{j=1}^{J} x_j + y \right) \right\} \Big/ \prod_{j=1}^{J+1} \Gamma(\theta_j)$$

where $\Gamma(\cdot)$ denotes the gamma function and $x_1, \ldots, x_J, y$ are all positive. Define

$$U_1 = \frac{X_1}{X_1 + Y}, \ldots, U_J = \frac{X_J}{X_J + Y} \ ,$$

so that $0 < U_j < 1$ for $j = 1, \ldots, J$, and correlation among $U_1, \ldots, U_j$ is established by the common dependence on Y. Through simple algebra one has

$$X_j = \frac{Y U_j}{1 - U_j} \quad \text{for } j = 1, \ldots J \ .$$

Hence

$$\frac{\partial x_j}{\partial u_j} = \frac{y}{(1 - u_j)^2} \quad \text{and} \quad \frac{\partial x_j}{\partial y} = \frac{u_j}{1 - u_j} \quad \text{for } j = 1, \ldots J \ ,$$
$$\frac{\partial x_j}{\partial u_k} = 0 \quad \text{for } j \neq k \ ,$$

while

$$\frac{\partial y}{\partial u_j} = 0 \quad \text{for } j = 1, \ldots J, \qquad \text{and} \qquad \frac{\partial y}{\partial y} = 1 \ ,$$

giving the Jacobian matrix

$$\begin{bmatrix} \frac{y}{(1-u_1)^2} & 0 & \cdots & 0 & \frac{u_1}{1-u_1} \\ 0 & \frac{y}{(1-u_2)^2} & \cdots & 0 & \frac{u_2}{1-u_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{y}{(1-u_J)^2} & \frac{u_J}{1-u_J} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} ,$$

whose determinant is just the product of the diagonal entries,

$$\frac{y^J}{\prod_{j=1}^J (1-u_j)^2} \ .$$

Therefore the joint density of $U_1, \ldots, U_J, Y$ is

$$\begin{aligned} f(u_1, \ldots, u_J, y) &= \frac{y^J}{\prod_{j=1}^J (1-u_j)^2} \prod_{j=1}^J \frac{(yu_j)^{\theta_j - 1}}{(1-u_j)^{\theta_j - 1}} y^{\theta_{J+1}-1} \\ &\qquad \times \exp\left\{ -\left(\sum_{j=1}^J \frac{yu_j}{1-u_j} + y \right) \right\} \Bigg/ \prod_{j=1}^{J+1} \Gamma(\theta_j) \\ &= \frac{\prod_{j=1}^J u_j^{\theta_j - 1} y^{\sum_{j=1}^{J+1} \theta_j - 1}}{\prod_{j=1}^J (1-u_j)^{\theta_j + 1} \prod_{j=1}^{J+1} \Gamma(\theta_j)} \exp\left\{ -y \left(1 + \sum_{j=1}^J \frac{u_j}{1-u_j} \right) \right\} \ . \end{aligned}$$

Let

$$q = 1 + \sum_{j=1}^J \frac{u_j}{1-u_j} \qquad \text{and} \qquad \theta = \sum_{j=1}^{J+1} \theta_j \ .$$

Then obtain the joint density of $U_1, \ldots, U_J$, which is the objective here, by integrating out y as follows:

$$\begin{aligned} f(u_1, \ldots, u_J) &= \frac{\prod_{j=1}^J u_j^{\theta_j - 1}}{\prod_{j=1}^J (1-u_j)^{\theta_j + 1} \prod_{j=1}^{J+1} \Gamma(\theta_j)} \int_0^\infty y^{\theta - 1} \mathrm{e}^{-qy} \, \mathrm{d}y \\ &= \frac{\Gamma(\theta)}{\prod_{j=1}^{J+1} \Gamma(\theta_j)} \frac{1}{q^\theta} \prod_{j=1}^J \frac{u_j^{\theta_j - 1}}{(1-u_j)^{\theta_j + 1}} \\ &= \frac{\Gamma\left(\sum_{j=1}^{J+1} \theta_j \right)}{\prod_{j=1}^{J+1} \Gamma(\theta_j)} \left(\prod_{j=1}^J \frac{u_j^{\theta_j - 1}}{(1-u_j)^{\theta_j + 1}} \right) \left(1 + \sum_{j=1}^J \frac{u_j}{1-u_j} \right)^{-\sum_{j=1}^{J+1} \theta_j} \ , \end{aligned}$$

where $0 < u_j < 1$ for $j = 1, \ldots, J$. This is the multivariate beta density given in (1.2.1).

Chapter 2

Slope estimation in structural line-segment heteroscedastic measurement error models

2.1 Introduction

Consider an observational study in which information on two variables is collected from random samples of n distinct groups within a population. Suppose a researcher is given only a set of summary statistics on the observed variables for each sample, along with the corresponding sampling errors. He wishes to determine whether there is a significant linear association between the two variables, and if so, to model that association with an accurate slope estimate. The sampling error involved with each variable eliminates the applicability of the simple linear regression approach, and calls

for the implementation of a measurement error model. Moreover, the variability of the error on each observation for each variable is likely to be different, due to different sample sizes and other influences, so that a model that accounts for heteroscedastic measurement error is required.

Heteroscedastic measurement-error (ME) models have been developed for estimating the slope in such scenarios. Kulathinal, et al., apply this approach in [40] to the data collected in the WHO MONICA Project (2000) on cardiovascular disease and its risk factors. Patriota, et al., also examine these data in [52], as well as astronomical data obtained from the *Chandra* observatory, accounting for the heteroscedastic measurement error that characterizes each data set. Under this approach, each observation (x_i, y_i) is modeled as

$$x_i = \chi_i + \varepsilon_i, \quad y_i = \varphi_i + \nu_i, \quad \text{with } \varphi_i = \alpha + \beta\chi_i + \gamma_i \ ,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\nu_i \sim \mathcal{N}(0, \tau_i^2)$ are the independently-distributed heteroscedastic measurement errors, $\gamma_i \sim \mathcal{N}(0, \zeta^2)$ is the independently-distributed equation error, and all three errors are mutually independent. To assure identifiability, it is assumed that the measurement error variances (σ_i, τ_i), $i = 1, \ldots, n$, are known or can be estimated independently. Under the structural model the χ_i are independent and distributed as $\mathcal{N}(\mu_\chi, \sigma_\chi^2)$, while the φ_i are also independent and distributed as $\mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$. Both maximum-likelihood and method-of-moments estimators of the slope have been derived under the structural heteroscedastic measurement error (ME) model, as presented in [40, 52, 6] using this conventional equation error model. However, as will be demonstrated in an analysis of the WHO MONICA data, this model

is not robust against misspecification, so that the inclusion of the equation error component makes it susceptible to underestimation of the slope when the unknown data-generating mechanism does not actually have the equation-error structure.

In this chapter, the line-segment model for homoscedastic measurement error, as introduced by Davidov in [11], is extended to this heteroscedastic ME scenario. This alternative approach omits the equation error component and is symmetric in the two variables. These features make the line segment model more robust against misspecification, and thus preferable when the data-generation mechanism is completely unknown. It is assumed in this setting that that the points $(\mathcal{E}x_i, \mathcal{E}y_i)$ are randomly distributed on a line segment having latent endpoints at (η_X, η_Y) and (ξ_X, ξ_Y), with $\eta_X \leq \xi_X$. Let $\delta_X = \xi_X - \eta_X \geq 0$ and $\delta_Y = \xi_Y - \eta_Y$. Then the model is

$$x_i = \eta_X + \lambda_i \delta_X + \varepsilon_i, \quad \text{and} \quad y_i = \eta_Y + \lambda_i \delta_Y + \nu_i, \quad i = 1, \ldots, n\ , \tag{2.1.1}$$

where $\lambda_1, \ldots, \lambda_n$ are independently drawn from a common distribution G and take values in the unit interval, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ while $\nu_i \sim \mathcal{N}(0, \tau_i^2)$ for each i. It is also assumed that the error variances $\sigma_1, \ldots, \sigma_n$ are independently drawn from a common distribution, and likewise for $\tau_1, \ldots, \tau_n$. Again, to assure identifiability, it is assumed that these error variances are known for both variables. In practice, the error variance is rarely known, but may be accurately estimated using repeated measurements, or, when the observations are statistics based on samples, by using sampling errors. Let μ_λ and σ_λ^2 denote the mean and variance of λ_i, respectively, which must be finite since λ_i is bounded. Finally, assume that λ_i, σ_i and τ_i are mutually independent for all i. Note that this model does not include equation error.

Figure 2.1 provides a diagram of the above model along with the equation error model. In this diagram, (x_i, y_i) is the observed data pair. The linear association between X and Y is represented by the dashed trendline $Y = \alpha + \beta X$, upon which the line segment having endpoints at (η_X, η_Y) and (ξ_X, ξ_Y) rests. The conditional expectation of (x_i, y_i), given χ_i, onto the line/segment is the point $(\chi_i, \alpha + \beta\chi_i)$ under the equation error model. This point is equivalent to the point $(\eta_X + \delta_X\lambda_i, \eta_Y + \delta_Y\lambda_i)$, which is the conditional expectation of (x_i, y_i), given λ_i, under the line-segment model. This point lies on the dotted line between (x_i, y_i) and the line segment from the perspective of the line-segment model. But under the equation error model, it lies on the vertical dotted line through the unobserved point (χ_i, φ_i).

Trivially, the path from the observed pair (x_i, y_i) to the line segment, corresponding to the line segment model, is always smaller than the path that goes from (x_i, y_i) to (χ_i, φ_i), and then to the line, corresponding to the equation error model. When there is a strong correlation between the two variables, this difference in path lengths is almost negligible when using either model to compute an estimate of the slope, since the errors are small. In such cases, the dominating influence on the variance of the slope estimate is the overall dispersion in X — the larger the spread, the smaller the variance. Since the equation error is a vertical displacement, its inclusion in the model attenuates the effect of the dispersion in X, giving it an advantage over the line segment model. But when the errors become larger, the path difference becomes the dominating influence on the variance of the slope estimate among the two models. Hence in any scenario where the variances of the measurement errors are

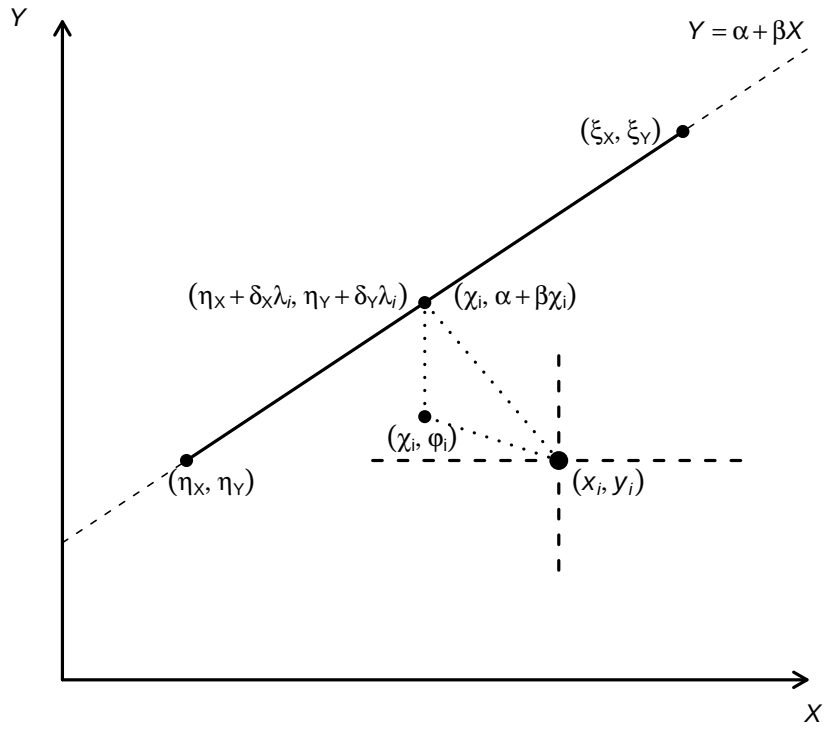


Figure 2.1: Diagram of equation error and line-segment models with heteroscedastic measurement error.

not small, the line segment model will provide a more precise estimate of the slope. This improvement is demonstrated using simulation studies.

The line segment model provides an additional benefit, whether or not there is measurement error or heteroscedasticity. The equation-error model assumes a specific structure for the underlying data-generation mechanism, a structure which may not not properly explain the association between the two variables. In this case, points that lie far from the line and toward the horizontal extremes exert an excessive influence on the slope estimate in equation-error models due to the effect of

including the equation errors. But in the line segment model, no such effect occurs. Consequently the influence of such points on the slope estimate is attenuated. Therefore the line segment model is more robust against outliers at the horizontal extremes, and is better able to explain the association between the two variables in situations in which the unknown data-generation mechanism does not have the equation-error structure. This benefit will be illustrated in the application of our model to the same WHO MONICA data, which contain influential points that have caused incorrect slope estimates under the equation-error models.

In the next section a method-of-moments estimate of the slope is derived under this extension of the line-segment model, and its asymptotic variance is obtained in Section 2.3 using the delta method. A large-sample estimate of the variance of this slope estimate is derived in Section 2.4 which may be used in real data analysis. In Section 2.5 simulation studies are performed which verify the accuracy and precision of these estimates when the data-generation mechanism conforms to the line-segment model, and which show that the precision of this slope estimate is superior in this setting to that of the method-of-moments estimate obtained through an equation error model, using assorted ranges of measurement error variances. In Section 2.6 the application of this slope estimation to real data is illustrated, and estimates derived under the line segment model are compared with those derived using several equation error methods, and using the line-segment model when homoscedastic errors are naïvely assumed. The results are summarized and briefly discussed in Section 2.8.

2.2 Point estimation of the slope

Given the structural line-segment model described in Section 2.1, one has

$$\mathrm{Var}(X_i) = \delta_X^2\sigma_\lambda^2 + \sigma_i^2, \quad \mathrm{Var}(Y_i) = \delta_Y^2\sigma_\lambda^2 + \tau_i^2, \text{ and } \mathrm{Cov}(X_i, Y_i) = \delta_X\delta_Y\sigma_\lambda^2 \,,$$

so that

$$\mathrm{Var}\left(\frac{X_i}{\sigma_i}\right) = \frac{\delta_X^2\sigma_\lambda^2}{\sigma_i^2} + 1 \quad \text{and} \quad \mathrm{Var}\left(\frac{Y_i}{\tau_i}\right) = \frac{\delta_Y^2\sigma_\lambda^2}{\tau_i^2} + 1 \,.$$

Hence

$$\frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(\frac{X_i}{\sigma_i}\right) = \delta_X^2\sigma_\lambda^2\sigma_n^\star + 1 \,, \qquad \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(\frac{Y_i}{\tau_i}\right) = \delta_Y^2\sigma_\lambda^2\tau_n^\star + 1 \,,$$

and

$$\frac{1}{n}\sum_{i=1}^n \mathrm{Cov}(X_i, Y_i) = \delta_X\delta_Y\sigma_\lambda^2 \,,$$

where

$$\sigma_n^\star = \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma_i^2} \qquad \text{and} \qquad \tau_n^\star = \frac{1}{n}\sum_{i=1}^n \frac{1}{\tau_i^2} \,.$$

One may then solve for δ_X (which is taken to be nonnegative) and δ_Y, and use the sign of the mean covariance to determine the sign of δ_Y, to get

$$\delta_X = \sigma_\lambda^{-1}\sqrt{\left[\frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(\frac{X_i}{\sigma_i}\right) - 1\right]_+ \Big/ \sigma_n^\star}$$

and

$$\delta_Y = \mathrm{sgn}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{Cov}(X_i, Y_i)\right)\sigma_\lambda^{-1}\sqrt{\left[\frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(\frac{Y_i}{\tau_i}\right) - 1\right]_+ \Big/ \tau_n^\star} \,,$$

where $[\cdot]_+ = \max(0, \cdot)$. Then the slope of the line segment is $\beta = \delta_Y/\delta_X$, provided $\delta_X > 0$. Note that in the ratio δ_Y/δ_X the common σ_λ^{-1} factor drops out, so that knowing the moments of λ is unnecessary for slope estimation. However, estimates of $\sum_{i=1}^n \mathrm{Var}(X_i/\sigma_i)/n$, $\sum_{i=1}^n \mathrm{Var}(Y_i/\tau_i)/n$, and $\sum_{i=1}^n \mathrm{Cov}(X_i, Y_i)/n$ are required.

Following the approach described in [12] and [11], method-of-moments estimators are derived, by equating sample moments with theoretical moments:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{X_i}{\sigma_i}\right) &\stackrel{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_i}\right)^2 = S^*_{xx}\ , \\
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{Y_i}{\tau_i}\right) &\stackrel{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{\tau_i}\right)^2 = S^*_{yy} \ \text{ and} \\
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(X_i,Y_i) &\stackrel{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})y_i = S_{xy}\ .
\end{aligned}$$

Hence

$$\widehat{\delta}_X = \sigma_\lambda^{-1}\sqrt{\left[S^*_{xx}-1\right]_+/\sigma^\star_n} \quad \text{and} \quad \widehat{\delta}_Y = \mathrm{sgn}(S_{xy})\sigma_\lambda^{-1}\sqrt{\left[S^*_{yy}-1\right]_+/\tau^\star_n}\ , \tag{2.2.1}$$

so that the estimated slope of the line segment is

$$\widehat{\beta} = \widehat{\delta}_Y/\widehat{\delta}_X = \mathrm{sgn}(S_{xy})\sqrt{\frac{\sigma^\star_n\left[S^*_{yy}-1\right]_+}{\tau^\star_n\left[S^*_{xx}-1\right]_+}}\ . \tag{2.2.2}$$

When the measurement errors are homoscedastic, (2.2.1) agrees with (2.12) in [11]. Note that this estimate is only valid when both $S^*_{xx} > 1$ and $S^*_{yy} > 1$, which is not a problem as long as the measurement error variances are not very large relative to the dispersions in their respective variables. In the case that they are, estimation of the slope becomes rather futile anyway.

2.3 Variance of the slope estimate

Derivation of the variance of this estimate also follows that in [11]. Since the slope estimate $\widehat{\beta}$ is location invariant, one may without loss of generality set $\eta_X = \eta_Y = 0$ in the derivation of $\mathrm{Var}(\widehat{\beta})$. Define

$$Z_i = \left(X_i,\ \frac{X_i}{\sigma_i^2},\ \frac{X_i^2}{\sigma_i^2},\ Y_i,\ \frac{Y_i}{\tau_i^2},\ \frac{Y_i^2}{\tau_i^2},\ X_iY_i\right)' = (z_1, z_2, z_3, z_4, z_5, z_6, z_7)'$$

and let

$$T_n = \frac{1}{n}\sum_{i=1}^{n} Z_i = \left(\overline{X},\ \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\sigma_i^2},\ \frac{1}{n}\sum_{i=1}^{n}\frac{X_i^2}{\sigma_i^2},\ \overline{Y},\ \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\tau_i^2},\ \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\tau_i^2},\ \frac{1}{n}\sum_{i=1}^{n}X_iY_i\right)' = (t_1, t_2, t_3, t_4, t_5, t_6, t_7)\ .$$

This structure on T_n is more complex than its five-dimensional counterpart in the homoscedastic case. Nevertheless, setting $\phi_\lambda = \mathcal{E}(\lambda_i^2) = \sigma_\lambda^2 + \mu_\lambda^2$, it is straightforward to show that

$$\boldsymbol{\mu} = \mathcal{E}(T_n) = (\delta_X\mu_\lambda,\ \delta_X\mu_\lambda\sigma^\star,\ \delta_X^2\phi_\lambda\sigma^\star + 1,\ \delta_Y\mu_\lambda,\ \delta_Y\mu_\lambda\tau^\star,\ \delta_Y^2\phi_\lambda\tau^\star + 1,\ \delta_X\delta_Y\phi_\lambda)'$$

based on the model assumptions in (2.1.1), and the assumed existence of $\sigma^\star = \mathcal{E}(\sigma_i^{-2}) = \lim_{n\to\infty}\sigma_n^\star$ and $\tau^\star = \mathcal{E}(\tau_i^{-2}) = \lim_{n\to\infty}\tau_n^\star$. The central limit theorem then assures that the distribution of $\sqrt{n}(T_n - \boldsymbol{\mu})$ converges to the $\mathcal{N}(0,\Sigma)$ distribution as $n\to\infty$, where Σ is the 7×7 covariance matrix of Z_i with entries $\Sigma_{ij} = \mathrm{Cov}(z_i, z_j)$, $1 \le i, j \le 7$.

Now define $S_n = (S_{xx}^*,\ S_{yy}^*,\ S_{xy})$. Expanding each of these components and applying some algebraic manipulation, one finds that

$$S_n = H(T_n) = (t_3 - 2t_1t_2 + t_1^2\sigma_n^\star,\ t_6 - 2t_4t_5 + t_4^2\tau_n^\star,\ t_7 - t_1t_4)'\ ,$$

so that

$$\boldsymbol{\phi} = \mathcal{E}(S_n) = H(\boldsymbol{\mu}) = (\delta_X^2\sigma_\lambda^2\sigma^\star + 1,\ \delta_Y^2\sigma_\lambda^2\tau^\star + 1,\ \delta_X\delta_Y\sigma_\lambda^2)'\ .$$

Applying the delta method and Slutzky's theorem to S_n, one has

$$\sqrt{n}(S_n - \boldsymbol{\phi}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, M\Sigma M')\ ,$$

where

$$M = \left.\frac{\partial H(T_n)}{\partial (T_n)}\right|_{T_n=\boldsymbol{\mu}} = \begin{bmatrix} 0 & -2\mu_\lambda\delta_X & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2\mu_\lambda\delta_Y & 1 & 0 \\ -\mu_\lambda\delta_Y & 0 & 0 & -\mu_\lambda\delta_X & 0 & 0 & 1 \end{bmatrix} .$$

The law of large numbers assures that $S_n \xrightarrow{P} \boldsymbol{\phi}$, and the continuous mapping theorem then implies that $\widehat{\beta} = \widehat{\beta}(S_n) \xrightarrow{P} \widehat{\beta}(\boldsymbol{\phi}) = \delta_Y/\delta_X = \beta$, so that $\widehat{\beta}$ is a consistent estimator of β. A second application of the delta method then gives

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, B'M\Sigma M'B) = \mathcal{N}(0, \omega) ,$$

where

$$B = \left.\frac{\partial \widehat{\beta}(S_n)}{\partial (S_n)}\right|_{S_n=\boldsymbol{\phi}} = \begin{bmatrix} \frac{-\delta_Y}{2\delta_X^3\sigma_\lambda^2\sigma^\star}, & \frac{1}{2\delta_X\delta_Y\sigma_\lambda^2\tau^\star}, & 0 \end{bmatrix}' .$$

Hence

$$\begin{aligned} \omega &= \frac{\delta_Y^2}{4\delta_X^6\sigma_\lambda^4(\sigma^\star)^2}(4\delta_X^2\mu_\lambda^2\Sigma_{22} + \Sigma_{33} - 4\delta_x\mu_\lambda\Sigma_{23}) \\ &\quad + \frac{1}{4\delta_x^2\delta_y^2\sigma_\lambda^4(\tau^\star)^2}(4\delta_Y^2\mu_\lambda^2\Sigma_{55} + \Sigma_{66} - 4\delta_Y\mu_\lambda\Sigma_{56}) \\ &\quad - \frac{1}{2\delta_X^4\sigma_\lambda^4\sigma^\star\tau^\star}(4\delta_X\delta_Y\mu_\lambda^2\Sigma_{25} - 2\delta_X\mu_\lambda\Sigma_{26} - 2\delta_Y\mu_\lambda\Sigma_{35} + \Sigma_{36}) \end{aligned}$$

is the desired asymptotic variance of $\sqrt{n}(\widehat{\beta} - \beta)$.

Substitute

$$\begin{aligned} \Sigma_{22} &= \delta_X^2\sigma_\lambda^2\sigma^{\star\star} + \sigma^\star & \Sigma_{33} &= \delta_X^4(\kappa_\lambda - \phi_\lambda^2)\sigma^{\star\star} + 4\delta_X^2\phi_\lambda\sigma^\star + 2 \\ \Sigma_{23} &= \delta_X^3(\gamma_\lambda - \mu_\lambda\phi_\lambda)\sigma^{\star\star} + 2\delta_X\mu_\lambda\sigma^\star & \Sigma_{55} &= \delta_Y^2\sigma_\lambda^2\tau^{\star\star} + \tau^\star \\ \Sigma_{66} &= \delta_Y^4(\kappa_\lambda - \phi_\lambda^2)\tau^{\star\star} + 4\delta_Y^2\phi_\lambda\tau^\star + 2 & \Sigma_{56} &= \delta_Y^3(\gamma_\lambda - \mu_\lambda\phi_\lambda)\tau^{\star\star} + 2\delta_Y\mu_\lambda\tau^\star \\ \Sigma_{25} &= \delta_X\delta_Y\sigma_\lambda^2(\sigma\tau)^\star & \Sigma_{26} &= \delta_X\delta_Y^2(\gamma_\lambda - \mu_\lambda\phi_\lambda)(\sigma\tau)^\star \\ \Sigma_{35} &= \delta_X^2\delta_Y(\gamma_\lambda - \mu_\lambda\phi_\lambda)(\sigma\tau)^\star & \Sigma_{36} &= \delta_X^2\delta_Y^2(\kappa_\lambda - \phi_\lambda^2)(\sigma\tau)^\star \end{aligned}$$

into this expression, where $\gamma_\lambda = \mathcal{E}(\lambda_i^3)$, $\kappa_\lambda = \mathcal{E}(\lambda_i^4)$, and, assuming the existence of each, we define

$$\sigma^{\star\star} = \mathcal{E}(\sigma_i^{-4}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma_i^4} = \lim_{n\to\infty} \sigma_n^{\star\star}, \quad \tau^{\star\star} = \mathcal{E}(\tau_i^{-4}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\tau_i^4} = \lim_{n\to\infty} \tau_n^{\star\star} ,$$

and $(\sigma\tau)^\star = \mathcal{E}(\sigma_i^{-2}\tau_i^{-2})$. Note that $(\sigma\tau)^\star = \sigma^\star\tau^\star$ by the independence of σ_i and τ_i for all i.

After much labor, one obtains

$$\begin{aligned} \omega \;=\; & \frac{1}{4\delta_X^6\delta_Y^2\sigma_\lambda^4}\left\{2\frac{\delta_X^4}{(\tau^\star)^2} + 4\delta_X^2\delta_Y^2\sigma_\lambda^2\left(\frac{\delta_X^2}{\tau^\star} + \frac{\delta_Y^2}{\sigma^\star}\right) + 2\frac{\delta_Y^4}{(\sigma^\star)^2}\right. \\ & \left. + \delta_X^4\delta_Y^4\Gamma_\lambda\left[\frac{\sigma^{\star\star}}{(\sigma^\star)^2} + \frac{\tau^{\star\star}}{(\tau^\star)^2} - 2\frac{(\sigma\tau)^\star}{\sigma^\star\tau^\star}\right]\right\} , \end{aligned} \tag{2.3.1}$$

where $\Gamma_\lambda = \kappa_\lambda - 4\mu_\lambda^4 + 8\mu_\lambda^2\phi_\lambda - 4\mu_\lambda\gamma_\lambda - \phi_\lambda^2$. The Cauchy-Schwartz inequality guarantees that $\sigma^{\star\star} \geq (\sigma^\star)^2$ and $\tau^{\star\star} \geq (\tau^\star)^2$, so that the expression $\sigma^{\star\star}/(\sigma^\star)^2 + \tau^{\star\star}/(\tau^\star)^2 - 2(\sigma\tau)^\star/\sigma^\star\tau^\star$ will be nonnegative, and will vanish in the case of homoscedasticity. Although the ratio $(\sigma\tau)^\star/\sigma^\star\tau^\star$ equals one, it is advantageous to keep it in the form of (2.3.1) because it will need to be estimated in the next section. Therefore, the large-sample variance of the estimated slope based on n independent observed pairs (x_i, y_i) with heteroscedastic measurement error is $\mathrm{Var}(\widehat{\beta}) = \omega/n$.

Note that in the case of homoscedastic error variance, (2.3.1) reduces to (3.3) given in [11].

2.4 Estimating the variance of the slope estimate

To obtain an estimate of $\mathrm{Var}(\widehat{\beta})$, replace each occurrence of δ_X and δ_Y in (2.3.1) with the corresponding estimates given in (2.2.2). Also replace $\sigma^\star$, $\tau^\star$, $\sigma^{\star\star}$, $\tau^{\star\star}$ and $(\sigma\tau)^\star$

with $\sigma_n^\star$, $\tau_n^\star$, $\sigma_n^{\star\star}$, $\tau_n^{\star\star}$ and $(\sigma\tau)_n^\star$, respectively. After some simplification, one obtains

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\beta}) &= \frac{\sigma_n^\star/\tau_n^\star}{n(S_{xx}^*-1)}\left\{1+\frac{1}{2(S_{yy}^*-1)}+\frac{S_{yy}^*-1}{S_{xx}^*-1}+\frac{S_{yy}^*-1}{2(S_{xx}^*-1)^2}\right. \\ &\qquad \left. + \frac{S_{yy}^*-1}{4}\left[\frac{\sigma_n^{\star\star}}{(\sigma_n^\star)^2}+\frac{\tau_n^{\star\star}}{(\tau_n^\star)^2}-2\frac{(\sigma\tau)_n^\star}{\sigma_n^\star\tau_n^\star}\right]\sigma_\lambda^{-4}\widehat{\Gamma}\right\} \quad (2.4.1)\end{aligned}$$

To derive the estimate $\sigma_\lambda^{-4}\widehat{\Gamma}$, an estimate of λ_i is required for $i=1,\ldots,n$. Suppose (x_i,y_i) is observed, while $\mathcal{E}(X_i,Y_i)=(\eta_X+\lambda_i\delta_X,\eta_Y+\lambda_i\delta_Y)$. Assume that λ_i is the unique value in $[0,1]$ that minimizes the Euclidean distance between (x_i,y_i) and the line segment whose endpoints are at (η_X,η_Y) and (ξ_X,ξ_Y). Simple calculus gives

$$\lambda_i=\frac{\delta_X(x_i-\eta_X)+\delta_Y(y_i-\eta_Y)}{\delta_X^2+\delta_Y^2}\ ,$$

so that

$$\begin{aligned}\widehat{\lambda}_i &= \frac{\widehat{\delta}_X(x_i-\widehat{\eta}_X)+\widehat{\delta}_Y(y_i-\widehat{\eta}_Y)}{\widehat{\delta}_X^2+\widehat{\delta}_Y^2} \\ &= \left[\frac{(x_i-\widehat{\eta}_X)\sqrt{[S_{xx}^*-1]_+/\sigma_n^\star}+(y_i-\widehat{\eta}_Y)\,\mathrm{sgn}(S_{xy})\sqrt{[S_{yy}^*-1]_+/\tau_n^\star}}{(S_{xx}^*-1)/\sigma_n^\star+(S_{yy}^*-1)/\tau_n^\star}\right]\sigma_\lambda \\ &= \psi(x_i,y_i)\sigma_\lambda=\psi_i\sigma_\lambda\ .\end{aligned}$$

One then has

$$\widehat{\mu}_\lambda=\sigma_\lambda\left(\frac{1}{n}\sum_{i=1}^n\psi_i\right),\quad \widehat{\phi}_\lambda=\sigma_\lambda^2\left(\frac{1}{n}\sum_{i=1}^n\psi_i^2\right),\quad \widehat{\gamma}_\lambda=\sigma_\lambda^3\left(\frac{1}{n}\sum_{i=1}^n\psi_i^3\right)\ ,$$

and

$$\widehat{\kappa}_\lambda=\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n\psi_i^4\right)\ .$$

Thus

$$\begin{aligned}\widehat{\Gamma}_\lambda &= \widehat{\kappa}_\lambda - 4\widehat{\mu}_\lambda^4 + 8\widehat{\mu}_\lambda^2\widehat{\phi}_\lambda - 4\widehat{\mu}_\lambda\widehat{\gamma}_\lambda - \widehat{\phi}_\lambda^2 \\ &= \sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i^4\right) - 4\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^4 + 8\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^2\left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right) \\ &\quad - 4\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)\left(\frac{1}{n}\sum_{i=1}^n \psi_i^3\right) - \sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right)^2,\end{aligned}$$

so

$$\begin{aligned}\sigma_\lambda^{-4}\widehat{\Gamma}_\lambda &= \left(\frac{1}{n}\sum_{i=1}^n \psi_i^4\right) - 4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^4 + 8\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^2\left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right) \\ &\quad - 4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)\left(\frac{1}{n}\sum_{i=1}^n \psi_i^3\right) - \left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right)^2. \qquad (2.4.2)\end{aligned}$$

Then (2.4.2) may be substituted into (2.4.1) to compute $\widehat{\mathrm{Var}}(\widehat{\beta})$.

However, the issue of estimating (η_X, η_Y), which is necessary for computing ψ_i, $i = 1, \ldots, n$, has been sidestepped. One remote possibility is that η_X and η_Y are known. Another approach, given in [12] under the assumption that μ_λ and σ_λ^2 are known, uses the Method of Moments to obtain $\widehat{\eta}_X = \overline{x} - \mu_\lambda\widehat{\delta}_X$ and $\widehat{\eta}_Y = \overline{y} - \mu_\lambda\widehat{\delta}_Y$. Thirdly, one may construct consistent estimators of η_X and η_Y using nonparametric estimation of G, as discussed in [12]. When information about G cannot be determined, a heuristic alternative is suggested here.

The line segment in our model rests on the line having slope β and passing through the point $(\eta_X + \mu_\lambda\delta_X, \eta_Y + \mu_\lambda\delta_Y)$. We estimate the location of this segment with a line having slope $\widehat{\beta}$ and passing through the point $(\overline{x}, \overline{y})$, whose equation is thus $y = \widehat{\beta}(x - \overline{x}) + \overline{y}$. Now, for $i = 1, \ldots, n$, consider the perpendicular line having slope $-1/\widehat{\beta}$ which passes through the point (x_i, y_i), whose equation is thus $y = -1/\widehat{\beta}(x - x_i) + y_i$. Then

both lines contain the orthogonal projection (x_i^*, y_i^*) of (x_i, y_i) onto the estimated line segment. We substitute (x_i^*, y_i^*) into each equation in place of (x, y), and set their right-hand sides equal, to get $\widehat{\beta}(x_i^* - \overline{x}) + \overline{y} = -1/\widehat{\beta}(x_i^* - x_i) + y_i$. Solving for x_i^* and simplifying gives us

$$x_i^* = \frac{\widehat{\beta}(y_i - \overline{y} + \widehat{\beta}\overline{x}) + x_i}{\widehat{\beta}^2 + 1} .$$

This value can then be used to compute $y_i^* = -1/\widehat{\beta}(x_i^* - x_i) + y_i$. If we let $x_0^* = \min\{x_1^*, \ldots, x_n^*\}$, we have $x_0^* \to \eta_X$ as $n \to \infty$. If $\widehat{\beta} \geq 0$, we let $y_0^* = \min\{y_1^*, \ldots, y_n^*\}$. Otherwise, we let $y_0^* = \max\{y_1^*, \ldots, y_n^*\}$. In either case, $y_0^* \to \eta_Y$ as $n \to \infty$. Hence (x_0^*, y_0^*) is a biased but consistent estimator of (η_X, η_Y), so we propose the use of $(\widehat{\eta}_X, \widehat{\eta}_Y) = (x_0^*, y_0^*)$ in the computation of ψ_i, $i = 1, \ldots, n$ when nothing is known about the distribution of λ_i.

2.5 Simulation study

To confirm the accuracy of these estimates, $n = 50$ data pairs are generated using (2.1.1) with $\eta_X = 0$, $\eta_Y = 0$, $\delta_X = 9$, and $\delta_Y = 6$, so that $\beta = 2/3$. For $i = 1, \ldots, n$, λ_i is drawn from a beta distribution with both parameters equal to 2, σ_i is drawn from a uniform distribution on $(a, 2a)$, and τ_i is drawn from a uniform distribution on $(b, 2b)$, where a and b are fixed but arbitrary positive numbers. Hence Γ_λ can be computed from the known moments of a beta distribution, while $\sigma^\star$, $\tau^\star$, $\sigma^{\star\star}$ and $\tau^{\star\star}$ are derived from the known moments of a uniform distribution. A range of values for a and b is selected, then (2.2.1), (2.3.1), (2.4.1) and (2.4.2) are used to compute $\widehat{\beta}$, its variance $\mathrm{Var}(\widehat{\beta})$, and the estimate $\widehat{\mathrm{Var}}(\widehat{\beta})$ of this variance. For this first simulation it

is assumed that η_X and η_Y are known. The procedure is repeated 500 times, and the estimates at each iteration are recorded.

For each run of the simulation, Table 2.1 presents the selected values of a and b, which control the magnitude of the error variances, in the first two columns. Column 3 gives the median value of $\widehat{\beta}$, computed over 500 iterations using (2.2.1), along with the sample variance of this estimate in column 4. Column 5 provides the expected value of $\mathrm{Var}(\widehat{\beta})$, based on (2.3.1) divided by n. Column 6 provides the median value of $\widehat{\mathrm{Var}}(\widehat{\beta})$, computed over 500 iterations using (2.4.1). Ideally, the value in column 3 matches the true value of the slope, i.e., 2/3, and the values in columns 4 and 6 are both close to the value in column 5 for each run.

Table 2.1: Simulation results for several choices of a and b, based on 500 iterations, with (η_X, η_Y) known and $\beta = 2/3$. (S.V. = sample variance)

a	b	median($\widehat{\beta}$)	(S.V.)	Var($\widehat{\beta}$)	median($\widehat{\mathrm{Var}}(\widehat{\beta})$)
0.05	0.03	0.669	(0.00086)	0.00087	0.00083
0.35	0.25	0.662	(0.00224)	0.00204	0.00219
0.50	0.55	0.663	(0.00564)	0.00550	0.00610
0.75	0.70	0.666	(0.01077)	0.00982	0.01085
0.85	0.90	0.657	(0.02037)	0.01618	0.01801

For all choices of a and b, the estimate of the line segment slope is centered very near the true value of 0.667. Moreover, the observed variance among 500 computed values of $\widehat{\beta}$ (in column 4) is consistently close to the expected variance of $\widehat{\beta}$ (in column 5), and the median estimate of the variance of $\widehat{\beta}$ over 500 iterations (in column 6) also

proves to be quite accurate, with a gradual loss of accuracy as the error variances grow. To the extent that these estimates of the variance are off-target, they are consistently a bit high, and hence give more conservative estimates. The precision of these estimates even as the error variance grows is remarkable given the need to estimate many parameters. Hence this first simulation confirms the reliability of the estimates under the line segment model when η_X and η_Y are known.

In a second simulation, set $\eta_X = 2$, $\eta_Y = 3$, $\delta_X = 9$, and $\delta_Y = -6$, so that $\beta = -2/3$. This time it is assumed η_X and η_Y are unknown, and the heuristic estimates of Section 2.4 are employed. All other conditions are the same as above. Table 2.2, which has the same structure as Table 2.1, displays a summary of these results.

Table 2.2: Simulation results for several choices of a and b, based on 500 iterations, with (η_X, η_Y) unknown and $\beta = -2/3$. (S.V. = sample variance)

a	b	median($\widehat{\beta}$)	(S.V.)	Var($\widehat{\beta}$)	median($\widehat{\text{Var}}(\widehat{\beta})$)
0.05	0.04	−0.667	(0.00086)	0.00087	0.00085
0.30	0.33	−0.665	(0.00228)	0.00239	0.00257
0.46	0.52	−0.660	(0.00581)	0.00490	0.00528
0.71	0.65	−0.673	(0.00976)	0.00849	0.00968
0.98	0.93	−0.664	(0.02160)	0.01871	0.02178

Despite the need to use biased estimates of η_X and η_Y to compute the values in column 6, these estimates of the variance of $\widehat{\beta}$ are centered at values only slightly larger than the expected values listed in column 5. The sample variance of $\widehat{\beta}$ over

500 iterations, given in column 4, also corresponds well with the expected variance in every run given in column 5, although it becomes inflated as the error variances grow. Moreover, the sample median of $\widehat{\beta}$ is consistently accurate even as the variance of the measurement errors increases. Hence the derivations in the preceding sections are strongly validated by these simulations.

The performance of the slope estimate under the line-segment model may be compared to the performance of the method-of-moments slope estimate proposed recently in [52], based on the structural heteroscedastic ME model with equation error described in the introduction. An EM algorithm to compute maximum-likelihood estimates for these model parameters was proposed in [40], and additional estimation methods were presented in [6]. In [52], Patriota, et al., derive both method-of-moments and maximum-likelihood estimates for the parameters and provide performance comparisons with the earlier approaches. The focus here is on their method-of-moments estimate, which is designated as MM-P, as a contrast to the estimate under the line segment model, since that estimate is provided in closed form, and the authors found its performance to be superior to that of alternate equation-error models when the Gaussian error assumption is valid.

Using notational equivalents, the MM-P slope estimate is $\widehat{\beta} = S_{xy}/(S_{xx} - \sigma_n^\star)$, where $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2/(n-1)$, $S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})/(n-1)$, and (needed below) $S_{yy} = \sum_{i=1}^n (y_i - \overline{y})^2/(n-1)$. The asymptotic variance corresponding to the MM-P slope estimate is $[2\beta^2(\sigma^{\star\star} - \sigma_\chi^4) + \pi]/\sigma_\chi^4$, where $\pi = \beta^2\sigma_\chi^2\sigma^\star + \zeta^2\sigma_\chi^2 + (\sigma\tau)^\star + \zeta^2\sigma^\star + \sigma_\chi^2\tau^\star + 2\beta^2\sigma_\chi^4$. For the line-segment model, $\sigma_\chi^2 = \delta_X^2\sigma_\lambda^2$ and the equation

error ζ^2 equals zero. Hence once may compute the expected large-sample variance of $\widehat{\beta}$ under MM-P for the above simulation scenarios using the known moments of the beta and uniform distributions, once a and b are specified. Moreover, using the proposed estimates of these parameters, the MM-P estimate of $\mathrm{Var}(\widehat{\beta})$ is

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \frac{2S_{xy}^2[\sigma_n^{\star\star} - (S_{xx} - \sigma_n^\star)^2]}{n(S_{xx} - \sigma_n^\star)^4} + \frac{S_{xy}^2 + S_{xx}S_{yy} + (\sigma\tau)_n^\star - \sigma_n^\star\tau_n^\star}{n(S_{xx} - \sigma_n^\star)^2} \,. \tag{2.5.1}$$

When the first simulation is repeated using the MM-P estimates, with the same choices for a and b, the median of the 500 point estimates of β is consistently close to the true value of 2/3. The first two columns of Table 2.3 show the expected value of $\mathrm{Var}(\widehat{\beta})$ under the MM-P model for each pair (a, b), along with the corresponding median of the 500 estimates of $\mathrm{Var}(\widehat{\beta})$ computed using (2.5.1). As the magnitude of the error variance grows, the difference between these two values grows, but remains within reason. But note that when these values are compared with those in the last two columns, which are the corresponding columns imported from Table 2.1, one finds that the MM-P expected variance and estimated variance of $\widehat{\beta}$ are much smaller than those obtained under the line segment model (MM-LS) when a and b are small (first two rows), about the same when a and b are close to 0.5 (middle row), and much larger when a and b are larger than 0.5 (last two rows). In other words, the MM-P estimates are more precise when the error variances are quite small, but the line segment model estimates have greater precision when the error variances become appreciable. Hence there is evidence that the line segment model will provide a better estimate of the slope when the measurement error variances are not small.

Of course, the estimates from Table 2.1 are based on the assumption that η_X and

Table 2.3: Simulation variance estimates for several choices of a and b, based on 500 iterations, using the MM-P model and the MM-LS model, with $\beta = 2/3$.

		MM-P Model		MM-LS	
a	b	$\text{Var}(\widehat{\beta})$	$\text{median}(\widehat{\text{Var}}(\widehat{\beta}))$	$\text{Var}(\widehat{\beta})$	$\text{median}(\widehat{\text{Var}}(\widehat{\beta}))$
0.05	0.03	0.00002	0.00002	0.00087	0.00083
0.35	0.25	0.00150	0.00151	0.00204	0.00219
0.50	0.55	0.00569	0.00555	0.00550	0.00610
0.75	0.70	0.01248	0.01259	0.00982	0.01085
0.85	0.90	0.02043	0.02003	0.01618	0.01801

η_Y are known. However, even in the second simulation scenario, in which (η_X, η_Y) must be estimated, the estimate of $\text{Var}(\widehat{\beta})$ under the line segment model still returns a smaller value than the MM-P version once the error variances become appreciable, as Table 2.4 shows. In this scenario it is observed that the variance estimate is indeed smaller under the MM-P model when the σ_i and the τ_i are restricted to small values, but when the σ_i are drawn from the interval $(0.46, 0.92)$ and the τ_i are drawn from $(0.52, 1.04)$, the estimates of the variance of $\widehat{\beta}$ are approximately the same under both methods. Then, when the σ_i are drawn from the interval $(0.71, 1.42)$ and the τ_i are drawn from $(0.65, 1.30)$, the variance estimate under the line segment method is significantly smaller, and when the σ_i are drawn from the interval $(0.98, 1.96)$ while the τ_i are drawn from $(0.93, 1.86)$, the line segment model estimate of the variance of $\widehat{\beta}$ is almost half of the corresponding estimate under the MM-P model. Hence the line segment approach is recommended for slope estimation when the heteroscedas-

tic measurement error variances are not small, as that method will yield narrower confidence intervals for the slope.

Table 2.4: Simulation variance estimates for several choices of a and b, based on 500 iterations, using the MM-P model and the line segment model, with $\beta = -2/3$.

		MM-P Model		MM-LS Model	
a	b	$\mathrm{Var}(\widehat{\beta})$	$\mathrm{median}(\widehat{\mathrm{Var}}(\widehat{\beta}))$	$\mathrm{Var}(\widehat{\beta})$	$\mathrm{median}(\widehat{\mathrm{Var}}(\widehat{\beta}))$
0.05	0.04	0.00003	0.00003	0.00087	0.00085
0.30	0.33	0.00184	0.00179	0.00239	0.00257
0.46	0.52	0.00488	0.00472	0.00490	0.00528
0.71	0.65	0.01057	0.01102	0.00849	0.00968
0.98	0.93	0.02660	0.02685	0.01871	0.02178

It should be noted that in this simulation the data-generation mechanism was specified according to the line-segment model, while the MM-P procedure is based on the equation-error model. Thus our estimation procedure had a built-in advantage. When the data-generation mechanism is unknown, one must take care in choosing a model, as the results from disparate models can be quite different. The next section demonstrates this issue.

2.6 Real data application

The World Health Organization (WHO) established the Multinational MONItoring of trends and determinants of CArdiovascular disease (MONICA) during the 1980s to study the association between known risk factors, like smoking and obesity, and

trends in cardiovascular disease. The linear association between data on the average annual change in the observed risk score (X) and the average annual change in event rate (Y), both given as percentages, was modeled in [40] and [52] using measurement error models with equation error, with the sampling error in the trend estimates taken as the heteroscedastic measurement error. Table 2.5 displays the estimated slope and its estimated standard error for each gender computed under several different models. The ordinary least squares (OLS) method disregards the measurement error. K-2000 and K-2002 represent maximum likelihood estimates provided in [40], while MM-P and ML-P represent the method-of-moments and maximum-likelihood estimates reported in [52], using measurement error models. Finally, the MM-LS estimates are those obtained using the line segment model.

Table 2.5: Estimates of the slope and standard errors of the estimates for the WHO MONICA data on males and females, based on seven models.

	Males		Females	
Model	$\widehat{\beta}$	$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$	$\widehat{\beta}$	$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$
OLS	0.31	0.20	0.51	0.33
K-2000	0.43	0.22	0.57	0.33
K-2002	0.47	0.23	0.68	0.24
MM-P	0.35	0.22	0.58	0.38
ML-P	0.47	0.23	0.68	0.41
MM-LS	1.76	0.23	2.10	0.38

Figure 2.2 displays the data for males ($N = 38$) and for females ($N = 36$) sep-

arately, with the magnitude of the measurement errors indicated by the crosshairs. Lines having the estimated slopes under the line segment model and under the ML-P and MM-P models are added to each plot in Figure 2.2. For the MM-LS results, the lines are passed through the means $(\overline{X}, \overline{Y})$ for each gender, while for the ML-P and MM-P results the estimated intercepts provided in that paper are used. It is quite

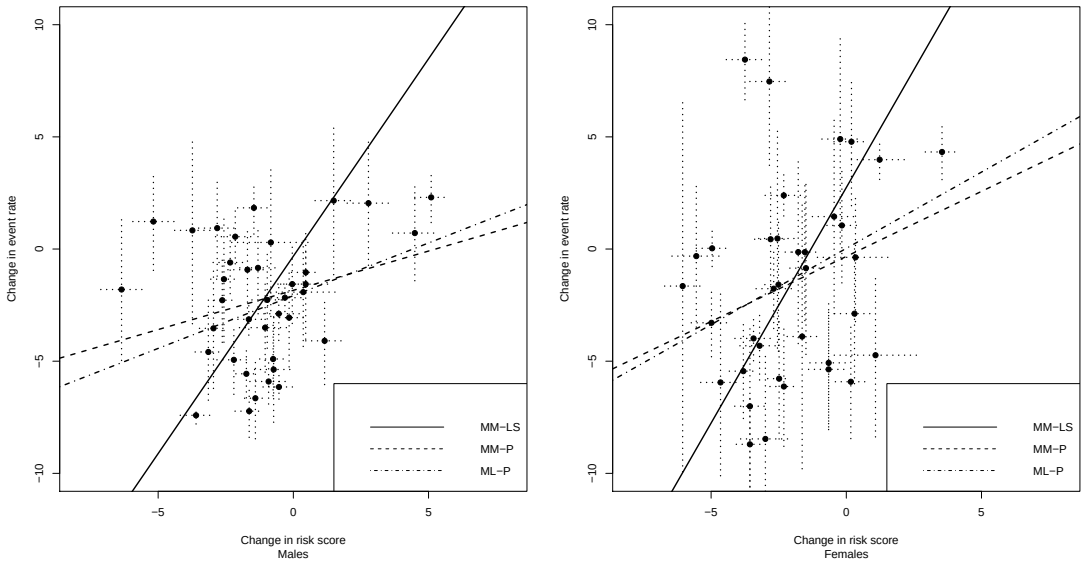


Figure 2.2: Scatterplot of change in event rate versus change in risk score, with standard errors, from WHO MONICA project, and lines having estimated slopes under three models, for males and females.

surprising that the estimated slopes under the line segment model are dramatically steeper than those obtained under all the other models — more than three times the size — while the standard errors of the slope estimates are about the same for all six models. While this initially makes the MM-LS results appear to be in error, inspection of the data scatter in each plot of Figure 2.2 reveals that these results are

more consistent with the observable trend. This discrepency illustrates the effect of using equation error in measurement error models when there is insufficient support for this assumption, as discussed in the introduction. In contrast, the line segment model omits the equation error component and produces a slope estimate which is able to capture the observed trend more accurately. Indeed, the steeper slopes send an even stronger message to the public about the urgency of maintaining cardiovascular health. While it is difficult to ascertain which model is most appropriate for the latent WHO MONICA data-generation mechanism, tools are available for assessing model fitness.

The robustness of the MM-LS method may be demonstrated by deleting each of several identified influential points one at a time and computing the slope estimate on each subset of the data. The same is done for the MM-P model, and the results are displayed in Table 2.6. Deletion of influential points alters the slope under the line segment model by 5% to 14% for males and by -8% to 14% for females. But under the MM-P model the slope is altered by -34% to 29% for males and by -24% to 24% for females. This illustrates the robustness of the line segment model to influential points and helps explain the disparity between the slope estimates from the two models. Given that the precision of the slope estimates is equivalent between the two models, the robustness property recommends the line segment model over the others considered.

Table 2.6: Estimates of the slope and standard errors of the estimates for the WHO MONICA data on males and females, based on two models, when no points are deleted, and when individual influential points are deleted.

Males

Model	Full Data	Point 1	Point 2	Point 3
MM-P	0.35(0.22)	0.45(0.24)	0.31(0.25)	0.23(0.25)
MM-LS	1.76(0.23)	1.85(0.25)	1.86(0.26)	2.01(0.27)

Females

Model	Full Data	Point 1	Point 2	Point 3	Point 4
MM-P	0.58(0.38)	0.72(0.35)	0.65(0.41)	0.63(0.35)	0.44(0.44)
MM-LS	2.10(0.38)	1.93(0.38)	2.22(0.41)	2.10(0.37)	2.39(0.50)

2.7 Illustration: Slope underestimation with the SLM

Consider again the WHO MONICA data discussed in Section 2.6. If one were to disregard the measurement error, and naïvely apply the ordinary least squares (OLS) method to estimate the parameters of the linear trend, the resulting best-fit line would not correspond with the trend. Figure 2.3 again displays the data scatterplots for males and for females separately, along with the best-fit lines derived under the OLS method. Clearly, the slope of the plotted OLS best-fit line is significantly smaller than the slope estimation under the line segment model, which more accurately captures the trend observed in the scatterplots. This is not a defect in the OLS method, but a result of applying a model for the data-generating mechanism that is not appropriate

in this case. Moreover, any other model which relies in part on structural specifications comparable to those in the simple linear model, including the measurement error models examined above, is susceptible to similar underestimation of the slope.

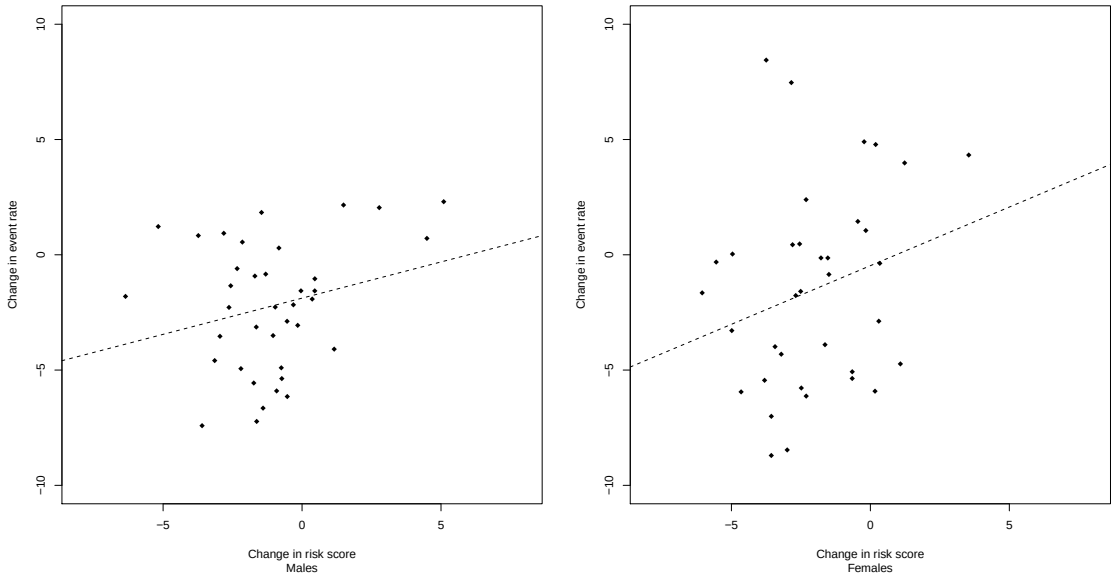


Figure 2.3: Scatterplot of change in event rate versus change in risk score, from WHO MONICA project, and best-fit lines based on the OLS method, for males and females.

To further illustrate this point, generate a random sample of 100 points scattered uniformly between -2 and 2 about the line $Y = 0$ in a thick cloud. In this case, the data are clearly not generated from the SLM. Then rotate the line and each point in the data cloud 51.3° about the origin, so that the resulting data cloud should properly be perceived to follow a linear trend whose slope is about 1.25, based on both the underlying data generation mechanism and the geometry of the cloud. However, if OLS is applied to estimate the trend, the OLS slope estimate is only 1.1, which is a significant underestimate. As the angle of rotation is increased, the underestimation

of the slope by OLS becomes more pronounced. Figure 2.4 displays the data cloud along with its actual linear trend and the OLS best-fit line for six different progressively steeper linear trends. Clearly, if a researcher attempted to estimate the linear association between the two variables in any of the displayed scenarios using OLS, the result would not guide the researcher to the truth about that association. While subsequent diagnostic tools might pick up the lack-of-fit in the most extreme cases, they may well miss it in more subtle cases. The structure of the data clouds must not be ignored, and models which can properly estimate the linear trend in light of that structure must be employed. In Figure 2.5, the slope of the data-generating trendline as it makes a complete rotation through 2π radians is shown, along with the corresponding estimated slope based on OLS regression on the generated data. The larger the magnitude of the slope of the actual trend, the greater the underestimation when OLS is misapplied. This illustrates a common outcome of misspecifying the SLM: the slope estimate is pulled toward zero. In a scenario in which the data are indeed generated from a SLM (with a reasonably small error variance) and the OLS slope estimate is computed as the plane rotates, the plot of the slope estimate follows that of the trendline slope very closely even when it becomes quite steep, as displayed in Figure 2.6. This is how the plot of slope estimates for rotating centered data should appear when the model is specified correctly—like the plot of the tangent function. Of course, as the error variance increases this ideal result deteriorates.

This illustration suggests a test for misspecification of the SLM. Suppose one is presented with bivariate data whose association might be represented by the SLM.

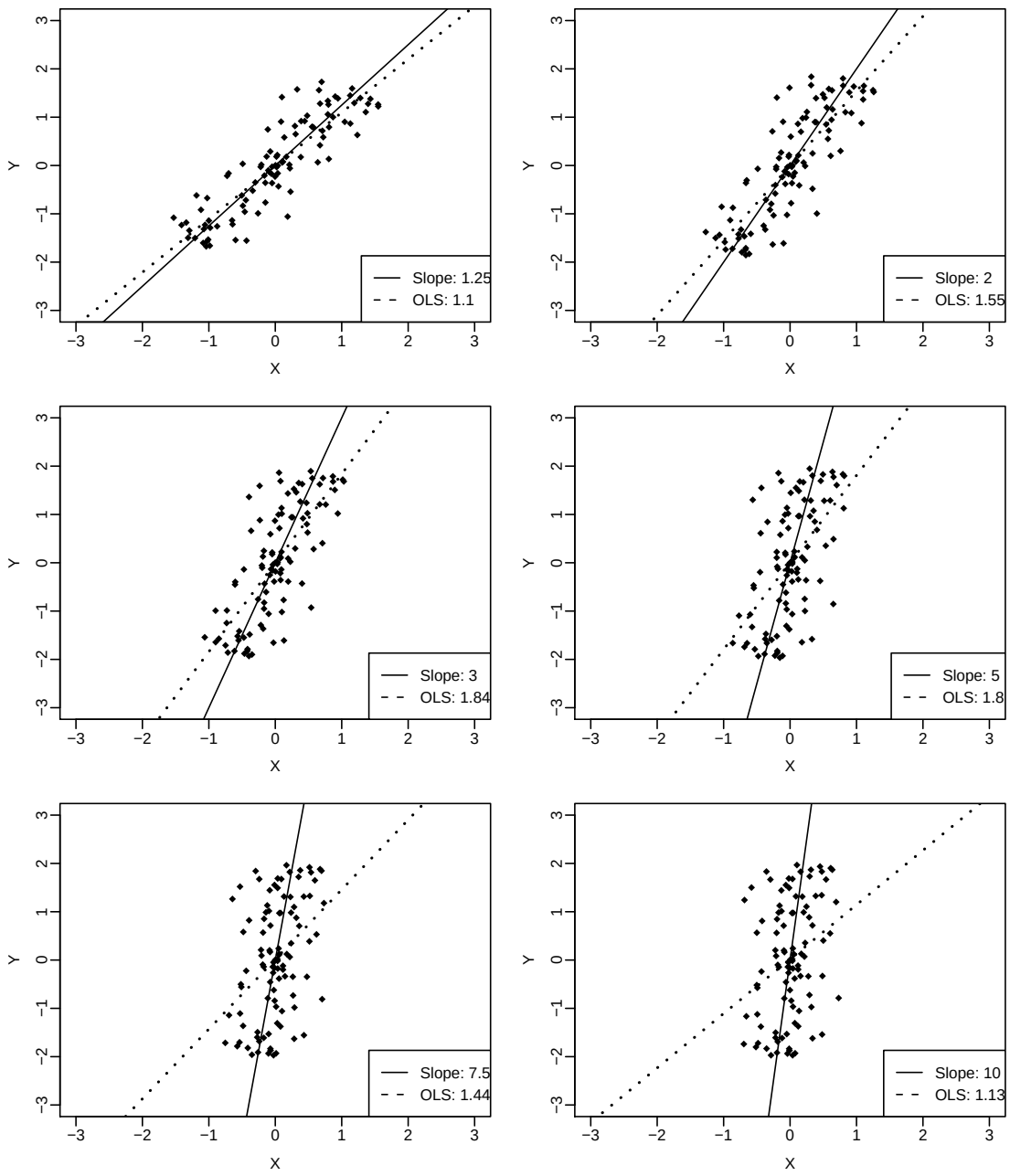


Figure 2.4: Rotating data cloud, with the underlying linear trend and the estimated trend under OLS regression.

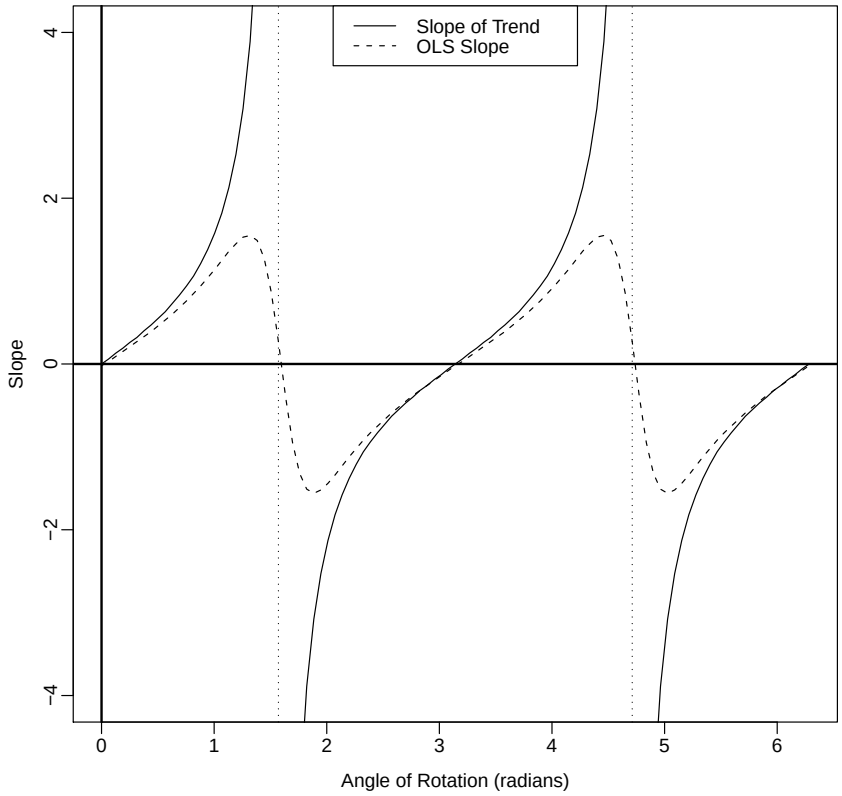


Figure 2.5: Progression of the slope of the trendline for the rotating centered data cloud, and the corresponding estimated slope using OLS regression when the data are not generated from the SLM, as the plane rotates about the origin.

Center the data about zero for both variables, select one variable as the covariate, apply OLS regression, and obtain a slope estimate $\widehat{\beta}$ and an estimate $\widehat{\sigma}^2$ for the error variance. Then rotate the plane completely about the origin in small increments and compute a new OLS slope estimate at each step. Plot the progression of OLS slope estimates against the angle of rotation, as in Figures 2.5 and 2.6. Meanwhile, generate a new data set from the SLM with normal error, using the same centered values of the covariate, and the initial estimates $\widehat{\beta}$ for the slope and $\widehat{\sigma}^2$ for the error variance. Then

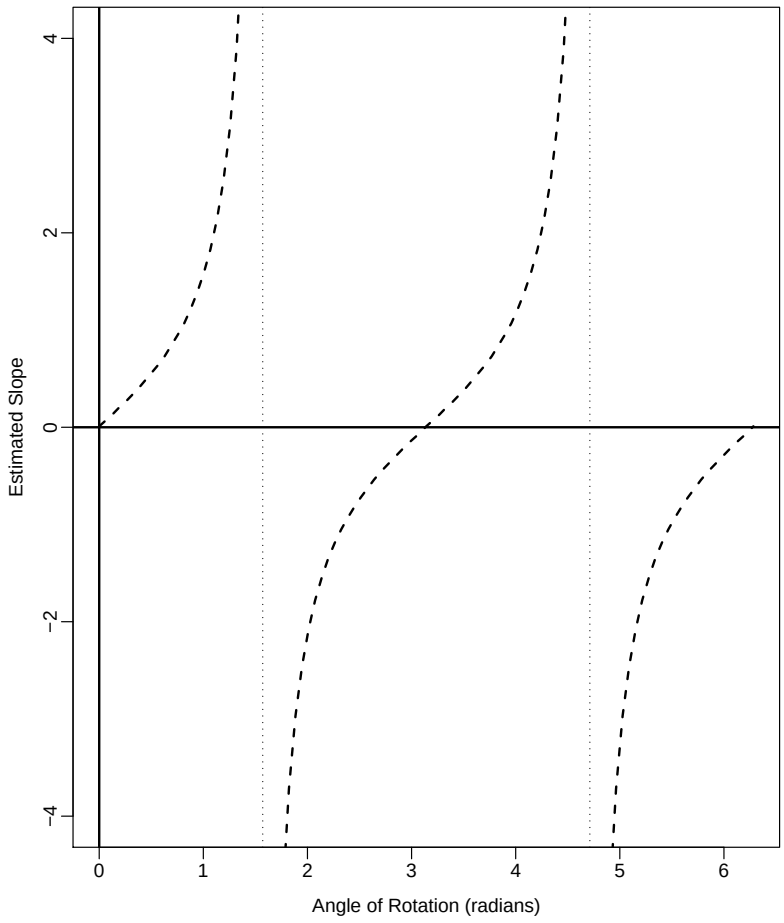


Figure 2.6: Progression of the value of the slope estimate using OLS regression as the plane rotates about the origin when the centered data are generated from the SLM with a small error variance.

compute OLS slope estimates for the generated data as the plane is rotated about the origin, and add the progression of these slope estimates to the previous plot. If the original data were generated in a manner consistent with the SLM, the two plots should be similar in amplitude, phase and shape. Of course, variability is expected in the generated data, especially when $\widehat{\sigma}^2$ is large. Thus N data sets are generated in the same way, and the rotating OLS slope estimation procedure is repeated on each

set. At each increment of the rotation there will be a distribution of N OLS slope estimates. Select an upper and lower quantile for this distribution at each angle and thereby obtain a pointwise confidence band for the path of the slope estimate if the original data may indeed be regarded as having been generated from the SLM. If the path of OLS slope estimates for the actual rotating centered data does not lie entirely within this band, it may be concluded at the corresponding significance level that the SLM is not the correct model for these data.

For instance, consider again the WHO MONICA data for the males. When the procedure described above is implemented, the results shown in Figure 2.7 are obtained. The progression of OLS slope estimates as the centered real data rotate about the origin is represented by the solid line. After $N = 10,000$ data sets are generated under the SLM, using the OLS slope estimate for the unrotated real data, and a robust estimate of the error variance based on the middle 50% of the ordered residuals, a distribution of slope estimates at each angle of rotation becomes available. The pair of dashed lines shown in Figure 2.7 represents the middle 50% of OLS slope estimates for the rotating generated data, and the pair of dotted lines represents the middle 90%. Clearly, the path for the real data does not fall within either band, since its amplitude and phase are quite different, so that there is strong evidence that the SLM would be a misspecification for these data. After the same experiment is performed with the centered WHO MONICA data for females, the progression of OLS slope estimates is also found to lie outside the confidence bands, as shown in Figure 2.8. Hence it is also inappropriate to specify the SLM for the female data. It is therefore

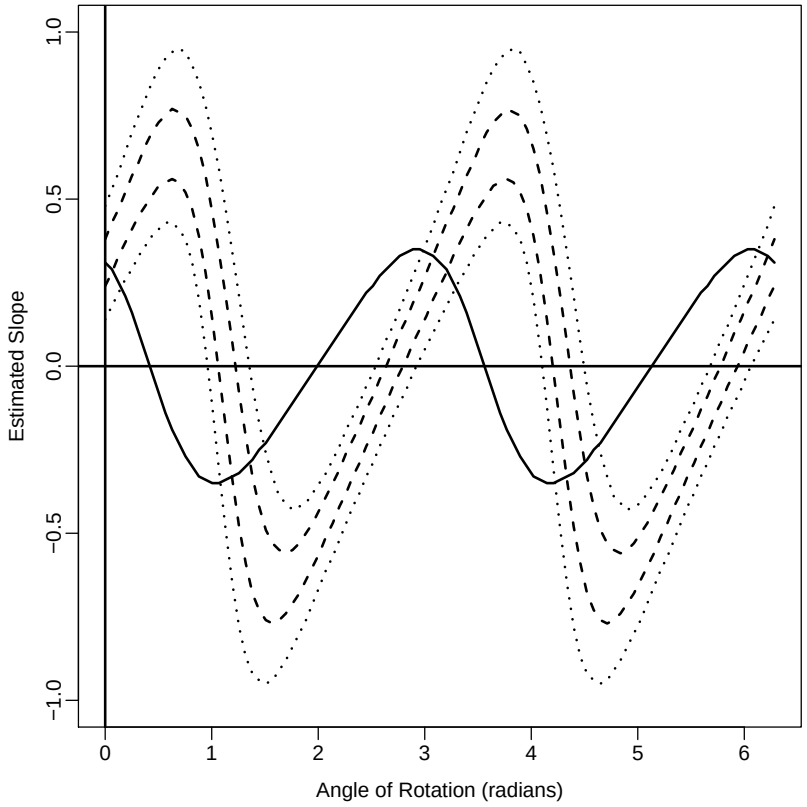


Figure 2.7: Progression of the value of the slope estimate using OLS regression on the centered WHO MONICA data for males as the plane rotates about the origin, along with 50% (dashed) and 90% (dotted) pointwise confidence bands for a specified SLM based on 10,000 generated data sets.

justified to suspect the suitability of any measurement error model which involves an equation error component for these data, as such models produce slope estimates very close to those obtained using OLS (as in [40] and [52]).

Meanwhile, suppose orthogonal least squares is used to estimate the slopes for each data set, based on the model which regards the observations as perpendicular deviations from the trendline. In this model, $Y_i = \beta X_i + \nu_i$ for $i = 1, \ldots, n$, with

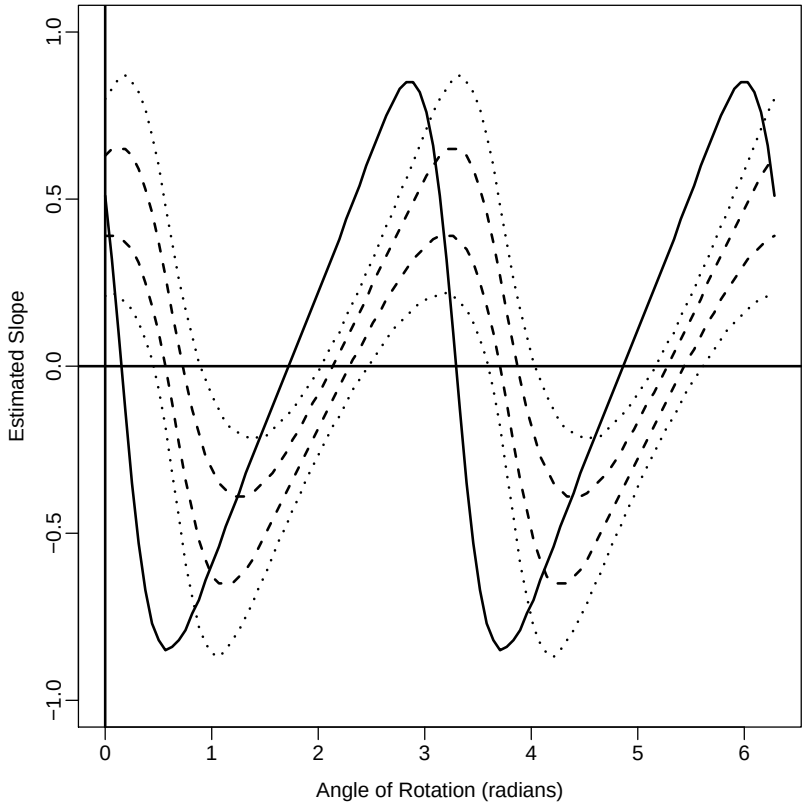


Figure 2.8: Progression of the value of the slope estimate using OLS regression on the centered WHO MONICA data for females as the plane rotates about the origin, along with 50% (dashed) and 90% (dotted) pointwise confidence bands for a specified SLM based on 10,000 generated data sets.

independent errors $\nu_i = \varepsilon_i\sqrt{\beta^2+1}$ with $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$. The progression of these slope estimates as the centered real data for males and for females rotate about the origin is represented by the solid line in each plot of Figure 2.9. Now generate $N = 1000$ data sets under this model, using the orthogonal least squares slope estimate for the unrotated real centered data, and an estimate of the error variance based on the

entire collection of residuals. The pairs of dashed lines represent the middle 50% and 90% of slope estimates for the rotating generated data. But since there is little variability over 1000 iterations, both of these confidence bands are very narrow. In this scenario the path for the real data comes very close to falling within the bands, despite the narrowness of the bands. Moreover, the paths of the estimates have the tangent function appearance that is expected when the model is specified well. Thus there is evidence that this model would be a more legitimate specification for

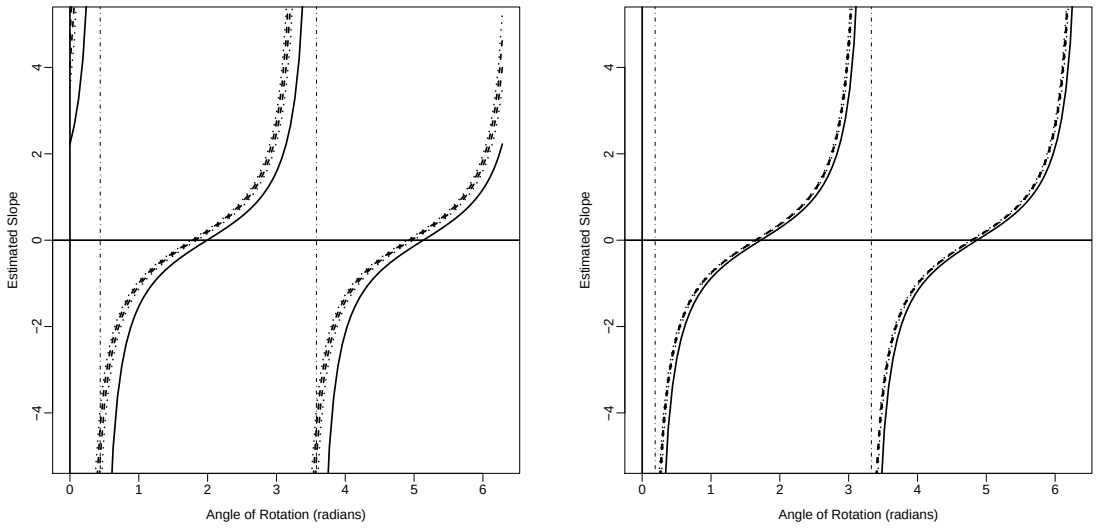


Figure 2.9: Progression of the value of the slope estimate using orthogonal least-squares regression on the centered WHO MONICA data for males (left) and for females (right) as the plane rotates about the origin, along with 50% (dashed) and 90% (dotted) pointwise confidence bands for a specified perpendicular-deviation model based on 1000 generated data sets.

these data. Hence when the measurement error for the observations on each variable is considered, it would be wise to implement a model which aims to minimize the sum of perpendicular distances between the observations and the trendline. The line

segment model proposed here approaches that general framework, and thus provides a slope estimate which is more consistent with the geometry of the data cloud.

2.8 Discussion

Many scientific investigations involve assessing the association between two variables when measurements recorded on both variables are subject to random error. When the variance of this error differs from one subject to another, a heteroscedastic measurement error model is appropriate. Conventional ME models incorporate an equation error component, which involves making potentially insupportable assumptions about the unknown data-generation mechanism. Misspecification of the chosen model can lead to significant underestimation of the slope, regardless of the estimation procedure used.

In this paper we have provided an alternate heteroscedastic ME model based on a line-segment parameterization. This model does not incorporate equation error and is symmetric in both variables. For any setting in which this model corresponds well with the underlying data-generation mechanism, we have provided an accurate estimate of the linear association between the two variables, signified by the slope of the line segment, along with a reliable estimate of its precision. We have demonstrated through simulations that, under conditions when the line-segment model is properly specified, the corresponding variance estimate will yield smaller confidence intervals than will equation-error models when the error variances are not small. This novel estimation procedure enables an investigator to make precise inferences about the

slope when heteroscedastic ME models are applied to scientific data, and thereby draw conclusions that will generally be more trustworthy than those derived using other ME models. Moreover, because the line-segment parameterization is robust against influential points which may plague equation-error models when their implementation is misspecified, the advantages of our model are further reinforced.

Chapter 3

Detecting physiological synchrony during dyadic interactions

The synchronization of oscillatory systems – or coupled oscillations – is widely studied in the biological and physical sciences (e.g., in [47], [53], and [65]), and also has applications in the social sciences, economics, and medicine (e.g., in [55]). The synchrony of these oscillations can provide information about the system not available from separate univariate analyses. Consider, for example, the investigation of several electroencephalographic (EEG) signals measured simultaneously from an individual's scalp during a particular task. Each signal could be analyzed separately, and those with the most activity would indicate an area of relative activation. However, various signals can show simultaneous activation, revealing communication between different areas of the brain during the task ([14], [20], and Chapter 1 of this dissertation). Furthermore, different types of such coherence – or synchrony – may be evident for

different mental processes, as is the case with epileptic seizures ([55]). Thus, the study of synchrony and oscillatory systems can provide a valuable means of studying psychophysiological processes, as well as possible changes in those processes as a function of different stimuli and conditions.

In the current study the application of two recently developed methodologies for examining the relations between two time series is proposed. The first technique is the Empirical Mode Decomposition (EMD), an algorithm to filter continuous time series data. The second method is the structural heteroscedastic measurement-error (SHME) model, which is adapted here for detecting a linear association between two discrete time series. These techniques are applied to physiological data from individuals in couples that participated in a laboratory-based social interaction task.

The chapter is organized as follows. First, some of the common synchronization measures and their rationale are briefly reviewed in the context of emotional processes in dyadic interactions. Second, the EMD and SHME methods are described, with details for each of the required steps. Third, an application of the proposed methods is provided for illustration. The chapter is ended with a brief discussion of the potential of these models in psychophysiological research.

3.1 Synchronization measures

Synchronization measures have become an important tool for exploring the associations between time series. Multiple methods now exist to identify and characterize synchronization, including indices of linear interdependence, such as cross-correlation,

coherence, and event-related coherence, and more recent measures of nonlinear interdependence, such as mutual information ([37]). In econometric research, for example, one of the most common methods to assess synchronization between two time series is co-integration ([25], [15]). In psychological research, perhaps the most standard method consists of cross-correlations (e.g., in [24], [46]). This method can be useful to examine concurrent and lagged relations between two time series, either the entire series or through windows of interest (e.g., [3]).

3.1.1 Synchronization of emotion in dyadic interactions

The theoretical rationale for this analyses comes from human and animal research suggesting that psychophysiological linkages between two conspecifics are an inherent element of social bonding and attachment ([7], [8], [16], [23], [26], [31], [32], [58]). The study of dyadic interactions indicates that emotional exchanges between the two members of a couple can be highly interdependent ([63], [10], [17], [18], [61]). This research shows, for example, that the adoption of one individual's emotion state by another promotes relationship longevity ([29]), that the length of the relationship between romantic and non-romantic partners corresponds to the level of emotional coherence that the pair maintains ([1]), and that the facial expression and emotional tone exhibited by romantic partners is a strong predictor of relationship dissolution ([44]).

Research in dyadic interactions using psychophysiological signals is scarcer. In a classic study of couples, Levenson and Gottman [43] found that, during a conver-

sation of disagreement, distressed couples showed significantly higher levels of synchrony between the partners' autonomic response signals than non-distressed couples. Moreover, this synchrony was predictive of current marital satisfaction in the same couples. This study notwithstanding, the relative absence of research on psychophysiological synchrony in couples is conspicuous, largely because most theories of human attachment and emotion regulation suggest that the emotional experiences of one member of a couple are highly related- if not dependent upon- the experiences of his or her partner (cf. [58]). The view taken here is that a large part of the problem is methodological; theoretical developments in this area greatly outpace methodological innovations. In order to fully understand dyadic emotion regulation and psychophysiological synchrony in couples, the field needs accessible methods that can capture and adequately represent the complexity in interdependent emotional regulatory systems ([9]).

3.1.2 Synchrony between continuous measures: Signal extraction using Empirical Mode Decomposition

The Empirical Mode Decomposition (EMD) was developed by Huang, et al. [34], to filter continuous data into any number of intrinsic mode functions (IMFs). Each IMF represents a particular frequency of the original data, and the most powerful frequencies are separated out of the original time series until a desired number of IMFs has been achieved (or until no powerful frequency can be detected within the residual series). These IMFs must satisfy two conditions. First, in each IMF, the

total number of extrema and the total number of zero crossings may differ at most by one. Second, at every point in the IMF, the mean value of the envelopes defined by the local maxima and the local minima must equal zero. These conditions are necessary for the purpose of defining the concept of instantaneous frequency in a meaningful way. The IMFs are extracted from a time series one-by-one beginning with the highest intrinsic frequency using an iterative process called *sifting*. The goal of this process is the empirical identification of intrinsic oscillatory modes in the data based on their characteristic time scales. The time lapse between successive extrema define this time scale.

In the sifting process, the local maxima of the original time series are identified and connected by a cubic spline line to form an upper envelope. A lower envelope is similarly formed. Adjustments at the signal boundaries must be implemented to eliminate boundary effects in forming the cubic spline. Then the mean of the two envelopes is computed and subtracted from the original time series to form a new series. If this new series satisfies the two IMF conditions, it is taken as the first IMF. Otherwise, the process is repeated on the new series, and so on, until the IMF conditions are satisfied. Once the first IMF is identified, it is subtracted from the original data, and the residual becomes the starting point for finding the next IMF. The procedure stops when the residual signal fails to yield any suitable IMF candidates, or the maximum number of desired IMFs is achieved. Kim and Oh [35] have developed an `R` package called `EMD` that implements this procedure very efficiently.

The input to the EMD is any continuous time series, without any consideration of stationarity. The output from the EMD consists of a residual signal and a set of n IMFs in decreasing-frequency order. The first few IMFs cumulatively carry any high-frequency noise in the original time series, while the latter k IMFs together carry the actual signal of interest. Summing the residual and the last k IMFs together, thus produces the denoised time series. An important goal here is determining the value of k. The Fast-Fourier Transform may be used to detect the most powerful frequencies within each IMF, and only those IMFs whose dominant frequencies are below a desired threshhold are selected. However, it is informative to compare the plot of the extracted denoised signal with that of the original signal in every case to determine whether more or fewer IMFs should be included. The synchrony between the resulting signals can be assessed using cross-correlations. These steps are illustrated with empirical data in subsequent sections.

3.1.3 Synchrony between discrete measures: Slope estimation using a Structural Heteroscedastic Measurement-Error Model

The SHME model may be adapted to detect a linear association between discrete time series. This approach is particularly suited for capturing the relationship between two time series when the variability within each time series is not constant. The first step in the application of the SHME model consists of filtering the raw signal. For example, if the observed time series consists of EKG data (as in the empirical application in this study), the raw signal is transformed into a heart rate in the form of, say, beats

per minute. This can be accomplished in various ways, as is illustrated in subsequent sections. A line-segment version of this model is presented in Chapter 2, where it is shown that the precision of the parameter estimates under the line segment model was superior to that of the equation error model of Patriota, et al. [52], when the variance of the measurement error was not small. However, in this application the latter model is used, since it has been discerned that the measurement error—in the form of the computed sampling error—is rather small.

Once the data are filtered, each of two time series are partitioned into n segments of some specified width, where n depends on the duration of the task. The choice of the segment width is a function of both detailed information and precision. Denote these segments $I_1, \ldots, I_n$. Consider, for example, a selected time of five seconds for the segments. Each segment I_i will consist of m_i distinct heart rate values $x_{j,i}$, $j = 1, \ldots, m_i$, for one of the series (e.g., one person's signal), each of which lasts for k_j milliseconds, and p_i distinct heart rate values $y_{j,i}$, $j = 1, \ldots, p_i$ for the other series (e.g., the other person's signal), each of which lasts for l_j milliseconds. Thus, for each segment I_i, the weighted mean heart rates are computed as

$$u_i = \frac{1}{5000} \sum_{j=1}^{m_i} k_j x_{j,i} \quad \text{and} \quad v_i = \frac{1}{5000} \sum_{j=1}^{p_i} l_j y_{j,i}$$

for each series, respectively. Similarly, the weighted variances of the mean heart rate for each segment are approximated as

$$\sigma_i^2 \approx s_i^2 \sum_{j=1}^{m_i} \left(\frac{k_j}{5000} \right)^2 \quad \text{and} \quad \tau_i^2 \approx t_i^2 \sum_{j=1}^{p_i} \left(\frac{l_j}{5000} \right)^2 ,$$

where s_i^2 and t_i^2 are the sample variances for each time series over I_i, respectively.

Since these $2n$ variances are potentially different across the two series (e.g., as in two individuals in a couple), any method for estimating the linear association between $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ must account for heteroscedastic measurement error on each variable.

As described in the second chapter, the SHME model with equation error assumes that

$$u_i = \chi_i + \varepsilon_i \ , \qquad v_i = \mu_i + \nu_i \qquad \text{and} \qquad \mu_i = \alpha + \beta\chi_i + \gamma_i \ ,$$

where the independent measurement errors are $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\nu_i \sim \mathcal{N}(0, \tau_i^2)$, and the equation error is $\gamma_i \sim \mathcal{N}(0, \sigma^2)$. Moreover, all error terms are mutually independent.

Under a structural model, both χ_i and μ_i are assumed to be random with unspecified but finite first and second moments. Note that the symmetry of this model would allow one to switch μ_i and χ_i in the latter model equation above, so that there is no implication of directionality. Techniques for estimating the slope β in this setting are available in the literature (e.g., in [6], [40], [52], and in Chapter 2 here). When the measurement error variance is small, as in the application here, the method-of-moments estimate of Patriota, et al. ([52]), provides an efficient estimate of the slope that is simple to compute. This approach will be used to estimate β and test whether it is significantly nonzero in the empirical application.

To this end, let

$$S_{uu} = \sum_{i=1}^{n} \frac{(u_i - \overline{u})^2}{n-1} \ , \quad S_{uv} = \sum_{i=1}^{n} \frac{(u_i - \overline{u})(v_i - \overline{v})}{n-1} \ , \quad S_{vv} = \sum_{i=1}^{n} \frac{(v_i - \overline{v})^2}{n-1} \ ,$$

$$\sigma_n^\star = \sum_{i=1}^{n} \frac{\sigma_i^2}{n}\,, \quad \tau_n^\star = \sum_{i=1}^{n} \frac{\tau_i^2}{n}\,, \quad \sigma_n^{\star\star} = \sum_{i=1}^{n} \frac{\sigma_i^4}{n}\,, \quad \text{and} \quad (\sigma\tau)_n^\star = \sum_{i=1}^{n} \frac{\sigma_i^2\tau_i^2}{n}\,.$$

Moreover, let $\sigma_\chi^2 = \mathrm{Var}(\chi)$, $\sigma^\star = \lim_{n\to\infty} \sigma_n^\star$, $\sigma^{\star\star} = \lim_{n\to\infty} \sigma_n^{\star\star}$, $\tau^\star = \lim_{n\to\infty} \tau_n^\star$, and $(\sigma\tau)^\star = \lim_{n\to\infty}(\sigma\tau)_n^\star$. Then, having established that the distribution of $\sqrt{n}(\widehat{\beta} - \beta)$ converges to $\mathcal{N}(0, \omega)$, the slope estimate β and its asymptotic variance ω under this model are

$$\widehat{\beta} = \frac{S_{uv}}{(S_{uu} - \sigma_n^\star)} \quad \text{and} \quad \omega = \frac{2\beta^2(\sigma^{\star\star} - \sigma_\chi^4) + \pi}{\sigma_\chi^4}\,,$$

where

$$\pi = \beta^2\sigma_\chi^2\sigma^\star + \sigma^2\sigma_\chi^2 + (\sigma\tau)^\star + \sigma^2\sigma^\star + \sigma_\chi^2\tau^\star + 2\beta^2\sigma_\chi^4\,.$$

Thus $\mathrm{Var}(\widehat{\beta}) \approx \omega/n$ for n large. Substituting the parameter estimates given in Patriota, et al. [52], and simplifying, the estimated variance of $\widehat{\beta}$ is

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \frac{2S_{uv}^2[\sigma_n^{\star\star} - (S_{uu} - \sigma_n^\star)^2]}{n(S_{uu} - \sigma_n^\star)^4} + \frac{S_{uv}^2 + S_{uu}S_{vv} + (\sigma\tau)_n^\star - \sigma_n^\star\tau_n^\star}{n(S_{uu} - \sigma_n^\star)^2}\,.$$

Then reject the hypothesis $H_0 : \beta = 0$ when the ratio $\widehat{\beta}/\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$ deviates significantly from zero with respect to the standard normal. This procedure is illustrated with empirical data in subsequent sections.

3.2 Empirical illustration

3.2.1 Procedures

The data in this study are from four couples who completed psychophysiological measurements as part of a study of dyadic interactions (see [18]). All four couples were heterosexual with ages across all participants ranging from 26 to 32 years. The

first three couples defined their relationship as “exclusively dating” and the fourth coupled as “married.”

3.2.2 Measures

Physiological measures were collected through the MP150 physiological data collection system (BIOPAC systems) and AcqKnowledge. Stimuli were administered in a computer monitor using E-prime (Psychology Software Tools, Inc.). Three autonomic response variables were recorded from each individual within the dyad simultaneously throughout the experiment. Respiration was recorded using an elastic belt that was attached to each of the participants. The belt was placed on each subject’s chest at the point of highest extension during inhalation and exhalation. The center of the belt contained a device that recorded the level of stretch within the belt at any moment, with greater stretch indicating inhalation and lower stretch indicating exhalation. Level of stretch within the belt was measured continuously at a rate of 1000hz.

Thoracic impedance was measured using four spot electrodes placed at the well of the neck, back of the neck, center of the chest, and center of the back. This configuration is known formally as the Qu, et al., configuration ([54]). Each spot electrode came prepared with Ag/AgCl paste, and had an adhesive collar to ensure both good conductivity as well as stationarity of the electrode during the experiment. Level of impedance was measured continuously at a rate of 1000hz. An electrocardiogram was recorded using a lead II configuration, with spot electrodes on the left and right torso

(bipolar leads), as well as the right collarbone (unipolar lead). All spot electrodes came prepared with Ag/AgCl paste, and had an adhesive collar (for the same reasons as the impedance measures). The electrocardiogram was measured continuously at a rate of 1000hz. All signals were recorded via the BIOPAC 150 and sent online to an external computer for storage using Acqknowledge. Though Acqknowledge contains a suite of processing techniques available at the researcher's disposal, all of the raw signals were exported to text files and processed in the software package `R` [56] for analysis.

Participants visited a laboratory for the physiological assessment in couples. They were instructed about the experiment and completed three tasks. During the first task (Baseline) participants were seated in comfortable armchairs and instructed to relax and refrain from making bodily movements or gestures for a period of five minutes. Blindfolds were placed over the participants' eyes and the overhead lights were turned off in order to induce an environment of relaxation. The purpose of this first task was to gain a baseline signal for each individual. During the second task (Gazing), participants were asked to gaze into one another's eyes without talking or touching each other for three minutes. The purpose of this task was to elicit a stressful interaction between the participants. During the third task (In-sync), they were instructed to try to become in-sync with each other for three minutes. The term in-sync was described to the participants as being analogous to becoming one individual, and therefore their goal would be to match their partner's physiological signal. After the completion of the three tasks, the participants were debriefed and

paid for their participation.

3.2.3 Application of EMD to respiration and impedance

The EMD was applied to two continuous signals — the respiration and the thoracic impedance — which were obscured by high-frequency components taken as background noise. The respiration signal is a measurement of the expansion and contraction of the rib cage as the subject breathes, and thus oscillates about a fairly constant value at a varying frequency. The impedance measures the cyclical changes in cardiopulmonary output and, thus is correlated with heartbeat and respiration. Figure 3.1 displays the raw impedance signal for one individual (i.e., male) in Couple 3 during the first minute of the gazing task. As depicted in the figure, this time series is very noisy.

The EMD of this impedance series produces 10 IMFs (displayed in Figure 3.2).

Only the last two IMFs are selected, and added to the residual, to obtain the smoother signal shown in Figure 3.3.

Preceding IMFs could be added to obtain more detail, but at the cost of gaining noise. Figure 3.4 displays the resulting denoised impedance signals for both members of each couple during the first minute of the baseline task.

After removing the noise from each individual's time series across the three tasks, occasions of synchrony were detected between the denoised signals for the two individuals in each of the couples. For this, each pair of extracted signals was examined using a sliding window of a fixed six-second width, which moved in two-second increments

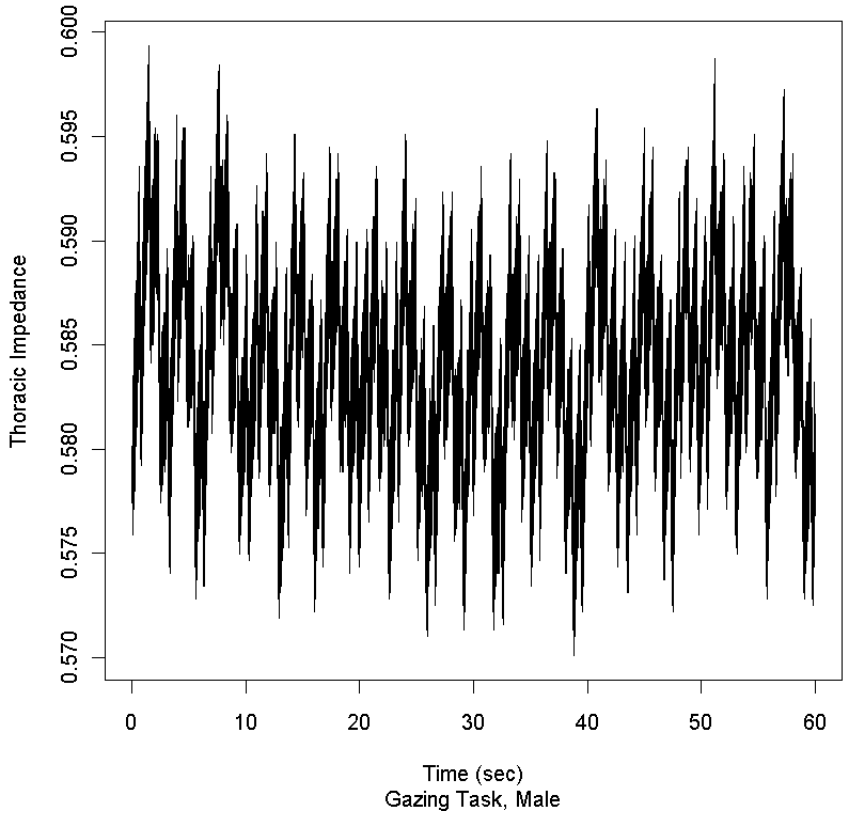


Figure 3.1: Male's impedance signal during gazing task, for Couple 3.

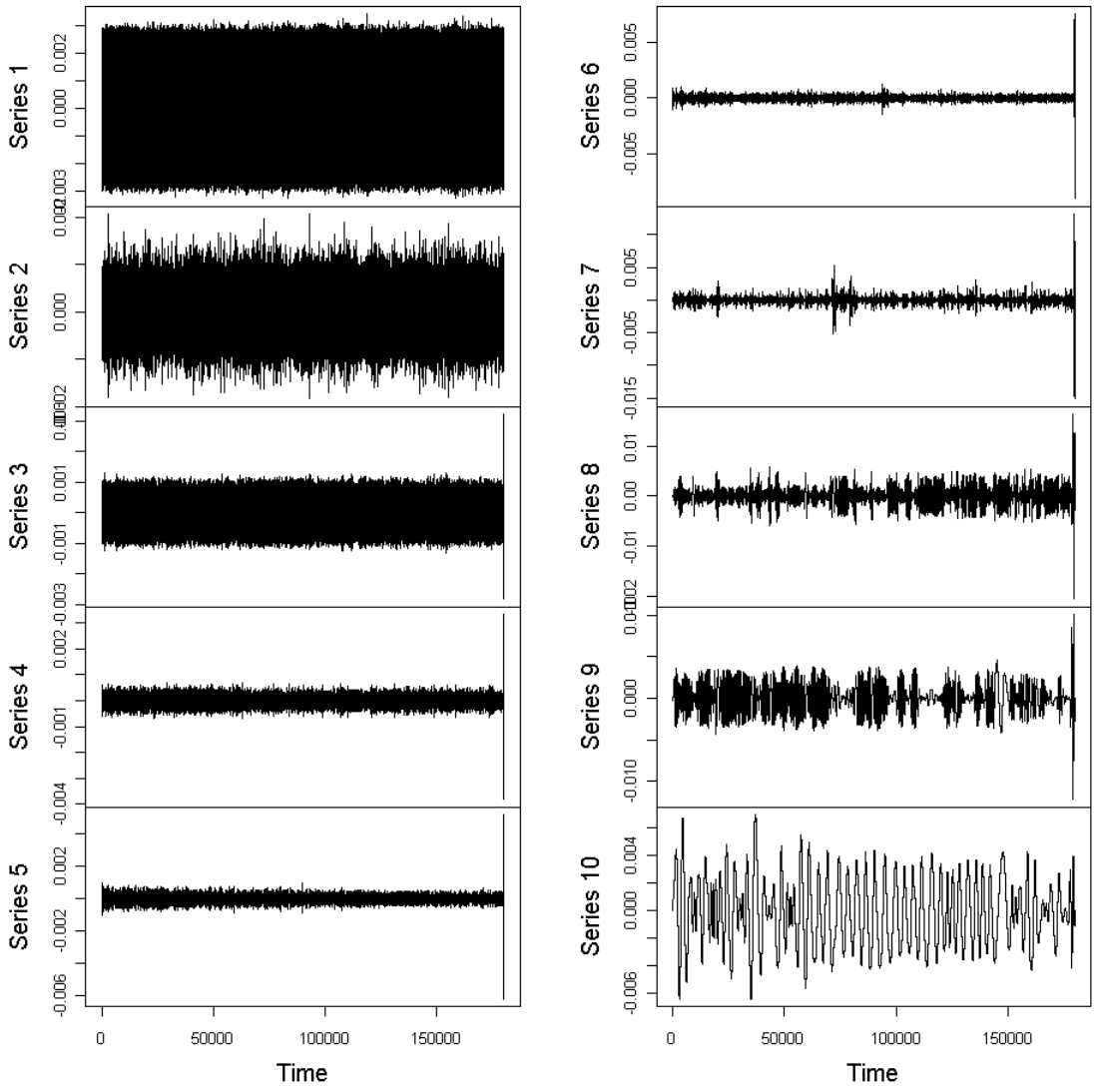


Figure 3.2: IMFs produced by EMD of male's impedance signal during gazing task.

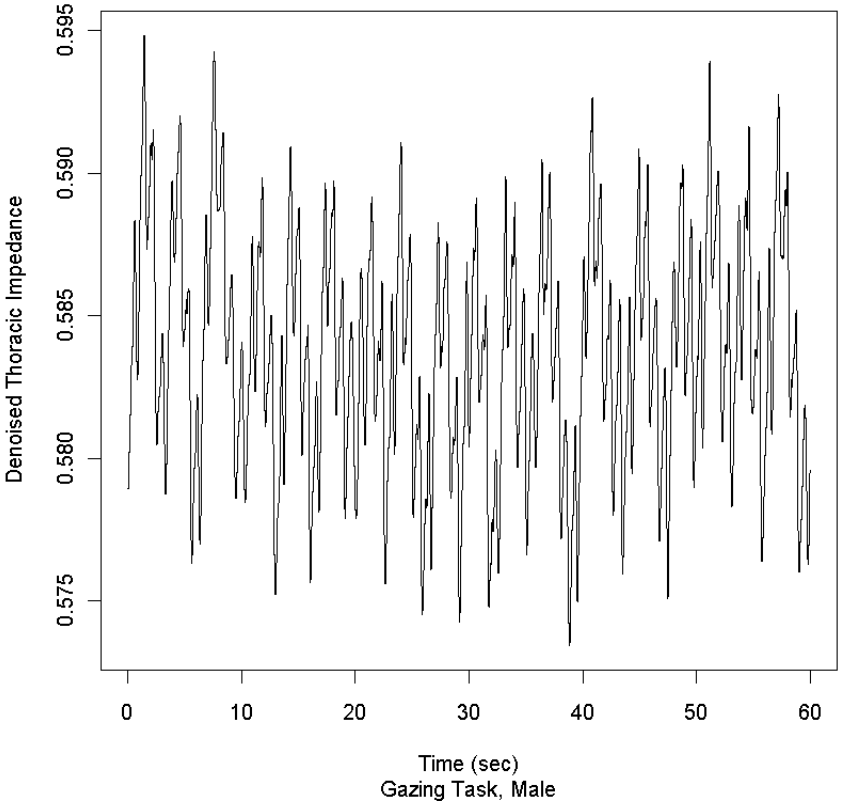


Figure 3.3: Denoised form of male's impedance signal during gazing task.

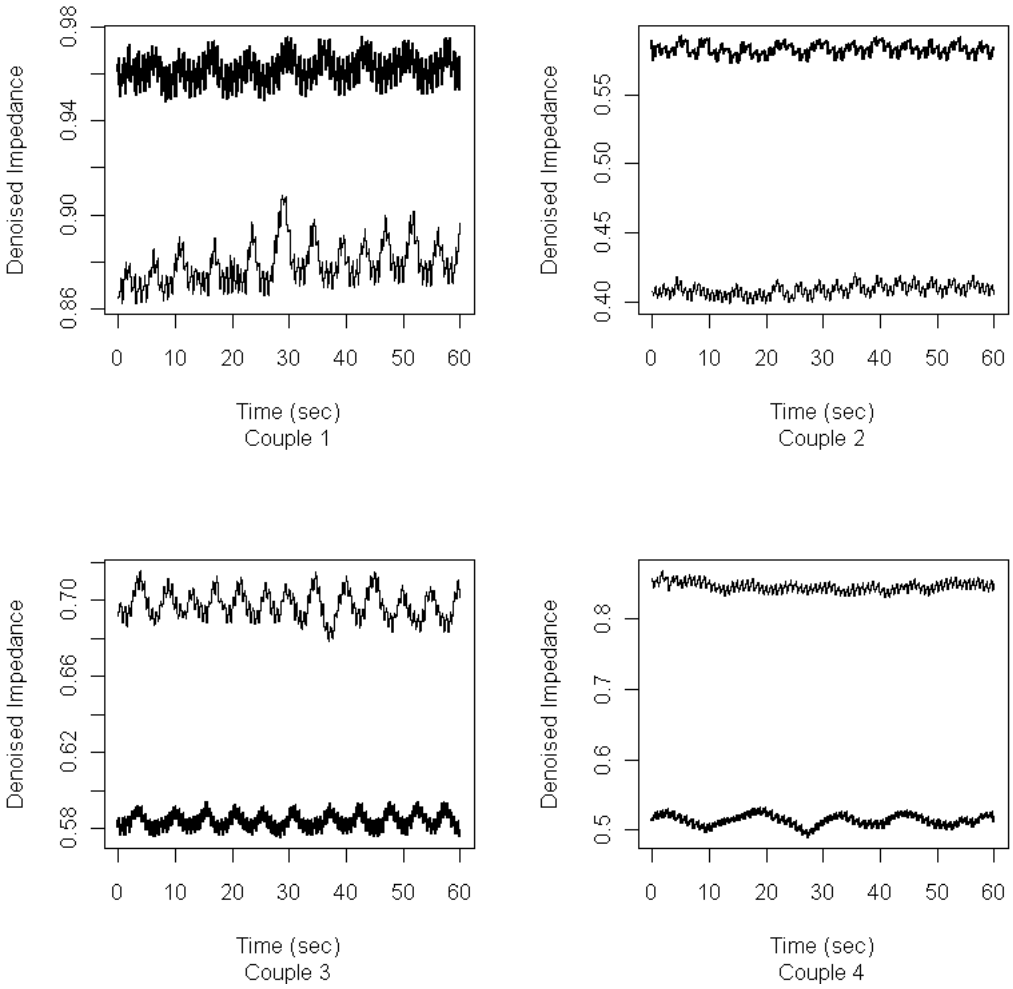


Figure 3.4: Denoised impedance for the male (dark) and the female (light) during the baseline task for each couple.

from the beginning to the end of each three- to five-minute task. This choice of the window width and the increment size is arbitrary; other choices result in equivalent outputs but with different details.

At each point, the cross-correlation was then computed between the signals over a range of lags, and the maximum computed value was selected as a measure of synchrony during that moment. The default lag range in `R` was used, which is $\pm\lfloor 10 \log_{10}(3000) \rfloor = \pm 34$. This measure is referred to as the *Instantaneous Coupling* (IC) strength, as in the first chapter. Figure 3.5 displays the IC series for the third couple during the baseline task with respect to their respiration (solid line) and their impedance (dashed line). Note that the two series are highly correlated, as one would expect. Moreover, there appear to be many occasions during this task when the couple's physiological responses appears to be highly synchronized in both variables.

For each of the three tasks in the experiment, the proportion $\widehat{\pi}$ of IC values that exceeded a given threshhold was then computed. Threshholds of 0.6 for the respiration and 0.5 for the impedance were chosen, as these values provided a reasonable baseline proportion (i.e., not too small). Finally, the proportions above the threshhold for the second and third tasks were compared with that from the baseline, and a routine hypothesis test was conducted to determine whether any subsequent proportion was significantly higher than the baseline proportion. If so, it was considered as evidence of synchronization between the individuals' physiological signals. Table 3.1 displays the results of these analyses for respiration and impedance for each of the four couples.

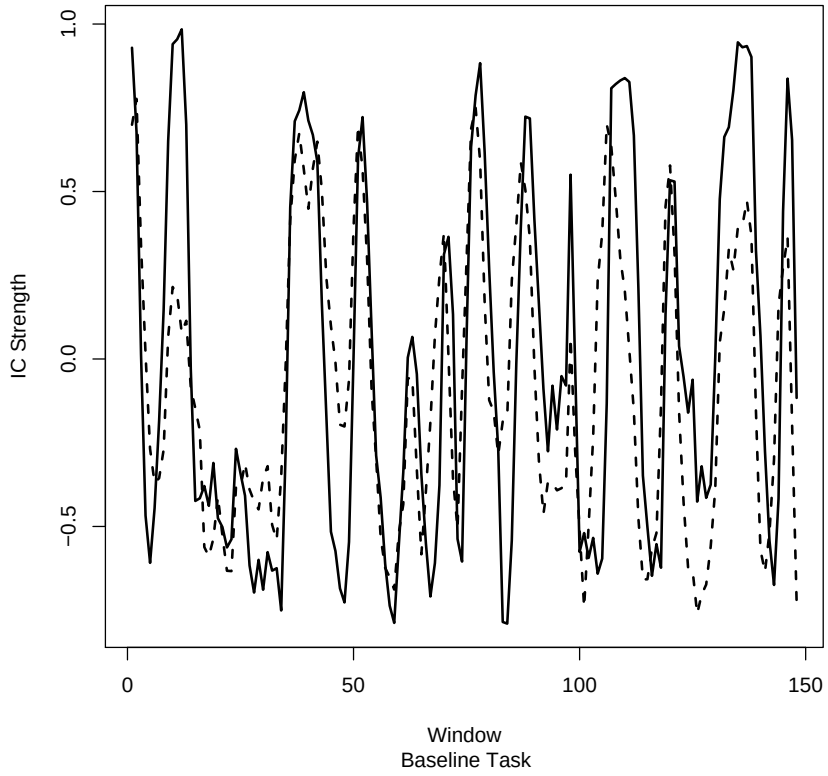


Figure 3.5: IC strength for Couple 3 during baseline task, with respect to respiration (solid line) and impedance (dashed line).

Table 3.1: Significant increase in relative frequency of strong Instantaneous Coupling across tasks

Couple	Task	Respiration $\widehat{\pi}$	P-value	Impedance $\widehat{\pi}$	P-value
1	baseline	0.149	—	0.020	—
	gazing	0.239	0.048 *	0.102	0.008 **
	in-sync	0.886	0.000 ***	0.011	0.709
2	baseline	0.068	—	0.007	—
	gazing	0.125	0.080	0.045	0.048 *
	in-sync	0.659	0.000 ***	0.364	0.000 ***
3	baseline	0.236	—	0.122	—
	gazing	0.125	0.988	0.045	0.986
	in-sync	0.818	0.000 ***	0.375	0.000 ***
4	baseline	0.216	—	0.027	—
	gazing	0.114	0.984	0.148	0.001 ***
	in-sync	0.841	0.000 ***	0.000	0.979

$0.05 < * < 0.01 < ** < 0.001 < ***$

For respiration, the results indicate a significant increase in synchrony from baseline between the partners' signals during the in-sync task, for all four couples. During the gazing task, such increase in synchrony was only evident for the first couple. With regard to impedance, the significant increase in synchrony between the partners was perceptible during the gazing task for three of the couples, and such amplification was also true for two couples during the in-sync task.

3.2.4 Application of SHME to heart rate

In the first step, the raw EKG signal during each task was transformed into a heart rate. For this, the duration of each R-R (peak-to-peak) interval (in milliseconds) was determined, and its reciprocal was used to compute the heart rate (in beats per millisecond). Then the obtained values were multiplied by 60,000 to convert them to beats per minute. Because the first recorded ventricular contraction usually does not occur in the EKG signal until after a few milliseconds, the beginning of the time series was padded with the first computed heart rate value. Similarly, because the last recorded ventricular contraction usually occurs a few milliseconds prior to the end of the EKG signal, the end of the heart rate time series was padded with the last computed value.

Figure 3.6 displays the resulting heart rate signals during the first 100 seconds of the baseline task for both individuals in the four couples. Note that each heart rate oscillates over a large range of values except for that of the male in the second couple, who has an almost constant heartbeat. In every case, the female's heart tends to

Table 3.2: Slope estimates for association between heart rates using the SHME model across tasks

Couple	Task	$\widehat{\beta}$	$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$	P-value
1	baseline	0.002	0.274	0.993
	gazing	1.071	0.212	0.000 ***
	in-sync	1.344	0.626	0.032 *
2	baseline	0.358	0.703	0.610
	gazing	0.504	0.436	0.248
	in-sync	0.579	0.473	0.221
3	baseline	−0.089	0.079	0.254
	gazing	0.171	0.099	0.083
	in-sync	0.369	0.149	0.013 **
4	baseline	−0.142	0.185	0.445
	gazing	−0.227	0.961	0.813
	in-sync	0.497	0.239	0.037 *
$0.05 < * < 0.01 < ** < 0.001 < ***$				

beat faster. The objective here is to examine whether there is a significant linear association between the two heart rates within any of the tasks.

For each of the tasks, the heart rate time series for both the male and female were partitioned into n segments of five seconds following the procedure described in previous sections. Then the SHME model was applied to the EKG data generated separately for each of the four couples. The results from these analyses are presented in Table 3.2.

The results from these analyses indicate that, during the gazing task, two of the

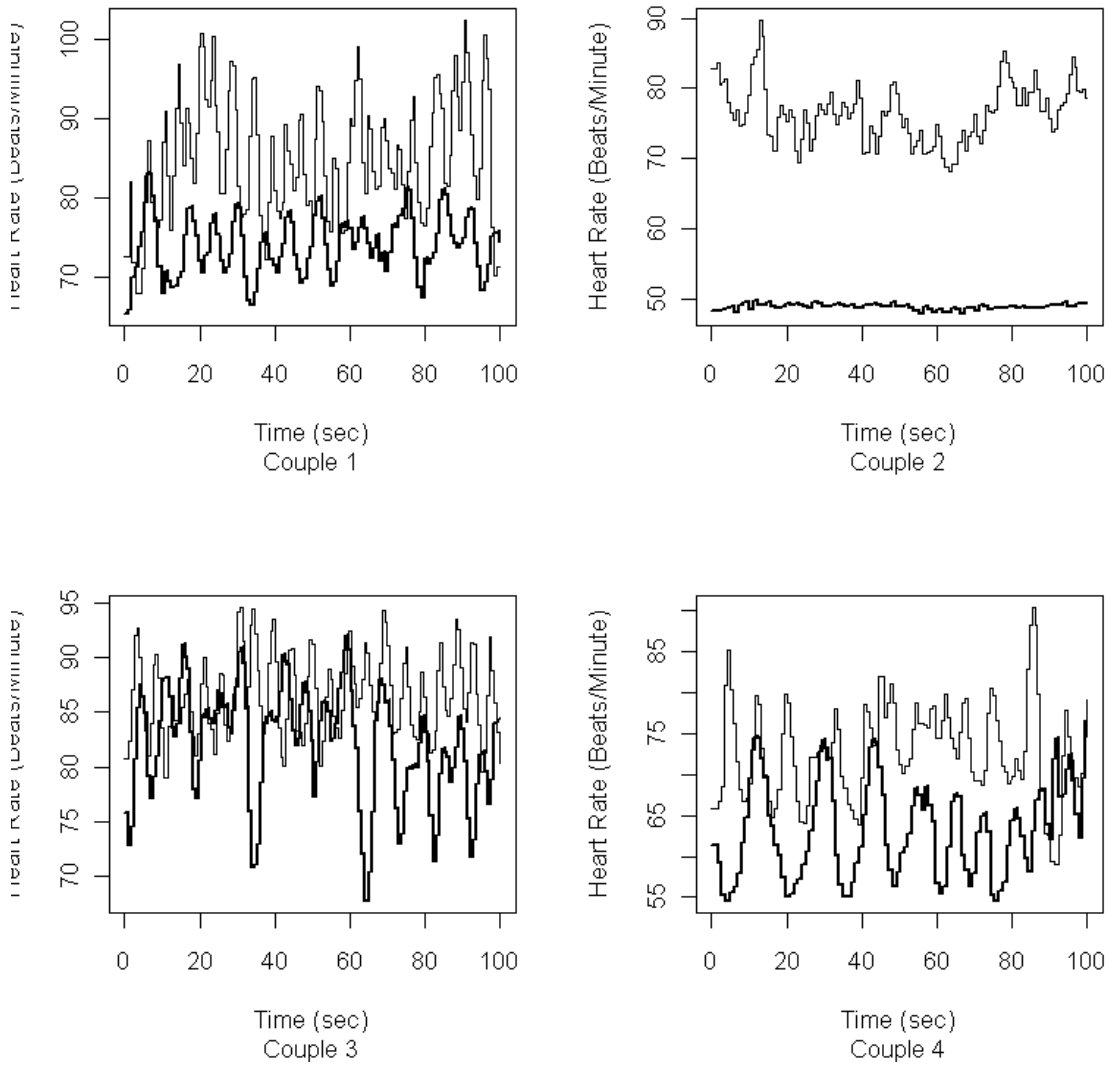


Figure 3.6: Heart rate for the male (dark) and the female (light) during the baseline task for each couple.

four couples showed a significant linear association between their heart rates. During the in-sync task, such synchrony between the partners' heart rates was evident for three couples. In contrast, and as expected, no synchrony was perceptible during the baseline task for any couple.

3.2.5 Cross-equivalence analysis

To confirm that the discovery of synchrony in heart rate, respiration, and thoracic impedance within each of the four couples using these methods was not artificial, the same methods were applied to two mismatched couples. For this, the male from Couple 1 was paired with the female from Couple 2 as one dyad, and the male from Couple 2 was paired with the female from Couple 3 as a separate dyad. Then the same procedures were implemented to detect synchrony in heart rate, respiration, or thoracic impedance during the three tasks. Table 3.3 reports the results from these cross-equivalence analyses. These results indicate no synchrony across couples and tasks (i.e., all P-values exceeding 0.1).

3.3 Discussion

3.3.1 Summary of results

In this chapter two techniques for assessing synchrony between psychophysiological time series are presented. For respiration and thoracic impedance, which are continuously oscillating signals, the EMD was used to filter the data and extract smooth versions of the time series. A moving window was applied to measure the maximum

Table 3.3: Measures of synchrony between heart rates, respiration and thoracic impedance for mismatched couples across tasks

Mismatched Couple	Task	$\widehat{\beta}$	$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$	P-value
1	baseline	-11.525	12.356	0.823
	gazing	0.250	0.206	0.117
	in-sync	-54.825	482.732	0.545
2	baseline	0.000	0.001	0.500
	gazing	0.023	0.022	0.151
	in-sync	0.000	0.021	0.500

Mismatched Couple	Task	Respiration $\widehat{\pi}$	P-value	Impedance $\widehat{\pi}$	P-value
1	baseline	0.095	—	0.041	—
	gazing	0.091	0.538	0.011	0.930
	in-sync	0.148	0.118	0.080	0.119
2	baseline	0.230	—	0.108	—
	gazing	0.091	0.999	0.045	0.968
	in-sync	0.216	0.598	0.114	0.448

cross-correlation between the signals of the two individuals in the couple within the window over a lag range, and to determine when this coupling exceeded a chosen threshold. The relative frequency of high coupling values during the baseline was then compared with those during the gazing and in-sync tasks. Synchronization in respiration or impedance was inferred when the proportion of coupling occurrences increased significantly from the baseline to the experimental tasks.

For heart rate, which is discrete in nature, the SHME model with equation error was applied to identify synchrony between the partners' signals. Using this approach, the slope representing the linear association between the heart rates of the two individuals in the couple during each of the three tasks was estimated. This slope was taken as an indicator of synchronization between the two individuals' heart rate. Importantly, a cross-validation analyses provided no evidence for synchrony when different members of a couple were randomly paired, thus providing evidence for the discriminative validity of the synchrony detection approaches.

Synchronization of the physiological signals was regarded as a reflection of emotional coherence between the two individuals in the couple. For example, during the in-sync task, participants might have concentrated on matching each other's breathing—as a way to mirror their partners' physiological state—thus resulting in an increase in synchrony for respiration. This effect might have carried over to the impedance. Similarly, matching each other's breathing could have resulted in an increase of the coupling between their heart rates. Accordingly, the methods used in these analyses appear to be useful to study emotional coregulation in dyadic inter-

actions (c.f. [58]). An explanation for the detection of synchrony during the gazing task for some of the couples, however, is elusive. It is plausible that these partners have developed a subconscious habit of synchronization, so that they matched their breathing rates without any prompting to do so.

3.3.2 Methodological considerations and future directions

The two approaches for assessing synchrony described in this chapter present a number of benefits. For example, the EMD algorithm, as a tool to parse out noise from continuous data, has two important advantages over other standard methods. First, it does not rely on assumptions of stationarity, assumptions required by methods such as the Fourier transform. Second, in the decomposition of the original series via EMD, there is no leakage of energy, which is common in techniques such as the Wavelet transform. Moreover, in many situations, heart rate data are analyzed using methods for continuous signals. The heart rate signal, however, constitutes a step function, since it is constant on intervals between contractions. Hence, analyzing this signal as a continuous measure is not appropriate. A smoothing method could be used to transform the step function into a continuous signal, but making inferences using an imputed signal is hard to justify statistically.

A fundamental hope for the proposed statistical methods is that they can be used profitably to better understand dyadic emotion regulation and human coregulation. When two individuals interact, we assume that emotional synchronization is a continuous process that is best studied in a manner that is as close to the raw data

as possible. The EMD and SHME approaches allow for this type of data analysis. One obvious extension of these analyses is the use of covariates to assess the extent to which psychophysiological synchronization is related to couple-level or individual difference variables of interest. For example, when studying intact couples, the approaches described here can be examined as a function of marital satisfaction or attachment styles, with the degree of synchronization evidenced across a study paradigm serving as a both an outcome variable (e.g., do more highly satisfied coupled evidence greater heart rate synchronization?) as well as a predictor of future relationship outcomes. Also, this study investigated associations between two time series. An important extension of this work would involve the use of multivariate time series. For example, a pertinent question here is how to identify synchronization among multiple physiological signals. In particular, emotion researchers would be interested in examining under which conditions, and to what extent, such multivariate coherence is most likely to emerge. These applications, of course, would require the inclusion of more couples in the sample.

Chapter 4

Optimal and robust design for efficient system-wide synchronization in networks of randomly-wired neuron-nodes

4.1 Introduction

This study was motivated in part by a study of network coherence of emotion variables [33] and a book on brain rhythms [4]. In the emotion study, tight connections with strong wiring potentials are found in several subgroups among a collection of emotion variables. In contrast, sparse connections and weak wiring are observed between subgroups. This result immediately leads to the following question: Does this empirical

network structure efficiently lead to system-wide synchronization as a phenomenon of emotion arousal? In addition, the system-wide synchronization of neuron-firing is believed to be closely related to memory reactivation in the animal brain when it is sleeping. However, many memories must be relocated from one part of the hippocampus to the other parts of the brain, and different memories are believed to be stored in different locations within the hippocampus. Thus a system of neurons responsible for one memory needs to be activated as one whole in a very efficient manner. Contemplating representations of the signal transmission in network models of these phenomena has led to the innovative approach presented here.

An artificial neural network (ANN) is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation. The study of ANNs is motivated by their comparability to biological neural networks. They are investigated in order to increase understanding about their biological counterparts, and to use the functional power of biological neural networks to guide the development of modern technology. In particular, technicians seek to adapt the brain's capacity for self-organization, learning, generalization and fault-tolerance. Introductions to ANNs and their applications may be found in [27], [30], [39], [42] and [60]. The use of ANNs to study the dynamics of neurological networks in the brain of a behaving animal is ubiquitous throughout the neuroscience literature, as in [21].

In this chapter, a simple ANN consisting of a fixed number of neuron-like hypothetical nodes (or "neurodes") is considered. A novel approach is proposed for

modeling signal transmission among these nodes as follows: a node becomes activated at a point in time when it receives a signal from any of its nearest neighbors, in a manner similar to the accumulation of action potential in a neuron. And, similar to the discharge of that accumulated action potential, every activated node transmits its signal to all its immediate neighbors at the next time point after activation and instantly becomes deactivated, until reactivated by a neighboring node. In this simulation of signal transmission in biological neural networks, nodes change activation states within a discrete time resolution.

One particular phase of interest pertaining to the network under consideration is called *system-wide synchronization* (SWS), which is specifically defined as that state in which all nodes simultaneously activate all their nearest neighbors and are activated by all their nearest neighbors. This SWS phase is employed to represent the synchronization mentioned in the above two motivating examples. In regard to this phase, the kind of wiring configuration on a deterministic network that produces an efficient SWS phase is studied. Then random networks are considered. Here the wiring between any two nodes is regarded to be governed by an independent Bernoulli random variable. This consideration reflects the fact that two emotion variables or two neurons may not be invariantly wired together at all times when responding to all stimuli. Furthermore, this randomness is a valid approach because the true dynamics underlying this wiring are still not yet understood, especially in neuroscience.

It is also known that not all neurons are wired together equally-well at all times, nor are the emotion variables. Thus it is important to consider the efficiency of achieving

SWS upon a class of random networks subject to a budget of total probabilities for all potential wirings. Furthermore, sometimes a neuron or emotion variable in a network may malfunction. Ideally, the remaining nodes in the network should still perform and achieve the SWS phase. This leads to the necessity of robustness with respect to a malfunctioning node in the network. Hence the following two optimal design problems are studied:

Q1: How can one allocate wiring probabilities subject to a budget constraint in order to maximize the probability of achieving SWS?

Q2: What impact would a robustness criterion have on this wiring structure?

To resolve the above two problems, the Simulated Annealing (SA) algorithm is employed as an optimization technique. No analytic solution is envisioned, but numerical ones are obtained, since the number of potential wirings grows as the square of the number of nodes in the network. For example, 15 nodes will give rise to 105 potential wirings in the network. Here the computations needed for optimizing a 105-dimensional function are overwhelming. However, results obtained in this chapter as the study pass from deterministic networks to random networks, and then to random networks with a robustness criterion, allow a reduction of the computational complexity to a rather manageable level. The energy functions employed in the SA algorithm are developed accordingly.

This chapter is organized as follows. First a mode of signal transmission in a connected deterministic network, in which the nodes change activation states as described above, is described. Then it is shown that one of two distinct phases will

occur: system-wide synchronization (SWS), or subgroup alternation (SGA). Next, both geometric and algebraic criteria are obtained by which one may determine which phase a network will generate. After this, random networks under a constraint on the sum of the edge-probabilities are considered, and search is conducted for an optimal allocation of the edge-probabilities such that SWS networks will occur with high probability, using the simulated annealing algorithm. Then a further goal is imposed: to find an optimal allocation that will frequently generate SWS networks that are also robust against loss of a node. The outcome of this search is presented. The concluding section consists of remarks on potential applications of these findings.

4.2 Deterministic networks and signal transmission

Consider a network $\mathcal{N}$ consisting of n nodes, arbitrarily labeled $1, 2, \ldots, n$, and let the set of nodes be denoted $\{\mathcal{N}\}$. An edge (or wiring) between nodes i and j may be denoted $e_{i,j}$, and the set of all edges in network $\mathcal{N}$ is thus denoted $\mathscr{E}\{\mathcal{N}\}$. Note that the cardinality of $\mathscr{E}\{\mathcal{N}\}$ is at most $n(n-1)/2$. In a neural network, nodes represent neurons and edges represent the connections between them. Regard a network as a system through which a signal travels from node to node by means of available edges. The definitions of all terms and notation used in the foregoing discussion, when not provided, may be found in the literature, e.g., in [2], [50] or [64].

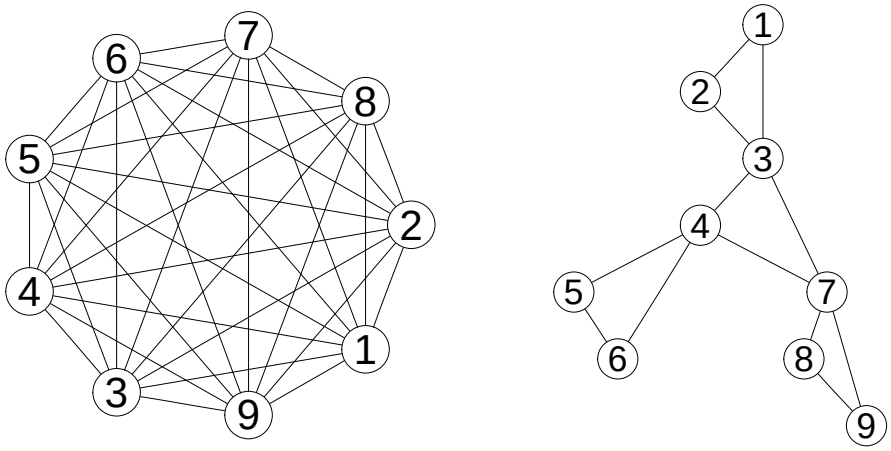
Initially, regard every node in a network $\mathcal{N}$ to be in an "off" state, that is, *deactivated*. Then at some moment, say, at step 0, node i switches to an "on" state, i.e., is *activated*. It may be that a signal was applied to node i from some external source, or

some internal process activated it. This signal will then be transmitted throughout the network in discrete time steps, activating other nodes as it reaches them. However, once the signal is transmitted from any node to its immediate neighbors, that node deactivates until the signal returns at a later step. The only exception occurs if the node receives the signal back from a neighbor during the same step in which it also transmits the signal. Hence, at step 1 each neighbor of node i will be activated, but node i will again become deactivated. At step 2, the neighbors of the neighbors of node i (which include node i itself) will be activated, while the neighbors of node i are deactivated. Assume that the signal then continues to be transmitted indefinitely in this manner at steps $3, 4, \ldots$. This mode of signal transmission, in which nodes switch to an "off" state until turned "on" by neighboring nodes, is intended to represent the behavior of neurons in the brain.

Note that, if $\mathcal{N}$ is a connected network, then its diameter with respect to node i (i.e., the shortest path between node i and the node most distant from it) cannot exceed $n-1$ for each $i = 1, \ldots, n$, so that the overall diameter $D(\mathcal{N})$ of the network may not exceed $n-1$ (a connected network consisting of a string of n consecutive nodes, with $n-1$ edges linking them, has the maximum possible diameter). Clearly then, a signal which originates at any node in a connected network $\mathcal{N}$ will have been propagated at least once to every other node within $n-1$ steps. The interest here focusses on network configurations which result in the simultaneous and sustained activation of all nodes after finitely many steps.

For example, suppose a network consists of nine nodes, as in Figure 4.1: Note that

Figure 4.1: Networks $\mathcal{N}_A$ and $\mathcal{N}_B$



network $\mathcal{N}_A$ is fully connected, while network $\mathcal{N}_B$ is connected, but not fully. Each network can be represented by a symmetric 9×9 connectivity matrix, with zeros on the main diagonal, and with m_{ij} indicating the presence of an edge between nodes i and j, as shown below:

$$M(\mathcal{N}_A) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

$$M(\mathcal{N}_B) = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Signal transmission within a network $\mathcal{N}$ is modeled by a variation of matrix multiplication. Define the *state* of node i at step k, denoted $s_k(i)$, such that $s_k(i) = 1$ if node i is activated at step k, and $s_k(i) = 0$ if node i is deactivated at step k. Let the *network state vector* $\mathbf{v} = (v_1, \ldots, v_n)$ at step k be a vector of length n such that, at step k, $v_i = s_k(i)$ for $i = 1, \ldots, n$. Thus at step 0 the network state vector consists of a 1 in the position corresponding to one of the nodes, and a 0 in every other position. At subsequent steps, the elements of the vector will change to reflect the changes in the state of the network. In this setting, the product $[M\mathbf{v}]$ of the matrix $M = M(\mathcal{N})$ with $\mathbf{v}$ at step k produces the network state vector $\mathbf{v}$ at step $k+1$, where $[M\mathbf{v}]$ is found by computing the usual product of a matrix with a vector, but with the restriction that all non-zero entries in the result are set equal to one. Hence if $\mathbf{v}$ represents the state of the network at step 0, then $[M^k\mathbf{v}]$ represents the state of the network at step k.

To illustrate, the evolution of the network state vector at steps 0 through 5 for both networks $\mathcal{N}_A$ and $\mathcal{N}_B$, if the signal originates at node 1, is displayed:

Network state vector for $\mathcal{N}_A$, steps 0 through 5

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \cdots$$

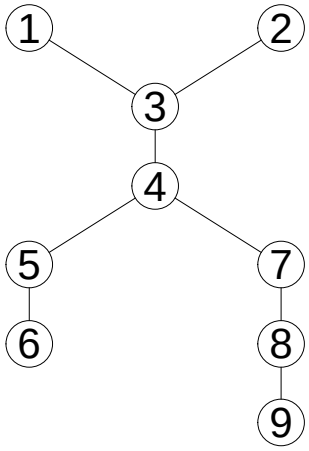
Network state vector for $\mathcal{N}_B$, steps 0 through 5

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \cdots$$

Notice that, by step 2, the network state vector for $\mathcal{N}_A$ consists of a 1 in each position, indicating that all nine nodes are simultaneously activated, and remains in this state thereafter. This phenomenon is described as system-wide synchronization (SWS). Hence network $\mathcal{N}_A$ has achieved SWS after two steps. The same result occurs with the network state vector for $\mathcal{N}_B$, but not until step 3. For either network, if instead the signal begins at a different node, the network still achieves SWS, but not necessarily after the same number of steps, because the signal will eventually reach node 1. In general, a connected network is SWS if and only if there exists some step k at which all of its nodes are concurrently activated. Define the *order* of a SWS network $\mathcal{N}$ *with respect to* node i, denoted $O_i(\mathcal{N})$, as the minimum number of steps from activation of node i at step 0 until SWS occurs. Then the *order* of a SWS network $\mathcal{N}$ having n nodes, denoted $O(\mathcal{N})$, may be defined as the maximum over $i = 1, \ldots, n$ of $O_i(\mathcal{N})$.

One may then inquire whether or not all connected networks are SWS. Consider network $\mathcal{N}_C$ in Figure 4.2, along with the corresponding evolution of the network state vector when node 1 is activated at step 0. Observe that SWS does not occur in this case:

Figure 4.2: Network $\mathcal{N}_C$



Network state vector for $\mathcal{N}_C$

$$\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix} \to \begin{bmatrix}0\\0\\1\\0\\0\\0\\0\\0\\0\end{bmatrix} \to \begin{bmatrix}1\\1\\0\\1\\0\\0\\0\\0\\0\end{bmatrix} \to \begin{bmatrix}0\\0\\1\\0\\1\\0\\1\\0\\0\end{bmatrix} \to \begin{bmatrix}1\\1\\0\\1\\0\\1\\0\\1\\0\end{bmatrix} \to \begin{bmatrix}0\\0\\1\\0\\1\\0\\1\\0\\1\end{bmatrix} \to \begin{bmatrix}1\\1\\0\\1\\0\\1\\0\\1\\0\end{bmatrix} \to \begin{bmatrix}0\\0\\1\\0\\1\\0\\1\\0\\1\end{bmatrix} \to \cdots$$

This time, the network state vector eventually begins to alternate indefinitely between two complementary states, so that SWS never occurs. Instead, the network is partitioned into two subgroups which are activated at alternating time steps. Hence

a network may be classified as a SWS network if SWS occurs after finitely many steps, and as a Subgroup Alternating (SGA) network if the above phenomenon occurs.

4.3 Criteria for System-Wide Synchronization in deterministic networks

Now consider whether there is some geometric feature of a network that determines whether or not SWS occurs. To address this question, first consider the following two lemmas, whose proofs are supplied in the appendix. Here, a node is periodic with period 2 if the node is activated at even-numbered time steps and deactivated at odd-numbered time steps, or vice versa, while a node is periodic with period 1 if it is activated at every time step. Also, $D_i(\mathcal{N})$ denotes the diameter of network $\mathcal{N}$ with respect to node i.

Lemma 1. *Every activated node in a connected network is periodic with period* $p \leq 2$.

Lemma 2. *If an activated node* i *in a connected network* $\mathcal{N}$ *becomes periodic with period* $p = 1$ *at some step* k*, then SWS will occur within* $D_i(\mathcal{N})$ *additional steps.*

Together, these lemmas infer that the key to the SWS of a connected network is that one of the nodes must become periodic with period 1 at some step in the signal transmission process. Note that this occurred in the evolution of the network state vector for $\mathcal{N}_B$ at step 1, since nodes 2 and 3 remained activated in going from the first to the second step. The geometric feature which allowed this to happen is the presence in the structure of the network of a loop consisting of an odd number of

edges. Refer to such a loop, in which there is a closed path consisting of an odd number of edges, as an *odd-length loop*.

Note that in network $\mathcal{N}_B$ there are four such loops of length 3, while in network $\mathcal{N}_A$ there are numerous such loops of lengths 3, 5, 7 or 9. But there are no odd-length loops in network $\mathcal{N}_C$. This leads to the following theorem and corollary, whose proofs are in the appendix:

Theorem 1. *System-wide synchronization occurs in a connected network if and only if its structure includes a loop consisting of an odd number of edges.*

Now, suppose one has a SWS network. One can then determine an upper bound on the order of the network, based on its diameter:

Corollary 1. *A SWS network $\mathcal{N}$ having n nodes has order $O(\mathcal{N}) \leq 2D(\mathcal{N})$, regardless of the node at which the signal originates.*

Hence the most efficient SWS network, in terms of having the smallest order, is one with the smallest diameter possible. For example, if $n-1$ nodes are each connected by a single edge to one hub node, the network will have a diameter of 2. Thus if this same network also includes a loop of length 3, it will be a SWS network of order at most 4. Network $\mathcal{N}_A$ has a diameter of 1, since each node is a neighbor to every other node. Thus $\mathcal{N}_A$ is a SWS network of order 2.

4.4 Subgroup Alternation in deterministic networks

Corollary 1 together with Lemmas 1 and 2 imply that if a connected network $\mathcal{N}$ is not SWS, then after at most $2D(\mathcal{N})$ steps every node will be periodic with period $p = 2$. Moreover, the network will itself have a period of 2 no later than step $2D(\mathcal{N})$, with one subset of nodes activated simultaneously at only the odd-numbered steps, and the remaining subset of nodes activated concurrently at only the even-numbered steps, as noted in the example of network $\mathcal{N}_C$. Of course, which subset corresponds to the odd-numbered steps depends on the choice of initial node. Hence if a network is not SWS, it will begin subgroup alternation (SGA) after finitely many steps.

One may then identify a feature of the matrix $M(\mathcal{N}) = M$ corresponding to a connected network $\mathcal{N}$ that establishes the network as SGA. This feature is the ability to partition M into two sub-matrices M_1 and M_2 such that M_1 consists of $m \geq 1$ columns from M and M_2 consists of the remaining $n - m \geq 1$ columns, and such that each sub-matrix consists of one or more rows of zeros, but in complementary rows.

As an example, consider the matrix corresponding to SGA network $\mathcal{N}_C$:

$$M(\mathcal{N}_C) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Take columns 1, 2, 4, 6 and 8 for M_1 and the remaining columns for M_2:

$$M_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that M_1 has rows of zeros in rows 1, 2, 4, 6 and 8, which correspond to one of the two subsets of nodes which are activated concurrently once the network achieves

SGA. Meanwhile M_2 has rows of zeros only in the other four rows, i.e., in rows 3, 5, 7 and 9, corresponding to the other subset of nodes.

Given this condition, one may simply relabel the nodes of $\mathcal{N}_C$ in such a way that $M(\mathcal{N}_C)$ takes on a block off-diagonal form. This is done by consecutively labeling the nodes corresponding to one of the two subsets, and then labeling the remaining nodes corresponding to the other subset. For instance, if in $\mathcal{N}_C$ one relabels node 3 as node 6, node 4 as node 3, node 5 as node 7, node 6 as node 4, node 7 as node 8, and node 8 as node 5, the matrix for network $\mathcal{N}_C$ becomes

$$M(\mathcal{N}_C) = \left[\begin{array}{ccccc|cccc} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]$$

The lower-left block is matrix M_1 with its zero-rows removed, while the upper-right block is matrix M_2 with its zero-rows removed. If the lower-left block is referred to as submatrix A, then the upper-right block is its transpose A'. In general, the matrix of every SGA network can be represented in this manner. This may be stated in a theorem:

Theorem 2. *A connected network $\mathcal{N}$ consisting of n nodes achieves subgroup alternation if and only if there is some $m \in \{1, \ldots, n-1\}$ and some permutation of the node labels $\{1, \ldots, n\}$ such that its matrix $M = M(\mathcal{N})$ can be partitioned into a block off-diagonal matrix of the form*

$$M = \begin{bmatrix} 0 & A' \\ A & 0 \end{bmatrix},$$

where A is an $(n-m) \times m$ matrix.

The proof is supplied in the appendix.

Now suppose M takes this block off-diagonal form, and suppose λ is an eigenvalue of M with corresponding eigenvector $\mathbf{v} = (v_1, \ldots, v_m, v_{m+1}, \ldots, v_n)'$. Since M is symmetric, λ is real. Then the block off-diagonal form of M and the relation $M\mathbf{v} = \lambda\mathbf{v}$ implies $A'(v_{m+1}, \ldots, v_n)' = \lambda(v_1, \ldots, v_m)'$ and $A(v_1, \ldots, v_m) = \lambda(v_{m+1}, \ldots, v_n)'$. If $\lambda \neq 0$, then $-\lambda$ must also be an eigenvector of M, with corresponding eigenvector $\mathbf{v} = (-v_1, \ldots, -v_m, v_{m+1}, \ldots, v_n)'$, since the relations

$$A'(v_{m+1}, \ldots, v_n)' = -\lambda(-v_1, \ldots, -v_m)' = \lambda(v_1, \ldots, v_m)'$$

and

$$A(-v_1, \ldots, -v_m) = -A(v_1, \ldots, v_m)' = -\lambda(v_{m+1}, \ldots, v_n)'$$

also hold. Hence, if n is even, the set of eigenvalues of M must be representable as $\{\pm\lambda_1, \ldots, \pm\lambda_{n/2}\}$. If n is odd, the requirement that nonzero eigenvalues appear in positive/negative pairs requires that at least one eigenvalue must be zero, so that the sum of the eigenvalues equals the trace of M, which is zero. Thus, if n is odd, the set

of eigenvalues of M must be representable as $\{0, \pm\lambda_1, \ldots, \pm\lambda_{(n-1)/2}\}$. In general, the matrix corresponding to a SGA network must have symmetry of its eigenvalues about zero. Therefore, if a network $\mathcal{N}$ is SWS, the eigenvalues corresponding to its matrix will *not* be symmetric about zero, i.e., there will be at least one nonzero eigenvalue whose additive inverse is not an eigenvalue.

Hence both a geometric criterion (odd-length loops) and an algebraic criterion (asymmetry of eigenvalues) have been obtained by which one may determine whether or not a deterministic network is SWS.

4.5 Optimization of a random network under a budget constraint

Armed with these criteria, this section directs its attention to a random network consisting of n nodes with $N = n(n-1)/2$ distinct potential edges. In a sequence of independent trials, each edge may or may not occur, so that any particular realization of the network may or may not be SWS (or even connected). Moreover, even when a SWS network is realized, it may not be relatively efficient, i.e., the average number of steps required to achieve SWS over the n nodes may be relatively large. Suppose that the probability that edge $e_{i,j}$ occurs between nodes i and j in any trial is some fixed number $p_{i,j} \in [0, 1]$. One may then form a vector $\mathbf{p}$ of length N consisting of

these probabilities, such that

$$\begin{aligned}\mathbf{p} &= (p_{1,2}, p_{1,3}, \ldots, p_{1,n}, p_{2,3}, \ldots, p_{2,n}, \ldots, \\ &\qquad p_{n-2,n-1}, p_{n-2,n}, p_{n-1,n}) \, .\end{aligned}$$

Suppose further that there exists some fixed budget constraint $B \leq N$ such that

$$B = \sum_{1 \leq i < j \leq n} p_{i,j} \, .$$

Then the question is: Are there optimal allocations of probabilities among the N components of $\mathbf{p}$ that conform to the budget constraint and that maximize the probability that a relatively efficient SWS network will be realized in any trial?

Certainly, if B is large enough (at least n), one may assign a probability of 1 to each of n edges chosen such that, when present, the resulting network is SWS and as efficient as possible. Thus only situations where B is relatively small are of interest, so that high probabilities cannot be assigned very liberally. In nature, systems are configured so as to allocate limited resources in an optimal manner. By requiring a low budget, this tendency is modeled.

One exhaustive method would require an examination of each of the 2^N possible networks to identify which of these are SWS and have a desired level of efficiency. Then one would need to search for those vectors in $[0,1]^N$ that conform to the budget constraint while maximizing the probability of producing one of these identified networks in any trial. This is clearly impractical.

As an alternative, first substitute $[0,1]^N$ with the lattice $\{0.0, 0.1, \ldots, 0.9, 1.0\}^N \subset [0,1]^N$, and require B to be a positive multiple of 0.1. If optimal vectors are found

in the lattice, it may be assumed that the optimal vectors in $[0,1]^N$ lie nearby. Thus consider now the space consisting of only those vectors $\mathbf{p}$ whose components $p_{i,j}$ lie in the set $\{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$ and sum to B. Nevertheless, an exhaustive exploration of this search space remains intractable. Hence it is necessary to employ a tool like the simulated annealing (SA) search algorithm, as developed independently by Kirkpatrick, et al. [36], and by Černý [5].

In the SA algorithm, one starts at some initial vector in the search space. Then a neighboring vector is selected, which is defined to be a vector whose components match those of the initial vector in all but two of the N positions, and which differ from the initial vector in those two positions by ± 0.1. For instance, if $\mathbf{p} = (0.2, 0.1, 0.3, 0.4, 0.0, 0.8)$, the vector $(0.2, 0.1, 0.2, 0.4, 0.1, 0.8)$ would be a neighbor. If the neighboring vector is more likely to produce an efficient SWS network, the search moves to that vector. Otherwise, the search may still move to it with a certain probability which gradually diminishes from one to zero as the algorithm progresses. Then, at the next iteration, a neighbor is selected and the decision to move to that neighbor is repeated.

To implement the SA algorithm, a positive real-valued *energy function* $E(\mathbf{p})$ that estimates the "energy" of each point $\mathbf{p}$ in the search space is required. Here the energy of $\mathbf{p}$ represents its likelihood to generate networks which are *not* SWS, and networks which are SWS but relatively inefficient. Hence the energy function is constructed such that it decreases toward zero as optimal solutions are encountered.

The energy function implemented here consists of I iterations. At each iteration

the function uses the probabilities $p_{i,j}$ to create a realization of the random network. The function then determines whether the realized network is SWS. If the network is SWS, the energy function determines the geometric mean order of the network starting at each of the n nodes, as a measure of its efficiency. Upon completing I iterations, the energy function computes the proportion φ of the I realized networks which were *not* SWS, and, among those realized networks which were SWS, the average ψ of the geometric mean order. Note that the intention is to minimize both φ and ψ. Then a weighted average $W = a\varphi + b\psi$ is formed, with a and b chosen such that φ dominates until it becomes very small, at which point ψ begins to have greater influence. To control the rate at which the energy function decreases as W decreases, W is passed to the sine function, ensuring that $0 \leq W \leq \pi/2$. In this implementation, the energy function is

$$E(\mathbf{p}) = \sin(0.5\varphi + 0.001\psi).$$

As anyone would surmise, this energy function is highly variable, no matter what value is chosen for I ($I = 1,000$ is used here), unless $\mathbf{p}$ consists almost exclusively of ones and zeros. To reduce the variability, the average function value over 10 applications to any vector $\mathbf{p}$ is computed. Nevertheless, the variability remains. While this does not prove to be a critical problem when n is small (say, 4 or 5), it attenuates the success of the algorithm when $n = 10$ or $n = 15$.

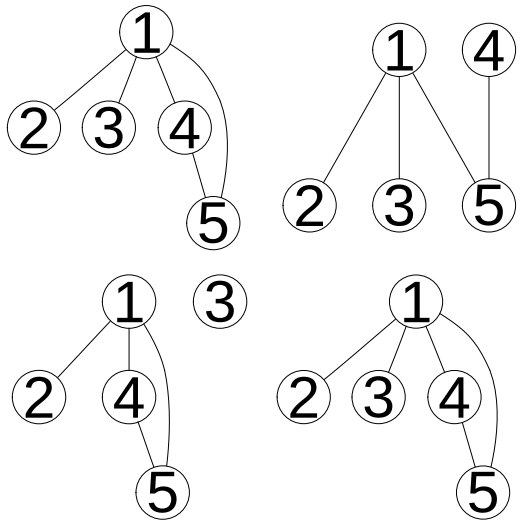
Throughout the SA algorithm, a record is maintained of the best vector the search has encountered, i.e., the vector whose computed energy is lowest. Once the algorithm terminates, the best vector is declared to be optimal in the sense of having the lowest

energy in the search space, and thus the greatest likelihood of producing efficient SWS networks among those having the same budget constraint. In practice, the vector identified by the algorithm may not be optimal, but it is usually quite good. Moreover, it is not unique, since permuting the labels of the nodes would change the labels of the potential edges among them, producing a different vector of probabilities.

The SA algorithm also requires a temperature function which decreases monotonically from one to zero as the algorithm progresses from its first iteration to its last. In this study, the function $T(z) = 1 - \mathrm{e}^{5(z-1)}$ is used, where z is the proportion of iterations completed. The algorithm relies on the temperature at any iteration to determine the probability of moving from a state whose energy is lower to a neighboring state whose energy is higher, as mentioned above. This probability decreases as the temperature decreases, so that the algorithm gradually narrows its focus to one convex region of the search space.

For example, consider a random network consisting of five nodes and thus 10 potential edges. Suppose $B = 4$, which is low enough to ensure that a SWS network cannot be guaranteed in any realization. If the search starts with an initial vector consisting of a probability of 0.4 for each potential edge (with corresponding energy 0.3536), the SA algorithm returns as the optimal result $(0.8, 0.8, 0.8, 0.8, 0.0, 0.0, 0.0, 0.0, 0.0, 0.8)$, whose energy is found to be 0.3301. Four realizations of the random network with this optimal allocation of probabilities to the 10 potential edges are displayed in Figure 4.3. Note that only three of the networks are connected, and of these only two are SWS (having a loop of length 3): The two SWS networks are identical, each of di-

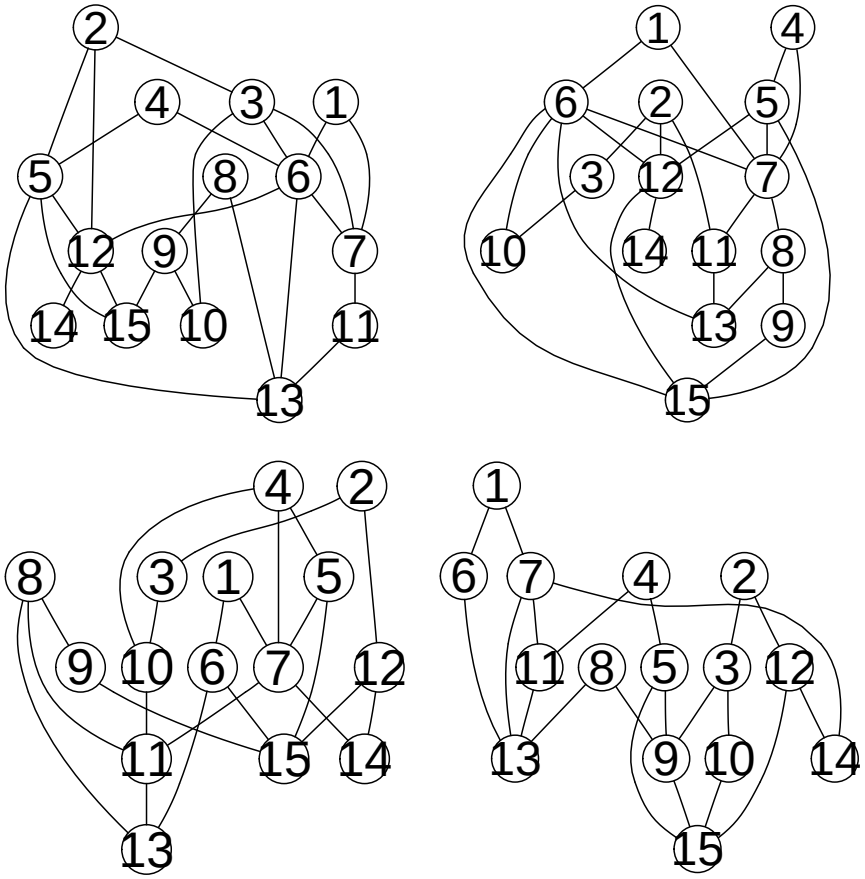
Figure 4.3: Four realizations of an optimized random five-node network



ameter 2. Given the low budget, this is the best allocation of probabilities the search can find.

As a second example, consider a random network consisting of 15 nodes and 105 potential edges, with a relatively high budget of $B = 21$. If the search starts with an initial vector consisting of a probability of 0.2 for each potential edge (with corresponding energy 0.2510), the SA algorithm returns an optimal result whose energy is computed to be 0.0077, which is remarkably low. Four realizations of this random network are displayed in Figure 4.4. Note that all four networks are SWS, as loops of length 3 may be easily identified among each set of edges, and odd-length loops of higher dimension are also evident. Hence the SA algorithm proves quite effective in finding an allocation of probabilities among potential edges that maximizes the likelihood that any realization of a random network will be SWS.

Figure 4.4: Four realizations of an optimized random fifteen-node network



4.6 Robust networks

A desirable feature of a SWS network is that it be *robust*. One way of defining robustness is in terms of the loss of wirings between nodes. An alternative definition of robustness involves the preservation of a SWS network if one node "malfunctions." This study focusses on networks which are robust under the latter definition, i.e., robust against the loss of a node. Observe that network $\mathcal{N}_A$ is robust against loss of a node, whereas if nodes 3, 4 or 7 were to be removed from network $\mathcal{N}_B$, the remaining sub-network would not even be connected.

The energy function $E(\mathbf{p})$ implemented in the SA algorithm may be modified so that it can also check each SWS network $\mathcal{N}$ among the I realized networks for robustness. This is achieved by successively removing the ith row and column from $M(\mathcal{N})$, $i = 1, \ldots, n$, and determining whether the resulting submatrix corresponds to a SWS network. This obviously slows down the algorithm considerably. Let χ denote the proportion of SWS networks which are *not* robust among I realized networks. This time the weighted average $W = a\phi + b\psi + c\chi$ is computed, with the constants a, b and c chosen to balance the influences of system-wide synchronization, robustness, and efficiency as the algorithm progresses. In this implementation, the energy function used is

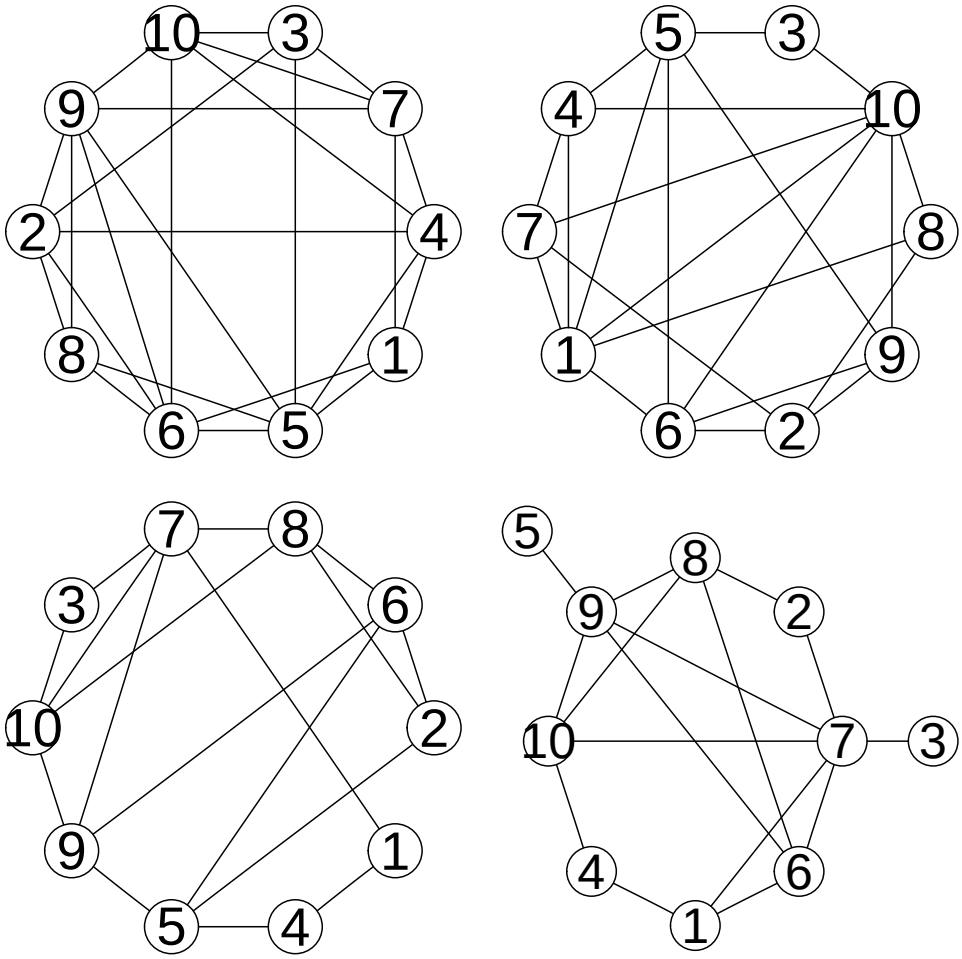
$$E(\mathbf{p}) = \sin(0.5\phi + 0.001\psi + 0.1\chi).$$

To make the problem more realistic, and hence more interesting, the probabilities $p_{i,j}$ comprising probability vector $\mathbf{p}$ are further restricted to the set $\{0.0, 0.1, \ldots, 0.9\}$, so that no edge in the network can ever be guaranteed. This reduces the search space slightly.

The SA algorithm is implemented with these modifications, using a network of 10 nodes and a budget of $B = 22.5$, beginning with an equidistribution of $p_{i,j} = 0.5$ among the 45 potential edges, yielding an energy of 0.100. The optimal probability vector returned by the procedure has a computed energy of 0.046, which is quite good given the additional restrictions. Then the search is repeated using a 15-node network with a budget of 31.5, starting with an equidistribution of $p_{i,j} = 0.3$ among the 105 potential edges, and an energy of 0.292. The SA algorithm returns an optimal

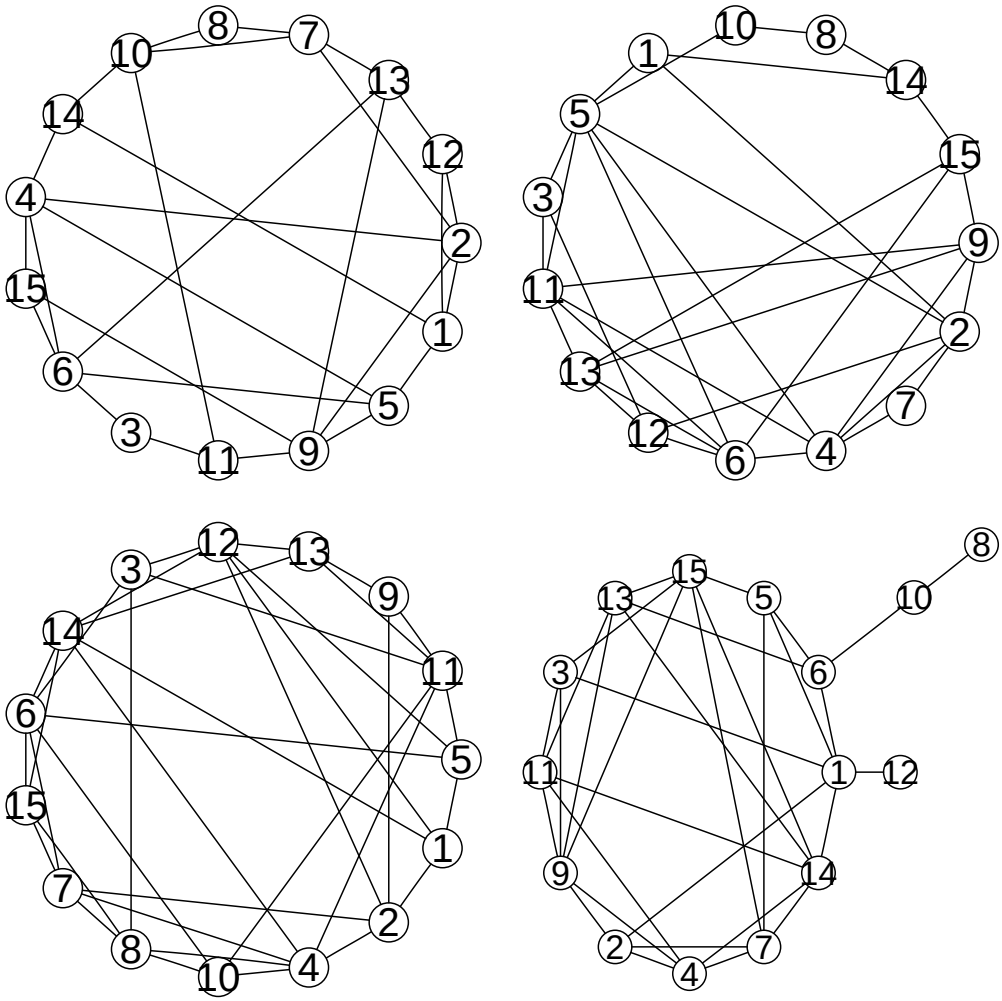
result whose energy is 0.180. Four realizations of random networks based on this optimal result are displayed in Figures 4.5 and 4.6. In Figure 4.5, the lower-right

Figure 4.5: Four realizations of a robust optimized random ten-node network



network is SWS, but not robust, since the loss of nodes 7 or 9 would render the remaining network unconnected. But the other three networks are both SWS and robust. Coincidentally, the lower-right network in Figure 4.6 is also SWS, but not robust, while the other three are both. One may observe that in the robust networks, odd-length loops (particularly, triangles) occur at multiple distinct locations in the

Figure 4.6: Four realizations of a robust optimized random fifteen-node network



network. Hence optimal robust configurations of random networks are those which connect smaller subnetworks which are themselves SWS networks. This of course requires a sufficient budget to permit such a configuration.

4.7 Conclusion

System-wide synchronization has been defined in networks based on the premise that, once a signal leaves a node, the node becomes inactive, and remains so until the signal returns to it, as is known of neurons. Working with this definition, this study discovers that a system which can be modeled by such a network either achieves a state of system-wide synchronization after an initial start-up period, or wavers indefinitely between two complementary states. This outcome ultimately depends only on the geometry of the network. Hence the functionality of such a system would require attention to whether or not its structure contains loops with an odd number of edges.

In a system where the connectivity among nodes is random, perhaps dependent on the correlations between them, one may further determine an optimal allocation of probabilities among the potential edges, subject to certain budget constraints, such that the functionality of the system is most likely. In application to such a system which is not functioning optimally, one might determine how to reallocate resources in order to improve performance. Moreover, to promote robustness of such a system, one should ensure that its subsystems are independently designed to function optimally.

The results obtained in this chapter may be summarized as follows:

S0 [Deterministic network:] A SWS phase can only be achieved in a connected network containing at least one substructure consisting of a closed loop comprised of an odd-number of edges.

S1 [Random network:] Any wiring probability allocation scheme achieves a high

potential of producing an efficient SWS phase when it frequently gives rise to several heavily-connected hubs.

S2 [Random network with robustness:] The robustness constraint requires a higher budget, and at the same time replaces the potential hubs with many well-scattered potential triangular substructures.

Result [S0] regarding deterministic networks affords a huge reduction in computations for Result [S1], which in turn partly anticipates Result [S2]. However the appearance of abundant triangular substructures in a robust random network is somewhat surprising.

Immediate implications of these results on the study of emotion and memory reactivation are posited as follows:

I1 : An emotion arousal can be effectively triggered when subgroups of emotion variables, such as behavioral, experiential and physiological variables, are well-wired, even though the subgroups themselves are sparsely connected with redundant wiring.

I2 : Even in the absence of an inhibiting mechanism, the feed-forward and feedback mechanisms of signal transmission are sufficient to efficiently and robustly generate SWS phases among a designated group of neurons with strong local connections and sparse global wiring.

4.8 Proofs

Proof of Lemma 1. Assume node i is activated at step k. At step $k+1$, each neighbor of i is activated, while i may either be deactivated or reactivated by a neighbor. At step $k+2$, each neighbor of node i reactivates node i. This cycle then repeats indefinitely, so that node i is never deactivated for more than one consecutive step. Hence node i is periodic with period at most two. □

Proof of Lemma 2. Assume activated node i is periodic with period $p = 1$ at step k. At step $k+1$, every neighbor of node i is activated, and node i remains activated since node i has period 1. Now the neighbors of node i have period 1 since their neighbor, node i, is activated at every step. At step $k+2$, every neighbor of the neighbors of node i is activated, node i remains activated, and every neighbor of node i remains activated. Now the neighbors of the neighbors of node i have period 1. By step $k + D_i(\mathcal{N})$, this effect will have been transferred to the nodes which are at the greatest distance from node i, so that every node in $\mathcal{N}$ is activated. Hence all nodes of $\mathcal{N}$ are simultaneously activated within $D_i(\mathcal{N})$ additional steps. □

Proof of Theorem 1. It is evident from the proof of Lemma 2 that an activated node i becomes periodic with period 1 if and only if the signal reaches both i and some neighbor of i in the same step. This requires the presence of a closed path in the network structure comprised of an odd number of nodes, i.e., an odd-length loop. Then, and only then, the signal will follow two branches which reach two neighboring nodes at the same step. By Lemma 2, such a network is SWS. □

Proof of Corollary. Assume $\mathcal{N}$ is a SWS network with n nodes. If node i is activated at step 0, the signal will propagate to every node of $\mathcal{N}$ within $D_i(\mathcal{N}) \leq D(\mathcal{N})$ steps. Since $\mathcal{N}$ is SWS, $\mathcal{N}$ must have at least one odd-length loop in its structure. Thus the signal must enter the loop and reach two neighboring nodes h and j at the same step while propagating throughout the network, so that h and j each become periodic with period 1. By Lemma 2, $\mathcal{N}$ will then achieve simultaneous activation of all nodes within $D_h(\mathcal{N}) \leq D(\mathcal{N})$ steps. Therefore SWS occurs within $D_i(\mathcal{N}) + D_h(\mathcal{N}) \leq 2D(\mathcal{N})$ steps. Hence $O_i(\mathcal{N}) \leq 2D(\mathcal{N})$. Since i is arbitrary, $O(\mathcal{N}) \leq 2D(\mathcal{N})$. □

Proof of Theorem 2. Assume $\mathcal{N}$ is a connected network which is SGA, with corresponding $n \times n$ matrix M. Then after finitely many steps every node of $\mathcal{N}$ is periodic with period 2. This implies that the network state vector $\mathbf{v} = (v_1, \ldots, v_n)$ eventually begins to alternate indefinitely between two states, call them α and β, in which each v_i alternates between one and zero for $i = 1, \ldots, n$ (at each step, $\mathbf{v}$ consists of at least one zero component and at least one nonzero component). Since each successive state results from the product $[M\mathbf{v}]$, the inner product of the ith row of M with $\mathbf{v}$ must be zero whenever $v_i = 1$ and be nonzero otherwise. Hence the ith row of M must have zeros in those columns corresponding to the positions of the nonzero components in $\mathbf{v}$ in one of the two alternating states, and a one in at least one of the remaining columns. One may thus let M_1 be the matrix consisting of the $m \geq 1$ columns of M which have zeros in the rows corresponding to all the positions of the nonzero components in $\mathbf{v}$ when it is in state α, and let M_2 be the matrix consisting of the remaining $n - m \geq 1$ columns of M which have zeros in the rows corresponding

to all the positions of the nonzero components in $\mathbf{v}$ when it is in state β. Then M_1 has one or more rows consisting only of zeros, while each corresponding row of M_2 contains at least one nonzero entry. Likewise, M_2 has one or more rows consisting only of zeros, while each corresponding row of M_1 contains at least one nonzero entry. By relabeling the nodes so that those which are activated in state α are consecutively numbered $1, \ldots, m$, and those which are activated in state β are consecutively numbered $m+1, \ldots n$, the matrix M will thus take the form

$$M = \begin{bmatrix} 0 & A' \\ A & 0 \end{bmatrix},$$

where A is an $(n-m) \times m$ matrix.

Moreover, if M may be partitioned in this way, then after finitely many steps every node of $\mathcal{N}$ must be periodic with period 2. Hence $\mathcal{N}$ is SGA. $\square$

Bibliography

[1] C. Anderson, D. Keltner, and O. P. John, *Emotional convergence over people over time*, Journal of Personality and Social Psychology **84** (2003), 1054–1068.

[2] N. Boccara, *Modeling Complex Systems*, Springer-Verlag, 2004.

[3] S. M. Boker, J. L. Rotondo, and K. King, *Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series*, Psychological Methods **7** (2002), 338–355.

[4] G. Buszáki, *Rhythms of the Brain*, Oxford University Press, 2006.

[5] V. Černý, *A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm*, Journal of Optimization Theory and Applications **45** (1985), 41–51.

[6] C. L. Cheng and J. Riu, *On estimating linear relationships when both variables are subject to heteroscedastic measurement errors*, Technometrics **48** (2006), 511–519.

[7] J. A. Coan, *Toward a neuroscience of attachment*, Handbook of attachment: Theory, research, and clinical applications (NY) (J. Cassidy and P. R. Shaver, eds.), Guildford Publications, 2008, pp. 241–265.

[8] J. A. Coan, H. S. Schaefer, and R. J. Davidson, *Lending a hand: Social regulation of the neural response to threat*, Psychological Science **17** (2006), 1032–1039.

[9] P. M. Cole, S. E. Martin, and T. A. Dennis, *Emotion regulation as a scientific construct: Methodological challenges and directions for child development research*, Child Development **75** (2004), 317–333.

[10] C. P. Cowan and P. A. Cowan, *When partners become parents: The big life change for couples*, Lawrence Erlbaum Associates, Mahwah, NJ.

[11] O. Davidov, *Estimating the slope in measurement error models—a different perspective*, Statistics & Probability Letters **71** (2005), 215–223.

[12] O. Davidov and A. Goldenshluger, *Fitting a line segment to noisy data*, Journal of Statistical Planning and Inference **119** (2004), 191–206.

[13] A. P. Dempster, N. M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statististical Society B **39** (1977), 1–38.

[14] A. K. Engel and W. Singer, *Temporal binding and the neural correlates of sensory awareness*, Trends in cognitive sciences **5** (2001), 16–25.

[15] R. F. Engle and C. W. J. Granger, *Co-integration and error correction: Representation, estimation, and testing*, Econometrica **55** (1987), 251–276.

[16] R. Feldman, *Parent-infant synchrony: Biological foundations and developmental outcomes*, Current Directions in Psychological Science **16** (2007), 340–345.

[17] E. Ferrer and J. R. Nesselroade, *Modeling affective processes in dyadic relations via dynamic factor analysis*, Emotion **3** (2003), 344–360.

[18] E. Ferrer and K. F. Widaman, *Dynamic factor analysis of dyadic affective processes with inter-group differences*, Modeling dyadic and interdependent data in the developmental and behavioral sciences (Hillsdale, NJ) (N. A. Card, J. P. Selig, and T. D. Little, eds.), Psychology Press, 2008, pp. 107–137.

[19] C. Fraley and A. E. Raftery, *How many clusters? Which clustering method? Answers via model-based cluster analysis*, The Computer Journal **41** (1998), no. 8, 578–588.

[20] P. Fries, *A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence*, Trends in Cognitive Sciences **9** (2005), 474–480.

[21] S. Fujisawa, A. Amarasingham, M. Harrison, and G. Buzsáki, *Behavior-dependent short-term assembly dynamics in the medial prefontal cortex*, 2008.

[22] W. Gersh and B. R. Tharp, *Spectral regression—amount of information analysis of seizures in humans*, Quantitative Analytic Studies in Epilepsy (NY) (P. Kellaway and I. Petersén, eds.), Raven Press, 1976, pp. 509–532.

[23] J. Gottman, C. Swanson, and K. Swanson, *A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction*, Personality and Social Psychology Review **6** (2002), 326–340.

[24] J. M. Gottman, *Time-series analysis applied to physiological data*, Principles of psychophysiology: Physical, social, and inferential elements (NY) (J. T. Cacioppo and L. G. Tassinary, eds.), Cambridge University Press, 1990, pp. 754–774.

[25] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica **37** (1969), no. 3, 424–438.

[26] S. J. Guastello, D. Pincus, and P. R. Gunderson, *Electrodermal arousal between participants in a conversation: Nonlinear dynamics and linkage effects*, Nonlinear Dynamics, Psychology, and Life Sciences **10** (2006), 365–399.

[27] K. Gurney, *An Introduction to Neural Networks*, London, Routledge, 1997.

[28] J. A. Hartigan and M. A. Wong, *A K-means clustering algorithm*, Applied Statistics **28** (1979), 100–108.

[29] E. Hatfield, J. T. Cacioppo, and R. L. Rapson, *Emotional contagion: Cambridge studies in emotion and social interaction*, Cambridge University Press, Cambridge, UK, 1994.

[30] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.

[31] M. A. Hofer, *Relationships as regulators: A psychobiological perspective on bereavement*, Psychosomatic Medicine **46** (1984), 183–197.

[32] ———, *Hidden regulators in attachment, separation, and loss*, The development of emotion regulation: Biological and behavioral considerations (N.A. Fox, ed.), vol. 59, Monographs of the society for research in child development, 1994, pp. 250–283.

[33] F. Hsieh, E. Ferrer, S. Chen, I.B. Mauss, and J.J. Gross, *A small-world network approach for evaluating coherence in multivariate systems: an application to psycho-physiological emotion data*, 2008.

[34] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, *The Empirical Mode Decomposition and Hilbert Spectrum for nonlinear and nonstationary time series analysis*, Proceedings of the Royal Society A **454** (1998), 903–995.

[35] D. Kim and H.-S. Oh, *EMD: A package for Empirical Mode Decomposition and Hilbert Spectrum*, The R Journal **1/1** (2009), 40–46.

[36] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, *Optimization by Simulated Annealing*, Science **220 (4598)** (1983), 671–680.

[37] M. A. Kramer, U. T. Eden, S. S. Cash, and E.D. Kolaczyk, *Network inference with confidence from multivariate time series*, Physical Review E **79** (2009), 061916.

[38] M. A. Kramer, E. D. Kolaczyk, and H. E. Kirsch, *Emergent network topolgy at seizure onset in humans*, Epilsepsy Research **79** (2008), 173–186.

[39] D. Kriesel, *A Brief Introduction to Neural Networks*, 2007.

[40] S. B. Kulathinal, K. Kuulasmaa, and D. Gasbarra, *Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations*, Statistics in Medicine **21** (2002), 1089–1101.

[41] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F.J. Varela, *Measuring phase synchrony in brain signals*, Human Brain Mapping **8** (1999), no. 4, 194–208.

[42] J. Lawrence, *Introduction to Neural Networks*, California Scientific Software Press, 1994.

[43] R. W. Levenson and J. M. Gottman, *Marital interaction: Physiological linkage and affective exchange*, Journal of Personality and Social Psychology **45** (1983), 587–597.

[44] ———, *Physiological and affective predictors of change in relationship satisfaction*, Journal of Personality and Social Psychology **49** (1985), 85–94.

[45] Mathworks, *MATLAB*, Natick, MA, 1984.

[46] I. B. Mauss, R. W. Levenson, L. McCarter, F. H. Wilhelm, and J. J. Gross, *The tie that binds? Coherence among emotional experience, behavior, and autonomic physiology*, Emotion **5** (2005), 175–190.

[47] R. E. Mirollo and S. H. Strogatz, *Synchronization of pulse-coupled biological oscillators*, SIAM Journal on Applied Mathematics **50** (1990), 1645–1662.

[48] Mitra Lab, *Chronux Analysis Software*.

[49] S. M. Montgomery and G. Buzsáki, *Gamma oscillations dynamically couple hippocampal CA3 and CA1 regions during memory task performance*, Proceedings of the National Academy of Sciences **104** (2007), 14495–14500.

[50] M.E.J. Newman, A.-L. Barabási, and D.J. Watts, *The Structure and Dynamics of Networks*, Princeton University Press, 2006.

[51] I. Olkin and R. Liu, *A bivariate beta distribution*, Statistics & Probability Letters **62** (2003), 407–412.

[52] A. G. Patriota, H. Bolfarine, and M. de Castro, *A heteroscedastic structural errors-in-variables model with equation error*, Statistical Methodology **6** (2009), 408–423.

[53] A. Pikovsky and M. Rosenblum, *Synchronization: A universal concept in nonlinear sciences*, Cambridge University Press, Cambridge, UK, 2001.

[54] M. Qu, Y. Zhang, J. G. Webster, and W. J. Tompkins, *Motion artifact from spot and band electrodes during impedance cardiography*, IEEE Transactions on Biomedical Engineering **BME-33** (1986), 1029–1036.

[55] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, *Performance of different synchronization measures in real data: A case study on electroencephalographic signals*, Physical Review E **65** (2002), 041903–14.

[56] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009.

[57] J. C. Rajapaske and J. Zhou, *Learning effective brain connectivity with dynamic bayesian networks*, NeuroImage **37** (2007), 749–760.

[58] D. A. Sbarra and C. Hazan, *Coregulation, dysregulation, self-regulation: An integrative analysis and empirical agenda for understanding adult attachment, separation, loss, and recovery*, Personality and Social Psychology Bulletin **12** (2008), 141–167.

[59] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications, With R Examples*, 2nd ed., Springer, 2006.

[60] M. Smith, *Neural Networks for Statistical Modeling*, Van Nostrand Reinhold, 1993.

[61] H. Song and E. Ferrer, *State-space modeling of dynamic psychological processes via the Kalman smoother algorithm: Rationale, finite sample properties, and applications*, Structural Equation Modeling **16** (2009), 338–363.

[62] P. Tass, M. G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler, and H. J. Freund, *Detection of $n : m$ phase locking from noisy data:*

Application to magnetoencephalography., Physical Review Letters **81** (1998), no. 15, 3291–3294.

[63] A. Thomson and N. Bolger, *Emotional transmission in couples under stress*, Journal of Marriage & the Family **61** (1999), 38–48.

[64] D.J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, 1999.

[65] C. Weishenbush, M. Nishioka, A. Ishikawa, and Y. Arakawa, *Observation of the coupled excitation-photon mode splitting in a semiconductor quantum microactivity*, Physical Review Letters **69** (1992), no. 23, 3314.