MARKOV MODELS FOR LONGITUDINAL COURSE OF YOUTH BIPOLAR DISORDER

by

Adriana Lopez

B.S. in Statistics, National University of Colombia, 2000

M.S. in Mathematics (Statistics), University of Puerto Rico, 2002

M.A. in Applied Statistics, University of Pittsburgh, 2004

Submitted to the Graduate Faculty of
the Department of Statistics in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF STATISTICS

This dissertation was presented

by

Adriana Lopez

It was defended on

March 28, 2008

and approved by

Satish Iyengar

Leon Gleser

Henry Block

Boris Birmaher

Dissertation Advisors: Satish Iyengar,

Leon Gleser

MARKOV MODELS FOR LONGITUDINAL COURSE OF YOUTH BIPOLAR DISORDER

Adriana Lopez, PhD

University of Pittsburgh, 2008

In this dissertation, mixture of first order Markov chains and Hidden Markov models were used to model variable length sequences in order to find longitudinal patterns. Data from the Course and Outcome of Bipolar Youth (COBY) study was used to estimate these models. A mixture of four first order Markov chains found patterns of movers and stayers. Cluster 4 is the stayers. Cluster 3 are movers among the depression, well and submania states. Cluster 2 are movers that tend to stay in the well state. Cluster 1 are movers that tend to go to the submania/subdepression state. On the other hand, a hidden Markov model with ten hidden states justifies the use of a scale with syndromal, subsyndromal and asymptomatic episodes defined by psychiatrists. The inclusion of covariates in hidden Markov models showed that: males move more than females, children move more than teenagers, and patients who live in another situation move more than patients who live with both natural parents. For bipolar diagnosis, BPII and BPNOS patients show similar transition patterns. Age of bipolar onset sheds light on the stability of patients with a childhood and an early adolescence onset. Thus, the possibility of an early diagnosis of the disorder would consequently lead to provide appropriate treatment, and that would lessen the impairment of bipolar youth. Socio-economic status showed patients with low socio-economic status staying more weeks with subsyndromal submanic and mixed episodes, and less weeks with subsyndromal depression and asymptomatic episodes. Quite the opposite behavior observed for their counterparts in with high socio-economic status. This is the first research using these two Markov models to analyze the longitudinal course of bipolar disorder in children and adolescents. No previous study has modeled the longitudinal course of bipolar disorder using Markov models that estimate the transitions among the different episodes of bipolar disorder. Furthermore, no previous study has modeled the effects of covariates consistently with the longitudinal nature of the disease.

Acknowledgements

I would like to thank my advisors, Satish Iyengar and Leon Gleser for their guidance and support during my years as a graduate student. Thank you for the brain-storming meetings from the beginning of my research and for all your comments throughout the writing process of this dissertation. And thank you to all members in my committee for the time they invested in reading, providing me with important feedback to make this document more complete and more readable.

I also would like to thank you all my professors at the University of Pittsburgh. For those in the statistics department, thank you for strengthening my statistical background through your lectures and for improving my skills as teacher and researcher by sharing your academic experiences in seminars and meetings. For those outside the statistics department thank you for providing me with knowledge in your own fields of study, your classes broadened my inter-disciplinary formation.

To Boris Birmaher, David Axelson and their group of collaborators in the Course and Outcome of Bipolar Youth study, thank you for providing me with such an interesting and important data set. I greatly appreciate the time we spent discussing COBY, its descriptive statistics and the interpretation of the results from the models.

Irene and Christos, my dear Cypriot friends, thank you for mingling together during the first two years of my graduate studies at Pitt. Those gatherings made adjustment to grad school in Pittsburgh easier. Thank you for sharing your culture and traditions and by inviting me to experience it by myself. Thank you to you and to Melissa, Lulu, Ana Maria, Ghideon, Scott, Jim, Kaleab, Christina and Mihaela for the chats in the office, in the labs or department's corridors.

Mary, Connie and Kim, thank you for being so attentive and diligent in sending reminders

with all the deadlines for the university procedures required for graduate students. Thank you for all the errands you run on my behalf during my life as graduate student at Pitt.

To all the Colombians I met in Pittsburgh, especially those close friends I made while living there, I want to thank you all for the good times we had, for helping me get through socially while I was doing this research.

Even though I have not have the time to socialize much in my new home since August 2007—Doha, Qatar—, I would like to thank the few friends I have met. Thank you for your cooperation in helping me adjust to Doha while I was finishing this work, you made the transition smoother.

Thank you to everyone I met throughout my graduate life, the interactions with all of you enhanced a more integral individual.

From the bottom of my heart, I thank my parents and family for their support from the distance, your prayers kept me going.

Alex, thank you for being beside me since the beginning of my graduate studies, for all your contributions while this thesis was taking shape and for every minute you put to make this happen.

Salomé, my precious girl, thank you for giving me the sweet touch to everyday of my life. Your smile was and is the best reward at the end of day.

To God,

To Siervo and Tili,

To Alex Leonardo and Salomé

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Real-world processes generally produce observable outputs which can be discrete or continuous, stationary or nonstationary, and pure or corrupted (Rabiner, 1989). The real-world process motivating this research is the follow-up of children and adolescents with bipolar disorder from the Course and Outcome of Bipolar Youth (COBY) study. COBY is a large longitudinal prospective study of 413 children and adolescents diagnosed with bipolar spectrum disorders (Bipolar I, Bipolar II or Bipolar Not Otherwise Specified) interviewed every 6 months. The follow-up period of these participants range from 6 months to 6 and a half years. In the interview, DSM-IV-TR (2000) criteria information is gathered for each week of the follow-up period, and then later translated into weekly ratings using the Psychiatric Status Rating (PSR). The PSR scale is ordinal with values 1 through 6, with 1 representing an asymptomatic episode and 6 representing a severe episode. Depression, mania and hypomania are three of the diagnoses measured in COBY using the PSR. The scores for depression, mania and hypomania are combined in a single scale with twelve categories as shown in Section 2.3.2. This 12 PSR scale is used to study the longitudinal course of youth bipolar disorder.

Figure 1. shows the longitudinal course of four patients chosen at random from the 413 COBY participants with complete intake and follow-up data until January 2008. Figure 1. pictures the variety and complexity of longitudinal profiles encountered in COBY patients. Such complexity poses a challenge in finding a model that best describes patterns observed in the longitudinal course of bipolar youth.

To date, the longitudinal course and outcome of bipolar disorder has been analyzed using descriptive statistics, logistic regression and survival analysis (Marneros and Brieger, 2002). Hennen (2003) reviews statistical methods for the analysis of longitudinal bipolar clinical

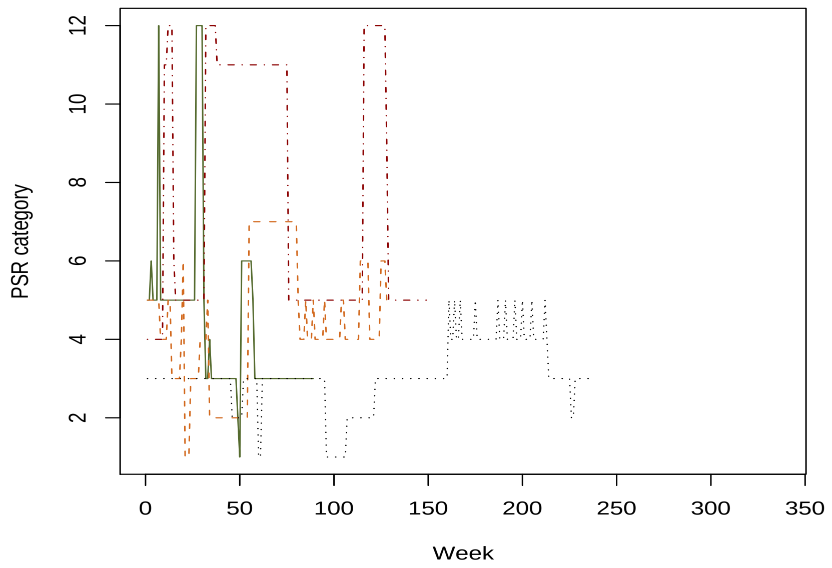


Figure 1: Profile plots of a sample of four patients

trials, recommending consideration of summary measures, random effects or GEE regression modeling and survival analysis. In a more recent and different venue, Pincus (2006) suggests what kind of psychiatric serial data can be analyzed with approximate entropy (ApEn), and Glenn et al. (2006) and Rao et al. (2006) present applications of ApEn to mood data from patients with bipolar disorder.

For better understanding of bipolar disorder in youth, it is important to construct models that help elucidate the dynamic process that these patients undergo. How does one construct models of complex systems to understand and analyze their dynamic behaviors? For processes like the one in COBY, models of dynamic behaviors are often best expressed in terms of a sequence of events or phenomena that occur over time. This is typically called a state-based modeling approach. This modeling approach defines the set of valid states of a dynamic process and describes the system dynamics in terms of stochastic transitions among these states. Markov models are one example of such models.

1.1 CONTRIBUTIONS

With the purpose of finding longitudinal course patterns, this dissertation studied two approaches to describe the dynamic behavior of bipolar youth: (i) clustering based on finite mixtures of first-order Markov chain models, and (ii) hidden Markov models.

This is the first use of these two Markov models to analyze the longitudinal course of bipolar disorder in children and adolescents. No previous study has modeled the longitudinal course of bipolar disorder using Markov models that estimate the transitions among the different episodes of bipolar disorder. Furthermore, no previous study has modeled the effects of covariates consistently with the longitudinal nature of the disease process.

A brief description of the two Markov approaches studied in this dissertation follows. More details about these models are presented in Chapter 3.

In clustering based on finite mixtures of first-order Markov chain models, clustering means the partition of the bipolar youth into meaningful subgroups determined by the longitudinal course of the mood data observed for each subject. It is assumed that the data are generated by a mixture of probability distributions in which each distribution corresponds to a different group or cluster. Here the distributions define first order Markov chains, and the parameters consist of initial probabilities and transition probabilities.

Specifically, the data is modeled as having been generated in the following fashion: 1) a subject is assigned to a particular cluster with some probability, and 2) the behavior of that subject over time is then generated from a first-order Markov chain model with parameters specific to that cluster. It is assumed that this model generates the longitudinal data that was observed, and that only the subject behaviors are seen and not the actual cluster assignments. The Expectation-Maximization (EM) algorithm is used to learn the proportion of subjects assigned to each cluster as well as the parameters of each first order Markov model.

On the other hand, hidden Markov models (HMMs), have become popular media for modelling phenomena such as speech (Rabiner, 1989; Juang and Rabiner, 1991). In HMMs, there is a set of quantities, x, representing some unobservable phenomenon, and a set of observables, y. Roughly speaking, y is a distorted version of x. In the context of speech recognition, x represents a time-sequence of configurations of an individual's vocal tract, and y represents the corresponding time-sequence of projected sounds. The Markovian assumption would be that the elements of x form a realization of a Markov chain (Archer and Titterington, 2002).

Using the notation $p(x)$ for the probabilities of all relevant x, and $p(y|x)$ for the probability model assumed for the process by which x is distorted so as to produce y, the prime interest is to use the observed y to make inferences about the unobserved x, and the normative statistical approach is to base such inferences on $p(x|y)$, the conditional distribution for x given y. By Bayes theorem:

$$p(x|y) \propto p(y|x)p(x), \tag{1.1}$$

where the constant of proportionality does not depend on x. Thus, the right-hand side of 1.1 is the key to defining such quantities as the maximum a posteriori (MAP) estimate of x, given y. The above notation is somewhat facile, however, because within the factors on the right-hand side of 1.1 lurk parameters that are very likely to be unknown. In fact,

$$p(x) = p(x|\beta), \qquad p(y|x) = p(y|x, \phi), \tag{1.2}$$

where β and ϕ denote parameters. The ultimate goal is to estimate these parameters. There are several approaches to estimating the parameters; some of these approaches are likelihood-based. For hidden Markov models, the observed data are y, the parameters are (β, ϕ), and the likelihood function is

$$L(\beta, \phi; y) = p(y|\beta, \phi) = \sum_x p(y|x, \phi)p(x|\beta) \tag{1.3}$$

The practical problem in dealing with data from hidden Markov models is that the right-hand side of 1.3 is not a simple function; the summation cannot be carried out explicitly in order to create a neat formula and, consequently, computation of the maximizing (β, ϕ) is not immediately straightforward. Were both x and y observed, the situation would be simpler in that the appropriate likelihood would be

$$L_c(\beta, \phi; x, y) = p(y|\beta, \phi) = \sum_x p(y|x, \phi)p(x|\beta) \tag{1.4}$$

and the two factors on the right-hand side of 1.4 can be maximized separately with respect to β and to ϕ, respectively. The subscript ‘c’ is for ‘complete’. The pair $(x; y)$ constitute complete data, whereas the real-life data, y, are incomplete, with x missing. This interpretation of data from hidden Markov models suggests the use of the EM algorithm, an iterative procedure for computing maximum likelihood estimates (MLEs) for incomplete data. For HMMs, the EM algorithm is usually known as the *Baum-Welch algorithm*.

In a hidden Markov model, the parameters are not strictly identifiable. For instance, the indices of the states of the Markov chain can be permuted without changing the law of the process $\{x\}$, and hence also the law of $\{y\}$. After defining an equivalence relation, Leroux (1992) shows that the equivalence classes are identifiable, i.e., that parameter values in different equivalence classes produce different stationary laws for the process $\{y\}$.

1.2 DISSERTATION OUTLINE

Chapter 2. of this dissertation presents relevant aspects of bipolar disorder, along with a description of the COBY data. Finite mixtures of first-order Markov chain models for clustering and hidden Markov models are described in Chapter 3. The results obtained with these two models are reported in Chapter 4. Chapter 5. contains a discussion of the results, along with the conclusions of this research. Also other approaches different to the two used in this dissertation that can be explored to characterize the longitudinal course of bipolar youth are proposed as future research.

2.0 PSYCHIATRIC BACKGROUND

2.1 BIPOLAR DISORDER

2.1.1 Definition

Bipolar disorder, previously known as manic-depressive illness, is a familial recurrent mental disorder characterized by periods of depression and mania (or hypomania, a less severe form of mania) (DSM-IV, 1994). This disorder is accompanied by severe problems in the person's psychosocial functioning and increases the risk for drug abuse, legal problems and suicide. Symptoms of depression, mania and hypomania are described below.

2.1.2 Diagnosis

One of the sources for psychiatric diagnosis is the Diagnostic and Statistical Manual (DSM) written by the American Psychiatric Association, currently in its fourth text revised edition (DSM-IV-TR, 2000). The DSM-based definition of bipolar disorder is built on the identification of individual mood episodes. Mood episodes are discrete periods of altered feeling, thought and behavior; they have a distinct onset and offset, beginning and eventually ending gradually after several weeks or months. In bipolar disorder, the cardinal symptoms are discrete periods of abnormal mood and activation that define depressive and manic (or hypomanic) episodes, respectively. Diagnosis of such episodes is based exclusively on *phenomenology*, the descriptive appearance of the syndrome of interest.

Major depressive episodes are defined by periods of depression or irritability, and/or loss of interest or pleasure in life, which typically endure for weeks. These symptoms are often accompanied by changes in sleep, appetite, energy, cognition, and judgment (First and Tasman, 2004). Particular signs and symptoms of depression include: lasting sad, anxious, or empty mood; feelings of hopelessness or pessimism; feelings of guilt, worthlessness, or helplessness; loss of interest or pleasure in activities once enjoyed; decreased energy, a feeling of fatigue or of being “slowed down”; difficulty concentrating, remembering, making decisions; restlessness or irritability; sleeping too much, or too little; change in appetite and/or unintended weight loss or gain; chronic pain or other persistent bodily symptoms that are not caused by physical illness or injury; thoughts of death or suicide attempts. A depressive episode is diagnosed if five or more of these symptoms last most of the day, nearly every day, for a period of two weeks or longer (NIMH, 2002).

Manic episodes are defined by periods of abnormally elevated, expansive, or irritable mood accompanied by marked impairment in judgement and social and occupational functioning. These symptoms are frequently accompanied by unrealistic grandiosity, excess energy, and increases in goal-directed activity that have a high potential for damaging consequences (First and Tasman, 2004). Symptoms and signs of mania include: increased energy, activity, and restlessness; excessively “high”, overly good, euphoric mood; extreme irritability; racing thoughts and talking very fast, jumping from one idea to another; distractibility, cannot concentrate well; little sleep needed; unrealistic beliefs in one’s abilities and powers; poor judgment; spending sprees; a lasting period of behavior that is unusual; increased sexual drive; abuse of drugs, particularly cocaine, alcohol, and sleeping medications; provocative, intrusive, or aggressive behavior; denial that anything is wrong. A manic episode is diagnosed if elevated mood occurs with three or more of the other symptoms most of the day, nearly every day, for one week or longer. If the mood is irritable, four additional symptoms must be present (NIMH, 2002).

Diagnosis of bipolar disorder derives from the occurrence of individual episodes over time. Those who experience major depressive and manic episodes are diagnosed with *bipolar I* disorder or *BPI* (DSM-IV-TR, 2000) and those with major depressive and milder or shorter episodes of mania, usually called “hypomanic episodes” are diagnosed with *bipolar II* disorder or *BPII* (DSM-IV-TR, 2000). DSM-IV (1994) is the first version of the DSM series to include a specific category for bipolar II disorder. The separation of type II from both type I and major depressive disorders was supported by evidence found in studies of bipolar disorder (First and Tasman, 2004). According to the DSM-IV-TR (2000), people who have significant manic, hypomanic and depressive symptoms, but who do not fulfill the criteria for BPI or BPII are classified as "*bipolar disorder not otherwise specified*" or *BPNOS*.

2.1.3 Children and adolescents

Literature concerning adult samples has noted that 20% to 40% of adults report that their onset was during childhood, with depression as the first episode (Geller and Luby, 1997). Several other studies report the onset of bipolar disorder occurring during youth for a large number of patients. Particularly, for studies on prepubertal and adolescent populations, there is a general consensus that bipolar disorder can and does exist in children and adolescents, and furthermore, that it leads to marked impairment in functioning—specifically, marked deterioration in: academic achievement; work effort; maternal, paternal and peer relationships and extracurricular involvement (Shulman et al., 2002).

However, bipolar disorder is difficult to recognize and diagnose in youth, because it does not fit precisely the symptom criteria established for adults, and because its symptoms can resemble or co-occur with those of other common childhood-onset mental disorders. Additionally, symptoms of bipolar disorder may be initially mistaken for normal emotions and behaviors of children and adolescents. (NIMH, 2000). Shulman et al. (2002) report about the consistency across studies in the findings that youths with bipolar disorder appear to have higher rates of mixed mania and rapid mood changes than adults. The most common mania symptoms among youths are psychomotor agitation (irritable and prone to destructive outbursts), reduced sleep duration and talkativeness. On the other hand, when they are depressed, bipolar youths have: a sad appearance, poor self-steem, hallucinations, frequent absences from school or poor performance in school, somatic complaints (headaches, muscles aches, stomachaches or tiredness), they talk of or make efforts to run away from home, they are irritable, they complain, cry unexplainably, isolate socially, communicate poorly, and are extremely sensible to rejection or failure (Shulman et al., 2002; NIMH, 2000).

Existing evidence indicates that bipolar disorder beginning in childhood or early adolescence may be a different, and possibly more severe, form of the illness than older adolescent- and adult-onset bipolar disorder. When bipolar disorder begins before or soon after puberty, it is often characterized by a continuous, rapid-cycling, irritable, and mixed symptom state that may co-occur with disruptive behavior disorders, such as attention deficit hyperactivity disorder (ADHD) or conduct disorder (CD), or may have features of these disorders as initial symptoms. In contrast, later adolescent- or adult-onset bipolar disorder tends to begin suddenly, often with a classic manic episode, and to have a more episodic pattern with relatively stable periods between episodes, and there is also less comorbidity. Studies comparing youth- versus adult-onset bipolar disorder attribute the differences in illness characteristics either to the possibility that an earlier age of onset indicates a more severe biological form of the illness, or that an earlier onset interrupts psychosocial development. Both factors are likely involved in explaining the consistent findings that an earlier age of onset tends to be associated with greater overall psychopathology and impairment (Shulman et al., 2002).

Summarizing, prepubertal onset manic-depressive disorder may not present with the sudden or acute onset and improved interepisode functioning characteristic of the disorder in older adolescents and adults. Rather, it may present with a picture of continuous, mixed, rapid cycling of multiple brief episodes. Therefore, future studies of the longitudinal course of bipolar children will be crucial for developing long-term, prophylactic treatments for implementation during the prepubertal years (Geller and Luby, 1997).

2.1.4 Longitudinal course

Both phenomenological types of data, *cross-sectional* and *longitudinal*, are essential for the definition of mood disorders and the proper diagnosis of bipolar disorder (First and Tasman, 2004). Very often the diagnosis of bipolar disorder can be correctly made only during the long-term course of the illness, because in the majority of cases the first episode of the disorder is depressive. Thus, the concept of bipolar disorder is fundamentally defined by its course (Marneros and Brieger, 2002).

The longitudinal course of any mental disorder includes all phenomena which occur after the onset of the illness. Features of major importance when studying longitudinal course are

(Marneros and Brieger, 2002):

- onset of the disorder(type of onset, age of onset)
- episodes (i.e., type of episode, number, frequency, length)
- cycles (i.e., number, length, frequency, intervals, persisting symptoms, stability of syndrome shift)
- activity of the episodes (i.e., re-manifestation during a defined period of time)
- outcome (the end-point of follow-up in a defined period of time)

Longitudinal course and outcome can be assessed with different methodologies. A compromise between the retrospective and prospective perspectives may be the concept of a *catch-up study*, i.e., information comes from case records and is retrospectively assessed, while present data are assessed by "catching up" with the former patients and actively examining them.

Another descriptive characteristic of a study is the observation time: long-term (10 or more years), medium-term (4-9 years) and short-term (1-3 years). Another distinction among studies is controlled vs. naturalistic studies. In the former, the researcher controls certain variable that may modify longitudinal course and outcome, such as treatment. In the latter, the researcher observes longitudinal course and outcome without interfering with the natural course of the illness.

2.2 STATISTICAL ANALYSIS OF LONGITUDINAL COURSE OF BIPOLAR DATA

Hennen (2003) reviews statistical methods for the analysis of longitudinal bipolar clinical trials involving relatively large samples, with outcome measures obtained repeatedly overtime. Special circumstances affecting choice of methods in bipolar disorder research include: (i) longitudinal study designs are preferable, with repeated measurements made at several time points; (ii) outcome measures that can be considered continuous or quasi-continuous, such as change-from-baseline scores on a psychiatric rating scale, may be preferable to or at least should supplement, binary outcome measures; (iii) it may be advisable to adjust for baseline severity levels, even with randomized, blinded assignment; (iv) missing data due to subject dropout and other reasons occur frequently, and data analytic methods need to accommodate missingness; (v) temporal variation in outcome measures is often considerable, requiring a large number of subjects to assure adequate statistical power.

Several methods are commonly used to assess outcomes obtained serially in longitudinal (repeated measures) bipolar disorder research. Seven of these are widely used (Hennen, 2003): 1. Endpoint analysis, 2. Endpoint analysis with last-observation-carried-forward (LOCF), 3. Summary measures (including slope estimation and area-under-the-curve estimation) (Senn et al., 2000), 4. Random effects/mixed effects regression modeling or generalized estimating equation (GEE) regression modeling (Laird and Ware, 1982; Diggle et al., 1994; Hardin and Hilbe, 2003), 5. Time-to-event (survival analysis) modeling (Klein and Moeschberger, 1997), 6. Multivariate analysis of variance (MANOVA), and 7. Analysis of variance (ANOVA) with repeated measures (Winer et al., 1991).

Of these seven alternatives, Method 1 is sometimes useful, especially when the number of repeated measures is small. Method 2 is very commonly used, especially in controlled treatment trials research. For Method 3, if the summary measure is selected appropriately, this methods yields readily interpretable results and missing data are typically tolerated acceptably well, although they may disrupt reliable estimation of the chosen summary measure. Method 4 has the flexibility to be extended in various ways by the selection of different modeling assumptions, all the available data are used, there is tolerance of missing data and covariate adjustments are readily incorporated. Method 5 is increasingly widely used because of its meaningful clinical and scientific results in their time-to-signal-events outcomes. Method 6 is of little practical use because a missing observation requires casewise deletion and it essentially disregards the time dimension. Method 7 may be useful in some situations in which the ϵ degrees of freedom adjustment does not reduce the statistical power substantially. All these methods can be applied to data obtained in observational studies acquiring information sequentially over time.

The choice of an appropriate method depends on several factors. If there are only two time periods, then for continuous data a summary statistic combining baseline and endpoint observations is likely to be the most useful approach, a t-test or a simple linear regression model can be done. For a binary outcome, logistic regression is recommended. When there are multiple time periods and time-to-event data are available , then survival analysis can be used. Alternatively, random effects or GEE methods can be chosen. After illustrating the seven methods in a clinical trial of medicines used to treat mania, Hennen (2003) recommends that bipolar disorder investigators doing longitudinal research consider summary measures, random effects or GEE regression modeling and survival analysis as potentially more useful.

Marneros and Brieger (2002) present a review of the longitudinal course and outcome of bipolar disorder studies, mainly for naturalistic studies. Among the results at follow-up reported by these studies are descriptive statistics such as the number of episodes, length of episodes, impairment, cycle count and length, switch rates, symptomatology changes, times to recovery, relapse rates, hospitalization percentage. The statistical analyses used in most of such studies are logistic regression and survival analysis, depending on the response variable of interest.

Other statistical analyses already applied in longitudinal studies of bipolar disorder is approximate entropy (ApEn). It was introduced as a model-independent quantification of the regularity (complexity) of data. This approach calibrates an ensemble extent of sequential interrelationships, quantifying a continuum ranging from totally ordered to completely random, with larger values corresponding to greater apparent process randomness or serial irregularity and smaller values corresponding to more instances of recognizable features or patterns in the data. For ApEn, discerning changes in order from apparently random to very regular is the primary statistical focus (Pincus, 2006).

ApEn assigns a non-negative number to a sequence or time-series. Two input parameters, a run length m and a tolerance window r must be specified to compute ApEn. It measures the logarithmic likelihood that runs of patterns that are close (within r) for m contiguous observations remain close (within the same tolerance width r) on next incremental comparisons. Theoretical analysis and clinical applications have estimated standard parameter values, 1 or 2 for m and a fixed value of 0.1 to 0.25 times the standard deviation of the individual subject time series. These input parameters produce good statistical reproducibility for ApEn for time series of lengths 60 or more. ApEn has been used when the time series for all subjects have the same length. Pincus (2006) suggests what kind of psychiatric serial data can be analyzed with ApEn and Glenn et al. (2006) and Rao et al. (2006) present applications of approximate entropy to mood data from patients with bipolar disorder.

2.3 COURSE AND OUTCOME OF BIPOLAR YOUTH STUDY

The Course and Outcome of Bipolar Youth (COBY) study is funded by the National Institute of Mental Health. COBY was designed to build on and extend the existing scientific database on the cross-sectional presentation and longitudinal course of pediatric bipolar disorder (Birmaher et al., 2006).

2.3.1 Demographics

Up to January 2008, there were 413 participants in COBY having intake and follow-up data. These participants were enrolled in outpatient and inpatient units at three university centers: Brown University (n=135), University of California at Los Angeles (n=74) and University of Pittsburgh Medical Center (n=204).

The 413 children and adolescents (mean age±SD: 12.63±3.26) were assessed by semi-structured interview and diagnosed as BPI (n=244), BPII (n=28) or BPNOS (n=141). There are several variables measured at intake, among them: demographic characteristics, age of onset and duration of bipolar spectrum illness and symptom severity (Birmaher et al., 2006). Despite the huge number of characteristics measured on each patient at intake, the interest of this dissertation is only in the demographic variables. Here is brief summary of the COBY patients in terms of the chosen variables. 185 patients experienced the bipolar onset during childhood, 123 during early adolescence and 105 during late adolescence. Social economical status, an ordinal variable with categories ranging from 1 (low SES) to 5 (high SES), has this distribution: 7.5%, 17.2%, 21.1%, 34.6% and 19.6%. The percentage of males in the study is 53.5% and 42.1% of the patients live with both natural parents.

2.3.2 Longitudinal data: Psychiatric Status Rating

COBY also measured longitudinal changes in psychiatric symptomatology, functioning, and treatment exposure. They were assessed using the Longitudinal Interval Follow-up Evaluation (LIFE) (Keller et al., 1987). The LIFE was administered to adolescents and parents separately. On the other hand, younger children were interviewed together with their parents, because often these children have problems determining the times of their symptomatology. Any discrepancies between the informants' responses were discussed and a summary score based on all available information was determined. The LIFE evaluates the longitudinal course of symptoms by identifying "change points", frequently anchored by memorable dates for the subject (e.g., holidays, beginning of school). The severity of ongoing symptoms, the onset of new symptoms and the episode polarity for bipolar disorder since the last appointment are tracked on a week-by-week basis using the LIFE Psychiatric Status Rating (PSR) scale.

For the data at hand, there are a total of 71,328 longitudinal entries that result from the follow-up of the 413 participants. The numbers of weeks of follow-up range from 26 to 337 weeks, with median 176 weeks, which puts COBY between a short and medium term study, according to the study characteristics described by Marneros and Brieger (2002). Regarding the design, COBY is a catch-up study and since treatment has been given to the patients throughout the follow-up without randomizing the subjects to any treatment, this does make COBY a naturalistic study.

The Psychiatric Status Rating (PSR) was developed to generate analyzable data about the longitudinal course of a subject's psychopathology. The PSR's are numeric values that have been operationally linked to the DSM-IV-TR (2000) criteria. DSM-IV criteria information is gathered in the interview, and then later translated into ratings for each week of the follow-up period. The ratings indicate the severity level of an episode, as well as whether the patient has recovered or relapsed. For DSM-IV mood disorders, the PSR scores range from 1 for no symptoms, 2 to 4 for varying levels of subthreshold symptoms and impairment, and 5 to 6 for full criteria with different degrees of severity or impairment. Comorbid disorders and psychosis are also rated on a weekly basis on a 3-point scale of 1 to 3, where 3 indicates threshold symptomatology. There are 22 diagnoses rated in the PSR for each child. Only three of them are under study here: depression, hypomania and mania. Table 1. contains the meaning and description of the the PSR scale used to rate the severity of these three diagnoses.

Table 1: Six-point rating scale for Psychiatric Status Rating

CODE	TERM	DESCRIPTION
6	Definite criteria (severe)	Meets DSM-IV criteria for definite episode and has either prominent psychotic symptoms or extreme impairment in functioning
5	Definite criteria	Meets DSM-IV criteria for definite, current episode, but has no prominent psychotic symptoms or extreme impairment in functioning
4	Marked	Does not meet definite DSM-IV criteria, but has major symptoms or impairment from the disorder
3	Partial remission	Considerably less psychopathology than full criteria with no more than moderate impairment in functioning, but still has obvious evidence of the disorder
2	Residual	Either patient claims not to be completely back to ”usual self” or rater notes the presence of one or more symptoms of this disorder in no more than a mild degree
1	Baseline	Patient returns to ”usual self” without any residual symptoms of this disorder, but may or may not have significant symptoms from other condition or disorder

The 6-point PSR scales for depression, mania and hypomania can be combined in a single 12-point PSR scale as shown in Table 2. Thus, a single sequence is generated for each COBY participant, instead of having three separated sequences. This scale is arranged in such a way that the lower two values correspond to depressive episodes, then the well episode, followed for the categories of the manic episodes (submania, hypomania and mania) with each of the following: pure, subdepression and MDD. This scale is the one used in the statistics reported for the PSR scores.

Table 2: The twelve PSR categories for follow-up data

Score	Episode	Depression	Mania	Hypomania
1	Major Depressive Disorder (MDD)-Pure	5-6	1	1-2
2	Subdepression only	3-4	1	1-2
3	Well	1-2	1	1-2
4	Submania only	1-2	1	3-4
5	Submania/Subdepression	3-4	1	3-4
6	Submania/MDD	5-6	1	3-4
7	Hypomania-Pure	1-2	1	5-6
8	Hypomnia/Subdepression	3-4	1	5-6
9	Hypomania/MDD	5-6	1	5-6
10	Mania-Pure	1-2	5-6	1-2
11	Mania/Subdepression	3-4	5-6	1-2
12	Mixed state	5-6	5-6	1-2

2.3.3 Descriptive statistics of the PSR

A fairly simplified characterization of the COBY follow-up data at hand is: (i) the three longitudinal sequences (one for each: depression, mania and hypomania) have been coded into one sequence for each COBY participant, (ii) each sequence is represented as a list of discrete symbols, and (iii) each symbol represents one of twelve possible PSR categories.

As an exploratory analysis, the initial states probabilities and the transition matrix of a first order Markov chain have been computed: (i) using the data from all 413 COBY participants and (ii) by bipolar diagnosis, i.e., for BPI, BPII and BPNOS. The results are reported int he next two sections.

2.3.3.1 Initial state probabilities Regarding the episode observed in the first week of follow-up for each of the COBY patients, Table 3 contains the distribution of the initial state for all patients and by bipolar diagnosis.

Clearly, the most frequent initial state overall and in each of the bipolar groups is Well (3). About a third of COBY patients were in a well state when entering the study. This is also observed within diagnoses.

However, the picture changes a little when talking about the least frequent initial state. Overall, Hypomania/MDD (9) is the least frequent. Hypomania/Subdepression (8) and Hypomania/MDD (9) and are the least frequent initial state for BPI. For BPII, states 7 to 12—those related to hypomania and mania—were not observed as initial states; among the states that were observed, Submania/Subdepression (5) is the least frequent initial state. Lastly, states 7 and 10 to 12—Hypomania pure and the mania states—were not observed as initial states for the children and adolescents diagnosed with BPNOS. Hypomania/MDD (9) is the least frequent initial state among the initial states observed for BPNOS.

Table 3: Initial state distribution (in percentages)

PSR	1	2	3	4	5	6	7	8	9	10	11	12
All	8.7	9.0	31.5	13.1	17.0	5.8	1.2	1.0	0.7	4.8	2.4	4.8
BPI	7.4	7.8	31.6	10.3	13.1	5.7	2.1	0.8	0.8	8.2	4.1	8.2
BPII	25.0	10.7	35.7	10.7	7.1	10.7						
BPNOS	7.8	10.6	30.5	18.4	25.5	5.0		1.4	0.7			

2.3.3.2 Transition matrices As a description of the episode transitions experienced by the COBY patients, the transition probability matrix of a first order Markov chain was computed by averaging the frequency of the transitions of all patients, specifically:

$$p_{jk} = \frac{n_{jk}}{\sum_{k=1}^{12} n_{jk}}$$

where,

$$n_{jk} = \sum_{i=1}^{413} \sum_{t=1}^{L_i-1} I(\text{PSR} = k \text{ at week } t+1 | \text{PSR} = j \text{ at week } t), j, k = 1, \ldots, 12$$

where, $I(\cdot)$ is an indicator function that equals 1 when its argument is true and 0 otherwise.

The transition matrices by bipolar diagnoses were also computed, again averaging the frequency of the transitions of patients sharing the same diagnosis. In the four transition matrices below, a zero means that no transition was observed in COBY from the PSR score in the row to the PSR score in the column.

Table 4. presents a summary of the overall transition probabilities among the twelve PSR scores. Note that the transitions on the diagonal have values above 0.71 and the transition from Well to Well (3,3) has the highest probability. All transitions off the diagonal are below 0.10, except for Submania only to Well (4,3) and for Hypomania-Pure to Well (7,3). This transition probability matrix suggest that COBY patients tend to stay on the same episode, they do not transition to another episode too often.

Table 4: Overall transition probabilities among the 12 PSR scores

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
 [1,] 0.80
 [2,]      0.79
 [3,]           0.91
 [4,]           0.12 0.80                     0
 [5,]                     0.86
 [6,]                          0.83                       0
 [7,]           0.13                0.74                  0
 [8,]                                    0.74
 [9,]                0                        0.82 0      0
[10,]                                         0    0.71
[11,]                                         0           0.84
[12,]                                                           0.84
```

`0` means a zero-transition probability and `blank` means probability between 0 and 0.10

Besides, there are eight transitions that were not observed overall in COBY patients: Submania/MDD → Mania/Subdepression (6,11), Hypomania pure → Mania/Subdepression

(7,11), Submania only ←→ Hypomania/MDD (4,9 and 9,4), Hypomania/MDD ←→ Mania pure (9,10 and 10,9) and Hypomania/MDD ←→ Mania/Subdepression (9,11 and 11,9).

A similar summary of the transition probabilities for patients diagnosed with BPI is presented in Table 5. Again the highest transition probability is observed for the Well to Well (3,3) transition, and all the probabilities on the diagonal are greater than 0.73. In contrast, there are only three transitions off the diagonal with probabilities above 0.10. Those transitions correspond to Submania only to Well (4,3), Hypomania-Pure to Well and Hypomania/Subdepression to Subdepression only (8,2). There are twelve transitions that were not observed for BPI patients.

Regarding the transition probability matrix for BPII patients, there are 57 transitions that did not occur in COBY (see Table 6), not surprising given that there are only 28 individuals in this group. Most of the observed transitions are in the upper left corner of the matrix. For BPII, the transitions on the diagonal are between 0.44 and 0.88, with the probabilities for Well-Well (3,3) being the highest. This time there are several transitions off the diagonal with probabilities above 0.10.

For BPNOS, the transition probabilities on the diagonal of the transition probability matrix are above 0.66 and below 0.91. The highest probabilities are for transitions Well to Well (3,3) and Mixed state to Mixed state (12,12). Only three probabilities off the diagonal are greater than 0.10: Hypomania-Pure to Well (7,3), Mania-Pure to MDD-Pure (10,1) and Mania-Pure to Subdepression only (10,2). There are 31 zero-probability transitions.

Thus from the results obtained at this exploratory stage, it could be said that COBY patients, overall and by bipolar diagnosis, tend to stay on the same episode and do not move to another episode too often. However, there are some slight differences among the bipolar diagnosis. They differ in transitions observed off the diagonal. And BPII seems to be the group with patients that transition more frequently.

Table 5: Transition probabilities among the 12 PSR scores for BPI patients

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
 [1,] 0.80
 [2,]      0.80
 [3,]           0.91
 [4,]           0.14 0.79                0    0
 [5,]                     0.86
 [6,]                          0.80                      0
 [7,]           0.11           0    0.79                 0
 [8,]      0.10               0         0.75
 [9,]                0                        0.83 0     0
[10,]                          0              0    0.73
[11,]                                         0          0.83
[12,]                                                          0.83
```

`0` means a zero-transition probability and `blank` means probability between 0 and 0.10

Table 6: Transition probabilities among the 12 PSR scores for BPII patients

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
 [1,] 0.75                     0.10                      0
 [2,]      0.75 0.13                          0    0    0           0
 [3,]           0.88                               0          0
 [4,]           0.32 0.59                          0          0     0
 [5,]                     0.79      0              0    0           0
 [6,] 0.20                     0.72 0                   0     0
 [7,]      0.21 0.29      0    0    0.44           0    0     0     0
 [8,] 0    0              0.18           0.67           0           0
 [9,] 0.16 0    0    0              0    0.13 0.63 0    0     0
[10,] 0    0    0.20 0.10 0    0.10 0    0    0    0.50 0     0.10
[11,] 0              0    0         0    0    0    0    0.84  0
[12,] 0.18 0    0.14 0    0         0    0    0    0    0     0.63
```

`0` means a zero-transition probability and `blank` means probability between 0 and 0.10

Table 7: Transition probabilities among the 12 PSR scores for BPNOS patients

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	0.81										0	0
[2,]		0.78									0	
[3,]			0.91									
[4,]				0.84					0			0
[5,]					0.87					0		
[6,]						0.88	0				0	
[7,]			0.12				0.67		0		0	0
[8,]						0		0.75		0		0
[9,]		0		0			0		0.87	0	0	
[10,]	0.12	0.13						0	0	0.66		0
[11,]	0					0	0		0		0.86	
[12,]	0	0		0			0	0				0.91

0 means a zero-transition probability and blank means probability between 0 and 0.10

3.0 STATISTICAL LITERATURE REVIEW

A review of two Markov model techniques found in the literature to find patterns in discrete longitudinal sequences of varying length are presented in the following sections.

3.1 MIXTURES OF FIRST ORDER MARKOV CHAIN MODELS FOR CLUSTERING

Cadez et al. (2003) present a mixture of first-order Markov chain models to cluster Internet users. The description of their model is given here in terms of the COBY follow-up data. The data is modeled as having been generated in the following fashion: 1) a subject is randomly assigned to a cluster with unknown probabilities of cluster assignment, and 2) the behavior of that subject is then generated from a Markov chain model with parameters specific to that cluster.

This approach to clustering is sometimes called a model-based (or mixture model) approach. The clustering model described above is a finite mixture of Markov models.

Let $\mathbf{X}_n$ be a multivariate random variable taking on values $\mathbf{x}_n$ corresponding to the behavior of the nth child or adolescent, $n = 1, \ldots, N$. Let C be a discrete-valued variable taking on values $c_1, c_2, \ldots, c_K$. The value of C corresponds to the unknown cluster assignment for a child. A *mixture model for* $\mathbf{X}$ *with* K *components* has the form:

$$p(\mathbf{x}_n|\theta) = \sum_{k=1}^{K} p(c_k|\theta) p_k(\mathbf{x}_n|c_k, \theta) \tag{3.1}$$

where $p(c_k|\theta)$is the marginal probability of the kth cluster satisfying $\sum_k p(c_k|\theta) = 1$, and

$p_k(\mathbf{x}|c_k, \theta)$ is the statistical model describing the distribution of $\mathbf{X}_n$ for subjects in the kth cluster. θ denotes the *parameters* of the model. Details on θ for the first order Markov chain model are given below.

In this research, $\mathbf{X}_n = (X_1, X_2, \ldots, X_{L_n})$ is an arbitrarily long sequence of variables describing the bipolar conditions of the nth child who has been followed during L_n weeks. The variable X_i takes on some value x_i from among the M possible PSR categories or states representing the child's bipolar conditions. The assumption here is that each model component is a *first-order Markov chain model*:

$$p_k(\mathbf{x}_n|c_k, \theta) = p(x_{n1}|\theta_k^I) \prod_{i=2}^{L_n} p(x_{ni}|x_{n(i-1)}, \theta_k^T)$$

where θ_k^I denotes the parameters of the probability distribution over the PSR category at intake among subjects in cluster k, and θ_k^T denotes the parameters of the probability distributions over transitions from one category to the next by a subject in cluster k. This model captures (to some degree) the nature of the child's bipolar conditions. Specifically, it captures the child's condition at intake, the dependency between two consecutive conditions and the last condition observed in the follow-up.

Explicitly, θ in the model described above is $\theta = \{\pi, \theta^I, \theta^T\}$, where:

- π is a vector of K mixture weights, $\pi = \{\pi_1, \ldots, \pi_K\}$, $\sum_{k=1}^{K} \pi_k = 1$
- θ^I is a set of K initial state probability vectors, $\theta^I = \{\theta_1^I, \ldots, \theta_K^I\}$ where the per-component initial state probabilities θ_k^I, $1 \leq k \leq K$ are vectors of length M: $\theta_k^I = (\theta_{k,1}^I, \ldots, \theta_{k,M}^I)$, $\sum_{j=1}^{M} \theta_{k,j}^I = 1$
- θ^T is a set of K transition matrices, $\theta^T = \{\theta_1^T, \ldots, \theta_K^T\}$ where the per-component transition probability matrices θ_k^T, $1 \leq k \leq K$ are square matrices of order M: $\theta_k^T = \{\theta_{k,j,l}^T\}$, $\sum_{l=1}^{M} \theta_{k,j,l}^T = 1$

After estimating the model parameters given the data, the model can be used to assign subjects to clusters as follows. Given the observed behavior $\mathbf{x}_n$ of subject n, the probability distribution over the hidden variable C corresponding to the cluster assignment of the subject, can be computed by Bayes' rule:

$$p(c_k|\mathbf{x}_n, \theta) = \frac{\pi_k p_k(\mathbf{x}_n|c_k, \theta)}{\sum_{j=1}^{K} \pi_j p_j(\mathbf{x}_n|c_j, \theta)}. \tag{3.2}$$

The probabilities $p(c_k|\mathbf{x}_n, \theta)$ are sometimes called *membership probabilities.* Once these probabilities have been computed, the subject is assigned to the cluster with the highest probability.

3.1.1 Estimating the parameters in the model

3.1.1.1 Bayesian estimation To encode prior knowledge about the domain and/or to smooth the maximum likelihood estimates, a prior probability distribution over the parameter values, denoted $p(\theta)$, can be introduced. A criterion for estimating the parameters is to identify those parameter values that maximize the posterior probability of θ given the training data:

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \arg\max_{\theta} \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)p(\theta)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)} = \arg\max_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)p(\theta)$$

Closed-form solutions for θ_{MAP} do not always exist and iterative algorithms such as the EM are used to search for maxima. Each variable X_i is finite, $p(x_{n1}|\theta_k^I)$ is a multinomial distribution, and $p(x_{ni}|x_{n(i-1)}, \theta_k^T)$ is a set of multinomial distributions. A prior distribution used often for the parameters of a multinomial distribution is the Dirichlet distribution. A Dirichlet distribution for the multinomial distribution with parameters $\phi = (\phi_1, \ldots, \phi_a)$ is given by

$$p(\phi_1, \ldots, \phi_a|\alpha_1, \ldots, \alpha_a) = \frac{\Gamma(\sum_{i=1}^{a} \alpha_i)}{\prod_{i=1}^{a} \Gamma(\alpha_i)} \prod_{j=1}^{a} \phi_j^{\alpha_j - 1} \tag{3.3}$$

subject to $\sum_{i=1}^{a} \phi_i = 1$, $0 < \phi_i < 1$, $\alpha_i > 0$. Given this Dirichlet prior for ϕ, suppose we observe data (a multinomial sample) such that there are n_i occurrences of state i for $i = 1, \ldots, a$. Then, the posterior distribution for ϕ is another Dirichlet distribution with hyperparameters $(\alpha_1 + n_1, \ldots, \alpha_a + n_a)$. Thus, the Dirichlet distribution is a conjugate distribution for multinomial sampling.

3.1.1.2 The EM algorithm The expected value of the objective function over the class-posterior distribution using a fixed set of "current" parameters, denoted by $\mathcal{Q}$, is the key quantity of the EM. For the log-posterior (MAP) function, $lP_{\mathbf{x}_1,\ldots,\mathbf{x}_N}(\theta) = \log p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta) + \log p(\theta)$,

$$\mathcal{Q}(\theta, \theta_{old}) = \langle lP_{\mathbf{x}_1,\ldots,\mathbf{x}_N}(\theta)\rangle_{P(\theta_{old})} = \sum_{n=1}^{N}\sum_{k=1}^{K} P_{n,k}(\theta_{old}) \log[\pi_k p(\mathbf{x}_n|c_k, \theta)] + \log p(\theta) \tag{3.4}$$

where $P_{n,k}(\theta)$, $1 \le n \le N$, $1 \le k \le K$ is the class-posterior probability distribution given in (3.2). Maximizing $\mathcal{Q}$ with respect to each subset of parameters θ, the update rules for each set of parameters are:

- Mixture weights:
$$\pi_k = \frac{\sum_{n=1}^{N} P_{n,k}(\theta_{old}) + \alpha_k^{\pi}}{\sum_{k'=1}^{K}[\sum_{n=1}^{N} P_{n,k'}(\theta_{old}) + \alpha_{k'}^{\pi}]} \tag{3.5}$$
where α_k^{π} is the hyperparameter associated with π_k, $k = 1, \ldots, K$.
- Initial state probabilities:
$$\theta_{k,j}^{I} = \frac{\sum_{n=1}^{N} P_{n,k}(\theta_{old}) I(\mathbf{x}_{n1} = j) + \alpha_{k,j}^{I}}{\sum_{j'=1}^{M}[\sum_{n=1}^{N} P_{n,k}(\theta_{old}) I(\mathbf{x}_{n1} = j') + \alpha_{k,j'}^{I}]} \tag{3.6}$$
where $\alpha_{k,j}^{I}$ is the hyperparameter associated with $\theta_{k,j}^{I}$ and $I(\mathbf{x}_{n1} = j)$ is an indicator function that equals to 1 if the arguments are equal and 0 otherwise, $k = 1, \ldots, K$ and $j = 1, \ldots, M$.
- Transition probabilities:
$$\theta_{k,j,l}^{T} = \frac{\sum_{n=1}^{N} P_{n,k}(\theta_{old}) \delta_{j,l}(\mathbf{x}_n) + \alpha_{k,j,l}^{T}}{\sum_{l'=1}^{M}[\sum_{n=1}^{N} P_{n,k}(\theta_{old}) \delta_{j,l'}(\mathbf{x}_n) + \alpha_{k,j,l'}^{T}]} \tag{3.7}$$
where $\alpha_{k,j,l}^{T}$ is the hyperparameter associated with $\theta_{k,j,l}^{T}$ and $\delta_{j,l}(\mathbf{x}_n)$ denotes the number of transitions from state j to state l in $\mathbf{x}_n$, $k = 1, \ldots, K$ and $j, l = 1, \ldots, M$.

3.1.1.3 Initialization of the EM algorithm The likelihood function in Equation (3.1) usually has multiple local maxima; thus, a search for the overall maximum requires the application of the EM algorithm multiple times from a wide selection of starting values. Even when starting from several points, all of the EM runs may fail to converge to the global maximum, or the algorithm may get trapped in a flat likelihood area. Therefore, it is of special interest to obtain a reasonable set of initial values. The usual approach to specifying an initial set of starting values is to generate θ randomly.

3.1.1.4 Stopping rule of the EM algorithm One may be naïvely tempted to use a stopping rule for the EM algorithm based on the changes in the parameters or based on the log-likelihood being sufficiently small. Unfortunately, taking small EM-steps does not imply that the algorithm is getting close to the global maximum. It may be possible that the algorithm has been trapped in a flat log-likelihood area in which case either the relative change in the parameter or the log-likelihood measures only lack of progress in the estimation process. To avoid drawing this wrong conclusion, we need a more appropriate stopping rule. Böhnig et al. (1994) proposed the use of the following stopping rule criterion, which would force the algorithm to stop when the solution is near a local maximum:

$$\text{stop if } l_j^{stop} - l_j < \texttt{tol} \tag{3.8}$$

where l_j is the log likelihood value at the jth iteration of the EM algorithm,

$$l_j^{stop} = l_{j-2} + \frac{1}{1-a_j}(l_{j-1} - l_{j-2}),$$

with

$$a_j = \frac{l_j - l_{j-1}}{l_{j-1} - l_{j-2}},$$

and `tol` is the desired tolerance.

3.1.2 Choosing the number of components or clusters

The fit of a mixture model to a given data set can only improve (and the likelihood can only increase) as more components are added to the model. Hence likelihood cannot be used directly in assessment of models for cluster analysis (Fraley and Raftery, 1998). Here are the evaluation criteria used to choose the number of clusters or components in the mixture of first order Markov chain models.

3.1.2.1 Bayesian Information Criterion (BIC) An advantage of the mixture-model approach to clustering is that it allows the use of approximate Bayes factors to compare models. This gives a systematic means of selecting not only the parametrization of the model, but also the number of clusters. The Bayes factor is the posterior odds for one model against the other assuming neither is favoured a priori. When EM is used to find the maximum mixture likelihood, a reliable approximation to minus twice the log Bayes factor called the BIC (Schwarz, 1978) is applicable. The smaller (more negative) the value of the BIC, the stronger the evidence for the model. In the BIC, a term is added to the loglikelihood penalizing the complexity of the model, so that it may be maximized for more parsimonious parameterizations and smaller numbers of groups than the loglikelihood. The BIC can be used to compare models with differing parameterizations, differing numbers of components, or both. Although standard regularity conditions for the BIC to give consistent estimators of the parameters do not hold for mixture models, there is considerable theoretical and practical support for its use in this context (Fraley and Raftery, 1998).

3.1.2.2 Normalized Entropy Criterion (NEC) Celeux and Soromenho (1996) proposed the normalized entropy criterion as a criterion to be minimized to assess the number of components in a mixture model.

$$NEC(K) = \frac{E(K)}{L(K) - L(1)} \tag{3.9}$$

where $L(\cdot)$ corresponds to the log likelihood of the mixture model with $(\cdot)$ components and

$$E(K) = -\sum_{j=1}^{K}\sum_{n=1}^{N} p(c_k|\mathbf{x}_n,\theta)\ln p(c_k|\mathbf{x}_n,\theta) \geq 0$$

with $p(c_k|\mathbf{x}_n,\theta)$ as defined in Equation (3.2). This criterion determines K^* which minimizes $NEC(K)$, $2 \leq K \leq K_{sup}$. K^* is then chosen as the number of components in the mixture if $NEC(K^*) < 1$ otherwise no clustering structure is declared in the data (Biernacki et al., 1999).

3.1.2.3 Markov Chain Cross-Validation (MCCV) Cross-validation is a well-known technique to select a model from a family of candidate models. In probabilistic model-based clustering–i.e., finite mixture models, where each component can be considered as a cluster–, any score function which measures the quality of fit of the density also provides a candidate function for model selection. Smyth (2000) presents cross-validated likelihood as an appropriate score function for choosing the number of components in finite mixture models.

Here are some details of this measure. Let $l_k^{train} = \log p(\hat{\theta^k}(\mathbf{x}_1,\ldots,\mathbf{x}_N)|\mathbf{x}_1,\ldots,\mathbf{x}_N)$ denote the log-likelihood of the fitted model with k components, with $\hat{\theta^k}$ estimated using the data set $\mathbf{x}_1,\ldots,\mathbf{x}_N$ and the log-likelihood evaluated on that same data. l_k^{train} is a non-decreasing function of k since the increment on mixture components allows better fit to the data. Hence, l_k^{train} do not provide clues about the number of components.

Instead, assuming that there is a large test data set $\mathbf{x}_{N+1},\ldots,\mathbf{x}_{N+R}$ that was not used to fit the model, the log-likelihood $l_k^{test} = \log p(\hat{\theta^k}(\mathbf{x}_1,\ldots,\mathbf{x}_N)|\mathbf{x}_{N+1},\ldots,\mathbf{x}_{N+R})$, corresponds to the log-likelihood evaluated on the $\mathbf{x}_{N+1},\ldots,\mathbf{x}_{N+R}$ data set, but with the parameters of the model determined by $\mathbf{x}_1,\ldots,\mathbf{x}_N$. l_k^{test} or test log-likelihood or log predictive score can be interpreted as a function of k, keeping $\mathbf{x}_1,\ldots,\mathbf{x}_N$ fixed. Smyth (2000) showed a property of the test log-likelihood that motivates its use as a model selection criterion in this context.

However, a large independent data set $\mathbf{x}_{N+1},\ldots,\mathbf{x}_{N+R}$ is not always available. A practical alternative for model selection is to use a cross-validated estimate of the test log-likelihood: l_k^{cv}. In one version of cross-validation, the available data are repeatedly partitioned into two sets, one is used to build the model and the other is used to evaluate the statistic of interest. Let M be the number of partitions. For the ith partition, denote by

$\mathbf{e}_i$ the data subset used for evaluation of the log-likelihood and $\mathbf{x} - \mathbf{e}_i$ be the remaining data, which is used for model estimation. Therefore, the cross-validated estimate of the test log-likelihood for the kth model is:

$$l_k^{cv} = \frac{1}{M} \sum_{i=1}^{M} \log p(\hat{\theta^k}(\mathbf{x} - \mathbf{e}_i | \mathbf{e}_i) \tag{3.10}$$

Depending on how one chooses the partitions, different cross-validation methodologies have been proposed Smyth (2000). One of them, "v-fold" cross-validation, partitions $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in v disjoint test subsets: $\{\mathbf{s_1}, \ldots \mathbf{s_v}\}$, each of size N/v. Two special cases are $v = N$, known as "leave-one-out" and $v = 10$, known as the ten-fold cross-validation. For model selection in linear regression, several authors have proposed a particular CV procedure generating M independent partitions, each with a fixed fraction β of the data used as the test sample and $1 - \beta$ used to estimate the parameters. This last procedure is known as "Repeated Learning Testing" (RLT) or "Monte Carlo Cross Validation" (MCCV).

3.1.2.4 Score The number of clusters (or components in the mixture) is chosen by finding the model that accurately predicts R new "test" cases $\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+R}$. That is, a model with K clusters that minimizes the out-of-sample predictive log score is chosen:

$$\text{Score}(K, \mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+R}) = -\frac{\sum_{h=1}^{R} \log_2 p(\mathbf{x}_{n+h} | \theta_K(\mathbf{x}_1, \ldots, \mathbf{x}_N))}{\sum_{h=1}^{R} L_h} \tag{3.11}$$

3.2 HIDDEN MARKOV MODELS

The basic theory of hidden Markov models (HMMs) was published in a series of classic papers by Baum and his colleagues in the late 1960s and early 1970s. Rabiner (1989) provides an overview of the basic theory of HMMs (as originated by Baum and his colleagues) and provides practical details on methods of implementation of the theory.

A Hidden Markov Model (HMM) is a probabilistic model of the joint probability of a collection of random variables $\{O_1, \ldots, O_T, q_1, \ldots, q_T\}$, where $\{O_1, \ldots, O_T\}$ denotes an observed sequence and $\{q_1, \ldots, q_T\}$ denotes a hidden sequence. The O_t variables are either continuous or discrete and the q_t variables are always discrete (Bilmes, 1997). Under an HMM, there are two conditional independence assumptions:

1. The t^{th} hidden variable, given the $(t-1)^{\text{th}}$ hidden variable, is independent of previous variables (both observed and hidden):

$$P(q_t|q_{t-1}, O_{t-1}, \ldots, q_1, O_1) = P(q_t|q_{t-1})$$

2. The t^{th} observation, given the t^{th} hidden variable, is independent of all other variables (past and future):

$$P(O_t|q_T, O_T, q_{T-1}, O_{T-1}, \ldots, q_{t+1}, O_{t+1}, q_t, q_{t-1}, O_{t-1}, \ldots, q_1, O_1) = P(O_t|q_t)$$

3.2.1 Elements of an HMM

When $\{O_1, \ldots, O_T\}$ are categorical, an HMM is characterized by the following:

1. The number of states in the model (K). Although the states are hidden, for many practical applications there is often some physical significance attached to the states or to sets of states of the model. The individual states will be denoted as $S = \{S_1, S_2, \ldots, S_K\}$, and the state at time t as q_t.
2. The number of distinct observation symbols per state (M). The observation symbols correspond to the physical output of the system being modeled. The individual symbols will be denoted as $V = \{v_1, v_2, \ldots, v_M\}$.
3. The state transition probability distribution $A = \{a_{ij}\}$ where $a_{ij} = P[q_{t+1} = S_j|q_t = S_i]$, $1 \leq i, j \leq K$. When any state can reach any other state in a single step, we have $a_{ij} > 0$ for all i, j. For other types of HMMs, $a_{ij} = 0$ for one or more (i, j) pairs.
4. The observation symbol probability distribution in state j, $B = \{b_j(m)\}$, where $b_j(m) = P[O_t = v_m|q_t = S_j]$, $1 \leq j \leq K$, $1 \leq m \leq M$. Here $b_j(m) \geq 0$ and $\sum_{m=1}^{M} b_j(m) = 1$.
5. The initial state distribution $\boldsymbol{\pi} = \pi_i$ where $\pi_i = P[q_1 = S_i]$, $1 \leq i \leq K$. $\pi_i \geq 0$ and $\sum_{i=1}^{K} \pi_i = 1$.

Given appropriate values of K, M, A, B and $\boldsymbol{\pi}$, the HMM can be used as a generator to give an observation sequence: $O = O_1 O_2 \dots O_T$, (O_t is one of the symbols from V and T is the number of observations in the sequence) as follows:

1. Choose an initial state $q_1 = S_i$ according to the initial state distribution π
2. Set $t = 1$
3. Choose $O_t = v_m$ according to the symbol probability distribution in state S_i, i.e., $b_i(m)$
4. Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i, i.e., a_{ij}
5. Set $t = t + 1$; return to step 3 if $t < T$; otherwise terminate the procedure

A complete specification of an HMM requires specification of two model parameters (K and M), specification of observation symbols, and the specification of the three probability measures A, B and $\boldsymbol{\pi}$. The complete parameter set of the model will be indicated by $\lambda = (A, B, \boldsymbol{\pi}$.

3.2.2 The three basic problems for HMMs

For the model presented above there are three basic problems of interest that must be solved for the model to be useful in real-world applications:

- Problem 1

  Given the observation sequence $O = O_1 O_2 \dots O_T$, and a model $\lambda = (A, B, \boldsymbol{\pi})$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?

- Problem 2

  Given the observation sequence $O = O_1 O_2 \dots O_T$, and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 \dots q_T$, which is optimal in some meaningful sense (i.e., best "explains" the observations)?

- Problem 3

  How do we select the model parameters $\lambda = (A, B, \boldsymbol{\pi})$ to maximize $P(O|\lambda)$ for the observation sequence O?

Problem 1 is the evaluation problem, how to compute the probability that the observed sequence was produced by the model (scoring how well a given model matches a given observation sequence). In problem 2 the goal is to uncover the hidden part of the model, i.e., find the "correct" state sequence. For practical situations, an optimality criterion is used to solve this problem as best as possible. There are several reasonable optimality criteria that can be imposed, and hence the choice of criterion is a strong function of the intended use for the uncovered state sequence. For problem 3, the objective is to optimize the model parameters so as to best describe how a given observation sequence comes about.

3.2.3 Solutions to the three basic problems for HMMs

3.2.3.1 Solution to problem 1: probability evaluation The most straightforward way of calculating the probability of the observation sequence $O = O_1O_2 \dots O_T$, given the model λ, is through enumerating every possible state sequence of length T. This involves on the order of $2TK^T$ calculations, which is computationally unfeasible, even for small values of K and T. Thus, a more efficient procedure is required to solve Problem 1. Fortunately such a procedure exists and is called the forward-backward procedure. Only the forward procedure is needed to solve Problem 1. Consider the forward variable, defined as the probability of the partial observation sequence $O_1O_2 \dots O_t$, (until time t) and state S_i at time t, given the model λ:

$$\alpha_t(i) = P(O_1O_2 \dots O_t, q_t = S_i|\lambda)$$

Forward procedure: $\alpha_t(i)$ can be solved inductively as follows:

1) Initialization: $\alpha_1(i) = \pi_i b_i(O_1)$, $1 \le i \le K$

2) Induction: $\alpha_{t+1}(j) = \left[\sum_{i=1}^{K} \alpha_t(i)a_{ij}\right] b_i(O_{t+1})$, $1 \le t \le T-1$, $1 \le j \le K$

3) Termination: $P(O|\lambda) = \sum_{i=1}^{K} \alpha_T(i)$

In a similar manner, consider a backward variable, defined as the probability of the partial observation sequence from $t+1$ to the end, given state S_i at time t and the model λ:

$$\beta_t(i) = P(O_{t+1}O_{t+2}\dots O_T|q_t = S_i, \lambda)$$

Backward procedure: $\beta_t(i)$ can be solved inductively as follows:

1) Initialization: $\beta_T(i) = 1,$ $\quad 1 \le i \le K$

2) Induction: $\beta_t(i) = \sum_{j=1}^{K} a_{ij} b_j(O_{t+1})\beta_{t+1}(j),$ $\quad t = T-1, T-2, \dots, 1,$ $\quad 1 \le i \le K$

3.2.3.2 Solution to problem 2: optimal state sequence

The most widely used criterion is to find the single best state sequence, i.e., to maximize $P(Q|O, \lambda)$, which is equivalent to maximize $P(Q, O|\lambda)$. A formal technique for finding this single best state sequence exists, based on dynamic programming methods, and is called the Viterbi algorithm.

Viterbi algorithm: To find the single best state sequence $Q = q_1q_2\dots q_T$, for the given observation sequence $O = O_1O_2\dots O_T$, the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state S_i needs to be defined:

$$\delta_t(i) = \max_{q_1q_2\cdots q_T} P[q_1q_2\dots q_t = i, O_1O_2\dots O_t|\lambda]$$

By induction we have: $\delta_{t+1}(j) = [\max_i \delta_t(i)a_{ij}]b_j(O_{t+1})$. To actually retrieve the state sequence, it is necessary to keep track of the argument which maximized $\delta_{t+1}(j)$, for each t and j. It is done via the array $\psi_t(j)$. The complete procedure for finding the best state sequence can now be stated as follows:

$$
\begin{array}{lll}
\text{1) Initialization:} & \delta_1(i) = \pi_i b_i(O_1), & 1 \leq i \leq K \\
 & \psi_1(i) = 0 & \\
\text{2) Recursion:} & \delta_t(j) = \max\limits_{1\leq i\leq K}[\delta_{t-1}(i)a_{ij}]b_j(O_t), & 2 \leq t \leq T,\, 1 \leq j \leq K \\
 & \psi_t(i) = \arg\max\limits_{1\leq i\leq K}[\delta_{t-1}(i)a_{ij}] & 2 \leq t \leq T,\, 1 \leq j \leq K \\
\text{3) Termination:} & P^* = \max\limits_{1\leq i\leq K}[\delta_T(i)] & \\
 & q_T^* = \arg\max\limits_{1\leq i\leq K}[\delta_T(i)] & \\
\text{4) Path backtracking:} & q_T^* = \psi_{t+1}(q_{t+1}^*) & t = T-1, T-2, \ldots, 1
\end{array}
$$

3.2.3.3 Solution to problem 3: parameter estimation The third, and by far the most difficult, problem of HMMs is to determine a method to adjust the model parameters $(A, B, \boldsymbol{\pi})$ to maximize the probability of the observation sequence given the model. There is no known way to analytically solve for the model which maximizes the probability of the observation sequence. However, $\lambda = (A, B, \boldsymbol{\pi})$ can be chosen such that $P(O|\lambda)$ is locally maximized using an iterative procedure such as the Baum-Welch method (or equivalently the EM algorithm). Rabiner (1989) discusses an iterative procedure, based primarily on the classic work of Baum and his colleagues, for choosing model parameters.

In order to describe the procedure for iterative update and improvement of HMM parameters, the probability of being in state S_i at time t, and state S_j at time $t+1$, given the model and the observation sequence, denoted by $\xi_t(i,j)$ needs to be defined:

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

From the definitions of the forward and backward variables, $\xi_t(i,j)$ can be written in the form:

$$
\begin{aligned}
\xi_t(i,j) &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}
\end{aligned}
$$

where the numerator is just $P(q_t = S_i, q_{t+1} = S_j, O|\lambda)$, and dividing by $P(O|\lambda)$ gives the desired probability. Now, letting $\gamma_t(i)$ represent the probability of being in state S_i at time t, given the observation sequence and the model: $\gamma_t(i) = P(q_t = S_i|O, \lambda)$, $\gamma_t(i)$ can be related to $\xi_t(i,j)$ by summing over j: $\gamma_t(i) = \sum_{j=1}^{K} \xi_t(i,j)$. Now,

$$\sum_{t=1}^{T-1} \gamma_t(i) \qquad \text{can be interpreted as the expected number of transitions from } S_i \text{, and}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) \quad \text{can be interpreted as the expected number of transitions from } S_i \text{ to } S_j$$

Then, a set of reasonable update formulas for $\boldsymbol{\pi}$, A and B are:

$$\pi_i = \gamma_1(i), \qquad a_{ij} = \frac{\displaystyle\sum_{t=1}^{T-1} \xi_t(i,j)}{\displaystyle\sum_{t=1}^{T-1} \gamma_t(i)}, \qquad b_j(m) = \frac{\displaystyle\sum_{t=1, O_t = v_m}^{T} \gamma_t(j)}{\displaystyle\sum_{t=1}^{T} \gamma_t(j)}$$

.

Hence, π_i is the expected frequency in state S_i at time $t = 1$, a_{ij} is the expected number of transitions from state S_i to state S_j over the expected number of transitions from S_i, and $b_j(m)$ is the expected number of times in S_j and observing symbol v_m over the expected number of times in S_j.

3.2.3.4 Parameter estimation for independent sequences

When data on N individuals is available to estimate the parameters in the hidden Markov model, and those N observed sequences are independent of each other, i.e.,

$$P(\mathbf{O}|\lambda) = \prod_{n=1}^{N} P(O_n|\lambda)$$

The update formulas for $\boldsymbol{\pi}$, A and B are:

$$\pi_i = \frac{1}{N}\sum_{n=1}^{N}\gamma_1^{(n)}(i), \qquad a_{ij} = \frac{\displaystyle\sum_{n=1}^{N}\sum_{t=1}^{T_n-1}\xi_t^{(n)}(i,j)}{\displaystyle\sum_{n=1}^{N}\sum_{t=1}^{T_n-1}\gamma_t^{(n)}(i)}, \qquad b_j(m) = \frac{\displaystyle\sum_{n=1}^{N}\sum_{t=1,O_t^{(n)}=v_m}^{T_n}\gamma_t^{(n)}(j)}{\displaystyle\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_t^{(n)}(j)}$$

3.2.4 Implementation issues for HMMs

3.2.4.1 Scaling Since $\alpha_t(i)$ is the sum of a large number of terms, all of them products of probabilities, then as t starts to get big ($t \geq 10$), each term of $\alpha_t(i)$ starts to head exponentially to zero. For sufficiently large t ($t \geq 100$), the dynamic range of the $\alpha_t(i)$ computation will exceed the precision range of essentially any machine. Hence, the only reasonable way of performing the computation is by incorporating a scaling procedure. The basic scaling procedure is to multiply $\alpha_t(i)$ by a scaling coefficient that is independent of i. A similar scaling is done to the $\beta_t(i)$ coefficients, since these also tend to zero exponentially fast, and then, at the end of the estimation of the HMM parameters, the scaling coefficients are canceled out exactly (Rabiner, 1989). The recommended scaling factor for the forward and backward variables performed at every observation time is $1/\sum_{i=1}^{K}\alpha_t(i)$ (Devijver, 1985).

3.2.4.2 Choosing the model size and type Whiting and Pickett (1988) present three information criteria to choose the number of hidden states in an HMM. It assumes an ergodic model, i.e., for the hidden Markov model $\lambda = (A, B, \boldsymbol{\pi})$, with K hidden states and M possible categories for the observed sequences, the elements of B are strictly positive and no elements of A and $\boldsymbol{\pi}$ are constrained a priori to be zero. Hence, the total number of independent parameters is $(K-1)(K+1) + K(M-1)$. Notice that this quantity is a function of K, the number of hidden states, since M is fixed a priori.

Each of the criteria involves minimization of a function of the maximum likelihood and the number of independent model parameters with the following general form:

$$IC(k) = -\max_{\lambda_k}\log Pr(O;\lambda_k) + f(k, L) \tag{3.12}$$

where k denotes the number of independent parameters in λ, and L denotes the observation sample length. The function $f(k, L)$ is a non-decreasing function of k and L. The three criteria are:

a. AIC(k) (Akaike, 1974): $f(k, L) = k$

b. BIC(k) (Schwarz, 1978; Rissanen, 1978): $f(k, L) = \frac{1}{2}k \log L$

c. CIC(k) (Hannan and Quinn, 1979): $f(k, L) = k \log \log L$

4.0 RESULTS

These results were obtained after fitting the models by running codes in the statistical package R, version 2.6.1., in an Intel Core Duo processor with a speed of 2.8 GHz and memory of 4 GB 667 MHz. Due to the complexity of the models and the size of the longitudinal data in COBY, the computation time ranged from few hours to whole days.

4.1 FINDING PATTERNS IN LONGITUDINAL COURSE

4.1.1 Mixture of first order Markov chains

The estimation of the parameters of the mixture of first order Markov chain model was done using the EM algorithm as described in Equations (3.5), (3.6) and (3.7). The statistical package R version 2.6.1. was used to write and run codes to fit this model.

EM initializations were random using a Dirichlet distribution as in Equation 3.3, with $\phi_i = 1$ for: the vector of k mixture proportions, each of the k vectors of 12 initial state probabilities and each row of the k 12×12 transition probability matrices. Between 20 to 40 random initializations were used each time that the EM algorithm was run.

The convergence of the EM algorithm was determined as described in Equation (3.8), with a tolerance of 0.001. The EM algorithm was run for k from 2 to 10, 16 times, each time fixing 20 partitions. In each partition, 50% of the patients were allocated in a training set and the remaining 50% were allocated in a test set. The training set was used to fit the model and the test set was used to evaluate the MCCV and Score criteria. After all the runs, $M = 320$ in Equation (3.10).

The number of free parameters in the model depends on the number of number of components k and number of PSR categories, which is 12. Values of the number of components ranged from 2 to 10. Specifically, for each model the number of free parameters is $(k-1) + k \cdot 11 + k \cdot 12 \cdot 11 = 144k - 1$.

Table 8. presents the summary of the evaluation criteria to choose the number of components in the mixture of first order Markov chains model.

Table 8: Evaluation criteria for assessing the number of components in the mixture of first order Markov chains

k	AIC	BIC	NEC	MCCV*	Score*
2	84126.64	87046.88	0.00107	21716.47 (899.96)	0.88127 (0.03340)
3	81526.07	85911.52	0.00133	21378.57 (908.86)	0.86752 (0.03306)
4	79969.49	**85820.14**	**0.00073**	21298.72 (912.78)	0.86430 (0.03364)
5	78747.95	86063.81	0.00105	21245.95 (946.47)	0.86216 (0.03516)
6	78025.75	86806.81	0.00234	**21198.76** (923.16)	**0.86024** (0.03417)
7	77508.38	87754.65	0.00358	21213.97 (932.72)	0.86084 (0.03424)
8	77093.60	88805.07	0.00252	21208.45 (945.32)	0.86061 (0.03458)
9	76983.37	90160.05	0.00377	21208.86 (957.30)	0.86063 (0.03518)
10	76465.81	91107.70	0.00344	21237.51 (939.48)	0.86180 (0.03448)

*The standard deviation of the estimate of the criterion is given in parentheses

It can be appreciated that AIC does not help to take a decision, since it does not reach a minimum for the values of k. On the other hand, the other criteria are minimized, however not at the same value of k. As can be seen in boldface, BIC and NEC are both minimized for $k = 4$, while the MCCV and the Score criteria suggest a mixture model with 6 components. Therefore there was no consensus in the evaluation criteria to determine the number of components of the mixture of first order Markov chains model.

Furthermore, only the BIC criterion follows a monotone pattern. The normalized entropy criterion (NEC) oscillates between 0.00073 and 0.00377 in the range of k values. MCCV and Score have a flat behavior, and also there are some ups and downs throughout the range of the number of components in the mixture. These findings lead to the selection of the mixture model with four components, as estimated by the BIC. Nevertheless, the models with three or with five components are also reasonable, given that the values for the BIC criterion are close for $k = 3, 4, 5$. However, since the minimum value was attained at $k = 4$, only the results for that model are reported.

4.1.1.1 Mixture of four first order Markov chains Appendix A. contains the parameters estimated for the mixture model with four components that had the largest likelihood among the models with that number of clusters. Tables 23. and 24. show respectively, the vectors of initial state probabilities and the transition probability matrix for the mixture of four components. The percentage of individuals in components 4 through 1 are 69.41, 15.76, 10.71 and 4.12, respectively.

Figures 2., 3., 4. and 5. show the profiles of the patients in each of the four clusters, by bipolar diagnosis. In cluster 1 there are ten, two and five patients respectively from BPI, BPII, and BPNOS. Those numbers are 25, 3 and 16 in cluster 2; 34, 7 and 24 in cluster 3 and the most frequent cluster, cluster 4, has 175 BPI patients, 16 BPII patients and 96 BPNOS patients. From this distribution of bipolar diagnosis in the clusters, it can be concluded that regarding the bipolar diagnosis this model does not help in identifying longitudinal patterns.

However, when looking Figure 2. in more detail, it can be appreciated that cluster 4 is characterized by individuals that move among the twelve episodes in the PSR scale in Table 2 and also have several long stretches on a same PSR category. This observation can be corroborated by the estimates in Table 23., where all twelve PSR categories have a positive initial state probability. Also, in Table 24., where all the elements in the diagonal are greater than 82% and all elements off-the-diagonal less than 6.5%. This cluster could then be labeled as the cluster of the stayers.

Figure 3. shows that cluster 3 is distinguished by spikiness and periods on a same episode during several weeks. From the parameters of this cluster in Table 23., episodes Hypomania-Pure (7), Hypomania/Subdepression (8) and Mania/Subdepression (11) have a very small probability as initial state, and most of the patients started their follow-up with a Subdepression only (2), a Well (3) or a Submania/Subdepression (5) episode.

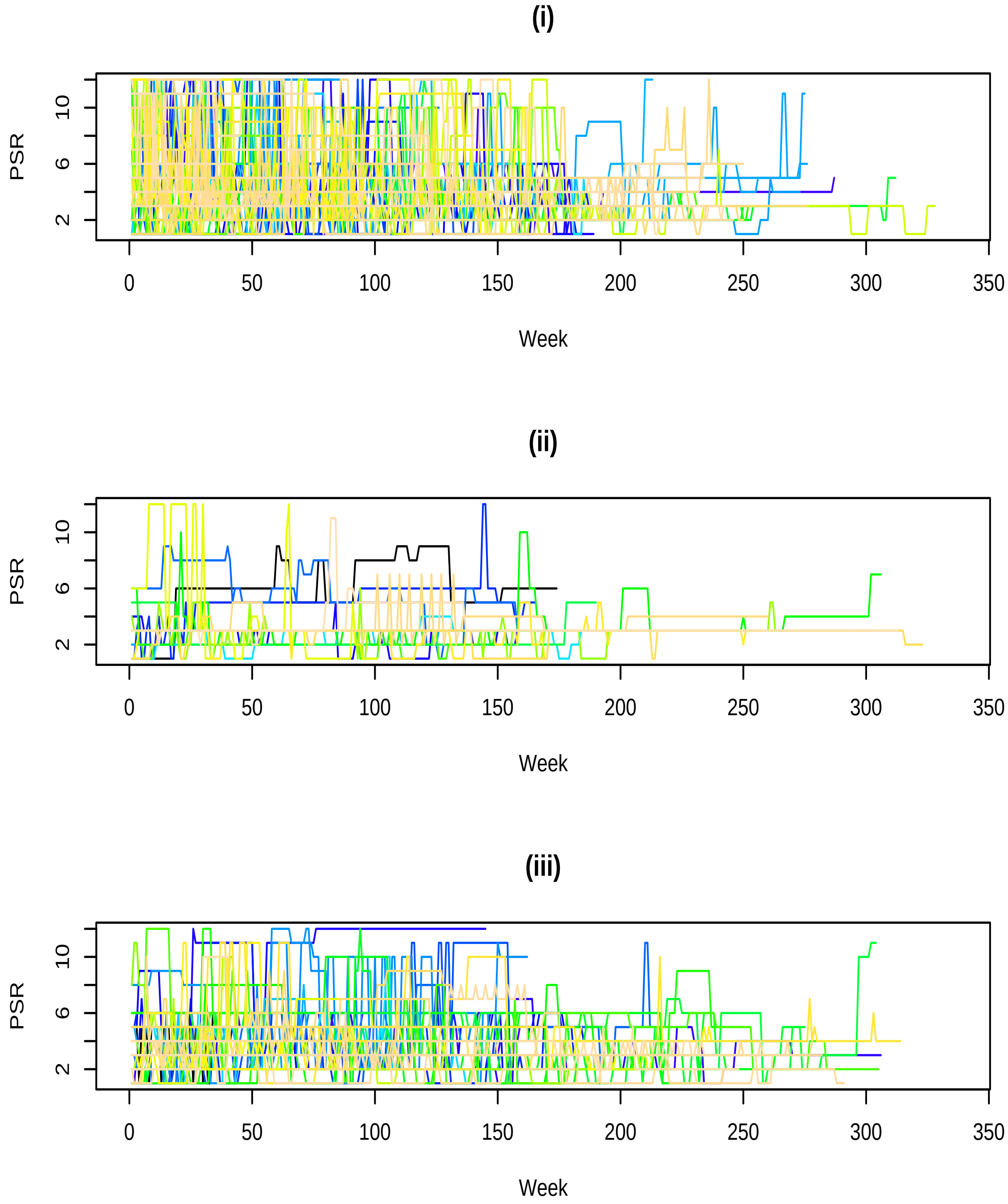
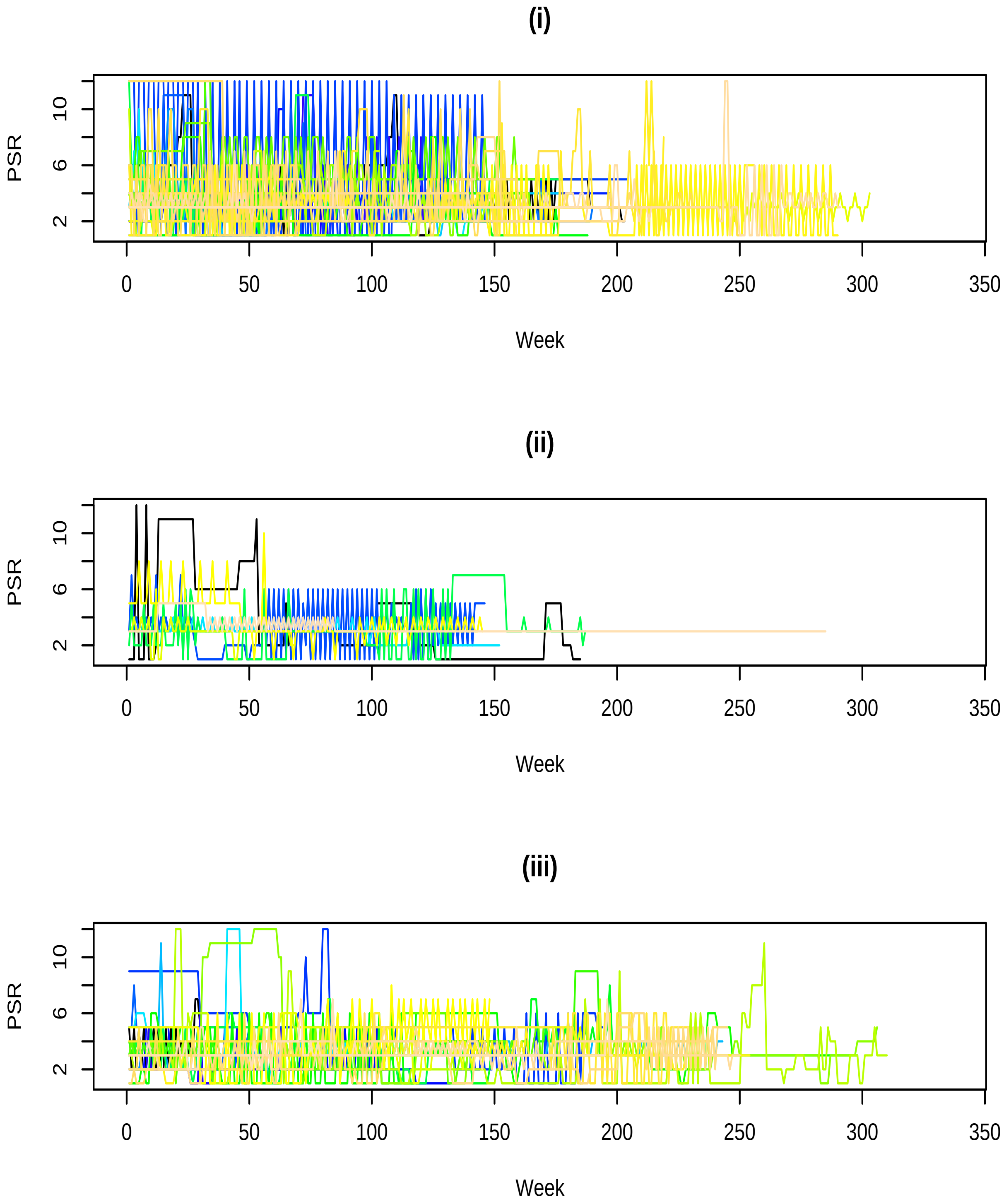


Figure 2: Profile plots of patients in cluster 4 in the mixture of four first order Markov chains by: (i) BPI, (ii) BPII, (iii) BPNOS



(i)

(ii)

(iii)

Figure 3: Profile plots of patients in cluster 3 in the mixture of four first order Markov chains by: (i) BPI, (ii) BPII, (iii) BPNOS

Table 24. shows that the probabilities on the diagonal of the transition matrix vary between 0.433 and 0.835. Among the transitions off-the-diagonal, those with a probability greater than 0.1 are: MDD-Pure (1) to Submania/MDD (6) and viceversa; Subdepression only (2) to Submania/Subdepression (5) and viceversa; Well (3) to Submania only (4) and viceversa; Hypomania-Pure (7) to Well (3) and Submania only (4); Hypomania/Subdepression (8) to Subdepression only (2) and Submania/Subdepression (5); Mania-Pure (10) to Well (3) and Submania only (4); Mania/Subdepression (11) to Subdepression only (2) and Mixed state (12) to MDD-Pure (1).

Note that the transition probabilities are higher when moving from all states to states 1 through 6, therefore the individuals in this cluster are characterized by moving toward a depression state (1 or 2), a well state (3) or a submania state (4, 5 or 6).

Cluster 2 shows even more spikiness and less stays on the same episode for several weeks as can be appreciated in Figure 4. This time episodes Hypomania-Pure (7), Hypomania/Subdepression (8), Mania/Subdepression (11) and Mixed State (12) have small probabilities as initial states, as can be seen in Table 23.

In terms of transitions, Table 24. shows that the probabilities of staying on the same episode are between 30.0% and 86.3%. Off-the-diagonal probabilities greater than 10% are observed for entering a Well (3) state from: (1) MDD-Pure, (2) Subdepression only, (4) Submania only, (7) Hypomania-Pure, (8) Hypomania/Subdepression, (10) Mania-Pure and (11) Mania/Subdepression. Also for entering a Subdepression only (2) state from: (3) Well, (8) Hypomania/Subdepression and (10) Mania-Pure. Other high probabilities are observed from transitions Submania only (4) to Submania/Subdepression (5), Mania-Pure (10) to MDD-Pure (1) and Mania/Subdepression (11) to Mixed State (12). It is also notorious than in this cluster, when a patient presents a Mania/Subdepression episode, he/she experiences a Well (3) or a Mixed state (12) episode in the following week. In conclusion, cluster 2 is characterized mainly for patients who move but also spend several weeks in the Well state.

Figure 5. presents even a spikier pattern for patients classified in cluster 1. This cluster contains patients who have several periods of frequent ups and downs. For cluster 1, Table 24. indicates that states involving MDD (1, 6 and 9) and the mania states (10 and up) have an almost null probability of being initial states.

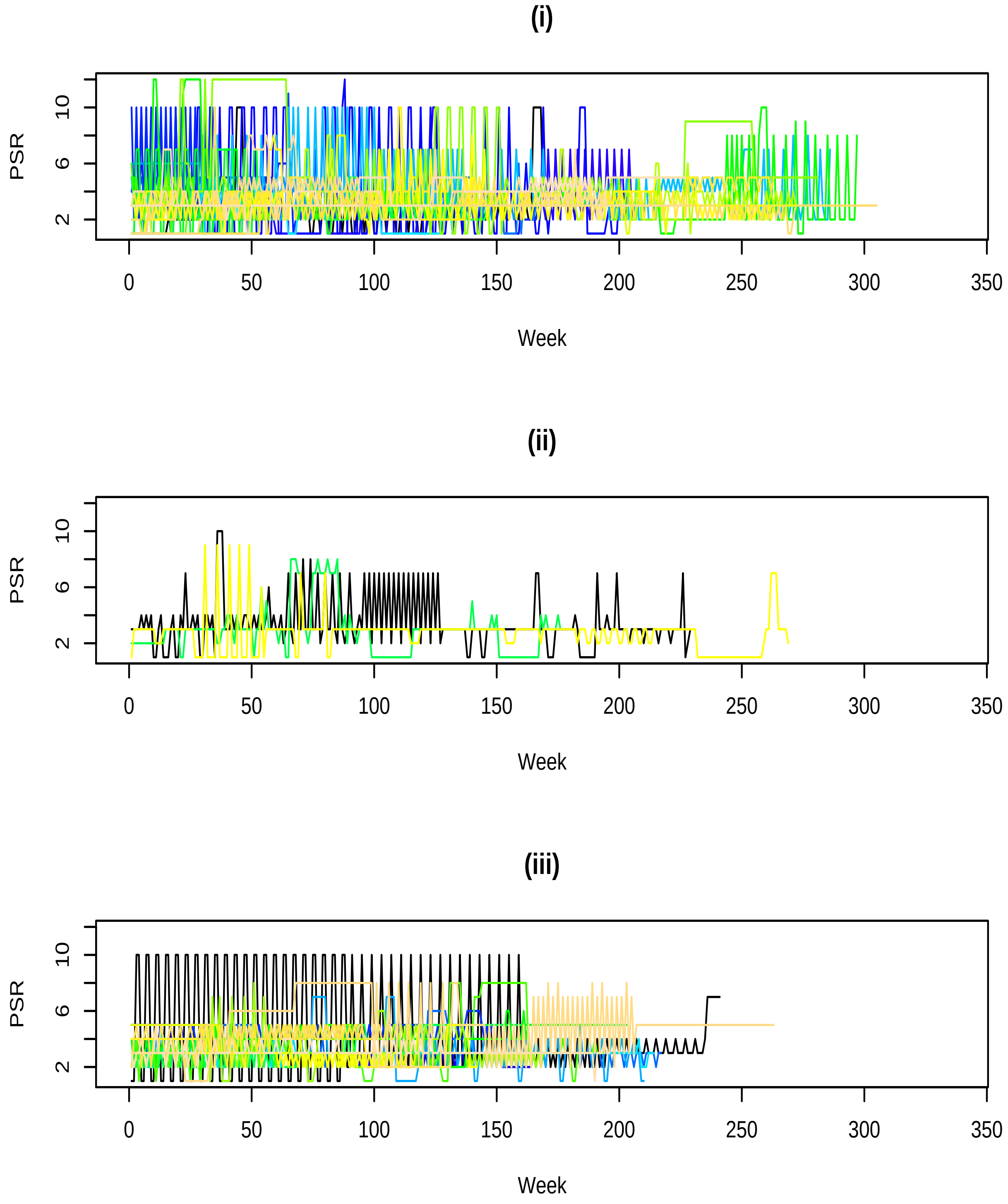


Figure 4: Profile plots of patients in cluster 2 in the mixture of four first order Markov chains by: (i) BPI, (ii) BPII, (iii) BPNOS

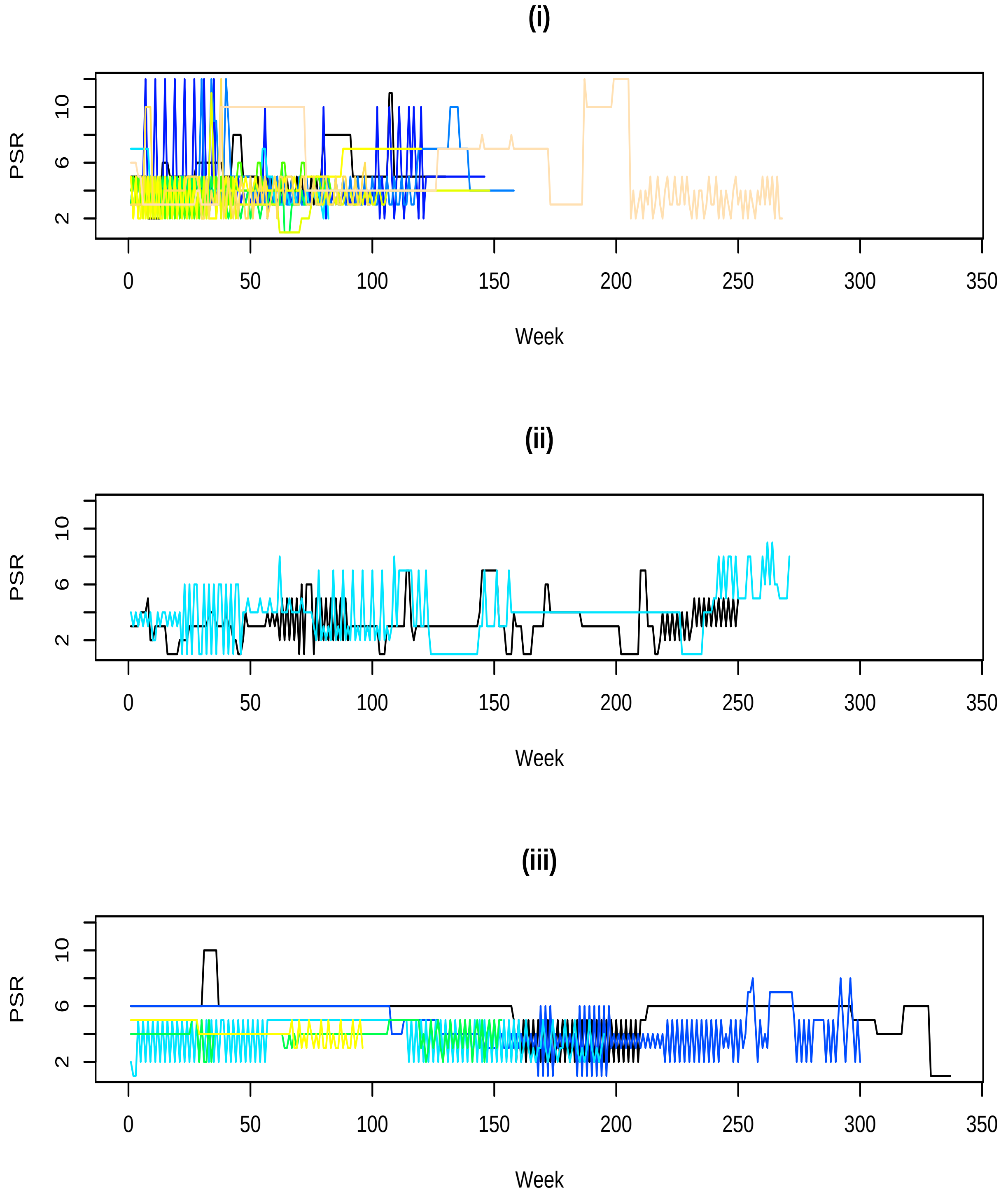


Figure 5: Profile plots of patients in cluster 1 in the mixture of four first order Markov chains by: (i) BPI, (ii) BPII, (iii) BPNOS

Table 24. shows that the transitions in the diagonal vary from 9% for Subdepression only (2) to 90.3% for Submania/MDD (6). This time there are more transitions off-the-diagonal with entries greater than 10%, the one for Mania/Subdepression (11) to Submania/Subdepression (5) is the largest, estimated at 66.5%. From MDD-Pure (1), the most frequent transition is to Submania/MDD (6). In the same sense, other transitions that should be highlighted are: Subdepression only (2) and Well (3) to Submania/Subdepression (5) and viceversa, Hypomania with Subdepression and with MDD (8 and 9) to Submania/Subdepression (5), Hypomania/MDD (9) to Submania/MDD (6) and Mixed state (12) to Well (3). These individuals clearly transition more than those in the previous three clusters. These are movers, who tend to transition more often to a Subdepression only (2) state, a Well state or a Submania/Subdepression (5) state, and stay in those three states for several weeks.

To end the description of the findings of the mixture of four first order Markov chains model, when looking to the three plots labeled (i), (ii) and (iii) in each figure of the four clusters, it is observed that the range of PSR categories differs by bipolar diagnosis. That shows that even though the clusters are identifying similar characteristics (spikiness and/or long stretches on the same episode) for the bipolar longitudinal courses of children and adolescents, the range of episodes varies by bipolar diagnosis.

4.1.2 Hidden Markov models

Hidden Markov models were estimated as described in Section 3.2, which was found implemented in the RHmm package in R (Taramasco, 2007). The number of hidden states k was initially varied from 2 to 10, but since 10 was found by the evaluation criteria to be the best hidden Markov model, the value of k was extended up to 12. The estimation was done several times for each k, with tolerance 0.00001 for stopping the algorithm and with 100 random initializations for the parameters of each HMM. For the evaluation criteria in COBY, $L = \sum_{n=1}^{413} T_n = 71,328$ in Equation (3.12).

Table 9. reports the information criteria BIC and CIC. The model that maximizes both of these criteria is the one with ten hidden states, as is highlighted with boldface.

Table 9: Evaluation criteria for assessing the number of components in hidden Markov models

k	BIC	CIC
2	186350.47	186191.77
3	151238.58	150978.32
4	129568.66	129194.15
5	111507.58	111006.11
6	100680.24	100039.12
7	96342.06	95548.60
8	89511.71	88553.21
9	89627.66	88491.42
10	**82994.46**	**81667.79**
11	87029.61	85499.82
12	84628.94	82883.33

The parameters of the best hidden Markov model with ten hidden states expressed as percentages are reported in Table 10. The initial state probabilities of the hidden states are presented in the first row, followed by the transition probability matrix among the hidden states, and finally the distribution of the PSR categories conditional on each of the hidden states are given in each of the columns in the last portion of the table.

Birmaher et al. (2006) reported results in eight categories, that was derived by grouping the episodes in the twelve PSR scale as follows:

- Asymptomatic: 3
- DSM-IV syndromal episode:
  - Pure MDD: 1
  - Pure mania/hypomania: 7, 10
  - Pure mixed: 12
  - Cycling: 6, 8, 9, 11
- Subsyndromal episode:
  - Subsyndromal pure depression: 2
  - Subsyndromal pure mania: 4
  - Subsyndromal Mixed: 5

From this classification, and the conditional distributions of observed PSR on the ten hidden states in Table 10.—where each of the hidden states can be characterized by the most frequent PSR category in the state as highlighted in boldface—the ten hidden states can be labeled as:

- State 1: DSM-VI syndromal hypomania/mania (PSR's 7 and 10: Hypomania pure and Mania pure)
- State 2: Subsyndromal pure mania (PSR 4: Submania only)
- State 3: DSM-VI syndromal cycling (PSR's 8 and 9: Hypomania/Subdepression and Hypomania/MDD)
- State 4: DSM-VI syndromal cycling (PSR 11: Mania/Subdepression)
- State 5: Subsyndromal pure depression (PSR 2: Subdepression only)
- State 6: DSM-VI syndromal cycling (PSR 6: Submania/MDD)
- State 7: (PSR's 2, 3, 4: Subdepression only, Well, Submania only)
- State 8: DSM-VI syndromal MDD and Pure mixed(PSR's 1 and 12: MDD pure and Mixed state)

- State 9: Asymptomatic (PSR 3: Well)
- State 10: Subsyndromal mixed (PSR 5: Submania/Subdepression)

There is a striking similarity between these two classifications. States 1, 2, 5, 9 and 10 in the hidden Markov model with ten hidden states are identified as proposed by the psychiatrists. States 3, 4 and 6 correspond to the group labeled as cycling by the psychiatric team. State 8 fusions the MDD-Pure and Mixed state categories and State 7 which does not appear in the psychiatrists classification, is a mixture of subsyndromal pure depression, asymptomatic and subsyndromal pure mania.

Now that the hidden states had been labeled, it can be observed from the distribution of the initial state in Table 10. that the asymptomatic state (9) is the most frequent. Thus, most of the patients start the follow-up without bipolar symptoms. The least frequent states is 4, one of the syndromal cycling states.

Note that the transition probability matrix in Table 10, shows probabilities over 88% on the diagonal. Off the diagonal, only transitions from syndromal cycling (3 and 4) to subsyndromal mixed (10) have probabilities over 5%.

Table 10: Parameters of the HMM with ten hidden states (in percentages)

Hidden State		1	2	3	4	5	6	7	8	9	10
Initial state probabilities		5.9	11.3	1.9	2.6	9.1	5.3	10.6	13.6	23.9	15.9
Transition probability matrix	1	**91.0**	3.7	0.4	0.4	0.4	0.1	0.8	0.4	2.4	0.4
	2	0.6	**94.3**		0.1	0.3	0.5	0.3	0.2	2.5	1.3
	3	0.7	0.1	**90.5**	0.7	1.0	0.5	0.1	0.5	0.6	**5.4**
	4	1.0	1.0	1.1	**88.0**	0.7	0.5	0.1	1.6	0.5	**5.5**
	5	0.1	0.3	0.1		**92.5**	0.2	0.4	1.6	3.8	1.1
	6	0.2	1.6	0.4		0.5	**91.1**		1.6	0.3	4.2
	7	0.3	0.4			0.5	0.1	**96.0**	1.2	1.2	0.3
	8	0.3	0.3	0.2	0.1	3.3	0.9	1.8	**90.3**	2.3	0.6
	9	0.3	0.6			1.1		0.3	0.5	**96.9**	0.2
	10		1.2	0.4	0.4	0.9	1.2	0.2	0.4	0.6	**94.6**
Conditional distribution of observed PSR on hidden states	1	2.2				0.1		0.6	**75.1**		0.1
	2	2.7				**94.3**		**12.0**	0.1	0.1	
	3	3.0		0.3	0.2		0.1	**59.7**	0.1	**99.6**	
	4		**96.7**					**20.1**		0.2	0.1
	5		3.1	1.3	0.5	4.8		4.3			**99.6**
	6						**99.5**	0.2	5.7		
	7	**54.8**	0.1					2.1			
	8	1.2		**65.5**		0.6		0.2			0.3
	9	0.1		**31.7**			0.2		0.2		
	10	**36.0**	0.1					0.6	1.7		
	11	0.1		0.1	**98.6**	0.2					
	12			1.0	0.7		0.2	0.2	**17.1**		

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

Assuming that the underlying process is a Markov process, this model can be represented using the transition intensity matrix $\mathbf{U}$:

$$\begin{pmatrix} -\sum_{h\neq 1} u_{1h} & u_{12} & \dots & u_{1l} \\ u_{21} & -\sum_{h\neq 2} u_{2h} & \dots & u_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ u_{l1} & u_{l2} & \dots & -\sum_{h\neq l} u_{lh} \end{pmatrix}$$

A relation between: the matrix of transition probabilities of the hidden process over a time interval t, $\mathbf{P}(t)$, and the transition intensity matrix $\mathbf{U}$ can be established with the Kolmogorov forward differential equations

$$\frac{\partial \mathbf{P}(t)}{\partial t} = \mathbf{P}(t)\mathbf{U} \tag{4.1}$$

where the element (i,j) in $\mathbf{P}(t)$ represents the probability of a transition from the state i to the state j in a time interval t, denoted as $p_{ij}(t)$. A solution to this system of differential equations can be expressed as

$$\mathbf{P}(t) = \mathbf{A}\mathrm{diag}\{e^{\rho_1^t}, e^{\rho_1^t}, \dots, e^{\rho_k^t}\}\mathbf{A}^{-1} \tag{4.2}$$

where $\mathbf{A}$ is the square matrix containing in column i the eigenvector associated with the eigenvalue ρ_i of the transition intensity matrix $\mathbf{U}$ (Marshall and Jones, 1995).

The model can be extended to introduce covariates as a proportional factor in the baseline transition intensities. Thus, the element (i,j) of the transition intensity matrix $\mathbf{U}$ is represented as

$$u_{ij}(\mathbf{z}) = u_{ij} \exp \beta'_{ij}\mathbf{z} \tag{4.3}$$

where β_{ij} is the vector of coefficients associated with the vector of covariates $\mathbf{z}$ for the transition between states i and j. Equation (4.3) for the transition intensity $u_{ij}(\mathbf{z})$ resembles the proportional hazard model with constant hazard function. The resulting transition intensity matrix $\mathbf{U}(\mathbf{z})$ for a subject with vector of covariates $\mathbf{z}$ in Equations (4.1) and (4.2) to compute the transition probability matrix $\mathbf{P}(t|\mathbf{z})$.

The `msm` package (Jackson, 2007) in R allows to fit hidden Markov models with covariates in the hidden states, as described above. In the following sections, the effect of gender, age, cohabitation, bipolar diagnosis, age of bipolar onset and socio-economic status in the hidden Markov model with ten states is analyzed separately, i.e., a model for each covariate.

4.2.1 Hidden Markov model with gender as covariate

Table 11. contains the parameters estimated for the hidden Markov model with 10 hidden states when gender is included in the analysis. When gender is included in the model, the estimates of the initial states probabilities even out throughout the hidden states.

Table 11: Parameters of the HMM with ten hidden states with gender as covariate (in percentages)†

Hidden State		1	2	3	4	5	6	7	8	9	10
Initial state probabilities		10.6	10.1	7.8	9.9	9.4	8.6	10.5	9.4	12.4	11.3
Transition probability matrix-Females	1	**65.9**	2.7	1.6	3.3	**5.2**	3.1	**5.6**	4.9	3.7	3.9
	2	0.7	***97.3***	0.5	0.2	0.1	0.2	0.2	0.3	0.4	0.1
	3	0.4	1.4	**57.5**	1.8	**9.5**	2.7	**6.5**	**5.9**	**5.8**	**8.4**
	4	0.2	3.3	1.7	**77.6**	3.1	2.4	3.0	3.6	2.3	2.7
	5	0.2	0.7	0.2	0.1	***95.4***	0.6	1.3	0.2	1.0	0.3
	6	0.5	2.3	1.2	1.6	1.6	***83.9***	1.8	2.4	1.3	3.5
	7	0.7	0.8	0.7	1.0	1.6	0.6	***93.4***	0.6	0.1	0.5
	8	0.5	1.3	0.9	1.4	2.5	1.3	1.7	***88.5***	0.9	1.0
	9	1.0	0.8	0.2						***97.8***	
	10	0.5	1.2	0.5	0.7	0.9	0.1	0.4	0.1	0.2	***95.4***

Table 11: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Transition probability matrix-Males	1	**60.3**	2.8	1.7	3.9	**5.5**	3.9	**6.5**	**5.5**	4.6	**5.4**
	2	0.8	***94.6***	0.8	0.5	0.4	0.6	0.6	0.6	0.9	0.2
	3	0.4	1.5	**54.1**	1.8	**9.4**	2.8	**7.6**	**6.5**	**7.3**	**8.5**
	4	0.2	3.7	2.4	**71.2**	3.5	2.9	4.1	4.0	3.9	4.2
	5	0.2	1.2	0.6	0.5	***90.5***	1.0	2.0	0.6	2.3	1.0
	6	0.5	2.8	1.8	2.4	3.1	**75.0**	2.8	3.5	3.2	4.8
	7	0.7	0.9	1.2	1.3	2.6	1.2	***89.2***	1.3	0.2	1.3
	8	0.6	2.0	1.4	2.2	4.2	2.4	3.3	**78.7**	3.0	2.5
	9	1.0	0.8	0.3	0.1	0.1		0.1	0.1	***97.5***	0.1
	10	0.8	1.3	0.6	0.9	1.4	0.3	0.9	0.3	0.6	***92.9***
Conditional distribution of observed PSR on hidden states	1	2.6				0.1		0.5	**76.4**		0.1
	2	4.3				**94.1**		**18.3**	0.1	0.1	
	3	4.4		0.3	0.2		0.1	**61.7**	0.1	**99.6**	
	4		**96.6**					**13.5**		0.2	0.1
	5		3.1	1.3	0.5	5.1		3.3			**99.6**
	6						**99.5**	0.2	4.4		
	7	**80.4**	0.1					1.6			
	8	1.7		**65.9**		0.5		0.1			0.2
	9	0.1		**31.3**			0.2		0.2		
	10	6.2	0.1					0.6	1.3		
	11	0.2		0.1	**98.7**	0.2					
	12			1.0	0.7		0.1	0.2	**17.4**		

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

Note that this time there are two transition matrices, one for females and one for males. In both of them, probabilities over 80% are in boldface and italics. Probabilities above 5% and below 80% are in boldface only. In both matrices, states 1 (Syndromal hypomania/mania) and 3 (Syndromal cycling) are the most affected with the inclusion of gender, because the probabilities of these states in the diagonal are between 54% and 66% and they also have several transitions off the diagonal with probabilities above 5%. When comparing the two transition probability matrices, note that the transitions on the diagonal are consistently smaller for males. That indicates that males tend to transition more than females. The conditional distribution of the observed PSR on the hidden states is similar to the one estimated for the HMM without covariates. The only difference is for hidden states 1 and 7. Hidden state 1 is mainly made up of PSR 7. And hidden state 7 has a different distribution that the one observed for the HMM in Table 10, but it is made up of the same mixtures of PSRs.

The observation above, about males moving more than females, is corroborated with the mean sojourn times in Table 12. "Sojourn time" refers to the total lengths of all PSR categories through the longitudinal course for a single patient. "Mean sojourn time" is the average sojourn times for all patients. For hidden states 2, 5 and 8, the mean sojourn time of female is double than the one for males. And consistently, the mean sojourn times in all other hidden states are always greater for females. That shows that males tend to spend less time in each of the states before moving to another state.

Table 12: Mean sojourn times by gender

Gender	1	2	3	4	5	6	7	8	9	10
Female	2.4	36.2	1.8	3.9	21.1	5.7	14.5	8.1	44.8	21.2
Male	2.0	17.7	1.6	2.9	9.8	3.4	8.5	4.1	38.2	13.3

In Appendix B, Figures 6. and 7. show the prevalence estimated with the hidden Markov model with ten hidden states for females and males, respectively. For females, the model overestimates state 2 (subsyndromal mania) during the whole follow-up span, and underestimates state 9 (asymptomatic). The observed and estimated prevalences agree in all other states. For males, the model seems to do better, because the discrepancies in states 2 and 9, between the observed and estimates prevalences, are smaller than those observed for females.

4.2.2 Hidden Markov model with age as covariate

Table 13. contains the estimates of the hidden Markov model with ten hidden states when including age as covariate. Age was measured in years, but for this analysis, it was categorized as 1 for age 13 or more and 0 otherwise. Thus, 1 corresponds to teenagers and 0 to children. The distribution of the initial state is almost uniform. The transition probability matrices show that children move more than teenagers do. All ten probabilities in the diagonal are consistently smaller for children than they are for teenagers.

Table 13: Parameters of the HMM with ten hidden states with age as covariate (in percentages)†

Hidden State		1	2	3	4	5	6	7	8	9	10
Initial state probabilities		12.4	9.9	7.9	9.8	9.3	8.6	10.1	9.3	11.9	10.8
Transition probability matrix-Children	1	**59.3**	0.7	1.2	4.1	**5.1**	4.6	**6.9**	4.2	**7.2**	**6.7**
	2	0.8	***93.3***	0.9	0.5	0.8	0.7	1.5	0.3	0.8	0.4
	3	1.5	4.5	**51.3**	2.7	**5.9**	4.5	**5.8**	4.3	**8.8**	**10.7**
	4	2.5	2.7	0.9	**68.1**	3.7	2.9	**5.0**	3.3	**5.1**	**5.8**
	5	0.7	1.3	0.9	1.0	***89.8***	0.7	2.2	1.0	1.7	0.7
	6	0.2	2.0	0.8	2.5	3.2	**74.5**	3.7	3.4	4.0	**5.7**
	7	1.6	0.8	1.2	1.7	2.1	1.5	***86.7***	2.2	0.7	1.6
	8	0.3	1.6	1.8	2.1	**5.0**	2.5	4.2	**75.9**	3.6	3.1
	9	0.3	0.9		0.2	0.7	0.1	0.7		***97.1***	
	10	0.8	1.5	0.8	0.7	1.5	1.3	1.2	0.6	1.0	***90.6***

Table 13: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Transition probability matrix-Teenagers	1	**67.3**	0.6	1.1	3.6	4.0	3.2	**5.9**	4.1	**5.6**	4.6
	2	0.7	***96.3***	0.5	0.2	0.4	0.2	1.2	0.1	0.3	0.1
	3	1.5	4.4	**54.5**	2.7	**5.5**	4.0	**5.3**	4.3	**7.5**	**10.3**
	4	2.1	2.7	0.8	**76.0**	2.8	2.3	3.4	3.6	3.0	3.3
	5	0.6	0.8	0.4	0.8	***94.8***	0.2	1.3	0.3	0.6	0.2
	6	0.2	1.9	0.7	1.7	2.1	***84.3***	1.8	2.6	1.9	2.7
	7	1.1	0.7	0.6	1.0	1.5	0.9	***92.3***	1.2	0.2	0.4
	8	0.3	1.3	0.8	1.6	2.6	1.6	2.2	***86.8***	1.7	1.0
	9	0.1	0.7			0.2		0.2		***98.8***	
	10	0.6	1.4	0.5	0.3	1.2	0.9	0.6	0.3	0.5	***93.7***
Conditional distribution of observed PSR on hidden states	1	2.7				0.1		0.6	**76.1**		0.1
	2	3.7				**94.3**		**17.6**	0.1	0.1	
	3	4.1		0.3	0.1		0.1	**64.6**	0.1	**99.6**	
	4		**96.8**					**11.0**		0.2	0.1
	5		2.9	1.3	0.5	4.9		3.3			**99.6**
	6						**99.4**	0.2	4.7		
	7	**79.6**	0.1					1.7			
	8	1.7		**65.9**		0.5		0.2			0.2
	9	0.1		**31.3**			0.2		0.2		
	10	7.8	0.1					0.6	1.4		
	11	0.2		0.1	**98.7**	0.2					
	12			1.0	0.7		0.1	0.2	**17.3**		

[†] An empty cell means that the estimated probability expressed in percentage is less than 0.1%

Also, for children, hidden states 1, 3, 4, and 6 (syndromal hypomania/mania and the three syndromal cycling states) have several transitions off the diagonal with probabilities over 5%. Below the diagonal the transition (8,5), syndromal MDD and Pure mixed to subsyndromal pure depression, has probability 5%. The conditional distribution of the PSR on the hidden state differs from the one estimated for the HMM with ten hidden states without covariates in hidden states 1 and 7.

The mean sojourn times of children are consistently smaller than the ones observed for teenagers, as seen in Table 14. In hidden states 2, 5, 7, 8, and 9 the mean sojourn times are almost double for children. States 1 and 3, where children and teenagers have similar behavior in the transition probability matrices, show similar mean sojourn times for children and adolescents.

Table 14: Mean sojourn times by age

Gender	1	2	3	4	5	6	7	8	9	10
Children	1.9	14.3	1.5	2.6	9.1	3.4	6.8	3.6	33.7	9.9
Teenagers	2.5	26.4	1.6	3.6	18.4	5.8	12.2	7.0	79.9	15.2

The goodness of fit of this model can be assessed with Figures 8. and 9. in Appendix B. The model for children does not seem to rise any objections. On the other hand, the model for teenagers shows overestimation for state 2 in the whole follow-up range, and underestimation for hidden state 10 between weeks 1 and 170.

4.2.3 Hidden Markov model with cohabitation as covariate

When covariate cohabitation (lives with both natural parents: yes or no) is considered in the hidden Markov model with ten hidden states, it is found that the initial states distribution gives the highest probability to hidden state 1, as can be seen in Table 15. Even though the effect of cohabitation seems the same for the two categories, it is observed a consistent pattern in the diagonal of the transition probability matrices. Those probabilities are smaller for the patients who live in another situation different to living with both natural parents.

This suggests that patients who live in another situation move more among the bipolar episodes.

Again the conditional distribution of the PSR on the hidden states differs for hidden states 1 and 7, from the ones estimated for the HMM without covariates.

Table 15: Parameters of the HMM with ten hidden states with cohabitation as covariate (in percentages)†

Hidden State		1	2	3	4	5	6	7	8	9	10
Initial state probabilities		13.9	9.6	7.5	9.6	9.0	8.3	10.0	9.1	12.1	10.8
Transition probability matrix-Parents	1	**63.0**	4.0	1.6	3.7	3.8	2.7	**6.0**	**6.7**	4.8	3.9
	2	0.2	***97.6***	0.1	0.4	0.4	0.5	0.1	0.4	0.1	0.1
	3	1.1	1.5	**58.9**	4.3	**7.2**	2.9	**5.8**	**5.0**	**5.4**	**8.0**
	4	1.1	2.1	1.9	**77.1**	3.0	2.2	3.5	3.5	2.7	3.0
	5	0.4	0.7	0.4	0.7	***94.4***	0.1	0.7	0.2	2.1	0.3
	6	0.5	2.0	1.0	1.7	2.0	***84.1***	2.0	2.2	2.0	2.5
	7	0.8	0.7	0.6	0.6	1.3	0.6	***93.9***	1.0	0.1	0.3
	8	1.2	0.7	0.9	1.5	2.8	1.5	2.1	***87.4***	1.2	0.8
	9	0.7	0.6	0.3		0.1				***98.0***	
	10	0.4	1.2	0.6	0.3	0.6	0.8	0.6	0.2	0.4	***94.9***
Transition probability matrix-Other	1	**60.8**	4.2	1.7	4.2	4.1	3.0	**5.8**	**6.3**	**5.3**	4.7
	2	0.2	***95.2***	0.2	0.8	0.8	0.9	0.4	0.7	0.4	0.4
	3	1.1	1.5	**54.9**	4.7	**7.6**	3.1	**6.3**	4.8	**6.2**	**9.8**
	4	1.1	2.1	2.3	**70.7**	4.0	2.8	4.4	4.0	3.9	4.6
	5	0.4	1.0	0.9	1.2	***89.8***	0.5	1.5	0.7	2.9	1.1
	6	0.5	2.0	1.2	2.1	2.7	**78.1**	2.9	3.3	2.8	4.3
	7	1.0	1.4	1.0	1.4	2.1	1.2	***88.7***	1.8	0.3	1.2
	8	1.8	0.9	1.5	1.8	3.7	2.5	3.2	***80.1***	2.4	2.2
	9	0.8	0.8	0.4	0.1	0.4		0.1	0.1	***97.2***	0.1

Table 15: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
	10	0.4	1.3	0.8	0.6	1.0	1.2	1.0	0.5	0.7	***92.4***

Table 15: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Conditional distribution of observed PSR on hidden states	1	2.6				0.1		0.5	**76.5**		0.1
	2	3.7				**93.7**		**20.1**	0.1	0.1	
	3	4.2		0.3	0.2		0.1	**57.9**	0.1	**99.6**	
	4		**96.5**					**15.9**		0.2	0.1
	5		3.2	1.3	0.5	5.5		3.1			**99.6**
	6						**99.5**	0.2	4.3		
	7	**81.1**	0.1					1.6			
	8	1.7		**66.0**		0.5		0.2			0.2
	9	0.1		**31.2**			0.2		0.2		
	10	6.4	0.1					0.5	1.3		
	11	0.2		0.1	**98.6**	0.2					
	12			1.0	0.7		0.2	0.1	**17.4**		

[†] An empty cell means that the estimated probability expressed in percentage is less than 0.1%

Observing the mean sojourn times of cohabitation in Table 16., we see that the times are consistently higher for patients who live with both natural parents. Furthermore, for states 2, 5, 7 and 8, patients who live with both natural parents spent double the number of weeks than those patients who live in another situation. Similar mean sojourn times are observed for both categories in states: 1, 3, 4 and 6, (the syndromal hypomania/mania and syndromal cycling states).

Plots of the prevalences for the two categories of cohabitation are in Appendix B. Figure 10. and 11 show that the model overestimates states 2, and underestimates states 9 from week 150, for patients who live with both natural parents, and for patients who live in another situation.

Table 16: Mean sojourn times by cohabitation

Cohabitation	1	2	3	4	5	6	7	8	9	10
With both natural parents	2.2	40.1	1.9	3.8	17.1	5.8	15.7	7.3	49.0	18.9
Other situation	2.0	20.0	1.7	2.9	9.2	4.0	8.2	4.4	34.3	12.4

4.2.4 Hidden Markov model with bipolar diagnosis as covariate

When bipolar diagnosis is included in the hidden Markov model with ten hidden states, the initial state distribution ranges from 0.074 (hidden state 3) and 0.139 (hidden state 1), as appears in Table 17.

Table 17: Parameters of the HMM with ten hidden states with bipolar diagnosis as covariate (in percentages)†

Hidden State		1	2	3	4	5	6	7	8	9	10
Initial state probabilities		13.9	9.7	7.4	9.6	9.0	8.2	10.3	9.0	12.1	10.9
Transition probability matrix-BPI	1	**72.9**	2.1	0.3	2.9	2.8	2.3	**5.0**	**5.1**	3.5	3.0
	2	0.5	***96.8***	0.3	0.8	0.4	0.1	0.4	0.1	0.6	0.1
	3	0.4	0.9	**65.0**	4.3	4.7	3.1	4.7	3.9	4.2	**8.9**
	4	0.1	2.3	1.6	***82.2***	2.2	2.0	2.6	3.1	1.4	2.6
	5	0.7	0.6	0.7	0.8	***94.3***	0.3	1.1	0.2	0.8	0.5
	6	0.7	1.7	1.3	1.7	1.6	***84.7***	1.8	2.2	1.4	2.8
	7	0.8	0.6	0.6	0.6	0.9	0.6	***94.2***	1.0	0.1	0.6
	8	0.9	0.8	0.9	1.5	2.3	1.4	1.6	***89.1***	0.5	1.0
	9	0.5	0.7	0.2		0.5				***98.0***	
	10	0.5	0.7	0.5	0.7	0.6	0.4	0.4	0.1	0.4	***95.7***

Table 17: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Transition probability matrix-BPII	1	**61.4**	2.1	0.4	4.0	**5.0**	2.9	**6.9**	**6.8**	**5.3**	**5.1**
	2	0.5	***93.5***	0.7	0.9	0.9	0.3	1.0	0.4	1.4	0.5
	3	0.4	1.0	**58.6**	4.4	**5.7**	3.5	**5.6**	4.0	**6.4**	**10.4**
	4	0.2	2.4	2.6	**67.5**	3.7	3.7	4.6	**5.8**	4.2	**5.3**
	5	0.9	0.7	0.8	1.0	***90.1***	0.7	2.1	0.6	1.9	1.1
	6	0.8	1.6	1.7	2.0	2.3	***80.0***	2.4	3.4	2.1	3.8
	7	1.3	0.7	0.9	1.1	1.7	1.1	***89.6***	2.2	0.2	1.4
	8	0.9	0.9	1.2	1.7	3.0	2.1	2.6	***84.6***	1.3	1.7
	9	0.7	0.9	0.4	0.1	1.0		0.2	0.1	***96.6***	0.1
	10	0.6	1.3	0.9	0.9	1.4	1.0	1.0	0.4	1.0	***91.5***
Transition probability matrix-BPNOS	1	**60.4**	2.1	0.5	4.4	**5.2**	3.0	**6.6**	**6.6**	**5.5**	**5.7**
	2	0.5	***93.2***	0.7	0.9	1.0	0.4	1.0	0.4	1.4	0.5
	3	0.4	1.0	**56.5**	4.8	**6.1**	3.5	**5.7**	3.9	**7.2**	**10.9**
	4	0.2	2.4	2.5	**67.8**	3.7	3.7	4.5	**5.5**	4.0	**5.5**
	5	0.9	0.8	0.9	1.2	***89.0***	0.9	2.2	0.7	2.0	1.4
	6	0.8	1.6	1.7	2.2	2.4	**78.6**	2.5	3.5	2.3	4.4
	7	1.4	0.7	1.0	1.4	1.9	1.4	***87.7***	2.3	0.2	1.9
	8	0.9	0.9	1.4	2.1	3.4	2.6	3.3	***81.4***	1.7	2.4
	9	0.7	1.1	0.4	0.1	1.1		0.3	0.1	***96.1***	0.1
	10	0.6	1.4	0.9	0.9	1.4	1.1	1.1	0.5	1.0	***91.0***

Table 17: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Conditional distribution of observed PSR on hidden states	1	2.6				0.1		0.5	**76.4**		0.1
	2	3.9				**93.0**		**21.9**	0.1	0.1	
	3	4.3		0.3	0.2		0.1	**55.6**	0.1	**99.6**	
	4		**96.3**					**16.3**		0.2	0.1
	5		3.4	1.3	0.5	6.2		3.4			**99.6**
	6						**99.5**	0.2	4.5		
	7	**80.8**	0.1					1.4			
	8	1.7		**65.7**		0.5		0.1			0.2
	9	0.1		**31.5**			0.2		0.2		
	10	6.4	0.1					0.5	1.2		
	11	0.2		0.1	**98.6**	0.2					
	12			1.0	0.7		0.2	0.1	**17.4**		

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

The transition probability matrices show that the highest probabilities on the diagonal are estimated for BPI. Those same probabilities consistently decreased for BPII, and even more for BPNOS. This observation and the fact that there are eleven transitions off the diagonal with probabilities above 5% for BPII and BPNOS, implicate that the BPI patients do not move to often among states, and than BPII and BPNOS patients have a similar transition pattern.

The conditional distribution of observed PSR given the hidden state are similar to those estimated for the HMM with ten hidden states without covariates, except for hidden states 1 and 7.

The mean sojourn times in Table 18. reiterates the observation made with the transition probability matrices. BPI patients spend almost twice the number of weeks than patients with BPII and BPNOS diagnosis, for states: 2, 4, 5, 7, 9 and 10 (that corresponds to PSR scores 1 through 5, and 12, i.e., pure depression episodes, well, submania pure and with subdepression, and mixed state). BPI patients spend consistently more weeks in all ten states than patients with either of the other two diagnoses. And BPII and BPNOS show the same pattern in the number of days spend in each of the hidden states.

Table 18: Mean sojourn times by bipolar diagnosis

Diagnosis	1	2	3	4	5	6	7	8	9	10
BPI	3.2	30.1	2.3	5.0	17.0	6.0	16.5	8.6	49.0	22.7
BPII	2.0	14.9	1.9	2.5	9.5	4.4	8.9	5.9	28.6	11.0
BPNOS	2.0	14.1	1.7	2.5	8.5	4.1	7.5	4.8	24.6	10.4

Appendix B. contains Figures 12., 13. and 14. that show the observed prevalence and the estimated prevalence of the HMM with ten hidden states with bipolar diagnosis as covariate. Figure 12. shows underestimation for state 2, and for state 10 only from week 150 and up. Overestimation is observed for state 9. For BPII and BPNOS patients, Figures 13. and 14. show overestimation for state 9 starting in week 150, and underestimation of state 10 starting around week 170.

4.2.5 Hidden Markov model with age of bipolar onset as covariate

The parameters of the HMM with ten hidden states when age of bipolar onset is included as covariate are presented in Table 19. The distribution of the initial state is almost uniform.

Table 19: Parameters of the HMM with ten hidden states with age of bipolar onset as covariate (in percentages)†

Hidden State	1	2	3	4	5	6	7	8	9	10
Initial state probabilities	8.0	10.4	8.2	10.2	9.7	9.0	10.7	9.7	12.5	11.4

Table 19: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Transition probability matrix-Childhood	1	**66.0**	1.7	2.3	2.8	4.2	3.0	**5.2**	4.2	**6.2**	4.5
	2	0.7	***97.5***	0.1	0.6	0.1	0.4	0.4		0.1	
	3	0.4	1.1	**61.5**	3.7	**6.7**	3.2	**5.8**	4.6	**5.8**	**7.3**
	4	0.9	3.6	1.6	**74.7**	3.3	2.2	4.4	1.5	4.0	3.6
	5	0.4	1.1	0.5	0.7	***93.8***	0.3	1.4	0.1	1.6	0.1
	6	1.0	2.7	1.5	1.7	2.3	***80.8***	2.4	2.2	2.3	3.2
	7	1.2	1.2	0.8	0.7	2.2	0.9	***90.4***	0.5	1.1	1.1
	8	1.5	1.9	1.1	1.7	4.1	1.9	4.1	**77.8**	3.8	2.1
	9	0.2	0.6	0.1				2.2		***96.7***	
	10	0.4	0.4	0.5	0.6	0.2	0.3	0.2		0.1	***97.2***
Transition probability matrix-Early adolescence	1	**63.3**	1.6	2.3	2.9	4.4	3.4	**5.2**	4.9	**7.0**	4.8
	2	0.7	***96.9***	0.2	0.7	0.1	0.6	0.6	0.1	0.1	0.1
	3	0.4	1.0	**55.9**	4.0	**7.3**	3.5	**6.3**	**6.4**	**6.5**	**8.6**
	4	0.9	3.5	1.6	**74.5**	3.2	2.3	4.5	2.0	4.1	3.4
	5	0.4	1.0	0.5	0.8	***93.2***	0.3	1.5	0.1	2.0	0.1
	6	0.9	2.5	1.4	1.7	2.2	***81.7***	2.3	2.7	2.0	2.6
	7	1.1	1.2	0.8	0.8	2.3	1.0	***89.4***	0.7	1.5	1.1
	8	1.2	1.6	0.9	1.6	3.1	1.7	2.4	***83.9***	2.6	0.9
	9	0.2	0.6	0.1				2.1		***96.9***	
	10	0.6	0.4	0.7	0.8	0.6	0.7	0.4	0.1	0.3	***95.4***

Table 19: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Transition probability matrix- Late adolescence	1	**58.6**	1.6	2.5	2.9	**5.0**	4.0	**5.7**	**5.3**	**8.7**	**5.8**
	2	0.9	***95.0***	0.4	0.8	0.3	0.9	1.1	0.2	0.2	0.2
	3	0.4	1.1	**53.3**	4.1	**8.1**	3.7	**6.9**	**7.0**	**7.2**	**8.3**
	4	0.9	3.9	1.8	**66.8**	4.2	3.2	**6.0**	3.5	**5.2**	4.6
	5	0.5	1.6	0.9	1.1	***89.1***	0.7	2.2	0.4	3.1	0.4
	6	1.0	2.7	1.8	1.8	2.8	**74.8**	3.3	3.5	3.9	4.3
	7	1.5	1.3	1.0	1.3	3.1	1.4	***83.9***	1.3	3.4	1.9
	8	1.5	1.8	1.1	1.8	4.4	2.6	4.4	**75.5**	4.4	2.5
	9	0.4	0.7	0.2		0.1	0.1	2.3		***96.2***	
	10	0.8	0.4	0.9	1.0	1.1	1.0	0.8	0.2	0.7	***93.0***
Conditional distribution of observed PSR on hidden states	1	2.7				0.1		0.6	**76.6**		0.1
	2	3.7				**95.2**		**14.2**	0.1	0.1	
	3	4.3		0.3	0.2		0.1	**66.6**	0.1	**99.7**	
	4		**96.9**					**12.8**		0.2	0.1
	5		2.9	1.3	0.5	4.1		2.9			**99.6**
	6						**99.5**	0.2	4.1		
	7	**80.9**	0.1					1.7			
	8	1.7		**66.1**		0.5		0.2			0.2
	9	0.1		**31.2**			0.2		0.2		
	10	6.4	0.1					0.6	1.3		
	11	0.2		0.1	**98.7**	0.2					
	12			1.0	0.7		0.2	0.2	**17.4**		

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

The transition matrices in Table 19., show the highest probabilities on the diagonal, however states 1 and 3 (syndromal hypomania/mania and one of the syndromal cycling states) have the lowest probabilities on the diagonal. Therefore, it is for those two states where transitions off the diagonal have probabilities higher than 5%. This time, the probabilities on the diagonal do not increase consistently through the three categories of age of bipolar onset. For states 6, 8 and 9, the probabilities on the diagonal are higher for patients with onset during early adolescence. In all the other states the probabilities on the diagonal are always higher for patients with onset during childhood. This indicates that the individuals who move more often are those whose onset was during late adolescence.

The conditional distributions of the PSR on the hidden states resemble the ones estimated in the HMM without covariates. The only differences are observed for states 1 and 7.

Table 20. presents the mean sojourn times by age of bipolar onset. Similar times are observed for states 1, 3, 4, 6, 8 and 9. In states 2, 5, 7 and 10, the patients with childhood onset spend almost double the time than patients with late adolescence onset. The times for patients with late adolescence onset are consistently lower throughout the states.

Table 20: Mean sojourn times by age of bipolar onset

Onset	1	2	3	4	5	6	7	8	9	10
Childhood	2.4	39.7	2.0	3.4	15.5	4.6	9.8	3.9	29.6	34.7
Early adolescence	2.2	31.0	1.7	3.4	14.0	4.9	8.8	5.7	31.6	21.0
Late adolescence	1.9	19.4	1.6	2.5	8.6	3.4	5.6	3.5	25.2	13.4

In Appendix B., Figures 15., 16. and 17. show the prevalences observed and estimated with the HMM including age of bipolar onset as covariate. For the model of patients with bipolar onset during childhood there is overestimation for states 2 and 10, and underestimation for state 9. For the model of patients with bipolar onset during early adolescence, the same pattern is observed, only that the underestimation of states 9 starts at week 100. In the model of patients with bipolar onset during late adolescence, the discrepancies between the observed and estimated prevalences is smaller.

4.2.6 Hidden Markov model with socio-economic status as covariate

When socio-economic status (SES) is included in the hidden Markov model with ten hidden states, one the main changes is observed in the distribution of the initial state. As can be seen in Table 21., the probability is highest for state 1.

For the probabilities on the diagonal of the transition probability matrices, three patterns are observed throughout the five categories of SES. The first pattern is consistently decreasing probabilities. This pattern happens for states 1, 2, 3, 4 6 and 10. Here is an example of what is meant with that statement: the probability on the diagonal for state 1, i.e., (1,1), decreases from the highest estimated for SES 1 to the smallest estimated for SES 5. The second pattern is consistently increasing, which is observed for states 5 and 9. And the third pattern is neither consistently increasing or consistently decreasing. This last pattern occurs in states 7 and 8.

Table 21: Parameters of the HMM with ten hidden states with socio-economic status as covariate (in percentages)†

Hidden State		1	2	3	4	5	6	7	8	9	10
Initial state probabilities		38.7	6.9	5.4	6.8	6.5	6.0	7.3	6.5	8.3	7.6
Transition probability matrix-SES 1	1	**69.5**	1.2	1.9	2.5	3.4	3.6	4.6	3.8	**5.2**	4.3
	2	0.1	***98.1***	0.7	0.2	0.1	0.2	0.2		0.3	0.1
	3	0.4	1.1	**66.9**	3.5	**5.0**	3.2	4.6	3.6	3.5	**8.3**
	4	0.3	1.1	1.3	***83.9***	2.5	1.4	2.0	2.5	2.4	2.6
	5	0.4	0.6	0.8	1.4	***89.3***	1.0	1.4	2.0	1.2	1.8
	6	1.3	2.0	0.6	1.6	1.5	***84.8***	1.7	2.8	1.5	2.1
	7	1.3	0.9	0.9	1.1	1.5	1.2	***83.9***	2.3	5.3	1.7
	8	1.2	0.7	0.9	2.2	3.1	2.2	2.4	***81.9***	3.9	1.5
	9	0.1	0.4	0.5	0.8	0.8	0.6	3.3	0.7	***92.0***	0.7
	10	0.5	0.9	0.5	0.1	0.7	0.7	0.9	0.1	0.1	***95.6***

Table 21: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
	1	**68.2**	1.3	1.9	2.5	3.8	3.7	**5.0**	4.1	**5.2**	4.5
	2	0.1	***97.4***	0.7	0.4	0.1	0.3	0.4		0.5	0.1
	3	0.4	1.1	**62.6**	3.9	**5.5**	3.5	**5.7**	4.4	4.0	**9.1**
Transition probability matrix-SES 2	4	0.3	1.3	1.7	**79.8**	2.6	1.9	2.8	3.3	2.9	3.4
	5	0.4	0.6	0.9	1.3	***91.0***	1.0	1.3	1.4	0.9	1.2
	6	1.5	2.3	0.6	1.8	1.9	***82.3***	2.1	2.9	1.9	2.7
	7	1.3	1.2	0.9	1.1	1.8	1.2	***88.4***	2.0	0.5	1.5
	8	1.2	0.8	0.9	2.2	3.2	2.2	2.6	***82.2***	3.2	1.6
	9	0.2	0.5	0.1	0.4	0.5	0.2	1.5	0.1	***96.3***	0.1
	10	0.6	1.1	0.6	0.2	0.8	0.8	1.0	0.2	0.1	***94.7***
	1	**66.7**	1.4	1.8	2.4	4.2	3.7	**5.3**	4.5	**5.2**	4.8
	2	0.1	***96.0***	0.7	0.6	0.2	0.4	0.8	0.1	0.9	0.1
	3	0.4	1.2	**57.8**	4.3	**6.0**	3.8	**7.0**	**5.3**	4.6	**9.8**
Transition probability matrix-SES 3	4	0.3	1.4	2.1	**74.7**	2.8	2.7	3.8	4.3	3.5	4.4
	5	0.4	0.6	0.9	1.1	***92.2***	1.0	1.2	1.0	0.7	0.9
	6	1.7	2.7	0.6	1.9	2.4	**79.3**	2.6	3.0	2.4	3.5
	7	1.2	1.6	1.0	1.2	2.1	1.2	***88.5***	1.7	0.2	1.4
	8	1.1	0.8	1.0	2.1	3.3	2.1	2.8	***82.3***	2.6	1.8
	9	0.2	0.6		0.2	0.3	0.1	0.7		***97.8***	
	10	0.7	1.3	0.7	0.3	1.0	1.0	1.1	0.3	0.2	***93.4***

Table 21: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
	1	**65.2**	1.5	1.8	2.3	4.6	3.7	**5.6**	4.9	**5.2**	**5.1**
	2	0.1	***93.4***	0.8	0.8	0.5	0.7	1.6	0.2	1.6	0.3
	3	0.4	1.3	**52.5**	4.6	**6.5**	4.1	**8.4**	**6.3**	**5.2**	**10.6**
Transition probability matrix-SES 4	4	0.4	1.6	2.6	**68.3**	2.9	3.7	**5.1**	**5.6**	4.2	**5.6**
	5	0.4	0.7	1.0	0.9	***93.1***	1.0	1.1	0.8	0.5	0.6
	6	1.9	3.0	0.7	2.1	3.0	**75.6**	3.1	3.0	3.0	4.5
	7	1.2	2.0	1.0	1.2	2.5	1.2	***88.0***	1.5	0.2	1.3
	8	1.1	0.9	1.0	2.0	3.4	2.1	3.0	***82.3***	2.2	2.0
	9	0.4	0.7		0.1	0.2		0.3		***98.2***	
	10	0.8	1.5	0.9	0.6	1.3	1.2	1.2	0.7	0.5	***91.5***
	1	**63.6**	1.6	1.8	2.1	**5.1**	3.8	**6.0**	**5.4**	**5.2**	**5.4**
	2	0.1	***88.2***	0.8	1.2	1.2	1.1	3.0	0.6	2.9	0.9
	3	0.5	1.3	**46.9**	4.9	**7.1**	4.3	**10.1**	**7.6**	**6.0**	**11.3**
Transition probability matrix-SES 5	4	0.4	1.7	3.1	**60.5**	3.1	**5.0**	**6.9**	**7.1**	**5.0**	**7.1**
	5	0.4	0.7	1.0	0.8	***93.7***	1.0	1.0	0.6	0.4	0.5
	6	2.1	3.4	0.7	2.2	3.7	**71.4**	3.8	3.1	3.8	**5.7**
	7	1.1	2.5	1.0	1.2	3.0	1.3	***87.2***	1.3	0.2	1.2
	8	1.1	0.9	1.1	1.9	3.5	2.0	3.2	***82.2***	1.8	2.3
	9	0.5	0.8		0.1	0.1		0.2		***98.2***	
	10	0.9	1.7	1.1	1.1	1.6	1.5	1.4	1.3	1.4	***88.1***

Table 21: (continued)

Hidden State		1	2	3	4	5	6	7	8	9	10
Conditional distribution of observed PSR on hidden states	1	2.5				0.1		0.5	**76.3**		0.1
	2	4.1				**93.4**		**21.2**	0.1	0.1	
	3	4.6		0.3	0.2		0.1	**54.2**	0.1	**99.6**	
	4		**96.4**					**18.8**		0.2	0.1
	5		3.3	1.3	0.5	5.8		3.0			**99.6**
	6						**99.5**	0.2	4.7		
	7	**80.3**	0.1					1.4			
	8	1.7		**65.9**		0.5		0.1			0.2
	9	0.1		**31.4**			0.2		0.2		
	10	6.4	0.1					0.5	1.2		
	11	0.2		0.1	**98.7**	0.2					
	12			1.0	0.7		0.2	0.1	**17.3**		

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

Off the diagonal, SES 1 has the fewest number of transitions with probabilities over 5%, and SES 5 has the most number of probabilities above 5%.

The conditional distributions of the PSR have the same patterns observed for the HMM with other covariates: only the estimates for state 1 and 7 differ from the ones obtained with the HMM without covariates.

Table 22. contains some interesting results. The mean sojourn times for states 1 (syndromal hypomania/mania), 3 (syndromal cycling, PSRs 8 and 9) and 8 (syndromal MDD and pure mixed) are almost the same no matter to what socio-economic status the patients belong to. State 7 (mixture of subsyndromal pure depression, asymptomatic and subsyndromal pure mania) does not have a monotone pattern throughout the range of SES, but the mean sojourn times are around 5 and 8 days.

For states 4 and 6 (syndromal cycling, PSRs 6 and 11), the mean sojourn times decrease as the socio-economic status increases, with the mean sojourn time for SES 1 doubling the mean sojourn time observed for SES 5. State 5 (subsyndromal pure depression) presents a monotonically increasing pattern in the mean sojourn times of the five categories of the socio-economic status, in this case the mean sojourn time of SES 5 doubles the mean sojourn time estimated for SES 1. The most striking patterns are observed for states 1, 9 and 10. The mean sojourn times in states 2 and 10 (subsyndromal pure mania and subsyndromal mixed) increase monotonically through the range of categories of socio-economic status. However, for state 2, the mean sojourn time of SES 1 is almost seven times the mean sojourn time observed for SES 5. While for state 10, the mean sojourn time of SES 1 is only three times the mean sojourn time observed for SES 5. On the other hand, state 9 (asymptomatic) shows a monotonically decreasing pattern, with the mean sojourn time of SES 5 being a little over four times the mean sojourn time estimated for SES 1.

Table 22: Mean sojourn times by socio-economic status

SES	1	2	3	4	5	6	7	8	9	10
1	2.7	53.0	2.5	5.7	8.7	6.0	5.6	4.9	11.8	21.9
2	2.6	37.6	2.1	4.4	10.5	5.1	8.0	5.0	26.6	18.0
3	2.5	24.4	1.8	3.4	12.2	4.3	8.0	5.1	43.7	14.4
4	2.3	14.5	1.5	2.6	13.7	3.5	7.6	5.1	54.4	11.0
5	2.2	7.9	1.3	2.0	15.0	2.9	7.1	5.0	53.5	7.7

In Appendix B., Figure 18. shows that the observed and the estimated prevalences of patients who belong to SES 1 differ for states 2, 9 and 10. There is overestimation in states 2 and 10 and underestimation in state 9. A similar observation can be made for patients who belong to SES 2, as shown in Figure 19., only that the overestimation in state 10 starts in week 150. Figure 20. shows overestimation for state 2 and underestimation for state 9, however, the latter is from week 150 and up. No discrepancies between the prevalences are notorious in Figure 21. Figure 22. shows underestimation in states 5 and 9 for the first 100 and 150 weeks, respectively. It also shows overestimation for state 10 up to week 150.

5.0 DISCUSSION, CONCLUSION AND FUTURE WORK

5.1 DISCUSSION

With mixtures of first order Markov chains, this study found:

1. The model with four components as the best model, according to the BIC criterion. Since the other evaluation criteria did not have a monotone pattern, only the BIC was used to make the final decision on the number of components. The AIC, BIC and NEC criteria are easily computed. On the other hand, the MCCV and Score criteria are computationally complex.
2. In this mixture model, the bipolar diagnosis itself does not shed light in the classification of the longitudinal courses of bipolar patients, but the range of episodes observed by bipolar type differs.
3. The estimated parameters of this mixture show the frequency of the patterns, the distribution of the episode with which patients in each cluster started their longitudinal follow-up, and the 144 transitions among the 12 PSR categories within each cluster. This information has never been used before in the psychiatric field. Neither have been plots of the longitudinal course of the patients. These clusters and their parameters certainly help to understand how bipolar youth transition from episode to episode and from week to week.
4. The mixture of first order Markov chains model with four components identified patterns in the longitudinal course of bipolar youth that are characterize for flat and spiky periods, i.e., some of the clusters are conformed by patients who present the same episode for several weeks, other clusters have individuals who move along the whole spectrum

of bipolar episodes and other clusters combines these two features. The four clusters were labeled as: stayers (cluster 4, 70%), movers to the depression, well and submania states (cluster 3, 16%), movers who also tend to stay several weeks in the well state (cluster 2, 11%), and movers who also tend to stay in the subdepression only, well and submania/subdepression states (cluster 1, 4%).

5. Cluster 2 identifies patients who present either a well or a mixed state episode after experiencing a mania/subdepression episode.

6. Cluster 1 groups patients whose longitudinal courses almost never start with an episode involving MDD or a mania state. Also in this cluster, a patient who experiences a mania/subdepression episode, either stays on that same state (probability 1/3) or move to a submania/subdepression episode (probability 2/3). Subjects experiencing a hypomania/subdepression episode will stay for one more week on that same state 20% of the time, or they will experience the following week either a submania/subdepression or a submania/MDD episode (probability 40% each).

On the other hand, these are the findings obtained with the hidden Markov model approach:

1. The hidden Markov model with ten hidden states was found as the best model by the evaluation criteria. The BIC and CIC estimated the same number of hidden states.

2. An eight PSR scale used by (Birmaher et al., 2006) was found to be similar to the hidden states, only in a different order. The eight PSR scale was used to label the hidden states. The two extra categories in the HMM are due to: (i) a separate identification of the cycling states—instead of only one cycling state there are three—, (ii) Pure MDD and Pure mixed—PSR 1 and 12 respectively—are merged in one state, and (iii) one hidden state is identified as the mixture of subsyndromal pure depression, asymptomatic and subsyndromal pure mania.

3. This HMM gives a statistical justification for the use of the eight PSR scale. With the extra consideration that cycling should probably not be joined in one category, because cycling behavior characterizes the longitudinal course of bipolar youth.

Finally, the inclusion of individual demographic covariates measured at intake provided these results:

1. For gender, males tend to transition more than females do. In average, female spend almost twice the number of weeks than males in these episodes: subsyndromal pure mania (PSR 4) , subsyndromal pure depression episode (PSR 2) and syndromal MDD and pure mixed (PSRs 1 and 12).
2. When including age, categorized as 1 for age 13 or more and 0 otherwise, i.e. 1 corresponds to teenagers and 0 to children, the hidden Markov model estimated that children move more than teenagers. The mean sojourn times of teenagers almost double those of children for episodes: subsyndromal pure mania, subsyndromal pure depression, syndromal MDD and pure mixed, and asymptomatic.
3. Patients who live in another situation different to living with both natural parents, move more among the bipolar episodes. This shows the effect of family stability in the mood of children and adolescents. Particularly, for episodes subsyndromal pure mania, subsyndromal pure depression, syndromal MDD and pure mixed, and asymptomatic, the mean sojourn times of patients who live in another situation are doubled the corresponding times of patients who live with both natural parents.
4. For the bipolar diagnosis, BPI patients have the behavior of stayers and patients with the other two diagnoses have the behavior of movers. The transition probability matrices and the mean sojourn times for BPII and BPNOS share similarities, suggesting that the transition patterns of patients with these two diagnoses agree in the ten bipolar episodes.
5. Age of bipolar onset showed childhood and early adolescence onsets with similar behaviors, opposed to more transient patients with a late adolescence onset. This indicates that the earlier a patient develops the disorder, the more stable the patient will become through the episodes, spending more time in each, specially those that are milder. This could also suggest that the disorder is diagnosed earlier, and therefore, the prescription of treatment will help to lessen the impairment from the disorder.
6. Socio-economic status turned out to be one of the most interesting covariates. The effects vary throughout the five categories of this covariate. It shows patients with low socio-economic status spend more time with subsyndromal pure mania and subsyndromal

mixed episodes and less time with subsyndromal pure depression and asymptomatic episodes, the opposite of the patients in high socio-economic status. Hence, the poorer a patient is the more impaired he/she can be by bipolar disorder. The lack of economic resources seems to lengthen the presence of submanic and mixed episodes. And the abundance of economic resources seem to imply more weeks without bipolar symptoms or more weeks with subsyndromal depression. Maybe the economic status could be affecting the access to treatment, are poor families with bipolar children and adolescence being able to afford treatments for the disorder?

7. In all the six models with covariates, the conditional distribution of the observed PSRs on the hidden states is similar to the one estimated for the HMM without covariates. The only difference is for hidden states 1 and 7. Hidden state 1 is mainly made up of PSR 7. And hidden state 7 has a different distribution that the one observed for the HMM without covariates, but it is made up of the same mixtures of PSRs.
8. Regarding the goodness of fit of these six HMMs, the prevalence estimated with the hidden Markov model with ten hidden states with each of the six covariates usually overestimates state 2 (subsyndromal mania) and underestimates state 9 (asymptomatic).

5.2 CONCLUSIONS

In summary:

1. Four clusters show patterns of movers and stayers. Cluster 4 is the stayers. Cluster 3 are movers among the depression, well and submania states. Cluster 2 are movers that tend to stay in the well state. Cluster 1 are movers that tend to go to the submania/subdepression state.
2. The range of transitions in each of the clusters vary by bipolar diagnosis.
3. MCCV and Score did not serve as criteria to choose the number of components. NEC does not behave well for this model.
4. Ten hidden states are labeled using syndromal, subsyndromal and asymptomatic episodes defined by the psychiatrists.

5. When including a binary covariate in hidden Markov models, patients in one of the categories transition more than the patients in the other category: males move more than females, children move more than teenagers, and patients who live in another situation move more than patients who live with both natural parents.
6. For bipolar diagnosis in the hidden Markov model, BPII and BPNOS patients show similar transition patterns.
7. Age of bipolar onset sheds light on the stability of patients with a childhood and an early adolescence onset. The possibility of an early diagnosis of the disorder, which consequently would lead to providing appropriate treatment, would lessen the impairment of bipolar youth.
8. Socio-economic status is by far the covariate in the hidden Markov model with the most interesting effect. It shows patients with low socio-economic status staying more weeks with subsyndromal submanic and mixed episodes, and less weeks with subsyndromal depression and asymptomatic episodes. Quite the opposite behavior observed for their counterparts in the high socio-economic status.

5.3 FUTURE WORK

The inclusion of longitudinal covariates like treatment remains to be studied in the future. Modeling covariates with the mixture of first order Markov chains model should also be considered, to study the effect of covariates in the clusters.

Also, at the end of the literature review of statistical models that could be appropriate for finding patterns of the longitudinal course of bipolar youth these two references were found and kept as ideas to work on in the future:

- Clustering Variable Length Sequences by Eigenvector Decomposition using HMM
  Porikli (2004) proposes a clustering method using HMM parameter space and eigenvector decomposition. This algorithm can cluster both constant and variable length sequences without requiring normalization of data.

- Mixed Memory Markov Models

  Saul and Jordan (1999) study Markov models whose state spaces arise from the Cartesian product of two or more discrete random variables. This models could be applied in COBY using the original 6-point scales of depression, hypomania and mania. Saul and Jordan (1999) show how to parameterize the transition matrices of these models as a convex mixture of simpler dynamical models. The parameters in these models admit a simple probabilistic interpretation and can be fitted iteratively by an EM algorithm. They derive a set of generalized Baum-Welch updates for factorial hidden Markov models that make use of this parameterization and also describe a simple iterative procedure for approximately computing the statistics of the hidden states.

BIBLIOGRAPHY

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

G. Archer and D. Titterington. Parameter estimation for hidden Markov chains. *Journal of Statistical Planning and Inference*, 108(1-2):365–390, 2002.

C. Biernacki, G. Celeux, and G. Govaert. An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20:267–272, 1999.

J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report 97021, International Computer Science Institute, Berkeley, CA, USA., 1997.

B. Birmaher, D. Axelson, M. Strober, M. Gill, S. Valeri, L. Chiappetta, N. Ryan, H. Leonard, J. Hunt, S. Iyengar, J. Bridge, and M. Keller. Clinical course of children and adolescents with bipolar spectrum disorders. *Archives of General Psychiatry*, 63:175–183, 2006.

D. Böhnig, E. Dietz, R. Schaub, P. Schlattman, and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one parameter exponentional family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388, 1994.

I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-based clustering and visualization of navigation patterns on a web site. *Data mining and knowledge discovery*, 7:399–424, 2003.

G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13:195–212, 1996.

P. Devijver. Baum's forward-backward algorithm revisited (STMA V28 1824). *Pattern Recognition Letters*, 3:369–373, 1985.

P. Diggle, K. Liang, and S. Zeger. *Analysis of Longitudinal Data.* Oxford University Press, 1994.

DSM-IV. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.* American Psychiatric Association, Washington, DC., 1994.

DSM-IV-TR. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision*. American Psychiatric Association, Washington, DC., 2000.

M. First and A. Tasman. *DSM-IV-TR Mental Disorders. Diagnosis, Etiology and Treatment*. Wiley, 2004.

C. Fraley and A. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998. URL http://comjnl.oxfordjournals.org/cgi/reprint/41/8/578.pdf.

B. Geller and J. Luby. Child and adolescent bipolar disorder: A review of the past 10 years. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36(9):1168–1176, 1997.

T. Glenn, P. Whybrow, N. Rasgon, P. Grof, M. Alda, C. Baethge, and M. Bauer. Approximate entropy of self-reported mood prior to episodes in bipolar disorder. *Bipolar disorders*, 8:424–429, 2006.

E. Hannan and B. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B: Methodological*, 41:190–195, 1979.

J. Hardin and J. Hilbe. *Generalized Estimating Equations*. Chapman & Hall Ltd, 2003.

J. Hennen. Statistical methods for longitudinal research on bipolar disorders. *Bipolar disorders*, 5:156–168, 2003.

C. Jackson. *Multi-state modelling with R: the* `msm` *package*. Comprehensive R Archive Network, 2007. URL http://rss.acs.unt.edu/Rdoc/library/msm/doc/msm-manual.pdf.

B. Juang and L. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33:251–272, 1991.

M. Keller, P. Lavori, B. Friedman, E. Nielsen, J. Endicott, P. McDonald-Scott, and N. Andreasen. The longitudinal interval follow-up evaluation: a comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry*, 44:540–548, 1987.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag Inc, 1997.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

B. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40:127–143, 1992.

A. Marneros and P. Brieger. Prognosis of bipolar disorders: a review. In M. Maj, H. Akiskal, J. Lopez-Ibor, and N. Sartorius, editors, *Bipolar disorder Vol. 5*. John Wiley and Sons Ltd, 2002.

Guillermo Marshall and Richard H. Jones. Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14:1975–1983, 1995.

NIMH. Bipolar disorder. NIH Publication Number: NIH 3679, 2002. URL http://www.nimh.nih.gov/publicat/bipolar.cfm.

NIMH. Child and adolescent bipolar disorder: An update from the national institute of mental health. NIH Publication Number: NIH 4778, 2000. URL http://www.nimh.nih.gov/publicat/bipolarupdate.cfm.

S. Pincus. Approximate entropy as a measure of irregularity for psychiatric serial metrics. *Bipolar disorders*, 8:430–440, 2006.

F. Porikli. Clustering variable length sequences by eigenvector decomposition using hmms. In *Proceedings of IEEE International Conference on Pattern Recognition (ICPR), Workshop SSPR*, pages 352–360, Portugal, 2004. Springer-Verlag.

L. Rabiner. A tutorial on hidden markov models and selected application in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

V. Rao, C. Rao, and V. Yeragani. A novel technique to evaluate fluctuations of mood: implications for evaluating course and treatment effects in bipolar/affective disorders. *Bipolar disorders*, 8:453–466, 2006.

J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

L. Saul and M. Jordan. Mixed memory markov models: decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, 37:7587, 1999.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

S. Senn, L. Stevens, and N. Chaturvedi. Repeated measures in clinical trials: Simple strategies for analysis using summary measures. *Statistics in Medicine*, 19(6):861–877, 2000.

K. Shulman, A. Schaffer, A. Levitt, and N. Herrmann. Effects of gender and age on phenomenology and management of bipolar disorders: a review. In M. Maj, H. Akiskal, J. Lopez-Ibor, and N. Sartorius, editors, *Bipolar disorder Vol. 5*. John Wiley and Sons Ltd, 2002.

P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

O. Taramasco. *RHmm: Hidden Markov Models simulations and estimations*. Comprehensive R Archive Network, 2007. URL http://cran.r-project.org/web/packages/RHmm/index.html.

R. Whiting and E. Pickett. On model order estimation for partially observed markov chains. *Automatica*, 24(4):569–572, 1988.

B. Winer, D. Brown, and K. Michels. *Statistical Principles in Experimental Design*. McGraw Hill Book Co, New York, 3rd edition, 1991.

APPENDIX A

PARAMETERS OF MIXTURE OF FOUR FIRST ORDER MARKOV CHAINS MODEL

For the model with four components that has the highest likelihood, the component weights were: 0.6941, 0.1576, 0.1071, 0.04116

A.0.1 Initial state probabilities

Table 23: Initial state probabilities of mixture of four first order Markov chains (in percentages)[†]

	PSR categorie											
Component	1	2	3	4	5	6	7	8	9	10	11	12
4	8.4	8.1	27.8	13.8	17.1	5.6	1.4	1.4	0.7	6.3	3.5	5.9
3	7.7	13.4	38.4	9.8	18.4	4.6			1.5	1.5		4.6
2	15.8	9.0	45.7	11.3	11.3	4.5				2.3		
1		5.9	29.4	17.6	23.5		17.6	5.9				

[†] An empty cell means that the estimated probability expressed in percentage is less than 0.1%

A.0.2 Transition probability matrices

Table 24: Transition probability matrices of mixture of four first order Markov chains (in pertecentages)†

		PSR categorie											
Cluster		1	2	3	4	5	6	7	8	9	10	11	12
4	1	**87.7**	**5.3**	3.9	0.2	1.0	1.3			0.1	0.3		
	2	1.4	**90.5**	**5.7**	0.4	1.4	0.1		0.2		0.2		
	3	0.4	1.4	**96.2**	1.4	0.2		0.2			0.1		
	4	0.2	0.3	4.6	**91.4**	2.0	0.6	0.4			0.4	0.1	0.1
	5	0.3	1.1	0.7	1.8	**93.8**	1.3	0.1	0.2			0.5	0.2
	6	1.2	0.4	0.4	1.6	4.7	**90.1**			0.5	0.1		0.9
	7	0.1	1.1	2.5	4.5	0.5	0.1	**87.6**	1.5	0.4	1.6		0.1
	8	0.4	3.3	0.4	0.4	4.0		2.5	**86.2**	1.3		0.8	0.6
	9	0.9		0.4		4.9	3.1	1.3	4.5	**83.4**			1.3
	10	1.1	2.4	3.0	**6.5**	0.3	0.2	2.0	0.2		**82.2**	1.4	0.6
	11		0.5	1.5	0.7	6.2	0.5	0.2	1.3		1.5	**86.3**	1.3
	12	0.8		1.5	1.0	2.8	2.2	0.3	0.4	0.8	0.3	1.1	**88.7**
3	1	**71.0**	3.5	3.3	0.5	0.7	**17.4**	0.2	0.1	0.2	0.1	0.2	3.0
	2	2.4	**77.1**	**5.4**	2.1	**10.9**	0.7		0.6		0.1	0.7	0.1
	3	0.8	2.2	**83.5**	**11.3**	0.8	0.2	0.9			0.1		
	4	0.6	1.5	**33.9**	**59.9**	1.7	0.4	1.7			0.3		
	5	0.6	**14.6**	2.9	1.9	**75.7**	1.1	0.1	3.0			0.2	
	6	**43.8**	1.9	2.3	1.5	3.8	**45.3**		0.6		0.2		0.6
	7	0.7		**27.2**	**14.6**	2.0		**51.2**		0.7	3.6		
	8		**11.4**		3.1	**38.1**			**43.3**			4.1	
	9	**5.5**		1.8		3.6	1.8			**87.3**			
	10	**5.3**	**5.3**	**18.5**	**15.9**		2.6	3.6			**46.1**	2.6	
	11	2.6	**19.7**	1.3	1.3	2.6	2.6					**68.4**	1.3
	12	**31.7**		3.3		1.7	1.7				0.8		**60.8**
2	1	**72.0**	2.6	**11.1**	1.7		0.3	2.9	0.3	1.0	**7.9**		
	2	0.6	**37.3**	**38.9**	**9.4**	2.8	0.1	4.1	2.7	0.2	3.8	0.1	
	3	1.4	**10.2**	**75.4**	**8.6**	1.2	0.2	2.0	0.3		0.7		
	4	1.4	**9.2**	**24.8**	**47.2**	**16.6**	0.5	0.2					0.2
	5		3.6	**6.4**	**26.3**	**63.2**	0.2	0.1					0.1
	6	4.6		**5.5**	4.6	2.8	**81.6**		0.9				
	7	**8.1**	**8.6**	**48.6**	0.5			**30.0**	3.8		0.5		
	8	0.9	**22.8**	**12.3**		2.6		**6.1**	**54.4**		0.9		
	9	**17.1**	2.9			2.9				**77.1**			
	10	**28.8**	**21.2**	**16.5**	0.6						**32.4**		0.6
	11			**49.8**									**49.8**
	12	2.3				**6.8**	2.3		2.3				**86.3**

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

Table 24: (continued)

		PSR categorie											
Cluster		1	2	3	4	5	6	7	8	9	10	11	12
1	1	**64.4**	3.8	3.8	3.8	1.9	**21.2**	1.0					
	2	0.7	**9.0**	**13.4**	**15.3**	**58.2**	0.4	2.6			0.4		
	3	1.4	**5.3**	**45.9**	**19.2**	**24.0**	0.4	1.4	0.2		0.7		1.4
	4	0.6	**5.0**	**15.8**	**72.3**	4.7	0.8	0.6	0.2		0.2		
	5	0.2	**26.9**	**22.0**	3.4	**43.1**	0.5	0.3	1.8		0.6	0.3	0.8
	6	**5.6**		0.5	1.5	1.5	**90.3**			0.5	0.2		
	7		4.8	**6.2**	2.1	1.4		**82.8**	2.1		0.7		
	8			3.2	3.2	**32.3**	3.2	6.5	**51.6**				
	9					**39.9**	**39.9**			**20.0**			
	10		**6.0**	1.5	1.5	**6.0**	1.5	1.5			**80.6**		1.5
	11					**66.5**						**33.3**	
	12		**10.0**	**40.0**		**5.0**				**10.0**	**5.0**		**30.0**

† An empty cell means that the estimated probability expressed in percentage is less than 0.1%

APPENDIX B

PREVALENCE PLOTS FOR HIDDEN MARKOV MODELS WITH COVARIATES

B.0.3 With gender

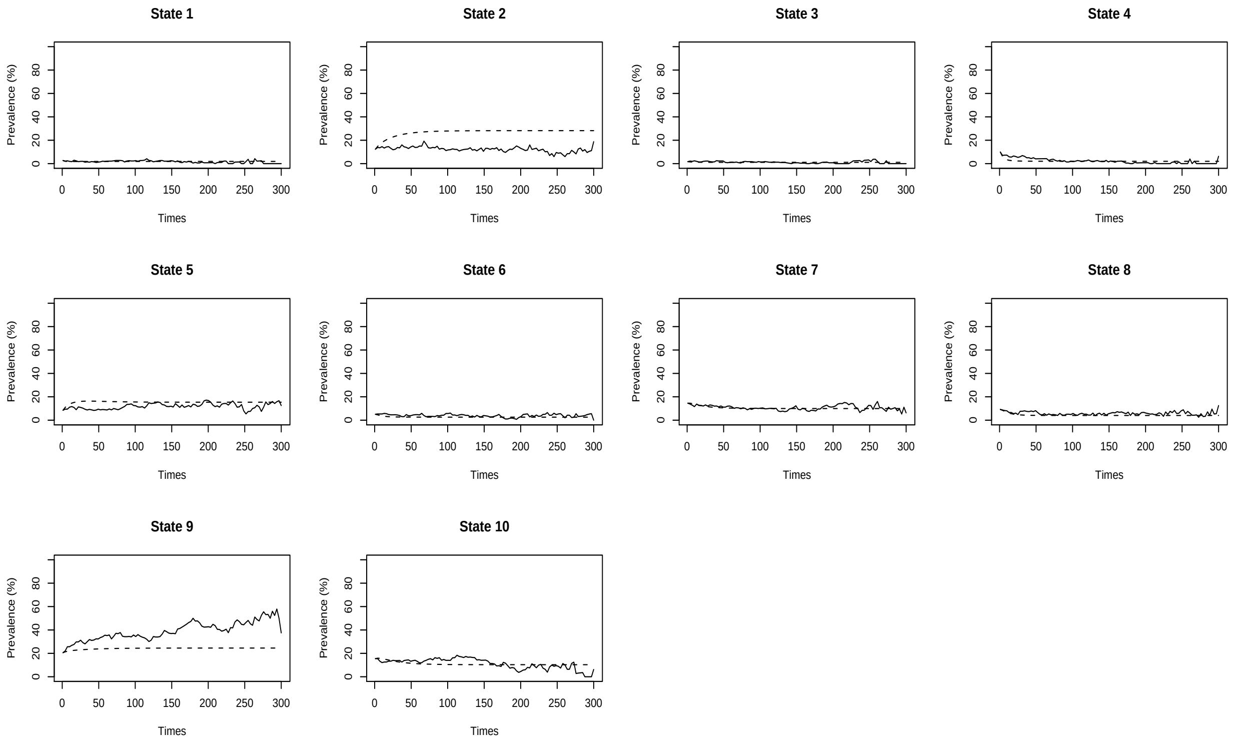


Figure 6: Prevalence of ten episodes of bipolar disorder in females: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

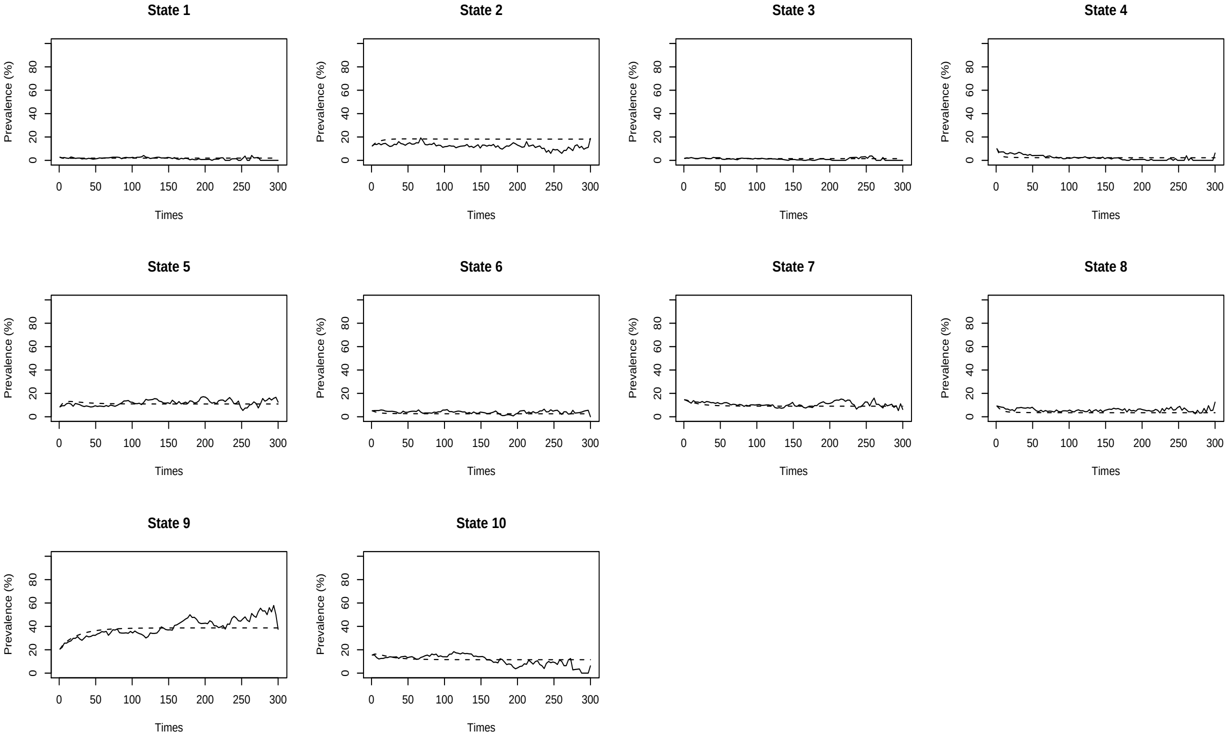


Figure 7: Prevalence of ten episodes of bipolar disorder in males: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

B.0.4 With age

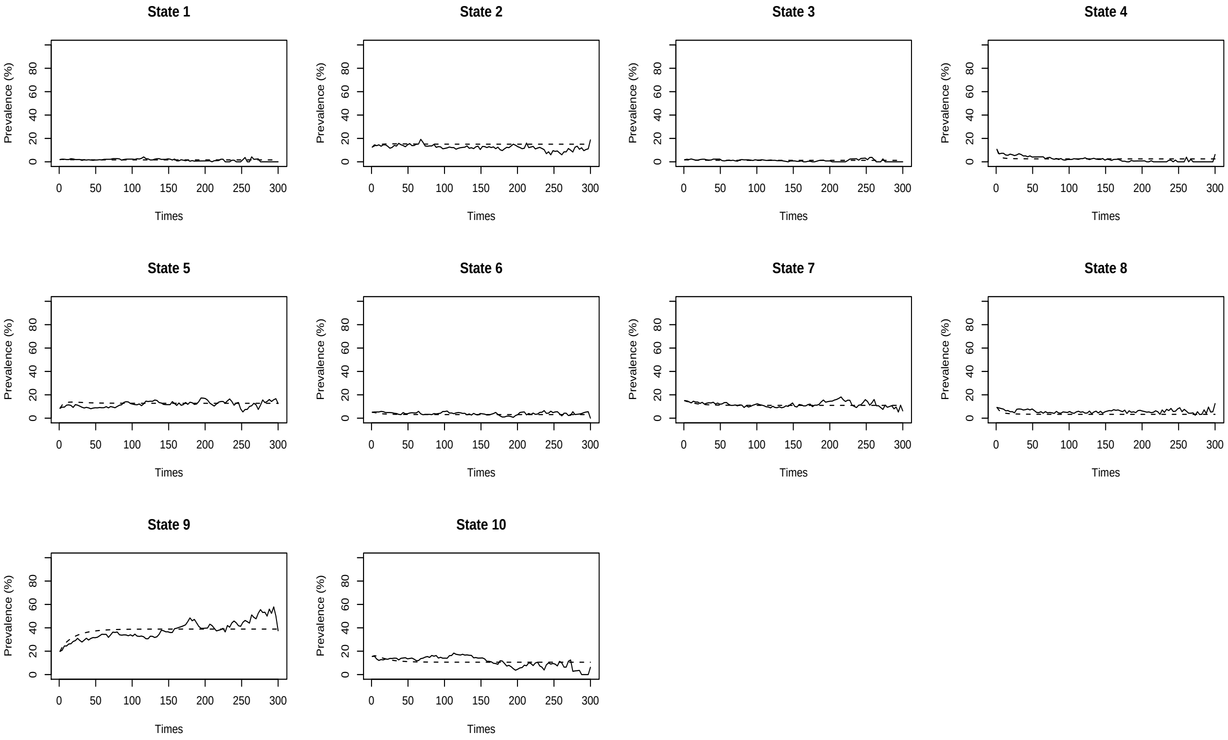


Figure 8: Prevalence of ten episodes of bipolar disorder in children (age<13): observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

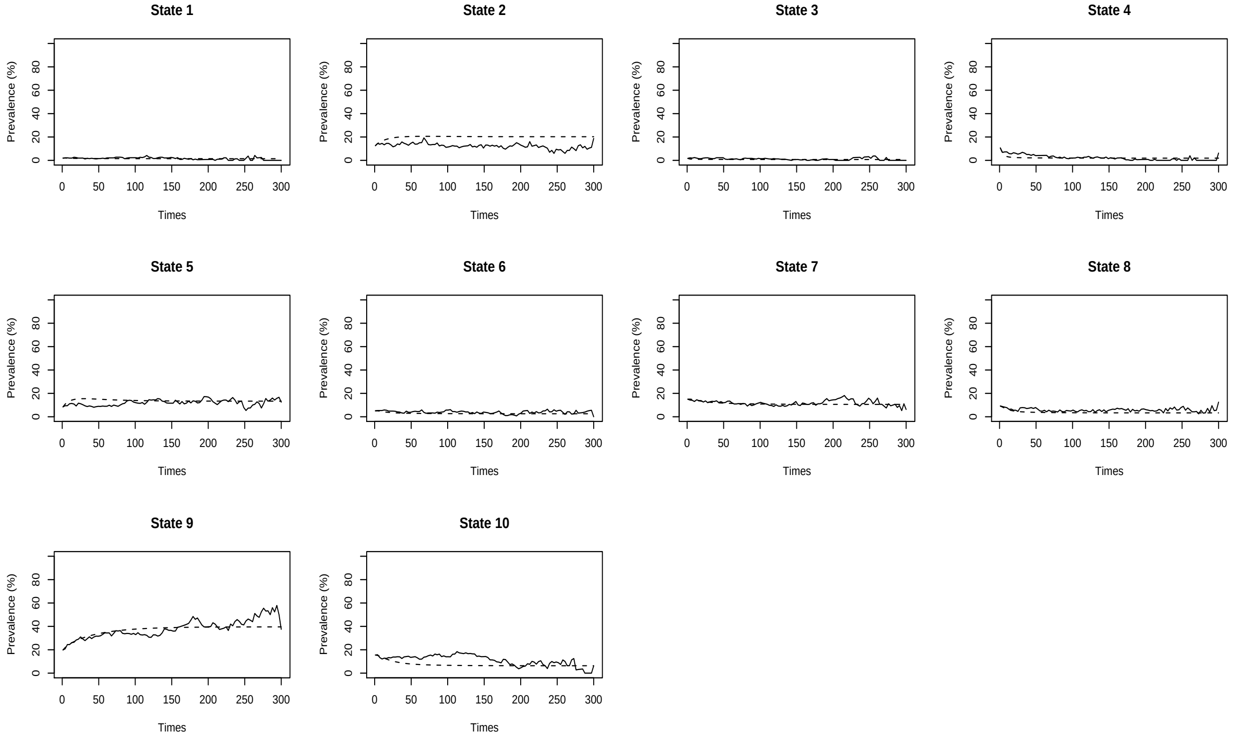


Figure 9: Prevalence of ten episodes of bipolar disorder in teenagers (age≥13): observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

B.0.5 With cohabitation

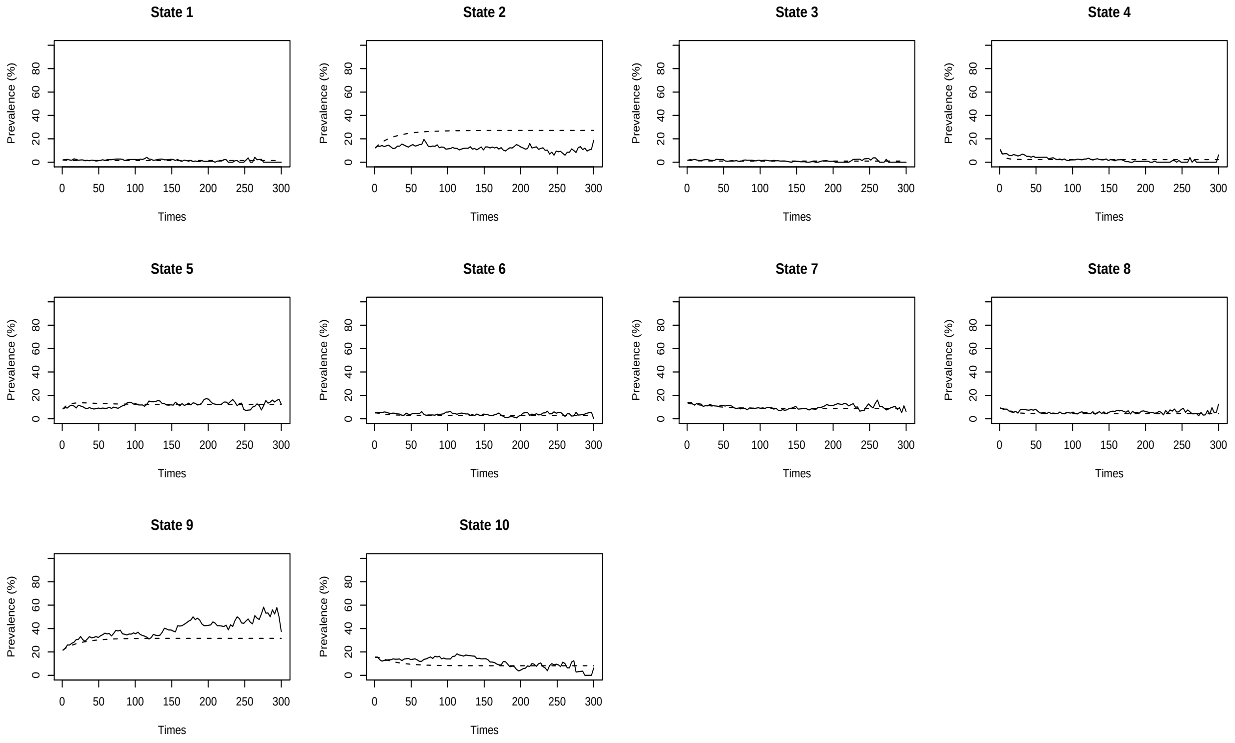


Figure 10: Prevalence of ten episodes of bipolar disorder in participants who live with both natural parents: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

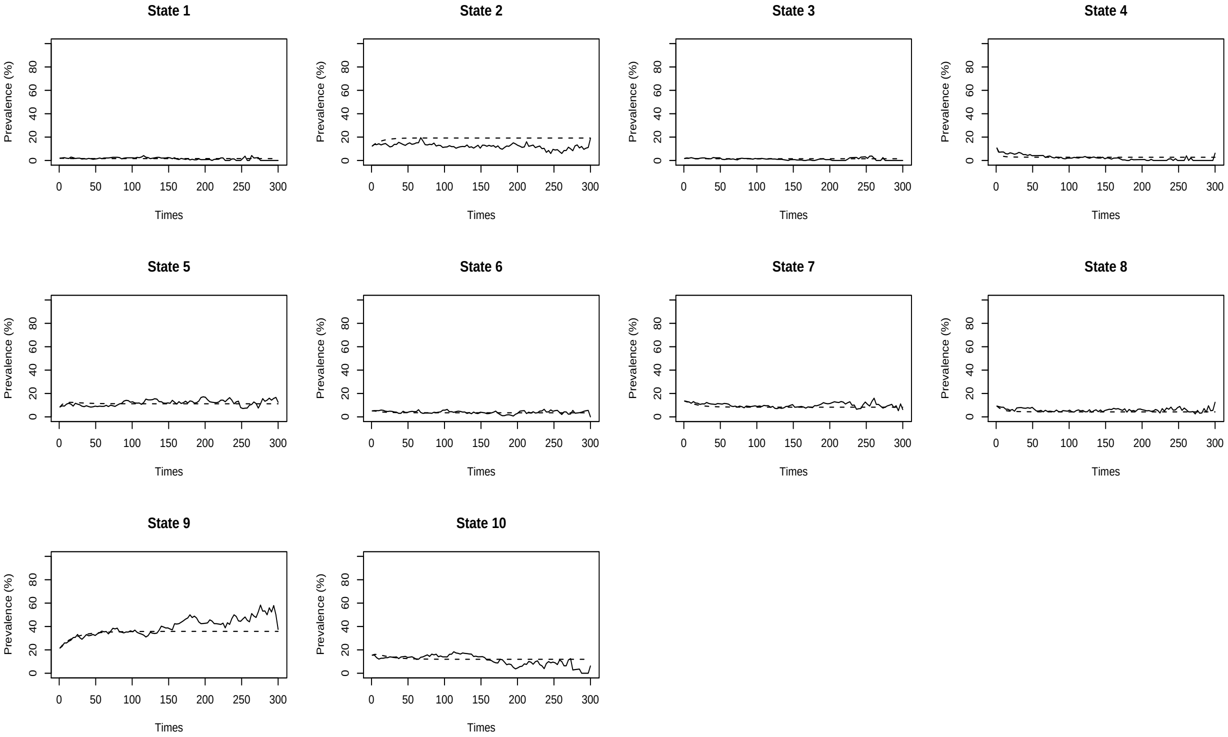


Figure 11: Prevalence of ten episodes of bipolar disorder in participants who live in another situation: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

B.0.6 With bipolar diagnosis

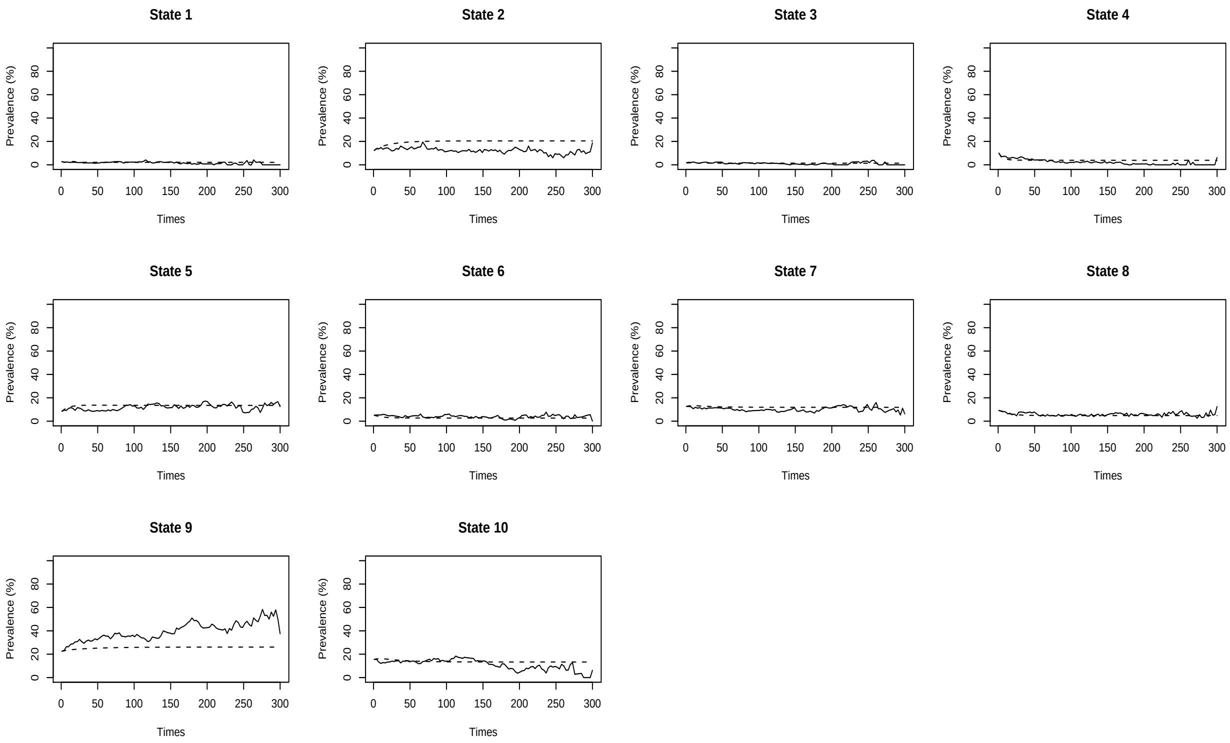


Figure 12: Prevalence of ten episodes of bipolar disorder in BPI patients: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

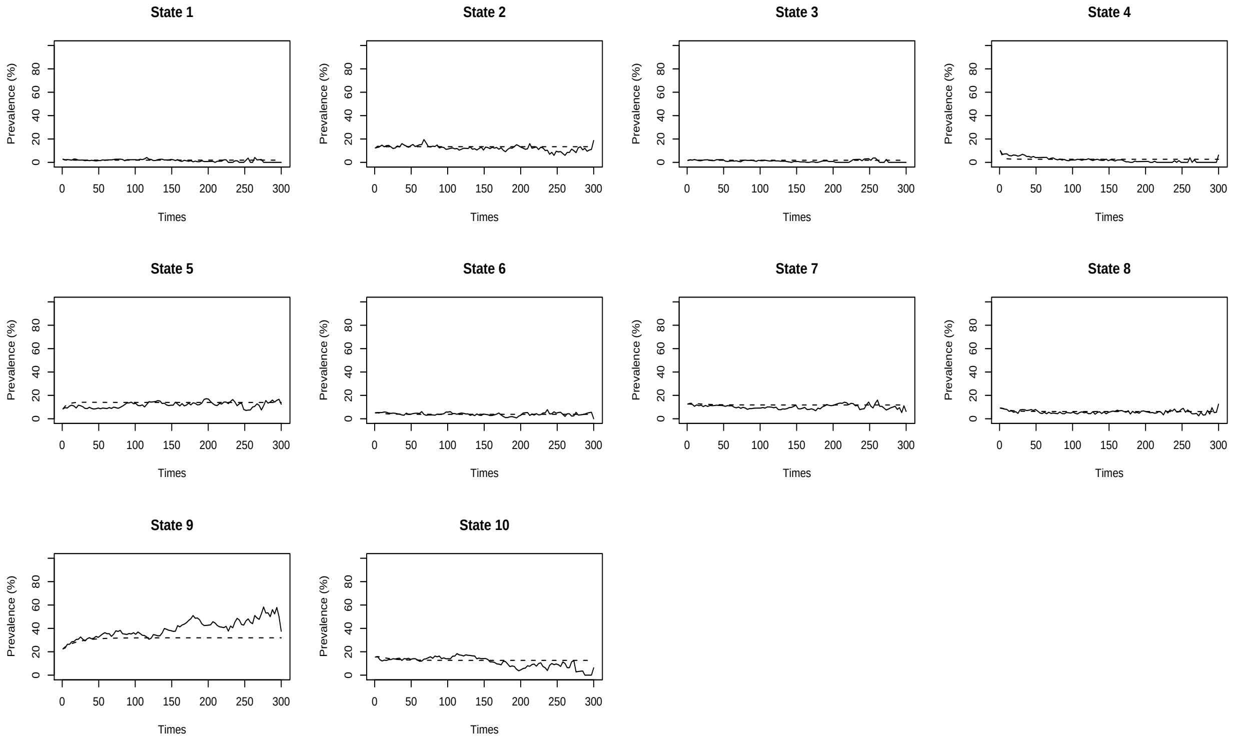


Figure 13: Prevalence of ten episodes of bipolar disorder in BPII patients: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

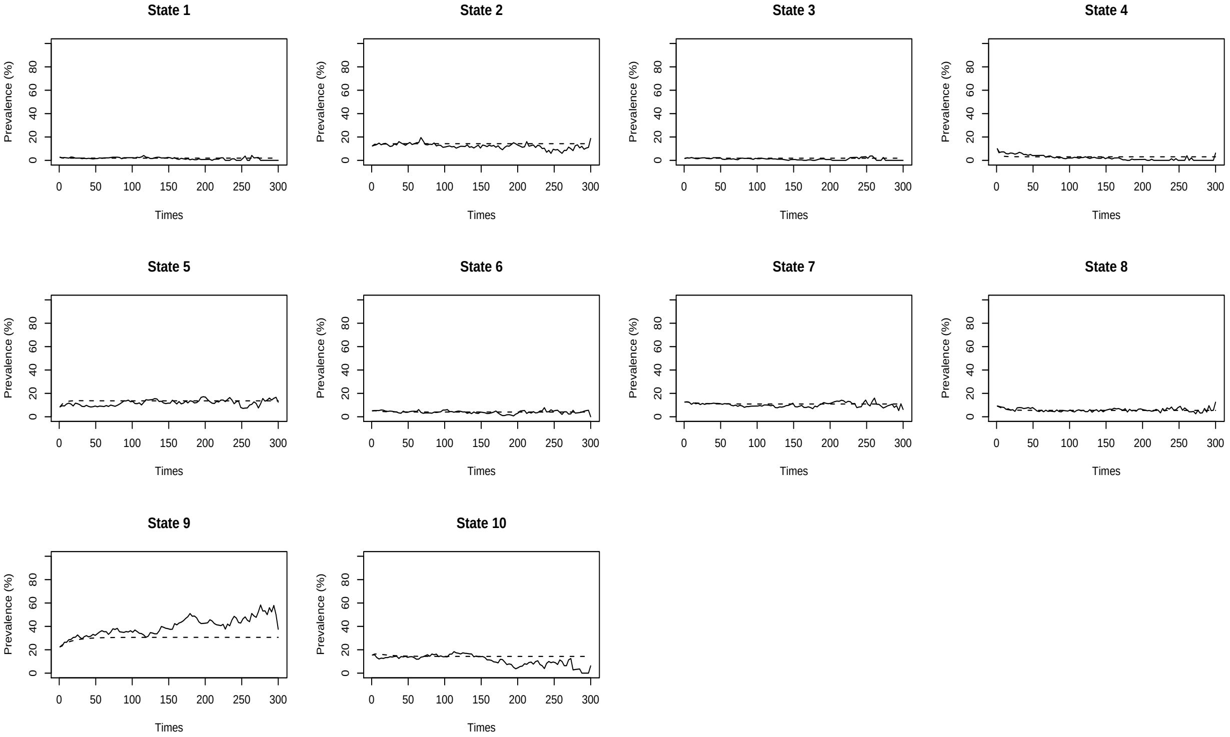


Figure 14: Prevalence of ten episodes of bipolar disorder in BPNOS patients: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

B.0.7 With age of bipolar onset

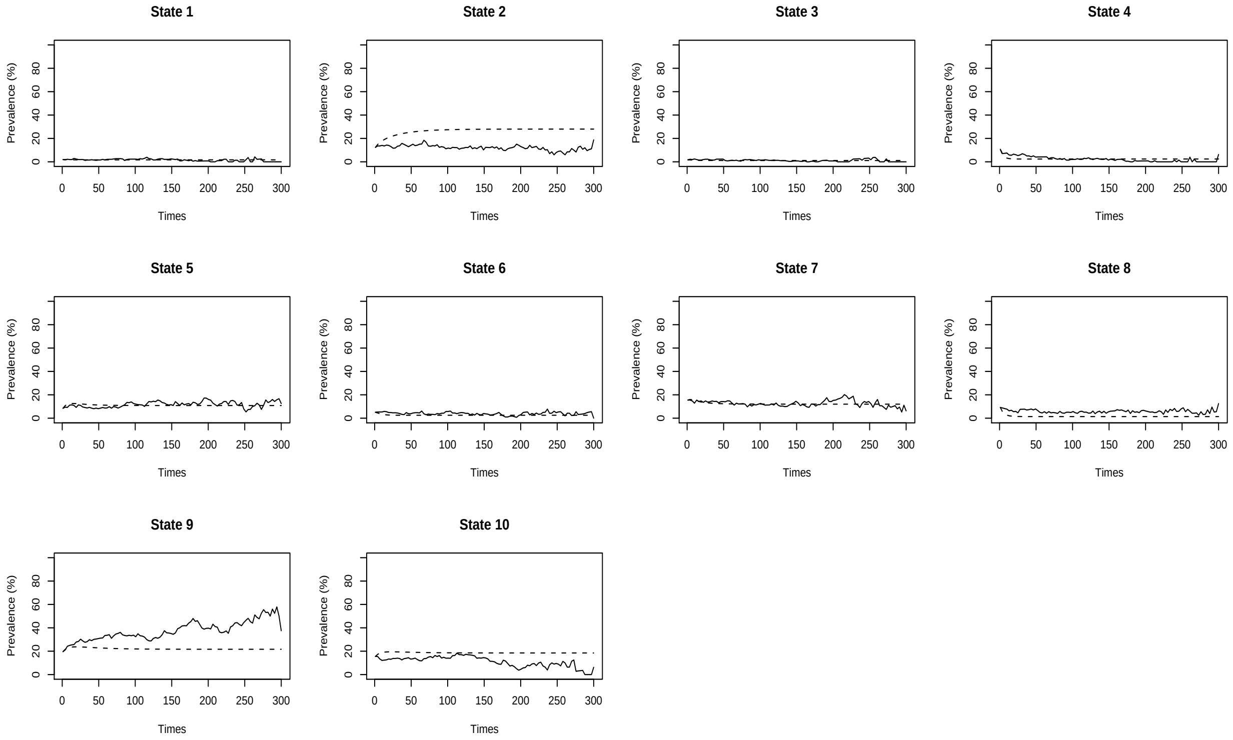


Figure 15: Prevalence of ten episodes of bipolar disorder in patients with bipolar onset during childhood: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

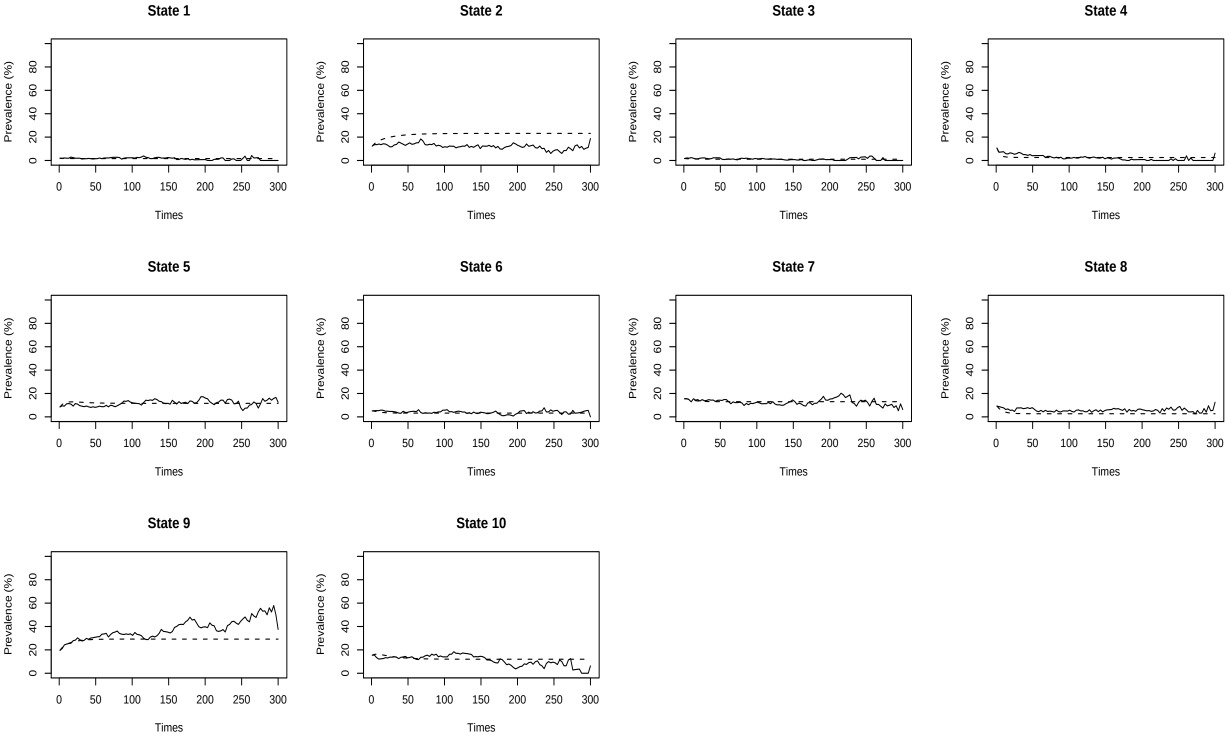


Figure 16: Prevalence of ten episodes of bipolar disorder in patients with bipolar onset during early adolescence: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

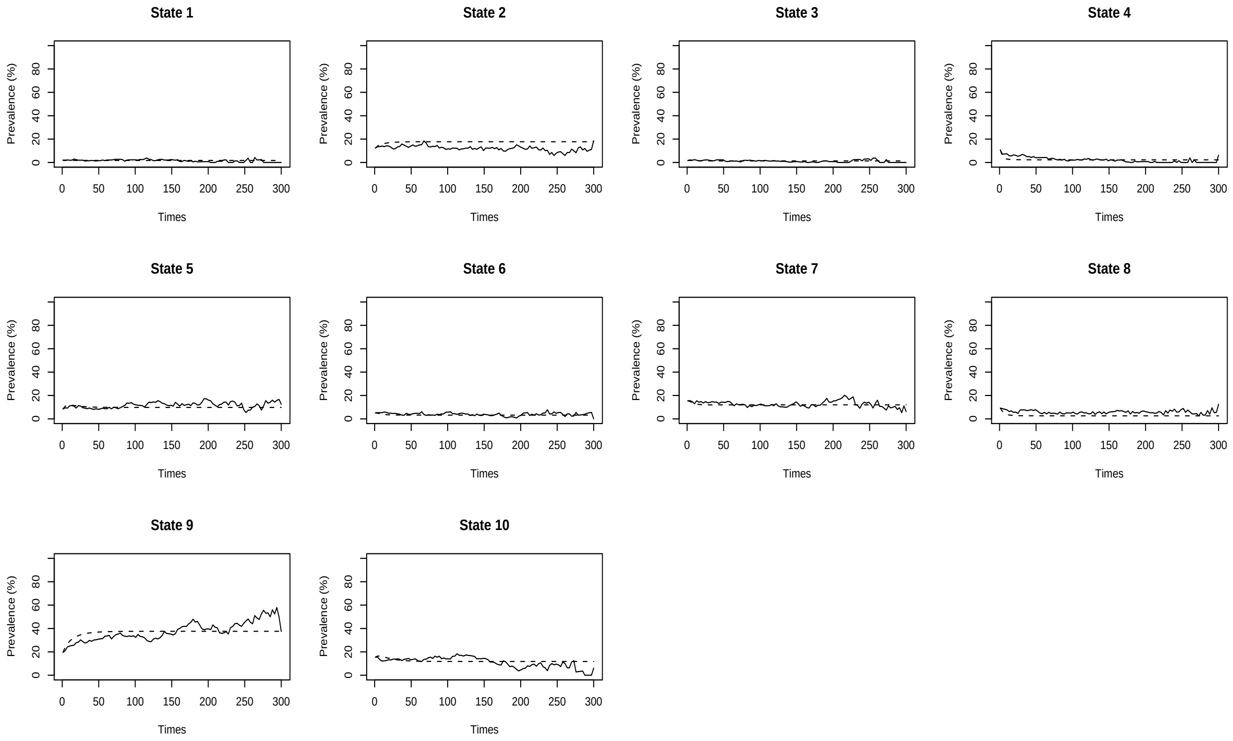


Figure 17: Prevalence of ten episodes of bipolar disorder in patients with bipolar onset during late adolescence: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

B.0.8 With socio-economic status

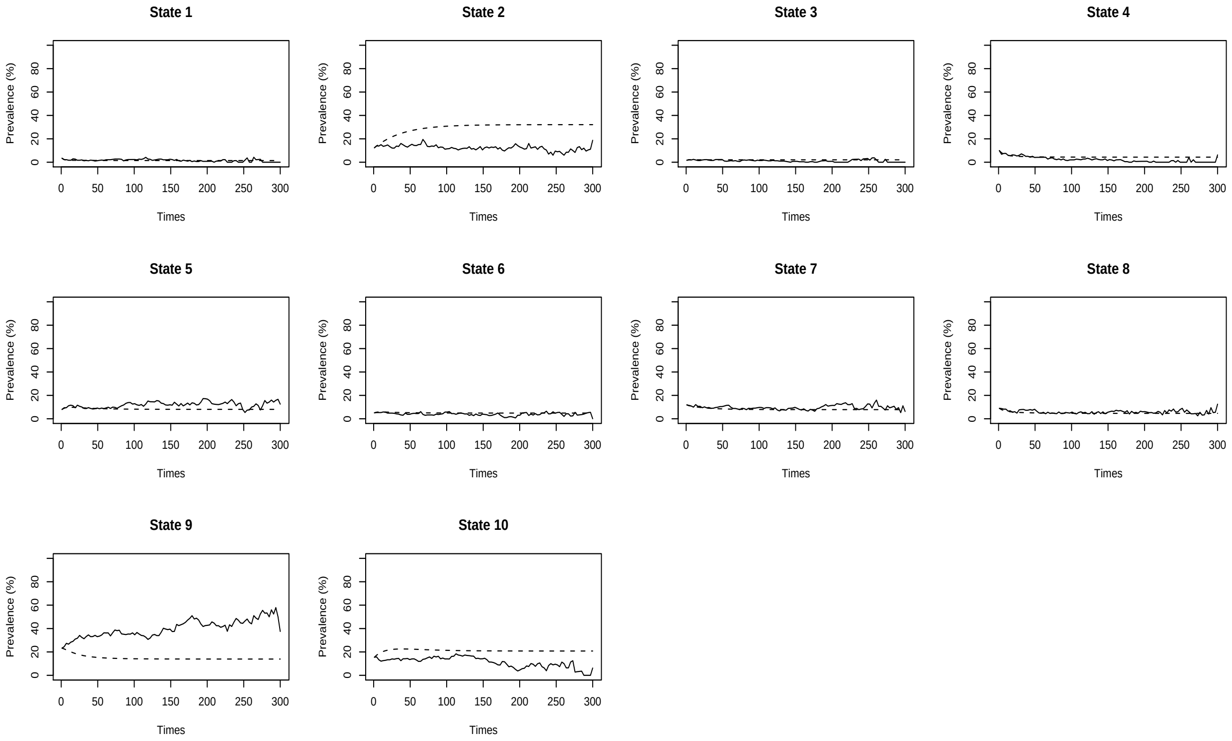


Figure 18: Prevalence of ten episodes of bipolar disorder in patients with socio-economic status 1: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

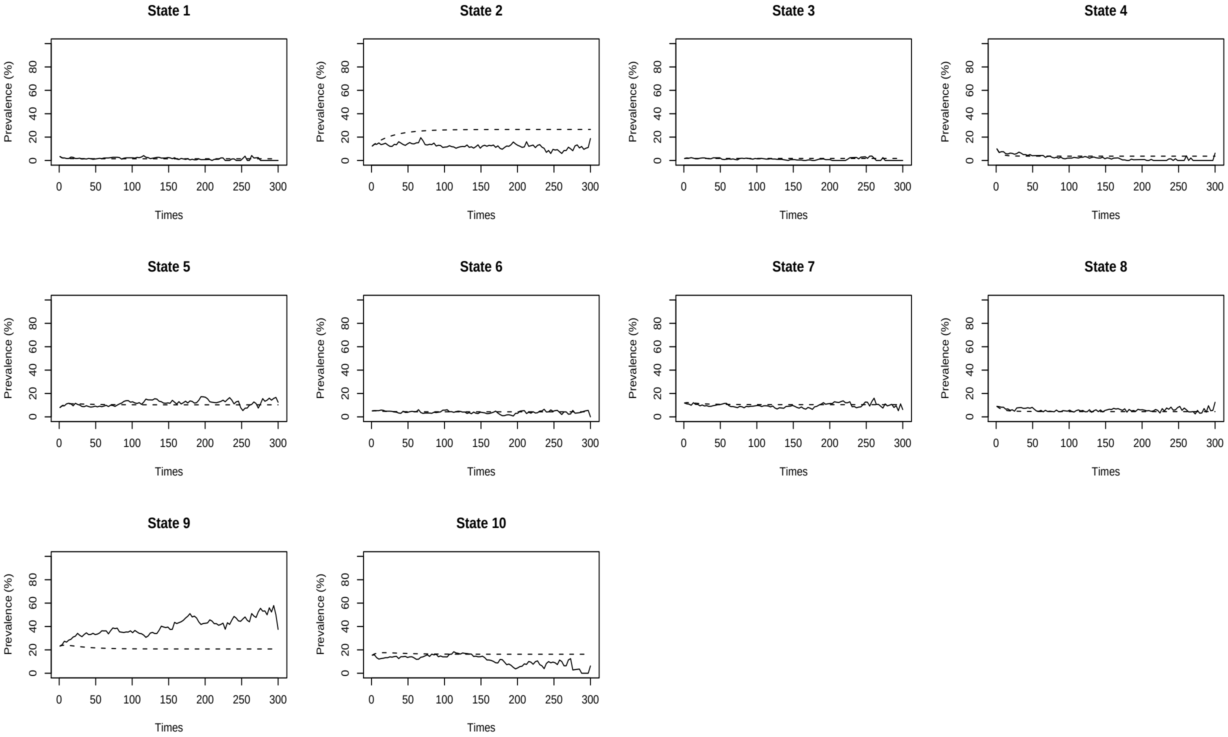


Figure 19: Prevalence of ten episodes of bipolar disorder in patients with socio-economic status 2: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

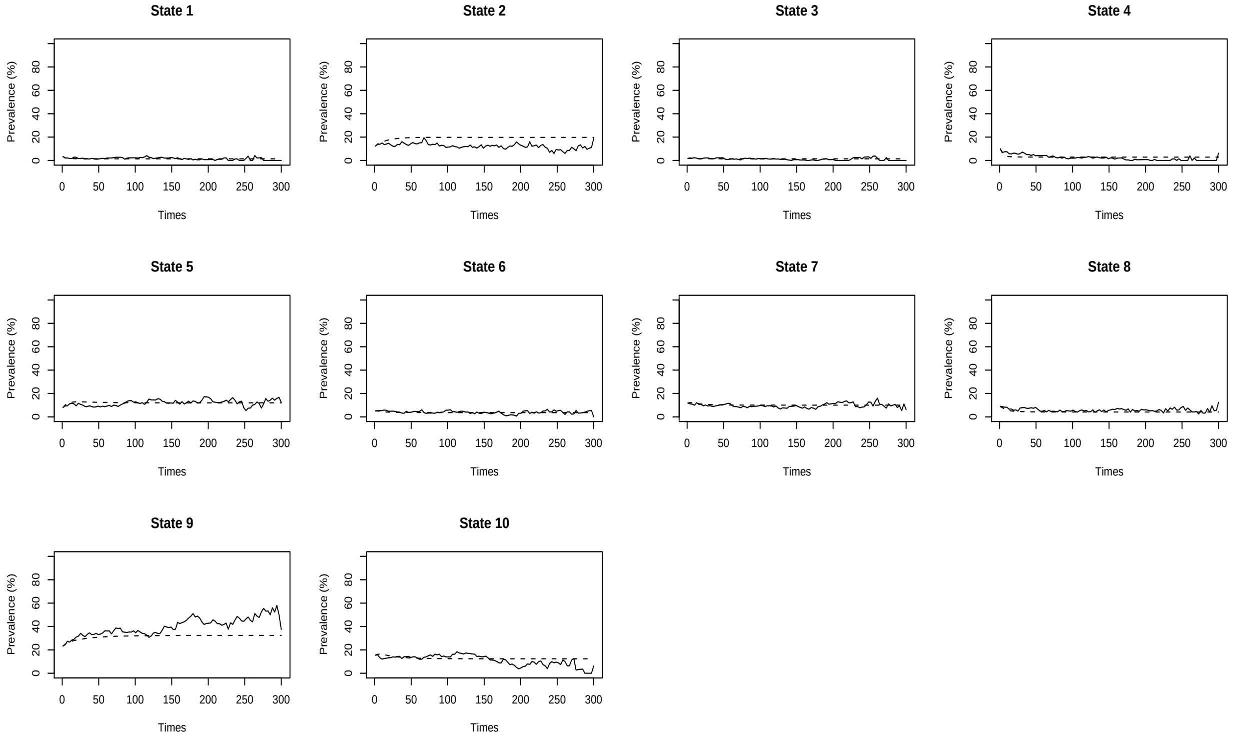


Figure 20: Prevalence of ten episodes of bipolar disorder in patients with socio-economic status 3: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

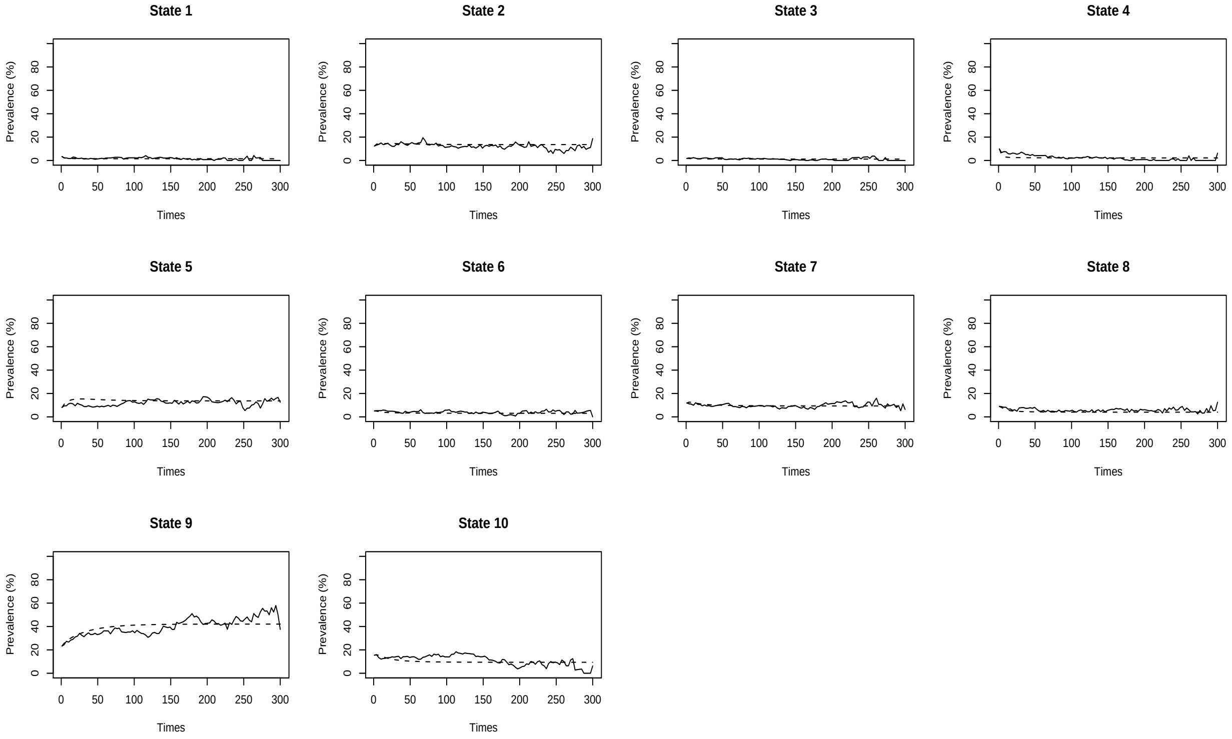


Figure 21: Prevalence of ten episodes of bipolar disorder in patients with socio-economic status 4: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)

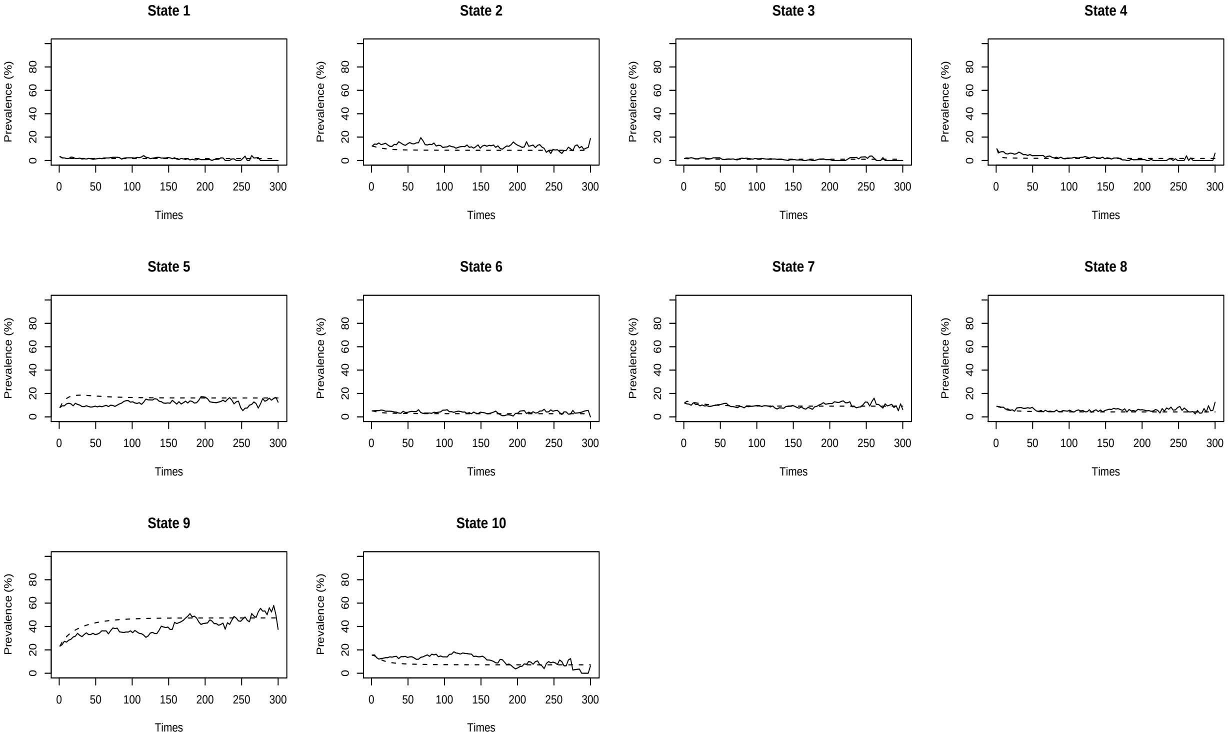


Figure 22: Prevalence of ten episodes of bipolar disorder in patients with socio-economic status 5: observed (solid line), estimated with the 10 states hidden Markov model (dashed line)