Dynamic Factor Analysis for Cognitive Trajectories*

Yorghos Tripodis† and Nikolaos Zirogiannis‡

Abstract

We propose a dynamic factor model appropriate for panel datasets and develop an estimation algorithm which can handle datasets with large number of subjects and short temporal information. The algorithm uses a two cycle iterative approach for model estimation in such a large dataset. Each iteration consists of two distinct cycles, both following an EM algorithm approach. This iterative process will continue until convergence is achieved. We utilized a dataset from the National Alzheimer Coordinating Center (NACC) to estimate underlying measures of cognition based on a battery of observed neuropsychological tests. We assess the goodness of fit and the precision of the dynamic factor model estimators and compare it with a non-dynamic version in which temporal information is not used. The dynamic factor model is superior to a non-dynamic version with respect to fit statistics shown in simulation experiments. Moreover, it has increased power to detect differences in the rate of decline for a given sample size.

Keywords: Dynamic Factor Models, EM algorithm, Panel Data, State-Space models, Cognition, Alzheimer's disease, neuropsychological performance

*This study was supported by U01 AG016976 through a Junior Investigator Grant and by a SAS-IIF grant award. Any remaining errors are ours.

†Department of Biostatistics, Boston University, yorghos@bu.edu.

‡School of Public and Environmental Affairs, Indiana University, nzirogia@indiana.edu.

1 Introduction

Alzheimers Disease (AD), the most common form of dementia, is a significant cause of disability and mortality among the elderly. The latest figures show that 5.2 million people in the US, approximately 14% of the population over age 70, are afflicted by AD (Alzheimer's-Association 2012). As the population ages over the next several decades, this number is expected to increase (Brookmeyer et al. 2007). The only definitive way to diagnose AD is post-mortem, but neuro-psychiatrists reach a pre-mortem diagnosis by reviewing and discussing the subject's clinical history, as well as scores from a variety of neuropsychological evaluation tests (Duara et al. 2010). Many observational and clinical trial studies of cognitive aging use neuropsychological test batteries to assess overall cognition and its specific domains (Snyder et al. 2011). The results of the neuropsychological tests which are part of the batteries can exhibit high within-subject variability (Behl et al. 2005) and may make diagnosis difficult. Moreover, the emphasis in Alzheimers disease clinical research has shifted to developing interventions before symptoms onset. In order to address this need, researchers are required to develop cognitive measures which discriminate between cognitively healthy subjects and individuals with small cognitive changes who will convert to mild cognitive impairment (MCI).

Statistical tools have been developed to extract information from these evaluation tests in order to estimate a single or multiple cognitive indices. Methods involving estimation of latent variables have been gaining attention in various fields of research. A common method for the estimation of such latent variables is confirmatory factor analysis (CFA) (Hayden et al. 2011; Park et al. 2012). The repeated nature of the tests is often ignored in these models, even though recent studies attempt to capture the temporal information in order to increase performance of measures for cognitive change (Proust et al. 2006; Locascio and Atri 2011; Johnson et al. 2012).

This article provides statistical tools which will advance our understanding of the longitudinal properties of cognitive trajectories in the normal and prodromal phase. Specifically, we develop an estimation algorithm for a longitudinal/dynamic factor analysis model which can be applied in studies with panel data. We apply the factor model to a variety of neuropsychological tests

using data from the National Alzheimers Coordinating Center (NACC) study and estimate a smooth cognitive measure for each individual's total cognition as well as measures for specific cognitive domains, such as memory, attention and language. We hypothesize that by incorporating longitudinal information into the factor models we increase the accuracy of the estimates of change over time and consequently increase power to detect differences between groups. We focus on a case-control sample using data where participants are selected to be cognitively normal. Cases in this study include participants that will convert to MCI after the period used in the analysis, while controls will continue having normal cognition for the next two follow-up visits after the end of the analytic period. In the next section we describe the dynamic factor model and its estimation method. In §3, we assess its performance in estimating the underlying cognition and its domains and compare it with a factor model which ignores any longitudinal information. We apply the dynamic and non-dynamic versions of the factor model to clinical data collected by NACC, and compare their power of detecting differences in the rate of cognitive change for various sample sizes. Finally in §4, we conclude with a discussion of the methods and results including limitations and directions for future studies.

2 Methods

In this section we describe the dynamic version of the factor model and its estimation process. The difference of the dynamic factor model for panel data from the non-dynamic version is that the former captures not only correlations between input variables but also autocorrelations and cross correlations of these variables of interest. We develop an iterative two-cycle estimation process, which is an extension of the ECME algorithm (Liu and Rubin 1998). This model is flexible enough to be applicable in studies with multiple individuals, short unequally spaced temporal information.

2.1 Model

We let $\mathbf{U}_t$ denote the $n \times 1$ vector containing the unobserved cognitive indices of n subjects at time t, with $t = 1, ..., T$. We assume that the dynamic properties of $\mathbf{U}_t$ can be captured by a Markov process. For illustration purposes, and without loss of generality, we first present the case where we have equally spaced observations and equal number of neuropsychological tests for each subject. We subsequently present the model for the general case with unequally spaced or missing observations. Hence, we form the following linear Gaussian state space model:

$$\mathbf{y}_t = \mathbf{B}\mathbf{U}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{D}), \tag{2.1}$$

$$\mathbf{U}_{t+1} = \mathbf{T}\mathbf{U}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{I}_n), \tag{2.2}$$

where $\mathbf{B}$ is the matrix of factor loadings with dimensions $np \times n$, with p denoting the number of observed variables, $\mathbf{y}_t$ is a $np \times 1$ vector of observed neuropsychological measures per individual, $\mathbf{T}$ is $n \times n$ transition matrix and $\mathbf{I}_n$ is a $n \times n$ identity matrix and $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ are error terms (Koopman 1993; Durbin and Koopman 2001). The state space formulation described in (2.1) and (2.2), models the behavior of the unobserved state vector $\mathbf{U}_t$ over time using the observed values $\mathbf{y}_1, ..., \mathbf{y}_n$. The state vector $\mathbf{U}_t$ is assumed to be independent of the error terms $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ for all t. In addition, the error terms $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ are assumed to be independent, identically distributed (i.i.d.) (Kohn and Ansley 1989; deJong 1991). In general, the model defined by equations (2.1) and (2.2) is not identifiable. Zirogiannis and Tripodis (2014) state the conditions for identifiability for a general dynamic factor model. In order for the model in (2.1) and (2.2) to be identifiable we must impose a certain structure. We first assume that the unobserved cognitive indices follow a multivariate random walk, so that $\mathbf{T} = \mathbf{I}_n$. This is a reasonable assumption when modeling cognition for an aging population where the spacing of the observation period is roughly annual. Similar non-stationary models for psychological constructs have been suggested by Molenaar and Campbell (2009) and used among others by Hekler et al. (2013) and Gu et al. (2014). We also impose a structure on the factor loading matrix $\mathbf{B}$ and the variance of the idiosyncratic errors $\mathbf{D}$.

We assume that the factor loadings for each observed variable are the same for each individual in the study. This assumption is necessary in order to have comparable estimated cognitive indices across individuals. We also assume that participants in the study are conditionally independent and that the variance of the idiosyncratic errors is the same for all individuals. These assumptions result in a block diagonal structure for $\mathbf{D}$. The imposed structure results to a model that is fully identifiable.

Unequally spaced and missing observations

It is very common in longitudinal observational studies to have unequally spaced or missing observations. Let τ_{it} be the distance between observations t and $t + \tau_{it}$ of the i^{th} subject, and $\boldsymbol{\tau}_t$ the vector with the distances between two subsequent observations at time t. Then we can re-write the state-space form of the multivariate random walks as:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{B}\mathbf{U}_t + \mathbf{e}_t, \qquad \mathbf{e}_{it} \sim N(\mathbf{0}, \mathbf{D}) \\ \mathbf{U}_{t+\tau} &= \mathbf{U}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\tau}_t) \end{aligned}$$

This time-varying model can be used for unequally spaced and missing observations, as well as for forecasting for any $\boldsymbol{\tau}_n$ steps ahead.

2.2 2-step modified ECME Algorithm

The high dimensionality of the data vector $\mathbf{y}_t$ makes estimation of our model rather problematic. Usual Newton-type gradient methods do not work in this situation creating the need for a novel estimation approach. We introduce a modified ECME algorithm that makes estimation of the model specified in (2.1) and (2.2), feasible through an iterative two-cycle process. The 2-cycle modified ECME algorithm is an extension of the ECME algorithm developed by Liu and Rubin (1998), which itself is an extension of the widely known EM algorithm (Dempster, Laird, and Rubin

1977). The modified ECME algorithm starts by partitioning the vector of unknown parameters $\boldsymbol{\Psi}$ into $(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$ where $\boldsymbol{\Psi}_1$ contains the elements of $\mathbf{D}$ that need to be estimated, while $\boldsymbol{\Psi}_2$ contains the relevant elements of $\mathbf{B}$. We use the term "cycle" as an intermediary between a "step" and an "iteration" as in Meng and Dyk (1997). In the case of our modified ECME algorithm, every iteration is comprised of two cycles. Each cycle includes one E-step and one M-step, where the first cycle estimates $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ given the estimates of $\boldsymbol{\Psi}$ of the previous *iteration*, while the second cycle estimates $\boldsymbol{\Psi}_2$ given the estimates of $\boldsymbol{\Psi}$ of the previous *cycle*.

The functional form of the complete-data log-likelihood at time period t is (McLachlan and Peel 2000):

$$\begin{aligned} \log\ell_t(\boldsymbol{\Psi}) &= \frac{1}{2}\log\{|\mathbf{D}^{-1}|\} - \frac{1}{2}\{(\mathbf{y}_t - \mathbf{B}\boldsymbol{u}_t)'\mathbf{D}^{-1}(\mathbf{y}_t - \mathbf{B}\boldsymbol{u}_t) \\ &\quad -(\boldsymbol{u}_{t+1} - \boldsymbol{u}_t)'(\boldsymbol{u}_{t+1} - \boldsymbol{u}_t) \end{aligned}$$

Since $\boldsymbol{u}_t$ is unobserved, we can consider it missing and use the EM algorithm framework. In order to find the MLE, we need to calculate the distribution of the latent variable $\boldsymbol{u}_t$ conditional on the observed values of $\mathbf{y}_t$. There is a long literature describing the EM procedure for factor analysis in cross-sectional data starting with Rubin and Thayer (1982). Applying the EM framework for longitudinal data we need to condition not only on the concurrent observed value of $\mathbf{y}_t$ but on all the previous observed history $\mathbf{y}_1, \ldots, \mathbf{y}_t$. As we will see in the following two subsections, we use the first cycle to obtain estimates for $\boldsymbol{u}_t$ by conditioning on the concurrent observed variables, $\mathbf{y}_t$, and the second to update these estimates by conditioning on the history of the observed variable, $\mathbf{y}_1, \ldots, \mathbf{y}_t$ using the Kalman filter (Kalman 1960). This iterative process will continue until the likelihood function stops increasing and convergence is achieved.

2.2.1 First cycle

During the k^{th} iteration of the first cycle, the E-step of the 2-cycle ECME algorithm is:

$$\mathbf{Z}_{\mathbf{\Psi}}(\mathbf{\Psi}_1, \mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) = \mathrm{E}_{\mathbf{\Psi}} \left\{ \sum_{t=1}^{T} \log \ell_t \left[(\mathbf{\Psi}_1, \mathbf{\Psi}_2) \, | \mathbf{y}_t, \, \mathbf{\Psi}_1^{(k-1)}, \, \mathbf{\Psi}_2^{(k-1)} \right] \right\}. \tag{2.3}$$

Following the notation presented in McLachlan and Peel (2000, p.242), the sufficient statistics are calculated in the $(k-1)$ iteration by the following equations:

$$\begin{aligned} \boldsymbol{\gamma}^{(k-1)} &= \left(\mathbf{B}^{(k-1)} \mathbf{B}^{(k-1)'} + \mathbf{D}^{(k-1)} \right)^{-1} \mathbf{B}^{(k-1)} \\ \boldsymbol{\omega}^{(k-1)} &= \mathbf{I} - \boldsymbol{\gamma}^{(k-1)'} \mathbf{B}^{(k-1)} \end{aligned} \tag{2.4}$$

The first M-step involves differentiating $\mathbf{Z}_{\mathbf{\Psi}}(\mathbf{\Psi}_1, \mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)})$ with respect to $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ in order to obtain $\mathbf{\Psi}_1^{(k)}$ and $\mathbf{\Psi}_2^{(k/2)}$:

$$\mathbf{Z}_{\mathbf{\Psi}}(\mathbf{\Psi}_1^{(k)}, \mathbf{\Psi}_2^{(k/2)}; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}) \geq \mathbf{Z}_{\mathbf{\Psi}}(\mathbf{\Psi}_1, \mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k-1)}), \tag{2.5}$$

The first-cycle M-step is identical to the M-step of the traditional EM algorithm for factor analysis models (McLachlan and Krishnan 2008):

$$\mathbf{B}^{(k/2)} = \mathbf{C}_{yy} \boldsymbol{\gamma}^{(k-1)} \left\{ \boldsymbol{\gamma}^{(k-1)'} \mathbf{C}_{yy} \boldsymbol{\gamma}^{(k-1)} + n \boldsymbol{\omega}^{(k-1)} \right\}^{-1}, \tag{2.6}$$

$$\mathbf{D}^{(k)} = n^{-1} \mathrm{diag} \left\{ \mathbf{C}_{yy} - \mathbf{C}_{yy} \boldsymbol{\gamma}^{(k-1)} \mathbf{B}' \right\}, \tag{2.7}$$

where $\mathbf{C}_{yy}$ is the sample unconditional covariance matrix of $\mathbf{Y}_T = (\mathbf{y}_1', ..., \mathbf{y}_T')' =$, i.e. $\mathrm{E}(\mathbf{Y}_T \mathbf{Y}_T{}') = \mathbf{C}_{yy}$. At the end of the first cycle we have updated estimates for all the elements of the variance-covariance matrix of the idiosyncratic errors, $\mathbf{D}$, and intermediate estimates for the matrix of factor loadings, $\mathbf{B}$. We use these estimates in the second cycle to get updated estimates for the factor loadings.

2.2.2 Second cycle

In the E-step of the second cycle we estimate $\mathbf{\Psi}_2^{(k)}$. We proceed by calculating:

$$\mathbf{Z}_{\mathbf{\Psi}_2}(\mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k)}, \mathbf{\Psi}_2^{(k/2)}) = \mathrm{E}_{\mathbf{\Psi}_2}\left\{\sum_{t=1}^{T} \ell_t \left[\mathbf{\Psi}_2 | \mathbf{Y}_{t-1}, \mathbf{\Psi}_1^{(k-1)}, \mathbf{\Psi}_2^{(k/2)}\right]\right\}. \tag{2.8}$$

The second E-step involves forming the expected complete-data log likelihood conditional on $\mathbf{Y}_{t-1}$, which is the set of past observations $\mathbf{y}_1, ..., \mathbf{y}_{t-1}$. The subsequent M-step involves differentiating $\mathbf{Z}_{\mathbf{\Psi}_2}(\mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k)}, \mathbf{\Psi}_2^{(k/2)})$ with respect to $\mathbf{\Psi}_2$. We choose $\mathbf{\Psi}_2^{(k)}$ such that:

$$\mathbf{Z}_{\mathbf{\Psi}_2}(\mathbf{\Psi}_2^{(k)}; \mathbf{\Psi}_1^{(k)}, \mathbf{\Psi}_2^{(k/2)}) \geq \mathbf{Z}_{\mathbf{\Psi}_2}(\mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k)}, \mathbf{\Psi}_2^{(k/2)}). \tag{2.9}$$

Upon maximization of $\mathbf{Z}_{\mathbf{\Psi}_2}$, the estimate $\mathbf{\Psi}_2^{(k)}$ is used in the E-step of the first cycle of the next iteration. We calculate and maximize $\mathbf{Z}_{\mathbf{\Psi}_2}(\mathbf{\Psi}_2; \mathbf{\Psi}_1^{(k)}, \mathbf{\Psi}_2^{(k/2)})$ by using the prediction error decomposition of the conditional likelihood (Harvey 1990):

$$\log \ell_t(\mathbf{\Psi}_2) = \log\frac{1}{2\pi} - \frac{1}{2}\left[\log|\mathbf{F}_t| + \boldsymbol{\upsilon}_t' \mathbf{F}_t^{-1} \boldsymbol{\upsilon}\right], \tag{2.10}$$

where $\boldsymbol{\upsilon}_t$ is the prediction error conditional on past history and $\mathbf{F}_t$ is its variance. Quantities, $\boldsymbol{\upsilon}_t$ and $\mathbf{F}_t$ can be estimated with the use of the Kalman filter, which is a set of recursions which allow information about the system to be updated every time an additional observation $\mathbf{Y}_t$ is introduced (Durbin and Koopman 2001, p.11). Once $\boldsymbol{\upsilon}_t$ and $\mathbf{F}_t$ are calculated, (2.10) is maximized with respect to $\mathbf{\Psi}_2$, as illustrated in (2.9).

3 Results

In the next section, we assess and apply the model and the estimation process described in §2. We first assess the performance of the 2-cycle ECME estimator using a simulation study. We then

apply the model in data from the NACC study. We also compare the dynamic factor model with a non-dynamic version in which temporal information is not used.

3.1 Simulation

The model from which we simulate is a variant used by Doz et al. (2011) which is based on a simulation scheme used by Stock and Watson (2002). We define:

$$\mathbf{B} = \begin{pmatrix} \mathbf{f} & 0 & \cdots & 0 \\ 0 & \mathbf{f} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \mathbf{f} \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} \mathbf{d} & 0 & \cdots & 0 \\ 0 & \mathbf{d} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \mathbf{d} \end{pmatrix}, \tag{3.1}$$

with

1. $\mathbf{f}$ a $p \times 1$ vector of factor loadings with $\mathbf{f}_{[k]} \sim \mathcal{U}(0,1)$ subject to $\sum_{k=1}^{p} \mathbf{f}_{[k]} = 1$,
2. $\mathbf{d}$ a $p \times p$ diagonal matrix of variances for the idiosyncratic elements, with $\mathbf{d}_{[k][k]} = \mathbf{f}_{[k]} \frac{\beta_k}{1-\beta_k}$ with $\beta_k \sim \mathcal{U}(0.1, 0.9)$

where $k = 1, \ldots, p$. We generate 1000 replicates from the model defined by (2.1), (2.2) and (3.1) with $\mathbf{U}_0 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_n)$, for different combination of sizes for observed tests, p, number of subjects n, and time points, T. Specifically, we use $p = 5, 10, 15$, $n = 10, 50, 100, 200, 300$ and $T = 3, 5, 7, 10, 15$. The choice of these values corresponds to our specific application. We specify factor loadings which are the same across individuals who do not share any familial or other relationship. The coefficient β_k is the ratio between the variance of the idiosyncratic component, $\mathbf{e}_t$, and the total variance of the corresponding observed variable, $\mathbf{Y}_t$. In the simulation, this ratio is drawn from a uniform distribution drawn from the interval $(0.1, 0.9)$. This interval was chosen in order to avoid parameters at the boundary of the parameter space.

Estimation was done using the 2-cycle modified ECME and we obtained estimates of the factor $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2, \ldots, \hat{\mathbf{U}}_T)'$. Performance was measured by the trace statistic:

$$\frac{\text{tr}(\mathbf{U}'\hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\mathbf{U})}{\text{tr}(\mathbf{U}'\mathbf{U})}.$$

The trace statistic is a multivariate version of the R^2 of the regression of the true factors on the estimated factors (Doz et al. 2011). A number close to 1 implies a good approximation of the estimated latent variable to the true factor. We used the trace statistic, TR_{DFA}, as a performance measure for the dynamic factor model. We also obtained estimates of the latent factor using a non-dynamic factor model defined only by equation (2.1). In order to have comparable results, we also used the 2-cycle modified ECME for the non-dynamic version of the factor model. We then calculated the equivalent trace statistic, TR_{CFA} for the non-dynamic model. We use the ratio of the two trace statistics, $\frac{TR_{DFA}}{TR_{CFA}}$, as a comparison measure of the two models. Values above 1 imply that the dynamic model has superior performance to the non-dynamic version, with respect to how well the estimated are close to the true factors.

Table 1 reports the results of the trace statistics from the simulation experiment. The numbers in the table refer to the average across 1000 replicates. As an example, we use the case of $T = 3$, $n = 300$, and $p = 5$; 86% of the variability of the true, simulated factor is explained by the factor estimated by DFA. The explained variability using DFA is 1% higher than the explained variability using CFA. The goodness of fit of the estimated factors, as measured with TR_{DFA}, increases with the size of individuals n in the sample, and the number of repeated observations per individual T. TR_{DFA} varies from 0.77 for a small n-small T sample to 0.97 for a moderate n-moderate T. The goodness of fit of the estimators does not improve as the number of observed tests p per individual increases for a given size n and T. Moreover, the dynamic factor model always performs at least as good as the non-dynamic version. The relative performance of the dynamic factor model increases with T. Based on $\frac{TR_{DFA}}{TR_{CFA}}$, the relative performance of the dynamic model ranges from 1% better goodness of fit compared to the non-dynamic version when $n = 10$, $p = 5$ and $T = 3$ to 9% when

$n = 300$, $p = 15$ and $T = 15$.

3.2 Application

We used the NACC dataset with visits from September 2005 to June 2013 for testing and evaluation. NACC serves as a repository for data collected at 34 past and present Alzheimer's Disease Centers (ADCs) throughout the United States. The ADCs conduct clinical and biomedical research on Alzheimer's disease and related disorders. Centers enroll their study subjects in various ways, including referral from clinicians, self-referral by patients themselves or concerned family members, active recruitment through community organizations, and volunteers who wish to contribute to research. Most centers also enroll volunteer control subjects. Study subjects at each center are best regarded as a case series, not necessarily representing all cases of disease in a defined population. For more information on the study see Morris et al. (2006).

We focus on a study sub-sample which includes cognitively healthy participants at initial visit. For all subjects we only considered their neuropsychological test results while cognitively healthy, even though some converted to mild cognitive impairment (MCI) state at a later visit. For those participants who did not convert to MCI during our observation period, we only considered those with at least 4 visits. To avoid the risk of healthy participants converting to MCI at a future visit beyond our observation period, we excluded the last two measurement occasions for that group. For those participants converted to MCI we considered those with at least 1 follow-up with normal cognition. We also excluded non-English speakers as well as subjects with a number of comorbidities such as stroke, Parkinson's disease, depression etc. We then created two balanced groups, with $n = 149$ each, matched by age, sex and education which differ only in their future cognitive state: one group will convert to MCI at the next visit (converters), while the other group will remain cognitively normal for at least the next two subsequent visits (non-converters). The description of the sample is given in figure 1. The mean (SD) age at initial visit is 75.7 (7.5) with 15.4 (2.5) average years of education. There are 170 (57.1%) women in the sample with 3.0 (1.2)

visits on average, and 2.3 (1.3) years of follow-up since the initial visit.

We considered four factor models using different neuropsychological measures according to their relation to a specific domain: i) memory, ii) attention-psychomotor speed, iii) language and iv) general cognition. For each factor model, we run both a dynamic and a non-dynamic version. For both versions of the model, the one step-ahead prediction errors were tested for normality and residual autocorrelation. Even though both the dynamic and the non-dynamic version of all four factor models indicated non-normal errors(e.g. for general cognition, p-value$<$0.0001 for Bowman-Shenton test for normality and for Box-Ljung portmanteau test for autocorrelation at lag 1), further investigation showed that this is caused by outliers from seven participants. These participants have significantly lower estimated factors at the last visit, which may indicate misdiagnosis or untimely diagnosis of MCI. For each neuropsychological test, we run a mixed effects regressions using PROC MIXED in SAS 9.3 with random intercepts and random slopes for time to test the hypothesis that there are significant differences in the rate of change by group (converters vs non-converters). We also used mixed effects regression on the factors estimated by the dynamic as well as the non-dynamic factor models. Table 2 shows the estimated annual rate of change for each of the neuropsychological tests and for the simple and dynamic factors. For ease of comparison, all outcomes have been standardized, using the mean and standard deviation of all cognitively healthy NACC participants. We note that there is no significant annual change for the group of non-converters for all neuropsychological measures, with the exception of logical memory: delayed, and for the estimated factors. For the converters, only MMSE, Trails B and Verbal Fluency Test: vegetables show significant decrease at the 5% level, while the factors from the simple (non-dynamic) factor model for attention and language show significant decrease over time. For the dynamic factor model estimates, all three domains and total cognition estimates show significant decreases over time for the group which progressed to MCI at the next follow-up period. Given that an important feature that leads to an MCI diagnosis is manifestation of significant cognitive decline, it is important to note that the the dynamic factor model estimates show evidence of decline even **before** conversion to MCI. We also note that both the dynamic and the non-dynamic version of

factor models show significant differences in the annual rate of change between groups. In general, the estimates of difference of the factor models are larger and have lower p-values than the estimates of input variables. Furthermore, the estimates of difference from the dynamic factor model are at least as high with larger p-values than the equivalent estimates of the simple factor models. This difference is due to the fact that DFM incorporates the longitudinal aspect of the psychometric results of every patient. This may be an indication of increased power for the dynamic factor model, which we explore in the next sub-section.

Power analysis

We also investigated the performance of the observed indicators and the estimated factors with respect to power. Our main aim remains to detect differences by group in the rate of change. In order to assess the power of each outcome, we follow a bootstrapping scheme using the data described in the previous section. We first assume that there is a difference in the annual rate of change between normal controls and MCI while they are both cognitively normal. For a given sample size n, we perform the following steps:

Simulation scheme

1. Select $n/2$ matched pairs with replacement.
2. Estimate factors for all domains and for total cognition using simple and dynamic factor models.
3. Run a mixed effect regression on the estimated factor using time since first visit, group (converters vs non-converters) and time×group interaction, along with age at initial visit as covariates.
4. Is the estimate of time×group interaction significant at the 5% error level?
5. Repeat for 1000 times.

Table 3 shows the power of detecting significant differences for different sample sizes for all outcomes. We note that power of the dynamic factor model estimates is higher than the power of the non-dynamic version. For the total summary index, the power for the dynamic version varies from 49.9% for $n = 120$ to 96.7% for $n = 240$. The power for the non-dynamic version is much lower and goes up to 76.9% for $n = 240$. These results indicate that a smaller sample size is required for a given power in order to find significant differences in the rate of change by groups. Using the results from table 3, we can calculate the required sample size for both the dynamic and the non-dynamic factor models for an 80% power at $\alpha = 5\%$. For the total cognition index, the DFA model requires a sample size of 187 while the non-dynamic version (CFA) requires a sample size of 252. We get similar results for the other domains: memory ($n_{DFA} = 370$, $n_{CFA} = 485$), attention ($n_{DFA} = 334$, $n_{CFA} = 546$), language ($n_{DFA} = 414$, $n_{CFA} = 1419$).

We also note that the power of the factor models is always higher than the power of the individual neuropsychological tests. This indicates that using factor models increases the power of detecting significant differences in the rate of change. One notable exception is the Boston Naming Test (BNT) in the language domain. BNT has a higher power for all sample sizes considered compared to the non-dynamic factor estimates. It also has a higher power than the dynamic factor model when $n = 120$. For larger sample sizes, the dynamic factor model estimates have higher power than BNT.

4 Conclusion

In this article, we developed an algorithm to estimate a dynamic factor model for latent cognitive variables. We compared it with equivalent factor models which do not use temporal information in the estimation, and showed that the dynamic factor model estimates are more accurate as reflected by comparison of fit statistics in simulation experiments. They are also more precise than the non-dynamic version estimates as shown by improved power to detect differences in the rate of decline. Since the estimated latent index is a weighted average of the concurrent observed

values, the reason for the improved performance of the dynamic factor model is due to the fact that the weighting scheme of DFA takes into account any within-subject variability over time and any cross-correlation of tests. In the non-dynamic version, weights depend on the correlation between tests as well as on between-subject variability. Measures that are highly correlated or have increased between-subject variability will receive higher weight. The main limitation with the non-dynamic approach is that we do not use any information from the within-subject variability over time. If we do not account for variability over time we may over(under)inflate the weights. In the dynamic factor model, the estimated latent variable is a weighted average of observed values from all time points. Concurrent values are weighted higher than observations further back into the past which will be discounted exponentially. The rate of discount will depend on the variability of each observed measure over time. For example, in the dynamic factor model, past observations of measures that are stable over time will be discounted less. Koopman and Harvey (2003) provide a general description of the weighting schemes for the model defined by equations (2.1) and (2.2).

The dynamic factor model can be extended to allow for observed variables loading to multiple factors or for studies where participants may be clustered due to familial or other relationship. The current model is applicable to data with short temporal component with unequally spaced observations. This is a particular strength of the estimation algorithm, since most of the observational studies on cognition have these specific characteristics. A limitation of the current study is that the estimated factor is not validated with changes in biomarkers, such as volumetric data from MRI scans. Additionally, even though NACC battery is well validated and we consider the tests which load to specific domains as known, this may not be true in other applications. Another limitation of this current study is the use of aggregate scores for each test rather than the scores of each specific item used in each test. Crane et al. (2008) show that it is advantageous to use the item scores to derive latent factors in longitudinal studies. Unfortunately, in the NACC study, as in many large studies, the data for items are not readily available for all participants. The methodology presented in this paper can be easily applied in most large studies where only the aggregate scores for each test are available. The dynamic factor model is particularly useful when we are interested

in finding differences in the rate of cognitive change between groups. This advantage can be used in future observational studies researching the heterogeneity in rates of progression of MCI and AD patients, or in future clinical trials that need to identify healthy participants at high risk of significant decline in cognition.

References

Alzheimer's-Association (2012). 2012 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association 8*(2), 131–168.

Behl, P., T. L. Stefurak, and S. E. Black (2005). Progress in clinical neurosciences: cognitive markers of progression in alzheimer's disease. *The Canadian journal of neurological sciences 32*(2), 140–151.

Brookmeyer, R., E. Johnson, K. Ziegler-Graham, and H. M. Arrighi (2007). Forecasting the global burden of alzheimer's disease. *Alzheimer's & Dementia 3*(3), 186–191.

Crane, P. K., K. Narasimhalu, L. E. Gibbons, D. M. Mungas, S. Haneuse, E. B. Larson, L. Kuller, K. Hall, and G. van Belle (2008, October). Item response theory facilitated co-calibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology 61*(10), 1018–27.e9.

deJong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics 19*(2), 1073–1083.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Doz, C., D. Giannone, and L. Reichlin (2011). A quasi maximum likelihood approach for large approximate dynamic factor models. *Review of Economics and Statistics*.

Duara, R., D. A. Loewenstein, M. Greig, A. Acevedo, E. Potter, J. Appel, A. Raj, J. Schinka, E. Schofield, W. Barker, Y. Wu, and H. Potter (2010). Reliability and validity of an algorithm for the diagnosis of normal cognition, MCI and dementia: Implications for multi-center research studies. *The American Journal of Geriatric Psychiatry 18*(4), 363–370.

Durbin, J. and S. Koopman (2001). *Time Series Analysis by State Space Methods*. Number 24 in Oxford Statistical Science Series. Oxford, U.K.: Oxford University Press.

Gu, F., K. Preacher, and E. Ferrer (2014). A state space modeling approach to mediation analysis. *Journal of Educational and Behavioral Statistics 39*(2), 117–143.

Harvey, A. (1990). *The Econometric Analysis of Time Series* (Second ed.). The MIT Press.

Hayden, K. M., R. N. S. Jones, C. Zimmer, B. L. Plassman, J. N. Browndyke, C. D. Pieper, L. H. Warren, and K. A. Welsh-Bohmer (2011). Factor structure of the national alzheimer's coordinating centers uniform dataset neuropsychological battery: An evaluation of invariance between and within groups over time. *Alzheimer Disease & Associated Disorders 25*(2), 128–137.

Hekler, E. B., M. P. Buman, N. Poothakandiyil, D. E. Rivera, J. M. Dzierzewski, A. A. Morgan, C. S. McCrae, B. L. Roberts, M. Marsiske, and P. R. Giacobbi (2013). Exploring behavioral markers of long-term physical activity maintenance a case study of system identification modeling within a behavioral intervention. *Health Education & Behavior 40*(1 suppl), 51S–62S.

Johnson, J. K., A. L. Gross, J. Pa, D. G. McLaren, L. Q. Park, J. J. Manly, and f. t. A. D. N. Initiative (2012). Longitudinal change in neuropsychological performance using latent growth models: a study of mild cognitive impairment. *Brain Imaging and Behavior 6*(4), 540–550.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*, 35–45.

Kohn, R. and C. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika 76*(1), 65–79.

Koopman, S. (1993). Disturbance smoother for state space models. *Biometrika 80*.

Koopman, S. J. and A. C. Harvey (2003). Computing observation weights for signal extraction and filtering. *Journal of Economic Dynamics and Control 27*(7), 1317–1333.

Liu, C. and D. Rubin (1998). Maximum likelihood estimation of factor analysis using the ecme alogorithm with complete and incomplete data. *Statistica Sinica 8*, 729–747.

Locascio, J. J. and A. Atri (2011). An overview of longitudinal data analysis methods for neurological research. *Dementia and Geriatric Cognitive Disorders 1*(1), 330–357.

McLachlan, G. and T. Krishnan (2008). *The EM Algorithm and Extensions* (2 edition ed.).

Hoboken, N.J: Wiley-Interscience.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models* (first ed.). Wiley-Interscience.

Meng, X. and D. V. Dyk (1997). The EM algorithm-an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(3), 511–567.

Molenaar, P. and C. Campbell (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science 18*(2), 112–117.

Morris, J. C., S. Weintraub, H. C. Chui, J. Cummings, C. DeCarli, S. Ferris, N. L. Foster, D. Galasko, N. Graff-Radford, E. R. Peskind, D. Beekly, E. M. Ramos, and W. A. Kukull (2006). The uniform data set (UDS): Clinical and cognitive variables and descriptive data from alzheimer disease centers:. *Alzheimer Disease & Associated Disorders 20*(4), 210–216.

Park, L. Q., A. L. Gross, D. McLaren, J. Pa, J. K. Johnson, M. Mitchell, and J. J. Manly (2012). Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging and Behavior 6*(4), 528–539.

Proust, C., H. Jacqmin-Gadda, J. M. G. Taylor, J. Ganiayre, and D. Commenges (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics 62*(4), 1014–1024.

Rubin, D. and D. Thayer (1982). EM algorithms for ML factor analysis. *Psychometrika 47*(1), 69–76.

Snyder, P. J., C. E. Jackson, R. C. Petersen, A. S. Khachaturian, J. Kaye, M. S. Albert, and S. Weintraub (2011). Assessment of cognition in mild cognitive impairment: A comparative study. *Alzheimer's & Dementia 7*(3), 338–355.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Zirogiannis, N. and Y. Tripodis (2014). Dynamic factor analysis for short panels: Estimating performance trajectories for water utilities. In *Proceedings AAEA Annual Meeting, Minneapolis, MN, July 27-29, 2014*.

Figure 1: Description of analytic sample

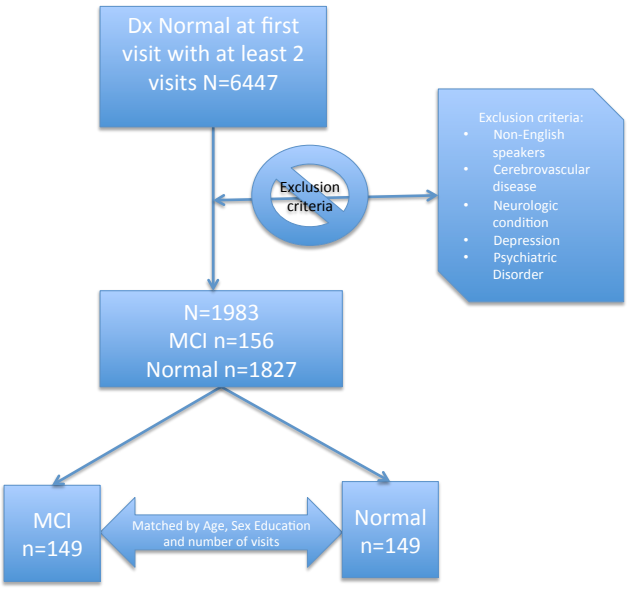

Table 1: Performance of factor estimators from 1000 simulations (using latest code)

		T=3		T=5		T=7		T=10		T=15	
		TR_{DFA}	$\frac{TR_{DFA}}{TR_{CFA}}$	TR_{DFA}	$\frac{TR_{DFA}}{TR_{CFA}}$	TR_{DFA}	$\frac{TR_{DFA}}{TR_{CFA}}$	TR_{DFA}	$\frac{TR_{DFA}}{TR_{CFA}}$	TR_{DFA}	$\frac{TR_{DFA}}{TR_{CFA}}$
n=10	p=5	0.77	1.00	0.80	1.02	0.83	1.02	0.86	1.02	0.88	1.01
	p=10	0.77	1.00	0.78	1.02	0.81	1.03	0.83	1.03	0.86	1.02
	p=15	0.77	1.00	0.77	1.02	0.78	1.03	0.81	1.03	0.84	1.02
n=50	p=5	0.85	1.01	0.87	1.02	0.90	1.02	0.92	1.02	0.93	1.01
	p=10	0.85	1.01	0.86	1.02	0.87	1.03	0.90	1.03	0.92	1.02
	p=15	0.85	1.01	0.85	1.02	0.86	1.03	0.88	1.03	0.91	1.02
n=100	p=5	0.86	1.01	0.89	1.02	0.91	1.02	0.93	1.02	0.95	1.01
	p=10	0.86	1.01	0.86	1.02	0.89	1.03	0.92	1.03	0.94	1.02
	p=15	0.87	1.01	0.86	1.02	0.88	1.03	0.90	1.03	0.93	1.03
n=200	p=5	0.88	1.01	0.90	1.02	0.92	1.02	0.94	1.02	0.96	1.02
	p=10	0.87	1.01	0.89	1.02	0.91	1.03	0.93	1.03	0.95	1.02
	p=15	0.89	1.01	0.88	1.02	0.89	1.03	0.92	1.04	0.93	1.03
n=300	p=5	0.86	1.01	0.91	1.02	0.93	1.03	0.95	1.05	0.96	1.03
	p=10	0.87	1.01	0.90	1.02	0.91	1.03	0.94	1.03	0.97	1.07
	p=15	0.90	1.01	0.89	1.02	0.91	1.03	0.93	1.04	0.96	1.09

Table 2: Parameter estimates for annual rate of change for all neuropsychological tests

Domain	Test	Did not progress to MCI		Progress to MCI		Difference	
		Estimate (SE)	p-value	Estimate (SE)	p-value	Estimate (SE)	p-value
Memory	MMSE	-0.02 (0.03)	0.390	-0.06 (0.03)	0.027	0.04 (0.04)	0.343
	Logical Memory: Immediate	0.04 (0.03)	0.202	-0.04 (0.03)	0.140	0.08 (0.04)	0.052
	Logical Memory: Delayed	0.06 (0.03)	0.035	-0.02 (0.03)	0.404	0.08 (0.04)	0.038
	Factor CFA	0.04 (0.03)	0.133	-0.04 (0.03)	0.148	0.08 (0.04)	0.037
	Factor DFA	0.05 (0.04)	0.253	-0.09 (0.04)	0.026	0.14 (0.06)	0.020
Attention-Psychomotor Speed	Digits Forward	-0.04 (0.02)	0.127	-0.03 (0.02)	0.268	-0.01 (0.03)	0.770
	Digits Backward	0.01 (0.03)	0.680	-0.03 (0.03)	0.201	0.04 (0.04)	0.231
	WAIS	-0.03 (0.02)	0.073	-0.03 (0.02)	0.124	0.01 (0.02)	0.839
	TRAILS A	0.03 (0.03)	0.196	-0.01 (0.03)	0.599	0.05 (0.04)	0.199
	TRAILS B	0.01 (0.02)	0.742	-0.05 (0.02)	0.039	0.03 (0.03)	0.089
	Factor CFA	0.00 (0.02)	0.852	-0.04 (0.02)	0.001	0.04 (0.02)	0.048
	Factor DFA	-0.00 (0.03)	0.914	-0.08 (0.03)	0.003	0.04 (0.04)	0.046
Language	Animals	-0.00 (0.02)	0.847	-0.04 (0.02)	0.056	0.04 (0.03)	0.222
	Vegetables	-0.02 (0.02)	0.493	-0.05 (0.03)	0.029	0.04 (0.04)	0.285
	Boston Naming Test	0.04 (0.03)	0.100	-0.02 (0.03)	0.285	0.06	0.076
	Factor CFA	-0.00 (0.02)	0.916	-0.04 (0.02)	0.014	0.04 (0.03)	0.094
	Factor DFA	-0.02 (0.03)	0.515	-0.09 (0.03)	0.001	0.07 (0.04)	0.069
Total	*Factor CFA*	0.05 (0.02)	0.058	-0.04 (0.02)	0.098	0.09 (0.03)	0.012
	Factor DFA	0.01 (0.03)	0.654	-0.12 (0.03)	<.001	0.13 (0.05)	0.004

Table 3: Power analysis for group differences

Domain	Test	$n = 120$	$n = 150$	$n = 180$	$n = 210$	$n = 240$
Memory	MMSE	4.9	3.00	3.10	1.30	0.90
	Logical Memory: Immediate	15.0	20.3	23.1	25.2	29.8
	Logical Memory: Delayed	14.4	20.2	24.9	28.4	34.3
	Factor-CFA	17.2	22.2	26.8	32.2	38.1
	Factor-DFA	21.9	28.7	33.7	40.1	51.0
Attention-Speed	Digits Forward	2.3	0.9	1.00	0.40	0.00
	Digits Backward	4.4	4.1	3.9	1.80	1.90
	WAIS	1.7	0.7	0.2	0.00	0.00
	TRAILS A	12.6	14.8	16.1	18.8	16.6
	TRAILS B	12.5	13.4	17.1	15.8	16.6
	Factor-CFA	15.2	17.7	23.7	28.3	32.8
	Factor-DFA	21.5	24.4	30.6	45.9	52.4
Language	Animals	5.6	5.4	5.7	5.60	2.70
	Vegetables	5.8	4.0	4.4	3.00	1.40
	Boston Naming Test	19.1	20.8	24.5	23.4	22.1
	Factor-CFA	10.3	13.2	12.6	15.9	15.7
	Factor-DFA	14.8	20.8	31.7	37.0	37.1
Total	*Factor CFA*	39.3	49.2	59.0	65.5	76.9
	Factord DFA	49.9	66.0	81.1	92.8	96.7