PSYCHOMETRIKA—VOL. 87, NO. 2, 403–431
JUNE 2022
https://doi.org/10.1007/s11336-021-09825-7

PENALIZED ESTIMATION AND FORECASTING OF MULTIPLE SUBJECT INTENSIVE LONGITUDINAL DATA

ZACHARY F. FISHER

THE PENNSYLVANIA STATE UNIVERSITY

YOUNGHOON KIM, BARBARA L. FREDRICKSON AND VLADAS PIPIRAS

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Intensive longitudinal data (ILD) is an increasingly common data type in the social and behavioral sciences. Despite the many benefits these data provide, little work has been dedicated to realize the potential such data hold for forecasting dynamic processes at the individual level. To address this gap in the literature, we present the *multi-VAR framework*, a novel methodological approach allowing for penalized estimation of ILD collected from multiple individuals. Importantly, our approach estimates models for all individuals simultaneously and is capable of adaptively adjusting to the amount of heterogeneity present across individual dynamic processes. To accomplish this, we propose a novel proximal gradient descent algorithm for solving the multi-VAR problem and prove the consistency of the recovered transition matrices. We evaluate the forecasting performance of our method in comparison with a number of benchmark methods and provide an illustrative example involving the day-to-day emotional experiences of 16 individuals over an 11-week period.

Key words: ILD, forecasting, time series, regularization, LASSO, proximal gradient descent, vector autoregression.

Intensive longitudinal data (ILD) is increasingly available to social and behavioral scientists. With this increased availability come new opportunities for modeling and predicting complex biological, behavioral and physiological phenomena. Despite these new opportunities psychological researchers have not taken full advantage of promising opportunities inherent to this data, the potential to forecast psychological processes at the individual level. To address this gap in the literature, we present a novel modeling framework which addresses a number of topical challenges and open questions in the psychological literature on modeling dynamic processes. First, how can we model and forecast ILD when the length of individual time series and the number of variables collected are roughly equivalent, or when time series lengths are shorter than what is typically required for time series analyses? Second, how can we best take advantage of the cross-sectional (between-person) information inherent to most ILD scenarios while acknowledging individuals differ both quantitatively (e.g., in parameter magnitude) and qualitatively (e.g., in structural dynamics)? Despite the acknowledged between-person heterogeneity in many psychological processes, is it possible to leverage group-level information to support improved forecasting at the individual level? In the remainder of the manuscript, we attempt to address these and other pressing questions relevant to the forecasting of multiple-subject ILD.

Forecasting in Psychology

Technological developments have significantly eased the burden of collecting intensive longitudinal data (ILD) for psychological researchers. This includes sensor-based physiological measurements, health and movement data, measures of behavioral and emotional states, as well as

Correspondence should be made to Zachary F. Fisher, The Pennsylvania State University, Pennsylvania, USA. Email: fish.zachary@gmail.com

data from many other noisy and complex systems. Increased availability has brought with it the realization that ILD presents unique opportunities for psychological scientists looking to model, forecast and modify complex time-dependent processes. Despite this realization the lion's share of methods development within psychology has focused exclusively on explanation. That is, psychological researchers have primarily been concerned with the characterization of dynamic processes using a combination of theoretical knowledge and measures of model fit to guide model construction.

Despite this focus on explaining the past over predicting the future the development of modern forecasting methods specifically tailored to psychological data hold great promise for the field. For example, the accurate prediction of emotional and physiological states would be an invaluable tool for clinicians tasked with monitoring and intervening on individual behavior. Furthermore, accurate forecasts are helpful for identifying when and to whom an intervention should be applied. Forecasting also presents psychologists with a practical framework to assess conflicting evidence from empirical studies and competing causal theories. In this paper, we will focus specifically on forecasting daily measures of emotion dynamics and psychopathology, addressing some of the unique challenges inherent to this type of data.

Vector Autoregressive Models in Psychology

In the social science and behavioral sciences, vector autoregressive (VAR) models and their many flavors (e.g., Structural-VAR, graphical-VAR, time-varying-VAR) have become a common approach for modeling ILD. VAR models have been used to model binge eating behaviors (Wild et al., 2010), dynamics among mother–infant dyads (Ji et al., 2020; Chen et al., 2020), substance use patterns (Zheng et al., 2013), and persistent depressive symptoms (Groen et al., 2019), to name a few. VAR models are a natural fit for many idiographic analyses as they provide a concise interpretation of inter-variable relations. They are also visualized easily using path or network connection diagrams, and allow for the inclusion of many *potentially relevant* variables. This is useful when theory does not give concrete guidance on whether a variable is related to the process under study.

VAR models are also a mainstay of forecasting in many fields. Consider econometrics, for example, where the widespread adoption of VAR models in the mid-1980s marked the beginning of a boon in forecasting practice (Allen & Morzuch, 2006). These are just a few reasons why VAR models represent a natural jumping off point for applied social science researchers looking to apply forecasting methodologies in their work. However, there are a number of features common to ILD research which deserve additional attention in the context of VAR modeling.

The first issue we address was described by Sims (1980) as the "profligate parameterization" of the unrestricted or canonical VAR model. Indeed, the number of VAR parameters grows quadratically with each component series added to the system of equations. In this way, the flexibility of the VAR model specification is also its Achilles' heel, and there are a large number of unknown coefficients relative to the information available from the data. This imbalance can lead to overfitting the sample data and poor forecasting performance (Robertson & Tallman, 2001). This presents a potential problem for many ILD scenarios where time series lengths typically fall between 30 and 100 measurement occasions and many variables are collected (e.g., a 10–20 item scale). In other words, employing the VAR model in applied research can be a delicate operation. One wants to include all relevant variables in a model to ensure the dynamics under study are well-captured; however, stringent theoretically motivated restrictions are generally required to obtain a useful model.

The second issue our proposed method aims to address is that of multiple-subject ILD, and more specifically how to best utilize cross-sectional information when modeling intra-individual processes. This is a fundamental question in both psychology and time series analysis. In psychol-

ogy, much attention has been paid to multilevel modeling as a means to synthesize time series data collected on multiple individuals (Bringmann et al., 2013; Epskamp et al., 2018). This approach is promising when the number of variables in the analysis is not large and individuals do not differ substantially in terms of their overall model structure. Another approach for leveraging cross-sectional information for multivariate time series modeling at the individual-level is Group Iterative Multiple Model Estimation (GIMME; Gates & Molenaar, 2012). The GIMME approach is built on the structural-VAR (S-VAR) framework and is available in the `gimme` package (Lane et al., 2019).

Foundations of the Proposed Approach

With our approach, we hope to retain the features of VAR modeling that are so attractive to social science researchers while confronting the problem of overparameterization. To accomplish this, we turn to methods that induce sparsity on the VAR parameter space through regularized estimation (Basu & Michailidis, 2015a, 2015b). Although it is also possible to address this issue by imposing some lower-dimensional structure on the data matrix, as in dynamic factor analysis (Molenaar, 1985; Stock & Watson, 2002), or by combining dimension reduction and VAR modeling (Bulteel et al., 2018b), we focus our attention on the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) and adaptive LASSO (Zou, 2006) frameworks. Although originally developed for cross-sectional data, these methods have been readily applied in the domain of multivariate time series analysis, and a number of authors have found these methods to be successful in forecasting applications. For example, Li and Chen (2014) found standard LASSO methods outperformed dynamic factor models in out-of-sample forecasting and Medeiros and Mendes (2016); Kock and Callot (2015) found the adaptive LASSO (Zou, 2006) to outperform standard forecasting approaches in both simulation studies and real-world data problems.

In the cross-sectional setting, a number of authors have considered applying LASSO methods to data that arises from some fixed number of groups (Gross & Tibshirani, 2016; Ollier & Viallon, 2017). These groups may represent different cohorts of individuals or different genres of movies; however, the underlying theme of these approaches is that we might learn more about each individual group or genre by structuring the combined data in some reasonable way. Here, a sensible approach should return a pooled solution if in fact the underlying relations are identical across units of analysis, and return strictly unit-specific results if the units share little in common. Most importantly, a sensible approach would be capable of operating in the gray area where some relations are common across units and others are unit-specific.

To the best of our knowledge, the multi-VAR approach presented here is the first work that combines regularized estimation of time series models (Basu & Michailidis, 2015a, 2015b) with the problem of supervised learning of multiple-group data (Gross & Tibshirani, 2016; Ollier & Viallon, 2017). We believe this combination is exceptionally well suited to many problems in the social sciences. In addition, we make a number of unique contributions to the existing literature. First, we prove a consistency result for our estimator in the proposed multiple-subject estimation problem. Second, we propose a proximal gradient descent approach for solving the multiple-unit LASSO (standard and adaptive) problem based on the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009). Third, we evaluate the performance of our proposed method in a simulation study and a real data example from Fredrickson et al. (2017) involving day-to-day emotional experiences. Finally, we provide a convenient R package for applied researchers looking to use the proposed methods (Fisher, 2021).

1. Estimating Vector Autoregressive Models

We focus our attention on the multivariate time series, $\{\mathbf{X}_t\}_{t\in\mathbb{Z}} = \{(X_{j,t})_{j=1,\ldots,d}\}_{t\in\mathbb{Z}}$. $\mathbf{X}_t$ is considered to follow a vector autoregressive model of order p, VAR(p), if

$$\mathbf{X}_t = \boldsymbol{\Phi}_1\mathbf{X}_{t-1} + \cdots + \boldsymbol{\Phi}_p\mathbf{X}_{t-p} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{1}$$

for some $d \times d$ matrices $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ and a white noise series $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}} \sim \text{WN}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ characterized by $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$ and $\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s^{'}) = 0$ for $s \neq t$. For simplicity, we assume $\mathbf{X}_t$ is of zero mean. Generally, a unique causal stationary solution to (1) can be ensured by satisfying the stability condition given by $\det(\boldsymbol{\Phi}(z)) \neq 0$, for $|z| \leq 1$, $z \in \mathbb{C}$, where $\boldsymbol{\Phi}(z) = \mathbf{I}_d - \boldsymbol{\Phi}_1 z - \cdots - \boldsymbol{\Phi}_p z^p$.

1.1. Estimation of Unrestricted VAR Models

It is common to estimate (1) using ordinary least squares (OLS) regression,

$$(\widehat{\boldsymbol{\Phi}}_1, \ldots, \widehat{\boldsymbol{\Phi}}_p) = \underset{\boldsymbol{\Phi}_1,\ldots,\boldsymbol{\Phi}_p}{\text{argmin}} \sum_{t=p+1}^{T} \|\mathbf{X}_t - \boldsymbol{\Phi}_1\mathbf{X}_{t-1} - \cdots - \boldsymbol{\Phi}_p\mathbf{X}_{t-p}\|_2^2, \tag{2}$$

where T is the sample size and $\|\cdot\|_2$ denotes the Frobenius (Euclidean) norm, which is equivalent to running component-wise regression on each of the d VAR equations. In this case, the estimate $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ is defined as the sample variance–covariance matrix of the residuals. When there are no restrictions on $\boldsymbol{\Phi}$, the OLS estimates are asymptotically equivalent to those produced by generalized least squares (GLS) (Zellner, 1962). Under the assumption that $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ are independent across t, the OLS estimates obtained by component-wise regression are also the maximum likelihood (ML) estimates (Lütkepohl, 2007). These estimators are asymptotically normal under mild assumptions with explicit variance–covariance matrices.

A drawback of the unrestricted VAR model is the large number of parameters that must be estimated. In fact, the number of parameters scales quadratically as the number of component series increases. Assuming no mean structure, $pd^2 + d(d-1)/2$ model parameters need to be estimated for the unrestricted VAR(p) model. This means that, for example, a VAR(1) model with 10 component series requires estimating 145 parameters. With such a large parameter space, it is likely that many of the estimated linear relationships in an unrestricted VAR model will be spurious and the regression matrix $\mathbf{X}^{'}\mathbf{X}$ ill-conditioned. Furthermore, when $(T-p)d < pd^2 + d(d-1)/2$, estimation via OLS is not possible.

1.2. Estimation of Sparse VAR Models

As a consequence of the dimensionality issues surrounding unrestricted VAR estimation, much attention has been paid to methods for reducing the VAR parameter space. Ideally, the selection of relevant series would be guided by theory. Unfortunately, the ease associated with many types of electronic data collection in the behavioral and social sciences has allowed for the collection of many repeated measures, all of which are hypothesized as relevant to the phenomena under study. For this reason, it is often difficult to prune variables *a priori* when theory points to their inclusion. Several data-driven approaches have been presented in the literature to overcome this issue of high-dimensionality (Basu & Michailidis, 2015a; Han & Liu, 2013). In our current approach, we assume sparsity of $\boldsymbol{\Phi}$ and use penalized estimation to recover the model parameters.

1.2.1. LASSO Estimation To set up the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), the VAR model and associated data are expressed in a regression form

$$\underbrace{\begin{bmatrix} \mathbf{X}_{p+1}' \\ \mathbf{X}_{p+2}' \\ \vdots \\ \mathbf{X}_{T}' \end{bmatrix}}_{\mathcal{Y}} = \underbrace{\begin{bmatrix} \mathbf{X}_{p}' & \cdots & \mathbf{X}_{1}' \\ \mathbf{X}_{p+1}' & \cdots & \mathbf{X}_{2}' \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{T-1}' & \cdots & \mathbf{X}_{T-p}' \end{bmatrix}}_{\mathcal{X}} \underbrace{\begin{bmatrix} \mathbf{\Phi}_{1}' \\ \mathbf{\Phi}_{2}' \\ \vdots \\ \mathbf{\Phi}_{p}' \end{bmatrix}}_{\mathcal{B}^*} + \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_{p+1}' \\ \boldsymbol{\varepsilon}_{p+2}' \\ \vdots \\ \boldsymbol{\varepsilon}_{T}' \end{bmatrix}}_{\mathcal{E}} \tag{3}$$

or, equivalently,

$$\operatorname{vec}(\mathcal{Y}) = (\mathrm{I}_d \otimes \mathcal{X})\operatorname{vec}(\mathcal{B}^*) + \operatorname{vec}(\mathcal{E}), \tag{4}$$

$$\underbrace{\mathbf{Y}}_{Nd\times 1} = \underbrace{\mathbf{Z}}_{Nd\times q}\ \underbrace{\mathbf{B}^*}_{q\times 1} + \underbrace{\mathbf{E}}_{Nd\times 1}, \tag{5}$$

where the star $*$ indicates the true parameter, $N = T - p$ and $q = pd^2$. Here, we assume that $\mathbf{B}^*$ is s-sparse (i.e., $\Sigma_{i=1}^{p}\|\operatorname{vec}(\mathbf{\Phi}_i)\|_0 = \|\mathbf{B}\|_0 = \sum_{i=1}^{q} 1_{\{B_i\neq 0\}} = s$ where $\|\cdot\|_0$ is the ℓ_0-norm). With this structure in place, we can write the LASSO estimator as

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N}\|\mathbf{Y} - \mathbf{Z}\mathbf{B}\|_2^2 + \lambda\|\mathbf{B}\|_1, \tag{6}$$

where $\|\mathbf{B}\|_1 = \Sigma_{i=1}^{q}|B_i|$ for $\mathbf{B} = (B_1, \ldots, B_q)'$ and $\lambda > 0$ is the regularization penalty parameter. In (6), the scaling constant $\frac{1}{N}$ (corresponding to λ) sometimes takes the values $\frac{1}{2N}$, 1 and 2 depending on the convention. Here, $N = T - p$ refers to the time series length of a given individual in the sample. Changing the scaling context corresponds to a reparameterization of λ and does not impact the estimation of (6). Large values of λ typically correspond to sparser solutions.

1.3. Multiple-Subject Penalized VAR

Up to this point, we have presented the VAR model and optimization problem in terms of a single multivariate time series. This was useful for describing the estimators; however, the majority of ILD and many psychophysiological applications involve observing the same variables across multiple subjects. With multivariate repeated measurements collected from multiple subjects, we are now interested in estimating the sparse parameter vectors $\hat{\mathbf{B}}_1, \ldots, \hat{\mathbf{B}}_K$ for K individuals. Rarely, if ever are the relationships among items strictly the same across any two individuals in the sample. However, it is certainly possible and maybe even expected that certain qualitative aspects of a dynamic process are similar across individuals. For this reason, strategies involving the estimation of K separate LASSO problems are generally suboptimal. To overcome this limitation, we propose the multi-VAR modeling framework for multivariate time series data collected from multiple subjects.

1.4. The multi-VAR Approach

The approach described herein relies on the following decomposition of $\mathbf{B}_k^*$:

$$\mathbf{B}_k^* = \boldsymbol{\mu}^* + \boldsymbol{\Delta}_k^*, \quad k = 1, \ldots, K, \tag{7}$$

where $\boldsymbol{\mu}^* \in \mathbb{R}^q$ corresponds to the common effects across K individuals and $\boldsymbol{\Delta}_k^* \in \mathbb{R}^q$ corresponds to the effects unique to individual k. Now, considering the regularization parameters λ_1 and $\lambda_{2,k}$, $k = 1, \ldots, K$, which govern the cross-sectional heterogeneity in our solution, we can write the revised optimization problem as

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Delta}}_1, \ldots, \hat{\boldsymbol{\Delta}}_K) = \underset{\boldsymbol{\mu}, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_K}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{K} \|\mathbf{Y}^{(k)} - \mathbf{Z}^{(k)}(\boldsymbol{\mu} + \boldsymbol{\Delta}_k)\|_2^2 + \lambda_1 \|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K} \lambda_{2,k} \|\boldsymbol{\Delta}_k\|_1. \tag{8}$$

As mentioned previously, we prefer a sensible approach to handling multivariate time series data arising from multiple individuals, specifically in the case where the cross-sectional heterogeneity of the individual processes is unknown. If the individuals share very little in common, in terms of their time series, this approach should return essentially independent solutions. That is the results should be similar to what would be obtained from estimating K separate VAR models. In (6), larger values of the penalty parameter λ will increasingly drive the corresponding coefficient matrix $\mathbf{B}$ to zero. Similarly, in (8), large enough values of λ_1 will shrink the common effect matrix, $\hat{\boldsymbol{\mu}}$ toward zero, and we will be left with the individual-specific effects $\hat{\mathbf{B}}_k = \hat{\boldsymbol{\Delta}}_k$. This would essentially produce results similar to those obtained from estimating K independent VAR models.

Likewise, if the individual-level processes are essentially homogenous, a sensible approach would return results similar to estimating a single pooled model for all individuals in the sample. In this case, we would expect that large enough values of $\lambda_{2,k}$ would drive the individual-specific transition matrices, $\hat{\boldsymbol{\Delta}}_k$, toward zero, leaving only the common effect transition matrix to explain the individual-level results, $\hat{\mathbf{B}}_k = \hat{\boldsymbol{\mu}}$. Finally, if an individual's process has both common and unique components, a sensible approach would attempt to balance these contributions. In this case, the penalty parameters, λ_1 and $\lambda_{2,k}$, are selected to optimally govern the contribution of both the common ($\hat{\boldsymbol{\mu}}$) and unique ($\hat{\boldsymbol{\Delta}}_k$) effects to each individual's dynamics ($\hat{\mathbf{B}}_k$).

Using the decomposition presented in (7), it is also possible to rewrite the right-hand side (RHS) of (8) such that $\boldsymbol{\mu}$ only appears in the penalty term as

$$\begin{aligned} &\underset{\boldsymbol{\mu}, \mathbf{B}_1, \ldots, \mathbf{B}_K}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{K} \|\mathbf{Y}^{(k)} - \mathbf{Z}^{(k)}\mathbf{B}_k\|_2^2 + \lambda_1 \|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K} \lambda_{2,k} \|\mathbf{B}_k - \boldsymbol{\mu}\|_1 \\ &= \underset{\boldsymbol{\mu}, \mathbf{B}_1, \ldots, \mathbf{B}_K}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{K} \|\mathbf{Y}^{(k)} - \mathbf{Z}^{(k)}\mathbf{B}_k\|_2^2 + \lambda_1 \left(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K} \frac{\lambda_{2,k}}{\lambda_1} \|\mathbf{B}_k - \boldsymbol{\mu}\|_1 \right). \end{aligned} \tag{9}$$

To simplify the following discussion, let $r_k = \lambda_{2,k}/\lambda_1$ for $k = 1, \ldots, K$. Now, it is important to note, as in Gross and Tibshirani (2016), that any choice of the regularization parameters, $\frac{\lambda_{2,1}}{\lambda_1}, \ldots, \frac{\lambda_{2,K}}{\lambda_1}$, and coefficients $B_{1,j}, \ldots, B_{K,j}$ identifies a specific solution for the common effects in $\boldsymbol{\mu}$ where μ_j is the weighted and shrunken median of $B_{1,j}, \ldots, B_{K,j}$ as in Ollier and Viallon (2017) .

Indeed, the penalty term in (9) is separable in its q parameters such that we can consider a single explanatory coefficient $B_{k,j}$ and associated weight r_k for $k = 0, \ldots, K$. Using this

specification, we can rewrite the penalty term in (9) as the sum of the generic one-dimensional unconstrained optimization problem

$$\underset{\mu_j}{\operatorname{argmin}} \sum_{k=0}^{K} r_k |B_{k,j} - \mu_j|. \tag{10}$$

Implicitly, we set $r_0 = 1$ and $B_{0,j} = 0$ to match the penalty construction in (9). Expressed as in (10), the solution $\hat{\mu}_j$ becomes a properly weighted median of $(B_{0,j} = 0, B_{1,j}, \ldots, B_{K,j})$. In this setting, a number of scenarios relevant to applied researchers are worth considering, as discussed in Gross and Tibshirani (2016). First, if $r_k = r$, $k = 1, \ldots, K$ and $r \in (\frac{1}{K}, \frac{1}{K-2})$, the group effect $\hat{\mu}_j$ will be nonzero if and only if all $B_{k,j}$ are of the same sign, in which case it will be equal to the minimum value of $(B_{1,j}, \ldots, B_{K,j})$. This means for the group effect to exist, it must be present for all individuals in the sample and only then will deviations from the group be captured in the individual $(B_{1,j}, \ldots, B_{K,j})$. Second, if $\Sigma_{k=1}^{K} r_k < 1$, we are guaranteed the group effect $\hat{\mu}_j$ will equal zero, and the problem will resemble fitting K individual penalized VAR models. Third, if $r_k = r, k = 1, \ldots, K$ and $r > 1$, the group effect $\hat{\mu}_j$ will effectively be the median of $(B_{1,j}, \ldots, B_{K,j})$. In general, the weights r_k in the above minimization problem can be understood as a penalty applied to idiosyncratic dynamics (coefficients) not shared by the entire sample.

1.5. The Adaptive multi-VAR Approach

It is also possible to develop an adaptive-LASSO (Zou, 2006) version of the multi-VAR approach for the VAR model by minimizing the objective function

$$\frac{1}{N} \sum_{k=1}^{K} \|\mathbf{Y}^{(k)} - \mathbf{Z}^{(k)}(\boldsymbol{\mu} + \boldsymbol{\Delta}^{(k)})\|_2^2 + \lambda_1 \left(\boldsymbol{\omega} \|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K} \frac{\lambda_{2,k}}{\lambda_1} \boldsymbol{v}_k \|\boldsymbol{\Delta}^{(k)}\|_1 \right), \tag{11}$$

where $\omega_j = 1/|\tilde{B}_{\ell_j,j}|$ and $v_{k,j} = 1/|\tilde{B}_{k,j} - \tilde{B}_{\ell_j,j}|$ with $\tilde{B}_{k,j}$ and $\tilde{B}_{\ell_j,j}$ defined next. For each of the k individuals in the sample, the estimate $\tilde{\mathbf{B}}_k = (\tilde{B}_{k,j})$ of $\mathbf{B}_k$ can be obtained using maximum likelihood or OLS when the number of time points for each individual (Nd) exceeds the number of variables $(pd^2 + d(d-1)/2)$, or from (9) when this condition is not met. In addition, $\tilde{B}_{\ell_j,j}$ can be taken as the median coefficient estimate for variable j across all K individuals such that $\tilde{B}_{\ell_j,j} = \text{median}(\tilde{B}_{1,j}, \ldots, \tilde{B}_{K,j})$. A benefit of the adaptive LASSO approach in comparison with (8) is that we are able to weight the ℓ_1 penalty. By weighting the ℓ_1 penalty we are able to help ensure coefficients we might expect to be prominent in the model (based on a consistent first stage estimator, such as OLS) receive smaller penalties. In certain contexts, this helps to reduce the bias of the LASSO estimator and can provide a number of benefits, including consistency in both variable selection and parameter estimation (Zou, 2006). A nice property of this approach is that we can reexpress (9) and (11) as a weighted LASSO problem, namely

$$\underbrace{\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \\ \vdots \\ \mathbf{Y}^{(K)} \end{bmatrix}}_{\mathscr{Y}} = \underbrace{\begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{Z}^{(1)} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{Z}^{(2)} & \mathbf{0} & \mathbf{Z}^{(2)} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}^{(K)} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{Z}^{(K)} \end{bmatrix}}_{\mathscr{Z}} \underbrace{\begin{bmatrix} \boldsymbol{\mu}^* \\ \boldsymbol{\Delta}^{(1)*} \\ \vdots \\ \boldsymbol{\Delta}^{(K)*} \end{bmatrix}}_{\boldsymbol{\theta}^*} + \underbrace{\begin{bmatrix} \mathbf{E}^{(1)} \\ \mathbf{E}^{(2)} \\ \vdots \\ \mathbf{E}^{(K)} \end{bmatrix}}_{\mathscr{E}}, \tag{12}$$

where the criterion we are now concerned with minimizing is given by

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \|\mathscr{Y} - \mathscr{Z}\boldsymbol{\theta}\|_2^2 + \lambda_1 \|\boldsymbol{\theta}\|_{1,\mathbf{w}} \tag{13}$$

and $\|\boldsymbol{\theta}\|_{1,\mathbf{w}} = \sum_i w_i|\theta_i|$ with $\mathbf{w}' = (\mathbf{1}_d', (\lambda_{2,1}/\lambda_1)\mathbf{1}_d', \ldots, (\lambda_{2,K}/\lambda_1)\mathbf{1}_d')$ for (9) and $\mathbf{w}' = (\boldsymbol{\omega}', (\lambda_{2,1}/\lambda_1)\boldsymbol{\nu}_1', \ldots, (\lambda_{2,K}/\lambda_1)\boldsymbol{\nu}_K')$ for (11).

It is worth nothing that the design matrix $\mathscr{Z}$ in (12) is not of full column rank even if the number of observations per subject will exceed the number of parameters. In particular, OLS for (12) is not feasible. Yet, under sparsistency, results on consistency and sparsistency for LASSO estimation are available as discussed in the appendix below.

2. Computational Algorithm and Estimation

Solving (13) requires iterative methods as the ℓ_1 penalty is not differentiable, and no analytic solutions exist. A popular schema for solving penalized regression problems is coordinate descent as popularized by Friedman et al. (2010). Coordinate descent has proved to be an exceedingly effective algorithm for exploiting the sparsity of the coefficient vector structure, partly because it moves parameters one at a time. Coordinate descent is easier to implement than many competing approaches, and this has likely also contributed to its popularity. Another class of methods for solving (13) falls under the umbrella of proximal gradient descent. Unlike coordinate descent, proximal gradient descent moves all the parameters of a model at once, and may provide efficiency gains for certain types of problems, such as the estimation of high-dimensional VAR models (Nicholson et al., 2017). We have chosen to implement our approach in the proximal gradient framework due to these desirable qualities, as well as the generality of the proximal framework to a wide-range of time series optimization problems applicable to the multi-VAR framework. In the remainder of this section, we introduce the proximal gradient descent algorithm we have implemented for solving (13) and describe a number of useful modifications for enhancing computational efficiency.

Proximal algorithms have proved incredibly useful in the fields of statistics, machine learning and image processing for solving complex optimization problems involving composite objective functions, such as the one presented in (9). In fact, many methods commonly used in psychometric research, such as expectation maximization (EM), majorization–minimization (MM) and iteratively reweighted least squares (IRLS), can be shown to be proximal algorithms (Polson et al., 2015). Broadly, a proximal algorithm refers to any algorithm where a proximal operator is applied to a subproblem of a larger optimization routine, often in a nonsmooth setting where the aim is simplifying the problem of interest. It is beyond the scope of the current work to describe the proximal operator itself in any generality; however, a detailed treatment of proximal operators and algorithms is given by Parikh and Boyd (2014). In the following section, we will present a proximal gradient descent algorithm for solving (9) and (11) in the form of (13).

To develop some intuition for the proximal gradient algorithm, let us first consider the unconstrained minimization of the convex differentiable function $f(\boldsymbol{\theta})$. At the global minimum of $f(\boldsymbol{\theta})$, a necessary and sufficient condition for the optimality of parameters $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is given by the zero-gradient condition $\nabla f(\boldsymbol{\theta}^*) = 0$. Typically, gradient descent methods require two primary decisions be made at each successive iteration. First, a direction of descent must be determined. This direction will be the direction of steepest descent $-\nabla f(\boldsymbol{\theta}^s)$ for $s = 0, 1, 2, \ldots$. Second, a step size (or scale factor) must be chosen to govern the size of the step taken. This step size is governed by a step size parameter γ^s, such that $\boldsymbol{\theta}^{s+1} = \boldsymbol{\theta}^s - \gamma^s \nabla f(\boldsymbol{\theta}^s)$ or equivalently

$$\boldsymbol{\theta}^{s+1} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ f(\boldsymbol{\theta}^s) + \langle \nabla f(\boldsymbol{\theta}^s), \boldsymbol{\theta} - \boldsymbol{\theta}^s \rangle + \frac{1}{2\gamma^s} \|\boldsymbol{\theta} - \boldsymbol{\theta}^s\|_2^2 \right\}. \tag{14}$$

Here, we can see the unconstrained minimization problem in (14) is simply the local linear approximation to $f(\boldsymbol{\theta})$ supplemented with a quadratic smoothness term.

Unlike the problem in (14), the optimization problems described in (9) and (11) are both nondifferentiable due to the presence of the weighted ℓ_1 penalty. At this point, it is helpful to consider the decomposition of $f(\boldsymbol{\theta})$ into separable components $g(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$ such that $f(\boldsymbol{\theta}) := g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ where $g(\boldsymbol{\theta})$ is convex and differentiable and $h(\boldsymbol{\theta})$ is convex but nondifferentiable. In doing so, we can define a gradient update where $g(\boldsymbol{\theta})$ is approximated as in (14), and we leave the nonsmooth $h(\boldsymbol{\theta})$ in its original form

$$\boldsymbol{\theta}^{s+1} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ g(\boldsymbol{\theta}^s) + \langle \nabla g(\boldsymbol{\theta}^s), \boldsymbol{\theta} - \boldsymbol{\theta}^s \rangle + \frac{1}{2\gamma^s} \|\boldsymbol{\theta} - \boldsymbol{\theta}^s\|_2^2 + h(\boldsymbol{\theta}) \right\}. \tag{15}$$

Now, for the weighted LASSO problem in (13), this decompositions takes the form

$$g(\boldsymbol{\theta}) = \frac{1}{N} \|\mathscr{Y} - \mathscr{Z}\boldsymbol{\theta}\|_2^2 \tag{16}$$

$$h(\boldsymbol{\theta}) = \lambda_1 \|\boldsymbol{\theta}\|_{1,\mathbf{w}} = \lambda_1 \sum_i w_i |\theta_i|, \tag{17}$$

where $\nabla g(\boldsymbol{\theta}) = \mathscr{Z}'(\mathscr{Y} - \mathscr{Z}\boldsymbol{\theta})$ and the composition of $\mathbf{w}$ is determined both by the similarity of individuals in the sample and the nature of the penalization scheme.

Fortunately, in the case of (13), the proximal operator for $g(\boldsymbol{\theta})$ has a closed-form solution whose evaluation is negligible in terms of computational costs. We can write the ith component of the proximal operator $\mathbf{prox}_{h,\lambda_1}$ as

$$(\mathbf{prox}_{h,\lambda_1}(\boldsymbol{\theta}))_i = \mathbf{prox}_{\lambda,w_i}(\theta_i) = \begin{cases} \theta_i + \lambda_1 w_i, & \text{if } \theta_i < -\lambda_1 w_i \\ 0, & \text{if } |\theta_i| \leq \lambda_1 w_i \\ \theta_i - \lambda_1 w_i, & \text{if } \theta_i > \lambda_1 w_i \end{cases} \tag{18}$$

due to the separable sum property and the definition of the weighted ℓ_1 norm. Using (18), we can now write gradient update given (15) as a proximal gradient update

$$\boldsymbol{\theta}^{s+1} = \mathbf{prox}_{h,\gamma^s} \left\{ \boldsymbol{\theta}^s - \gamma^s \nabla g(\boldsymbol{\theta}^s) \right\} \tag{19}$$

$$= \mathbf{prox}_{h,\gamma^s} \left\{ \boldsymbol{\theta}^s - \gamma^s \left(\mathscr{Z}'(\mathscr{Y} - \mathscr{Z}\boldsymbol{\theta}^s) \right) \right\}, \tag{20}$$

where the precomputation of $\mathscr{Z}'\mathscr{Z}$ and $\mathscr{Z}'\mathscr{Y}$ can further reduce the computational cost of each update as the objective functional value will only differ by a constant. A classic proximal gradient schemes for solving (13) is the iterative shrinkage-thresholding algorithm (ISTA). In the standard ISTA, formulation step size is treated as a constant across descent iterations, and no smoothing techniques are used to accelerate the descent. To overcome these limitations, Beck and Teboulle (2009) proposed a general fast iterative shrinkage-thresholding algorithm (FISTA) for solving gradient descent problems. In the remainder of this section, we describe the version of FISTA we have implemented for the multi-VAR problems described above.

Algorithm 1: Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) for Solving (13)

Input: Set $\boldsymbol{\theta}^0 = \boldsymbol{\psi}^0 = \mathbf{0}$, $c^0 = 1$, $\alpha = 0.5$, $\eta = 0.5$, $\gamma = 0.5$, choose an $\epsilon > 0$.
Output: A solution $\boldsymbol{\theta}^s$.
for $s = 0, \ldots, s_{\max}$ **do**
  Update $c^{s+1} := 0.5\left(1 + \sqrt{1 + 4(c^s)^2}\right)$ and terminate if
  $\|\boldsymbol{\theta}^{s+1} - \boldsymbol{\theta}^s\|_2 \leq \epsilon \max\{1, \|\boldsymbol{\theta}^s\|_2\}$
  while $f(\boldsymbol{\theta} - \nabla f(\boldsymbol{\theta})) > f(\boldsymbol{\theta}) - \gamma\alpha\|\nabla f(\boldsymbol{\theta})\|^2$ **do**

$$\begin{cases} \boldsymbol{\theta}^{s+1} = \mathbf{prox}_{\gamma^s, h}\left\{\boldsymbol{\psi}^s - \gamma^s \nabla g(\boldsymbol{\psi}^s)\right\} \\ \boldsymbol{\psi}^{s+1} = \boldsymbol{\theta}^{s+1} + \frac{c^{s-1}}{c^{s+1}}(\boldsymbol{\theta}^{s+1} - \boldsymbol{\theta}^s) \\ \gamma^s = \eta\gamma^s \end{cases} \tag{25}$$

  end
end

As mentioned previously, the choice of the step size parameter γ^s in gradient descent can have a large impact on the convergence rate of the estimator, and also whether a global minimum is reached. One convenient method for determining an approximately optimal step size is to perform a backtracking line search (Boyd & Vandenberghe, 2004) within each iteration. In this scheme, the step size is determined by iteratively rescaling γ by η where $\eta \in (0, 1)$ until

$$f(\boldsymbol{\theta} - \nabla f(\boldsymbol{\theta})) \leq f(\boldsymbol{\theta}) - \gamma\alpha\|\nabla f(\boldsymbol{\theta})\|^2, \tag{21}$$

where $\alpha \in (0, 0.5)$ is the second constant, in addition to η, used to govern the backtracking procedure. Based on previous experience, we have chosen a value of $\alpha = 0.5$, which corresponds to a maximum decrease in f between 1 and 50% and $\eta = 0.5$ which corresponds to a moderate value of granularity. As Boyd and Vandenberghe (2004, p. 466) suggest, η should be chosen within the range of 0.1 (more crude search) and 0.8 (less crude search).

A final improvement to the typical gradient descent procedures corrects the "zig-zagging" descent often observed during iterative computation of (19), which may slow convergence (Hastie et al., 2015). A solution initially proposed by Nesterov (2007) and incorporated into FISTA by Beck and Teboulle (2009) uses weighted combinations of the past gradient descent directions to smooth the global descent path. Another nice feature of proximal gradient descent is that the acceleration approach suggested by Nesterov (2007) can be integrated into the proximal operator such that the gradient step now involves

$$c^{s+1} := 0.5\left(1 + \sqrt{1 + 4(c^s)^2}\right) \tag{22}$$

$$\boldsymbol{\theta}^{s+1} = \mathbf{prox}_{\gamma^s, h}\left\{\boldsymbol{\psi}^s - \gamma^s \nabla g(\boldsymbol{\psi}^s)\right\} \tag{23}$$

$$\boldsymbol{\psi}^{s+1} = \boldsymbol{\theta}^{s+1} + \frac{c^{s-1}}{c^{s+1}}(\boldsymbol{\theta}^{s+1} - \boldsymbol{\theta}^s) \tag{24}$$

where the step size γ^s is chosen by the procedure while iterating until (21) is met. The constant c^s is updated at each iteration. Finally, we provide pseudocode describing our algorithm in full.

2.1. Forecasting from the Estimated VAR Process

Here, we provide a brief description of how forecasts are obtained from the estimated VAR(1) transition matrices. For the multi-VAR approach, *one*-step ahead linear prediction of $\mathbf{Y}_{T+1}^{(k)}$ for individual k is given by

$$\mathbf{Y}_{T+1}^{(k)} = \mathbf{Z}_{T}^{(k)}(\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\Delta}}_k). \tag{26}$$

From (26), the h-step ahead forecast can be computed recursively for any horizon h.

2.2. Selection of the Penalty Parameters

Performance of the proposed multi-VAR approach is dependent on the selection of the unknown penalty parameters λ_1 and $\lambda_{2,k}$, $k = 1, \ldots, K$. Here, we provide additional details on how the penalty parameters are chosen in the simulation studies and empirical example. The first step of our proposed procedure involves constructing a grid of plausible penalty values. Following Friedman et al. (2010), we first identify $\lambda_{1,\max}$, or the smallest value of λ_1 for which all the coefficients in the model will be zero. In the multi-VAR setting, $\lambda_{1,\max}$ is equal to $\max|\mathscr{Z}'\mathscr{Y}|$ where $\mathscr{Z}$ and $\mathscr{Y}$ are given in (13). From $\lambda_{1,\max}$, we construct a grid from $\lambda_{1,\max}$ to $\lambda_{1,\max}/1000$ using equally spaced values on a log-scale. Across a number of data contexts, 20 values of λ_1 and $\lambda_{2,k}$ were found to be sufficient. Following Ollier and Viallon (2014, p. 32), the ratio $\lambda_2/\lambda_{1,k}$ is chosen to vary on the interval $(0, \ldots, K)$, and this ratio is used to solve for $\lambda_{2,k,\max}$, from which another grid is constructed, $\lambda_{2,k,\max}$ to $\lambda_{2,k,\max}/1000$, also on a log-linear scale.

To identify the optimal penalty parameters from our grid of candidate values, we adapt the rolling-window cross-validation (RWCV) procedure for high-dimensional VAR models described by Banbura et al. (2010), Song and Bickel (2011), and Nicholson et al. (2017) to the multi-VAR problem. This procedure involves searching across the grid of predetermined values described above and choosing the combination of penalty parameters that minimize the h-step ahead mean-square forecast error (MSFE). Here, $h = 1, 2, 3, \ldots$ is the desired forecast horizon, and for all analyses in this paper, $h = 1$ is used to select the penalty parameters. We chose to use this forecast horizon as ILD is often collected daily, and we hypothesized that one-day-ahead predictions have utility in behavioral, health and social science applications. To implement the rolling-window cross-validation procedure, we divide each individual dataset into three periods. The first period is the initialization period beginning at the first time point and ending at T_1. Based on the literature above, we set $T_1 = T/3$, or approximately $1/3$ of the time series length. The second period is the training period, starting at T_1+1 and ending at T_2. We chose T_2 to be equal to $T-3$, leaving a hold-out-sample of three observations in the final period ($T_2 + 1$ to T) for our pseudo-out-of-sample forecast evaluation.

For each value of the λ_1 and $\lambda_{2,k}$ grid, we perform the following sequence. First, we solve the problem in (13) using timepoints $1, \ldots, T_1$ from each individual in the sample to obtain $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Delta}}_k$. Separately for each individual, these estimates are then used to forecast $\hat{\mathbf{Y}}_{T_1+1}^{(k)}$ and obtain the MSFE. We continue in this fashion, adding one observation at a time to the initialization period and forecasting ahead h-units until we reach $T_2 - h$, at which time the forecast performance is aggregated across the $(T_2 - T_1 - h + 1)$ forecasts for each combination of λ_1 and $\lambda_{2,k}$ as in

$$\text{MSFE}_{\lambda_1,\lambda_{2,k}}^{(k)} = \frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2-1} \|\hat{\mathbf{Y}}_{t+1}^{(k)} - \mathbf{Y}_{t+1}^{(k)}\|_2^2, \tag{27}$$

and the values of λ_1 and $\lambda_{2,k}$ which correspond to the smallest MSFE are chosen for evaluating the forecast performance in the hold-out sample.

3. Performance Evaluation

To better understand the finite sample properties of the proposed models and algorithm, we conduct a Monte Carlo simulation designed to replicate some of the basic features of ILD collected from multiple subjects.

3.1. Simulation Design

To evaluate the performance of the proposed approach for forecasting multiple subject ILD, we generated data according to a number of commonly encountered features: (1) individual time series lengths of $T = (30, 50, 100)$, (2) number of ILD variables collected per individual $d = (10, 20, 30)$, (3) total number of individuals in the sample $K = (10, 20)$, (4) the level of cross-sectional heterogeneity (low, medium and high) and (5) the type of penalized VAR model employed; (a) VAR fit by LASSO for each individual in the sample separately as in (6), (b) the multi-VAR as in (9), and (c) the adaptive multi-VAR(1) as in (11).

Across all design factors, the $d \times d$ sparse transition matrices for each individual were generated to have 5% nonzero entries. This means, for example, a multivariate time series with $d = 30$ would have 45 nonzero coefficients in the data generating model. The position of nonzero elements in each individual's transition matrix were selected randomly given the following constraints. In the low-heterogeneity condition, $2/3$ of paths were common to all individuals, and $1/3$ of paths were completely unique to each individual. In the medium heterogeneity condition, $1/2$ of each individual's paths were common and $1/2$ were unique. In the high-heterogeneity condition $1/3$ were common and $2/3$ were unique. Across all conditions the individual transition matrix elements were drawn from $\mathcal{U}(0.1, 0.9)$ until the stability condition from (1) was satisfied. For each of the $2 \times 3 \times 3 \times 3$ data generating conditions, we conducted 20 replications.

Across all conditions, the 3 final time points of each component series were withheld to evaluate forecast accuracy. For the synthetic data examples, cross-validation was performed in two different ways. First, we assumed the nonzero transition matrices were known and chose the penalty parameters that resulted in the smallest estimation error (e.g., $\|\hat{\mathbf{B}} - \mathbf{B}\|_F / \|\mathbf{B}\|_F$). This allowed us to compare the different approaches independent of the cross-validation method. Of course, in real data scenarios, $\mathbf{B}$ is unknown, and it is important to examine our proposed framework under realistic conditions. To this end, our second approach used the RWCV procedure described earlier to select optimal values of λ_1 and λ_2 in our simulation study.

3.2. Outcome Measures

To evaluate the performance of our approach, we looked at a number of measures relevant to forecast performance. These measures include (a) sensitivity, (b) specificity, and (c) root-mean-square forecast error (RMSFE). The mean sensitivity and the mean specificity were calculated as

$$\text{Mean sensitivity} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\Sigma_j(\hat{B}_{k,j} \neq 0 \text{ and } B_{k,j} \neq 0)}{\Sigma_j(B_{k,j} \neq 0)}\right), \tag{28}$$

$$\text{Mean specificity} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\Sigma_j(\hat{B}_{k,j} = 0 \text{ and } B_{k,j} = 0)}{\Sigma_j(B_{k,j} = 0)}\right) \tag{29}$$

where $B_{k,j}$ and $\hat{B}_{k,j}$ are the true and the estimated transition matrix elements, respectively, for individual k in a given design condition. Finally, the mean RMSFE is

$$\text{Mean RMSFE} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{d}\sum_{j=1}^{d}(\hat{\mathbf{Y}}^{(k)}_{j,t-3+h} - \mathbf{Y}^{(k)}_{j,t-3+h})^2} \tag{30}$$

where $(\hat{\mathbf{Y}}^{(k)}_{j,t-3+h} - \mathbf{Y}^{(k)}_{j,t-3+h})$ is the h-step ahead forecast error for individual k on variable j and $h \in \{1, 2, 3\}$.

3.3. Simulation Results

As expected, performance differences emerged when the penalty parameters were selected using the estimation error metric compared to the rolling-window cross-validation. This is clear when one compares Figures 1 and 2. An interesting pattern also emerges when comparing the performance of each method in aggregate. For example, the individual LASSO achieved a mean sensitivity of 0.88 and a mean specificity of 0.81 across all conditions when the true data generating matrices were used to select λ_1 and λ_2, using the estimation error metric. When the RWCV procedure was used mean sensitivity was 0.87 and a mean specificity 0.80. The standard multi-VAR achieved a mean sensitivity of 0.93 and a mean specificity of 0.78 for the estimation error condition and a sensitivity of 0.94 and specificity of 0.75 using RWCV. In aggregate, for the individual-level LASSO and standard multi-VAR approaches, sensitivity and specificity were similar across the two penalty selection procedures. On the other hand, performance of the adaptive multi-VAR approach differed considerably across the two approaches, with sensitivity increasing from 0.88 to 0.94 and specificity decreasing from 0.93 to 0.73. This suggests the adaptive multi-VAR approach suffered the most from the RWCV approach to selecting the penalty parameters. Regardless, the RWCV approach is what will be used in practice we will focus our discussion on this set of results.

3.3.1. Sensitivity and Specificity for $\mathbf{B}_k$ We first consider recovery of the total effect matrix for each individual. Although we would not expect the heterogeneity levels or number of individuals in the sample to impact the individual LASSO performance, we are certainly interested in their impact on the multi-VAR methods. Overall, the differing levels of heterogeneity had little impact on parameter recovery outside of small decrement in sensitivity for the multi-VAR approaches at the smallest time series length of $T = 30$. This is also consistent with the aggregate findings as the multi-VAR approaches both showed a decrement (0.03) in sensitivity and a slight increase (0.01–0.02) in specificity, as heterogeneity increased from low to high. In aggregate, both multi-VAR approaches showed no change in sensitivity or specificity when the number of individuals included in the sample was 10 or 20.

In contrast to the cross-sectional heterogeneity and the number of individuals conditions, we would expect the number of timepoints and the number of variables to impact the performance of all three approaches. For the standard multi-VAR, sensitivity increased as the time series length increased, from (0.88, 0.95, 0.99) for ($T = 30, 50, 100$), respectively, while specificity remained relatively constant, (0.75, 0.74.0.74). For the multi-VAR approach, both sensitivity and specificity increased as the time series length increased. Sensitivity and specificity ranged from (0.90, 0.95, 0.98) and (0.70, 0.74, 0.75), respectively. For the individual-level LASSO sensitivity also increased as the time series length increased, (0.73, 0.90, 0.98), while specificity decreased with larger sample sizes (0.83, 0.80, 0.78).

The standard multi-VAR showed slight increases in both specificity as the number of variables included in the analysis increased from 10 to 30. Specificity and sensitivity increased by 0.02

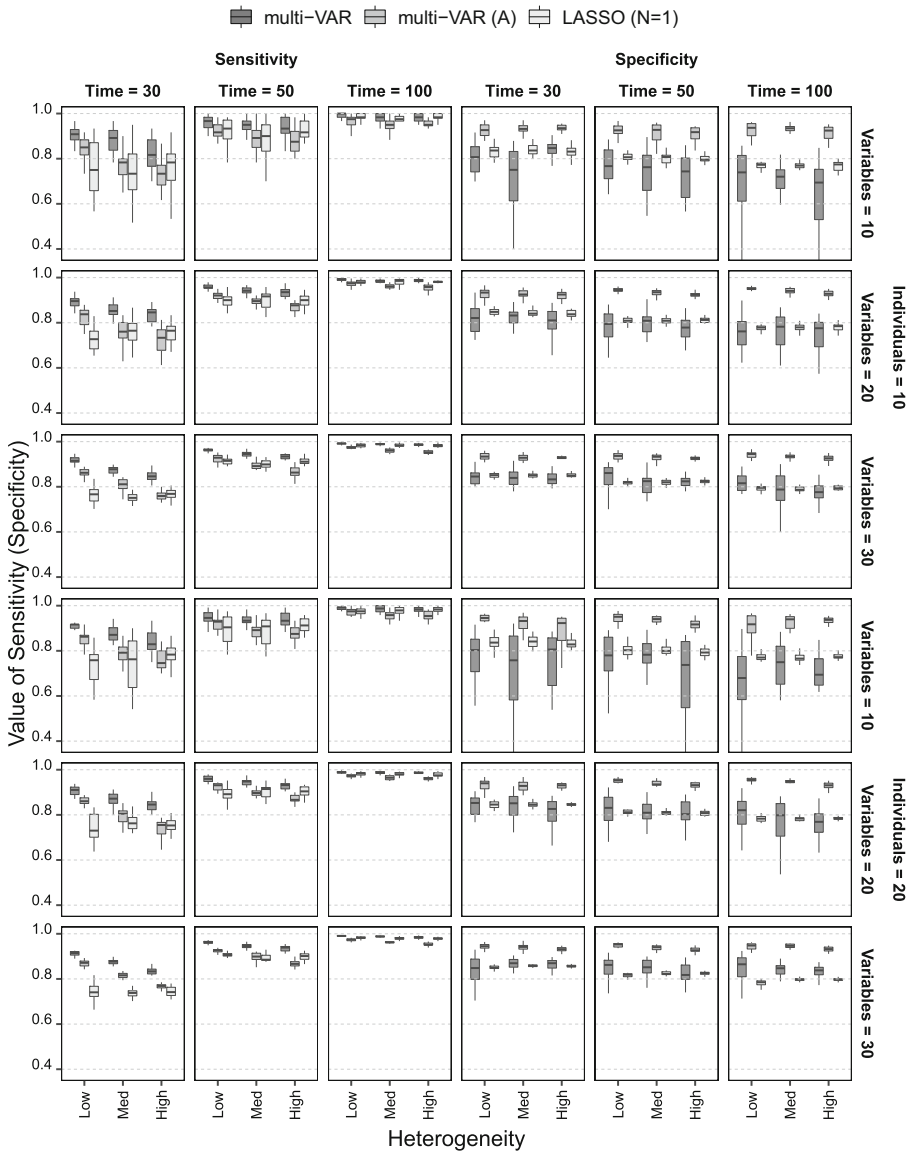

FIGURE 1.
Sensitivity and specificity for $\mathbf{B}_k$ (estimation error approach).

and 0.05, respectively. Similarly, the individual LASSO showed increases in sensitivity 0.04 and specificity 0.02 as the number of included variable increased. The multi-VAR approach showed no changes in sensitivity as the number of variables increased, and a small increase in specificity, from 0.69 to 0.76. It should be noted that for the individual LASSO, the time series lengths and variable dimensions considered here are quite small.

3.3.2. Sensitivity and Specificity for $\boldsymbol{\mu}$ In addition to looking at the recovery of $\mathbf{B}_k$, we are also interested in the recovery of the common effect matrix $\boldsymbol{\mu}$. Figure 3 shows the sensitivity and specificity for the two multi-VAR approaches in recovering the common effects. We do not

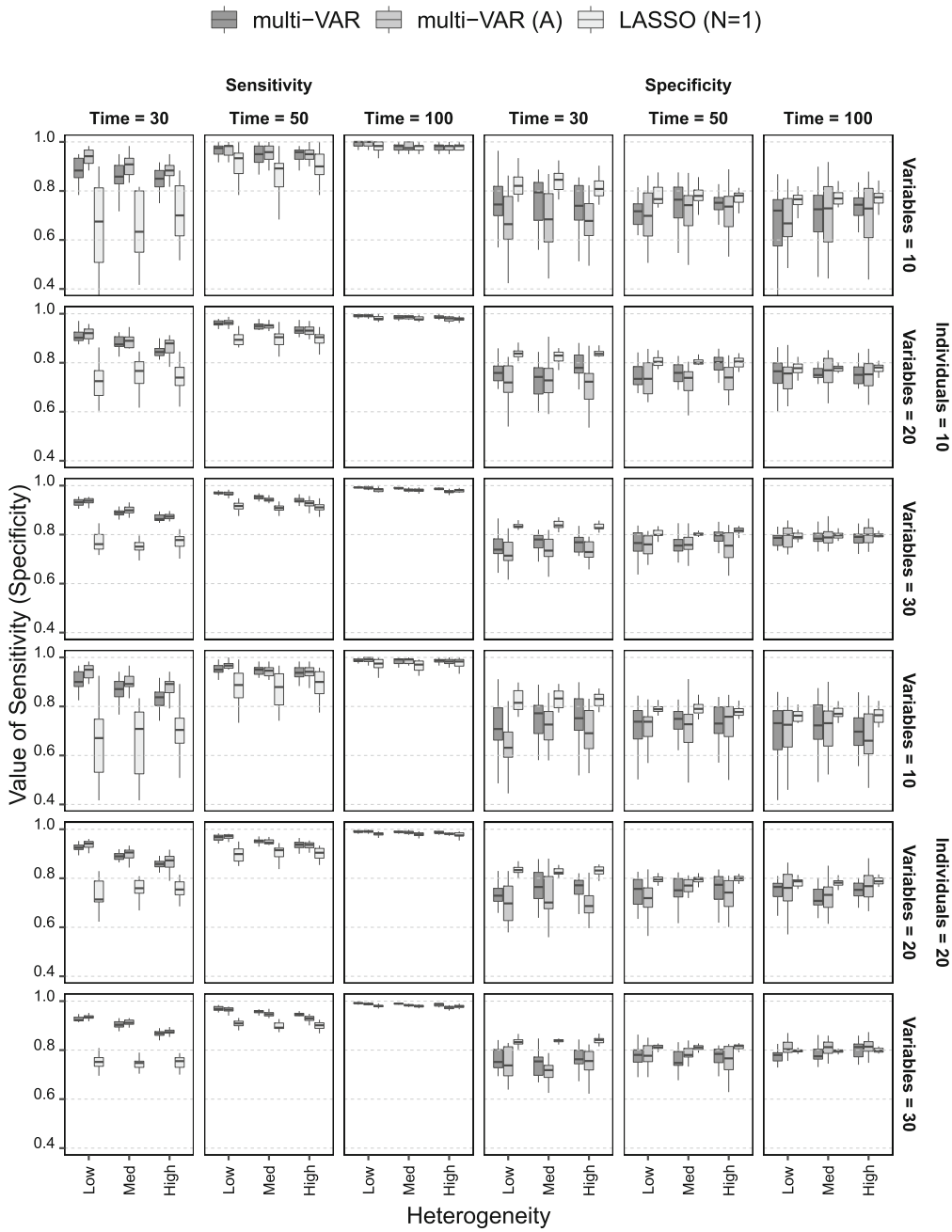

FIGURE 2.
Sensitivity and specificity for $\mathbf{B}_k$ (rolling window cross-validation).

include the individual-level LASSO here as it does not explicitly model group and unique model components. It is clear from the sensitivity plots in Figure 3 that both multi-VAR procedures do well in consistently capturing the common effects across all simulation blocks (sensitivity = 0.99). In terms of specificity, averaged over all simulation conditions, the standard multi-VAR obtains a specificity of 0.86 and the adaptive version 0.87. Importantly, although we see slight increases

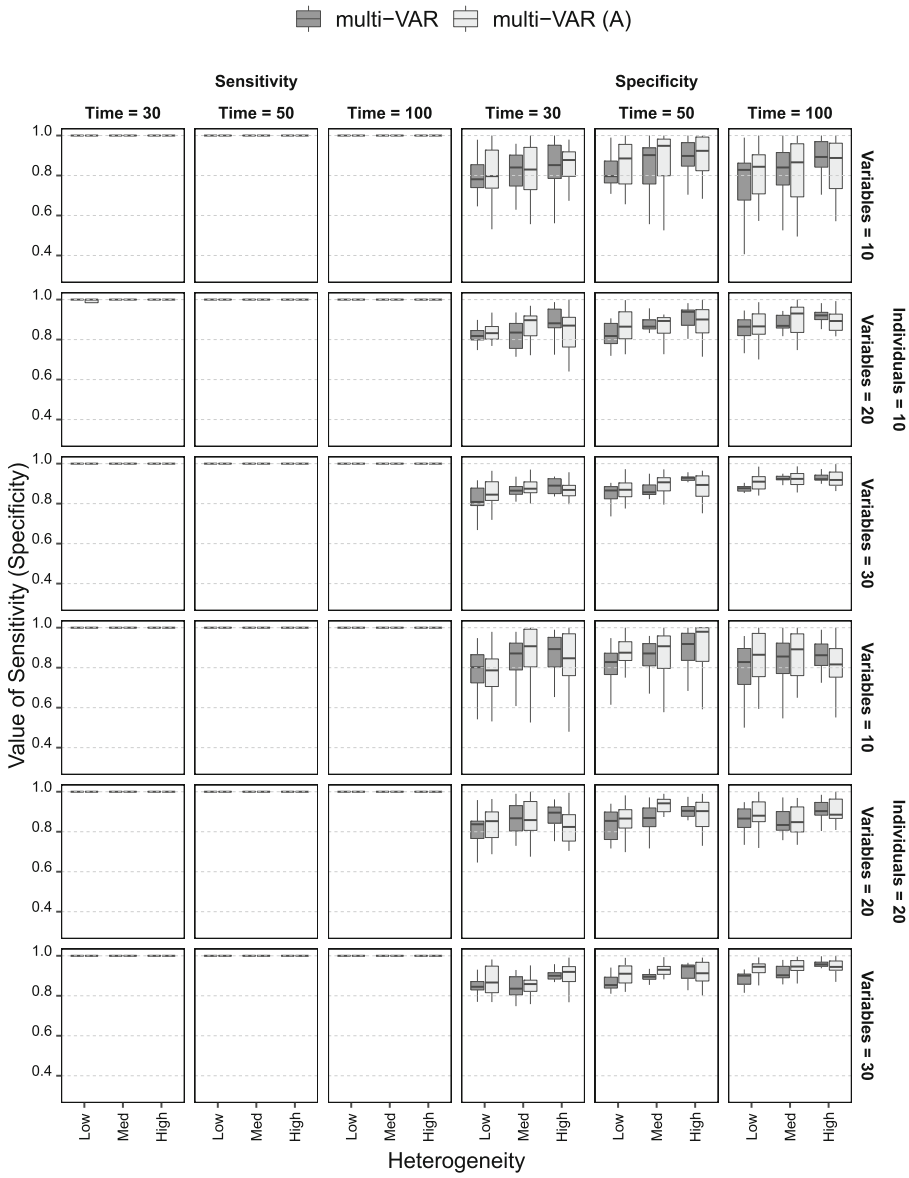

FIGURE 3.
Sensitivity and specificity for $\boldsymbol{\mu}$ (rolling window cross-validation).

in specificity as heterogeneity increases and $\boldsymbol{\mu}$ becomes sparser, both approaches perform well in terms of identifying the zero elements of $\boldsymbol{\mu}$.

It is also worth examining the recovery of the unique effect matrices $\boldsymbol{\Delta}_k$. Figure 4 shows the sensitivity and specificity for the two multi-VAR approaches in recovering the effects unique to each individual. As one might expect the sensitivity of our approaches for recovering $\boldsymbol{\Delta}_k$ is lower when compared to the common effect matrix, as these effects are less persistent in the parameter space. Here, we see a slight increase in sensitivity to recovering $\boldsymbol{\Delta}_k$ as $\boldsymbol{\mu}$ becomes sparser, but this is mostly at the smallest number of timepoints $T = 30$. Specificity of both approaches remains high

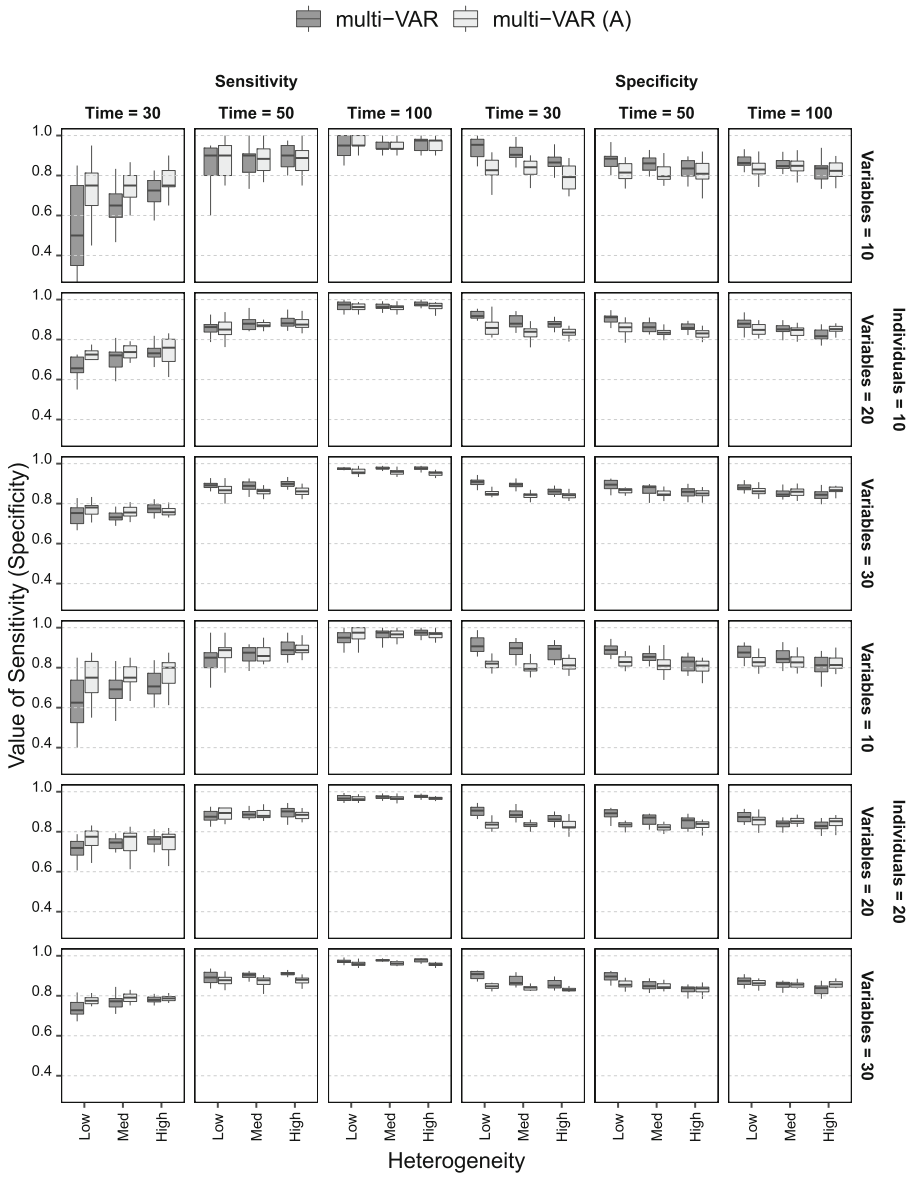

FIGURE 4.
Sensitivity and specificity for $\mathbf{\Delta}_k$ (rolling window cross-validation).

across all simulation conditions suggesting that recovery of the unique effects is quite balanced across the two performance measures.

3.3.3. Forecasting Performance The RMSFE across all simulation conditions and forecast horizons are given in Table 1. Again, these results are tabulated using the penalty parameters chosen with RWCW procedure. As mentioned previously, for the individual LASSO, we would not expect forecasting performance to be impacted by heterogeneity levels or the number of individuals in the sample. In aggregate, collapsing across the other simulation factors, this was also true for both multi-VAR approaches. In addition, for both time series length and the number

of variables, all three penalized regression procedures showed small decreases in RMSFE for the 1-unit forecast horizon as time series length and number of variables increased. For forecast horizons of 2 and 3, there was less variability across the approaches and simulation factors.

For the simulation conditions, we also examined the performance of various benchmark methods; (a) the series mean, (b) the AR(1) model for each component series, and (c) the VAR(1) model. The RMSFE for these benchmark methods are given in Table 2. For the forecast horizon of 1, the regularization methods outperformed the benchmark methods in terms of RMSFE, (LASSO = 0.89, multi-VAR = 0.87, multi-VAR (A) = 0.89, mean = 1.00, AR(1) = 1.00, VAR(1) = 1.26). For the forecast horizon of 2, the estimators performed more similarly, (LASSO = 0.95, multi-VAR = 0.95, multi-VAR (A) = 0.95, mean = 1.00, AR(1) = 1.00, VAR(1) = 1.31). This trend continued for the forecast window of 3, (LASSO = 0.97, multi-VAR = 0.97, multi-VAR (A) = 0.97, mean = 1.00, AR(1) = 1.00, VAR(1) = 1.34).

4. An Illustrative Example

We now present an empirical example based on Fredrickson et al. (2017) who examined the day-to-day emotional experiences of a nonclinical adult sample across an eleven week period. Each evening across an 11-week period participants evaluated their daily emotional experiences using the modified differential emotions scale (mDES) (Fredrickson, 2013). The mDES is a 20-item measure representing ten positive emotions and ten negative emotions. For the purpose of our current study, we included all 10 indicators for each of the emotion constructs. The question of how best to handle missing data within penalized estimation framework is an open question, and currently missing data routines are not supported in the multi-VAR framework. For this reason, we retained subjects with less than 10% missing data and imputed the missing values component-wise using the predicted values from a single run of the Kalman Filter. This procedure left us with 16 subjects on which to conduct our analysis.

For each of these 16 individuals in our sample, we partitioned the data matrix into a training and test set. The training set contained the first 77 days of the 82-day observation period and was used to estimate the various model parameters. The test set contained the final 5 days of the observation period and was used to evaluate the accuracy of the different methods. In addition to the individual-level LASSO and multi-VAR approaches (standard and adaptive), we also considered some benchmark forecasting methods. These methods include (a) the series *average* where all future forecasts are equal to the mean of the training data, (b) a *naïve* method where all forecasts are set to the value of the last observation in the training set, (c) a *drift* method which consists of drawing a straight line between the first and final observation of the training set, and extrapolating that trend line into the test set, (d) an *AR(1)* model fit to each component of the training series and (e) an unrestricted *VAR(1)* model. Root-mean-squared forecast error was used to evaluate forecasts for each of the 5 forecast horizons. Both the individual-level LASSO and multi-VAR approaches require tuning the λ regularization parameters. To select the optimal λ values, we used the rolling-window cross-validation approach described previously.

The 1–5-step ahead forecast accuracy for the individual methods is given in Table 3. The LASSO-based approaches performed similarly and obtained the smallest forecast error across the forecasting approaches we evaluated. Within the LASSO approaches, the two multi-VAR approaches performed the best in aggregate. In addition, Figure 5 provides a snapshot of the recovered transition matrices across the three approaches. In the first row of the figure, the transition matrices for Subject 1 from each algorithm are shown. The second row of Figure 5 provides a comparison of common effects resolved from the different approaches. For the individual LASSO, the matrix represents the median effects across all individuals. For the multi-VAR approaches, the transition matrices are the common effect matrices obtained from the algorithm directly. Lastly,

TABLE 1.
Root-mean-squared forecast error for simulation study conditions.

			H step-ahead forecast								
			$H = 1$			$H = 2$			$H = 3$		
Number of			Model			Model			Model		
Subjects	Variables	Time	LASSO	m-VAR	m-VAR (A)	LASSO	m-VAR	m-VAR (A)	LASSO	m-VAR	m-VAR (A)
					Heterogeneity: low						
10	10	30	0.96	0.97	0.98	0.94	0.94	0.94	0.98	0.99	1.00
		50	0.94	0.94	0.94	0.97	0.96	0.95	0.98	0.99	0.99
		100	0.96	0.96	0.94	0.96	0.95	0.95	1.00	0.98	1.00
	20	30	0.87	0.89	0.91	0.93	0.94	0.94	0.98	0.99	0.98
		50	0.85	0.87	0.89	0.97	0.97	0.97	0.99	0.99	0.97
		100	0.86	0.87	0.88	0.94	0.94	0.97	0.97	0.97	0.98
	30	30	0.85	0.88	0.89	0.96	0.98	0.98	0.99	1.00	1.00
		50	0.81	0.82	0.83	0.93	0.93	0.96	0.98	0.98	0.98
		100	0.79	0.80	0.79	0.91	0.91	0.89	0.95	0.95	0.95
20	10	30	0.91	0.92	0.94	0.95	0.95	0.95	0.96	0.96	0.96
		50	0.92	0.93	0.91	0.97	0.97	0.95	0.97	0.98	0.96
		100	0.92	0.92	0.93	0.95	0.96	0.95	0.98	0.98	0.96
	20	30	0.89	0.92	0.93	0.94	0.95	0.95	0.98	0.98	0.98
		50	0.86	0.87	0.88	0.93	0.94	0.95	0.95	0.95	0.96
		100	0.85	0.85	0.85	0.95	0.95	0.96	0.96	0.96	0.98
	30	30	0.84	0.87	0.89	0.94	0.96	0.96	0.96	0.97	0.97
		50	0.80	0.81	0.84	0.91	0.92	0.91	0.94	0.94	0.96
		100	0.81	0.82	0.83	0.92	0.92	0.92	0.97	0.97	0.97

TABLE 1.
continued

Number of			H step-ahead forecast								
			$H = 1$			$H = 2$			$H = 3$		
			Model			Model			Model		
Subjects	Variables	Time	LASSO	m-VAR	m-VAR (A)	LASSO	m-VAR	m-VAR (A)	LASSO	m-VAR	m-VAR (A)
					Heterogeneity: medium						
10	10	30	0.91	0.93	0.93	0.96	0.96	0.96	0.93	0.93	0.94
		50	0.93	0.94	0.91	0.99	0.99	0.98	0.99	0.99	1.02
		100	0.92	0.92	0.91	0.99	0.99	0.96	0.95	0.94	0.94
	20	30	0.90	0.92	0.91	0.96	0.97	0.96	0.98	0.98	0.98
		50	0.88	0.90	0.89	0.94	0.95	0.95	0.98	0.98	0.97
		100	0.84	0.85	0.89	0.94	0.94	0.95	0.96	0.96	0.98
	30	30	0.85	0.88	0.90	0.97	0.99	0.98	0.97	0.98	0.98
		50	0.83	0.85	0.83	0.94	0.95	0.94	0.99	1.00	0.97
		100	0.79	0.80	0.80	0.88	0.88	0.91	0.94	0.94	0.95
20	10	30	0.95	0.97	0.96	0.96	0.97	0.98	0.94	0.94	0.95
		50	0.94	0.95	0.90	0.96	0.95	0.95	0.98	0.98	0.96
		100	0.93	0.93	0.94	0.97	0.97	0.95	0.96	0.96	0.96
	20	30	0.88	0.91	0.91	0.94	0.95	0.95	0.97	0.97	0.97
		50	0.87	0.88	0.88	0.94	0.95	0.95	0.98	0.98	0.99
		100	0.86	0.86	0.87	0.97	0.98	0.97	0.99	0.99	0.99
	30	30	0.86	0.88	0.88	0.94	0.95	0.95	0.97	0.98	0.98
		50	0.82	0.84	0.84	0.93	0.94	0.95	0.97	0.97	0.97
		100	0.81	0.82	0.81	0.92	0.93	0.93	0.97	0.97	0.97

TABLE 1.
continued

Number of			H step-ahead forecast								
			$H = 1$			$H = 2$			$H = 3$		
			Model			Model			Model		
Subjects	Variables	Time	LASSO	m-VAR	m-VAR (A)	LASSO	m-VAR	m-VAR (A)	LASSO	m-VAR	m-VAR (A)
					Heterogeneity: high						
10	10	30	0.92	0.94	0.93	0.97	0.97	0.97	0.94	0.94	0.93
		50	0.91	0.92	0.95	0.95	0.95	0.94	0.95	0.95	0.95
		100	0.92	0.92	0.92	0.95	0.95	0.99	0.94	0.95	0.94
	20	30	0.89	0.92	0.90	0.98	0.99	0.97	0.93	0.94	0.93
		50	0.91	0.94	0.89	0.95	0.96	0.95	1.00	1.00	0.98
		100	0.88	0.88	0.86	0.94	0.94	0.93	0.99	0.99	0.94
	30	30	0.84	0.87	0.88	0.93	0.94	0.94	0.97	0.97	0.98
		50	0.81	0.83	0.82	0.91	0.92	0.93	0.97	0.97	0.97
		100	0.79	0.79	0.80	0.91	0.91	0.92	0.98	0.97	0.97
20	10	30	0.89	0.91	0.90	0.94	0.94	0.93	0.95	0.95	0.96
		50	0.91	0.91	0.89	0.95	0.95	0.97	0.97	0.97	0.98
		100	0.90	0.90	0.91	0.93	0.93	0.96	0.97	0.97	0.98
	20	30	0.90	0.93	0.92	0.95	0.96	0.96	0.98	0.98	0.98
		50	0.87	0.89	0.87	0.96	0.97	0.97	0.99	1.00	1.00
		100	0.88	0.88	0.89	0.95	0.95	0.95	0.98	0.98	0.98
	30	30	0.87	0.90	0.89	0.95	0.96	0.95	0.97	0.97	0.97
		50	0.84	0.86	0.84	0.92	0.93	0.94	0.96	0.97	0.97
		100	0.80	0.80	0.82	0.92	0.92	0.92	0.97	0.98	0.96

Note. (A) indicates adaptive version of the multi-VAR.

TABLE 2.
Root-mean-squared forecast error for benchmark methods across simulation conditions.

			H step-ahead forecast								
			$H = 1$			$H = 2$			$H = 3$		
Number of			Model			Model			Model		
Subjects	Variables	Time	Mean	AR(1)	VAR(1)	Mean	AR(1)	VAR(1)	Mean	AR(1)	VAR(1)
10	10	30	0.92	0.94	0.93	0.97	0.97	0.97	0.94	0.94	0.93
		50	0.91	0.92	0.95	0.95	0.95	0.94	0.95	0.95	0.95
		100	0.92	0.92	0.92	0.95	0.95	0.99	0.94	0.95	0.94
	20	30	0.89	0.92	0.90	0.98	0.99	0.97	0.93	0.94	0.93
		50	0.91	0.94	0.89	0.95	0.96	0.95	1.00	1.00	0.98
		100	0.88	0.88	0.86	0.94	0.94	0.93	0.99	0.99	0.94
	30	30	0.84	0.87	0.88	0.93	0.94	0.94	0.97	0.97	0.98
		50	0.81	0.83	0.82	0.91	0.92	0.93	0.97	0.97	0.97
		100	0.79	0.79	0.80	0.91	0.91	0.92	0.98	0.97	0.97
20	10	30	0.89	0.91	0.90	0.94	0.94	0.93	0.95	0.95	0.96
		50	0.91	0.91	0.89	0.95	0.95	0.97	0.97	0.97	0.98
		100	0.90	0.90	0.91	0.93	0.93	0.96	0.97	0.97	0.98
	20	30	0.90	0.93	0.92	0.95	0.96	0.96	0.98	0.98	0.98
		50	0.87	0.89	0.87	0.96	0.97	0.97	0.99	1.00	1.00
		100	0.88	0.88	0.89	0.95	0.95	0.95	0.98	0.98	0.98
	30	30	0.87	0.90	0.89	0.95	0.96	0.95	0.97	0.97	0.97
		50	0.84	0.86	0.84	0.92	0.93	0.94	0.96	0.97	0.97
		100	0.80	0.80	0.82	0.92	0.92	0.92	0.97	0.98	0.96

TABLE 3.
Root-mean-squared forecast error for Fredrickson et al. (2017) Data.

Method	Forecast window length				
	1	2	3	4	5
Mean	0.84	0.88	1.00	1.05	0.98
Naïve	0.89	1.15	1.34	1.31	1.35
Drift	0.89	1.17	1.36	1.34	1.39
AR(1)	0.79	0.88	1.01	1.05	0.98
VAR(1)	0.83	0.90	1.03	1.04	0.97
LASSO	0.82	0.86	1.00	1.02	0.94
multi-VAR	0.75	0.85	0.97	1.03	0.95
multi-VAR (A)	0.76	0.85	0.97	1.03	0.95

Note. (A) indicates adaptive multi-VAR.

for each method, the third row provides the path frequency counts across all individuals in the sample. Here, one can see a similar pattern of sparsity, as well as clustering within the positive and negative sub-scales. This is consistent with previous studies which have used a bivariate dynamic factor analysis approach to model positive and negative items from the mDES as representing distinct but interdependent constructs (Fisher et al., 2020).

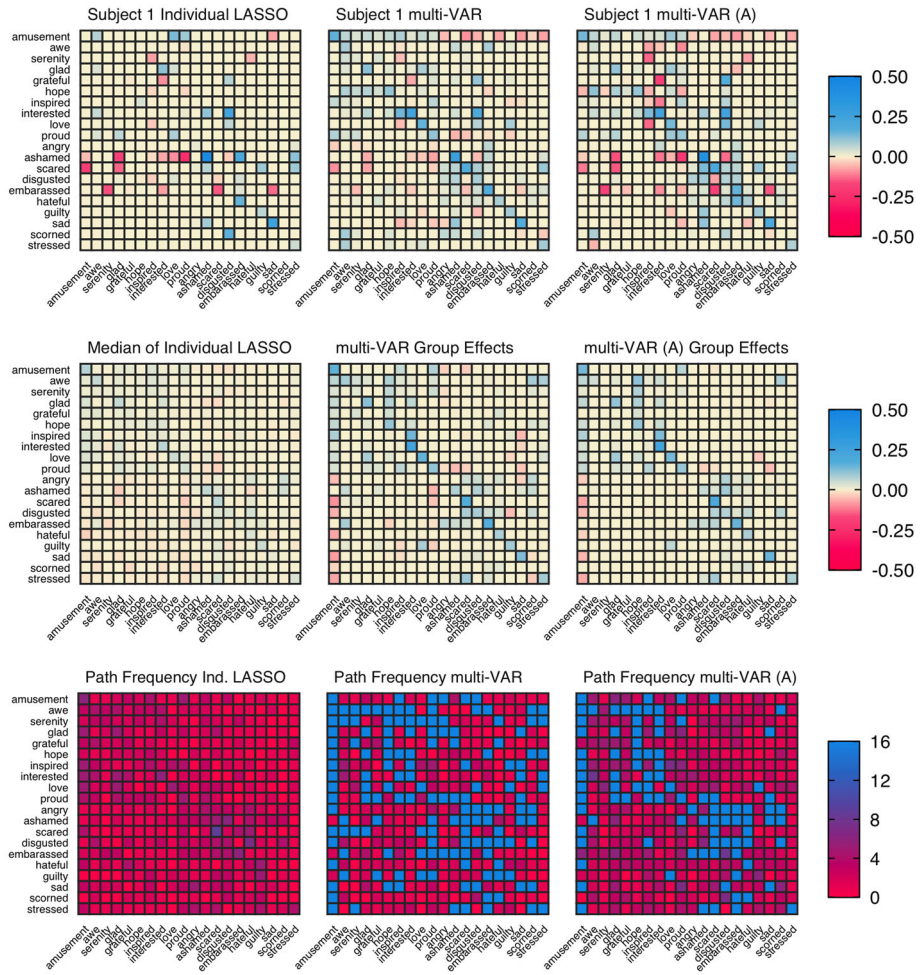

FIGURE 5.
Results from Fredrickson et al. (2017) data across approaches.

5. Discussion

This paper presents a novel approach for synthesizing multivariate time series obtained from multiple individuals. This method is especially well-suited to ILD paradigms when it is unclear how much individuals differ in terms of their dynamic processes. If individuals share little in common results from the proposed method resemble what would be obtained from fitting separate models to each individual. If individuals are homogenous results resemble what would be obtained from pooling the data and fitting a single model to the sample. Most importantly, if the truth lies somewhere in between these extremes—certain dynamics are shared while others are idiosyncratic—results will reflect this and provide researchers with new tools for isolating generalizable dynamics. Importantly, the simulation results presented here confirm that across three different levels of cross-sectional heterogeneity, the proposed methods perform well in recovering the model dynamics and forecasting compared to benchmark methods.

Despite these developments, a number of limitations and opportunities for future development are worth considering. First, although we addressed some limitations of the VAR modeling in

the context of ILD, others remain. For example, we assume the parameters themselves do not vary across time. This may be a strong assumption in the context of emotional dynamics. Second, in our simulation study and empirical example, we set the regularization parameter λ_2 to be constant across individuals. In practice, this is unlikely to hold and relaxing this assumption would potentially lead to better performance for the multi-VAR approaches, at an increased computational cost.

Relatedly, a limitation of our current work and an important area for future development involves identifying alternative methods for resolving the optimal multi-VAR penalty parameters. It was noted in the simulation study performance of the adaptive multi-VAR approach degraded considerably when the penalty parameters were chosen by RWCV. Although there appears to be little consensus on which performance estimation method works best in the case of time series data, two approaches are often considered: (1) Out-of-sample (OOS), and (2) cross-validation (CV) methods (Cerqueira et al., 2020). Choosing an appropriate method depends on the specific characteristics of the data. The difference between OOS and CV methods is that OOS methods always preserve the temporal order of the observations, and a model is never tested on historical data, relative to the training data. CV approaches, such as K-folds CV, often break the temporal order of time series and may produce poor estimates of predictive performance in real time series contexts.

However, recent work has empirically demonstrated that CV methods perform well for stationary time series data, even outperforming OOS approaches in some circumstances (Bulteel et al., 2018a; Bergmeir & Benítez, 2012; Bergmeir et al., 2014, 2018). The reasons for this are not entirely clear although a number of adaptations to traditional K-fold CV have been made to accommodate time dependence through block-sampling (see Bulteel et al. (2018a) for a review). One explanation may be that CV approaches more efficiently use the available data, without requiring hold-out or initialization samples. CV approaches may be highly relevant for the smaller sample sizes considered here and the adaptive multi-VAR framework. Future work should explore which if any of the existing OOS or CV approaches are particularly well-suited for the multi-VAR construction.

Another important area of development is to compare the multi-VAR approach to other frameworks capable of handling multiple-subject time series data. Two prominent methods are multilevel time series modeling (Bringmann et al., 2013; Epskamp et al., 2018) and GIMME (Gates & Molenaar, 2012). While the current procedure relies on the VAR model, GIMME is based on the structural-VAR, making a naive comparison difficult. Work is currently under way to extend multi-VAR to the structural- and graphical-VAR frameworks. It is our hope these extensions will provide researchers with more tools for flexibly accommodating cross-sectional dependence in time series data.

Based on the described results, approaches capable of accommodating individual idiosyncrasies while exploiting what is common hold great promise for improving our ability to characterize and forecast complex physical and mental health outcomes at the individual level. In this vein, we are optimistic that the continued adoption of forecasting methodology by social and behavioral science researchers will only help to further integrate the nomothetic and idiographic approaches.

Acknowledgments

Vladas Pipiras was supported in part by the NSF grant DMS-1712966.

Appendix

In this technical appendix, we discuss some theoretical aspect of LASSO estimation in the multi-VAR setting, namely concerning its consistency and sparsistency.

Consistency

Consistency of LASSO estimation for single (stable) VAR models was established in the seminal paper by Basu and Michailidis (2015a, 2015b), building upon such results in the regression setting by Loh and Wainwright (2012a, 2012b). In the multi-VAR setting, the model is inherently unidentifiable. It could be that the LASSO solution is consistent for some particular $\boldsymbol{\mu}^*$, $\boldsymbol{\Delta}_k^*$ in the model (7), or over a subset of such identifications, but this problem still appears largely unresolved. Some related result though can be found in the discussion on sparsistency below following Ollier and Viallon (2017). Here, we shall discuss a weaker form of consistency of $\hat{\mathbf{B}}_k = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\Delta}}_k$ for $\mathbf{B}_k^*$. The arguments are quite straightforward and shed some light on the problem, and also seemingly were not made in the related literature yet.

We first describe the basic result for a single VAR model expressed in the regression form (5), and then turn to a multi-VAR model. We index the model quantities with subscript k or superscript (k), $k = 1, \ldots, K$, representing the individual models in the multi-VAR setting. After expanding the quadratic term of the objective function (6), the estimation equation can be rewritten as in Basu and Michailidis (2015a) in terms of the quantities

$$\widehat{\Gamma}_k = \frac{1}{N}\mathbf{Z}^{(k)'}\mathbf{Z}^{(k)} = \frac{1}{N}(\mathrm{I}_d \otimes \mathcal{X}^{(k)}\mathcal{X}^{(k)'}), \quad \hat{\gamma}_k = \frac{1}{N}\mathbf{Z}^{(k)'}\mathbf{Y}^{(k)}. \tag{31}$$

Estimation consistency is proved under the following two conditions on these quantities:

- Restricted eigenvalue condition: The matrix $\widehat{\Gamma}_k$ is said to satisfy this condition with parameters $\alpha_k, \tau_k > 0$, if
$$\beta_k'\widehat{\Gamma}_k\beta_k \geq \alpha_k\|\beta_k\|_2^2 - \tau_k\|\beta_k\|_1^2, \quad \beta_k \in \mathbb{R}^q, \tag{32}$$
with $q = pd^2$.
- Deviation condition: This condition is satisfied if
$$\|\hat{\gamma}_k - \widehat{\Gamma}_k\mathbf{B}_k^*\|_\infty \leq Q_k(\mathbf{B}_k^*, \Sigma_{k,\varepsilon})\sqrt{\frac{\log q}{N}}, \tag{33}$$
for a deterministic function Q_k.

Let $s_k = \|\mathbf{B}_k^*\|_0$ denote the sparsity of the model. Under the conditions above and assuming $s_k\tau_k \leq \alpha_k/32$, Proposition 4.1 of Basu and Michailidis (2015a) states that any solution $\hat{\mathbf{B}}$ of (6) satisfies: for any $\lambda \geq 4Q_k(\mathbf{B}_k^*, \Sigma_{k,\varepsilon})\sqrt{\frac{\log q}{N}}$,

$$\|\hat{\mathbf{B}}_k - \mathbf{B}_k^*\|_1 \leq \frac{64s_k\lambda}{\alpha_k}, \quad \|\hat{\mathbf{B}}_k - \mathbf{B}_k^*\|_2 \leq \frac{16\sqrt{s_k}\lambda}{\alpha_k}. \tag{34}$$

Additionally, a result on the support of thresholded estimators of $\hat{\mathbf{B}}_k$ is also available. The consistency results in (34) apply to generic LASSO estimators as long as the quantities $\widehat{\Gamma}_k$, $\hat{\gamma}_k$ satisfy the restricted eigenvalue and deviation conditions.

Among the key contributions of Basu and Michailidis (2015a) are their results (Propositions 4.2 and 4.3) proving that $\widehat{\Gamma}_k$ and $\hat{\gamma}_k$ satisfy the restricted eigenvalue and deviation conditions with high enough probabilities, and expressing the various parameters involved in the conditions $(\alpha_k, \tau_k, Q_k(\mathbf{B}_k^*, \Sigma_{k,\varepsilon}))$ in terms of the VAR model parameters. Furthermore, in the restricted eigenvalue condition, τ_k can be chosen so that $s_k\tau_k \leq \alpha_k/32$. We also note that the right-hand side of the inequalities (34) are expected to be negligible for small λ and hence small $\log q/N$. The case when the logarithm of the dimension compares to the sample size through this way is the typical LASSO scenario.

In the multi-VAR setting, the optimization problem (9) can be expressed through the objective function

$$-\sum_{k=1}^{K} 2\mathbf{B}_k'\hat{\gamma}_k + \sum_{k=1}^{K} \mathbf{B}_k'\widehat{\Gamma}_k\mathbf{B}_k + \lambda_1\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K}\lambda_{2,k}\|\mathbf{B}_k - \boldsymbol{\mu}\|_1. \tag{35}$$

A consistency bound for the minimizer $\hat{\mathbf{B}}_k$ of (35) can still be obtained similarly as for single VAR models if one is willing to make the assumption

$$\|\hat{\boldsymbol{\mu}}\|_0 \leq s_0. \tag{36}$$

The constraint (36) could be imposed while optimizing (35) or choosing λ_1 appropriately large, or inferred to hold (with high enough probability) from sparsistency result, if available. Indeed, under (36), a consistency bound can be derived easily as in the proof of Proposition 3.3 in Basu and Michailidis (2015a, 2015b). That is, observe first that

$$\begin{aligned}
&-\sum_{k=1}^{K} 2\hat{\mathbf{B}}_k'\hat{\gamma}_k + \sum_{k=1}^{K} \hat{\mathbf{B}}_k'\widehat{\Gamma}_k\hat{\mathbf{B}}_k + \lambda_1\|\hat{\boldsymbol{\mu}}\|_1 + \sum_{k=1}^{K}\lambda_{2,k}\|\hat{\mathbf{B}}_k - \hat{\boldsymbol{\mu}}\|_1 \\
&\leq -\sum_{k=1}^{K} 2\mathbf{B}_k^{*'}\hat{\gamma}_k + \sum_{k=1}^{K} \mathbf{B}_k^{*'}\widehat{\Gamma}_k\mathbf{B}_k^* + \lambda_1\|\hat{\boldsymbol{\mu}}\|_1 + \sum_{k=1}^{K}\lambda_{2,k}\|\mathbf{B}_k^* - \hat{\boldsymbol{\mu}}\|_1
\end{aligned}$$

and rearranging the terms and setting $\mathbf{v}_k = \hat{\mathbf{B}}_k - \mathbf{B}_k^*$, we deduce

$$\sum_{k=1}^{K}\mathbf{v}_k\widehat{\Gamma}_k\mathbf{v}_k \leq \sum_{k=1}^{K} 2\mathbf{v}_k'(\hat{\gamma}_k - \widehat{\Gamma}_k\mathbf{B}_k^*) + \sum_{k=1}^{K}\lambda_{2,k}\left(\|\mathbf{B}_k^* - \hat{\boldsymbol{\mu}}\|_1 - \|\mathbf{B}_k^* - \hat{\boldsymbol{\mu}} + \mathbf{v}_k\|_1\right).$$

With $\hat{J}_k = \mathrm{supp}\{\mathbf{B}_k^* - \hat{\boldsymbol{\mu}}$ being the index support of $\mathbf{B}_k^* - \hat{\boldsymbol{\mu}}\}$, repeating the argument in Basu and Michailidis (2015a, 2015b), we get

$$0 \leq \sum_{k=1}^{K}\mathbf{v}_k'\widehat{\Gamma}_k\mathbf{v}_k \leq \sum_{k=1}^{K}\left(\frac{3\lambda_{2,k}}{2}\|(\mathbf{v}_k)_{\hat{J}_k}\|_1 - \frac{\lambda_{2,k}}{2}\|(\mathbf{v}_k)_{\hat{J}_k^c}\|_1\right) \tag{37}$$

as long as $\lambda_{2,k} \geq 4Q_k(\mathbf{B}_k^*, \Sigma_{k,\varepsilon})\sqrt{\frac{\log q}{N}}$ (with the function Q_k from the deviation condition), where $(\cdot)_{\hat{J}}$ and $(\cdot)_{\hat{J}^c}$ denote restrictions to the index sets $\hat{J}$ and $\hat{J}^c$, respectively. Then,

$$\sum_{k=1}^{K} \lambda_{2,k} \|(\mathbf{v}_k)_{\hat{J}_k^c}\|_1 \leq 3 \sum_{k=1}^{K} \lambda_{2,k} \|(\mathbf{v}_k)_{\hat{J}_k}\|_1$$

and one also has

$$\begin{aligned} \sum_{k=1}^{K} \lambda_{2,k} \|\mathbf{v}_k\|_1 &\leq 4 \sum_{k=1}^{K} \lambda_{2,k} \|(\mathbf{v}_k)_{\hat{J}_k}\|_1 \\ &\leq 4 \sum_{k=1}^{K} \lambda_{2,k} (s_0 + s_k)^{1/2} \|\mathbf{v}_k\|_2 \leq 4 \sqrt{\sum_{k=1}^{K} \lambda_{2,k}^2 (s_0 + s_k)} \|\mathbf{v}\|_2, \end{aligned} \tag{38}$$

by Cauchy–Schwarz inequality (twice) and the fact that $|\text{supp}\{\hat{J}_K\}| \leq s_0 + s_k$, where $\|v\|_2^2 = \sum_{k=1}^{K} \|v_k\|_2^2$. Similarly, by the restricted eigenvalue condition (32) for each model and assuming $s_k \tau_k \leq \alpha_k/32$, we have

$$\sum_{k=1}^{K} \mathbf{v}_k' \widehat{\Gamma}_k \mathbf{v}_k \geq \sum_{k=1}^{K} \frac{\alpha_k}{2} \|\mathbf{v}_k\|_2^2 \geq \frac{\min\{\alpha_k\}}{2} \|\mathbf{v}\|_2^2. \tag{39}$$

A combination of (37)–(39) yields, e.g.,

$$\frac{\min\{\alpha_k\}}{2} \|\mathbf{v}\|_2^2 \leq 6 \sqrt{\sum_{k=1}^{K} \lambda_{2,k}^2 (s_0 + s_k)} \|\mathbf{v}\|_2 \tag{40}$$

or

$$\|\mathbf{v}\|_2 \leq \frac{12 \sqrt{\sum_{k=1}^{K} \lambda_{2,k}^2 (s_0 + s_k)}}{\min\{\alpha_k\}}. \tag{41}$$

This is the multi-VAR analogue of the second consistency bound in (34). One can similarly obtain a bound on $\|\mathbf{v}\|_1$ analogous to the first one in (34).

Sparsistency

We comment here briefly on the possibility of recovering the supports of $\boldsymbol{\mu}^*$ and $\boldsymbol{\Delta}_k^*$. The same issue of (non)identifiability is fundamental here as well. Some result nevertheless are available in the literature for special cases. Assuming effectively that $s\lambda_1/\lambda_{2,k} = cK^{1/2}$, Ollier and Viallon (2017) gave conditions for identifiability and sparsistency with the limiting common parameter of interest $\boldsymbol{\mu}^*$ defined as the entrywise median of $\boldsymbol{B}_k^*$. Their approach goes through verifying a particular well-known irrepresentability condition on a design matrix. It could in principle be adapted to the multi-VAR context but the value of this effort might be questionable. First, irrepresentability conditions are quite restrictive and difficult to verify, and as a result, adaptive LASSO versions are advocated for. The setting where the limiting parameter of interest is necessarily related to the median could also be viewed restrictive.

References

Allen, P. G., & Morzuch, B. J. (2006). Twenty-five years of progress, problems, and conflicting evidence in econometric forecasting. What about the next 25 years? *International Journal of Forecasting, 22*(3), 475–492.

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics, 25*(1), 71–92.

Basu, S., & Michailidis, G. (2015a). Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics, 43*(4), 1535–1567.

Basu, S., & Michailidis, G. (2015b). Supplement to "Regularized estimation in sparse high-dimensional time series models". *Annals of Statistics, 43*(4), 1535–1567.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences, 2*(1), 183–202.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences: An International Journal, 191*, 192–213.

Bergmeir, C., Costantini, M., & Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis, 76*, 132–143.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis, 120*, 70–83.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8*(4), e60188.

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). Var(1) based models do not always outpredict AR(1) models in typical psychological applications. *Psychological Methods, 23*(4), 740.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2018). Improved insight into and prediction of network dynamics by combining Var and dimension reduction. *Multivariate Behavioral Research, 53*(6), 853–875.

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning, 109*(11), 1997–2028.

Chen, M., Chow, S.-M., Hammal, Z., Messinger, D. S., & Cohn, J. F. (2020). A person- and time-varying vector autoregressive model to capture interactive infant–mother head movement dynamics. *Multivariate Behavioral Research*, *56*(5), 739–767.

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research, 53*(4), 453–480.

Fisher, Z. F. (2021). multivar: Penalized estimation and forecasting of multiple subject vector autoregressive (multi-VAR) models. R package version 1.0.0. https://CRAN.R-project.org/package=multivar.

Fisher, Z. F., Chow, S.-M., Molenaar, P. C. M., Fredrickson, B. L., Pipiras, V., & Gates, K. M. (2020). A square-root second-order extended Kalman filtering approach for estimating smoothly time-varying parameters. *Multivariate Behavioral Research*, 1–19.

Fredrickson, B. L. (2013). Chapter One—Positive emotions broaden and build. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 1–53). Academic Press.

Fredrickson, B. L., Boulton, A. J., Firestine, A. M., Van Cappellen, P., Algoe, S. B., Brantley, M. M., Kim, S. L., Brantley, J., & Salzberg, S. (2017). Positive emotion correlates of meditation practice: A comparison of mindfulness meditation and loving-kindness meditation. *Mindfulness, 8*(6), 1623–1633.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*(1), 310–319.

Groen, R. N., Snippe, E., Bringmann, L. F., Simons, C. J. P., Hartmann, J. A., Bos, E. H., & Wichers, M. (2019). Capturing the risk of persisting depressive symptoms: A dynamic network investigation of patients' daily symptom experiences. *Psychiatry Research, 271*, 640–648.

Gross, S. M., & Tibshirani, R. (2016). Data shared lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis, 101*, 226–235.

Han, F., & Liu, H. (2013). Transition matrix estimation in high dimensional time series. In *International conference on machine learning* (pp. 172–180).

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press.

Ji, L., Chow, S.-M., Crosby, B., & Teti, D. M. (2020). Exploring sleep dynamic of mother–infant dyads using a regime-switching vector autoregressive model. *Multivariate Behavioral Research, 55*(1), 150–151.

Kock, A. B., & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics, 186*(2), 325–344.

Lane, S., Gates, K., Fisher, Z., Arizmendi, C., & Molenaar, P. (2019). *gimme: Group iterative multiple model estimation*. R package version 0.6-1.

Li, J., & Chen, W. (2014). Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting, 30*(4), 996–1015.

Loh, P.-L., & Wainwright, M. J. (2012a). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics, 40*(3), 1637–1664.

Loh, P.-L., & Wainwright, M. J. (2012b). Supplement to "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity". *Annals of Statistics, 40*(3), 1637–1664.
Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Springer.
Medeiros, M. C., & Mendes, E. F. (2016). 1-Regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics, 191*(1), 255–271.
Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181–202.
Nesterov, Y. (2007). *Gradient methods for minimizing composite objective function*. Technical Report 2007, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).
Nicholson, W. B., Matteson, D. S., & Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting, 33*(3), 627–651.
Ollier, E., & Viallon, V. (2014). Joint estimation of K related regression models with simple L_1-norm penalties. arXiv:1411.1594 [stat].
Ollier, E., & Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika, 104*(1), 83–96.
Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization, 1*(3), 127–239.
Polson, N. G., Scott, J. G., & Willard, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statistical Science, 30*(4), 559–581.
Robertson, J. C., & Tallman, E. W. (2001). Improving federal-funds rate forecasts in VAR models used for policy analysis. *Journal of Business & Economic Statistics, 19*(3), 324–330.
Sims, C. A. (1980). Macroeconomics and reality. *Econometrica, 48*(1), 1–48.
Song, S., & Bickel, P. J. (2011). Large vector auto regressions. arXiv:1106.3915 [q-fin, stat].
Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association, 97*(460), 1167–1179.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288.
Wild, B., Eichler, M., Friederich, H.-C., Hartmann, M., Zipfel, S., & Herzog, W. (2010). A graphical vector autoregressive modeling approach to the analysis of electronic diary data. *BMC Medical Research Methodology, 10*(1), 28.
Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*(298), 348–368.
Zheng, Y., Wiebe, R. P., Cleveland, H. H., Molenaar, P. C. M., & Harris, K. S. (2013). An idiographic examination of day-to-day patterns of substance use craving, negative affect, and tobacco use among young adults in recovery. *Multivariate Behavioral Research, 48*(2), 241–266.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Manuscript Received: 1 JUL 2020
Final Version Received: 13 SEP 2021
Published Online Date: 21 JAN 2022