Covariate-Dependent Vector Auto-Regresion and Universal Structural Equation Models for Explainable Idiographic Structure in Time Series

José Á. Sánchez Gómez Holly O'Rourke Gene Brewer *

Abstract

Keywords: Low-Rank Matrix Approximation; Principal Component Analysis; Spectral Methods; Weighted Networks.

*José Á. Sánchez Gómez is Assistant Professor, Department of Statistics, University of California Riverside, Riverside, CA 92521, USA. E-mail: josesa@ucr.edu. Holly O'Rourke is Associate Professor, Department of Psychology, University of California Riverside, Riverside, CA 92521, USA. Email: holly.orourke@ucr.edu. Gene Brewer is Professor, Department of Psychology, University of California Riverside, Riverside, CA 92521, USA.
Corresponding Authors: José Á. Sánchez Gómez (josesa@ucr.edu), Holly O'Rourke (holly.orourke@ucr.edu).

1 Introduction

- There is a growing availability of longitudinal data, which allows for an increase in the breath and depth of research regarding how specific psychometric measurements influence behavior, well-being and daily progress.

- Often, this data takes the form of a set of time-series, each associated with an individual patient. When modeling, we must take into account both the temporal dependency of the data (prior psychological states are likely to influence future states), and the dependencies among the psychometric measurements.

- Many methods have been proposed for the estimation of such dependencies such as sparse vector auto-regression models [1], dynamic factor analysis [8], and the universal structural equation modeling (USEM) [5].

- In psychometric research, it is vital to recognize the effect that individual characteristics of each subject can have in the overall fit of the model and interpretation. From this, a variety of works have focused on the estimation of a common (or nomothetic) structure across all individuals, as well as individual (or idiographic) structure that is exclusive to each subject. The GIMME method [4] performs the estimation of the network of psychometric relationships via the estimation of multiple USEM model, with a variable selection procedure that estimates model overlaps. More recently, the MultiVAR method [3] estimates the individual and common structure through a penalized optimization approach. Further extensions to consider the potential structure of subgroups has been considered for both approaches [6; 2].

- Often, in psychometric or MRI studies, our subject-specific time series data is accompanied by additional clinical data. The incorporation of this additional information may provide further insight into behavior, but its use has been largely ignored.

- One important omission from these approaches is to explain *why* each individual has a particular idiographic structure. While the common network has a straightforward interpretation, the individual structures are not necessarily easy to interpret or process. Therefore, the output of GIMME or MultiVAR can only measure the existence

of this idiographic structure, without explaining in any way the origin or motivation of potential underlying patterns.

- In the current writeup, we propose a method for estimating the nomothetic and idiographic structure of multi-subject time series, where the time dependence of each individual depends on a set of underlying covariates. This way, the idiographic effects are not simply unexplained individual structure, but are instead linked to the patients' individual characteristics. This allows for an actual interpretation of the idiographic structure.

- **Main advantages**: while other methods allow for modeling nomothetic and idiographic structure, the motivation for why such variations occur remains largely unknown. The explanation is simply: there is variation across individuals. This method connects the individual variation to underlying features, allowing us to determine the nature of the nomothetic structure, and to provide potential explanations for the origin of the idiographic structure. This can then be used as useful knowledge for further research or downstream statistical analysis.

2 Description of the Model

- Consider we have a study with N subjects. For each subject $1 \leq k \leq N$, we have access to two data modes: (i) a p-dimensional vector of covariates $\boldsymbol{Y}_k \in \mathbb{R}^p$, and (ii) a d-dimensional time series of length T_k, given as $\{\boldsymbol{X}_t^{(k)} \in \mathbb{R}^d \,:\, 1 \leq t \leq T_k\}$.

- In the usual GIMME or MultiVAR frameworks [4; 3], the aim is to model the time-series with a common+individual decomposition, usually of the form:

$$\boldsymbol{X}_t^{(k)} = \sum_{\ell=1}^{q} (\Psi_\ell^c + \Psi_\ell^{(k)}) \boldsymbol{X}_{t-\ell}^{(k)} + \boldsymbol{e}_t^{(k)}. \qquad \text{(MultiVAR);}$$

$$\boldsymbol{X}_t^{(k)} = (\mathrm{A}^c + \mathrm{A}^{(k)}) \boldsymbol{X}_t^{(k)} + \sum_{\ell=1}^{q} (\Phi_\ell^c + \Phi_\ell^{(k)}) \boldsymbol{X}_{t-\ell}^{(k)} + \varepsilon_t^{(k)}. \qquad \text{(GIMME).}$$

- While these approaches are useful at extracting common and individual structure, they miss the opportunity to exploit the potentially useful additional information contained in $\{\boldsymbol{Y}_k\}_{k=1}^N$.

- To exploit it, we assume that the time-series model coefficients depend on the values of Y. We consider the following models:

$$\boldsymbol{X}_t^{(k)} = \sum_{\ell=1}^{q} \Psi_\ell(\boldsymbol{Y}_k)\boldsymbol{X}_{t-\ell}^{(k)} + \boldsymbol{e}_t^{(k)}. \qquad \text{(MultiVAR);} \qquad (1)$$

$$\boldsymbol{X}_t^{(k)} = \mathrm{A}(\boldsymbol{Y}_k)\boldsymbol{X}_t^{(k)} + \sum_{\ell=1}^{q} \Phi_\ell(\boldsymbol{Y}_k)\boldsymbol{X}_{t-\ell}^{(k)} + \varepsilon_t^{(k)}. \qquad \text{(GIMME).} \qquad (2)$$

  Notice that, instead of being fixed, the set of coefficients $\Psi_\ell(Y_k), \mathrm{A}(Y_k), \Phi_\ell(Y_k)$ now depend on the individual covariate data. From this, the individual structure of each cannot be arbitrary, but it now depends on the specific features Y_k.

- As a first approach, lets assume that these parameters have a *additive* relationship to the underlying covariates. For a vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_p) \in \mathbb{R}^p$

$$\begin{aligned}
\Psi_\ell(\boldsymbol{y}) &= \Psi_{\ell 0} + y_1\Psi_{\ell 1} + y_2\Psi_{\ell 2} + \ldots + y_p\Psi_{\ell p};\\
\mathrm{A}(\boldsymbol{y}) &= \mathrm{A}_0 + y_1\mathrm{A}_1 + y_2\mathrm{A}_2 + \ldots + y_p\mathrm{A}_p;\\
\Phi_\ell(\boldsymbol{y}) &= \Phi_{\ell 0} + y_1\Phi_{\ell 1} + y_2\Phi_{\ell 2} + \ldots + y_p\Phi_{\ell p}.
\end{aligned}$$

  Here, the matrices $\{\Psi_{\ell 0}\}_{\ell=1}^q, \mathrm{A}_0, \{\Phi_{\ell 0}\}_{\ell=1}^q$ represent the structure that is common (nomothetic) among all subjects. The remaining parameters are affected by y, so they are dependent on the individual characteristics of each subject.

- **Example:** We consider a VAR model, where there is a single lagged relationship, *i.e.* $q = 1$, and two subject-level covariates $p = 2$. In that case, given that $\boldsymbol{Y}_k = (Y_{k1}, Y_{k2})^\top$ our equations simplify to:

$$\begin{aligned}
\boldsymbol{X}_t^{(k)} &= \Psi(\boldsymbol{Y}_k)X_{t-1}^{(k)} + \boldsymbol{e}_t^{(k)}\\
&= [\Psi_0 + Y_{k1}\Psi_1 + Y_{k2}\Psi_2] \cdot \boldsymbol{X}_{t-1}^{(k)} + \boldsymbol{e}_t^{(k)}
\end{aligned}$$

  Notice that each subject k will have their own VAR structure. This structure will incorporate a common structure shared across all subjects $\Psi_0 \in \mathbb{R}^{d\times d}$. It also has individual structure. Now, instead of having this individual structure vary arbitrarily, it has the form $Y_{k1}\Psi_1 + Y_{k2}\Psi_2$. Therefore, it depends on known information $Y_k = (Y_{k1}, Y_{k2})'$ about each subject.

- In Figure 2, we provide an example of how common and individual effects in a VAR time series model can relate to underlying subject-level covariates. Here, the dimen-

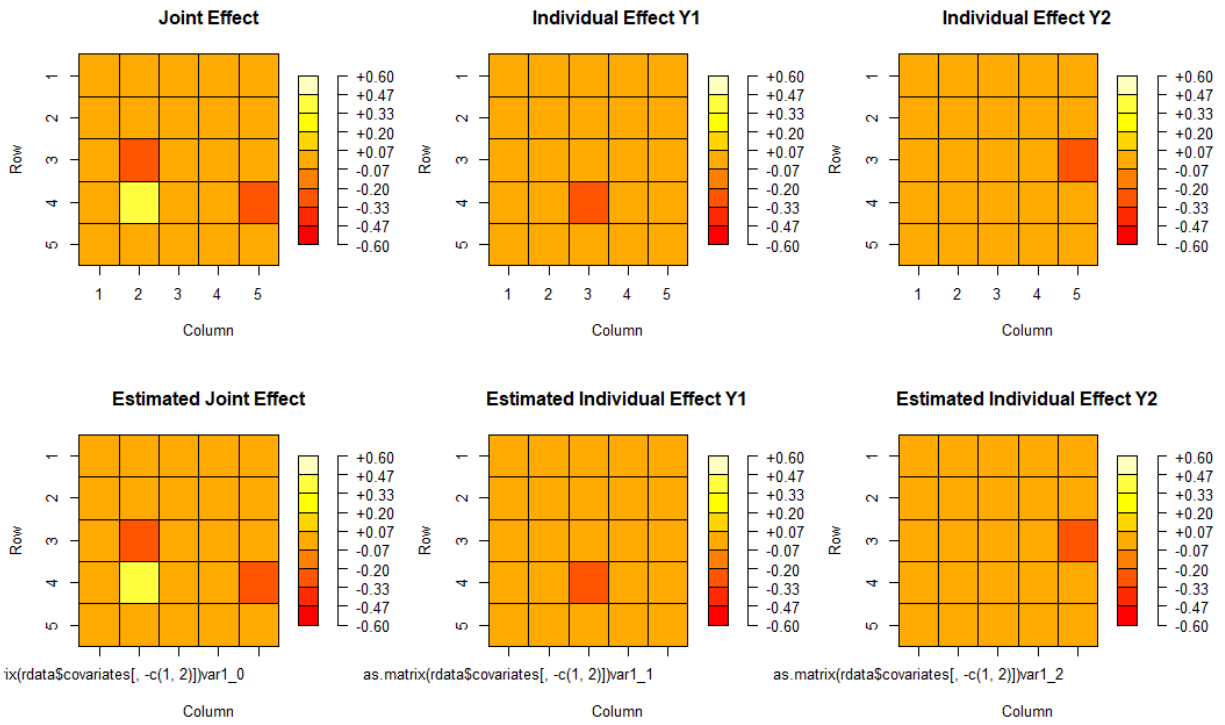

Figure 1: Illustration of the true and estimated Regression VAR parameters. For very large sample sizes, the traditional OLS method is successful. It would now be interesting to explore how regularization or model selection can help in higher dimension.

sion of our time series is $d = 5$, the number of subject-level covariates is $p = 2$, and the lag in the model is $q = 1$. We generate data for $N = 50$ subjects, all with a total of $T = 100$ time points.

- For this same example, we performed a traditional LS fitting, to see if we could successfully recover the common/individual structure of the time series. As we see in Figure 1, our estimation for very large sample sizes is successful.

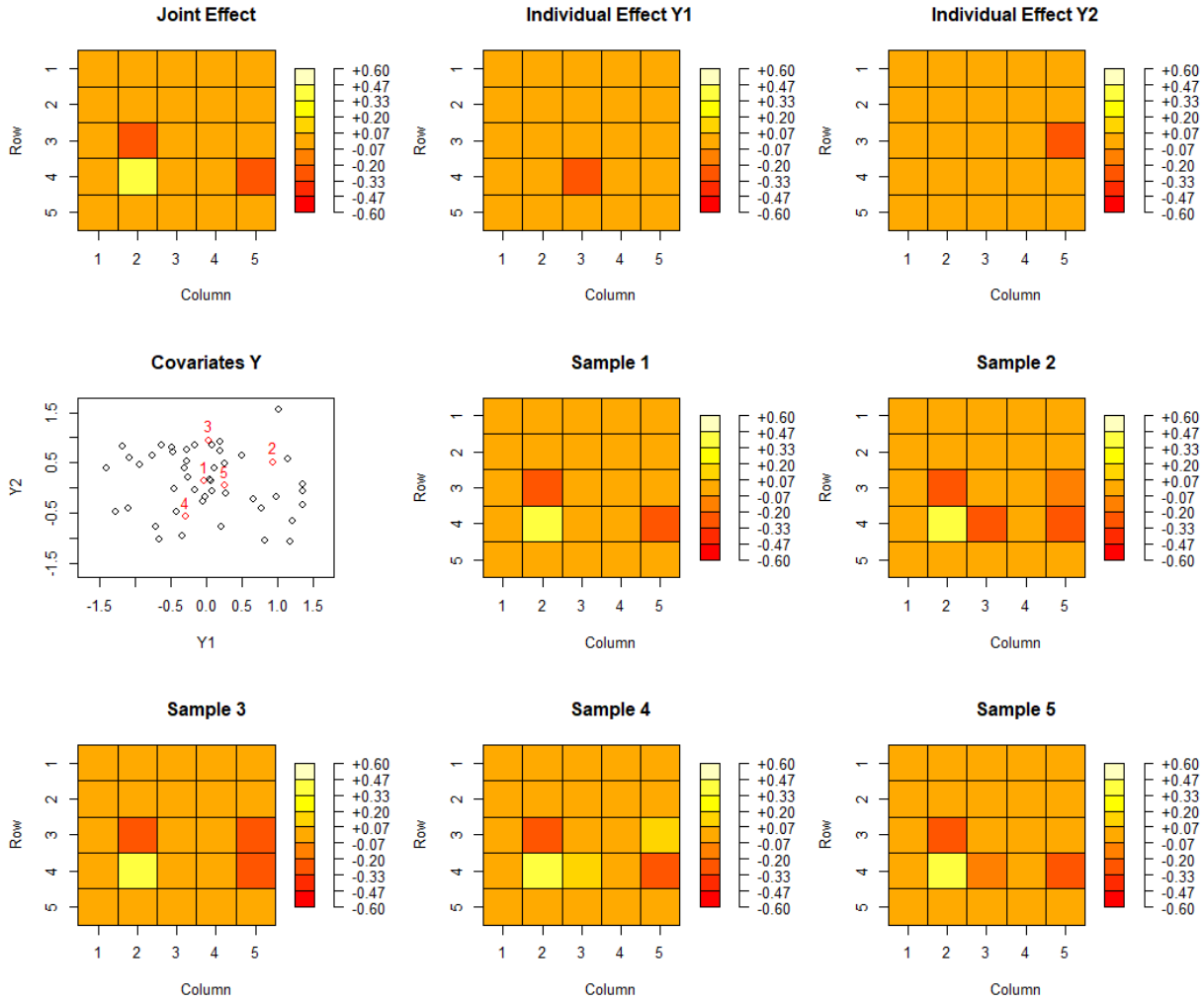

Figure 2: Illustration of the Regression VAR framework with 2 covariate-dependent effects. Top-panels: visualization of the joint effects that are common across all subjects. We also visualize the effects that are individual to variables Y_1 and Y_2. Center-left panel: Visualization of the covariate information for $N = 50$ subjects. We highlight the covariate information of the first 5 subjects. Remaining panels: Visualization of the effects for the first 5 subjects. We observe that, as subject 1 has nearly zero for both Y_1 and Y_2, the effects are mostly just the joint effects. Subject 3 has a high Y_2 covariate, but low Y_1 covariate, and therefore its effects are a mixture of the joint and Y_2 effects. Subject 2 has large covariates for both Y_1 and Y_2, so its time series effects are a mixture of all.

3 Fitting the RVAR Model with LASSO Penalty

In this section, we describe the fitting of the RVAR model of the form (1), with 1-lag relationships, *i.e.* $q = 1$. Assuming that, for each individual $k = 1, ..., N$, we can capture the VAR relationships in matrix form as:

$$\underbrace{\begin{bmatrix} (X^k_{T_k})' \\ (X^k_{T_k-1})' \\ \vdots \\ (X^k_2)' \end{bmatrix}}_{\mathbf{Z}^k} = \underbrace{\begin{bmatrix} (X^k_{T_k-1})' & Y_{k1}(X^k_{T_k-1})' & Y_{k2}(X^k_{T_k-1})' & \dots & Y_{kp}(X^k_{T_k-1})' \\ (X^k_{T_k-2})' & Y_{k1}(X^k_{T_k-2})' & Y_{k2}(X^k_{T_k-2})' & \dots & Y_{kp}(X^k_{T_k-2})' \\ \vdots & \vdots & \ddots & \vdots & \\ (X^k_1)' & Y_{k1}(X^k_1)' & Y_{k2}(X^k_1)' & \dots & Y_{kp}(X^k_1)' \end{bmatrix}}_{\mathbf{W}^k} \underbrace{\begin{bmatrix} \Psi_0' \\ \Psi_1' \\ \vdots \\ \Psi_p' \end{bmatrix}}_{\Psi} + \underbrace{\begin{bmatrix} (\varepsilon^k_T)' \\ (\varepsilon^k_{T-1})' \\ \vdots \\ (\varepsilon^k_{T-q+1})' \end{bmatrix}}_{\mathbf{E}^k}.$$

This can be simplified to matrix form as

$$\mathbf{Z}^k = \mathbf{W}^k \mathbf{\Psi} + \mathbf{E}^k \tag{3}$$

The equation (3) represents the VAR relationships for individual k. Then, we can model the behavior for all individuals via the equation

$$\underbrace{\begin{bmatrix} \mathbf{Z}^1 \\ \mathbf{Z}^2 \\ \vdots \\ \mathbf{Z}^N \end{bmatrix}}_{\mathbf{Z}} = \underbrace{\begin{bmatrix} \mathbf{W}^1 \\ \mathbf{W}^2 \\ \vdots \\ \mathbf{W}^N \end{bmatrix}}_{\mathbf{W}} \mathbf{\Psi} + \underbrace{\begin{bmatrix} \mathbf{E}^1 \\ \mathbf{E}^2 \\ \vdots \\ \mathbf{E}^N \end{bmatrix}}_{\mathbf{E}}$$

We can finally summarize the VAR equation model for all individuals simultaneously with the reduced matrix equation:

$$\mathbf{Z} = \mathbf{W}\mathbf{\Psi} + \boldsymbol{E}. \tag{4}$$

To solve for the parameter $\mathbf{\Psi} \in \mathbb{R}^{d(p+1)\times d}$, we propose performing the following optimization,

$$\widehat{\mathbf{\Psi}} := \underset{\mathbf{\Psi}\in\mathbb{R}^{d(p+1)\times d}}{\operatorname{argmin}} \|\!|\mathbf{Z} - \mathbf{W}\mathbf{\Psi}|\!\|_F^2 + \lambda_1 \|\!|\Psi_0|\!\|_1 + \lambda_2 \|\!|[\Psi_1\ \Psi_2\ \dots\ \Psi_p]|\!\|_1, \tag{5}$$

where the hyperparameter λ_1 serves to penalize the common structure of all VAR models, and λ_2 penalizes the Y covariate dependent portion of the model. We can interpret (5) as a matrix regression problem. To solve the optimization problem (5), we use the `glmnet` package, which allows for solving matrix regression problems directly. The `glmnet` implementation solves the problem by reducing the matrix regression problem to solving the

regression problems of their individual columns.

$$\widehat{\mathbf{\Psi}}^{\ell}(\lambda_1, \lambda_2) = \underset{\mathbf{\Psi}^{(\ell)} \in \mathbb{R}^{d(p+1)}}{\operatorname{argmin}} \left\| \mathbf{Z}_{\cdot \ell} - \mathbf{W}\mathbf{\Psi}^{(\ell)} \right\|_F^2 + \lambda_1 \|[\mathbf{\Psi}^{(\ell)}]_{1:d}\|_1 + \lambda_2 \|[\mathbf{\Psi}^{(\ell)}]_{(d+1):(dp)}\|_1.$$

Finally, we set the estimated parameter $\widehat{\mathbf{\Psi}}(\lambda_1, \lambda_2) := \left[\widehat{\mathbf{\Psi}}^1 \, \widehat{\mathbf{\Psi}}^2 \, \ldots \, \widehat{\mathbf{\Psi}}^d\right](\lambda_1, \lambda_2) \in \mathbb{R}^{d(p+1) \times d}$. Figure 1 in the previous section was derived with $\lambda_1 = \lambda_2 = 0$.

The current implementation only allows for a single lagged relationship, *i.e.* $q = 1$. We aim to create a more general implementation with lag $q > 1$ in the future.

4 Fitting RUSEM

The implementation of the RVAR model described in Section 3 is simply a generalization of the optimization problem considered in the Fisher et al. [3]. Notice that the RVAR focuses the estimation on the lagged effects $\mathbf{\Psi}$. For the GIMME model, we are also interested in recovering contemporaneous effects. The GIMME model cannot easily be extended to the regression setting. Therefore, we need to explore a different procedure for recovering the USEM structure in the regression setting. This will require thought and research...

I explored the original GIMME reference Gates and Molenaar [4], which points towards the modification index described in Sörbom [7]. I am still trying to understand how the GIMME adapts the estimation of USEM described in Sörbom [7] to scenarios with time dependence. This will be the next task.

References

[1] Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models.

[2] Crawford, C. M., Park, J. J., Chow, S.-M., Ernst, A. F., Pipiras, V., and Fisher, Z. F. (2024). Penalized subgrouping of heterogeneous time series. *arXiv preprint arXiv:2409.03085*.

[3] Fisher, Z. F., Kim, Y., Fredrickson, B. L., and Pipiras, V. (2022). Penalized estimation and forecasting of multiple subject intensive longitudinal data. *psychometrika*, 87(2):403–431.

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

[5] Kim, J., Zhu, W., Chang, L., Bentler, P. M., and Ernst, T. (2007). Unified structural equation modeling approach for the analysis of multisubject, multivariate functional mri data. *Human brain mapping*, 28(2):85–93.

[6] Lane, S. T., Gates, K. M., Pike, H. K., Beltz, A. M., and Wright, A. G. (2019). Uncovering general, shared, and unique temporal patterns in ambulatory assessment data. *Psychological methods*, 24(1):54.

[7] Sörbom, D. (1989). Model modification. *Psychometrika*, 54(3):371–384.

[8] Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.