SUPPLEMENTARY MATERIAL OF THE PAPER "COVARIANCE PATTERN MIXTURE MODELS FOR THE ANALYSIS OF MULTIVARIATE HETEROGENEOUS LONGITUDINAL DATA"

By Laura Anderlucci and Cinzia Viroli

Department of Statistical Sciences, University of Bologna

1. Simulation Study 1 - Model Validation. The model validation is examined under two different perspectives:

1. the robustness of the estimation algorithm; i.e. given k, if there is a temporal structure, is the algorithm able to recognize it?

2. the quality of the clustering; i.e. if there is a temporal structure that we do not take into account would the clustering be worsened (i.e. the algorithm is less able to recover the true classification of the units)?

1.1. *Robustness of the estimation algorithm.* In order to address the first point, we conducted a simulation study whom scheme is a full factorial design of the following characteristics:

- a number k of mixture components equal to 3;
- a 4×4 unconstrained within covariance matrix Ω;
- a 6×6 between covariance matrix Φ with structures equal to $GAR(1)$ and $GAR(3)$;
- a sample size n equal to 50, 100 and 500.

For each pattern we generated 100 datasets and we fitted 5 different models, according to different structures for Φ: GAR(1), GAR(2), GAR(3), GAR(4) and GAR(5). The best model was chosen according to the BIC and the AIC criteria. Results are shown in Table 1.

The two information criteria look equivalent in terms of performance: in some cases the BIC returns a better model selection, in some other cases AIC outperforms the BIC. When the true model has a temporal covariance matrix Φ with a $GAR(1)$ structure the algorithm shows a good performance in recovering it. In fact, even when the sample size is small, i.e. equal to 50, the algorithm picked up the correct model 89 times over 100; this proportion increases with the sample size, reaching the 96% when n is equal to 100 and 100% when n equal to 500.

TABLE 1

Ability in recovering the 'true' temporal structure given k: number of times (over 100) that a model with a particular structure for Φ was chosen according to BIC and AIC criteria. Shaded columns highlight the true model structure.

Sample size	Model Selection Criteria	GAR(1)	GAR(2)	GAR(3)	GAR(4)	GAR(5)
True Model: GAR(1), k=3, p=4, T=6						
$n = 50$	BIC	89	4	2	3	2
	AIC	85	5	2	2	6
$n = 100$	BIC	96	1	2	0	1
	AIC	96	0	2	1	1
$n = 500$	BIC	100	0	0	0	0
	AIC	100	0	0	0	0
True Model: GAR(3), k=3, p=4, T=6						
$n = 50$	BIC	31	29	34	3	3
	AIC	31	16	45	4	4
$n = 100$	BIC	34	17	49	0	0
	AIC	33	15	50	2	0
$n = 500$	BIC	34	16	50	0	0
	AIC	34	16	48	1	1

TABLE 2
Ability in recovering the 'true' temporal structure given k: number of times (over 100) that a model with a particular structure for Φ was chosen according to BIC and AIC. Shaded columns highlight the true model structure.

Sample size	Model Selection Criteria	GAR(1)	GAR(2)	GAR(3)	GAR(4)	GAR(5)	GAR(6)	GAR(7)
$n = 50$	BIC	8	30	55	3	2	0	2
	AIC	5	5	76	7	3	2	2
$n = 100$	BIC	8	10	73	4	3	1	1
	AIC	4	4	74	10	5	2	1
$n = 500$	BIC	2	8	80	6	0	3	1
	AIC	2	8	79	6	1	3	1

The performance is not as good when the model that generated the data has a temporal covariance structure equal to $GAR(3)$. In fact, when the order of the autoregressive process is higher, the algorithm is less able to choose the true model: on average it does in about the 45% of the cases. This less robust result can be due to the fact that there has not been an increase in the available data as well (i.e. there are only 6 time points), leading to have the same amount of information to estimate a larger number of parameters.

In order to verify the latter assumption, we performed a simulation study by considering further 100 data sets (for each sample size, $n = 50, 100, 500$) that have an increased number of time points T equal to 8 but with an autoregressive order is still equal to 3.

Results are showed in Table 2.

As expected, by increasing the number of time points T, the performance of the model improves. In fact, the algorithm was able to recover the true model in the 75% of the cases on average; as the sample size n increases, the proportion increases too.

This last results allows us to conclude that the model is quite good in recovering the true temporal structure, provided that a fair number of time points have been observed. Performances improve with larger sample sizes.

1.2. *Quality of the clustering.* In order to evaluate the quality of the clustering (i.e. the ability of recovering the true classification of the units in case we do not take into account a temporal structure), we conducted another simulation study where we considered the following features:

- a number k of mixture components equal to 3;

TABLE 3

Ability in recovering the 'true' classification of the units: average values of MISC over the 100 dataset generated for each settings; values in brackets represent the corresponding standard error. Shaded columns highlight the true model structure.

Sample size	GAR(1)	GAR(2)	GAR(3)	GAR(4)	GAR(5)
Expected separated clusters					
$n = 50$	0.009 (0.004)	0.019 (0.006)	0.022 (0.007)	0.019 (0.006)	0.024 (0.006)
$n = 100$	0.005 (0.003)	0.006 (0.003)	0.012 (0.005)	0.005 (0.003)	0.006 (0.003)
$n = 200$	0.002 (0.000)	0.005 (0.003)	0.002 (0.001)	0.010 (0.005)	0.014 (0.006)
Expected overlapped clusters					
$n = 50$	0.120 (0.013)	0.148 (0.014)	0.161 (0.014)	0.168 (0.014)	0.164 (0.014)
$n = 100$	0.101 (0.014)	0.108 (0.014)	0.112 (0.015)	0.111 (0.015)	0.124 (0.015)
$n = 200$	0.073 (0.012)	0.101 (0.014)	0.101 (0.015)	0.095 (0.014)	0.096 (0.014)

- a 4×4 unconstrained within covariance matrix Ω;
- a 6×6 between covariance matrix Φ with structure equal to $GAR(1)$;
- a sample size n equal to 50, 100 and 200.
- expected separation of the mixture components: expected well separated (different mean parameters for the three components) vs. expected overlapped clusters (same mean parameters for the three components);

For each setting we generated 100 data sets and we estimated 5 different models, according to different structures for Φ: GAR(1), GAR(2), GAR(3), GAR(4) and GAR(5). For each data set and model we computed the Misclassification Error Rate (MISC). Table 3 contains the average values of MISC over the 100 data set generated for each settings. Values in brackets represent the corresponding standard error.

The quality of the classification yielded by the model is very good: the MISC is indeed low, even when clusters are not expected to be well separated. Furthermore these results tell us that, in terms of clustering, the algorithm is robust: deviations from the true temporal structures do not lead to a totally wrong classification.

However, the MISC for a $GAR(1)$ temporal covariance structure is lower than the one referring to a model with an unrestricted covariance structure (i.e. $GAR(5)$), which means that modelling the occasions' variance substantially improves the classification.

2. Simulation Study 2 - Accuracy. Accuracy of the estimated growth curves within each class is one of the main targets of inference. In order to measure it, we performed a simulation study by generating several data sets from different parameterizations and structures (i.e. 100 data sets for each setting). The model's parameter values were fixed according to a full factorial design that allowed for examining the impact of the following aspects:

- a number k of mixture components equal to 3;
- expected separation of the mixture components: expected overlapping clusters (same mean parameters for the three components) vs. expected non-overlapping clusters (different and fairly spaced mean parameters for the three components);
- a 4×4 unconstrained within covariance matrix Ω;
- a 6×6 between covariance matrix Φ with structure equal to $GAR(1)$ and to $GAR(3)$;
- a sample size n equal to 50, 100 and 500.

From the combination of all these specific features 12 different settings were obtained; on the so generated data the corresponding mixture model was estimated and a measure of accuracy γ for the covariance matrices is given by the following expressions:

$$(1) \quad \gamma_U = \frac{\sum_{i=1}^{k} \|\hat{U}_i - U\|}{k \cdot \phi} \quad \gamma_D = \frac{\sum_{i=1}^{k} \|\hat{T}_i - T\|}{k \cdot T} \quad \gamma_\Omega = \frac{\sum_{i=1}^{k} \|\hat{\Omega}_i - \Omega_i\|}{k \cdot p^2}$$

where

$$\phi = \frac{T(T-1)}{2} - \frac{(T-m-1)(T-m)}{2}.$$

The lower γ, the higher the accuracy of the estimates. Accuracy for the between covariance matrix is not directly measured on Φ since the latter is the product of matrices U and D (recall: $\Phi^{-1} = U^\top D^{-1} U$). Table 4 contains the average accuracy over the 100 replications of each setting and its corresponding standard error. Please note that the accuracy measures γs defined above take into account the number of parameters to be estimated for each matrix.

Results from Table 4 show that the parameter estimates are fairly accurate. Precision increases as the sample size n increases and as the expected

TABLE 4

Average accuracy values for the covariance matrices and their corresponding standard error over 100 replications of each setting.

Sample size	$GAR(1)$			$GAR(3)$		
	γ_U (s.e.)	γ_D (s.e.)	γ_Ω (s.e.)	γ_U (s.e.)	γ_D (s.e.)	γ_Ω (s.e.)
Expected overlapping clusters						
$n = 50$	0.156 (0.011)	0.189 (0.011)	0.027 (0.001)	0.178 (0.012)	0.282 (0.018)	0.034 (0.002)
$n = 100$	0.121 (0.009)	0.157 (0.012)	0.021 (0.001)	0.117 (0.015)	0.173 (0.013)	0.017 (0.001)
$n = 500$	0.081 (0.007)	0.111 (0.011)	0.012 (0.001)	0.071 (0.009)	0.113 (0.012)	0.009 (0.001)
Expected non-overlapping clusters						
$n = 50$	0.092 (0.003)	0.117 (0.004)	0.020 (0.001)	0.111 (0.008)	0.170 (0.008)	0.024 (0.002)
$n = 100$	0.062 (0.002)	0.079 (0.003)	0.014 (0.001)	0.065 (0.003)	0.105 (0.004)	0.015 (0.001)
$n = 500$	0.032 (0.003)	0.036 (0.003)	0.006 (0.000)	0.030 (0.001)	0.048 (0.002)	0.006 (0.000)

cluster separation becomes more evident. The accuracy is not worsened by considering a larger autoregressive order structure for Φ, i.e. results for a $GAR(1)$ and a $GAR(3)$ look equivalent.

3. Simulation Study 3 - Accuracy of the regression coefficients and of the covariance matrices when the model is misspecified. Data were generated according to the following features:

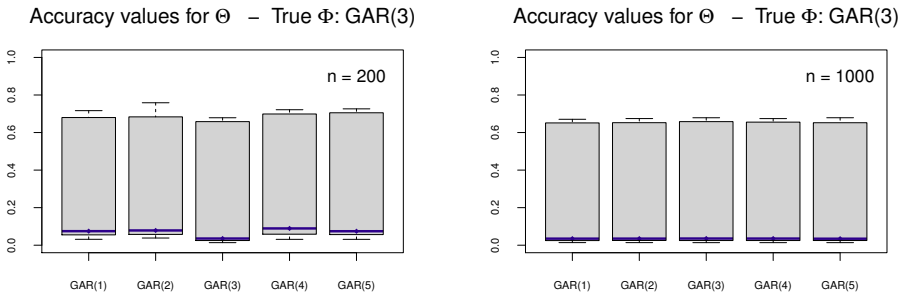
- $k = 3$ components;
- $p = 3$ responses;
- $T = 6$ time points;
- two different sample sizes: $n = 200$ and $n = 1000$;
- $q = 2$ covariates, randomly generated according to a Normal distribution (with zero mean and standard deviation equal to 2);
- the intercept terms were fixed equal to one, while the regression coefficients were randomly generated from a Normal distribution, with parameters set different according to the mixture components;
- unrestricted within covariance matrices Ωs (i.e. VVV);
- between covariance matrices Φ with temporal structure equal to $GAR(3)$.

For each setting 100 data sets were generated. The objective was to evaluate how the estimates change when the estimated model is misspecified in terms of temporal structure for Φ. In order to investigate this aspect, on the generated data 5 different models were estimated, according to different orders of the autoregressive process: from m equal to 1 to m equal to 5.

In order to measure the changes, the inaccuracy γ for each estimated parameter was estimated (see equations 1). Figure 1 shows the boxplot of the accuracy of Θ entries across 100 replications, according to the different temporal specification of matrix Φ . The plot shows clearly that a misspecification in the temporal structure does not really affect the growth estimates: accuracy values lie within a fairly small interval that is about the same for every fitted model.

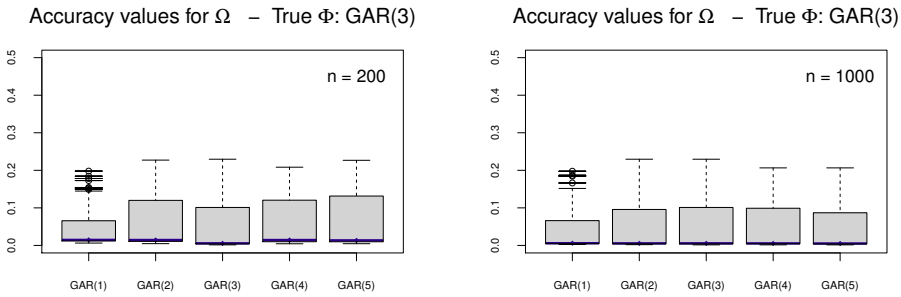
The distribution of the accuracy for Ω entries when the model is misspecified in terms of Φ is plot in Figure 2. Also in this case, there are no evident differences in the boxplots that correspond to different specification of Φ.

Figures 3, 4 show the distribution of the accuracy measured on the temporal components across the different fitted models. Again the model proofs to provide good estimates of the unknown parameters. However, this is the only case where the misspecification of Φ has a higher, even if limited, impact on the accuracy. In particular, the effect is more important if a much simpler structure is chosen.



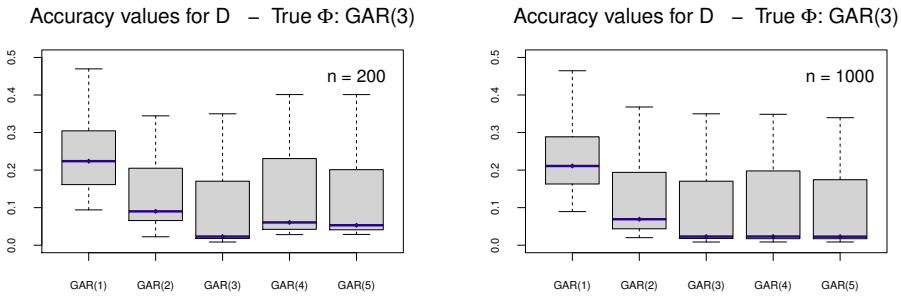
(a) Sample size $n = 200$ (b) Sample size $n = 1000$

Fig 1: Accuracy of matrix Θ entries according to different temporal specification for matrix Φ



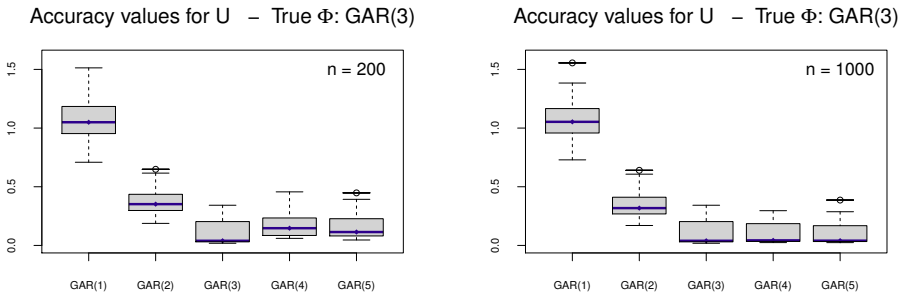
(a) Sample size $n = 200$ (b) Sample size $n = 1000$

Fig 2: Accuracy of matrix Ω entries according to different temporal specification for matrix Φ



(a) Sample size $n = 200$ (b) Sample size $n = 1000$

Fig 3: Accuracy of matrix D entries according to different temporal specification for matrix Φ



(a) Sample size $n = 200$ (b) Sample size $n = 1000$

Fig 4: Accuracy of matrix U entries according to different temporal specification for matrix Φ

4. Simulation Study 4 - Estimation of the number k of cluster when the model is misspecified. Another point that we investigated is how the model misspecification of matrix Φ impacts the estimate of the number of components.

Data were generated according to the following features:

- $k = 3$ components;
- $p = 3$ responses;
- $T = 6$ time points;
- a sample sizes n equal to 200;
- unrestricted within covariance matrices Ωs (i.e. VVV);
- unrestricted between covariance matrices Φs (i.e. GAR(5)) and a GAR(3) structure.

For each setting we generated 100 data sets. In order to investigate the choice of k when simpler model are specified we estimated model with smaller orders of the autoregressive process (i.e. from m=1 to $m = 5$) and we considered a different number of mixture components, namely $k = 2$, $k = 3$, $k = 4$.

Table 5 contains the number of times -over 100 replications- that the information criteria BIC and AIC preferred a particular model in terms of k over the others.

The right number of component was selected in the majority of cases, even when the temporal structure of data was misspecified. In the other cases, the heuristic parsimony of BIC and AIC makes the algorithm select a smaller number of cluster.

TABLE 5

Ability in recovering the 'true' number k of clusters according to different structures for Φ: number of times (over 100) that a model with a particular number k of mixture components and a specific structure for Φ was chosen according to BIC and AIC. Shaded columns highlight the true number of clusters.

matrix Φ		$k=2$	$k=3$	$k=4$
True structure for Φ: GAR(5)				
$GAR(1)$	BIC	14	85	1
	AIC	10	54	36
$GAR(2)$	BIC	31	69	0
	AIC	28	54	18
$GAR(3)$	BIC	42	58	0
	AIC	34	63	3
$GAR(4)$	BIC	45	54	0
	AIC	37	60	2
$GAR(5)$	BIC	48	51	0
	AIC	41	58	0
True structure for Φ: GAR(3)				
$GAR(1)$	BIC	16	82	3
	AIC	12	55	33
$GAR(2)$	BIC	35	65	0
	AIC	33	60	7
$GAR(3)$	BIC	45	55	0
	AIC	39	61	0
$GAR(4)$	BIC	44	55	0
	AIC	41	58	0
$GAR(5)$	BIC	44	55	0
	AIC	41	58	0