Bayesian Dynamic Tensor Regression Models

Monica Billio and Roberto Casarin and Matteo Iacopini

Abstract In this paper we introduce the literature on regression models with tensor variables and present a Bayesian linear model for inference, under the assumption of sparsity of the tensor coefficient. We exploit the CONDECOMP/PARAFAC (CP) representation for the tensor of coefficients in order to reduce the number of parameters and adopt a suitable hierarchical shrinkage prior for inducing sparsity. We propose a MCMC procedure via Gibbs sampler for carrying out the estimation, discussing the issues related to the initialisation of the vectors of parameters involved in the CP representation.

Key words: Tensor regression, Sparsity, Bayesian Inference, Hierarchical Shrinkage Prior

1 Bayesian Tensor Regression Model

Define a tensor as a generalisation of a matrix into a D-dimensional space, namely: $\mathscr{X} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$, where D is the order of the tensor and d_j is the length of dimension j. Matrices, vectors and scalars are particular cases of tensor variables, of order 2, 1 and 0, respectively. The common operations defined on matrices and vectors in linear algebra can be applied also to tensors via generalisations of their definition. For a remarkable survey on this subject, see [5].

The general tensor linear regression model (see [1], [2] for greater details) we present here can manage covariates and response variables in the form of vectors, matrices or tensors. It is given by:

Monica Billio
Department of Economics – Ca' Foscari University of Venice, Sestiere Cannaregio 872 – 30125 Venice (Italy), e-mail: billio@unive.it

Roberto Casarin
Department of Economics – Ca' Foscari University of Venice, Sestiere Cannaregio 872 – 30125 Venice (Italy), e-mail: r.casarin@unive.it

Matteo Iacopini
Department of Economics – Ca' Foscari University of Venice, Sestiere Cannaregio 872 – 30125 Venice (Italy) and Université Paris 1 - Panthéon-Sorbonne, 106-112 Boulevard de l'Hôpital – 75013 Paris (France), e-mail: matteo.iacopini@unive.it

$$\mathscr{Y}_t = \mathscr{A} + \mathscr{B} \times_{D+1} \text{vec}(\mathscr{X}_t) + \mathscr{C} \times_{D+1} \mathbf{z}_t + \mathscr{D} \times_n W_t + \mathscr{E}_t, \quad \mathscr{E}_t \overset{iid}{\sim} \mathscr{N}_{d_1,\ldots,d_D}(0, \Sigma_1, \ldots, \Sigma_D) \tag{1}$$

where the tensor response and errors are given by $\mathscr{Y}_t$, $\mathscr{E}_t \in \mathbb{R}^{d_1 \times \ldots \times d_D}$; while the covariates are $\mathscr{X}_t \in \mathbb{R}^{d_1^X \times \ldots \times d_M^X}$, $W_t \in \mathbb{R}^{d_n \times d_2^W}$ and $\mathbf{z}_t \in \mathbb{R}^{d_z}$. The coefficients are: $\mathscr{A} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$, $\mathscr{B} \in \mathbb{R}^{d_1 \times \ldots \times d_D \times p}$, $\mathscr{C} \in \mathbb{R}^{d_1 \times \ldots \times d_D \times d_z}$, $\mathscr{D} \in \mathbb{R}^{d_1 \times \ldots \times d_{n-1} \times d_2^W \times d_{n+1} \ldots \times d_D}$ where $p = \prod_i d_i^X$. The symbol $\times_n$ stands for the mode-n product between a tensor and a vector, as defined in [5]. This model extends several well-known econometric linear models, among which univariate and multivariate regression, VAR, SUR and Panel VAR models and matrix regression model (see [1] for formal proofs).

We focus on the particular case where both the regressor and the response variables are square matrices of size $k \times k$ and the error term is assumed to be distributed according to a matrix normal distribution:

$$Y_t = \mathscr{B} \times_3 \text{vec}(X_t) + E_t \qquad E_t \overset{iid}{\sim} \mathscr{N}_{k,k}(\mathbf{0}, \Sigma_c, \Sigma_r). \tag{2}$$

To significantly reduce the number of parameters we assume a CONDECOM/PARAFAC (CP) representation (more details in [5]) for the tensor. Let the vectors $\beta_j^{(r)} \in \mathbb{R}^{d_j}$, $j = 1, \ldots, D$, also called marginals of the CP representation, and R be the CP-rank of the tensor (assumed known and constant), then:

$$\mathscr{B} = \sum_{r=1}^{R} \mathscr{B}^{(r)} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \ldots \circ \beta_D^{(r)}, \tag{3}$$

2 Bayesian Inference

We follow the Bayesian approach for inference, thus we need to specify a prior distribution for all the parameters of the model. The adoption of the CP representation for the tensor of coefficients is crucial from this point of view, as it allows to reduce the problem of specifying a prior distribution on a multi-dimensional tensor, for which very few possibilities are available in the literature, to the standard multivariate case. Building from [3], we define a prior for each of the CP marginals of the tensor coefficient $\mathscr{B}$ by means of the following hierarchy:

$$\pi(\beta_j^{(r)} | \mathbf{W}, \phi, \tau) \sim \mathscr{N}_{d_j}(\mathbf{0}, \tau \phi_r W_{j,r}) \qquad \forall r \;\; \forall j \tag{4}$$

$$\pi(w_{p,j,r}) \sim \mathscr{E}xp(\lambda_{j,r}^2/2) \quad \forall r \;\; \forall j \;\; \forall p \tag{5}$$

$$\pi(\tau) \sim \mathscr{G}a(a_\tau, b_\tau) \quad \pi(\phi) \sim \mathscr{D}ir(\alpha_\phi) \quad \pi(\lambda_l) \sim \mathscr{G}a(a_\lambda, b_\lambda) \quad \forall j \;\; \forall r. \tag{6}$$

We complete the prior specification by assuming a hierarchical prior for the covariance matrices:

$$\Sigma_1 | \gamma \sim \mathscr{IW}(\gamma \Psi_1, \nu_1) \quad \Sigma_2 | \gamma \sim \mathscr{IW}(\gamma \Psi_2, \nu_2) \quad \gamma \sim \mathscr{G}a(a_\gamma, b_\gamma). \tag{7}$$

Given a sample $(\mathbf{Y}, \mathbf{X}) = \{Y_t, X_t\}_{t=1}^T$ and defining $\mathbf{x}_t := \text{vec}(X_t)$, the likelihood function of the model (2) is given by (see [1] for details of the Gibbs sampler):

$$L\left(\mathbf{Y},\mathbf{X}|\theta\right)=\prod_{t=1}^{T}(2\pi)^{-\frac{k^2}{2}}\left|\Sigma_2\right|^{-\frac{k}{2}}\left|\Sigma_1\right|^{-\frac{k}{2}}\exp\left\{-\frac{1}{2}\Sigma_c^{-1}(Y_t-\mathscr{B}\times_3\mathbf{x}_t)'\Sigma_1^{-1}(Y_t-\mathscr{B}\times_3\mathbf{x}_t)\right\}. \tag{8}$$

3 Simulation and Application

We performed a stimulation study by drawing a sample of $T=60$ couples $\{\mathbf{Y}_t,\mathbf{X}_t\}_t$ of square matrices of size varying from 10 to 50. The regressor is built by entry-wise independent Gaussian AR(1) processes with unitary noise variance. We initialised the marginals of the tensor $\mathscr{B}$ by simulated annealing and run the Gibbs sampler for $N=3000$ iterations. Fig. 1 shows the estimated coefficient tensor against the true (matricized form), for sizes 10 to 50. The plots suggest good performance of the proposed sampler in recovering the true value of the parameter, with slight tendence to overshrink, as common for local-global priors. See [1] for more details.

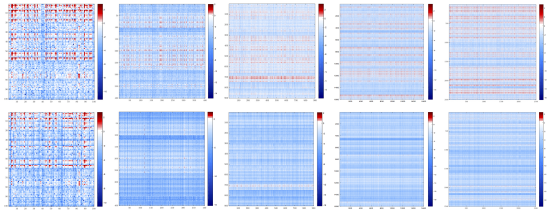

Fig. 1 Logarithm of the absolute value of the coefficient tensors: true $\mathscr{B}$ (*top*) and estimated $\hat{\mathscr{B}}$ (*bottom*). Sizes 10×10 (*left*) to 50×50 (*right*).

Denote $\text{vecr}(\cdot)$ the reverse vectorization operator and $\tilde{E}$ a binary matrix of unit shocks. We study the effects of the propagation of a shock via the matrix-valued impulse response function obtained as:

$$Y_h=\text{vecr}\left([\mathbf{B}'_{(3)}]^h\cdot\text{vec}\left(\tilde{E}\right)\right). \tag{9}$$

We apply the model to the study of $T=13$ yearly international trade networks of size 10×10, with $X_t=Y_{t-1}$. The results are in Fig. 2. The estimated coefficient tensor (10^4 entries) is rather sparse, with some regular patterns indicating that the trade flows depend on their past and on the "neighbouring" countries. The covariance matrices indicate higher values of the variances with respect to cross-sectional covariances. Fig. 3 shows the impulse response functions obtained by shocking the 10 most and 10 least relevant edges, respectively, and show significant level of propagation in space and persistence in time of the shock, with different magnitudes in the two cases.

4 Conclusions

We propose a matrix linear regression model (a reduced form of a tensor regression) which extends standard econometric models and allows each entry of the covariate to exert a different effect on each entry of the response. We exploit the PARAFAC decomposition to reduce the parameter space and follow a Bayesian approach to

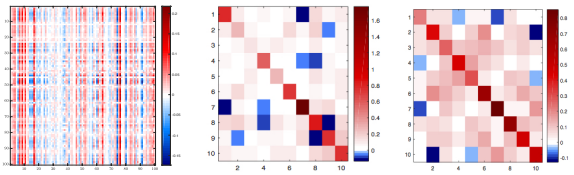

Fig. 2 Mode-3 matricization of the estimated coefficient tensor (*left*); estimated error covariance matrices: Σ_1 (*left*) and Σ_2 (*right*).

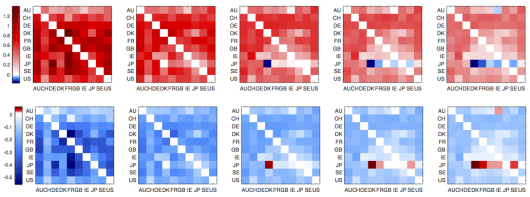

Fig. 3 Impulse response for $h = 1,\dots,5$ periods. Unitary shock on the 10 most (*top row*) and least (*bottom row*) relevant edges (sum of absolute values of all coefficients). Countries' labels on axes.

inference by means of a hierarchical local-global priors to allow efficient estimation of large sparse tensor coefficients. The accuracy of the model has been successfully tested both on synthetic and a real datasets, allowing the efficient estimate of large coefficient tensor.

Acknowledgements This research has benefited from the use of the Scientific Computation System of Ca' Foscari University of Venice (SCSCF) for the computational for the implementation of the inferential procedure.

References

1. Billio, M., Casarin, R., Iacopini, M.: Bayesian Dynamic Tensor Regression, arXiv preprint arXiv:1709.09606, (2017)
2. Billio, M., Casarin, R., Iacopini, M.: Bayesian Tensor Regression Models, Proceedings SIS 2017, (2017)
3. Guhaniyogi, R., Qamar, S., Dunson, D. B.: Bayesian Tensor Regression, Journal of Machine Learning Research, **18**:79, 1–31 (2017)
4. Hoff, P. D.: Multilinear Tensor Regression for Longitudinal Relational Data, The Annals of Applied Statistics, **9**, 1169–1193 (2015)
5. Kolda, T. G., Bader, B. W.: Tensor Decompositions and Applications, SIAM Review, **51**, 455–500 (2009)
6. Wang, H, West, M.: Bayesian Analysis of Matrix Normal Graphical Models, Biometrika, **96**, 821–834 (2009)
7. Zhou, H., Li, L., Zhu, H.: Tensor Regression with Applications in Neuroimaging Data Analysis, Journal of the Americal Statistical Association, **108**, 540–552 (2013)